FACULTY OF SCIENCE

SCHOOL OF MATHEMATICS AND STATISTICS

MATH2901

HIGHER THEORY OF STATISTICS

Contents

Probability

Statistics is all about making decisions based on data in the presence of **uncertainty**. In order to deal with uncertainty we need to develop a language for discussing it – **probability theory**. We will also derive in this chapter some fundamental results from probability theory that will be useful to us later on.

References: Kroese and Chan (2013) chapter 1, Hogg *et al* (2005) sections 1.1-1.4, Rice (2007) chapter 1.

Experiment, Sample Space, Event

Probability theory is about modelling and analysing **random experiments**: experiments whose outcome cannot be determined in advance, but is nevertheless still subject to analysis.

Definition
An **experiment** is any process leading to recorded observations.

Examples of random experiments are:

1. Tossing a die.
2. Measuring the lifetime of a machine.
3. Counting the number of calls arriving at a telephone exchange during a fixed time period.
4. Selecting a random sample of fifty people and observing the number of left-handers.
5. Observing the size of a population over time.

Mathematically, we can model a random experiment by specifying an appropriate *probability space*, which consists of three components: a *sample space*, a set of *events* and a *probability*. We will now describe each of these objects.

Definitions
An **outcome** is a possible result of an experiment.
The set Ω of all possible outcomes is the **sample space** of an experiment.
Ω is **discrete** if it contains a countable (finite or countably infinite) number of outcomes.

Examples of random experiments with their sample spaces are:

1. Cast two dice consecutively,

$$\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,6)\}.$$

2. The lifetime of a machine (in days),

$$\Omega = \mathbb{R}_+ = \{\text{ positive real numbers }\}\,.$$

3. The number of arriving calls at an exchange during a specified time interval,

$$\Omega = \{0,1,\cdots\} = \mathbb{Z}_+\ .$$

Notice that for modelling purposes it is often easier to take the sample space larger than necessary. For example the actual lifetime of a machine would certainly not span the entire positive real axis.

Definitions
An **event** is a set of outcomes (a subset of Ω). An **event occurs** if the result of the experiment is one of the outcomes in that event.

Thus, an event is a collection of some possible outcomes of the experiment. Events will be denoted by capital letters $A, B, C, \ldots$.

Examples of events are:

1. The event that the sum of two dice is 10 or more,

$$A = \{(5,5),(5,6),(6,5),(6,6)\}.$$

2. The event that a machines lives less than 1000 days,

$$A = [0,1000)\ .$$

3. The event that out of fifty selected people, five are left-handed,

$$A = \{5\}\,.$$

Definition

Events are **mutually exclusive** (disjoint) if they have no outcomes in common; that is, if they cannot both occur.

If A and B are mutually exclusive, we can say that $A \cap B = \emptyset$, where $\emptyset$ denotes the empty set.

Example

Experiment - toss a coin 3 times, record results. Let H denote 'head', T denote 'tail'.

$$\Omega = \{\mathrm{HHH}, \mathrm{HHT}, \mathrm{HTH}, \mathrm{THH}, \mathrm{TTH}, \mathrm{THT}, \mathrm{HTT}, \mathrm{TTT}\}$$

Ω is discrete.

Let A be the event 'at least one head';

$$A = \{\mathrm{HHH}, \mathrm{HHT}, \mathrm{HTH}, \mathrm{THH}, \mathrm{TTH}, \mathrm{THT}, \mathrm{HTT}\}.$$

Let B be the event 'exactly two heads';

$$B = \{\mathrm{HHT}, \mathrm{HTH}, \mathrm{THH}\}.$$

Let C be the event 'at least two tails';

$$C = \{\mathrm{TTH}, \mathrm{THT}, \mathrm{HTT}, \mathrm{TTT}\}.$$

Which of A, B and C are mutually exclusive? Then $B \subset A$, so if B occurs then A must occur. B and C are mutually exclusive, so they cannot both occur.

Let's say HTT is the result of the experiment. Which of A, B and C have occurred?

..., then A and C have occurred, but not B.

Advanced material only for interested students

Having [1] defined an event and the set of all possible events Ω, the second ingredient in the model is to specify the collection $\mathcal{F}$, say, of all events "of interest". That is,

[1] This section is not compulsory/assessable and should be read only by interested students to deepen their understanding.

the collection of all events to which we wish to assign a "probability" (which we will define next). It is most tempting to take $\mathcal{F}$ equal to the collection of *all* subsets of Ω (the power set). When, Ω is *countable* this okay, and this is the reason why you will never encounter $\mathcal{F}$ in an elementary exposition of probability theory. However, when Ω is uncountable, the power set of Ω is in general so large that one *cannot* assign a proper "probability" to all subsets! The same difficulty arises when we wish to assign a natural "length" to all subsets of $\mathbb{R}$.

Thus, for an uncountable Ω we have to settle for a smaller collection $\mathcal{F}$ of events. This collection should have nice properties. For example,

1. With A and B events, the set $A \cup B$ (A union B) should also be an event, namely the event that A *or* B *or* both occur.

2. With A and B events, the set $A \cap B$ (A intersection B) should also be an event, namely the event that A *and* B both occur.

3. With A and event, the event A^c (A complement) should also be an event, namely the event that A does *not* occur.

4. The set Ω itself should be an event, namely the "certain" event. Similarly $\emptyset$ should be an event, namely the "impossible" event.

Thinking this over, the minimal assumption that we impose on $\mathcal{F}$ is that it should be an object called a σ-algebra:

Definition 0.1 A **σ-algebra** $\mathcal{F}$ on Ω is a collection of subsets of Ω that satisfies

1. $\Omega \in \mathcal{F}$,

2. If $A \in \mathcal{F}$ then also $A^c \in \mathcal{F}$,

3. If $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_n A_n \in \mathcal{F}$.

<u>**Exercise 1**</u> Prove that if $\mathcal{F}$ is a σ-algebra then, with $A_1, A_2, \ldots \in \mathcal{F}$ also $\bigcap_n A_n \in \mathcal{F}$.

Events $A_1, A_2, \ldots$ are called **exhaustive** if their union is the whole sample space Ω. A sequence $A_1, A_2, \ldots$ of disjoint and exhaustive events is called a **partition** of Ω.

<u>**Exercise 2**</u> Let A, B, C be a partition of Ω. Describe the *smallest* σ-algebra containing the sets A, B and C.

<u>**Exercise 3**</u> Let Ω be a sample space with n elements. If $\mathcal{F}$ is the collection of all subsets of Ω, how many sets does $\mathcal{F}$ contain?

Borel σ-algebra

The most important example of a non-trivial σ-algebra is the **Borel** σ-algebra on $\mathbb{R}$, denoted $\mathcal{B}$.

This is defined as the smallest σ-algebra on $\mathbb{R}$ that contains all the intervals of the form $(\infty, x]$, for $x \in \mathbb{R}$. We say that $\mathcal{B}$ is *generated* by the collection of intervals $(\infty, x]$. This σ-algebra of sets is big enough to contain all important sets, and small enough to allow us to assign a natural "length measure" to all sets. This is called the **Lebesgue measure**, often denoted by Leb or m or λ. The proof of the existence and uniqueness of the Lebesgue measure is one of the great achievements of measure theory.

For $\mathbb{R}^n$ we can do something similar. The smallest σ-algebra on $\mathbb{R}^n$ that contains all the "rectangles" of the form

$$(-\infty, x_1] \times \cdots \times (-\infty, x_n],$$

with $(x_1, \ldots, x_n) \in \mathbb{R}^n$ is called the Borel-σ-algebra on $\mathbb{R}^n$; we write $\mathcal{B}^n$. The corresponding natural "volume" measure is again called the Lebesgue measure. For example, the Lebesgue measure of the unit disc is π.

Exercise 4 Show that because $(-\infty, x]$, $x \in \mathbb{R}$ are elements of $\mathcal{B}$ also the sets of the form $(a, b]$, (a, b) and $\{a\}$ are in $\mathcal{B}$. What other sets can you think of?

Exercise 5 What is the Lebesgue measure of $\mathbb{Q}$, the set of rational numbers?
Exercise 6 Consider the set that is constructed in the following way. We start with the interval $[0, 1)$ divide it into three parts $[0, \frac{1}{3})$, $[\frac{1}{3}, \frac{2}{3})$ and $[\frac{2}{3}, 1)$, and take away the middle interval. Next, we divide the remaining intervals (each) into three equal parts, and remove the middle part, and continue this procedure recursively, ad infinitum. This set is called the *Cantor set.* Convince yourself that C has as many points as $\mathbb{R}$ and yet has Lebesgue measure 0.

Extended real line and Borel σ-algebra

Instead of working with the real line, it will be convenient to work with the **extended real line** $\bar{\mathbb{R}} := [-\infty, \infty]$. The natural extension of $\mathcal{B}$ is the σ-algebra $\bar{\mathcal{B}}$ which is generated by the intervals $[-\infty, x]$. It is called the Borel σ-algebra on $\bar{\mathbb{R}}$.

Similarly, the Borel σ-algebra $\bar{\mathcal{B}}^n$ on $\bar{\mathbb{R}}^n$ (the meaning should be obvious) is defined as the σ-algebra that is generated by the rectangles of the form

$$[-\infty, x_1] \times \cdots \times [-\infty, x_n] \, .$$

Axioms and Basic Results

Before we proceed you are encouraged to recall the definition of union, intersection, and complement notation given in Figure 1 below and explained in detail in Section 1.3 of the Kroese Chan, 2013 recommended textbook.

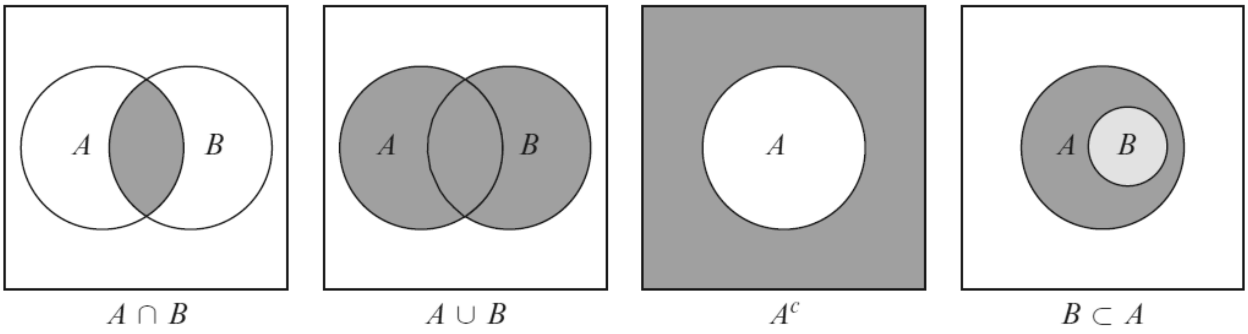


Figure 1: *Venn diagrams illustrating set operations. Each square represents the sample space* Ω

Given a sample space $(\Omega, \mathcal{F})$, a probability function $\mathbb{P}$ can be defined in the following way. To every event $A \in \mathcal{F}$ we assign a number $\mathbb{P}(A)$, the **probability that A occurs**. The function $\mathbb{P}$ must satisfy the axioms

(i) for each $A \subset \Omega$, $\mathbb{P}(A) \geqslant 0$

(ii) $\mathbb{P}(\Omega) = 1$

(iii) if $A_1, A_2, \dots$ are mutually exclusive (or disjoint)

$$(A_i \cap A_j = \emptyset \text{ for all } i, j \text{ with } i \neq j)$$

then $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$

Notation

Note that we will use the following interchangeable notation for the complement of event A

$$A^c \equiv \overline{A}$$

It follows that

1. if $A_1, A_2, \dots, A_k$ are mutually exclusive,

$$\mathbb{P}\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} \mathbb{P}(A_i),$$

2. $\mathbb{P}(\emptyset) = 0$,
3. for any $A \subseteq \Omega$, $0 \leq \mathbb{P}(A) \leq 1$ and $\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A)$.
4. if $B \subset A$, then $\mathbb{P}(B) \leq \mathbb{P}(A)$. Thus if B occurs $\Longrightarrow$ A occurs then $\mathbb{P}(B) \leq \mathbb{P}(A)$.

   Proof: Suppose $A \subset B$. Then, we can write B as $B = A \cup (A^c \cap B)$, where A and $A^c \cap B$ are disjoint events. Hence, according to the third and first axiom

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(A^c \cap B) \geqslant \mathbb{P}(A),$$

□

Rule for calculating probabilities: if Ω is discrete, $\mathbb{P}(A)$ = sum of probabilities for outcomes in A.

This follows from axiom (iii) since, for example in Figure 2, if $\Omega = \{s_1, s_2, \dots\}$ and $A = \{s_2, s_5, s_9, s_{11}, s_{14}\}$, then $A = \{s_2\} \cup \{s_5\} \cup \{s_9\} \cup \{s_{11}\} \cup \{s_{14}\}$, a mutually exclusive union, so

$$\mathbb{P}(A) = \mathbb{P}(\{s_2\} \cup \{s_5\} \cup \{s_9\} \cup \{s_{11}\} \cup \{s_{14}) = \mathbb{P}(\{s_2\}) + \mathbb{P}(\{s_5\}) + \mathbb{P}(\{s_9\}) + \mathbb{P}(\{s_{11}\}) + \mathbb{P}(\{s_{14}\}).$$

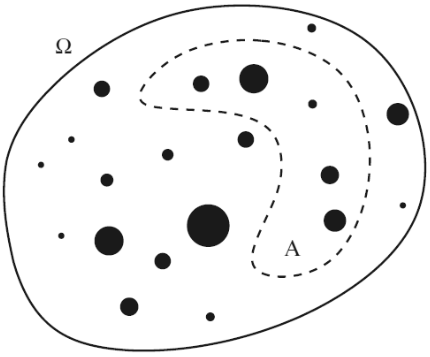


Figure 2: *Computation of probabilities in a discrete sample space Ω. Each blob represents the relative weight assigned to it by the probability measure $\mathbb{P}$.*

Thus, in the previous coin toss example, if we assume that the 8 possible outcomes are equally likely,

$\mathbb{P}(A) = \mathbb{P}(\text{at least one head}) = \frac{7}{8}$,
$\mathbb{P}(B) = \mathbb{P}(\text{exactly two heads}) = \frac{3}{8}$ and
$\mathbb{P}(C) = \mathbb{P}(\text{at least two tails}) = \frac{4}{8} = \frac{1}{2}$.

Monotonic sequences of events

Theorem 0.1 (Continuity property of $\mathbb{P}$) *If $A_1, A_2, \ldots$ is an increasing sequence of events, i.e., $A_1 \subset A_2 \subset \cdots$, then*

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right). \tag{1}$$

Property (1) is a kind of continuity property. We say that $\mathbb{P}$ is **continuous from below**.

Proof:

Suppose $A_1, A_2, \ldots$ is an increasing sequence of events. Define the event

$$A \stackrel{\text{def}}{=} \cup_n A_n = \cup_{n=1}^{\infty} A_n\ .$$

Now consider Figure 3.

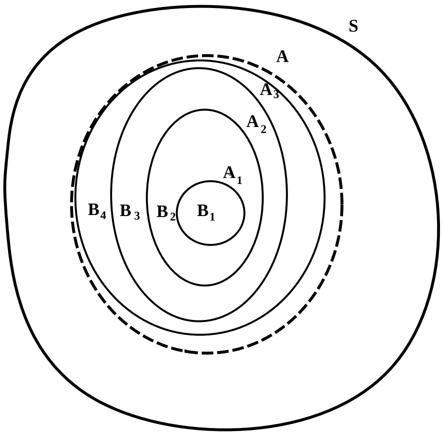


Figure 3: *Continuity from below, Note that here $S \equiv \Omega$*

Define the events $B_1, B_2, \ldots$ as

$$\begin{aligned} B_1 &= A_1 \\ B_2 &= A_2 \cap A_1^c \\ &\vdots \\ B_n &= A_n \cap A_{n-1}^c, \qquad n = 2, 3, \ldots \end{aligned}$$

The "rings" $B_1, B_2, \ldots$ are disjoint and

$$\bigcup_{i=1}^{n} B_i = \bigcup_{i=1}^{n} A_i = A_n \quad \text{and} \quad \bigcup_{i=1}^{\infty} B_i = A.$$

Hence, from axiom three it follows that

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) \stackrel{\text{axiom (iii)}}{=} \sum_{i=1}^{\infty} \mathbb{P}(B_i) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(B_i) = \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{i=1}^{n} B_i\right) = \lim_{n\to\infty} \mathbb{P}(A_n), \end{aligned}$$

which proves Property (1). □

In exactly the same way we have the following obverse result.

Theorem 0.2 (Continuity from above) *If $A_1, A_2, \ldots$ is an decreasing sequence of events, i.e., $A_1 \supseteq A_2 \supseteq \cdots$, then*

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right). \tag{2}$$

Counting Rules

Counting Rule 1
If there are k experiments with n_i possible outcomes in the ith ($i = 1, 2, \ldots, k$), then the total number of possible outcomes for the k experiments is $n_1 n_2 \ldots n_k = \Pi_{i=1}^{k} n_i$.

Example

Toss a 6-sided die 3 times.

What are k and n_i in this case?
What is the number of outcomes in the sample space?

An ordered arrangement of a set of distinct objects is a **permutation**.

Counting Rule 2
The number of possible permutations of r objects selected from n distinct objects is ${}^nP_r = \frac{n!}{(n-r)!}$, where $n! = n(n-1)(n-2)\ldots 3\cdot 2\cdot 1$ for integers $n \geq 1$ and $0! \equiv 1$.

Proof: ${}^nP_r = n(n-1)(n-2)\ldots(n-(r-1))$

$$\begin{aligned} &= n(n-1)(n-2)\ldots(n-r+1)\,\frac{(n-r)(n-r-1)\ldots 3\cdot 2\cdot 1}{(n-r)(n-r-1)\ldots 3\cdot 2\cdot 1} \\ &= \frac{n(n-1)(n-2)\ldots 3\cdot 2\cdot 1}{(n-r)(n-r-1)\ldots 3\cdot 2\cdot 1} = \frac{n!}{(n-r)!} \end{aligned}$$

Example

A particular committee has four members. One member must chair the committee, and a different committee member must take minutes from meetings.

How many different ways are there of choosing a Chair and a Minute-taker for this committee?

First label the four people as a, b, c, d.

The number of permutations of the letters a, b, c, d taken 2 at a time:

$$\begin{array}{ll} ab\ ac\ ad\ bc\ bd\ cd & n = 4,\ r = 2 \\ ba\ ca\ da\ cb\ db\ dc & {}^4P_2 = \tfrac{4!}{2!} = 12. \end{array}$$

Note: the number of possible permutations of r objects $(n = r)$ is ${}^rP_r = \frac{r!}{0!} = r!$

Counting Rule 3

The number of ways of choosing r objects from n distinct objects is

$$\frac{n!}{r!(n-r)!} \equiv \left(\begin{array}{c} n \\ r \end{array}\right) \quad (n \text{ choose } r),\ \ 0 \leq r \leq n.$$

Proof: the number of permutations (ordered arrangements) is ${}^nP_r$. But now we are ignoring order, so that a particular set of r objects, which contributes $r!$ permutations to the total, now counts as one selection and so the number of ways of choosing r from n is $\dfrac{{}^nP_r}{r!} = \dfrac{n!}{r!(n-r)!}$.

Example

From a committee of four people, two committee members will need to present the committee's recommendations to the board of directors.

How many ways are there of choosing two committee members to report to the board of directors?

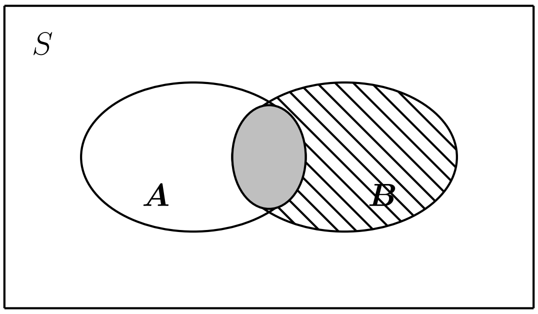
Choose 2 letters from a, b, c, d. We ignore order, so that ab and ba, etc. each count as one selection. The possibilities are: *ab ac ad bc bd cd*

$$n = 4,\ r = 2,\ \binom{4}{2} = \frac{4!}{2!2!} = 6.$$

Conditional Probability

Definition
The **conditional probability** that an event A occurs, given that an event B has occurred is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \text{ if } \mathbb{P}(B) \neq 0.$$



Given that B has occurred, the total probability for possible results of the experiment equals $\mathbb{P}(B)$, so that the probability that A occurs equals the total probability for outcomes in A (only those in $A \cap B$) divided by the total probability, $\mathbb{P}(B)$.

Lemma 1: $\mathbb{P}(A|B) = \mathbb{P}(A) \iff \mathbb{P}(B|A) = \mathbb{P}(B)$.

Proof: $\Longrightarrow$ if $\mathbb{P}(A|B) = \mathbb{P}(A)$, then

$$\begin{aligned} \mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \\ = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} &= \mathbb{P}(B). \end{aligned}$$

$\Longleftarrow$ interchange A and B in the above proof.

Independent Events

Events A and B are **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. For any two events A and B, $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$, so A and B are independent

$$\begin{aligned} &\iff \mathbb{P}(A|B) = \mathbb{P}(A) \\ &\iff \mathbb{P}(B|A) = \mathbb{P}(B) \quad \text{(lemma 1)} \end{aligned}$$

Example

Toss two fair dice. There are 36 outcomes in the sample space Ω, each with probability $\frac{1}{36}$. Let:

- A be the event '4 on first die'
- B be the event 'sum of numbers is 7'
- C be the event 'sum of numbers is 8'.

Intuition: A and B are independent, A and C are not independent. (B and C are not independent since

$$\mathbb{P}(B \cap C) = 0 \text{ and } \mathbb{P}(B) > 0,\ \mathbb{P}(C) > 0.)$$

We can write Ω as

$$\Omega = \{(1,1), (1,2), \ldots, (1,6), (2,1), \ldots, (2,6), \ldots, (6,6)\}$$

or display it as follows;

2nd die

$$\begin{array}{lr|cccccccc}
 & & 1 & 2 & 3 & 4 & 5 & 6 & & \\
 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \odot & & \\
 & 2 & \cdot & \cdot & \cdot & \cdot & \odot & \boxdot & & \\
\textbf{1st die} & 3 & \cdot & \cdot & \cdot & \odot & \boxdot & \cdot & A & \times \\
 & 4 & \times & \times & \otimes & \boxtimes & \times & \times & B & \odot \\
 & 5 & \cdot & \odot & \boxdot & \cdot & \cdot & \cdot & C & \square \\
 & 6 & \odot & \boxdot & \cdot & \cdot & \cdot & \cdot & &
\end{array}$$

Show that A and B are independent.
Show that A and C are dependent.

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

and $\frac{1}{36} = \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Thus A and B are independent.

$$\mathbb{P}(C) = \frac{5}{36}, \quad \mathbb{P}(A \cap C) = \frac{1}{36}, \quad \text{so}$$

$$\frac{1}{36} = \mathbb{P}(A \cap C) \neq \mathbb{P}(A)\mathbb{P}(C) = \frac{1}{6} \times \frac{5}{36}.$$

Thus A and C are not independent.

Also, $\mathbb{P}(A|C) = \dfrac{\mathbb{P}(A \cap C)}{\mathbb{P}(C)} = \dfrac{1/36}{5/36} = \dfrac{1}{5} \neq \mathbb{P}(A)$, again confirming A and C are not independent.

For a countable sequence of events $\{A_i\}$, the events are **pairwise independent** if

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j) \text{ for all } i \neq j$$

and the events are **(mutually) independent** if for any collection

$$A_{i_1}, A_{i_2}, \ldots, A_{i_n},$$

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_n}) = \mathbb{P}(A_{i_1}) \ldots \mathbb{P}(A_{i_n}).$$

Clearly, independence $\Longrightarrow$ pairwise independence, but not vice versa, as in the following example.

Example

A coin is tossed twice. Let A be the event 'head on the first toss', B the event 'head on the second toss' and C the event 'exactly one head turned up'.

A, B and C are pairwise independent (can you show this?), but

$$\mathbb{P}(A \cap B \cap C) = 0 \neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = (.5)^3,$$

so A, B and C are not independent.

Example

A ball is drawn at random from an urn containing 4 balls numbered 1,2,3,4. Let

$$\begin{aligned} A &= \{1,2\} \text{ (ball 1 or ball 2 is drawn)}, \\ B &= \{1,3\},\ C = \{1,4\}. \end{aligned}$$

Show that A, B, and C are pairwise independent but not independent.

Then A, B and C are pairwise independent (e.g. $\mathbb{P}(A \cap B) = \mathbb{P}(\{1\}) = \mathbb{P}(A)\mathbb{P}(B) = \frac{1}{4}$), but $\frac{1}{4} = \mathbb{P}(A \cap B \cap C) \neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$, so A, B and C are not independent.

Some Probability Laws

The Multiplicative Law

For events A_1, A_2

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_2 \cap A_1) = \mathbb{P}(A_2|A_1)\mathbb{P}(A_1).$$

For events A_1, A_2, A_3

$$\begin{aligned} \mathbb{P}(A_1 \cap A_2 \cap A_3) &= \mathbb{P}(A_3 \cap A_2 \cap A_1) \\ &= \mathbb{P}(A_3|A_2 \cap A_1)\mathbb{P}(A_2 \cap A_1) \\ &= \mathbb{P}(A_3|A_1 \cap A_2)\mathbb{P}(A_2|A_1)\mathbb{P}(A_1). \end{aligned}$$

The same pattern applies to higher numbers of events.

This law is particularly useful when we have a sequence of dependent trials.

Example

To gain entry to a selective high school students must pass 3 tests. 20% fail the first test and are excluded. Of the 80% who pass the first, 30% fail the second and are excluded. Of those who pass the second, 60% pass the third.

Question 1: What proportion of students pass the first two tests? Use the multiplicative law to answer this question.

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_2|A_1)\mathbb{P}(A_1) = 0.7 \times 0.8 = 0.56$$

Question 2: What proportion of students gain entry to the selective high school? Let A_1 be the event 'pass first test'. Similarly A_2, A_3. Then

$$\mathbb{P}(A_1) = 0.8\ ,\ \mathbb{P}(A_2|A_1) = 0.7\ ,\ \mathbb{P}(A_3|A_1 \cap A_2) = 0.6.$$

$\therefore\ \mathbb{P}(\text{student gains entry}) = \mathbb{P}(A_1 \cap A_2 \cap A_3)$

$$= \mathbb{P}(A_3|A_1 \cap A_2)\mathbb{P}(A_2|A_1)\mathbb{P}(A_1) = 0.6 \times 0.7 \times 0.8 = 0.336.$$

Question 3: What proportion pass the first two tests, but fail the third?

$$\begin{aligned}\mathbb{P}(A_1 \cap A_2 \cap \overline{A}_3) &= \mathbb{P}(\overline{A}_3|A_1 \cap A_2)\mathbb{P}(A_2|A_1)\mathbb{P}(A_1) \\ &= 0.4 \times 0.7 \times 0.8 = 0.224.\end{aligned}$$

$\lrcorner$

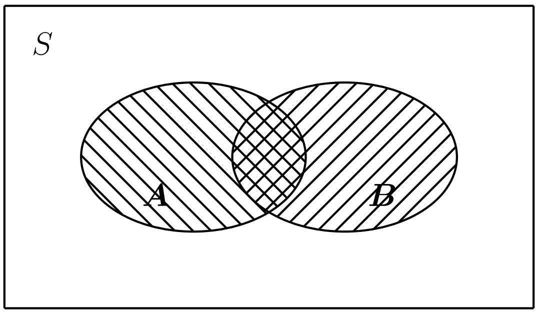
The Additive Law

For events A and B,

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

Proof: From the accompanying Venn diagram

$$A \cup B = A \cup (\bar{A} \cap B) \quad \text{and} \quad B = (A \cap B) \cup (\bar{A} \cap B).$$



Note that A and $\bar{A} \cap B$ are mutually exclusive, and that $A \cap B$ and $\bar{A} \cap B$ are mutually exclusive. So from the axioms

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(\bar{A} \cap B)$$

and

$$\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(\bar{A} \cap B).$$

The first of these equations gives $\mathbb{P}(\bar{A} \cap B) = \mathbb{P}(A \cup B) - \mathbb{P}(A)$. Substitution of this expression into the second equation leads to the required result.

This is one of the original axioms:

Corollary to Additive Law

For events A and B, if A and B are mutually exclusive

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$

Example

3 letters are placed at random into 3 addressed envelopes.

What is the probability that none is in the correct envelope?

Let A, B, C be the events that envelopes 1,2 and 3 contain the correct letters.

Then $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{3}$ and $\mathbb{P}(A \cap B) = \mathbb{P}(A \cap C) = \mathbb{P}(B \cap C) = \frac{1}{6}$ since, for example, $\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A) = \frac{1}{2} \times \frac{1}{3}$.

Also, $\mathbb{P}(A \cap B \cap C) = \frac{1}{6}$ since all 3 envelopes must contain the correct letters if any 2 envelopes contain the correct letters; that is,

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A \cap B) = \mathbb{P}(A \cap C) = \mathbb{P}(B \cap C).$$

Thus, $\mathbb{P}$(none is in the correct envelope)

$$\begin{aligned}
&= 1 - \mathbb{P}\text{ (at least one is in the correct envelope)} \\
&= 1 - \mathbb{P}(A \cup B \cup C) \\
&= 1 - \{\mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) \\
&\qquad -\mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C)\} \\
&= 1 - \{\frac{1}{3} + \frac{1}{3} + \frac{1}{3} - \frac{1}{6} - \frac{1}{6} - \frac{1}{6} + \frac{1}{6}\} = \frac{1}{3}.
\end{aligned}$$

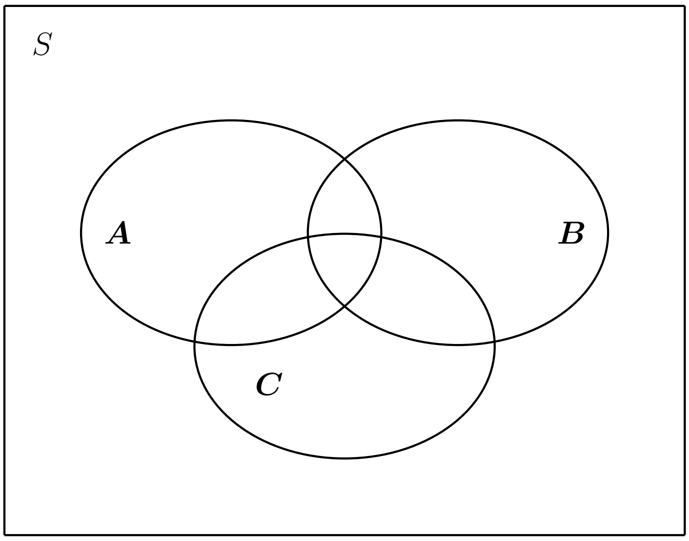


Figure 4: $A \cap B = A \cap C = B \cap C = A \cap B \cap C$

The Law of Total Probability

Suppose $A_1, A_2, \ldots, A_k$ are mutually exclusive ($A_i \cap A_j = \emptyset$ for all $i \neq j$) and **exhaustive** ($\bigcup_{i=1}^{k} A_i = \Omega$ =sample space) events; that is, $A_1, \ldots, A_k$ form a **partition** of Ω. Then, for any event B,

$$\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i).$$

Proof:

$B = \bigcup_{i=1}^{k}(B \cap A_i)$ (disjoint union since the A_i's are disjoint)

By axiom (iii) (in the finite case)

$$\begin{aligned}
\mathbb{P}(B) &= \mathbb{P}\left(\bigcup_{i=1}^{k}(B\cap A_i)\right) \\
&= \sum_{i=1}^{k}\mathbb{P}(B\cap A_i) \\
&= \sum_{i=1}^{k}\mathbb{P}(B|A_i)\mathbb{P}(A_i) \\
\text{since}\quad \mathbb{P}(B|A) &= \frac{\mathbb{P}(B\cap A)}{\mathbb{P}(A)}.
\end{aligned}$$

Example

Urn I contains 3 red and 4 white balls. Urn II contains 2 red balls and 4 white. A ball is drawn from Urn I and placed unseen into Urn II. A ball is now drawn at random from Urn II.

What is the probability that this second ball is red?

Let A_1 be the event '1st ball drawn red'
A_2 be the event '1st ball drawn white'
and let B be the event '2nd ball drawn red'.
A_1 and A_2 are mutually exclusive (they cannot both occur) and exhaustive (one of them must occur) and so

$$\begin{aligned}
\mathbb{P}(B) &= \mathbb{P}(B|A_1)\mathbb{P}(A_1)+\mathbb{P}(B|A_2)\mathbb{P}(A_2) \\
&= \frac{3}{7}\times\frac{3}{7}+\frac{2}{7}\times\frac{4}{7}=\frac{17}{49}.
\end{aligned}$$

Bayes' Formula

Bayes' Formula calculates conditional probabilities when the ordering of conditioning is reversed. In the simple two event situation Bayes' Formula is:

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

The most general version is:

Bayes' Formula

For a partition $A_1, A_2, \ldots, A_k$ and an event B,

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i=1}^{k}\mathbb{P}(B|A_i)\mathbb{P}(A_i)} = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\mathbb{P}(B)}$$

Example

Recall the previous example.

Given that the second ball drawn is red, what is the probability that the first ball was white?

$$\mathbb{P}(A_2|B) = \frac{\mathbb{P}(B|A_2)\mathbb{P}(A_2)}{\mathbb{P}(B)} = \frac{\frac{2}{7}\times\frac{4}{7}}{\frac{17}{49}} = \frac{8}{17}.$$

Example

A diagnostic test for a certain disease is claimed to be 90% accurate because, if a person has the disease, the test will show a positive result with probability 0.9 while if a person does not have the disease the test will show a negative result with probability 0.9. Only 1% of the population has the disease.

If a person is chosen at random from the population and tests positive for the disease, what is the probability that the person does in fact have the disease?

Let A be the event 'person has disease' and B be the event 'person tests positive'.

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

since A and $\overline{A}$ form a partition (they are mutually exclusive and exhaustive).

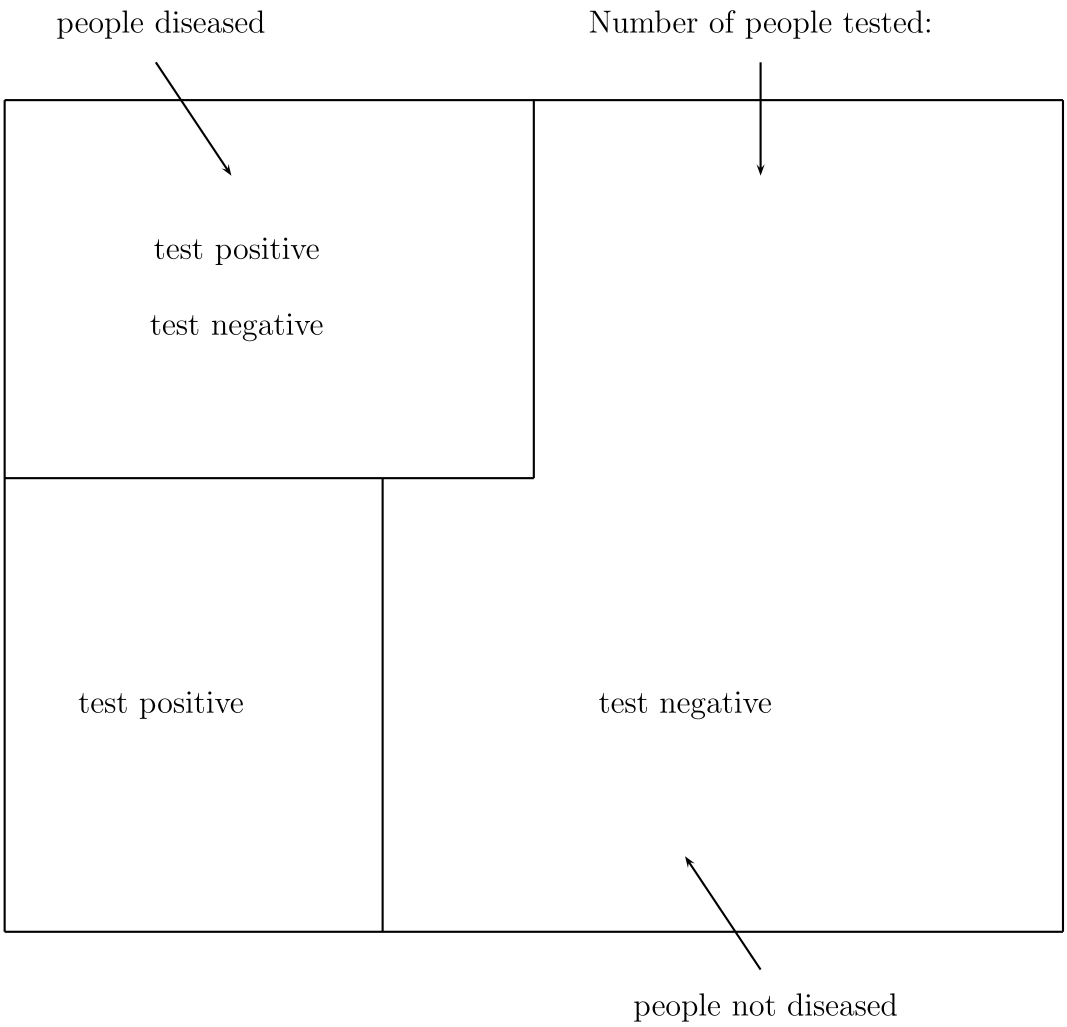
Now $\mathbb{P}(B|A) = 0.9$ $\qquad\mathbb{P}(A) = 0.01,$

And $\mathbb{P}(B|\bar{A}) = 0.1$ $\qquad\mathbb{P}(\bar{A}) = 0.99,$

$$\therefore\ \mathbb{P}(A|B) = \frac{.9\times .01}{.9\times .01 + .1\times .99} = \frac{1}{12};$$

that is, given that the person's test result is positive the probability that a person has the disease is $\frac{1}{12}$.

Alternatively, consider 1000 randomly selected and tested people.



$\mathbb{P}$ (disease)$= 0.01 = \frac{10}{1000}$

$\mathbb{P}$ (test positive|diseased) $= 0.9$,

$\mathbb{P}$ (test negative|not diseased) $= 0.9$

$\mathbb{P}$(diseased|test positive) $= \frac{9}{9+99} = \frac{9}{108} = \frac{1}{12}$

Part One —

Probability and Distribution Theory

Chapter 1

Random Variables

Consider a situation in which we want to take measurements of some variable of interest across multiple subjects. For example, we may be interested in measuring the weight loss achieved by participants in a weight loss program.

Inevitably, our measurements always vary from one subject to another due to factors that are beyond our control, or beyond our knowledge. For this reason, we treat the measurements as **random variables**.

In this chapter we will explore random variables and some properties of random variables that are important in their study.

Definition

For a discrete sample space Ω, a **random variable** X is a function defined on Ω with

$$\mathbb{P}(X = x) = \sum_{s:X=x} \mathbb{P}(\{s\})$$

being the probability that X takes the value x.

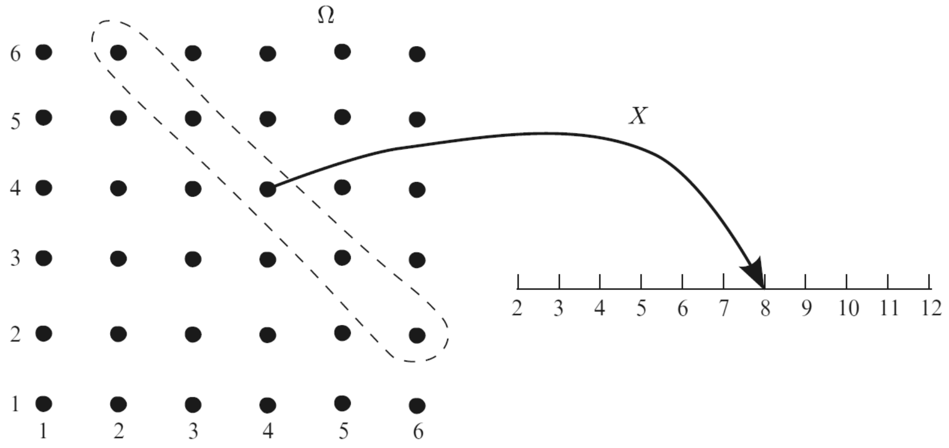
More generally, suppose we are given a sample space Ω with a collection of all subsets of Ω denoted by $\mathcal{F}$ (loosely $\mathcal{F}$ is a big bag containing all possible events) to which we can assign a probability specified by the set function $\mathbb{P}$. We call the triple $(\Omega, \mathcal{F}, \mathbb{P})$ a *probability space* and define a random variable in this space as follows.

Definition A numerical random variable X is a function $X(\omega)$ that takes any $\omega \in \Omega$ as an input and maps it onto the real line $\mathbb{R}$ with the property that for all $x \in \mathbb{R}$

$$A(x) = \{\omega \in \Omega : X(\omega) \leqslant x\}$$

is an event (sometimes written as $A(x) \in \mathcal{F}$ to indicate that it is in the bag of all possible events to which we can assign a probability).

The definition is depicted pictorially below, where $\Omega = \{(1,1), \ldots, (6,6)\}$ is the sample space generated by two dice thrown consecutively and $X(\{i,j\}) = i + j$ is the random variable indicating the sum of the faced of the two dice.



We now define the simplest non-trivial random variable.

Definition Let A be an event. Define the **indicator function** of A as the function $\mathrm{I}_A : \Omega \to \mathbb{R}$ (that is, it takes $\omega \in \Omega$ as an input and maps it to $\mathbb{R}$) such that for all $\omega \in \Omega$

$$\mathrm{I}_A = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases} .$$

Sometimes we may write $\mathrm{I}\{A\}$ or $\mathrm{I}_{\{A\}}$, depending on which happens to be the most economical, but unambiguous notation.

Example

Toss a fair coin 3 times.

$$\Omega = \{\mathrm{HHH}, \mathrm{HHT}, \mathrm{HTH}, \mathrm{THH}, \mathrm{TTH}, \mathrm{THT}, \mathrm{HTT}, \mathrm{TTT}\}$$

Let X denote the number of heads turned up. Then

$$\mathbb{P}(X=0) = \frac{1}{8}\ ,\ \mathbb{P}(X=1) = \frac{3}{8}\ ,\ \mathbb{P}(X=2) = \frac{3}{8}\ ,\ \mathbb{P}(X=3) = \frac{1}{8}.$$

Cumulative Distribution Function

The **cumulative distribution function** (cdf) of the random variable X is

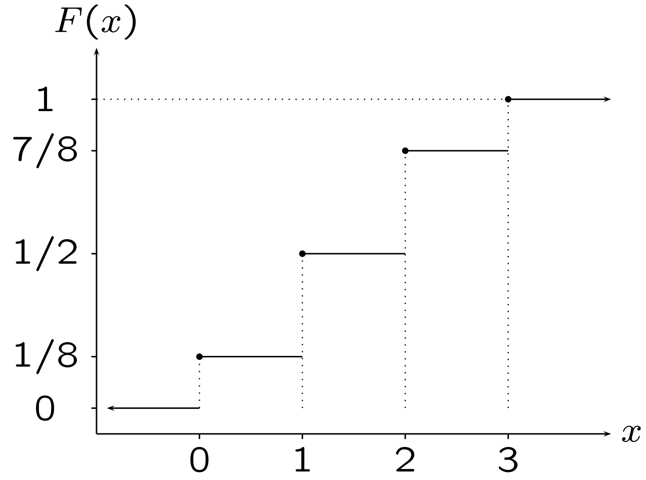
$$F_X(x) \overset{\text{def}}{=} \mathbb{P}(A(x)),$$

where $A(x) = \{\omega \in \Omega : X(\omega) \leqslant x\}$. Note that we can write this abbreviated as follows

$$F_X(x) = \mathbb{P}(X \leq x) \ .$$

Note that we may write $F(x)$ if there is no ambiguity about whose cdf we refer to.

The following figure shows the cdf for the previous example:



Theorem 1.1 (Properties of the Cdf)

1. *F is bounded between 0 and 1, and such that*

$$\lim_{x \downarrow -\infty} F(x) = 0, \qquad \lim_{x \uparrow \infty} F(x) = 1 \ .$$

2. *F is non-decreasing: if $x < y$, then $F(x) \leqslant F(y)$.*

3. *For any pair of numbers $x < y$*

$$\mathbb{P}(x < X \leq y) = F(y) - F(x)$$

4. *F is right-continuous:*

$$\lim_{n \uparrow \infty} F(x + 1/n) = F(x)$$

and

$$\mathbb{P}(X = x) = F(x) - \lim_{n \uparrow \infty} F(x - 1/n)$$

Proofs:

1. Define the event $A_n = \{\omega : X(\omega) \leqslant -n\}$ and note that $A_1 \supseteq A_2 \supseteq, \ldots$ is a sequence of decreasing events with limit the empty set:

$$\cap_{n=1}^{\infty} A_n = \emptyset$$

It follows from the **continuity from above** property that

$$\lim_{n\uparrow\infty} F(-n) = \lim_{n\uparrow\infty} \mathbb{P}(X \leqslant -n) = \lim_{n\uparrow\infty} \mathbb{P}(A_n) \overset{\text{continuity}}{=} \mathbb{P}(\cap_{n=1}^{\infty} A_n) = \mathbb{P}(\emptyset) = 0 \ .$$

Similarly for the case $F(\infty) = 1$.

2. Since $\{X \leqslant y\} = \{X \leqslant x\} \cup \{x < X \leqslant y\}$ and the events $\{X \leqslant x\}$ $\{x < X \leqslant y\}$ are disjoint (think two non-overlapping blobs in a Venn diagram), then we have

$$\mathbb{P}(X \leqslant y) = \mathbb{P}(X \leqslant x) + \mathbb{P}(x < X \leqslant y) \geqslant \mathbb{P}(X \leqslant x)$$

3. Rearrange the last equation above to obtain the result:

$$\mathbb{P}(x < X \leqslant y) = \mathbb{P}(X \leqslant y) - \mathbb{P}(X \leqslant x) = F(y) - F(x)$$

4. Set $B_n = \{x - 1/n < X \leqslant x\}$ and note that $B_1 \supseteq B_2 \supseteq \cdots$ is a decreasing sequence of events with limit $B = \cap_{n=1}^{\infty} B_n$ equivalent to $\{X = x\}$. Hence, by the continuity from above property:

$$\lim_{n\uparrow\infty} \mathbb{P}(B_n) = \mathbb{P}(B) = \mathbb{P}(X = x) \ .$$

Now note that

$$\underbrace{\mathbb{P}(x - 1/n < X \leqslant x)}_{\mathbb{P}(B_n)} = F(x) - F(x - 1/n)$$

and taking limits on both sides yields

$$\mathbb{P}(X = x) = F(x) - \lim_{n\uparrow\infty} F(x - 1/n) \ .$$

Now for the other case set $A_n = \{x < X \leqslant x + 1/n\}$ and note that

$$\mathbb{P}(A_n) = \mathbb{P}(x < X \leqslant x + 1/n) = F(x + 1/n) - F(x)$$

with $A_1 \supseteq A_2 \supseteq \cdots$ a decreasing sequence of events with limit $A_\infty \equiv \emptyset$. Hence,

$$\lim_{n\uparrow\infty} F(x + 1/n) - F(x) = \lim_{n\uparrow\infty} \mathbb{P}(A_n) = \mathbb{P}(A_\infty) = 0$$

Discrete Random Variables and Probability Functions

Definition
The random variable X is **discrete** if there are countably many values x for which $\mathbb{P}(X = x) > 0$.

The probability structure of X is most commonly described by its **probability function** (sometimes referred to as its "probability mass function", as in Hogg *et al* 2005).

Definition
The **probability function** of the discrete random variable X is the function f_X given by

$$f_X(x) = \mathbb{P}(X = x).$$

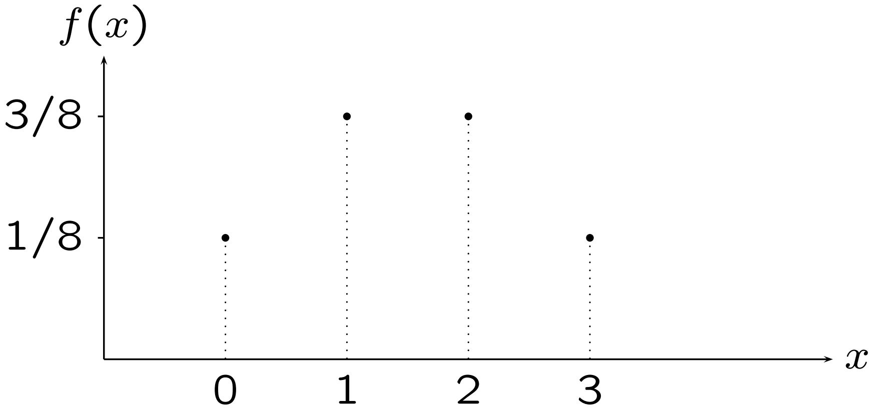
Results
The probability function of a discrete random variable X has the following properties:

1. $f_X(x) \geq 0$ for all $x \in \mathbb{R}$.

2. $\sum_{\text{all } x} f_X(x) = 1$.

Example

Below is the probability function for the number of heads in three coin tosses:



When X is discrete, $F_X(x)$ is the sum of the probabilities for possible values of X less than or equal to x.

Example

A coin, with p=probability of a head on a single toss, is tossed until a head turns up for the first time. Let X denote the number of tosses required.

Find the probability function and the cumulative distribution function of X.

If the tosses are independent, then

$$\begin{aligned} f_X(x) = \mathbb{P}(X = x) &= \mathbb{P}(x-1 \text{ tails then } 1 \text{ head}) \\ &= (1-p)^{x-1} \times p; \end{aligned}$$

that is, $f_X(x) = (1-p)^{x-1}p, \; x = 1, 2, \ldots; \; 0 < p < 1.$

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \leq x) = \sum_{a \leq x} \mathbb{P}(X = a) \\ &= \sum_{a=1}^{x} (1-p)^{a-1} p = 1 - (1-p)^x, \; x = 1, 2, \ldots \; . \end{aligned}$$

Continuous Random Variables and Density Functions

When a random variable has a continuum of possible values it is **continuous** (e.g. the lifetime of a light bulb has possible values in $[0, \infty)$). The analogue of the probability function for continuous random variables is the **density function** (sometimes called the "probability density function").

Definition

The **density function** of a continuous random variable is a real-valued function f_X on $\mathbb{R}$ with the property

$$\int_A f_X(x)\, dx = \mathbb{P}(X \in A)$$

for any (measurable) set $A \subseteq \mathbb{R}$.

Results

The density function of a continuous random variable X has the following properties:

1. $f_X(x) \geq 0$ for all $x \in \mathbb{R}$.
2. $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.

Regardless of whether a random variable X is continuous or discrete, its cumulative distribution function (cdf) is defined by

$$F_X(x) = \mathbb{P}(X \leq x).$$

The next two results show how F_X may be found from f_X and vice versa.

Result

The cumulative distribution function (cdf) F_X of a continuous random variable can be found from the density function f_X via

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

Result

The density function f_X of a continuous random variable can be be found from the cumulative distribution function (cdf) F_X via

$$f_X(x) = F_X'(x).$$

Result

For any continuous random variable X and pair of numbers $a \leq b$

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx = \text{area under } f_X \text{ between } a \text{ and } b.$$

These results demonstrate the importance of $f_X(x)$ and $F_X(x)$: if you know or can derive either $f_X(x)$ or $F_X(x)$, then you can derive any probability you want to about X, and hence any property of X that is of interest.

Some particularly important properties of X are described later in this chapter.

Example

The lifetime (in thousands of hours) X of a light bulb has density $f_X(x) = e^{-x}, x > 0$.

Find the probability that a bulb lasts between 2 thousand and 3 thousand hours.

The probability that a bulb lasts between 2 thousand hours and 3 thousand hours is

$$\int_2^3 e^{-x}\,dx = [-e^{-x}]_2^3 = e^{-2} - e^{-3} = (e-1)/e^3 \simeq 0.0855.$$

Alternatively, X has cdf

$$F_X(x) = 1 - e^{-x}, \quad x > 0.$$

and

$$\mathbb{P}(2 < X < 3) = F_X(3) - F_X(2) = 1 - e^{-3} - (1 - e^{-2}) = (e-1)/e^3 \simeq 0.0855.$$

Continuous random variables X have the property

$$\mathbb{P}(X = a) = 0 \quad \text{for any } a \in \mathbb{R}. \tag{1.1}$$

It only 'makes sense' to talk about the probability of X lying in some subset of $\mathbb{R}$.

A consequence of (1.1) is that, with continuous random variables, we don't have to worry about distinguishing between $<$ and $\leq$ signs. The probabilities are not affected. For example,

$$\mathbb{P}(2 < X < 3) = \mathbb{P}(2 \leq X < 3) = \mathbb{P}(2 < X \leq 3) = \mathbb{P}(2 \leq X \leq 3).$$

This is *not the case* for discrete random variables.

Definition

If F_X is strictly increasing in some interval, then F_X^{-1} is well defined and, for a specified $p \in (0,1)$, the ***p*th quantile** of F_X is x_p, where

$$F_X(x_p) = p \text{ or } x_p = F_X^{-1}(p).$$

$x_{0.5}$ is the **median** of F_X(or f). $x_{0.25}$, $x_{0.75}$ are the lower and upper **quartiles** of F_X (or f_X).

Example

Let X be a random variable with cumulative distribution function $F_X(x) = 1 - e^{-x}, x > 0$.

Find the median and quartiles of X.

$F_X(x) = 1 - e^{-x}, x > 0$, then $y = F_X(x) \Longrightarrow x = F_X^{-1}(y) = -\ln(1-y)$ and

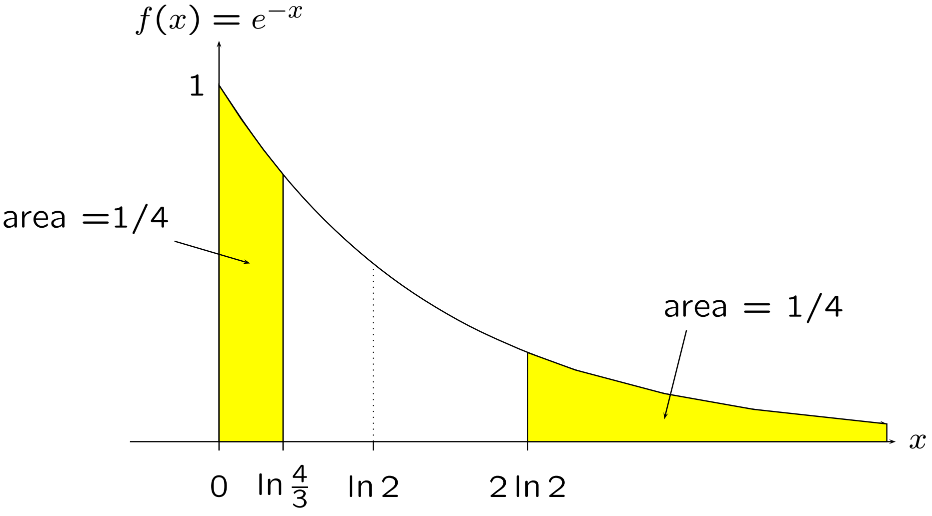
$$x_p = -\ln(1-p)$$

so

$$x_{0.5} = -\ln(0.5) = \ln 2$$

.

The lower quartile is $-\ln(0.75) = \ln 4/3$ and the upper quartile is $-\ln(0.25) = 2\ln 2$.



Area under f_X up to $\ln 2$ is $1/2$.

Expectation and Moments

Expectation

The mean or average of the numbers $a_1, a_2, \dots, a_n$ is

$$\frac{a_1 + \cdots + a_n}{n} = a_1 \cdot \frac{1}{n} + \cdots + a_n \cdot \frac{1}{n}.$$

Consider a random variable X with $\mathbb{P}(X = 5) = \frac{1}{5}$, $\mathbb{P}(X = 10) = \frac{4}{5}$. If we observed the values of, say, 100 random variables with the same distribution as X, we would expect to observe about 20 5's and about 80 10's so that the mean or average of the 100 numbers should be about

$$\frac{5 \times 20 + 10 \times 80}{100} = 5 \cdot \frac{1}{5} + 10 \cdot \frac{4}{5} \ (= 9);$$

that is, the sum of the possible values of X weighted by their probabilities.

Definition
The **expected value** or **mean** of a discrete random variable X is

$$\mathbb{E}X = \mathbb{E}[X] \overset{\text{def}}{=} \sum_{\text{all } x} x \times \mathbb{P}(X = x) = \sum_{\text{all } x} x f_X(x),$$

where f_X is the probability function of X.

By analogy, in the continuous case we have:

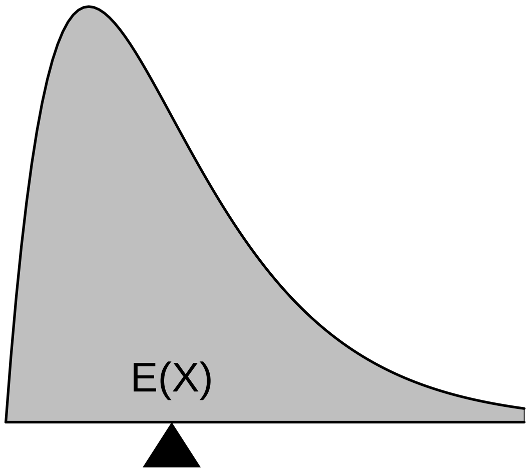
Definition
The **expected value** or **mean** of a continuous random variable X is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx,$$

where f_X is the density function of X.

In both cases, $\mathbb{E}(X)$ has the interpretation of being the *long run average* of X – in the long run, as you observe an increasing number of values of X, the average of these values approaches $\mathbb{E}(X)$.

In both cases, $\mathbb{E}(X)$ has the physical interpretation of the *centre of gravity* of the function f_X. So if a piece of thick wood or stone was carved in the shape of f_X; it would balance on a fulcrum placed at $\mathbb{E}(X)$ (see accompanying figure).



Example

Let X be the number of females in a committee with three members. Assume that there is a 50:50 chance of each committee member being female, and that committee members are chosen independently of each other.

Find $\mathbb{E}[X]$.

x	0	1	2	3
$f_X(x) = \mathbb{P}(X = x)$	1/8	3/8	3/8	1/8

$$\mathbb{E}[X] = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{3}{2}$$

The interpretation of $\frac{3}{2}$ is not that you expect X to be $\frac{3}{2}$ (it can only be 0,1,2 or 3) but that if you repeated the experiment, say, 100 times, then the average of the 100 numbers observed should be about $\frac{3}{2}$ $(= \frac{150}{100})$.

That is, we expect to observe about 150 females in total in 100 committees. We don't expect to see exactly 1.5 females on each committee!

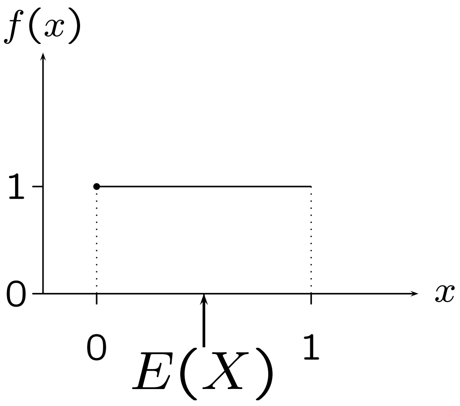
Example

Suppose X is a standard uniform random number generator (such as can be found on most hand-held calculators).

X has density $f_X(x) = 1, \quad 0 < x < 1$.

Find $\mathbb{E}[X]$.

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^1 x \cdot 1 dx = \frac{1}{2}$$



Note that in both of the above examples, f_X is symmetric and it is symmetric about $\mathbb{E}[X]$. This is not always the case, as in the following examples:

Example*

X has probability function

$$f_X(x) = \mathbb{P}(X = x) = (1-p)^{x-1} \cdot p, \ \ x = 1, 2, \ldots; \ 0 < p < 1.$$

Find $\mathbb{E}(X)$.

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{x=1}^{\infty} x(1-p)^{x-1} \cdot p \\
&= -p \sum_{x=1}^{\infty} \frac{d}{dp}(1-p)^x = -p\frac{d}{dp} \sum_{x=1}^{\infty}(1-p)^x \\
&= -p\frac{d}{dp}\left(\frac{1}{p} - 1\right) = \frac{1}{p}.
\end{aligned}$$

Example

X has density

$$f_X(x) = e^{-x}, x > 0$$

Find $\mathbb{E}(X)$.

$$\begin{aligned}
\mathbb{E}(X) &= \int_0^{\infty} xe^{-x} dx \\
&= -[xe^{-x}]_0^{\infty} + \int_0^{\infty} e^{-x} dx = 1
\end{aligned}$$

Special case: if X is degenerate, that is, $X = c$ with probability 1 for some constant c, then X is in fact just a constant and $\mathbb{E}[X] = \sum_{\text{all } x} xP(X = x) = c \cdot 1 = c$. Thus the expected value of a constant is the constant; that is, $\mathbb{E}[c] = c$.

Expectation of transformed random variables

Sometimes we are interested in a transformation of a random variable. Examples:

- Circumference of a tree trunk is measured, but we want to know the cross-sectional area of the trunk. The variable of interest is $\pi \left(\frac{X}{2\pi}\right)^2$.
- We count the number of consecutive times (X) a pair of coins land "odds" (one tail and one head) in a game of two-up. The variable of interest is whether or not the casino wins all bets (which happens for $X > 4$).

Transformations are also of interest when studying the properties of a random variable. For example, in order to understand X, it is often useful to look at the ***r*th moment** of X about some constant a, defined as $\mathbb{E}[(X-a)^r]$. This is another example of a transformation of X, for which we wish to find $\mathbb{E}[g(X)]$, for some function $g(x)$.

The following is an important result regarding the expectation of a transformation of a random variable:

Result

The expected value of a function $g(X)$ of a random variable X is

$$\mathbb{E}g(X) = \begin{cases} \sum_{\text{all } x} g(x) f_X(x) & X \text{ discrete} \\ \\ \int_{-\infty}^{\infty} g(x) f_X(x) dx & X \text{ continuous} \end{cases}$$

proof: We prove this only in the discrete case. Set $Y = g(X)$ and $f_X(x) = \mathbb{P}(X = x)$, then

$$\begin{aligned} \mathbb{E}g(X) = \mathbb{E}Y &\overset{\text{def}}{=} \sum_y y\, \mathbb{P}(Y = y) \\ &= \sum_y y\, \mathbb{P}(g(X) = y) = \sum_y y\, \mathbb{P}(X \in \{x : g(x) = y\}) \\ &= \sum_y y \sum_{x:g(x)=y} \mathbb{P}(X = x) = \sum_y \sum_{x:g(x)=y} y\mathbb{P}(X = x) \\ &= \sum_y \sum_{x:g(x)=y} g(x)\mathbb{P}(X = x) = \sum_x g(x)\mathbb{P}(X = x) \end{aligned}$$

where the last line follows from the fact that if x takes on all values in its domain, then $y = g(x)$ takes on all values in its range and vice-versa. □

Example

Let I denote the electric current, through a particular circuit, and I has density function

$$f_I(x) = \begin{cases} \frac{1}{2} & 1 \leq x \leq 3 \\ \\ 0 & \text{elsewhere} \end{cases}$$

so the expected value of the current is

$$\mathbb{E}(I) = \int_1^3 x\frac{1}{2}dx = 2.$$

Power $\mathbb{P}$ is a function of I and resistance. For a circuit with resistance 3 Ohms,

$$P = 3I^2$$

What is the expected value of P through this circuit?

$$\begin{aligned} \mathbb{E}(P) &= \int_1^3 3x^2 f(x)dx \\ &= 3\int_1^3 \frac{1}{2}x^2 dx \\ &= \frac{3}{2}\left[\frac{1}{3}x^3\right]_1^3 \\ &= \frac{3}{2}\frac{1}{3}(3^3 - 1) \\ &= 13 \end{aligned}$$

Note: In most situations,

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

Expectation of a variable under changes of scale

Often a change of scale is required, when studying a random variable. An example is when a change of measurement units is required (g→kg, $^0F \rightarrow ^0 C$).

Results

If a is a constant,

$$\begin{aligned} \mathbb{E}[X+a] &= \sum_{\text{all } x} (x+a)\mathbb{P}(X=x) \\ &= \mathbb{E}[X]+a \\ \text{and } \mathbb{E}[aX] &= \sum_{\text{all } x} ax\mathbb{P}(X=x) = a\mathbb{E}(X). \end{aligned}$$

Similarly for X continuous.

Also, $\mathbb{E}[g_1(X)+\cdots+g_n(X)] = \mathbb{E}[g_1(X)]+\ldots+\mathbb{E}[g_n(X)]$.

Standard Deviation and Variance

The standard deviation of a random variable is a measure of its *spread*. It is closely tied to the variance of a random variable, defined below:

Definition

If we let $\mu = \mathbb{E}(X)$, then the **variance** of X denoted by $\mathrm{Var}(X)$ is defined as

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2]$$

(which is the second moment of X about μ).

Definition

The **standard deviation** of a random variable X is the square-root of its variance:

$$\text{standard deviation of } X = \sqrt{\mathrm{Var}(X)}.$$

Standard deviations are more readily interpreted, because they are measured in the same units as the original variable X. So standard deviations are more commonly used as measures of spread in applied statistics, and in reporting the results of quantitative research.

Variances are a bit easier to work with theoretically, and so are more commonly used in mathematical statistics (and in this course).

Result

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mu^2.$$

Proof:

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}(X-\mu)^2\\ &= \mathbb{E}(X^2-2\mu X+\mu^2)\\ &= \mathbb{E}(X^2)-2\mu\mathbb{E}(X)+\mathbb{E}(\mu^2)\\ &= \mathbb{E}(X^2)-2\mu^2+\mu^2\\ &= \mathbb{E}(X^2)-\mu^2.\end{aligned}$$

Example

Assume the lifetime of a lightbulb (in thousands of hours) has density function $f_X(x) = e^{-x}$.

Calculate $\mathrm{Var}(X)$

We will use $\mathrm{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$.

Recall from page 14 that $\mathbb{E}(X) = 1$.

$$\begin{aligned}\mathbb{E}(X^2) &= \int_0^\infty x^2e^{-x}dx\\ &= -\left[x^2e^{-x}\right]_0^\infty - \int_0^\infty -2xe^{-x}dx\\ &= 0 + 2E(X) = 2\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2\\ &= 2-1^2 = 1\end{aligned}$$

Example

Consider two random variables A and B.

A:

x	1	2	3	4	5
$f_A(x)$	0.15	0.25	0.2	0.25	0.15

B:

x	1	2	3	4	5
$f_B(x)$	0.1	0.1	0.6	0.1	0.1

Which of A and B is more variable?

$\mathbb{E}[A] = \mathbb{E}[B] = 3$
$\mathbb{E}[A^2] = 10.7$, $\mathbb{E}[B^2] = 10$
$\mathrm{Var}(A) = 1.7, \mathrm{Var}(B) = 1$

Results

$$\mathrm{Var}(X + a) = \mathrm{Var}(X)$$

$$\mathrm{Var}(aX) = a^2\mathrm{Var}(X).$$

Moment Generating Functions

Definition

The **moment generating function** (mgf) of a random variable X is

$$m_X(u) = \mathbb{E}(\mathrm{e}^{uX}).$$

We say that the moment generating function of X exists if $m_X(u)$ is finite in some interval containing zero.

The name "moment generating function" comes from the following result concerning the rth moment of X, that is $\mathbb{E}(X^r)$:

Result

In general, $\mathbb{E}(X^r) = m_X^{(r)}(0)$ for $r = 0, 1, 2, \ldots$

Derivation: Without going into details, the condition that the moment generating function is finite on some interval around zero guarantees the existence of all moements and that the derivative and the expectation can be interchanged and

$$\begin{aligned} m_X^{(r)}(u) &= \partial_u^{(r)}\mathbb{E}(e^{uX}) \\ &= \mathbb{E}(\partial_u^{(r)}e^{uX}) \\ &= \mathbb{E}(X^r e^{uX}) \end{aligned}$$

given that the r-th derivative $m_X^{(r)}(u)$ is smooth enough around zero, the r-th moment can be computed by

$$\mathbb{E}(X^r) = \mathbb{E}(X^r e^{uX}\big|_{u=0}) = \mathbb{E}(X^r e^{uX})\big|_{u=0}$$

Example

X has mgf

$$m_X(u) = \begin{cases} (1-u)^{-1} & , u < 1 \\ \infty & , u \geq 1. \end{cases}$$

Find an expression for the rth moment of X.

$$\begin{aligned} \mathbb{E}(X) = m'_X(0) &= (1-u)^{-2}|_{u=0} = 1, \\ \mathbb{E}(X^2) &= m_X^{(2)}(0) = 2(1-u)^{-3}|_{u=0} = 2 \\ & \qquad (\mathrm{Var}(X) = 2 - 1 = 1) \\ \mathbb{E}(X^r) &= m_X^{(r)}(0) \\ &= 1 \cdot 2 \cdot 3 \ldots r(1-u)^{-r-1}|_{u=0} = r! \end{aligned}$$

Example*

X has probability function

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

Find the mgf of X. Hence find $\mathbb{E}(X)$ and $\mathrm{Var}(X)$. X has mgf

$$\begin{aligned} m_X(u) = \mathbb{E}(e^{uX}) &= \sum_{x=0}^{\infty} e^{ux} \cdot \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^u)^x}{x!} \\ &= e^{-\lambda} \cdot e^{\lambda e^u} = e^{\lambda(e^u - 1)}. \\ \mathbb{E}(X) &= m'_X(0) = e^{\lambda(e^u-1)} \cdot \lambda e^u|_{u=0} = \lambda \\ \mathbb{E}(X^2) &= m_X^{(2)}(0) \\ &= \lambda\left\{e^{\lambda(e^u-1)} \cdot e^u + e^{\lambda(e^u-1)} \cdot \lambda e^u \cdot e^u\right\}|_{u=0} \\ &= \lambda(1+\lambda) \\ \therefore \mathrm{Var}(X) &= \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2 = \lambda. \end{aligned}$$

Example*

X has probability function

$$f_X(x) = \mathbb{P}(X = x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}, x = 0, 1, \ldots, n;\ 0 < p < 1.$$

The Binomial Theorem states that

$$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x \, b^{n-x}$$

Use this result, when required, in order to:

1. Show that $f_X(x)$ is a probability function.
2. Find the mgf of X.
3. Find $\mathbb{E}(X)$ and $\mathrm{Var}(X)$ using the mgf.

$$\begin{aligned}
(a+b)^n &= \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x}. \\
m_X(u) &= \mathbb{E}\mathrm{e}^{uX} \\
&= \sum_{x=0}^{n} \mathrm{e}^{ux} \binom{n}{x} p^x (1-p)^{n-x} \\
&= (1 - p + pe^u)^n. \\
\mathbb{E}(X) &= m'_X(0) \\
&= m_X(1 - p + pe^u)^{n-1} \cdot pe^u|_{u=0} = np \\
\mathbb{E}(X^2) &= m_X^{(2)}(0) \\
&= np\{(1 - p + p\mathrm{e}^u)^{n-1} u + (n-1)(1 - p + pe^u)^{n-2} \cdot pe^u \cdot e^u\}|_{u=0} \\
&= np\{1 + (n-1)p\}. \\
\therefore \mathrm{Var}(X) &= np(1-p).
\end{aligned}$$

Example

X has density

$$f_X(x) = \mathrm{e}^{-x}, x > 0$$

Find the moment generating function of X.

$$\begin{aligned} m_X(u) &= \mathbb{E}(e^{uX}) = \int_0^\infty e^{ux} \cdot e^{-x} dx \\ &= \int_0^\infty e^{(u-1)x} dx = \begin{cases} (1-u)^{-1} & , u < 1 \\ +\infty & , u \geq 1. \end{cases} \end{aligned}$$

Then, as in (1), $\mathbb{E}(X^r) = r!, \ r = 0, 1, 2, \ldots$.

Properties of moment generating functions

The following results on uniqueness and convergence for moment generating functions will be particularly important later on.

Result

Let X and Y be two random variables all of whose moments exist. If

$$m_X(u) = m_Y(u)$$

for all u in a neighbourhood of 0 (i.e. for all $|u| < \varepsilon$ for some $\varepsilon > 0$) then

$$F_X(x) = F_Y(x) \quad \text{for all } x \in \mathbb{R}.$$

(*i.e.* the mgf of a random variable is *unique*)

Result

Let $\{X_n : n = 1, 2, \ldots\}$ be a sequence of random variables, each with moment generating function $m_{X_n}(u)$. Furthermore, suppose that

$$\lim_{n\to\infty} m_{X_n}(u) = m_X(u) \quad \text{for all } u \text{ in a neighbourhood of } 0$$

and $m_X(u)$ is a moment generating function of a random variable X. Then

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \quad \text{for all } x \in \mathbb{R}.$$

(*i.e.* convergence of mgf's implies convergence of cdf's)

These two results are stated as theorems in the book

Casella, G. and Berger, R.L. (1990). *Statistical Inference*, Duxbury.

The proofs rely on the theory of *Laplace transforms* but are not given in this reference. Instead, the reader is referred to

Widder, D.V. (1946). *The Laplace Transform.* Princeton, New Jersey: Princeton University Press.

Location and scale families of densities

The unknown density function of a continuous random variable may belong to a family of density functions that all have a similar form. Sometimes, such a family has a form which identifies it as being generated by location or scale changes. This can be useful in Statistics, where the unknown location or scale constants are called *parameters*, and have a clear meaning.

Result

Consider a random variable U with density function $f_U(x)$.
A *location family* of densities based on the random variable U is the family of densities $f_X(x)$ where $X = U + c$ for all possible c. $f_X(x)$ is given by:

$$f_X(x) = f_U(x - c)$$

A *scale family* of densities based on the random variable U is the family of densities $f_X(x)$ where $X = cU$ for all possible c. $f_X(x)$ is given by:

$$f_X(x) = c^{-1} f_U(x/c)$$

Proof: Let $X = U + c$. Then the cdf of X is given by

$$F_X(x) = \mathbb{P}(X \leqslant x) = \mathbb{P}(U + c \leqslant x) = \mathbb{P}(U \leqslant x - c) = F_U(x - c)$$

Differentiating both sides we find that $f_X(x) = f_U(x - c)$.

Letting $X = cU$ we can use a similar approach to show that $f_X(x) = c^{-1} f_U(x/c)$.

When specific distributions are introduced in chapter 2, you may be able to identify some which are generated by location or scale changes.

Example

Consider the following families of density functions:

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

2. $f_X(x) = \frac{1}{2\beta} e^{|(x-\mu)/\beta|}, \ -\infty < x < \infty$

3. $f_X(x) = kx^{k-1} e^{-x^k}, \ x > 0$

For each of the above families of distributions, explain whether it is a location family, a scale family, a location and scale family, or neither.

Example

Consider the random variable U which has density function

$$f_U(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \infty < x < \infty$$

Find the density function of the scale family based on U.

Bounding Probabilities

The **inclusion-exclusion** identity states that

$$\mathbb{P}(\cup_{i=1}^n A_i) = \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots$$
$$\cdots + (-1)^{n+1} \mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n)$$

It is frequently used to approximate or bound the probabilities of compound events such as $\cup_{i=1}^n A_i$.

Proof: Define $\mathrm{I}\{B\}$ to be the indicator function for the event $B = \cup_{i=1}^n A_i$. By De

Morgan's laws we have that:

$$\begin{aligned}
\mathrm{I}\{B\} &= 1 - \mathrm{I}\{\overline{B}\} \\
&= 1 - \mathrm{I}\{\overline{\cup_{i=1}^{n} A_i}\} \\
&= 1 - \mathrm{I}\{\cap_{i=1}^{n} \overline{A}_i\} \qquad \text{by De Morgan law} \\
&= 1 - \prod_{i=1}^{n} \mathrm{I}\{\overline{A}_i\} \qquad \text{since } \mathrm{I}\{A \cap B\} = \mathrm{I}\{A\}\mathrm{I}\{B\} \\
&= 1 - \prod_{i=1}^{n} (1 - \mathrm{I}\{A_i\})
\end{aligned}$$

Finally, the inclusion-exclusion identity follows by applying the expectation operator $\mathbb{E}[\,\cdot\,]$ on both sides of the expansion:

$$\begin{aligned}
\mathrm{I}\{B\} &= 1 - \prod_{i=1}^{n} (1 - \mathrm{I}\{A_i\}) \\
&= \sum_i \mathrm{I}\{A_i\} - \sum_{i<j} \mathrm{I}\{A_i\}\mathrm{I}\{A_j\} + \cdots + (-1)^{n+1} \prod_i \mathrm{I}\{A_i\} \\
&= \sum_i \mathrm{I}\{A_i\} - \sum_{i<j} \mathrm{I}\{A_i \cap A_j\} + \cdots + (-1)^{n+1} \mathrm{I}\{\cap_i A_i\}\ .
\end{aligned}$$

Without proof we now state what is commonly known as **Boole's inequalities**, which will be use in MATH2931:

$$\begin{aligned}
\mathbb{P}(\cup_{i=1}^{n} A_i) &\leqslant \sum_i \mathbb{P}(A_i) \\
&\leqslant \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) \\
&\vdots \\
\mathbb{P}(\cup_{i=1}^{n} A_i) &\geqslant \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) \\
&\geqslant \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i, A_j, A_k) - \sum_{i<j<k<m} \mathbb{P}(A_i, A_j, A_k, A_m) \\
&\vdots
\end{aligned}$$

Thus, the inclusion-exclusion formula is a kind of 'Taylor' expansion for probabilities and can be useful for obtaining upper and lower bounds on the unknown probability.

Chebychev's Inequality

Chebychev's Inequality is a fundamental result concerning tail probabilities of general random variables. It is useful for derivation of convergence results given later in the notes.

Chebychev's Inequality:
If X is any random variable with $\mathbb{E}(X) = \mu, \mathrm{Var}(X) = \sigma^2$ then

$$\mathbb{P}(|X - \mu| > k\sigma) \leq \frac{1}{k^2}.$$

The probability statement in Chebychev's Inequality is often stated verbally as:

the probability that X is more than k standard deviations from its mean.

Note that Chebychev's Inequality makes no assumptions about the distribution of X. This is a handy result – in practice, we usually do not know what the distribution of X is. But using Chebychev's inequality, we can make specific probabilistic statements about a random variable given only its mean and standard deviation!

Proof (continuous case):

$$\begin{aligned}\sigma^2 = \mathrm{Var}(X) &= \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)dx \\ &\geq \int_{|x-\mu|>k\sigma} (x-\mu)^2 f_X(x)dx \\ &\geq \int_{|x-\mu|>k\sigma} (k\sigma)^2 f_X(x)dx\end{aligned}$$

since $|x-\mu| > k\sigma \implies (x-\mu)^2 f_X(x) > (k\sigma)^2 f_X(x)$

$$\begin{aligned}\therefore \sigma^2 &\geq k^2\sigma^2 \int_{|x-\mu|>k\sigma} f_X(x)dx \\ &= k^2\sigma^2 P(|X-\mu| > k\sigma)\end{aligned}$$

$$\therefore \mathbb{P}(|X-\mu| > k\sigma) \leq \frac{1}{k^2}.$$

Example

The number of items a factory produces in 1 day has mean 500 and variance 100.

What is a lower bound for the probability that between 400 and 600 items will be produced tomorrow?

Let X denote the number of items produced tomorrow.

$$\mu = 500 \ , \ \ \sigma = 10 \ . \text{ Put } k = 10.$$

$$\mathbb{P}(|X - 500| > 10.10) = \mathbb{P}(X < 400)$$

$$+\mathbb{P}(X > 600) \leq \frac{1}{10^2}$$

$$\text{or } \mathbb{P}(400 \leq X \leq 600) \geq 1 - \frac{1}{100} = 0.99.$$

$\lrcorner$

Jensen's Inequality

Another important inequality is called Jensen's inequality and it states the following.

If $h(x)$ is a convex function and X is a random variable, then

$$\mathbb{E}h(X) \geqslant h(\mathbb{E}X)$$

For example, suppose that $h(x) = x^2$, then Jensen's inequality implies that

$$\mathbb{E}X^2 \geqslant (\mathbb{E}X)^2$$

which implies that

$$\text{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 \geqslant 0,$$

as we already suspected. As another example consider the case where $g(x) > 0$ and $f(x) > 0$ are densities on $\mathbb{R}$ with $X \sim f$ (that is, X having pdf f), then we have

$$\mathbb{E}\ln\frac{f(X)}{g(X)} = -\mathbb{E}\ln\frac{g(X)}{f(X)} \geqslant -\ln\mathbb{E}\frac{g(X)}{f(X)} = -\ln\int_{\mathbb{R}} g(x)\mathrm{d}x = -\ln(1) = 0$$

proof of Jensen's inequality: Since h is convex we have that for some constants a and b:

$$h(x) \geqslant ax + b$$

with $h(\mu) = a\mu + b$ for $x = \mu$ (line may be tangent to h at this point). Now put $\mu = \mathbb{E}X$, $x = X$ and eliminate $b = h(\mu) - a\mu$ to get

$$h(X) \geqslant a(X - \mu) + h(\mu)$$

Hence,

$$\mathbb{E}h(X) \geqslant h(\mu) + a\mathbb{E}[X - \mu] = h(\mu)$$

Describing a variable using data

Given a sample of data, $\{x_1, x_2, \ldots, x_n\}$, how would you summarise it, graphically or numerically? Below we will quickly review some key tools.

You will not be expected to construct the following numerical and graphical summaries by hand, but you should understand how they are produced, know how to produce them using the statistics package R, and know how to interpret such summaries.

Two steps to data analysis

The first two things to think about in data analysis are:

1. What is the research question? Descriptive statistics should primarily focus on providing insight into this question.
2. What are the properties of the variables of primary interest?

The most important property to think about when constructing descriptive statistics is whether each variable is **categorical** or **quantitative**.

Any variable measured on subjects is either categorical or quantitative.
Categorical – responses can be sorted into a finite set of (unordered) categories, e.g. gender.
Quantitative – Responses are measured on some sort of scale, e.g. height.

If the sample $\{x_1, x_2, \ldots, x_n\}$ comes from a quantitative variable, then the x_i are real numbers, $x_i \in \Re$. If it comes from a categorical variable, then each x_i comes from a finite set of categories or "levels", $x_i \in \{C_1, C_2, \ldots, C_K\}$.

Example

Consider the following questions:

1. Will more people vote the Liberal party ahead of Labour at the next election?
2. Does whether or not pregnant guinea pigs are given a nicotine treatment affect the number of errors made in a maze by their offspring?
3. Is whether or not a Titanic passenger survived related to their gender?
4. How does brain mass change in dinosaurs, as body mass increases?

What are the variables of interest in these questions? Are each of these variables categorical or quantitative?

We will work through each of the methods mentioned in the above table.

Example

Summary of descriptive methods

Useful descriptive methods for when we wish to summarise one variable, or the association between two variables, depend on whether these variables are categorical or quantitative.

Does the research question involve:

	One variable		Two variables		
Data type:	Categorical	Quantitative	Both categorical	One of each	Both quantative
Numerics:	Table of frequencies	Mean/sd Median/quantiles	Two-way table	Mean/sd per group	Correlation
Graphs:	Bar chart	Dotplot Boxplot Histogram	Clustered bar chart	Scatterplot Boxplots Histograms etc.	Scatterplot

Consider again the the research questions of the previous example.

What method(s) would you use to construct a graph to answer each research question?

Categorical data

We will simultaneously treat the problems of summarising one categorical variable and studying the association between two categorical variables, because similar methods are used for these problems.

Numerical summaries of categorical data

The main tool for summarising categorical data is a table of frequencies (or percentages).

A **table of frequencies** consists of the counts of how many subjects fall into each level of a categorical variable.
A **two-way table** (of frequencies) counts how many subjects fall into each combination of levels from a pair of categorical variables.

Example

We can summarise the NSW election poll as follows:

Party	Liberal	Labour
Frequency	237	128

Example

Consider the question of whether there is an association between gender and whether or not a passenger on the Titanic survived.

We can summarise the results from passenger records as follows:

		Outcome	
		Survived	Died
Gender	Male	142	709
	Female	308	154

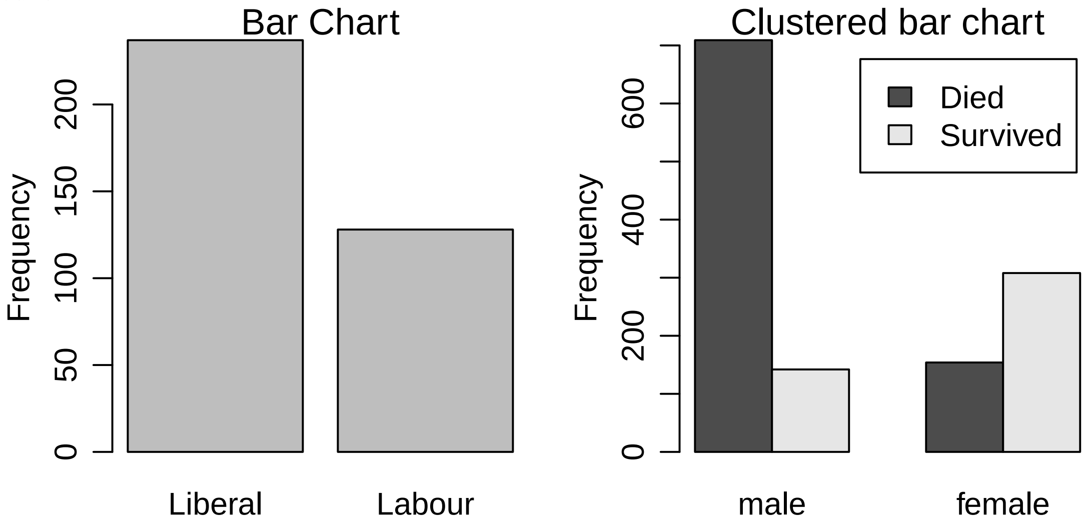
which suggests that a much higher proportion of females survived: their survival rate was 67% vs 17%!

In the Titanic example, an alternative summary was the percentage survival for each gender. Whenever one of the variables of interest has only two possible outcomes a list (or table) of percentages is a useful alternative way to summarise the data.

If you are interested in an association between more than two categorical variables you can extend the above ideas, *e.g.* construct a three-way table...

Graphical summaries of categorical data

A **bar chart** is a graph of a table of frequencies. A **clustered bar chart** graphs a two-way table, spacing the "bars" out as clusters to indicate the two-variable structure:



Pie charts are often used to graph categorical variables, however these are not generally recommended. It has been shown that readers of pie charts find it more difficult to understand the information that is contained in them, *e.g.* comparing the relative size of frequencies across categories. (For details, see the Wikipedia entry on pie charts and references therein http://en.wikipedia.org/wiki/Pie_chart)

Quantitative data

When summarising a quantitative variable, we are usually interested in three things:

- **Location** or "centre" – a value around which most of the data lie
- **Spread** – how variable the values are around their centre.
- **Shape** – other information about a variable apart from location and spread. Skewness is an important example.

Numerical summaries of quantitative data

The most commonly used numerical summaries of a quantitative variable are the sample mean, variance and standard deviation:

The **sample mean**

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

is a natural measure of location of a quantitative variable.

The **sample variance**

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

is a common measure of spread.

The **sample standard deviation** is defined as $s = \sqrt{s^2}$.

The variance is a useful quantity for theoretical purposes, as we will see in the coming chapters. The standard deviation however is of more practical interest because it is on the same scale as the original variable and hence is more readily interpreted.

The sample mean and variance are very widely used and we will derive a range of useful results about these estimators in this course.

Let's say we order the n values in the dataset and write them in increasing order as $\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}$. For example, $x_{(3)}$ is the third smallest observation in the dataset.

The **sample median** is

$$\tilde{x}_{0.5} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)}\right) & \text{if } n \text{ is even} \end{cases}$$

More generally, the ***p*th sample quantile** of the data x is

$$\tilde{x}_p = x_{(k)} \quad \text{where} \quad p = \frac{k - 0.5}{n}$$

for $k \in \{1, 2, \dots, n\}$. We can estimate the sample quantile for other values of p by linear interpolation.

The median is sometimes suggested as a measure of location, instead of $\bar{x}$, because it is much less sensitive to unusual observations (outliers). However, it is much less widely used in practice.

There are a number of alternative (but very similar) ways of defining sample quantiles. A different method again is used as the default approach on the statistics package `R`.

Example

The following (ordered) dataset is the number of mistakes made when ten subjects are each asked to do a repetitive task 500 times.

$$2 \quad 4 \quad 5 \quad 7 \quad 8 \quad 10 \quad 14 \quad 17 \quad 27 \quad 35$$

Find the 5th and 15th sample percentiles of the data. Hence find the 10th percentile.

There are ten observations in the dataset, so the 5th sample percentile is

$$\tilde{x}_{(1-0.5)/10} = \tilde{x}_{0.05} = 2$$

Similarly, the 15th sample percentile is 4. The 10th sample percentile is the average of these two. So $\tilde{x}_{0.1}$ can be estimated as

$$\tilde{x}_{0.1} = \frac{1}{2}\left(x_{(1)} + x_{(2)}\right) = \frac{1}{2}(2+4) = 3$$

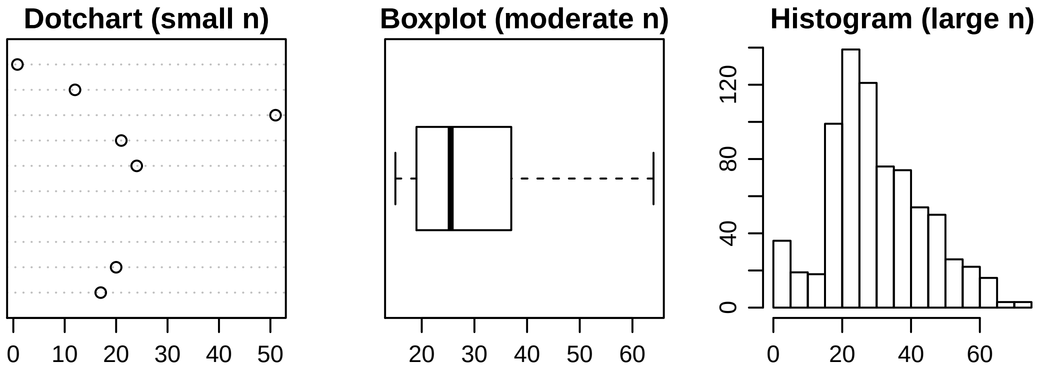
Apart from $\tilde{x}_{0.5}$, the two important quantiles are the **first and third quartiles**, $\tilde{x}_{0.25}$ and $\tilde{x}_{0.75}$ respectively. These terms are used to define the **interquartile range**

$$IQR = \tilde{x}_{0.75} - \tilde{x}_{0.25}$$

which is sometimes suggested as an alternative measure of spread to the sample standard deviation, because it is much less sensitive to unusual observations (outliers).

Graphical summaries of quantitative data

There are many ways to summarise a variable, and a key thing to consider when choosing a graphical method is the sample size (n). Some common plots:



A **dotchart** is a plot of each variable (x-axis) against its observation number, with data labels (if available). This is useful for small samples ($e.g.$ $n < 20$).

A **boxplot** concisely describes location, spread and shape via the median, quartiles and extremes:

- The line in the middle of the box is the median, the measure of centre.
- The box is bounded by the upper and lower quartiles, so box width is a measure of spread (the interquartile range, IQR).
- The whiskers extend until the most extreme value within one and a half interquartile ranges ($1.5IQR$) of the nearest quartile.
- Any value farther than $1.5IQR$ from its nearest quartile is classified as an extreme value (or "outlier"), and labelled as a dot or open circle.

Boxplots are most useful for moderate-sized samples (*e.g.* $10 < n < 50$).

A **histogram** is a plot of the frequencies or relative frequencies of values within different intervals or *bins* that cover the range of all observed values in the sample. Note that this involves breaking the data up into smaller subsamples, and as such it will only find meaningful structure if the sample is large enough (*e.g.* $n > 30$) for the subsamples to contain non-trivial counts.

An issue in histogram construction is choice of number of bins. A useful rough rule is to use

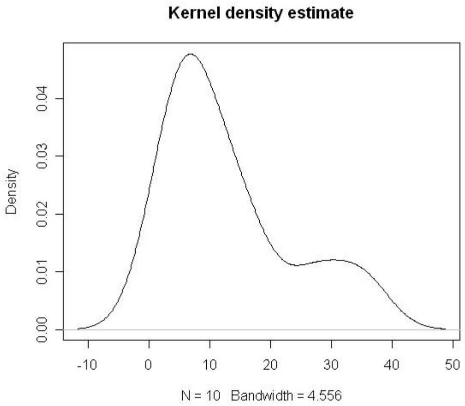
$$\text{number of bins} = \sqrt{n}$$

A histogram is a step-wise rather than smooth function. A quantitative variable that is continuous (*i.e.* a variable that can take any value within some interval) might be better summarised by a smooth function. So an alternative estimator that often has better properties for continuous data is a **kernel density estimator**:

$$\widehat{f_h}(x) = \frac{1}{n}\sum_{i=1}^{n} w_h(x - x_i)$$

for some choice of weighting function $w_h(x)$ which includes a "bandwidth parameter" h.

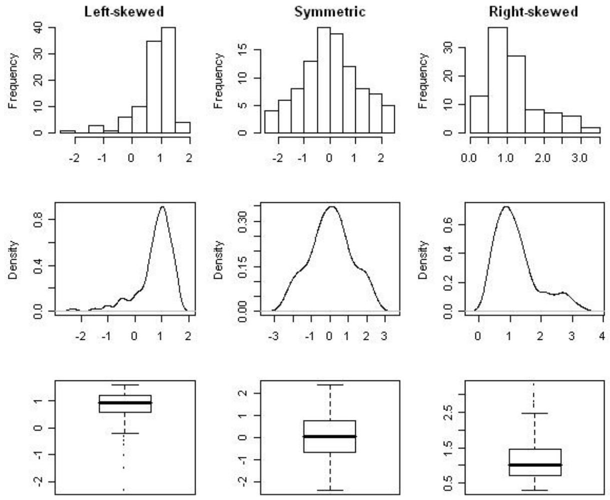
Usually, $w(x)$ is chosen to be the normal density (defined in Chapter 4) with mean 0 and standard deviation h. A lot of research has studied the issue of how to choose a bandwidth h, and most statistics packages are now able to automatically choose an estimate of h that usually performs well. The larger h is, the larger the bandwidth that is used *i.e.* the larger the range of observed values x_i that influence estimation of $\widehat{f_h}(x)$ at any given point x.



Shape of a distribution

Something we can see from a graph that is hard to see from numerical summaries is the **shape** of a distribution. Shape properties, broadly, are characteristics of the distribution apart from location and spread.

An example of an important shape property is **skew** – if the data tend to be asymmetric about its centre, it is skewed. We say data are "left-skewed" if the left tail is longer than the right, conversely, data are right-skewed if the right-tail is longer.



There are some numerical measures of shape, *e.g.* the coefficient of skewness κ_1:

$$\widehat{\kappa}_1 = \frac{1}{(n-1)s^3}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

but they are rarely used – perhaps because of extreme sensitivity to outliers, and perhaps because shape properties can be easily visualised as above.

Another important thing to look for in graphs is **outliers** – unusual observations that might carry large weight in analysis. Such values need to be investigated – are they errors, are they "special cases" that offer interesting insights, how dependent are results on these outliers.

Summarising associations between variables

We have already considered the situation of summarising the association between categorical variables, which leaves two possibilities to consider...

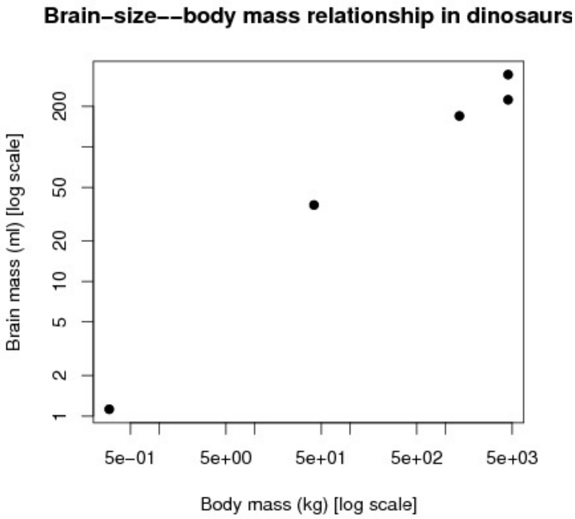
Associations between quantitative variables

Consider a pair of samples from two quantitative variables, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. We would like to understand how the x and y variables are related.

An effective graphical display of the relationship between two quantitative variables is a **scatterplot** – a plot of the y_i against the x_i.

Example

How did brain mass change as a function of body size in dinosaurs?



An effective numerical summary of the relationship between two quantitative variables is the correlation coefficient (r):

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

where $\bar{x}$ and s_x are the sample mean and standard deviation of x, similarly for y.

Results

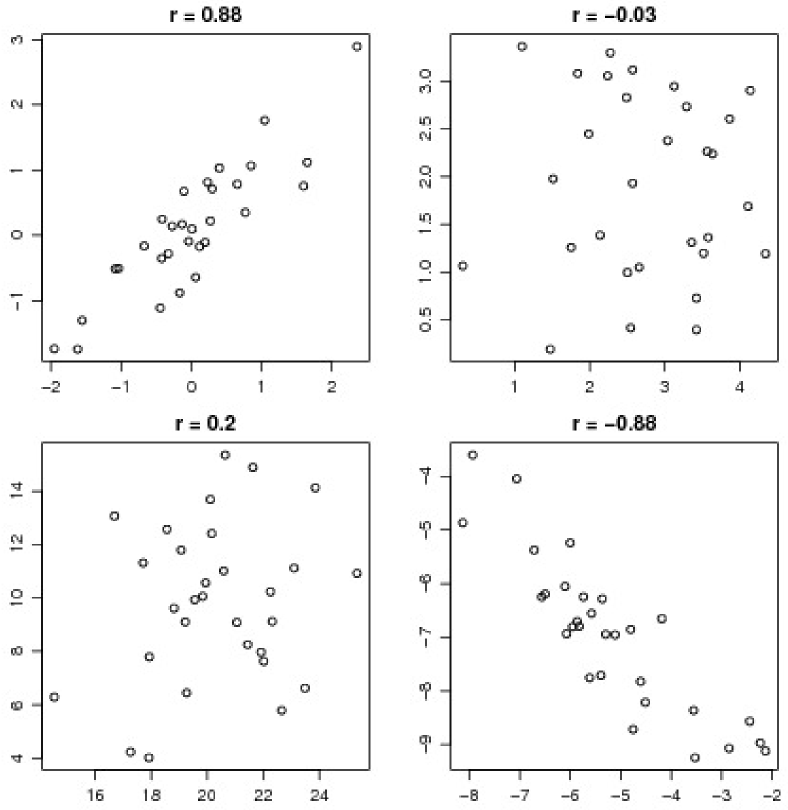
1. $|r| \leq 1$
2. $r = -1$ if and only if $y_i = a + bx_i$ for each i, for some constants a, b such that $b < 0$.
3. $r = 1$ if and only if $y_i = a + bx_i$ for each i, for some constants a, b such that $b > 0$.

Can you prove these results? (Hint: consider the square of $\frac{x_i-\bar{x}}{s_x} + \frac{y_i-\bar{y}}{s_y}$)

These results imply that r measures the strength and direction of associations between x and y:

- Strength of (linear) association – values closer to 1 or -1 suggest that the relationship is closer to a straight line
- Direction of association – values less than one suggest a decreasing relationship, values greater than one suggest an increasing relationship

Examples:



Associations between categorical and quantitative variables

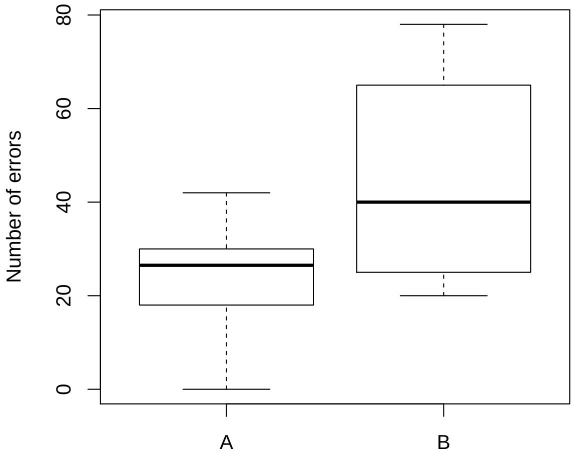
When studying whether categorical and quantitative variable are associated, an effective strategy is to summarise the quantitative variable(s) separately for each level of the categorical variable(s).

Example

Recall the guinea pig experiment – we want to explore whether there is an association between a nicotine treatment (categorical) and number of errors made by offspring (quantitative).

To summarise number of errors, we might typically use mean/sd and a boxplot. To instead look at the association between number of errors and nicotine treatment, we calculate mean/sd of number of errors for each of the two levels of treatment (nicotine and no nicotine), and construct a boxplot for each level of treatment:

	$\bar{x}$	s
Sample A	23.4	12.3
Sample B	44.3	21.5



Note that in the above example the boxplots are presented on a common axis – sometimes this is referred to as **comparative boxplots** or “side-by-side boxplots”. An advantage of boxplots over histograms is that they can be quite narrow and hence readily compared across many samples by stacking them side-by-side. Some interesting extensions are reviewed in the article “40 years of boxplots” by Hadley Wickham and Lisa Stryjewski at Rice University.

Producing numerical and graphical summaries on R

R is a free statistics package that can be used for data analysis, which will be useful to us in this course. You can download R from `cran.r-project.org`.

To open R, learn how to upload data, and for more detailed instructions on some of the below, please refer to the notes “A Brief Introduction to R” available on My eLearning.

To enter the example dataset we considered in the previous section:

```
> x=c(2, 4, 5, 7, 8, 10, 14, 17, 27, 35)
```

To calculate the sample mean, variance, standard deviation and median (respectively:)

```
> mean(x)
> var(x)
> sd(x)
> median(x)
```

To calculate the 5th, 10th and 15th percentiles (for example):

```
> quantile(x,c(0.05, 0.1, 0.15))
```

These are different to those we calculated above, because R uses a different default method of quantile estimation. Our calculations were done using a method R refers to as type=5:

```
> quantile(x,c(0.05, 0.1, 0.15), type=5)
```

To calculate a five-number summary use one of:

```
> quantile(x)
> summary(x)
```

The latter command will calculate the five-number summary together with the mean.

To construct a graph of the empirical cumulative distribution function of the dataset x:

```
> ex = ecdf(x) #To estimate the empirical distribution function
> plot(ex) #To plot it
```

To produce a histogram that bins observations by 10's:

```
> hist(x,breaks=c(0,10,20,30,40))
```

If you use hist(x) instead then R will automatically choose the bins for you, but it often makes poor choices (especially for small datasets).

To produce a kernel density estimate, use the following two lines:

```
> d = density(x)  #To estimate the kernel density
> plot(d)  # To plot it.
```

Finally, for a boxplot:

```
> boxplot(x)
```

Extended definition of Expectation

It [1] can be quite annoying to to always deal with two separate definitions for discrete and continuous random variables. This is especially the case if, as is often the case, a random variable can be a mix of the two. For this reason, in this section we look to define expectation in a unique way that does not discriminate between the discrete and continuous nature of the variable, and yet gives the same results as our previous definition. The result of this exercise is better understanding and notation when writing probability arguments.

Before we proceed we recall the definition of the Riemman integral:

[1] This section is again non-assessable material only for the interested students.

Riemman integral

The Riemman integral of the function g over the closed interval $[a, b]$ is defined as the limit of the Riemman sums:

$$\int_a^b g(x)\mathrm{d}x = \lim_{\delta_n \downarrow 0} \sum_{k=0}^{n-1} g(\tilde{x}_k)(x_{k+1} - x_k),$$

where $a = x_0 < x_1 < x_2 \cdots < x_n = b$, $\delta_n = \max_{0 \leqslant k \leqslant n}\{x_{k+1} - x_k\}$, and $\tilde{x}_k$ is any point in the interval $[x_{k+1}, x_k]$, typically $\tilde{x}_k = x_k$. We also know that the improper integral with $b = \infty$ is defined as the limit (possibly equal to ∞)

$$\int_a^\infty g(x)\mathrm{d}x = \lim_{b \uparrow \infty} \int_a^b g(x)\mathrm{d}x \ .$$

Lebesgue-Stieltjes integral

Now let $F_X(x) = F(x)$ be a cdf of some random variable $X \in \mathcal{X} \subseteq \mathbb{R}$. Suppose X is a continuous random variable (that is, $\mathbb{P}(X \in C) = 0$ for any countable set $C \subseteq \mathcal{X}$) with pdf

$$f_X(x) = f(x) = \lim_{\delta \downarrow 0} \frac{F(x+\delta) - F(x)}{\delta} \overset{\text{def}}{=} F'(x) \ .$$

Then, similarly to the Riemman integral above we can define the expectation $\mathbb{E}[g(X)]$ as follows

$$\begin{aligned}
\mathbb{E}[g(X)] &= \int_{\mathcal{X}} g(x) f(x)\mathrm{d}x \\
&\overset{\text{def}}{=} \lim_{\delta_n \downarrow 0} \sum_{k=0}^{n-1} g(\tilde{x}_k) f(\tilde{x}_k)(x_{k+1} - x_k) \\
&= \lim_{\delta_n \downarrow 0} \sum_{k=0}^{n-1} g(\tilde{x}_k)[F(x_{k+1}) - F(x_k)]
\end{aligned}$$

where in the last line we have used the Mean Value Theorem, which states that for a given continuous $F(x)$, there exists a $\tilde{x}_k \in (x_k, x_{k+1})$ such that:

$$\frac{F(x_{k+1}) - F(x_k)}{x_{k+1} - x_k} = F'(\tilde{x}_k) = f(\tilde{x}_k) \ .$$

Recall that

$$F(x_{k+1}) - F(x_k) = \mathbb{P}(x_{k+1} < X \leqslant x_k) = \mathbb{P}(\{\omega : X(\omega) \in (x_k, x_{k+1}]\}) \ .$$

Suppose now we introduce the equivalent notations

$$\begin{aligned}
\mu(x_k, x_{k+1}] &\overset{\text{def}}{=} F(x_{k+1}) - F(x_k) \\
F\{(x_k, x_{k+1}]\} &\overset{\text{def}}{=} F(x_{k+1}) - F(x_k) \\
\mathrm{d}F(x_k) &\overset{\text{def}}{=} F(x_{k+1}) - F(x_k)
\end{aligned}$$

to denote the length/size/probability (or simply 'measure') of the interval or set $(x_k, x_{k+1}]$. Just like $(x_{k+1} - x_k) \downarrow 0$ turns into the suggestive differential notation $\mathrm{d}x$ in the Riemman integral, we can then formally write

$$\mu(\mathrm{d}x) = F\{\mathrm{d}x\} = \lim_{(x_{k+1}-x_k)\downarrow 0} F(x_{k+1}) - F(x_k)$$

and the definition of expectation is then written variously in mathematics books as

$$\begin{aligned}\mathbb{E}[g(X)] &= \int_{\mathcal{X}} g(x)\mu(\mathrm{d}x) \\ &= \int_{\mathcal{X}} g(x)F\{\mathrm{d}x\} \\ &= \int_{\mathcal{X}} g(x)\mathrm{d}F(x)\end{aligned}$$

These definitions are frequently referred to as the **Lebesgue-Stieltjes** definition of an integral.

Remark 1.1 Recalling that a random variable $X(\omega)$, $\omega \in \Omega$ is a function from Ω to $\mathcal{X} \subseteq \mathbb{R}$, you may even encounter the notation

$$\mathbb{E}[g(X)] = \int_{\Omega} g(X(\omega))\mathbb{P}(\mathrm{d}\omega),$$

where $\mathbb{P}(\mathrm{d}\omega)$ is suggestive differential notation for

$$\mathbb{P}(\{\omega : X(\omega) \in (x_k, x_{k+1}]\}) = \mathbb{P}(\{\omega : X(\omega) \in \mathrm{d}x\}) .$$

You may now ask what the point of the new notation is. The answer is that the definition

$$\mathbb{E}[g(X)] \overset{\text{def}}{=} \lim_{\delta_n \downarrow 0} \sum_{k=0}^{n-1} g(\tilde{x}_k)[F(x_{k+1}) - F(x_k)]$$

works in both the continuous and discrete (or mixed!) cases. As a result, there is no need to treat the discrete or continuous cases differently by giving two separate formulas for calculating probabilities and expectations. We saw above that the definition works in the continuous case. We now show that it also works in the discrete case.

Suppose now X is a discrete random variable, that is, $\mathcal{X}$ is countable and $\mathbb{P}(X = x) > 0$ for all $x \in \mathcal{X}$. Then, letting $n \uparrow \infty$, we can choose a subsequence

$$\tilde{x}_{j_0} < \tilde{x}_{j_1} < \tilde{x}_{j_2} < \cdots$$

of $\tilde{x}_0 < \tilde{x}_1 < \tilde{x}_2 < \cdots$ to equal all the possible values in the countable set $\mathcal{X}$. By properties of the cdf F in the discrete case[2], we then have

$$F(x_{k+1}) - F(x_k) = \mathbb{P}(x_k < X \leqslant x_{k+1}) = \begin{cases} \mathbb{P}(X = \tilde{x}_k), & k \in \{j_0, j_1, \ldots\} \\ 0, & \text{otherwise} \end{cases}$$

[2] The cdf is a right-continuous step/staircase function with countable number of jumps.

Thus,

$$\begin{aligned}\mathbb{E}[g(X)] &= \lim_{\delta_n \downarrow 0} \sum_{k=0}^{n-1} g(\tilde{x}_k)[F(x_{k+1}) - F(x_k)] \\ &= \sum_{k=0}^{\infty} g(\tilde{x}_{j_k}) \mathbb{P}(X = \tilde{x}_{j_k}) \\ &= \sum_{x \in \mathcal{X}} g(x) \mathbb{P}(X = x)\end{aligned}$$

and we recover the definition of expectation in the discrete case.

Uniform versus Pointwise Convergence

Some students inquired about when we can interchange the order of differentiation and summation/integration as in the following example for computing the mean.

If X has pdf

$$f_X(x) = \mathbb{P}(X = x) = (1-p)^{x-1} \cdot p,\ x = 1, 2, \ldots;\ 0 < p < 1,$$

then

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x=1}^{\infty} x(1-p)^{x-1} p \\ &= -p \sum_{x=1}^{\infty} \frac{\mathrm{d}}{\mathrm{d}p}(1-p)^x = -p \frac{\mathrm{d}}{\mathrm{d}p} \sum_{x=1}^{\infty} (1-p)^x \\ &= -p \frac{\mathrm{d}}{\mathrm{d}p} \left(\frac{1}{p} - 1 \right) = \frac{1}{p}\end{aligned}$$

To answer this question we now give the definition of uniform convergence and then point to its consequences. Recall the following definition

Definition 1.1 (Pointwise Convergence) We say that the sequence of functions $f_n(x)$ converges to $f(x)$ pointwise for every x in its domain if for every $\varepsilon > 0$ we can find an integer $N(x)$ (possibly dependent on x) such that

$$|f_n(x) - f(x)| < \varepsilon, \qquad \text{whenever } n \geqslant N(x).$$

In contrast to pointwise convergence we also have the following concept.

Definition 1.2 (Uniform Convergence) We say that the sequence of functions $f_n(x)$ converges to $f(x)$ uniformly for every $x \in \mathcal{X}$ if for every $\varepsilon > 0$ we can find an integer N (independent on x) such that

$$|f_n(x) - f(x)| < \varepsilon, \qquad \text{whenever } n \geqslant N.$$

We can emphasize the independence of N on x by writing

$$\sup_{x \in \mathcal{X}} |f_n(x) - f(x)| < \varepsilon, \qquad \text{whenever } n \geqslant N.$$

The discrepancy value on the left-hand-side is then independent of x.

Example

Consider the continuous functions $f_n(x) = x^n$ on $\mathcal{X} = [0, 1]$. We have $f_n(x) \to 0$ for $0 \leqslant x < 1$ and $f_n(x) \to 1$ for $x = 1$. Thus, the limit depends on the value of x within $\mathcal{X}$ and we do not have uniform convergence on $\mathcal{X}$. This can also be seen from $\sup_{x \in \mathcal{X}} |f_n(x) - 0| = 1 \not< \varepsilon$.

We have the following criterion for uniform convergence.

Theorem 1.2 (Weierstrass M-test) *Suppose that we can find an upper bound to f_n such that*

$$|f_n(x)| \leqslant M_n, \qquad x \in \mathcal{X}$$

and $\sum_{n=1}^{\infty} M_n < \infty$ (that is, converges). Then,

$$S_m(x) \overset{def}{=} \sum_{n=1}^{m} f_n(x)$$

converges uniformly (and absolutely) on $x \in \mathcal{X}$ as $m \uparrow \infty$.

Example

The series $\sum_{n=1}^{\infty} (1-p)^x$ converges uniformly on $p \in (0, 1)$, because we can find a $0 < \varrho < p$ such that $(1-p)^x \leqslant (1-\varrho)^x$ for $p \in (0, 1)$ and $\sum_{x=1}^{\infty} (1-\varrho)^x = (1-\varrho)/\varrho < \infty$.

Consequences of uniform convergence

The main consequence of uniform convergence is that it allows us to exchange the order of limiting operations like integration with infinite summation or differentiation etc.

We now list these consequences without proofs.

Theorem 1.3 (Preservation of continuity) *Suppose $f_n(x)$ converges uniformly on $x \in \mathcal{X}$ to $f(x)$. Suppose further that $f_n(x)$ is continuous at $x = y \in \mathcal{X}$ for all n. Then, $f(x)$ is continuous at $x = y$. In other words, we can flip the order of limits as follows:*

$$\lim_{x \to y} \lim_{n \uparrow \infty} f_n(x) = \lim_{n \uparrow \infty} \lim_{x \to y} f_n(x) = \lim_{n \uparrow \infty} f_n(y)$$

Thus, if $f_n(x)$ are continuous functions, then so is the limiting function $f(x)$.

Theorem 1.4 (Exchanging a limit with integration) Suppose that $f_n(x)$ converges uniformly to $f(x)$ on $x \in \mathcal{X}$. Then,

$$\lim_{n \uparrow \infty} \int_{\mathcal{X}} f_n(x) \mathrm{d}x = \int_{\mathcal{X}} \underbrace{\lim_{n \uparrow \infty} f_n(x)}_{f(x)} \mathrm{d}x$$

From the last result we also have the following.

Theorem 1.5 (Exchanging a sum with integral) Suppose that

$$S_m(x) \overset{\text{def}}{=} \sum_{n=1}^{m} f_n(x)$$

converges uniformly to $S_\infty(x)$ on $x \in \mathcal{X}$ and $\int_{\mathcal{X}} f_n(x) \mathrm{d}x < \infty$ for all n. Then,

$$\int_{\mathcal{X}} \sum_{n=1}^{\infty} f_n(x) \mathrm{d}x = \sum_{n=1}^{\infty} \int_{\mathcal{X}} f_n(x) \mathrm{d}x \; .$$

This result justifies the exchange of expectation with Taylor expansion of the moment generating function of a **bounded** random variable X, in which we write

$$\begin{aligned} m_X(u) &= \mathbb{E}[\mathrm{e}^{uX}] \\ &= \mathbb{E}\left[1 + \frac{uX}{1!} + \frac{(uX)^2}{2!} + \dots\right] \\ &= 1 + \mathbb{E}[X] \cdot \frac{u}{1!} + \mathbb{E}[X^2] \cdot \frac{u^2}{2!} + \mathbb{E}[X^3] \cdot \frac{u^3}{3!} + \dots \end{aligned}$$

Theorem 1.6 (Exchanging summation with differentiation)

Let $S_m(x) = \sum_{n=1}^{m} f_n(x)$ and suppose

- *$\frac{d}{dx} S_m(x)$ converges uniformly on $x \in \mathcal{X}$;*
- *$S_m(x)$ converges.*

Then

$$\lim_{m\uparrow\infty}\frac{d}{dx}S_m(x)=\frac{d}{dx}\lim_{m\uparrow\infty}S_m(x)$$

which is the same as

$$\sum_{n=1}^{\infty}\frac{d}{dx}f_n(x)=\frac{d}{dx}\sum_{n=1}^{\infty}f_n(x)\ .$$

Example

If we let $S_m(p)=-\sum_{k=1}^m(1-p)^k$, then $\frac{\mathrm{d}}{\mathrm{d}p}S_m(p)=\sum_{k=1}^m k(1-p)^{k-1}$. To see that $\frac{\mathrm{d}}{\mathrm{d}p}S_m(p)$ converges uniformly, we recall the following result:

For a power series power series $\sum_{k=0}a_kx^k$, the radius of convergence R is defined to be $R:=\sup\{x,\ (a_kx^k)<\infty\}$ and for any $0<r<R$, the power series converges uniformly on every closed interval $[-r,r]$.

To proceed, we note that for every $p\in(0,1]$ the function $(k+1)(1-p)^{k-1}$ is bounded in k since

$$\begin{aligned}\lim_{k\to\infty}ke^{(k-1)\ln(1-p)}&=\lim_{k\to\infty}\frac{k}{e^{(k-1)\ln(1-p)}}\\ (\mathrm{L'Hopital's\,rule})&=\lim_{k\to\infty}\frac{(1-p)^{k-1}}{-\ln(1-p)}=0\end{aligned}$$

Therefore the radius convergence $R=1$ and for every $p\in[0,1)$, the sequence of partial sum $\frac{\mathrm{d}}{\mathrm{d}p}S_m(p)$ converges uniformly. Hence, we are justified in differentiating term by term:

$$-\frac{\mathrm{d}}{\mathrm{d}p}\sum_{k=1}^{\infty}(1-p)^k=\sum_{k=1}^{\infty}\frac{\mathrm{d}}{\mathrm{d}p}(1-p)^k\ .$$

Chapter 2

Common Distributions

In the previous chapter we saw that random variables may be characterized by probability functions (discrete case) and density functions (continuous case). Any non-negative function that sums to 1 is a legal probability function. Any non-negative function that integrates to 1 is a legal density function.

There are certain families of probability functions and density functions that are particularly useful in statistics. This chapter covers the most common ones.

This material is summarised concisely in Hogg *et al* (2005) Chapter 3 and in Rice (2007) Chapter 2.

Bernoulli Distribution

The Bernoulli distribution is very important in statistics, because it can be used to model response to any Bernoulli trial, defined as below:

> **Definition**
> A **Bernoulli trial** is an experiment with 2 possible outcomes. The outcomes are often labelled 'success' and 'failure'.

Example

The tossing of a coin is a Bernoulli trial. We may define:

$$\begin{aligned}\text{`success'} &= \text{heads}\\ \text{`failure'} &= \text{tails}\end{aligned}$$

or vice versa.

Some more examples of Bernoulli trials:

- Dead or alive?
- Sick?
- Flowering or not flowering?
- Sold?
- Faulty?

There are only two possible responses to any of the questions in the above example. Hence the above variables can all be modelled using the Bernoulli distribution, defined below:

Definition

For a Bernoulli trial define the random variable

$$X = \begin{cases} 1 & \text{if the trial results in success} \\ 0 & \text{otherwise} \end{cases}$$

Then X is said to have a **Bernoulli distribution**.

Result

If X is a Bernoulli random variable defined according to a Bernoulli trial with success probability $0 < p < 1$ then the probability function of X is

$$f_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases}$$

An equivalent way of writing this is $f_X(x) = p^x(1-p)^{1-x}, \quad x = 0, 1.$

Example

Consider coin-tossing as in the previous example. If the coin is fair $p = \frac{1}{2}$ and

$$f_X(x) = \left(\frac{1}{2}\right)^x \left(1 - \frac{1}{2}\right)^{1-x} = \frac{1}{2} \quad , x = 0, 1.$$

This is consistent with the notion that heads and tails are equally likely for a fair coin.

Definition

A constant like p above in a probability function or density is called a **parameter**.

The Bernoulli Distribution is a special case of the Binomial Distribution, covered in the next section.

Binomial Distribution

The **Binomial distribution** arises when several Bernoulli trials are repeated in succession.

Definition

Consider a sequence of n independent Bernoulli trials, each with success probability p. If

$$X = \text{total number of successes}$$

then X is a **Binomial** random variable with parameters n and p. A common shorthand is:

$$X \sim \text{Bin}(n, p).$$

The previous definition uses the symbol "$\sim$". In Statistics this symbol has a special meaning:

Notation

The symbol "$\sim$" is commonly used in statistics for the phrase

"is distributed as" or "has distribution".

The mathematical expression

$$Y \sim \text{Bin}(29, 0.72)$$

is usually read:

Y has a Binomial distribution with parameters $n = 29$ and $p = 0.72$.

Whenever summing the number of times we observe a particular binary outcome, across n independent trials, we have a binomial distribution.

Example

Write down a distribution that could be used to model X when X is:

1. The number of patients who survive a new type of surgery, out of 12 patients who each have 95% chance of surviving.

2. The number of patients who visit a doctor who are in fact sick, out of the 36 who visit the doctor. Assume that in general, 70% of people who visit their doctor are actually sick.

3. The number of plants that are flowering, out of 52 randomly selected plants in the wild, when the proportion of plants in the wild flowering at this time is p.

4. The number of plasma screen televisions that a store sells in a month, out of the store's total stock of 18, each with probability 0.6 of being sold.

5. The number of plasma screen televisions that are returned to the store due to faults, out of the 9 that are sold, when the return rate for this type of TV is 8%.

Note that in all of the above, we require the assumption of independence of responses across the n units in order to use the binomial distribution. This assumption is guaranteed to be satisfied if we randomly select units from some larger population (as was done in the above plant example).

Result

If $X \sim \text{Bin}(n, p)$ then its probability function is given by

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, \ldots, n.$$

This result follows from the fact that there are $\binom{n}{x}$ ways by which X can take the value x, and each of these ways has probability $p^x(1-p)^{n-x}$ of occurring.

Results

If $X \sim \text{Bin}\,(n, p)$ then

1. $\mathbb{E}(X) = np$,

2. $\text{Var}(X) = np(1-p)$,

3. $m_X(u) = (pe^u + 1 - p)^n$.

Example

Adam pushed 10 pieces of toast off a table. Seven of these landed butter side down.

- What distribution could be used to model the number of slices of toast that land butter side down? Assume that there is a 50:50 chance of each slice landing butter side down.

- What is the expected number of pieces of toast that land butter side down, and the standard deviation?
- What is the probability that exactly 7 slices land butter side down?
- Is this unusual? (Use a *tail probability* to answer this question.)

As mentioned in the previous section, the Binomial distribution generalises the Bernoulli distribution. The next result makes this explicit:

Results
X has a Bernoulli distribution with parameter p if and only if

$$X \sim \text{Bin}(1, p).$$

Geometric Distribution

The **Geometric distribution** arises when a Bernoulli trial is repeated until the first 'success'. In this case

$$X = \text{number of trials until first success}$$

and X is said to have a geometric distribution with parameter p, where p is the probability of success on each trial.

Result
If X has a Geometric distribution with parameter $0 < p < 1$ then X has probability function

$$f_X(x; p) = p(1-p)^{x-1}, x = 1, 2, \ldots$$

Results
If X has a Geometric distribution with parameter p then

1. $\mathbb{E}(X) = \frac{1}{p}$,
2. $\text{Var}(X) = \frac{1-p}{p^2}$.

Alternative definitions of the geometric distribution are possible. For example, a common definition is $X =$ number of failures before the first success. This leads to a distribution on $x = 0, 1, \ldots$, with a different mean than is given above, but the variance is unchanged.

Hypergeometric Distribution

Hypergeometric random variables arise when counting the number of binary responses, when objects are sampled independently from finite populations, and the total number of "successes" in the population is known.

Suppose that a box contains N balls, m are red and $N-m$ are black. n balls are drawn at random. Let

$$X = \text{number of red balls drawn.}$$

Then X has a **Hypergeometric distribution** with parameters N, m and n. We write

$$X \sim \mathsf{Hyp}(n, m, N)$$

Note that this can be thought of as a finite population version of the binomial distribution. Instead of assuming some constant probability p of "success" in the population, we say that there are N units in the population of which m are successes.

Result

If X has a **Hypergeometric distribution** with parameters N, m and n then its probability function is given by

$$f_X(x; N, m, n) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}} \quad 0 \leq x \leq \min(m, n).$$

Example

Lotto A machine contains 45 balls, and you select 6. Seven winning numbers are then drawn (6 main, one supplementary), and you win a major prize ($10,000+) if you pick six of the winning numbers.

What's the chance that you win a major prize from playing one game?

Let X be the number of winning numbers. X is hypergeometric with $N = 45$, $m = 6$, $n = 7$.

$$\begin{aligned}\mathbb{P}(X = x) &= f(x; 45, 6, 7) \\ &= \frac{\left(\begin{array}{c}6\\x\end{array}\right)\left(\begin{array}{c}39\\7-x\end{array}\right)}{\left(\begin{array}{c}45\\7\end{array}\right)} \text{ and} \\ \mathbb{P}(\text{win major prize}) &= \mathbb{P}(X = 6) \\ &= \frac{\binom{6}{6}\binom{39}{1}}{\binom{45}{7}}\end{aligned}$$

which is less than 1 in a million (it's 1 in 1,163,580).

Example

Write down a distribution that could be used to model X when X is:

1. The number of patients a town doctor sees who are in fact sick, when 800 people want to see a doctor, 500 of these are actually sick, and when the doctor only has time to see 32 of the 800 people (who are selected effectively at random from those that want to see the doctor).

   If X is the number of sick patients, then

   $$X \sim \mathsf{Hyp}(32, 500, 800)$$

2. The number of plants that are flowering, out of 52 randomly selected plants in the wild, from a population of 800 plants of which 650 are flowering.

   Answer:

   $$X \sim \mathsf{Hyp}(52, 650, 800)$$

3. The number of faulty plasma screen televisions that are returned to a store, when 5 of the store's 18 TV's are faulty, and 6 of the 18 TV's were sold.Answer:

   $$X \sim \mathsf{Hyp}(6, 5, 18)$$

Results

If X has a Hypergeometric distribution with parameters N, m and n then

1. $\mathbb{E}(X) = n \cdot \frac{m}{N}$,
2. $\text{Var}(X) = n \cdot \frac{m}{N}\left(1 - \frac{m}{N}\right)\left(\frac{N-n}{N-1}\right)$,

It can be shown that as N gets large, a hypergeometric distribution with parameters N, m and n approaches $Y \sim \text{Bin}(n, \frac{m}{N})$. A suggestion of this can be seen in the above formulae: $\mathbb{E}(X)$ has the form of a binomial expectation with $p = \frac{m}{N}$, and $\text{Var}(X)$ only differs from the corresponding binomial variance formula by a "finite population correction factor" $\frac{N-n}{N-1}$ which tends to one as N gets large.

Poisson Distribution

Definition

The random variable X has a **Poisson** distribution with parameter $\lambda > 0$ if its probability function is

$$f_X(x;\lambda) = \mathbb{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots.$$

A common abbreviation is

$$X \sim \text{Poisson}(\lambda).$$

The Poisson distribution often arises when the variable of interest is a *count*. For example, the number of traffic accidents in a city on any given day could be well-described by a Poisson random variable.

The Poisson is a standard distribution for the occurrence of rare events. Such events are often described by a *Poisson process*. A Poisson process is a model for the occurrence of point events in a continuum, usually a time-continuum. The occurrence or not of points in disjoint intervals is independent, with a uniform probability rate over time. If the probability rate is λ, then the number of points occurring in a time interval of length t is a random variable with a Poisson(λt) distribution.

Results

If $X \sim \text{Poisson}(\lambda)$ then

1. $\mathbb{E}(X) = \lambda$,
2. $\text{Var}(X) = \lambda$,
3. $m_X(u) = e^{\lambda(e^u - 1)}$.

Example

Suggest a distribution that could be useful for studying X in each of the following cases:

1. The number of workplace accidents in a month (when the average number of accidents is 1.4).

2. The number of people calling a help line per day.

3. The number of ATM customers overnight when a bank is closed (when the average number is 5.6).

Example

If, on average, 5 servers go offline during the day, what is the chance that no more than 1 will go offline? (assume independence of servers going offline)

Exponential Distribution

The **Exponential distribution** is the simplest common distribution for describing probability structure of *positive* random variables, such as lifetimes.

Definition

A random variable X is said to have an **exponential distribution** with parameter $\beta > 0$ if X has density function:

$$f_X(x;\beta) = \frac{1}{\beta}e^{-x/\beta} \ , \ x > 0.$$

Result

If X has an Exponential distribution with parameter β then

$$\mathbb{E}(X) = \beta \ , \ \mathrm{Var}(X) = \beta^2.$$

The exponential distribution is closely related to the Poisson distribution of the previous section. We know from previously that if a variable follows a Poisson process, counts of the number of times a particular event happens has a Poisson distribution with parameter λ. It can be shown that the *time until the next event* has an exponential distribution with parameter $\beta = 1/\lambda$.

Example

If, on average, 5 servers go offline during the day, what is the chance that no servers will go offline in the next hour? (*Hint:* Note that an hour is $\frac{1}{24}$ of a day.)

An important property of the exponential distribution is *lack of memory*: if X has an exponential distribution, then

$$\mathbb{P}(X > s + t | X > s) = \mathbb{P}(X > t)$$

In words, if the waiting time until the next event is exponential, then the waiting time until the next event is independent of the time you've already been waiting.

Note that the exponential distribution is a special case of the Gamma distribution described in a later section.

Uniform Distribution

The **uniform distribution** is the simplest common distribution for *continuous* random variables.

Definition
A continuous random variable X that can take values in the interval (a, b) with equal likelihood is said to have a **uniform distribution on (a, b)**. A common shorthand is:

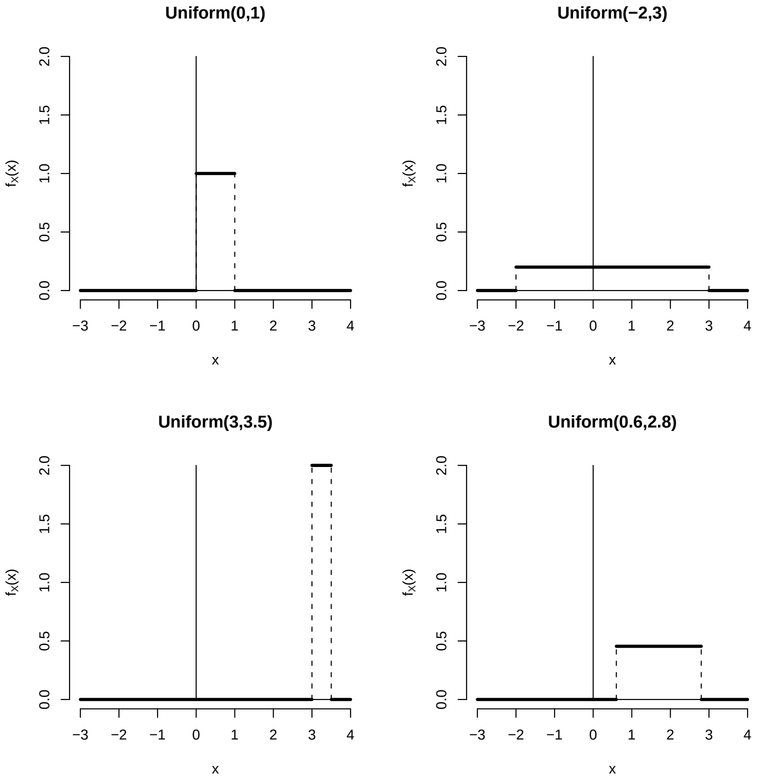
$$X \sim \text{Uniform}(a, b).$$

Definition
If $X \sim \text{Uniform}(a, b)$ then the density function of X is

$$f_X(x; a, b) = \frac{1}{b - a}, \quad a < x < b;\; a < b.$$

Note that $f_X(x; a, b)$ is simply a constant function over the interval (a, b), and zero otherwise.

The following figure shows four different uniform density functions.



Results

If $X \sim \text{Uniform}(a, b)$ then

1. $\mathbb{E}(X) = (a + b)/2$,
2. $\text{Var}(X) = (b - a)^2/12$,
3. $m_X(u) = \frac{e^{bu} - e^{au}}{(b-a)u}$.

Note that there is also a discrete version of the uniform distribution, useful for modelling the outcome of an event that has k equally likely outcomes (such as the roll of a die). This has different formulae for its expectation and variance than the continuous case does, which can be derived from first principles.

Special Functions Arising In Statistics

There are three more special distributions that we will consider in this chapter – the normal, gamma and beta distributions. But before discussing these distributions, we will need to define some *special functions* that are closely related to these distributions.

The Gamma Function

The *Gamma function* is essentially an extension of the factorial function (e.g. 4!=24) to general real numbers.

Definition

The **Gamma function** at $x \in \mathbb{R}$ is given by

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

Results

Some basic results for the Gamma function are

1. $\Gamma(x) = (x-1)\Gamma(x-1)$
2. $\Gamma(n) = (n-1)! \quad n = 1, 2, 3, \ldots$
3. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

The following result follows from (2.) of the above results. It is a useful result in a variety of statistical applications.

Result

If m is a non-negative integer then

$$\int_0^\infty x^m e^{-x}\, dx = m!$$

Example

$f_X(x) = \frac{1}{120} x^5 e^{-x} \quad x > 0$

What is $\mathbb{E}(X)$?

Find the answer using the gamma function.

$$\mathbb{E}(X) = \int_{-\infty}^\infty x f_X(x)\, dx = \int_0^\infty \tfrac{1}{120} x^6 e^{-x}\, dx = \tfrac{6!}{120} = 6.$$

The Beta Function

Definition

The **Beta function** at $x, y \in \mathbb{R}$ is given by

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt.$$

Result

For all $x, y \in \mathbb{R}$,

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

Example

$f_X(x) = 168x^2(1-x)^5\,dx, \quad 0 < x < 1.$

What is $\mathbb{E}(X^2)$? Use the beta function to find the answer.

$$\begin{aligned}\mathbb{E}(X^2) &= \int_0^1 x^2 f_X(x)\,dx = \int_0^1 168x^4(1-x)^5\,dx = 168 \times B(5,6)\\ &= 168 \times \frac{\Gamma(5)\Gamma(6)}{\Gamma(11)} = 168 \times \frac{4!5!}{10!} = \frac{168}{1260} = \frac{2}{15}.\end{aligned}$$

The Digamma and Trigamma Functions

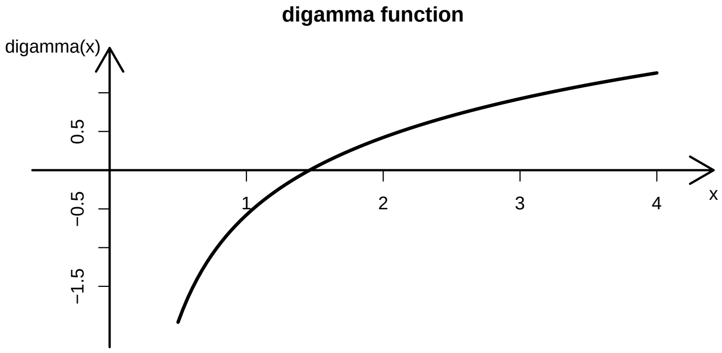
The *digamma* and *trigamma functions* will be required later in the notes, but we will give their definition here since they are closely related to the Gamma function:

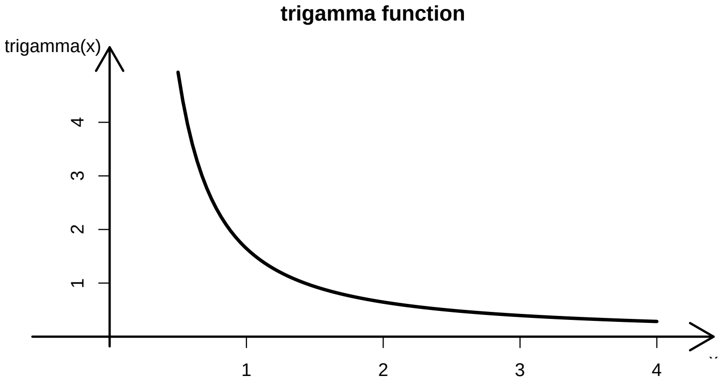
Definition

For all $x \in \mathbb{R}$,

$$\begin{aligned}\text{digamma}(x) &= \frac{d}{dx}\ln\{\Gamma(x)\},\\ \text{trigamma}(x) &= \frac{d^2}{dx^2}\ln\{\Gamma(x)\}.\end{aligned}$$

Unlike some other mathematical functions (such as sin, $\tan^{-1}$, $\ln_{10}, \ldots$), the digamma and trigamma functions are not available on ordinary hand-held calculators. However, they are just another set of special functions and are available in mathematics and statistics computer software such as `Maple`, `Matlab` and `R`. The following figure (constructed using `R` software) displays them graphically.





The Φ Function and its Inverse

The final special function we will consider is routinely denoted by the ‘capital phi’ symbol Φ. It is defined as follows:

Definition

For all $x \in \mathbb{R}$,

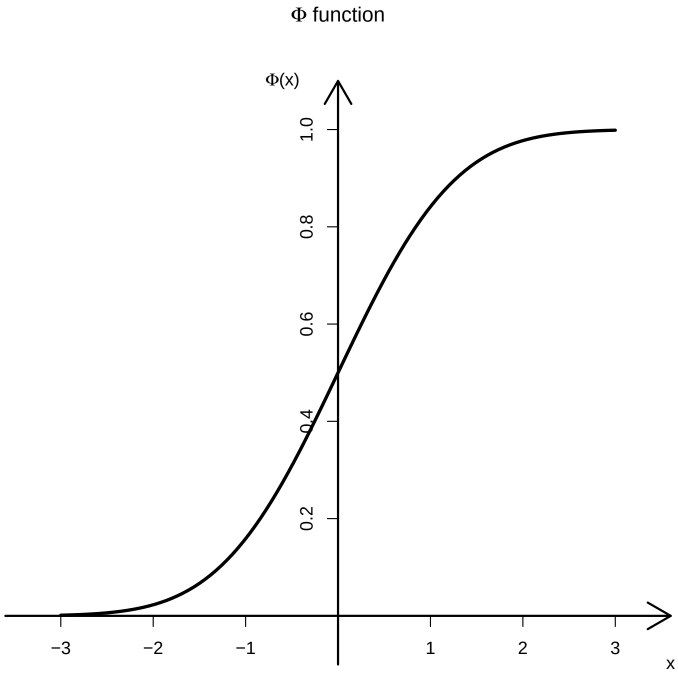
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt.$$

Note that $\Phi(x)$ cannot be simplified any further than in the above expression since $e^{-t^2/2}$ does not have a closed-form ‘anti-derivative’.

This function gives the cumulative distribution function of the standard normal distribution, considered in the following section.

Result
The Φ function has the following properties:

1. $\lim\limits_{x \to -\infty} \Phi(x) = 0$,
2. $\lim\limits_{x \to \infty} \Phi(x) = 1$,
3. $\Phi(0) = \frac{1}{2}$,
4. Φ is monotonically increasing over $\mathbb{R}$.



It follows from the previous result that the inverse of Φ, $\Phi^{-1}(x)$, is well-defined for all $0 < x < 1$. Examples are

$$\Phi^{-1}\left(\frac{1}{2}\right) = 0 \quad \text{and} \quad \Phi^{-1}(0.975) = 1.95996\ldots \simeq 1.96.$$

We will see later (in chapters 8-10) that the Φ^{-1} function plays a particularly important role in statistical inference.

Normal Distribution

A particularly important family of continuous random variables is those following the **normal distribution**:

Definition

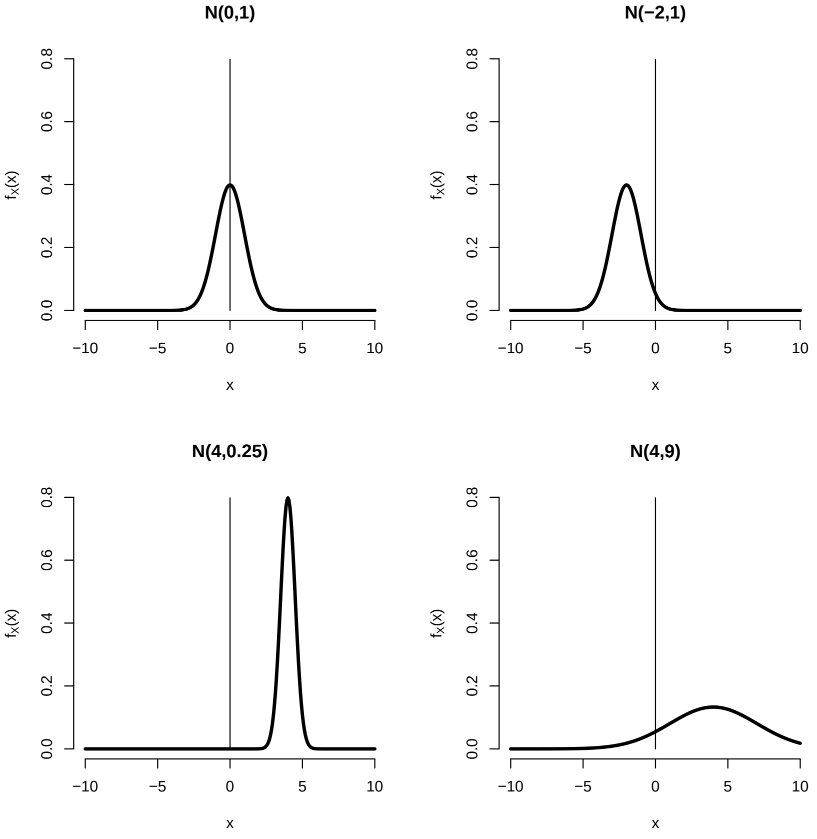
The random variable X is said to have a **normal distribution** with parameters μ and σ^2 (where $-\infty < \mu < \infty$ and $\sigma^2 > 0$) if X has density function

$$f_X(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

A common shorthand is

$$X \sim N(\mu, \sigma^2).$$

Normal density functions are symmetric “bell-shaped” curves symmetric about μ. The following figure shows four different normal density functions.



The normal distribution is the most important distribution in statistical practice. One reason is that many real-life random variables are observed to have normal, or nearly normal distributions. An explanation of this striking empirical phenomenon is given by the central limit theorem, which will be discussed in chapter 5. An even more powerful influence of this distribution in Statistics is through another application of the central limit theorem, stating that the approximate distribution of sample sums and averages is normal, even if the original observations come from non-normal distributions. This topic is described in chapters 7-9.

Results
If $X \sim N(\mu, \sigma^2)$ then

1. $\mathbb{E}(X) = \mu$,
2. $\text{Var}(X) = \sigma^2$,
3. $m_X(u) = e^{\mu u + \frac{1}{2}\sigma^2 u^2}$.

The special case of $\mu = 0$ and $\sigma^2 = 1$ is known as the **standard normal distribution**. It is common to use the letter Z to denote standard normal random variables:

$$Z \sim N(0, 1).$$

The standard normal density function is

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

Computing Normal Distribution Probabilities

Consider the problem:

$$\mathbb{P}(Z \leq 0.47) = \int_{-\infty}^{0.47} \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx.$$

The standard normal density function does not have a closed form anti-derivative and cannot be solved in the usual way. Note, however:

Result
If $Z \sim N(0, 1)$ then

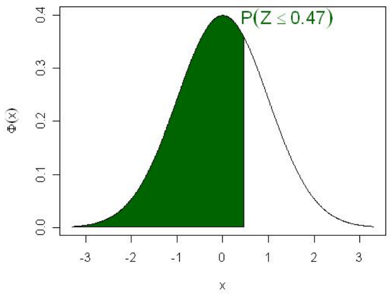
$$\mathbb{P}(Z \leq x) = F_Z(x) = \Phi(x).$$

In other words, the Φ function is the cumulative distribution function of the $N(0, 1)$ random variable.

This result means that probabilities concerning $Z \sim N(0, 1)$ can be computed whenever the function Φ is available. Tables for Φ are available in the back of the Course Pack (and on UNSW Blackboard). This can be used, for example, to show that:

$$\mathbb{P}(Z \leq 0.47) = \Phi(0.47) \simeq 0.6808.$$

The shaded area in the following figure corresponds to $\Phi(0.47)$.



Some other examples:

$$\mathbb{P}(Z \leq 1) = \Phi(1) \simeq 0.8413 \quad \text{and} \quad \mathbb{P}(Z \leq 1.54) = \Phi(1.54) \simeq 0.9382.$$

For finding a probability such as $\mathbb{P}(Z > 0.81)$, we need to work with the complement $\mathbb{P}(Z \leq 0.81)$:

$$\begin{aligned} \mathbb{P}(Z > 0.81) &= 1 - \mathbb{P}(Z \leq 0.81) \\ &= 1 - \Phi(0.81) \\ &\simeq 1 - 0.7910 = 0.2090 \end{aligned}$$

How about probabilities concerning non-standard normal random variables? For example, how do we find

$$\mathbb{P}(X \leq 12) \quad \text{where} \quad X \sim N(10, 9)?$$

Result

If $X \sim N(\mu, \sigma^2)$ then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

The preceding result is an example of an operation called *standardisation.* The variable Z is said to have been *standardised.* Other forms of transformation are discussed in Chapter 5. Thanks to standardisation, we can use tables of the Φ distribution to calculate any probability to do with a normally distributed random variable, as in the examples below.

Example

Find $\mathbb{P}(X \leq 12)$ where $X \sim N(10, 9)$.

$$\begin{aligned}
\mathbb{P}(X \leq 12) &= \mathbb{P}\left(\frac{X-10}{3} \leq \frac{12-10}{3}\right) \\
&= \mathbb{P}(Z \leq 0.67) \quad \text{where } Z \sim N(0,1) \\
&= \Phi(0.67) \simeq 0.7486
\end{aligned}$$

Example

The distribution of young men's heights is approximately normally distributed with mean 174 cm and standard deviation 6.4 cm.

1. What percentage of these men are taller than six foot (182.9 cm)?

2. What's the chance that a randomly selected young man is 170-something cm tall?

3. Find a range of heights that contains 95% of young men.

Gamma Distribution

Definition

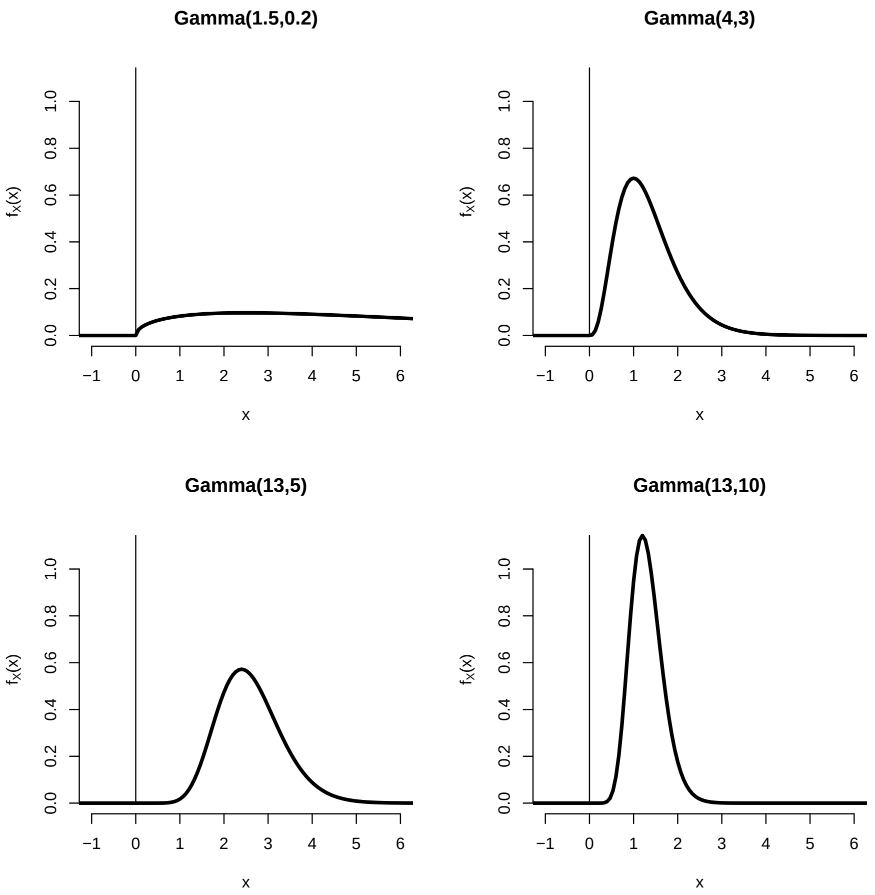
A random variable X is said to have a **Gamma distribution** with parameters α and β (where $\alpha, \beta > 0$) if X has density function:

$$f_X(x; \alpha, \beta) = \frac{e^{-x/\beta} x^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}}, \quad x > 0.$$

A common shorthand is:

$$X \sim \text{Gamma}(\alpha, \beta).$$

Gamma density functions are skewed curves on the positive half-line. The following figure shows four different Gamma density functions.



Results

If $X \sim \text{Gamma}(\alpha, \beta)$ then

1. $\mathbb{E}(X) = \alpha\beta$,
2. $\text{Var}(X) = \alpha\beta^2$,
3. $m_X(u) = \left(\frac{1}{1-\beta u}\right)^{\alpha}, \quad u < 1/\beta.$

The Gamma distribution generalises the Exponential distribution as below:

Result

X has an Exponential distribution if and only if

$$X \sim \text{Gamma}(1, \beta)$$

Beta Distribution

For completeness, we conclude with the **Beta distribution**. This generalises the Uniform(0,1) distribution, which can be thought of as a beta distribution with $a = b = 1$.

Definition

A random variable X is said to have a **Beta distribution** with parameters $\alpha, \beta > 0$ if its density function is

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

Results

If X has a Beta distribution with parameters α and β then

1. $\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta}$,
2. $\mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$,

Quantile-quantile plots of data

Consider the situation in which we have a sample of size n from some unknown random variable $\{x_1, x_2, \dots, x_n\}$ and we want to check if these data appear to come from a random variable with cumulative distribution function $F_X(x)$. This can be achieved using a *quantile-quantile plot* (sometimes called a *Q-Q plot*) as defined below.

Quantile-quantile plots

To check how well the sample $\{x_1, x_2, \dots, x_n\}$ approximates the distribution with cdf $F_X(x)$, plot the n sample quantiles against the corresponding quantiles of $F_X(x)$. That is, plot the points

$$\left(F^{-1}(p),\, x_{(k)} \right) \qquad \text{where } p = \frac{k-0.5}{n}, \quad \text{for all } k \in \{1, 2, \dots n\}$$

If the data come from the distribution $F_X(x)$, then the points will show no systematic departure from the one-to-one line.

According to the above definition, we need to know the exact cdf $F_X(x)$ to construct a quantile-quantile plot. However, for a location-scale family of distributions, a family that does not change its essential shape as its parameters change, we can construct the quantile-quantile plot using an arbitrary choice of parameters. In this case, we only need to check for systematic departures from a straight line rather than from the one-to-one line when assessing goodness-of-fit. This is the most common application of quantile-quantile plots. It allows us to see how well data approximates a whole family of location-scale distributions, without requiring any knowledge of what the values of the parameters are.

Among the special distributions introduced in this chapter are several examples of location or scale families.

Example

Recall the example dataset from Chapter 1:

$$2 \quad 4 \quad 5 \quad 7 \quad 8 \quad 10 \quad 14 \quad 17 \quad 27 \quad 35$$

Use a quantile-quantile plot to assess how well these data approximate a normal distribution.

There are ten values in this dataset, so the values of p we will use to plot the data are $\frac{k-0.5}{10}$ for all $k \in \{1, 2, \ldots, 10\}$, that is, for all

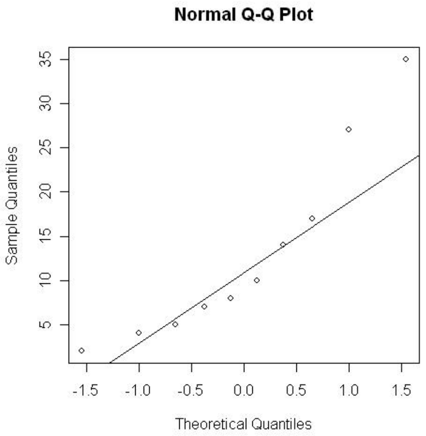
$$p \in \{\ 0.05,\ 0.15,\ 0.25,\ 0.35,\ 0.45,\ 0.55,\ 0.65,\ 0.75,\ 0.85,\ 0.95\ \}$$

We want to find the quantiles corresponding to these values of p from a normal distribution, and compare those to the observed values. We have tables for the standard normal distribution, so we will use these to obtain quantiles. That is, we will plot the $x_{(k)}$ against $\Phi^{-1}(p)$ for the ten values of p displayed above.

Using tables, we can show that the corresponding standard normal quantiles are

$$\{\ -1.64,\ -1.04,\ -0.67,\ -0.39,\ -0.13,\ 0.13,\ 0.39,\ 0.67,\ 1.04,\ 1.64\ \}$$

and so we plot these values against our ordered example dataset. This results in the following plot:



This plot does not follow a straight line – it has a systematic concave-up curve, so the data are clearly not normally distributed. In fact, because the curve is concave-up, the data are right-skewed (since the larger values in the dataset are much larger than expected for a normal distribution).

Special distributions on R

R has some useful functions for working with special distributions, as below.

Consider the normally distributed random variable with mean 2 and standard deviation 3, $X \sim N(2, 9)$.

To take a random sample of size 20 from X and store it in x:

```
> x = rnorm(20,2,3)
```

To find the density function of X at the value $x = 4$, *i.e.* to find $f_X(4)$:

```
> dnorm(4,2,3)
```

To find $F_x(4) = \mathbb{P}(X \leqslant 4)$:

```
> pnorm(4,2,3)
```

To find the 95th percentile of X:

```
> qnorm(0.95,2,3)
```

We can perform similar calculations for all of the special distributions considered in this chapter, using the following functions:

Binomial	rbinom, dbinom, pbinom, qbinom
Geometric	rgeom, dgeom, pgeom, qgeom
Hypergeometric	rhyper, dhyper, phyper, qhyper
Poisson	rpois, dpois, ppois, qpois
Exponential	rexp, dexp, pexp, qexp
Uniform	runif, dunif, punif, qunif
Normal	rnorm, dnorm, pnorm, qnorm
Gamma	rgamma, dgamma, pgamma, qgamma
Beta	rbeta, dbeta, pbeta, qbeta

Obtaining a normal quantile-quantile plot on R

Normal quantile-quantile plots can be calculated easily on R using the function qqnorm:

```
> x=c(2, 4, 5, 7, 8, 10, 14, 17, 27, 35)
> qqnorm(x)
```

To add a line to this plot to help with interpretation (as for the plot given above):

```
> qqline(x)
```

A quantile-quantile plot for any other family of distributions needs to be calculated manually, *i.e.* find the quantiles and store them in q, then use plot(q,sort(x)) to plot the data x, sorted from smallest to largest, against the quantiles q.

Chapter 3

Bivariate Distributions

Observations are often taken in pairs, leading to bivariate observations (X, Y), *i.e.* observations of two variables measured on the same subjects. For example, (height, weight) can be measured on people, as can (age, blood pressure), (gender, promotion) for employees, (sales, price) for supermarket products...

Often we are interested in exploring the nature of the relationship between two variables that have been measured on the same set of subjects. In this chapter we develop a notation for the study of relationship between two variables, and explore some key concepts.

For further reading, consider Hogg *et al* (2005) Chapter 2 or Rice (2007) Chapter 3 and Chapter 3 of Kroese and Chan (2014).

Joint Probability Function and Density Function

Definition
If X and Y are discrete random variables then the **joint probability function** of X and Y is

$$f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y),$$

the probability that $X = x$ **and** $Y = y$.

Why study joint probabilities?

Recall that if two variables are dependent, then

$$\mathbb{P}(A \cap B) \neq \mathbb{P}(A) \cdot \mathbb{P}(B)$$

In the context of two discrete random variables X and Y,

$$\mathbb{P}(X = x, Y = y) \neq \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$$

So when we want to calculate $\mathbb{P}(X = x, Y = y)$, or any joint probability involving both X and Y, we cannot find it using the probability functions of X and Y, which give us $\mathbb{P}(X = x)$ and $\mathbb{P}(Y = y)$. We instead need to know the joint probability function $f_{X,Y}(x, y)$.

Example

Suppose that X and Y have joint probability function as tabulated below.

$f_{X,Y}(x, y)$

		y		
		-1	0	1
x	0	1/8	1/4	1/8
	1	1/8	1/16	1/16
	2	1/16	1/16	1/8

Find $\mathbb{P}(X = 0 \text{ and } Y = 0)$. Show that $\mathbb{P}(X = 0 \text{ and } Y = 0) \neq \mathbb{P}(X = 0) \cdot \mathbb{P}(Y = 0)$.

$$\mathbb{P}(X = 0 \text{ and } Y = -1) = 1/8.$$

Note that

$$\mathbb{P}(X = 0) \cdot \mathbb{P}(Y = -1) = \frac{1}{2} \times \frac{5}{16} = \frac{5}{32}$$

Note that this means that we wouldn't have been able to get the correct answer without looking at the *joint* probability distribution of X and Y.

Example

Let X = number of successes in the first of two Bernoulli trials each with success probability p and let Y = total number of successes in the two trials. Then, for example,

$$f_{X,Y}(1, 1) = \mathbb{P}(X = 1 \text{ and } Y = 1) = \mathbb{P}(X = 1)\mathbb{P}(Y = 1) = p(1 - p).$$

$f_{X,Y}(x,y)$

		y		
		0	1	2
x	0	$(1-p)^2$	$p(1-p)$	0
	1	0	$p(1-p)$	p^2

Joint density functions

Definition

The **joint density function** of continuous random variables X and Y a bivariate function $f_{X,Y}$ with the property

$$\int\int_A f_{X,Y}(x,y)\,dx\,dy = \mathbb{P}((X,Y)\in A)$$

any (measurable) subset A of $\mathbb{R}^2$.

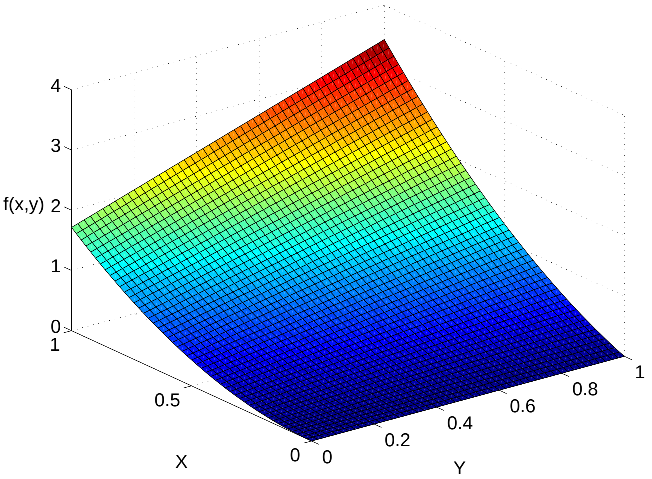
For two continuous random variables, X and Y, probabilities have the following geometrical interpretation: $f_{X,Y}$ is a surface over the plane $\mathbb{R}^2$ and probabilities over subsets $A \subseteq \mathbb{R}^2$ correspond to the *volume* under $f_{X,Y}$ over A.

For example, if

$$f(x,y) = \frac{12}{7}(x^2+xy) \qquad \text{for } x,y \in (0,1)$$

then the joint density function looks like this:

c]



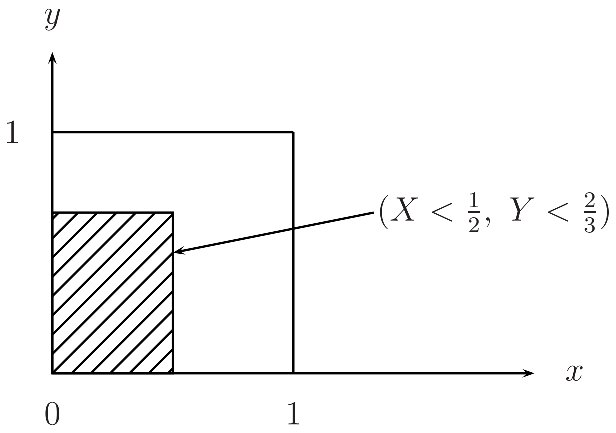
Example

(X, Y) have joint density function

$$f(x,y) = \frac{12}{7}(x^2 + xy) \qquad \text{for } x, y \in (0,1)$$

Find $\mathbb{P}(X < \frac{1}{2}, Y < \frac{2}{3})$.

First, what region of the (X, Y) plane do we want to integrate over?



We want to integrate $f_{X,Y}(x,y)$ over this region:

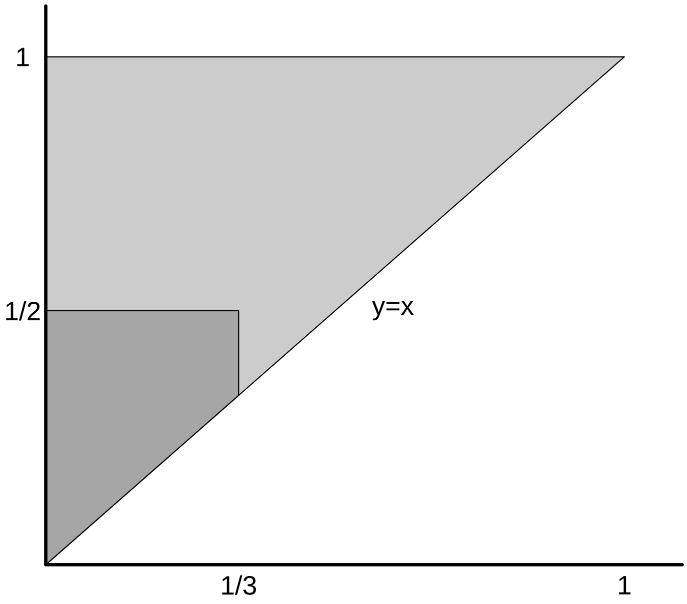
$$\begin{aligned}
\mathbb{P}(X < 1/2, Y < 2/3) &= \int_0^{1/2}\int_0^{2/3} f_{X,Y}(x,y)\,dy\,dx \\
&= \int_0^{1/2}\int_0^{2/3} \frac{12}{7}(x^2 + xy)\,dy\,dx \\
&= \frac{12}{7}\int_0^{1/2}\left[x^2 y + \frac{xy^2}{2}\right]_0^{2/3} dx \\
&= \frac{12}{7}\int_0^{1/2} x^2\frac{2}{3} + \frac{x}{2}\left(\frac{2}{3}\right)^2 dx \\
&= \frac{8}{7}\int_0^{1/2} x^2 + \frac{x}{3}\,dx \\
&= \frac{8}{7}\left[\frac{x^3}{3} + \frac{x^2}{6}\right]_0^{1/2} \\
&= \frac{8}{7} \times \left(\frac{1}{8\cdot 3} + \frac{1}{4\cdot 6}\right) \\
&= \frac{2}{21}
\end{aligned}$$

Example*

$$f_{X,Y}(x,y) = 2(x+y), \quad 0 < x < y, \ 0 < y < 1.$$

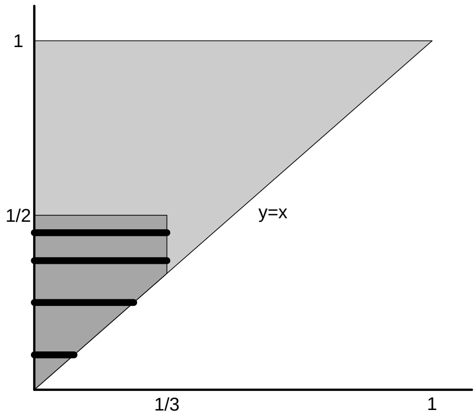
What is $\mathbb{P}(X < 1/3, Y < 1/2)$?

This is a challenging question because the limits for x are a function of y. This means that the domain of this density function has a triangular shape, and the area over which we need to integrate $f_{X,Y}(x,y)$ is trapezoidal, as shown (in dark grey) in the following figure:



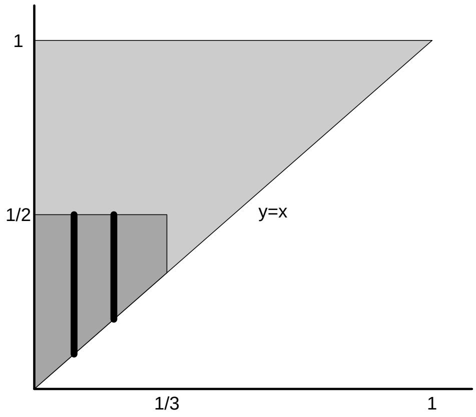
If we use horizontal strips as shown in the next figure then the integral needs to be broken up into two terms; corresponding to the triangle and the rectangle components of the trapezium.

$$\mathbb{P}(X < 1/3, Y < 1/2) = \int_0^{1/3}\int_0^y 2(x+y)\,dx\,dy + \int_{1/3}^{1/2}\int_0^{1/3} 2(x+y)\,dx\,dy = \frac{11}{108}.$$



If we use *vertical* strips as shown in the following figure then the integral can be done in one piece as follows:

$$\mathbb{P}(X < 1/3, Y < 1/2) = \int_0^{1/3} \int_x^{1/2} 2(x+y)\, dy\, dx = \frac{11}{108}$$



Other results for $f_{X,Y}(x,y)$

Many of the definitions and results for random variables, considered in chapter 1, generalise directly to the bivariate case. We consider some of these in this section.

Essentially, all that changes from chapter 1 to here is that instead of doing a single summation or integral, we now do a double summation or double integral, because there are now two variables under consideration.

In each of the following cases, think of what the univariate (one-variable) version of the result is, as given in chapter 1.

Firstly, we have the following property for joint probability functions.

Result

If X and Y are discrete random variables then

$$\sum_{\text{all } x}\sum_{\text{all } y} f_{X,Y}(x,y) = 1.$$

If X and Y are continuous random variables then

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1.$$

Next, we will consider a definition of the joint cumulative distribution function (cdf).

Definition

The **joint cdf** of X and Y is

$$\begin{aligned} F_{X,Y}(x,y) &= \mathbb{P}(X \le x, Y \le y) \\ &= \begin{cases} \displaystyle\sum_{u \leqslant x}\sum_{v \leqslant y} \mathbb{P}(X = u, Y = v) & (X \text{ discrete}) \\ \\ \displaystyle\int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)\,du\,dv & (X \text{ continuous}) \end{cases} \end{aligned}$$

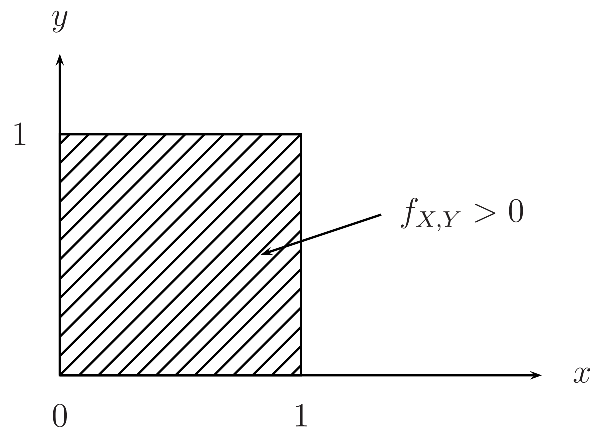
Example

consider again

$$f_{X,Y}(x,y) = \frac{12}{7}(x^2 + xy), \quad 0 < x < 1,\ 0 < y < 1.$$

Find $F_{X,Y}(x,y)$.

First, what region of the (X,Y) plane do we want to integrate over?



For $0 < x < 1$ and $0 < y < 1$,

$$\begin{aligned} F_{X,Y}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)\,dudv &= \frac{12}{7}\int_0^y\int_0^x u^2+uv\,du\,dv \\ &= \frac{12}{7}\int_0^y \left[\frac{u^3}{3}+\frac{u^2v}{2}\right]_0^x dv \\ &= \frac{12}{7}\int_0^y \left(\frac{x^3}{3}+\frac{x^2v}{2}\right) dv \\ &= \frac{12}{7}\left[\frac{x^3v}{3}+\frac{x^2v^2}{4}\right]_0^y = \frac{12}{7}\left(\frac{x^3y}{3}+\frac{x^2y^2}{4}\right). \end{aligned}$$

Thus, for example, $\mathbb{P}(X < \frac{1}{2}, Y < \frac{1}{3}) = F_{X,Y}\left(\frac{1}{2},\frac{1}{3}\right) = \frac{12}{7}\left(\frac{1}{72}+\frac{1}{144}\right) = \frac{3}{84}$

Finally, we will consider expectations. We define the expectation of some joint function of X and Y, $g(X,Y)$, as below.

Result

If g is any function of X and Y,

$$\mathbb{E}\{g(X,Y)\} = \begin{cases} \sum_{\text{all } x}\sum_{\text{all } y} g(x,y)\,\mathbb{P}(X=x,Y=y) & \text{(discrete)} \\ \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\,f_{X,Y}(x,y)\,dx\,dy & \text{(continuous)} \end{cases}$$

Note that this formula has the same form as that of $\mathbb{E}\{g(X)\}$ from chapter 1.

Example

$f_{X,Y}(x,y)$

		y		
		0	1	2
x	0	0.1	0.2	0.2
	1	0.2	0.2	a

where a is a constant.

For the above bivariate distribution,

1. Find a, if $f_{X,Y}(x,y)$ is a joint probability function.
2. Find $F_{X,Y}(1,1)$.
3. Find $\mathbb{E}(XY)$.

Marginal probability/density functions

Result

If X and Y are discrete, then $f_X(x)$ and $f_Y(y)$ can be calculated from $f_{X,Y}(x,y)$ as follows:

$$f_X(x) = \sum_{\text{all } y} f_{X,Y}(x,y)$$

$$f_Y(y) = \sum_{\text{all } x} f_{X,Y}(x,y)$$

$f_X(x)$ is sometimes referred to as the **marginal probability function** of X.

Example

			y		
		0	1	2	$f_X(x)$
x	0	0.1	0.2	0.2	
	1	0.2	0.2	0.1	
	$f_Y(y)$				

Find the marginal distributions of X and Y.

$$f_X(0) = \mathbb{P}(X=0) = \mathbb{P}(X=0, Y=0) + \mathbb{P}(X=0, Y=1)$$

$$+\mathbb{P}(X=0, Y=2) = 0.5.$$

Thus $f_X(0) = \mathbb{P}(X=0) = \sum\limits_{\text{all } y} \mathbb{P}(X=0, Y=y)$.

And in fact for any value x,

$$\begin{aligned} f_X(x) = \mathbb{P}(X=x) &= \sum_{\text{all } y} \mathbb{P}(X=x, Y=y) \\ &= \sum_{\text{all } y} f_{X,Y}(x,y). \end{aligned}$$

Result

If X and Y are continuous, then $f_X(x)$ and $f_Y(y)$ can be calculated from $f_{X,Y}(x,y)$ as follows:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$$

$f_X(x)$ is sometimes referred to as the **marginal density function** of X.

Example

$$f_{X,Y}(x,y) = \frac{12}{7}(x^2+xy),\ 0<x<1,\ 0<y<1.$$

Find $f_X(x)$ and $f_Y(y)$.

$$f_X(x) = \frac{12}{7}\int_0^1 x^2+xy\ dy = \frac{12}{7}\left(x^2+\frac{x}{2}\right),\ 0<x<1.$$

$$f_Y(y) = \frac{12}{7}\int_0^1 x^2+xy\ dx = \frac{12}{7}\left(\frac{1}{3}+\frac{y}{2}\right),\ 0<y<1.$$

Example

$$f_{X,Y}(x,y) = 2(x+y),\ 0<x<y,\ 0<y<1.$$

Find $f_X(x)$ and $f_Y(y)$.

$$f_X(x) = \int_x^1 2(x+y)dy = 1+2x-3x^2,\ 0<x<1.$$

$$f_Y(y) = \int_0^y 2(x+y)dx = 3y^2,\ 0<y<1.$$

Conditional Probability and Density Functions

Definition

If X and Y are discrete, the **conditional probability function** of X given $Y=y$ is

$$f_{X|Y}(x|y) = \mathbb{P}(X=x|Y=y) = \frac{\mathbb{P}(X=x,Y=y)}{\mathbb{P}(Y=y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

Similarly,

$$f_{Y|X}(y|x) = \mathbb{P}(Y=y|X=x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Note that this is simply an application of the definition of conditional probability from chapter 0.

Definition

If X and Y are continuous, the **conditional density function** of X given $Y = y$ is

$$f_{X|Y}(x|Y = y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Similarly,

$$f_{Y|X}(y|X = x) = \frac{f_{X,Y}(x, y)}{f_X(x)}.$$

Often we write $f_{Y|X}(y|x)$ as shorthand for $f_{Y|X}(y|X = x)$.

Example

			y		
		0	1	2	$f_X(x)$
x	0	0.1	0.2	0.2	0.5
	1	0.2	0.2	0.1	0.5
$f_Y(y)$		0.3	0.4	0.3	1

Find $f_{X|Y}(x|2)$ and $f_{Y|X}(y|0)$.

$$\mathbb{P}(X = 0|Y = 0) = \frac{0.1}{0.3} = \frac{1}{3}$$

$$\mathbb{P}(X = 1|Y = 0) = \frac{0.2}{0.3} = \frac{2}{3}$$

x	0	1
$\mathbb{P}(X = x\|Y = 0)$	$\frac{1}{3}$	$\frac{2}{3}$

Also,

y	0	1	2
$\mathbb{P}(Y = y\|X = 0)$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{2}{5}$

Example

$$f_{X,Y}(x,y) = \frac{12}{7}\left(x^2 + xy\right),\ 0 < x < 1,\ 0 < y < 1$$

Find $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$.

$$f_{X|Y}(x|y) = \frac{x^2 + xy}{\frac{1}{3} + \frac{y}{2}},\ 0 < x < 1$$

$$f_{Y|X}(y|x) = \frac{x^2 + xy}{x^2 + \frac{x}{2}},\ 0 < y < 1.$$

Let X and Y be continuous. For a given value of x, $f_{Y|X}(y|x)$ is an ordinary density function and has the usual properties such as:

Result

If X and Y are continuous then

$$\mathbb{P}(a \le Y \le b | X = x) = \int_a^b f_{Y|X}(y|x)\,dy.$$

Similar results apply to discrete X and Y:

Result

If X and Y are discrete then

$$\mathbb{P}(Y \in A | X = x) = \sum_{y \in A} f_{Y|X}(y|X = x).$$

Conditional Expected Value and Variance

The **conditional expected value** of X given $Y = y$ is

$$\mathbb{E}(X|Y = y) = \begin{cases} \displaystyle\sum_{\text{all } x} x\,\mathbb{P}(X = x|Y = y) & \text{if } X \text{ is discrete} \\ \\ \displaystyle\int_{-\infty}^{\infty} x\, f_{X|Y}(x|y)\,dx & \text{if } X \text{ is continuous} \end{cases}$$

Similarly,

$$\mathbb{E}(Y|X=x) = \begin{cases} \displaystyle\sum_{\text{all } y} y\,\mathbb{P}(Y=y|X=x) & \text{if } Y \text{ is discrete} \\[2em] \displaystyle\int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)\,dy & \text{if } Y \text{ is continuous} \end{cases}$$

Note that this can be thought of as an application of the definition of $\mathbb{E}(X)$ from chapter 1.

Example

Recall the example on page 82, in which

			y		
		0	1	2	$f_X(x)$
	0	0.1	0.2	0.2	0.5
x					
	1	0.2	0.2	0.1	0.5
	$f_Y(y)$	0.3	0.4	0.3	1

Find $\mathbb{E}(X|Y=2)$ and $\mathbb{E}(Y|X=0)$.

Recall the example from previous section for which it was established that:

x	0	1
$\mathbb{P}(X=x \mid Y=0)$	$\frac{1}{3}$	$\frac{2}{3}$

Then

$$\mathbb{E}(X|Y=0) = 0\left(\frac{1}{3}\right) + 1\left(\frac{2}{3}\right) = \frac{2}{3}$$

Example

Recall the example on page 83, in which

$$f_{X,Y}(x,y) = \frac{12}{7}\left(x^2+xy\right), \;\; 0<x<1, \;\; 0<y<1$$

Find $\mathbb{E}(X|Y)$.

Recall the example from previous section for which it was established that:

$$f_{X|Y}(x|y) = \frac{x^2+xy}{\frac{1}{3}+\frac{y}{2}}, \;\; 0<x<1$$

Then for any y,

$$\mathbb{E}(X|Y=y) = \int_0^1 x \cdot \frac{x^2+xy}{\frac{1}{3}+\frac{y}{2}}\,dx = \frac{3+4y}{4+6y}.$$

The **conditional variance** of X given $Y=y$ is

$$\text{Var}(X|Y=y) = \mathbb{E}(X^2|Y=y) - \{\mathbb{E}(X|Y=y)\}^2$$

where

$$\mathbb{E}(X^2|Y=y) = \begin{cases} \displaystyle\sum_{\text{all } x} x^2 P(X=x|Y=y) \\ \\ \displaystyle\int_{-\infty}^{\infty} x^2 f_{X|Y}(x|y)dx. \end{cases}$$

Similarly for $\text{Var}(Y|X=x)$.

Note that these definitions can be thought of as an application of the definitions of $\text{Var}(X)$ from chapter 1.

Example

Find $\text{Var}(X|Y=2)$ for the discrete data example on page 82.

Recall the example from previous section for which it was established that:

x	0	1
$\mathbb{P}(X=x|Y=0)$	$\frac{1}{3}$	$\frac{2}{3}$

Then: $\mathbb{E}(X^2|Y=0) = \frac{2}{3}$, $\text{Var}(X|Y=y) = \frac{2}{9}$

Example

Find $\text{Var}(X|Y)$ for the continuous data example on page 83.

Recall the example from previous section for which it was established that:

$$f_{X|Y}(x|y) = \frac{x+y}{y+\frac{1}{2}},\ 0<x<1$$

Then $\mathbb{E}(X^2|Y=y) = \displaystyle\int_0^1 x^2 \frac{(x+y)}{y+\frac{1}{2}} dx = \frac{4y+3}{12y+6}$,

$$\mathrm{Var}(X|Y=y) = \frac{6y^2+6y+1}{18(2y+1)^2}.$$

Independent Random Variables

Definition

Random variables X and Y are **independent** if and only if for all x, y

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

Result

Random variables X and Y are **independent** if and only if for all x, y

$$f_{Y|X}(y|x) = f_Y(y)$$

or

$$f_{X|Y}(x|y) = f_X(x).$$

This result allows an interpretation that conforms with the 'every day' meaning of the word *independent.* If X and Y are independent, then the probability structure of Y is unaffected by the 'knowledge' that X takes on some value x (and vice versa).

Result

If X and Y are independent,

$$F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y).$$

Example

			y		
		-1	0	1	$f_X(x)$
	0	0.01	0.02	0.07	0.1
x	1	0.04	0.13	0.33	0.5
	2	0.05	0.05	0.3	0.4
	$f_Y(y)$	0.1	0.2	0.7	1

Are X and Y independent?

X and Y are independent if

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) \text{ for all } x,y;$$

that is, every entry in the body of the table equals the product of the corresponding row and column totals.

X and Y are not independent if $f_{X,Y}(x,y) \neq f_X(x) f_Y(y)$ for at least one pair of values x, y.

Thus X and Y are not independent in this case since, for example

$$0.04 = \mathbb{P}(X = 1, Y = -1)$$

$$\neq \mathbb{P}(X = 1)\mathbb{P}(Y = -1) = (.5)(.1) = 0.05.$$

Example

X and Y have joint probability function $f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$

$$= p^2(1-p)^{x+y}, x = 0, 1, \ldots, y = 0, 1, \ldots;\ 0 < p < 1.$$

Are X and Y independent?

$$f_X(x) = \mathbb{P}(X = x) = \sum_{\text{all } y} f_{X,Y}(x,y)$$

$$= \sum_{y=0}^{\infty} p^2(1-p)^{x+y} = p(1-p)^x, x = 0, 1, \ldots$$

Similarly, $f_Y(y) = \sum_{x=0}^{\infty} p^2(1-p)^{x+y}$
$= p(1-p)^y, y = 0, 1, \ldots$

$$\therefore f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y) \text{ for all } x, y,$$

so X and Y are independent.

Example

X and Y have joint density

$$f_{X,Y}(x,y) = 6xy^2,\ 0 < x < 1,\ 0 < y < 1.$$

Are X and Y independent?

$$f_X(x) = \int_0^1 6xy^2 dy = 2x,\ 0 < x < 1$$

$$f_Y(y) = \int_0^1 6xy^2 dx = 3y^2,\ 0 < y < 1.$$

$\therefore f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$, so X and Y are independent.

Example

X and Y have joint density

$$f_{X,Y}(x,y) = 10xy^2,\ 0 < x < y,\ 0 < y < 1.$$

Are X and Y independent? Note that X cannot be larger than Y, so the value X takes depends on Y (and vice versa).

$$f_X(x) = \int_x^1 10xy^2 dy = \frac{10x}{3}(1 - x^3),\ 0 < x < 1$$

$$f_Y(y) = \int_0^y 10xy^2 dx = 5y^4,\ 0 < y < 1.$$

Thus $f_{X,Y}(x,y) \neq f_X(x) \cdot f_Y(y)$, so X and Y are dependent.

Result

If X and Y are independent,

$$\mathbb{E}(XY) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$$

and more generally, for any functions $g(X)$ and $h(Y)$,

$$\mathbb{E}\{g(X) \cdot h(Y)\} = \mathbb{E}\{g(X)\} \cdot \mathbb{E}\{h(Y)\}$$

Covariance and Correlation

Covariance

Definition

The **covariance** of X and Y is

$$\text{Cov}(X,Y) = \mathbb{E}\{(X-\mu_X)(Y-\mu_Y)\}$$

where $\mu_X = \mathbb{E}(X)$, $\mu_Y = \mathbb{E}(Y)$.

$\text{Cov}(X,Y)$ measures not only how X and Y vary about their means, but also how they vary together *linearly.* $\text{Cov}(X,Y) > 0$ if X and Y are positively associated, *i.e.* if X is likely to be large when Y is large and X is likely to be small when Y is small. If X and Y are negatively associated, $\text{Cov}(X,Y) < 0$.

Results

1. $\text{Cov}(X,X) = \text{Var}(X)$
2. $\text{Cov}(X,Y) = \mathbb{E}(XY) - \mu_X\mu_Y$.

Result

If X and Y are independent then $\text{Cov}(X,Y) = 0$.

Proof (continuous case):

$$\begin{aligned}
\mathbb{E}(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)\cdot f_Y(y)dxdy \\
&= \int_{-\infty}^{\infty} xf_X(x)dx \cdot \int_{-\infty}^{\infty} yf_Y(y)dy \\
&= \mathbb{E}(X)\cdot\mathbb{E}(Y).
\end{aligned}$$

Results

1. For arbitrary constants a, b,

   $$\text{Var}(aX+bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X,Y).$$

   Hence

2. $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X,Y)$,
3. $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$ when X and Y are independent,
4. $\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y)$ when X and Y are independent.

Correlation

Definition

The **Correlation** between X and Y is

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\cdot\text{Var}(Y)}}.$$

Corr(X,Y) measures the strength of the linear association between X and Y.

Definition

If Corr$(X,Y)=0$, then X and Y are said to be **uncorrelated**.

Independent random variables are uncorrelated, but uncorrelated variables are not necessarily independent; for example, if X has a distribution which is symmetric about zero and $Y = X^2$,

$$\mathbb{E}(XY) = \mathbb{E}(X^3) = 0 \text{ and } \mathbb{E}(X) = 0, \text{ so}$$

Cov$(X,Y)=0$ and Corr$(X,Y)=0$, but since $Y = X^2$, X and Y are dependent.

Results

1. $|\text{Corr}(X,Y)| \leq 1$
2. Corr$(X,Y) = -1$ if and only if $\mathbb{P}(Y = a+bX) = 1$ for some constants a, b such that $b<0$.
3. Corr$(X,Y) = 1$ if and only if $\mathbb{P}(Y = a+bX) = 1$ for some constants a, b such that $b>0$.

Proof: 1. Let $\varrho = \text{Corr}(X,Y)$.

$$0 \leq \text{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right)$$

where $\sigma_X^2 = \text{Var}(X)$ and $\sigma_Y^2 = \text{Var}(Y)$

$$\begin{aligned} &= 2 + 2\ \text{Cov}\left(\frac{X}{\sigma_X}\ ,\ \frac{Y}{\sigma_Y}\right) \\ &= 2(1+\varrho) \\ \therefore \varrho &\geq -1. \end{aligned}$$

Also, $0 \leq \text{Var}\ \left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 2(1-\varrho)$, so $\varrho \leq 1$.

2. If $\varrho = -1$, Var $\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) = 2(1+\varrho) = 0$.

This means that $\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}$ is a constant, *i.e.* $\mathbb{P}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y} = c\right) = 1$ for some constant c.

But

$$\frac{X}{\sigma_X}+\frac{Y}{\sigma_Y}=c \Longleftrightarrow Y=\frac{-X\sigma_Y}{\sigma_X}+c\,\sigma_Y$$

so $\mathbb{P}(Y=a+bX)=1$ for some constants $a=c\sigma_Y$ and $b=\frac{-\sigma_Y}{\sigma_X}<0$.

3. Similarly, for $\varrho=1, \mathbb{P}(Y=a+bX)=1$ for some constant a and $b=\frac{\sigma_Y}{\sigma_X}>0$.

The Bivariate Normal Distribution

The most commonly used special type of bivariate distribution is the bivariate normal.

X and Y have the bivariate normal distribution if

$f_{X,Y}(x,y)=$

$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\varrho^2}}\exp\left\{-\frac{1}{2(1-\varrho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2-2\varrho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)+\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}$$

$$-\infty<x<\infty,\ -\infty<y<\infty;\ -\infty<\mu_X<\infty,\ -\infty<\mu_Y<\infty,$$

$$\sigma_X>0,\ \sigma_Y>0,\ -1<\varrho<1.$$

Result

1. $X\sim N(\mu_X,\sigma_X^2)$
2. $Y\sim N(\mu_Y,\sigma_Y^2)$.

Result

$\varrho=\mathrm{Corr}(X,Y)$

Another useful result is that the conditional distributions $Y|X=x$ and $X|Y=y$ are normal, as shown in a Chapter 3 tutorial exercise.

Visualisation of the Bivariate Normal Density Function

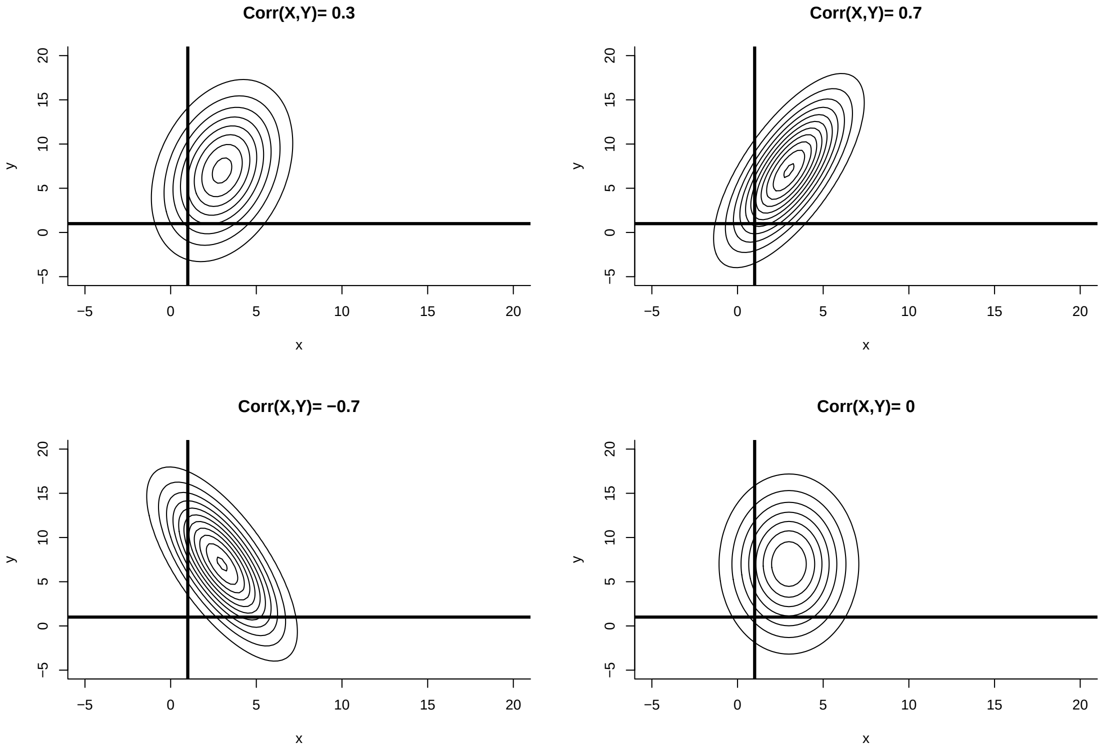
The bivariate normal density is a bivariate function $f_{X,Y}(x,y)$ with elliptical contours. The following figure provides contour plots of the bivariate normal density for

$$\mu_X=3,\ \mu_Y=7,\ \sigma_X=2,\ \sigma_Y=5$$

in all cases, but with

$$\varrho=\mathrm{Corr}(X,Y)$$

taking 4 different values: 0.3, 0.7, -0.7 and 0.



These, respectively, correspond to

- moderate positive correlation between X and Y,
- strong positive correlation between X and Y,
- strong negative correlation between X and Y,
- X and Y uncorrelated.

Result

If X and Y are uncorrelated jointly normal variables, then X and Y are independent.

Note that is a special exception to the rule given on page 90 in the case of normal random variables. You will prove this result in a Chapter 3 exercise.

Extension to n Random Variables

All of the definitions and results in this chapter extend to the case of more than two random variables. For the general case of n random variables we now give some of the most fundamental of these.

Definition

The **joint probability function** of $X_1, \dots, X_n$ is

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$$

Definition

The **joint cdf** in both the discrete and continuous cases is

$$F_{X_1,\dots,X_n}(x_1, \dots, x_n) = \mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n).$$

Definition

The **joint density** of $X_1, \dots, X_n$ is

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) = \frac{\partial^n}{\partial x_1 \dots \partial x_n} F_{X_1,\dots,X_n}(x_1, \dots, x_n).$$

Definition

$X_1, \dots, X_n$ are **independent** if

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) = f_{X_1}(x_1) \dots f_{X_n}(x_n)$$

or

$$F_{X_1,\dots,X_n}(x_1, \dots, x_n) = F_{X_1}(x_1) \dots F_{X_n}(x_n).$$

Chapter 4

Transformations

Definition
If X is a random variable, $Y = h(X)$ for some function h is a **transformation** of X.

Often we wish to transform a variable, for one of a few different reasons. Often we wish to transform a random variable or a set of random variables in order to calculate some summary statistic of interest to us. An example we will consider in later chapters is the average of a sample (the "sample mean"): we observe n random variables $X_1, X_2, \ldots, X_n$ and we are interested in their average $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. This can be considered as a transformation of the $X_1, X_2, \ldots, X_n$. Another reason we might want to transform a random variable is because the variable we measure is not the one of primary interest. An example of this introduced in a previous chapter is the measurement of tree circumference X, when what we are primary interested in is basal area, $\pi\left(\frac{X}{2\pi}\right)^2 = \frac{X^2}{4\pi}$. Yet another example, often encountered in applied statistics, is to transform a random variable in order to improve its properties (*e.g.* to reduce skewness).

In this chapter, we will learn key methods for deriving the probability or density function of a transformation of a random variable.

For further reading, consider Hogg *et al* (2005) sections 1.6.1, 1.7.1 and 2.2, or Rice (2007) sections 2.3 and 3.6.

First, we will start with the discrete case.

Result
For discrete X,

$$f_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}\{h(X) = y\} = \sum_{x:h(x)=y} \mathbb{P}(X = x).$$

Example

x	-1	0	1	2
$f_X(x)$	$1/8$	$1/4$	$1/2$	$1/8$

Find $f_Y(y)$ where $Y = X^2$.

y	0	1	4
$f_Y(y)$	$1/4$	$5/8$	$1/8$

since, for example $\mathbb{P}(Y = 0) = \mathbb{P}(X = 0) = 1/4$,

$$\mathbb{P}(Y = 1) = \mathbb{P}(X = -1) + \mathbb{P}(X = 1) = 5/8.$$

Now we will consider the continuous case. The density function of a transformed continuous variable is simple to determine when the transformation is *monotonic*.

Definition
Let h be a real-valued function defined over the set A where A is a subset of $\mathbb{R}$. Then h is a **monotonic** transformation if h is either strictly increasing or strictly decreasing over A.

Example

For the following functions, classify them as monotonic or non-monotonic, for the domain $x \in \mathbb{R}$. If the function is non-monotonic, specify a domain over which this function will be monotonic.

1. $h_1(x) = x^2$

2. $h_2(x) = 7(x - 4)^3$

3. $h_3(x) = \sin(x)$

Consider the function $h_1(x) = x^2$. Then h_1 is not monotonic over the whole real line $\mathbb{R}$. However h_1 is monotonic over the positive half-line $\{x : x > 0\}$. It is also monotonic over the negative half-line $\{x : x < 0\}$

The function $h_2(x) = 7(x - 4)^3$ is monotonic over $\mathbb{R}$.

The function $h_3(x) = \sin(x)$ is (very!) non-monotonic over $\mathbb{R}$, but it is monotonic over the interval $(\pi/4, \pi/2)$, for example.

Result

For continuous X, if h is monotonic over the set $\{x : f_X(x) > 0\}$ then

$$\begin{aligned} f_Y(y) &= f_X(x)\left|\frac{dx}{dy}\right| \\ &= f_X\left\{h^{-1}(y)\right\}\left|\frac{dx}{dy}\right| \end{aligned}$$

for y such that $f_X\left\{h^{-1}(y)\right\} > 0$.

Proof: $F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}\left\{h(X) \le y\right\}$

$$= \begin{cases} \mathbb{P}\left\{X \le h^{-1}(y)\right\} = F_X\left\{h^{-1}(y)\right\} & \text{if } h\uparrow \\ \\ \mathbb{P}\left\{X \ge h^{-1}(y)\right\} = 1 - F_X\left\{h^{-1}(y)\right\} & \text{if } h\downarrow \end{cases}$$

$$\therefore f_Y(y) = \begin{cases} f_X\left\{h^{-1}(y)\right\}\frac{dh^{-1}(y)}{dy} = f_X(x)\frac{dx}{dy} & \text{if } h\uparrow \\ \\ -f_X\left\{h^{-1}(y)\right\}\frac{dh^{-1}(y)}{dy} = -f_X(x)\frac{dx}{dy} & \text{if } h\downarrow \end{cases}$$

$$\text{Now } \frac{dy}{dx}\begin{cases} > 0 & \text{if } h\uparrow \\ < 0 & \text{if } h\downarrow \end{cases} \text{ and so}$$

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|.$$

Example

$f_X(x) = 3x^2,\ 0 < x < 1.$

Find $f_Y(y)$ where $Y = 2X - 1$.

$x = \frac{y+1}{2}, \frac{dx}{dy} = \frac{1}{2}$ and

$$\left|\frac{1}{2}\right| = \frac{3}{8}(y+1)^2,\ -1 < y < 1.$$

Example

Let $X \sim \text{exponential}(\beta)$, *i.e.* $f_X(x) = \frac{1}{\beta}e^{-x/\beta}, x > 0; \beta > 0.$

Find $f_Y(y)$ where $Y = X/\beta$. Let $Y = \lambda X$. Then $x = \frac{y}{\lambda}$, $\frac{dx}{dy} = \frac{1}{\lambda}$ and $f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = \lambda \mathrm{e}^{-y}\left|\frac{1}{\lambda}\right| = \mathrm{e}^{-y}, y > 0.$

Note that this result shows that any exponential variable can be transformed to the exponential distribution with parameter 1, by dividing by the parameter β. Hence the exponential distribution with parameter β is a *scale family* as defined in Chapter 1, and β can be interpreted as a *scale parameter* (*i.e.* a parameter that does not change the shape of the distribution, it just changes the scale of X).

Example

$f_X(x) = \sqrt{\frac{2}{\pi}} e^{-x^2/2}, x > 0.$

Let $Y = \frac{X^2}{2}$. Find the density function of Y.

Then $x = \sqrt{2y}$, $\frac{dx}{dy} = (2y)^{-\frac{1}{2}}$ and
$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = \sqrt{\frac{2}{\pi}} e^{-y}|(2y)^{-\frac{1}{2}}| = \frac{e^{-y}y^{-\frac{1}{2}}}{\sqrt{\pi}}$, $y > 0.$
Note: $\Gamma\left(\frac{1}{2}\right) = \int_0^\infty e^{-y}y^{-\frac{1}{2}}dy = \sqrt{\pi}$.

Linear Transformations

The simplest monotonic transformations are linear transformations:

$$h(x) = ax + b \quad \text{for } a \neq 0.$$

A common example of when a linear transformation is required is when applying a *change of scale*, *e.g.* changing temperature measurements from degrees Fahrenheit to degrees Celsius, $Y = \frac{5}{9}(X - 32)$. Another example is when we are interested in calculating a summary statistic (such as the sample mean) which can be written as a linear transformation of a set of observed random variables.

The linear transformation leads to the following special case.

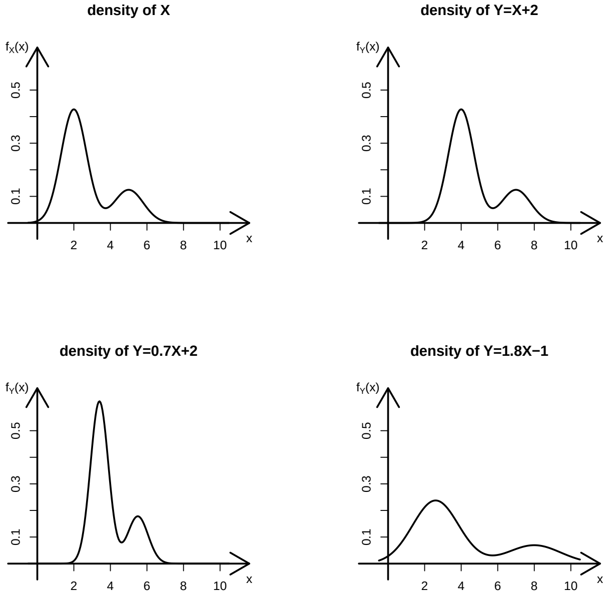
Result

For a continuous random variable X, if $Y = aX+b$ is a linear transformation of X with $a \neq 0$, then

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

for all y such that $f_X\left(\frac{y-b}{a}\right) > 0$.

This result is the formula for densities of location and scale families, discussed in chapter 1. It follows directly from the result in the previous section for general monotonic transformations. The implication is that linear transformations only change the *location* and *scale* of a density function. They do not change its shape. The following figure illustrates this for a bimodal density function f_X, and different choices of the linear transformation parameters.


density of X

density of Y=X+2

density of Y=0.7X+2

density of Y=1.8X−1

Probability Integral Transformation

Probability Integral Transformation

If X has density $f_X(x)$ and cdf $F_X(x)$, then $Y = F_X(X) \sim$ Uniform $(0, 1)$.

Proof: Let $Y = F_X(X)$. Then $y = F_X(x)$ and $\frac{dy}{dx} = f_X(x)$.

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = f_X(x) \cdot \frac{1}{f_X(x)} = 1,\ 0 < y < 1.$$

Alternatively, from first principles,

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}\left\{F_X(X) \leq y\right\} \\ &= \mathbb{P}\left\{X \leq F_X^{-1}(y)\right\} \text{ since } F_X \uparrow \\ &= F_X\left\{F_X^{-1}(y)\right\} = y. \\ \therefore \ & f_Y(y) = 1, \ 0 < y < 1. \end{aligned}$$

Example

$$f_X(x) = e^{-x}, x > 0$$

Find the distribution of Y where $Y = 1 - e^{-X}$.

$F_X(x) = 1 - e^{-x}, \ x > 0$ and so $Y = 1 - e^{-X} \sim$ Uniform $(0, 1)$.

The Probability Integral Transformation allows for easy simulation of random variables from any distribution for which the inverse cdf F_X^{-1} is easily computed.

A computer can be used to generate U where $U \sim$ Uniform $(0, 1)$. If we require an observation X where X has cdf F_X, then

$$U = F_X(X) \sim \text{Uniform } (0,1) \iff X = F_X^{-1}(U).$$

Hence we have the following result.

A universal random number generator

To generate a sample from any distribution X:

1. Use a computer to generate a random sample u from $U \sim$ Uniform $(0, 1)$.
2. Calculate your random sample from X as $x = F^{-1}(u)$.

A notable exception is when the cdf cannot be written in closed form, as is the case for the normal distibution, for example.

Example

We require an observation from the distribution with cdf

$$F_X(x) = 1 - e^{-x}, x > 0.$$

Explain how we can generate a random observation from this distribution. Let $y = F_X(x)$. Then $x = F_X^{-1}(y)$ and $y = 1 - e^{-x} \implies x = -\ln(1-y)$, so $F_X^{-1}(y) = -\ln(1-y)$. Thus, if $X = F_X^{-1}(U) = -\ln(1-U)$ where $U \sim$ Uniform $(0,1)$, then X has cdf $F_X(x) = 1 - e^{-x}, x > 0$.

Bivariate Transformations

If X and Y have joint density $f_{X,Y}(x,y)$ and U is a function of X and Y, we can find the density of U by calculating $F_U(u) = \mathbb{P}(U \leq u)$ and differentiating.

Example

$f_{X,Y}(x,y) = 1,\ 0 < x < 1,\ 0 < y < 1.$

Let $U = X + Y$. Find $f_U(u)$.

$F_U(u) = \mathbb{P}(X + Y \leq u)$

$$= \begin{cases} \displaystyle\int_0^u \int_0^{u-y} 1 dx dy & , 0 < u < 1 \\ \displaystyle 1 - \int_{u-1}^1 \int_{u-y}^1 1 dx dy & , 1 < u < 2 \end{cases}$$

$$= \begin{cases} \frac{u^2}{2} & , 0 < u < 1 \\ 2u - \frac{u^2}{2} - 1 & , 1 < u < 2. \end{cases}$$

Thus

$$f_U(u) = \begin{cases} u & , 0 < u < 1 \\ 2 - u & , 1 < u < 2. \end{cases}$$

An alternative way to find the density of U is by way of a bivariate transformation, using the following result:

Result

If U and V are functions of continuous random variables X and Y, then

$$f_{U,V}(u,v) = f_{X,Y}(x,y) \cdot |J|$$

where

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$$

is a determinant called the **Jacobian** of the transformation.

The full specification of $f_{U,V}(u,v)$ requires that the range of (u,v) values corresponding to those (x,y) for which $f_{X,Y}(x,y) > 0$ is determined.

The Jacobian has an interpretation similar to, although more detailed than, that of the factor $\left|\frac{dx}{dy}\right|$ in the univariate density transformation formula. The transformations U, V transform a small rectangle, with area $\delta x \delta y$ in the x, y plane, into a small parallelogram with area $\delta x \delta y / J$ in the u, v plane.

To find $f_U(u)$ by bivariate transformation:

1. Define some bivariate transformation to (U,V).
2. Find $f_{U,V}(u,v)$.
3. We want the marginal distribution of U. So now find $f_U(u) = \int_{-\infty}^{\infty} f_{U,V}(u,v)dv$.

Using a bivariate transformation to find the distribution of U is often more convenient than deriving it via the cumulative distribution function. Using the cdf requires double integration, which we can avoid when we use a bivariate transformation.

Example*

X, Y independent Uniform (0,1) variables.

$$f_{X,Y}(x,y) = 1, \;\; 0 < x < 1, \;\; 0 < y < 1.$$

Let $U = X + Y$ and $V = Y$. Use a bivariate transformation to (U,V) to find the density function of U.

First, note that $X = U - V$ and $Y = V$.

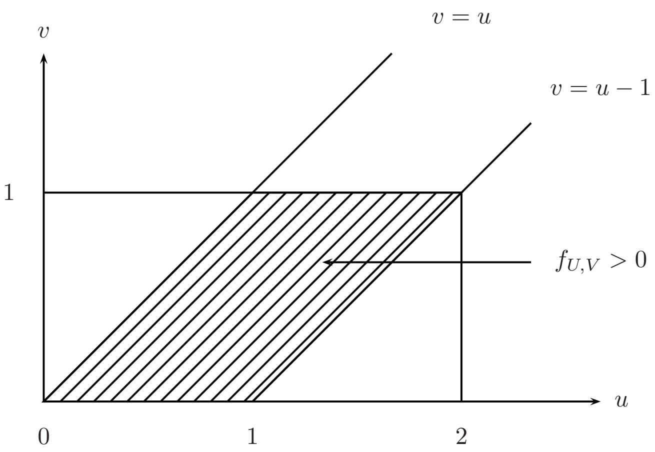
$$x = u - v, \;\; y = v \text{ gives } \frac{\partial x}{\partial u} = 1, \;\; \frac{\partial x}{\partial v} = -1,$$

$$\frac{\partial y}{\partial u} = 0, \;\; \frac{\partial y}{\partial v} = 1 \text{ and } J = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 1.$$

Now $0 < x < 1 \Leftrightarrow 0 < u - v < 1$ and $0 < y < 1 \Leftrightarrow 0 < v < 1$

$$\therefore f_{U,V}(u,v) = f_{X,Y}(x,y)|J|$$

$$= 1, \; v < u < 1 + v, \; 0 < v < 1.$$



$$f_U(u) = \begin{cases} \displaystyle\int_0^u 1 \cdot dv = u & , 0 < u < 1 \\ \displaystyle\int_{u-1}^1 1 \cdot dv = 2 - u & , 1 < u < 2. \end{cases}$$

Note that $V = Y$, so $f_V(v) = 1, 0 < v < 1$.

Example

$$f_{X,Y}(x,y) = 3y \;, \; 0 < x < y < 1.$$

Let $U = X + Y, \; V = Y - X$. Use a bivariate transformation to (U, V) to find the density function of F.

$$X = \frac{U - V}{2} \;,_1 Y = \frac{U + V}{2}.$$

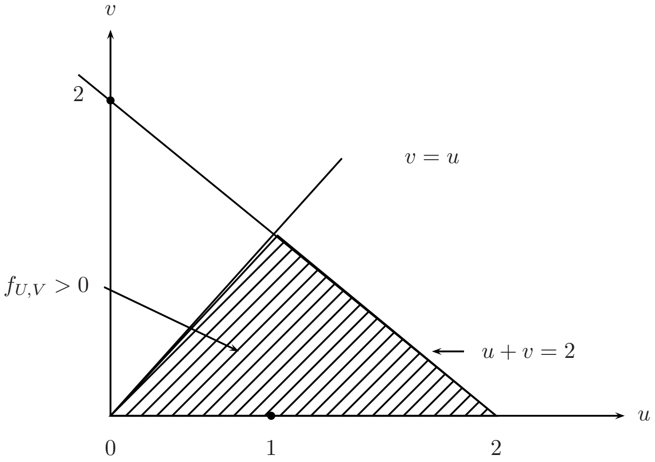
Now $x = \frac{u-v}{2}$, $y = \frac{u+v}{2}$ give

$$\frac{\partial x}{\partial u} = \frac{1}{2} \;, \; \frac{\partial x}{\partial v} = -\frac{1}{2} \;, \; \frac{\partial y}{\partial u} = \frac{1}{2} \;, \; \frac{\partial y}{\partial v} = \frac{1}{2}.$$

$$\therefore J = \begin{vmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{vmatrix} = \frac{1}{2} \text{ and}$$

$$0 < x < y \iff 0 < \frac{u-v}{2} < \frac{u+v}{2} \iff 0 < v < u,$$

$$y < 1 \iff \frac{u+v}{2} < 1 \iff u+v < 2.$$



$$\therefore f_{U,V}(u,v) = \frac{3(u+v)}{2}|\frac{1}{2}| = \frac{3}{4}(u+v), \quad 0 < v < u,\ u+v < 2.$$

$$\text{For } 0 < u < 1,\ f_U(u) = \int_0^u \frac{3}{4}(u+v)dv = \frac{9u^2}{8}.$$

$$\text{For } 1 < u < 2, f_U(u) = \int_0^{2-u} \frac{3}{4}(u+v)dv = \frac{3}{2} - \frac{3u^2}{8}.$$

$$\text{As an aside, note that } f_V(v) = \int_v^{2-v} \frac{3}{4}(u+v)du = \frac{3}{2}(1-v^2), 0 < v < 1.$$

Example

The lifetimes X and Y of two brands of components of a system are independent with

1

$$f_X(x) = xe^{-x}, x > 0 \text{ and } f_Y(y) = e^{-y}, y > 0.$$

The relative efficiency of the components, is measured as $U = \frac{Y}{X}$. Find the density function of the relative efficiency, using a bivariate transformation.

$$f_{X,Y}(x,y) = xe^{-(x+y)}, x > 0, y > 0.$$

Now $U = \frac{Y}{X}$. Let $V = X$. Then

$$X = V\ ,\ Y = UV \text{ and if } x = v,\ y = uv,$$

$$\frac{\partial x}{\partial u} = 0 \ , \ \frac{\partial x}{\partial v} = 1 \ , \ \frac{\partial y}{\partial u} = v \ , \ \frac{\partial y}{\partial v} = u, \text{ so}$$

$$J = \begin{vmatrix} 0 & 1 \\ v & u \end{vmatrix} = -v$$

Now $x > 0 \Longleftrightarrow v > 0$ and $y > 0 \Longleftrightarrow uv > 0 \Longrightarrow u > 0$ since $v > 0$.

$$\therefore f_{U,V}(u,v) = f_{X,Y}(x,y)|J|$$

$$= ve^{-(v+uv)}|-v| = v^2e^{-v(1+u)}, u > 0, v > 0.$$

$$\therefore f_U(u) = \int_0^\infty v^2 e^{-v(1+u)} dv = \frac{2}{(1+u)^3}, \ \ u > 0$$

Note also, $\mathbb{E}(U) = \displaystyle\int_0^\infty u \cdot \frac{2}{(1+u)^3} du = 1.$

Example

Suppose X denotes the total time from arrival to exit from a service queue and Y denotes the time spent in the queue before being served. Suppose also that we want the density of $U = X - Y$, the amount of time spent being served when

$$f_{X,Y}(x,y) = e^{-x}, 0 < y < x < \infty.$$

Now $U = X - Y$. Let $V = Y$.

Find the density function of U, using a bivariate transformation.

Then

$$X = U + V \ , \ Y = V \text{ and if } x = u + v, \ y = v,$$

$$\text{then } \frac{\partial x}{\partial u} = 1 \ , \ \frac{\partial x}{\partial v} = 0 \ , \ \frac{\partial y}{\partial u} = 1 \ , \ \frac{\partial y}{\partial v} = 1.$$

$$\therefore J = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} = 1 \text{ and}$$

$$0 < y < x < \infty \Longleftrightarrow 0 < v < u + v < \infty$$

$$\Longrightarrow v > 0 \ , \ u > 0 \text{ and } u + v < \infty \Longrightarrow u < \infty, \ v < \infty.$$

$$f_{U,V}(u,v) = \mathrm{e}^{-(u+v)}, u > 0, v > 0$$

Thus the time spent being served, U, and the time spent in the queue before service, V, are independent random variables. Also,

$$f_U(u) = e^{-u}, u > 0 \text{ and } f_V(v) = e^{-v}, v > 0.$$

Multivariate Transformations

Recall the change of variable formula for one-dimensional integration. Suppose we are given the integral

$$\int_{\mathcal{R}} f(x)\mathrm{d}x$$

If $g : \mathcal{R} \to \mathcal{S}$ is an invertible transformation that maps the region/interval $\mathcal{R}$ into $\mathcal{S}$, then for the change of variable $z = g(x)$ or $x = g^{-1}(z)$, we have

$$\int_{\mathcal{R}} f(x)\mathrm{d}x = \int_{\mathcal{S}} f(g^{-1}(z))\frac{1}{g'(g^{-1}(z))}\mathrm{d}z$$

Note that since

$$\frac{1}{g'(x)} = \frac{1}{g'(g^{-1}(z))} = (g^{-1}(z))'$$

the last bit can be written as

$$\int_{\mathcal{R}} f(x)\mathrm{d}x = \int_{\mathcal{S}} f(g^{-1}(z))(g^{-1}(z))'\mathrm{d}z$$

Something similar holds for higher dimensions. We give the formula without proof. Suppose we are given the n-dimensional integral

$$\int \cdots \int_{\mathcal{R}} f(\mathbf{x})\mathrm{d}\mathbf{x}$$

and the invertible transformation

$$\mathbf{z} = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} = \mathbf{g}(\mathbf{x}) = \begin{pmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_n(\mathbf{x}) \end{pmatrix}$$

which maps the region $\mathcal{R}$ into $\mathcal{S}$. Then, if we change the variable from $\mathbf{x}$ to $\mathbf{z} = \mathbf{g}(\mathbf{x})$ we obtain

Change of variable formula

$$\int \cdots \int_{\mathcal{R}} f(\mathbf{x})\mathrm{d}\mathbf{x} = \int \cdots \int_{\mathcal{S}} f(\mathbf{g}^{-1}(\mathbf{z}))\ |\det(J_{\mathbf{g}^{-1}})(\mathbf{z})|\ \mathrm{d}\mathbf{z},$$

where $|\det(J_{\mathbf{g}^{-1}})(\mathbf{z})|$ stands for the absolute value of the determinant of the Jacobian matrix of the transformation $\mathbf{g}^{-1}$ evaluated at $\mathbf{z}$.

The Jacobian matrix of a transformation $\mathbf{g}$ is defined as the matrix of partial derivatives

$$J_{\mathbf{g}}(\mathbf{x}) = \begin{bmatrix} \frac{\partial g_1}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial g_1}{\partial x_n}(\mathbf{x}) \\ \vdots & \vdots & \vdots \\ \frac{\partial g_n}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial g_n}{\partial x_n}(\mathbf{x}) \end{bmatrix}$$

Note that $|\det(J_{\mathbf{g}^{-1}})(\mathbf{z})|$ plays the same role as $(g^{-1}(z))'$ in the one dimensional case and satisfies

$$|\det(J_{\mathbf{g}^{-1}})(\mathbf{z})| = \frac{1}{|\det(J_{\mathbf{g}})(\mathbf{x})|}$$

just like $\frac{1}{g'(x)} = \frac{1}{g'(g^{-1}(z))} = (g^{-1}(z))'$ in the one-dimensional change of variable case. As a result of this

Example

Suppose $\mathbf{X} \in \mathcal{R}$ has pdf

$$f_{\mathbf{X}}(\mathbf{x})$$

and we transform the variable from $\mathbf{X}$ to $\mathbf{Z}$ using

$$\mathbf{Z} = \mathbf{g}(\mathbf{X})$$

We wish to express the pdf of $\mathbf{Z} \in \mathcal{S}$

$$f_{\mathbf{Z}}(\mathbf{z})$$

in terms of the pdf of $\mathbf{X}$ above. The formula for this is

Result

Suppose $\mathbf{X} \in \mathbf{R}^n$ has pdf $f_{\mathbf{X}}(\mathbf{x})$ and we transform $\mathbf{X}$:

$$\mathbf{Z} = \mathbf{g}(\mathbf{X}) \ ,$$

where $\mathbf{g} : \mathbf{R}^n \to \mathbf{R}^n$ is invertible and continuously differentiable. Then, the pdf of $\mathbf{Z}$ is

$$f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{z}))\ |\det(J_{g^{-1}}(\mathbf{z}))|$$

Proof: We can write for any sufficiently nice region $\mathcal{A} \subseteq \mathcal{S}$ that maps to $\mathcal{B} = \mathbf{g}^{-1}(\mathcal{A})$ under the transformation $\mathbf{g}$:

$$\begin{aligned}
\int_{\mathcal{A}} f_{\mathbf{Z}}(\mathbf{z})\mathrm{d}\mathbf{z} &= \mathbb{P}(\mathbf{Z} \in \mathcal{A}) \quad \text{by definition of pdf of } \mathbf{Z} \\
&= \mathbb{P}(\mathbf{g}(\mathbf{X}) \in \mathcal{A}) \\
&= \mathbb{P}(\mathbf{X} \in \mathbf{g}^{-1}(\mathcal{A})) \\
&= \int_{\mathbf{g}^{-1}(\mathcal{A})} f_{\mathbf{X}}(\mathbf{x})\mathrm{d}\mathbf{x} \quad \text{by definition of pdf of } \mathbf{X} \\
&= \int_{\mathcal{A}} f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{z}))|\ \det(J_{\mathbf{g}^{-1}}(\mathbf{z}))|\ \mathrm{d}\mathbf{z} \quad \text{by change of variable formula}
\end{aligned}$$

Hence it follows that for any nice region $\mathcal{A} \subseteq \mathcal{S}$ we have:

$$\int_{\mathcal{A}} f_{\mathbf{Z}}(\mathbf{z})\mathrm{d}\mathbf{z} = \int_{\mathcal{A}} f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{z}))|\ \det(J_{\mathbf{g}^{-1}}(\mathbf{z}))|\ \mathrm{d}\mathbf{z}$$

This suggests that the integrands are equivalent in some sense, because they give us the same integral for any region $\mathcal{A}$. Without getting into technicalities, we can see why it is plausible that

$$f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{X}}(\mathbf{g}^{-1}(\mathbf{z}))|\ \det(J_{\mathbf{g}^{-1}}(\mathbf{z}))|$$

Example

Suppose again that $\mathbf{X} \in \mathcal{R}$ has pdf

$$f_{\mathbf{X}}(\mathbf{x})$$

and we consider the pdf of $\mathbf{Z} = A\mathbf{X}$, where A is an $n \times n$ invertible matrix. Setting $\mathbf{g}(\mathbf{x}) = A\mathbf{x}$ in the result above, we obtain that

$$\mathbf{g}^{-1}(\mathbf{z}) = A^{-1}\mathbf{z}$$

and

$$J_{g^{-1}}(\mathbf{z}) = A^{-1}$$

so that

Result

$$f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{X}}(A^{-1}\mathbf{z})|\det(A^{-1})| = \frac{1}{|\det(A)|} f_{\mathbf{X}}(A^{-1}\mathbf{z})$$

You may use this transformation formula in your assignment without giving a proof there. Note that we already know the special case of this formula when $A = a$ is a scalar and $Z = aX$:

$$f_Z(z) = f_X(a^{-1}z)|\det(a^{-1})| = \frac{1}{|a|} f_X\left(\frac{z}{a}\right)$$

Sums of Independent Random Variables

Often we wish to sum random variables, particularly when calculating statistics to summarise data (such as $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$). In this chapter we will explore some key tools for deriving the distribution of the sum of two or more random variables.

Probability function/density function approach

Result

Suppose that X and Y are independent random variables taking only non-negative integer values, and let $Z = X + Y$. Then

$$f_Z(z) = \sum_{y=0}^{z} f_X(z-y) f_Y(y), \quad z = 0, 1, \ldots$$

This formula is called a (discrete) *convolution formula.*

Proof:

$$\begin{aligned} f_Z(z) &= \mathbb{P}(X+Y=z) \\ &= \mathbb{P}(X=z, Y=0) + \mathbb{P}(X=z-1, Y=1) + \ldots + \mathbb{P}(X=0, Y=z) \\ &= \mathbb{P}(X=z)\mathbb{P}(Y=0) + \mathbb{P}(X=z-1)\mathbb{P}(Y=1) + \ldots + \mathbb{P}(X=0)\mathbb{P}(Y=z) \\ &= f_X(z)f_Y(0) + f_X(z-1)f_Y(1) + \ldots + f_X(0)f_Y(z) \\ &= \sum_{y=0}^{z} f_X(z-y)f_Y(y) \end{aligned}$$

Example

$$f_X(k) = f_Y(k) = (1-\theta)\theta^k, \ k = 0, 1, 2, \ldots \ ; \ 0 < \theta < 1$$

Find the probability function of $Z = X + Y$.

$$\begin{aligned} f_{X+Y}(z) &= \sum_{k=0}^{z} (1-\theta)\theta^k (1-\theta)\theta^{z-k} \\ &= (z+1)(1-\theta)^2\theta^z, \ z = 0, 1, 2, \ldots \end{aligned}$$

Example*

$X \sim$ Poisson (λ_1), $Y \sim$ Poisson (λ_2).

$$\mathbb{P}(X=k) = \frac{e^{-\lambda_1}\lambda_1^k}{k!}, \ k = 0, 1, 2, \ldots$$

$$\mathbb{P}(Y=k)=\frac{e^{-\lambda_2}\lambda_2^k}{k!}\;,\;\;k=0,1,2,\ldots$$

Find the probability function of $Z = X + Y$.

$$\begin{aligned} f_{X+Y}(z) &= \sum_{k=0}^{z} \frac{e^{-\lambda_1}\lambda_1^k}{k!}\cdot\frac{e^{-\lambda_2}\lambda_2^{z-k}}{(z-k)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{z!}\sum_{k=0}^{z}\begin{pmatrix} z \\ k\end{pmatrix}\lambda_1^k\lambda_2^{z-k} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^z}{z!}\;,\;\;z=0,1,2,\ldots\;. \end{aligned}$$

Thus $X + Y \sim$ Poisson $(\lambda_1 + \lambda_2)$.

It follows by induction that if $X_1, X_2, \ldots, X_n$ are independent with $X_i \sim$ Poisson (λ_i), then

$$\sum_{i=1}^{n} X_i \sim \text{ Poisson }\left(\sum_{i=1}^{n}\lambda_i\right).$$

This is an important and useful property of Poisson random variables.

Result

Suppose X and Y are independent continuous variables with $X \sim f_X(x)$ and $Y \sim f_Y(y)$. Then $Z = X + Y$ has density

$$f_Z(z) = \int_{\text{all possible } y} f_X(z-y)f_Y(y)dy.$$

This formula is called a (continuous) convolution formula.

Proof:

$$\begin{aligned} F_Z(z)=\mathbb{P}(Z\le z) &= \mathbb{P}(X+Y\le z) \\ &= \underset{x+y\le z}{\int\cdots\int} f_{X,Y}(x,y)\,dx\,dy \\ &= \int_{\text{all possible } y}\int_{-\infty}^{z-y} f_X(x)f_Y(y)\,dx\,dy \\ &= \int_{\text{all possible } y} F_X(z-y)f_Y(y)\,dy \end{aligned}$$

To complete the proof we differentiate wrt z in order to obtain the density function $f_Z(z)$:

$$f_Z(z)=\frac{d}{dz}F_Z(z)=\int_{\text{all possible } y} f_X(z-y)f_Y(y)\,dy$$

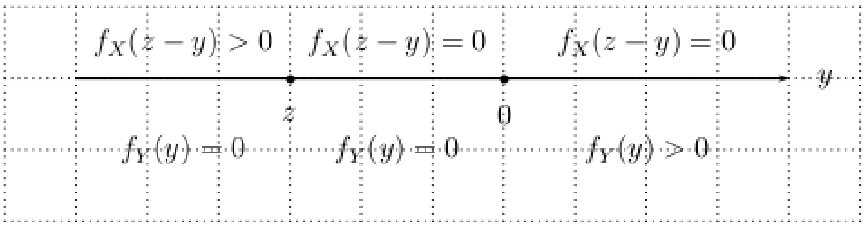
Example

X and Y are independent variables and $f_X(x) = e^{-x}$, $x > 0$, $f_Y(y) = e^{-y}$, $y > 0$.

Find the density function of $Z = X + Y$.

$$Z = X + Y \sim f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy.$$

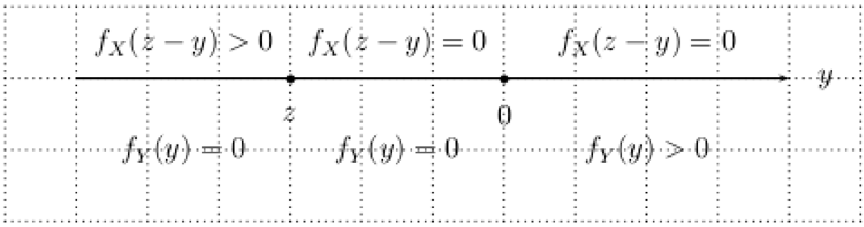
Now $f_X(z-y) = e^{-(z-y)}$ for $z - y > 0$ or $y < z$. If $z < 0$



$$\therefore \text{ If } z < 0, f_X(z-y) f_Y(y) = 0 \text{ for all } y$$

$$\Longleftrightarrow f_Z(z) = 0 \text{ for } z < 0, \text{ as expected.}$$

If $z > 0$



$$\therefore f_X(z-y) f_Y(y) = \begin{cases} 0 \text{ for } y < 0 \text{ and } y > z \\ \\ e^{-(z-y)} \cdot e^{-y} = e^{-z} \text{ for } 0 < y < z. \end{cases}$$

$$\therefore f_Z(y) = \int_0^z e^{-z} dy = z e^{-z}, z > 0.$$

Note that the answer is the density function of a Gamma(2,1) random variable. Thus the sum of two independent exponential(1) random variables is a Gamma(2,1) variable.

Example*

Y_1 and Y_2 are independent variables and $Y_1 \sim N(0,1),\ Y_2 \sim N(0,1)$.

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty$$

Find the distribution of $Z = Y_1 + Y_2$.

Z has density

$$f_Z(z) = \int_{-\infty}^{\infty} f_{Y_1}(z-y) f_{Y_2}(y) dy$$

where

$$f_{Y_1}(z-y) = \frac{1}{\sqrt{2\pi}} e^{-(z-y)^2/2}, \quad -\infty < z-y < \infty \quad \text{and} \quad f_{Y_2}(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}, \quad -\infty < y < \infty$$

For any fixed $z \in (-\infty, \infty)$, when considered as a function of y,

$$f_{Y_1}(z-y) = \frac{1}{\sqrt{2\pi}} e^{-(z-y)^2/2}, \quad -\infty < y < \infty$$

$$\begin{aligned}
\therefore f_Z(z) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(z-y)^2/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-z^2/2 + zy - y^2} dy \\
&= \frac{e^{-z^2/2}}{2\pi} \int_{-\infty}^{\infty} e^{-\left(y^2 - zy + \frac{z^2}{4}\right) + \frac{z^2}{4}} dy \\
&= \frac{e^{-z^2/4}}{2\pi} \int_{-\infty}^{\infty} e^{-\left(y - \frac{z}{2}\right)^2} dy.
\end{aligned}$$

Now for any μ and σ,

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} dy = 1 \quad \text{and} \quad \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} dy = \sigma\sqrt{2\pi}.$$

Put $\sigma^2 = \frac{1}{2}$ and $\mu = \frac{z}{2}$. Then $\displaystyle\int_{-\infty}^{\infty} e^{-(y-\frac{z}{2})^2} dy = \sqrt{\pi}$.

$$\therefore f_Z(z) = \frac{e^{-z^2/4}}{2\pi}\sqrt{\pi} = \frac{1}{2\sqrt{\pi}} e^{-z^2/4}, -\infty < z < \infty.$$

Thus $Z \sim N(0,2)$. More generally, we can show that the sum of *any* two normal random variables is also normal.

Example*

$X \sim$ Gamma $(\alpha_1, 1), Y \sim$ independently of Gamma $(\alpha_2, 1)$.

$$f_X(x) = \frac{e^{-x}x^{\alpha_1-1}}{\Gamma(\alpha_1)} \quad , x > 0; \quad \alpha_1 \geq 1$$

$$f_Y(y) = \frac{e^{-y}y^{\alpha_2-1}}{\Gamma(\alpha_2)} \quad , y > 0; \quad \alpha_2 \geq 1$$

Find the distribution of $Z = X + Y$.

Z has density

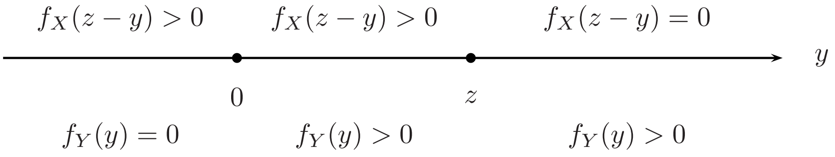
$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy$$

First we will find the limits on this integral, that is, the range of values of y for which $f_X(z-y)$ and $f_Y(y)$ are both non-zero. Both X and Y are only defined over positive values, so the range of values of y of interest to us satisfies both $z - y > 0$ and $y > 0$, *i.e.* $0 < y < z$.

$$f_X(z-y) = \frac{e^{-(z-y)}(z-y)^{\alpha_1-1}}{\Gamma(\alpha_1)} \quad \text{for } z - y > 0 \text{ or } y < z$$

$f_Z(z) = 0$ for $z < 0$ since X and Y are nonnegative random variables.

For $z > 0$



$$\therefore f_X(z-y) f_Y(y) = 0 \text{ for } y < 0 \text{ or } y > z$$

For $0 < y < z$,

$$\begin{aligned} f_X(z-y) f_Y(y) &= \frac{e^{-(z-y)}(z-y)^{\alpha_1-1}}{\Gamma(\alpha_1)} \cdot \frac{e^{-y}y^{\alpha_2-1}}{\Gamma(\alpha_2)} \\ &= \frac{e^{-z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} (z-y)^{\alpha_1-1} y^{\alpha_2-1} \end{aligned}$$

$$\begin{aligned}
\therefore f_Z(z) &= \int_0^z \frac{e^{-z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}(z-y)^{\alpha_1-1}y^{\alpha_2-1}dy \qquad \text{now substitute } t=\frac{y}{z}: \\
&= \frac{e^{-z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^1 \{z(1-t)\}^{\alpha_1-1}\{zt\}^{\alpha_2-1}zdt \\
&= \frac{e^{-z}z^{\alpha_1-1+\alpha_2-1+1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^1 (1-t)^{\alpha_1-1}t^{\alpha_2-1}dt \\
&= \frac{e^{-z}z^{\alpha_1+\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\cdot B(\alpha_2,\alpha_1),\ \text{where } B(\alpha_2,\alpha_1)=B(\alpha_1,\alpha_2)=\frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)} \\
\therefore f_Z(z) &= \frac{e^{-z}z^{\alpha_1+\alpha_2-1}}{\Gamma(\alpha_1+\alpha_2)} \quad ,z>0.
\end{aligned}$$

That is, $Z \sim$ Gamma $(\alpha_1+\alpha_2, 1)$.

Note: the Beta (α_1, α_2) density is $\dfrac{x^{\alpha_1-1}(1-x)^{\alpha_2-1}}{B(\alpha_1,\alpha_2)}$.

The arguments given in the previous example can be extended to derive:

Result

If $X_1, X_2, \ldots, X_n$ are independent with $X_i \sim \text{Gamma}(\alpha_i, \beta)$, then

$$\sum_{i=1}^n X_i \sim \text{Gamma}\left(\sum_{i=1}^n \alpha_i, \beta\right).$$

Moment generating function approach

Result

Suppose that X and Y independent random variables with moment generating functions m_X and m_Y. Then

$$m_{X+Y}(u) = m_X(u)m_Y(u).$$

Proof: If X and Y are independent, then $Z = X+Y$ has mgf

$$\begin{aligned}
m_Z(u) = \mathbb{E}(e^{u(X+Y)}) &= \mathbb{E}(e^{uX}\cdot e^{uY}) \\
&= \mathbb{E}(e^{uX})\cdot\mathbb{E}(e^{uY}) \\
&= m_X(u)\cdot m_Y(u)
\end{aligned}$$

so the mgf for the density

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy \text{ is } m_X(u)\cdot m_Y(u);$$

An alternative (and more complicated) proof is as follows:

$$\begin{aligned}
\mathbb{E}(e^{uZ}) = \int_{-\infty}^{\infty} e^{uz} f_Z(z)dz &= \int_{-\infty}^{\infty} e^{uz} \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\, dy\, dz \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{u(z-y)} f_X(z-y) dz \cdot e^{uy} f_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{ux} f_X(x)\, dx \cdot e^{uy} f_Y(y)\, dy, \quad \text{where } x = z-y \\
&= \int_{-\infty}^{\infty} m_X(u) \cdot e^{uy} f_Y(y)\, dy \\
&= m_X(u) \int_{-\infty}^{\infty} e^{uy} f_Y(y)\, dy = m_X(u) \cdot m_Y(u)
\end{aligned}$$

Example

Recall from Chapter 2 that if $X \sim N(\mu, \sigma^2)$ then $m_X(u) = e^{\mu u + \frac{1}{2}\sigma^2 u^2}$.

Use this result to find the distribution of $Z = Y_1 + Y_2$ where $Y_1 \sim N(0,1)$ independently of $Y_2 \sim N(0,1)$.

More generally,

Result

If $X_1, X_2, \ldots, X_n$ are independent random variables, then $\sum_{i=1}^n X_i$ has moment generating function

$$m_{\sum_{i=1}^n X_i}(u) = \prod_{i=1}^n m_{X_i}(u).$$

Proof:

$$\begin{aligned}
m_{\sum_{i=1}^n X_i}(u) &= \mathbb{E}(e^{u\sum_{i=1}^n X_i}) \\
&= \mathbb{E}\left(\prod_{i=1}^n e^{uX_i}\right) \\
&= \prod_{i=1}^n \mathbb{E}(e^{uX_i}) = \prod_{i=1}^n m_{X_i}(u).
\end{aligned}$$

This offers us a useful approach for deriving the distribution of the sum of independent random variables, using the 1-1 correspondence between distributions and moment generating functions. For this approach to work however we need to be able to recognise the distribution of the sum from its moment generating function.

Example

$X_1, X_2, \ldots, X_n$ independent Bernoulli (p) random variables.

Use moment generating functions to show that $\sum_{i=1}^n X_i \sim$ Binomial(n,p).

Each X_i has probability function

$$f_X(x) = p^x(1-p)^{1-x}, \;\; x = 0, 1; \;\; 0 < p < 1$$

and moment generating function (mgf)

$$\begin{aligned} m_X(u) = \mathbb{E}(e^{uX}) &= \sum_{k=0}^{1} e^{uX}\mathbb{P}(X = x) \\ &= e^{u\cdot 0}\mathbb{P}(X=0) + e^{u\cdot 1}\mathbb{P}(X=1) \\ &= 1 - p + pe^u \end{aligned}$$

Therefore, the moment generating function of the sum $\sum_{i=1}^n X_i$ is

$$m_{\sum_{i=1}^n X_i}(u) = \prod_{i=1}^n m_{X_i}(u) = (1-p+pe^u)^n$$

which is the mgf of a Binomial(n,p) random variable. Thus we can conclude that $\sum_{i=1}^n X_i \sim$ Binomial(n,p).

Note: if $Z \sim$ Binomial(n,p),

$$\begin{aligned} m_Z(u) = \mathbb{E}(e^{uZ}) &= \sum_{\text{all } z} e^{uz}\mathbb{P}(Z = z) \\ &= \sum_{z=0}^{n} e^{uz} \begin{pmatrix} n \\ z \end{pmatrix} p^z(1-p)^{n-z} \\ &= \sum_{z=0}^{n} \begin{pmatrix} n \\ z \end{pmatrix} (pe^u)^z(1-p)^{n-z} \\ &= (1-p+pe^u)^n \end{aligned}$$

Example

$X_1, X_2, \ldots, X_n$ independent random variables with $X_i \sim$Poisson (λ_i).

Find the mgf of X_i and hence deduce the distribution of $\sum_{i=1}^n X_i$.

If $X \sim$ Poisson (λ), then X has mgf

$$\begin{aligned} m_X(u) = \mathbb{E}(e^{uX}) &= \sum_{x=0}^{\infty} e^{ux} \cdot \frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^u)^x}{x!} \\ &= e^{-\lambda} \cdot e^{\lambda e^u} = e^{\lambda(e^u-1)}. \end{aligned}$$

Thus X_i above has mgf

$$m_{X_i}(u) = \mathbb{E}(e^{uX_i}) = e^{\lambda_i(e^u-1)}$$

and $\displaystyle\sum_{i=1}^{n} X_i$ has mgf

$$m_{\sum_{i=1}^n X_i}(u) = \mathbb{E}(e^{u\sum_{i=1}^n X_i}) = \prod_{i=1}^{n} \mathbb{E}(e^{uX_i}) = e^{\left(\sum_{i=1}^n \lambda_i\right)(e^u-1)}$$

This has the form of the Poisson mgf as derived above, but now with parameter $\sum_{i=1}^n \lambda$.

$$\therefore \sum_{i=1}^{n} X_i \sim \text{ Poisson } \left(\sum_{i=1}^{n} \lambda_i\right).$$

It can be shown (again using moment generating functions) that sums of independent normal random variables are also normal; with the means and variances added.

Result

If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are independent then

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

The following extension is for general weighted sums.

Result

If for $1 \leq i \leq n$, $X_i \sim N(\mu_i, \sigma_i^2)$ are independent then for any set of constants $a_1, \ldots, a_n$,

$$\sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i\mu_i, \sum_{i=1}^{n} a_i^2\sigma_i^2\right).$$

Can you prove these results?

Example

Suppose that

$$X_1 \sim N(3,8), \quad X_2 \sim N(-5,5), \quad X_3 \sim N(2,11).$$

and that $X_1.X_2$ and X_3 are independent. Let

$$U = 6X_1 + 3X_2 - 9X_3.$$

Using known results for sums of normal random variables, find the distribution of U.

Linear combinations of normal random variables are normal. So U is normal with mean

$$\mathbb{E}(U) = 6 \times 3 + 3 \times (-5) - 9 \times 2 = -15$$

and

$$\mathrm{Var}(U) = 6^2 \times 8 + 3^2 \times 5 + (-9)^2 \times 11 = 425$$

In summary,

$$U \sim \mathsf{N}(-15, 425).$$

An application of the multivariate probabilities

Now we've met some common methods for modelling data, it's time to think about how to collect data in such a way that we can use these models. Whether or not it is appropriate to do statistical modelling, using tools from the previous chapters, depends critically on how the data were collected. For example, we have been treating data as **random** variables. For this to be a reasonable thing to do, we need to sample randomly in order to introduce randomness.

In this section we will work through some key ideas concerning the collection of data. We will also meet the important idea that an appropriate study design enables the examination of the properties of the statistics we calculate from samples, using random variables.

For further reading on survey design, consider Rice (2007) Chapter 7.

Survey design

When collecting data in a survey it is critical to try to collect data that is **representative** and **random**.

Representativeness

When we collect a sample from a population, typically we would like to use this sample to make inferences about some property of the population at large. However, this is only reasonable if **the sample is representative of the population**. If this is not achieved then inferences about the population can be completely wrong.

We can formally define representativeness as below.

Definition

Consider a sample $X_1, \ldots, X_n$ from a random variable X which has probability or density function $f_X(x)$.
The sample is said to be representative if:

$$f_{X_i}(x) = f_X(x) \quad \text{for each } i$$

Representativeness is typically a **more important consideration than sample size** – it is better to have a small but representative sample than a large but unrepresentative sample.

Example

"The poll that changed polling" http://historymatters.gmu.edu/d/5168

The Literary Digest correctly predicted the outcomes of each of the 1916-1932 US presidential elections by conducting polls. These polls were a lucrative venture for the magazine – readers liked them; they got a lot of news coverage; and they were linked to subscription renewals. The 1936 postal card poll claimed to have asked one fourth of the nation's voters which candidate they intended to vote for. Based on more than 2,000,000 returned post cards, it issued its prediction: Republican presidential candidate Alfred Landon would win in a landslide. But this "largest poll in history" couldn't have been more wrong – the Democrat Roosevelt won by the election by the largest margin in history! (Roosevelt got more than 60% of the vote, but was predicted to get only 43%.) The Literary Digest lost a lot of credibility from this result and was soon discontinued.

The result was correctly predicted by a new pollster, George Gallup, based on just 50,000 voters selected in a representative fashion and interviewed face-to-face. Gallup not only predicted the election result, but before the Literary Digest poll was released, he correctly predicted that it would get it wrong! This election made "Gallup polls" famous, and formed a template for polling methods ever since.

What went wrong in the Literary Digest poll?

While some have claimed that more Republicans were Literary Digest subscribers, an American Statistician article claimed that voluntary response was the main issue.

How do you ensure a sample is representative? One way to ensure this is to take a simple random sample from the population of interest, as below.

Random samples

Definition
A **random sample** of size n is a set of random variables

$$X_1, \ldots, X_n$$

with the properties

1. the X_i's each have the same probability distribution.
2. the X_i's are independent.

We often say that the X_i are **iid** (independently and identically distributed).

Example

Sampling with replacement
Consider sampling a variable X in a population of 10 subjects, which take the following (sorted) values:

$$2 \quad 4 \quad 5 \quad 7 \quad 8 \quad 10 \quad 14 \quad 17 \quad 27 \quad 35$$

We sample three subjects randomly (with equal sampling probability for each subject), with replacement. Let these values be X_1, X_2 and X_3.

Show that X_1, X_2 and X_3 are iid.

$f_{X_1}(x) = \mathbb{P}(X_1 = x) = \frac{1}{10}$ for $x \in \{2, 4, 5, 7, 8, 10, 14, 17, 27, 27, 35\}$ $f_{X_2}(x) = f_{X_3}(x) = f_{X_1}(x)$ hence identically distributed. Since with replacement, $f_{X_2|X_1}(x) = f_{X_2}(x)$ (similarly for other combinations) hence the X_i are independently distributed.

It is more common however to sample without replacement. The most common method of obtaining such a random sample is to take a **simple random sample**:

Definition
A **simple random sample** of size n is a set of subjects sampled in such a way that all possible samples of size n are equally likely.

To obtain a random sample using `R`:

- Obtain a list of all subjects in the population, and assign each subject a number from 1 to N
- Use `sample(N,n)` to take a simple random sample of size n.

(The `sample` function generates N random numbers, assigns one to each subject, then includes in the sample the n subjects with smallest n random numbers.)

Strictly speaking, a simple random sample does not consist of iid random variables – they are identically distributed, but they are dependent, since knowledge that $X_i = x_i$ makes it less likely that $X_j = x_i$ because the ith subject can only be included in the sample once. However this dependence is very weak when the population size N is large compared to the sample size n (*e.g.* if $N > 100n$) and so in most instances it can be ignored. See MATH3831 for "finite sample" survey methods when this dependence cannot be ignored.

It is important in surveys, wherever possible, to ensure sampling is random. This is important for a few reasons:

- It ensures the n values in the sample are iid, which is an important assumption of most methods of statistical inference (as in the coming chapters).
- Random sampling removes selection bias – the choice of who is included in the study is taken away from the experimenter, hence it is not possible for them to (intentionally or otherwise) manipulate results through choice of subjects.
- Randomly sampling from the population of interest guarantees that the sample is representative of the population.

Unfortunately, it is typically very hard to obtain a simple random sample from the population of interest, so the best we can hope for is a "good approximation".

Example

Consider polling NSW voters to predict the result of a state election. You do not have access to the list of all registered voters (for privacy reasons).

How would you sample NSW voters?
It is difficult to think of a method that ensures a representative sample!

Statistics calculated from samples

Definition
Let $X_1, \dots, X_n$ be a random sample. A **statistic** is any real-valued function of the random sample.

While any real-valued function can in theory be considered as a statistic, in practice we focus on particular functions which measure something of interest to us about the sample. Important examples of statistics are:

- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, the **sample mean**
- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, the **sample variance**
- $\tilde{X}_{0.5}$, the **sample median**

A key advantage of random sampling is that the fact that the sample is random implies that **any statistic calculated from the sample is random**. Hence we can treat statistics as random variables and study their properties. Further, the iid property makes it a lot easier to derive some important properties of statistics, as in the important examples below.

Properties of the sample mean
If $X_1, \dots, X_n$ is a random sample from a variable with mean μ and variance σ^2, then the sample mean $\bar{X}$ satisfies:

$$\mathbb{E}(\bar{X}) = \mu \quad \text{and} \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Properties of the sample proportion
If $X_1, \ldots, X_n$ is a random sample from a Bernoulli(p) variable, then the sample proportion $\widehat{p}$ satisfies:

$$\mathbb{E}(\widehat{p}) = p \quad \text{and} \quad \text{Var}(\widehat{p}) = \frac{p(1-p)}{n}$$

Note that while the variance results given above require the variables to be iid (hence a random sample), the expectation results only requires the observations in the sample to be identically distributed.

Because of the property that $\mathbb{E}(\bar{X}) = \mu$, we say that sample means of random samples are **unbiased** (we will come back to this later on). Similarly, $\widehat{p}$ is unbiased.

Methods of survey sampling

There are many methods of survey sampling beyond taking a simple random sample. Key considerations when choosing a sampling scheme are **efficiency** and **effort** – ideally we would like an estimate that gives us a good estimate relative to the effort invested in sampling.

Definition
Consider two unbiased alternative statistics, denoted as $g(X_1, \ldots, X_n)$ and $h(Y_1, \ldots, Y_m)$. We say that $g(X_1, \ldots, X_n)$ is more **efficient** than $h(Y_1, \ldots, Y_m)$ if:

$$\text{Var}[g(X_1, \ldots, X_n)] < \text{Var}[h(Y_1, \ldots, Y_m)]$$

Note that the above notation implies that not only can the statistics that we are using differ ($g(\cdot)$ vs $h(\cdot)$), but the observations used to calculate the statistics can also differ ($X_1, \ldots, X_n$ vs $Y_1, \ldots, Y_m$). This reflects that there are two ways to achieve efficiency – use a different statistic (discussed in later chapters) or by sampling differently. The most obvious way that sampling differently can increase efficiency is by increasing the sample size, but even for a fixed sample size ($n = m$) efficiency varies with sampling method. Below are three common methods of sampling – for more, and a deeper study of their properties, see MATH3831.

Simple random sample Weaknesses are that it can be difficult to implement in practice, requiring a high effort, and it can be inefficient.

Stratified random sample If the population can be broken into subpopulations (or "strata") which differ from each other in the variable of interest, it is more efficient to sample separately within each stratum than to sample once across the whole population. *e.g.* Estimating average taxable income – this varies considerably with age, so a good survey design would involve sampling separately within age strata (if possible).

Cluster sampling This is useful when subjects in the population arise in clusters, and it takes less effort to sample within clusters than across clusters. Effort-per-subject can be reduced by sampling clusters and then measuring all (or sampling many) subjects within a cluster. *e.g.* Face-to-face interviews with 100 NSW household owners – it is easier logistically to sample ten postcodes, then sample ten houses in each postcode, than to travel to a random sample 100 households spread across (potentially) 100 NSW postcodes!

Example

Consider estimating the average heart rate of students, μ. Males and females are known to have different heart rates, μ_{M} and μ_{F}, but the same variance σ^2.

Consider estimating the mean μ using a stratified random sample, as follows:

- Take a random sample of size n of each gender.
- Calculate the sample mean of each gender $\bar{X}_{\mathrm{M}}$ and $\bar{X}_{\mathrm{F}}$.
- Since males and females occur with (approximately) equal frequency in the student population, we can estimate the overall mean heart rate as

$$\bar{X}_s = \frac{1}{2}\left(\bar{X}_{\mathrm{M}} + \bar{X}_{\mathrm{F}}\right)$$

1. Find $\mathrm{Var}(\bar{X}_s)$.
2. Show that the marginal variance of heart rate across the student population (ignoring gender) is $\mathrm{Var}(X) = \sigma^2 + (\mu_{\mathrm{M}} - \mu_{\mathrm{F}})^2/4$.
3. Hence show that stratified random sampling is more efficient than using a simple random sample of size $2n$, if $\mu_{\mathrm{M}} \neq \mu_{\mathrm{F}}$.

We are given that $\mathbb{E}\bar{X}_{\mathrm{M}} = \mu_{\mathrm{M}}$ and $\mathbb{E}\bar{X}_{\mathrm{F}} = \mu_{\mathrm{F}}$ and $\mathrm{Var}(\bar{X}_{\mathrm{M}}) = \mathrm{Var}(\bar{X}_{\mathrm{F}}) = \frac{1}{n}\sigma^2$ Therefore,

$$\mathrm{Var}(\bar{X}_s) = 1/4[\sigma^2/n + \sigma^2/n] = \frac{\sigma^2}{2n}$$

But if X is the heart rate of any arbitrary selected student (so that there is equal chance of being male or female), then we are given that $\mathbb{E}[X \mid \mathrm{M}] = \mu_{\mathrm{M}}$ and $\mathbb{E}[X \mid \mathrm{F}] = \mu_{\mathrm{F}}$ so that

$$\mathbb{E}[X] = \mathbb{E}[X \mid \mathrm{M}]\,\mathbb{P}(\mathrm{M}) + \mathbb{E}[X \mid \mathrm{F}]\,\mathbb{P}(\mathrm{F}) = \frac{\mu_{\mathrm{M}} + \mu_{\mathrm{F}}}{2} = \mu$$

and similarly for the variance

$$\begin{aligned}
\mathrm{Var}(X) = \mathbb{E}(X-\mu)^2 &= \mathbb{E}[(X-\mu)^2 \mid \mathrm{M}]\,\mathbb{P}(\mathrm{M}) + \mathbb{E}[(X-\mu)^2 \mid \mathrm{F}]\,\mathbb{P}(\mathrm{F}) \\
&= 1/2\mathbb{E}(X_M-\mu)^2 + 1/2\mathbb{E}(X_F-\mu)^2 \\
&= 1/2\mathbb{E}(X_M-\mu_M+\mu_M-\mu)^2 + 1/2\mathbb{E}(X_M-\mu_F+\mu_F-\mu)^2 \\
&= \sigma^2 + 1/2\mathbb{E}(\mu_M-\mu)^2 + 1/2\mathbb{E}(\mu_F-\mu)^2 \\
&= \sigma^2 + (\mu_M-\mu_F)^2/4
\end{aligned}$$

Hence, if $\bar{X} = \frac{1}{2n}\sum_{i=1}^{2n} X_i$ (that is, sample mean based on $2n$ observations)

$$\mathrm{Var}\left(\bar{X}\right) = \frac{\sigma^2}{2n} + \frac{(\mu_M-\mu_F)^2}{8n} \geqslant \frac{\sigma^2}{2n} = \mathrm{Var}(\bar{X}_s)$$

Design of experiments

Often in science we would like to demonstrate causation, *e.g.* does smoking causes learning difficulties? Does praying for a patient cause a better recovery following heart surgery?

While surveying a population often provides valuable information, it is very difficult to demonstrate **causation** based on just observing an association between two variables. The reason for this is that **lurking variables** can induce an association between X and Y when there is actually no causal relationship, or when the causal relationship has a completely different nature to what we observe.

Example

Student survey results demonstrate that students who walk to UNSW take a lot less time than students who use public transport!

Does this mean that walking to UNSW is faster than using public transport?
i.e. Should we all walk to UNSW to save time?

Definitions

An **observational study** (or survey) is a study in which we observe variables (X, Y) on subjects without manipulating them in any way.

An **experiment** is a study in which subjects are manipulating in some way (changing X) and we observe their response (Y).

The purpose of an experiment is to demonstrate that changes in X cause changes in Y.

Example

For each of the following experiments, what is X (the treatment variable) and Y (the response variable)?

1. The great prayer experiment (popularised in Richard Dawkins' book "The God Delusion")
Does praying for a patient influence their recovery from illness? A clinical trial was conducted to answer this question (Benson *et al.* 2006, published in the *American Heart Journal*), in which 1201 patients due to undergo coronary bypass surgery were randomly assigned to one of two groups – a group to receive daily prayers for 14 days following surgery, and a group who received no prayers. The study was double-blind, meaning that neither the patient nor anyone treating them knew whether or not they were being prayed for. The outcome variable of interest was whether or not each patient had any complications during the first 30 days following surgery.

2. A guinea pig experiment
Does smoking while pregnant affect the cognitive development of the foetus?

Johns et al (1993) conducted a study to look at this question using guinea pigs as a model. They injected nicotine tartate in a saline solution into ten pregnant guinea pigs, injected saline solution with no nicotine into ten other pregnant guinea pigs, and compared the cognitive development of offspring by getting them to complete a simple maze where they look for food. "Cognitive development" was measured as the number of errors made by the offspring when looking for food in a maze.

Note that both the above experiments (indeed any good experiment) is designed so that the only thing allowed to vary across groups is the treatment variable of interest (X) – so if a significant effect is detected in Y, the only plausible explanation would be that it was caused by X.

Key considerations in experimental design

Any experiment should compare, randomise and repeat:

Compare to demonstrate that changes in X cause changes in Y, we need to compare across suitable designed "treatment" groups (for which we have introduced changes in the value of X). These groups needs to be carefully designed so that the **only thing that differs** across groups is the treatment variable X. Double-blinding is often used for this reason (*e.g.* the prayer experiment), as is a "placebo" or "sham treatment" (*e.g.* saline-only injections in the guinea pig experiment).

Randomise the allocation of subjects to treatment groups. This ensures that any differences across groups, apart from those caused by treatment, are governed by chance (which we can then model!).

Repeat the **application of the treatment** to the different subjects in each treatment group. It is important that application of treatment is replicated (rather than applied once, "in bulk") in order that we can make inferences about the effect of the treatment in general.

The above points may seem relatively obvious, but they can be difficult to implement correctly – errors in design can be hard to spot.

Example

What error has been made in each of the following studies?

1. Consider the Mythbusters' "Is yawning contagious?" episode:
   http://www.yourdiscovery.com/video/mythbusters-top-10-is-yawning-contagious/
   The first attempt to answer this question involved sitting nine subjects together in a room for ten minutes, counting the number of yawns, and comparing results to when there was a "seed yawner" in the room with them who pretended to yawn for ten minutes.

   (Results were inconclusive!)

2. Greg was studying how mites affect the growth of cotton plants. He applied a mite treatment to eight (randomly chosen) plants by crumpling up a mite-infested leaf and leaving it at the base of each plant. He applied a no-mite treatment by not putting any leaves at the base of each of eight "control" plants.

   (Surprisingly, plants in the mite treatment had *faster* growth!?)

3. The success of a course on road rules was assessed by using the RTA's practice driving test:
   http://www.rta.nsw.gov.au/licensing/tests/driverknowledgetest/demonstrationdriverknowledgetest
   Participants were asked to complete the test before the course, then again afterwards, and results were compared.

(There was a significant improvement in scores on the test.)

Answers may be roughly along the lines:

1. Replicate the application of treatment: need to separately seed yawn to (groups of) subjects, as in second Mythbusters experiment. Also randomise!

2. Compare to an appropriate control: in this case, that would be crumpled-up leaves which are not mite-infested.

3. Compare to an appropriate control: in this case, people who do the test twice without attending a road rules course. (To account for learner effects.) And randomise!

Common experimental designs

Below are a few common experimental designs.

Randomised comparative experiment. Define K treatment groups (each with different levels of the variable X) and randomly assign subjects to each group.

Randomised blocks design. If there is some "blocking" variable known to be important to the response variable, break subjects into blocks according to this variable and randomise allocation of subjects to treatment groups separately within each block. This controls for the effects of the blocking variable.

Matched pairs design. A common special case of a randomised blocks design, where the blocks come in pairs. Common examples are "before-after" experiments (a pair of measurements is taken on a subject before and after treatment application), which control for subject-to-subject variation, and twins experiments (a pair of identical twins are studied, with one assigned to each of two treatment groups), which control for genetic variation.

For more details on the above and other common types of experiment, see MATH3851.

There is an analogy between randomised comparative experiments and simple random samples (which treat "all subjects as equal"), and between stratified random sampling and randomised blocks experiments (which break subjects into blocks/strata which are expected to differ in response). The terminology used is different but the concept is similar!

Example

Does regularly taking vitamins guard against illness? Consider two experiments on a set of $2n$ subjects:

A. Randomly assign subjects to one of two groups, each consisting of n subjects. The first group are given a vitamin supplement to take daily over the study period (three months), the second are given a placebo tablet (with no vitamins in it), to take daily. Number of illnesses are recorded over the study period.

B. All subjects are given a set of tablets (vitamins or placbeo) and asked to take them daily for three months. They are then given a different set of tablets (placebo or vitamin, whichever they didn't have last time) and are asked to take these for three months. Number of illnesses are recorded and compared over the two periods.

Let the number of illnesses in the two treatment groups be Y_v and Y_p. We are interested in the mean difference in number of illnesses between takers of vitamin tablets and takers of a placebo, estimated using the sample mean difference $\bar{Y}_\text{v} - \bar{Y}_\text{p}$. Assume $\text{Var}(Y_\text{v}) = \text{Var}(Y_\text{p}) = \sigma^2$.

1. What type of experiment has been done in each of A and B above?
2. Find $\text{Var}(\bar{Y}_\text{v} - \bar{Y}_\text{p})$ for experiment A.
3. It is noted in analysis that there is a correlation between number of illnesses in the two study periods (because some people get sick more often than others). Find $\text{Var}(\bar{Y}_\text{v} - \bar{Y}_\text{p})$ for experiment B, assuming that the correlation in measurements (and in sample means) across the two study periods is 0.5.
4. Which experiment gives a more efficient estimate of the treatment effect?

1. A is a randomized comparative experiment and B is a matched pairs design.

2. Here $\bar{Y}_\text{v}$ and $\bar{Y}_\text{p}$ are assumed independent (due to the randomized comparative experiment setting) and each is a sample mean of n subjects, hence $\text{Var}(\bar{Y}_\text{v}) = \text{Var}(Y_\text{v})/n = \sigma^2/n$ and

$$\text{Var}(\bar{Y}_\text{v} - \bar{Y}_\text{p}) = \text{Var}(\bar{Y}_\text{v}) + \text{Var}(\bar{Y}_\text{p}) = \frac{2\sigma^2}{n}$$

3. Here in scenario B, the Y_{p} and Y_{v} cannot be assumed to be independent. For this reason we will include the Cov in the calculation of the variance:

$$\begin{aligned}
\mathrm{Var}(Y_{\mathrm{v}} - Y_{\mathrm{p}}) &= \mathrm{Cov}(Y_{\mathrm{v}} - Y_{\mathrm{p}}, Y_{\mathrm{v}} - Y_{\mathrm{p}}) \\
&= \mathrm{Var}(Y_{\mathrm{v}}) + \mathrm{Var}(Y_{\mathrm{p}}) - 2\mathrm{Cov}(Y_{\mathrm{v}}, Y_{\mathrm{p}}) \\
&= 2\sigma^2 - 2\mathrm{Corr}(Y_{\mathrm{v}}, Y_{\mathrm{p}})\sqrt{\mathrm{Var}(Y_{\mathrm{v}})\mathrm{Var}(Y_{\mathrm{p}})} \\
&= 2\sigma^2 - 2\mathrm{Corr}(Y_{\mathrm{v}}, Y_{\mathrm{p}})\sigma^2 \\
&= 2\sigma^2(1 - \mathrm{Corr}(Y_{\mathrm{v}}, Y_{\mathrm{p}}))
\end{aligned}$$

Therefore,

$$\mathrm{Var}(\bar{Y}_{\mathrm{v}} - \bar{Y}_{\mathrm{p}}) = \frac{2\sigma^2}{n}(1 - \mathrm{Corr}(Y_{\mathrm{v}}, Y_{\mathrm{p}}))$$

4. Compared to study A, here we get a reduction in the variance if there is positive correlation $(1 - \mathrm{Corr}(Y_{\mathrm{v}}, Y_{\mathrm{p}})) < 1$ across the study periods. Note that the variance is increase if there is a negative correlation $(1 - \mathrm{Corr}(Y_{\mathrm{v}}, Y_{\mathrm{p}})) > 1$ across the study periods.

Chapter 5

Convergence of Random Variables

The previous chapter dealt with the problem of finding the density function (or probability function) of some transformation to a function of one or two random variables. In practice we are usually interested in some function of many variables – not just one or two. However, the calculations often become mathematically intractable in this situation. An example is the problem of finding the exact density function of the sum of 100 independent $\mathsf{U}(0,1)$ random variables.

Because of the difficulties in obtaining exact results, a large portion of mathematical statistics is concerned with *approximations* to density functions. These are based on convergence results for sequences of random variables.

In this chapter, we will focus on some key convergence results that are useful in statistics. Some of these (such as the law of large numbers and the central limit theorem) relate to sums or averages of random variables. These results are particularly useful, because sums and averages are typically used as summary statistics in quantitative research.

Suggested further reading for this chapter is Chapter 3.7 in Kroese & Chan (2014), Hogg *et al* (2005) Sections 4.2-4.4 or Rice (2007) Chapter 5.

Modes of Convergence

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space from Chapter 1, and let $X_1, X_2, \ldots, X$ be random variables on this space. In other words, each $X : \Omega \to \mathcal{X}$ is a function $X(\omega)$ mapping $\omega \in \Omega$ to $\mathcal{X} \subseteq \mathbb{R}$.

We have the following different modes/types of convergence of random variables.

Sure convergence

Recall that a random variable $X : \Omega \to \mathbb{R}$ is a function. Hence, we can consider the pointwise limit in standard first year Calculus:

$$\lim_{n \uparrow \infty} X_n(\omega) = X(\omega)$$

This is the familiar type of convergence as that of a real function $f_n(x) \to f(x)$ for some point x. Here we only replace x with ω.

Sure convergence

We say that $X_1, X_2, \ldots$ converges **surely** to X if

$$\lim_{n \uparrow \infty} X_n(\omega) = X(\omega), \quad \text{for all } \omega \in \Omega$$

Example

If $X_n(\omega) = X(\omega) + \frac{1}{n}$, then X_n converges to X surely.

Almost Sure convergence

Sure convergence is no different from the convergence in ordinary first year Calculus. We now introduce a special type of convergence that actually uses the concept of probability in its definition.

Almost Sure convergence

The sequence of numerical random variables $X_1, X_2, \ldots$ is said to converge **almost surely** to a numerical random variable X, denoted $X_n \xrightarrow{\text{a.s.}} X$, if

$$\mathbb{P}\left(\omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)\right) = 1 \; .$$

The last statement is equivalent to $\mathbb{P}(\omega \in \Omega \; : \; \lim_{n\to\infty} X_n(\omega) \neq X(\omega)) = 0$. In contrast to sure convergence here $X_\infty(\omega)$ may not always be equal to $X(\omega)$ for all events $\omega \in \Omega$, but the probability of any event ω for which $X_\infty(\omega) \neq X(\omega)$ is zero. Sure convergence implies almost sure convergence, but the converse is not true. Thus, almost sure convergence is weaker than the familiar sure convergence.

The definition of almost sure convergence above can be unwieldy to work with and for our subsequent analysis we will use the alternative, but equivalent definition.

Almost Sure convergence

$$X_n \xrightarrow{\text{a.s.}} X$$

if and only if for every $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{k\geqslant n} |X_k - X| > \varepsilon\right) = 0 \ .$$

Hence, the last limit can be used as an alternative definition of almost sure convergence.

We give a proof for the interested student.

proof: Define the event

$$A_n(\varepsilon) \overset{\text{def}}{=} \cup_{k\geqslant n}\{\omega : |X_k - X| > \varepsilon\}$$

and convince yourself of the following properties.

1. $A_1(\varepsilon), A_2(\varepsilon), \cdots$ is a decreasing sequence of events (with increasing n we remove more and more sets from a big union) with limit

$$A(\varepsilon) = \cap_{n=1}^{\infty} A_n(\varepsilon) = \{\omega : |X_\infty - X| > \varepsilon\}$$

2. $A_n(\varepsilon) \subseteq A_n(\varepsilon')$ for $\varepsilon > \varepsilon'$. In other words, $A_n(\varepsilon)$ is a decreasing function of ε.

3. $$\{\omega : \lim_{n\uparrow\infty} X_n \neq X\} = \cup_{m=1}^{\infty} A(1/m)$$

This is because $\cup_{m=1}^{n} A(1/m)$ is an increasing sequence of events with limit $A(0)$ corresponding to the ω for which $X_\infty \neq X$.

4. $$A_n(\varepsilon) = \{\omega : \sup_{k\geqslant n} |X_n - X| > \varepsilon\},$$

where sup is the smallest upper bound and is same as max when the maximum of the set exits [1]. Note that $\sup_{k\geqslant n} Y_k$ is a decreasing sequence as seen from Figure 5.1.

Using these properties we can now show the following. Assume

$$\mathbb{P}\left(\omega : \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\right) = 0,$$

[1] For example, sup of $[0,1]$ is 1 and is the same as the maximum of the set. The maximum does not exist for the set $(0,1)$, because 1 does not below to the set, but the sup is still 1, because this is the smallest number which bounds the set $(0,1)$ from above.

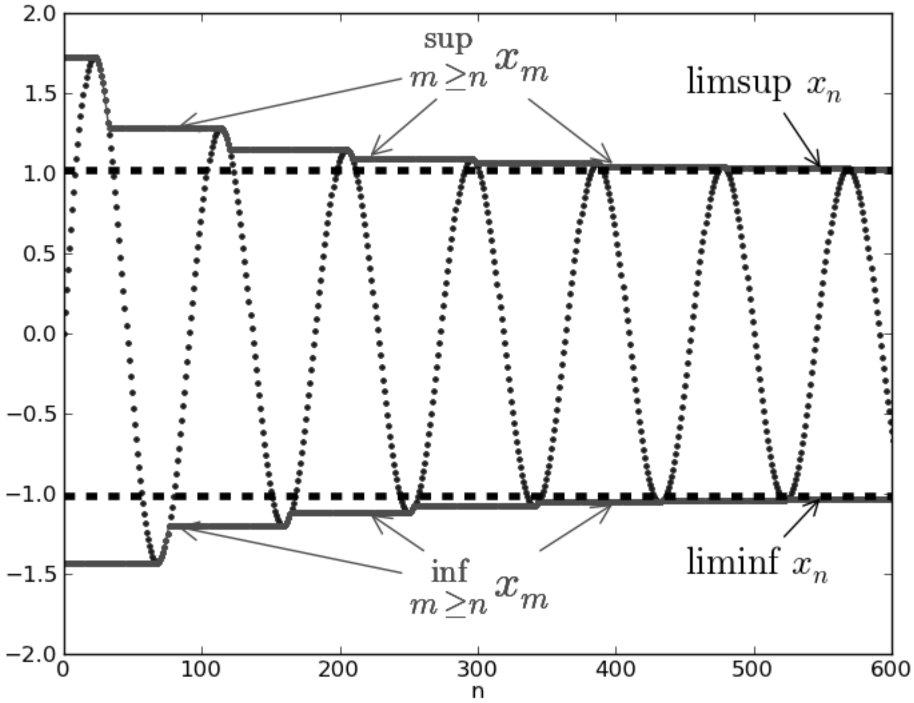


Figure 5.1: Supremum and infimum of the sequence x_m (source Wikipedia).

then

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\left(\sup_{k\geqslant n} |X_k - X| > \varepsilon\right) &= \lim_{n\to\infty} \mathbb{P}\left(A_n(\varepsilon)\right) \\
&= \mathbb{P}\left(\lim_{n\to\infty} A_n(\varepsilon)\right) \\
&= \mathbb{P}(A(\varepsilon)) \\
&\leqslant \mathbb{P}(A(1/m),\ m > \varepsilon^{-1}) \\
&\leqslant \mathbb{P}(\cup_{m=1}^{\infty} A(1/m)) = \mathbb{P}\left(\lim_{n\to\infty} X_n \neq X\right) = 0
\end{aligned}$$

Conversely, assume that for every $\varepsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{k\geqslant n} |X_k - X| > \varepsilon\right) = \mathbb{P}(A(\varepsilon)) = 0,$$

then

$$\begin{aligned}
\mathbb{P}\left(\lim_{n\to\infty} X_n \neq X\right) &= \mathbb{P}(\cup_{m=1}^{\infty} A(1/m)) \\
&\leqslant \sum_{m=1}^{\infty} \mathbb{P}(A(1/m)) = \sum_{m=1}^{\infty} 0 = 0
\end{aligned}$$

where in the last equation we used $\mathbb{P}(A(1/m)) = \mathbb{P}(A(\varepsilon_m)) = 0$ for all $\varepsilon_m > 0$.

□

Convergence in Probability

We now consider a convergence which is weaker than almost sure convergence.

Definition

The sequence of random variables $X_1, X_2, \ldots$ **converges in probability** to a random variable X if, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

This is usually written as

$$X_n \xrightarrow{\mathbb{P}} X.$$

Example

$X_1, X_2, \ldots$ are independent Uniform $(0, \theta)$ variables. Let $Y_n = \max(X_1, \ldots, X_n)$ for $n = 1, 2, \ldots$. Then it can be shown that

$$F_{Y_n}(y) = \begin{cases} \left(\frac{y}{\theta}\right)^n, & 0 < y < \theta \\ 1, & y \geq \theta \end{cases}$$

and

$$f_{Y_n}(y) = \frac{ny^{n-1}}{\theta^n}, \quad 0 < y < \theta.$$

Show that $Y_n \xrightarrow{\mathbb{P}} \theta$.

For $0 < \varepsilon < \theta$,

$$\begin{aligned} \mathbb{P}(|Y_n - \theta| > \varepsilon) &= \mathbb{P}(Y_n < \theta - \varepsilon) \\ &= \left(\frac{\theta - \varepsilon}{\theta}\right)^n \\ &\to 0 \text{ as } n \to \infty \end{aligned}$$

For $\varepsilon > \theta$, $\mathbb{P}(|Y_n - \theta| > \varepsilon) = 0$ for all $n \geq 1$, so

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - \theta| > \varepsilon) = 0$$

$$\therefore Y_n \xrightarrow{\mathbb{P}} \theta$$

In the last example it is actually easy to show that $Y_n \xrightarrow{\text{a.s.}} \theta$ as well. Consider the fact that for any n

$$\theta \geqslant Y_{n+1} \geqslant Y_n,$$

that is Y_n is monotonically increasing, but bounded from above by θ. Hence,

$$|Y_n - \theta| = \theta - Y_n \geqslant \theta - \sup_{k \geqslant n} Y_k = \sup_{k \geqslant n} |Y_k - \theta|$$

Hence, the event $\sup_{k \geqslant n} |Y_k - \theta| > \varepsilon$ implies the event $|Y_n - \theta| > \varepsilon$ and

$$\mathbb{P}(|Y_n - \theta| > \varepsilon) \geqslant \mathbb{P}(\sup_{k \geqslant n} |Y_k - \theta| > \varepsilon)$$

so that $\mathbb{P}(\sup_{k \geqslant n} |Y_k - \theta| > \varepsilon)$ is squashed to zero from above as $n \uparrow \infty$.

Example

For $n = 1, 2, \ldots, Y_n \sim N(\mu, \sigma_n^2)$ and suppose $\lim\limits_{n \to \infty} \sigma_n = 0$.

Show that $Y_n \xrightarrow{P} \mu$.

For any $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}(|Y_n - \mu| > \varepsilon) &= \mathbb{P}(Y_n < \mu - \varepsilon) + \mathbb{P}(Y_n > \mu + \varepsilon) \\ &= \int_{-\infty}^{-\varepsilon/\sigma_n} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y^2/2} \mathrm{d}y + \int_{\varepsilon/\sigma_n}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y^2/2} \mathrm{d}y\end{aligned}$$

$\to 0$ as $n \to \infty$. Thus $Y_n \xrightarrow{\mathbb{P}} \mu$.

We mentioned that convergence in probability is weaker than almost sure convergence. The reason for this is that almost sure convergence, that is, $\mathbb{P}(\sup_{k \geqslant n} |X_k - X| > \varepsilon) \downarrow 0$, implies convergence in probability as seen from:

$$\sup_{k \geqslant n} |X_k - X| \geqslant |X_n - X| \qquad \Rightarrow \qquad \{\sup_{k \geqslant n} |X_k - X| > \varepsilon\} \supseteq \{|X_n - X| > \varepsilon\}$$

The last in turn implies that $\mathbb{P}(|X_n - X| > \varepsilon)$ is squeezed to zero:

$$\mathbb{P}(|X_n - X| > \varepsilon) \leqslant \mathbb{P}(\sup_{k \geqslant n} |X_k - X| > \varepsilon) \downarrow 0, \qquad n \uparrow \infty\,.$$

Thus, we have the result.

Almost sure convergence implies convergence in probability

$$X_n \xrightarrow{\text{a.s.}} X \Longrightarrow X_n \xrightarrow{\mathbb{P}} X$$

and

$$X_n \xrightarrow{\text{a.s.}} 0 \Longleftrightarrow \sup_{k \geqslant n} |X_k| \xrightarrow{\mathbb{P}} 0$$

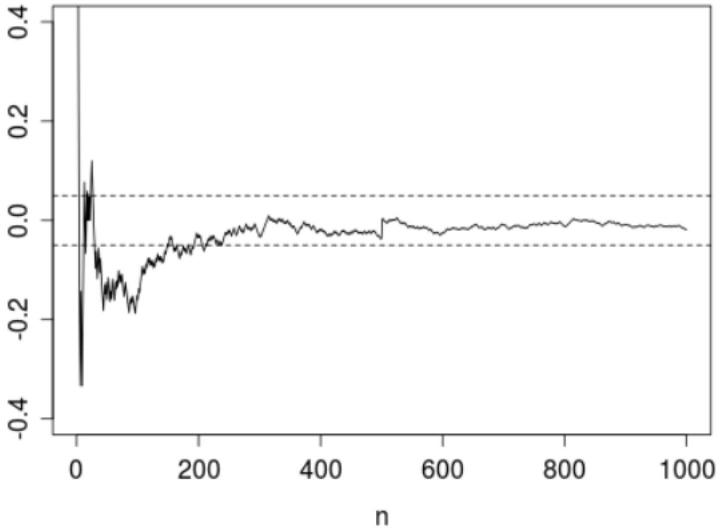


Figure 5.2: Almost sure convergence

The essential difference between almost sure convergence and convergence in probability is that almost sure convergence is a property of the entire sequence $X_1, X_2, \ldots$, because the distribution of $\sup_{k \geqslant n} |X_k - X|$ depends on the joint pdf

$$f_{X_n, X_{n+1}, \ldots}(x_n, x_{n+1}, \ldots),$$

while the convergence in probability is a property of X_n only and hence of the marginal (as opposed to the joint) pdf

$$f_{X_n}(x_n)$$

The difference can be visualized as follows. On Figure 5.2 we have depicted $X_n \xrightarrow{\text{a.s.}} 0$. Here as n gets larger and larger the probability of the random noodle/snake/path straying far away from the strip (the band $-\varepsilon < X_n < \varepsilon$) vanishes as $n \uparrow \infty$.

In contrast, on Figure 5.3 we have depicted many different realizations of the random path/snake $X_1, X_2, \ldots$. Here $X_n \xrightarrow{\mathbb{P}} 0$ means that the proportion of noodles leaving the strip goes to zero as $n \uparrow \infty$. This does not prevent a particular noodle from straying far away from the bands. We only want the proportion of these rogue noodles to get smaller and smaller with increasing n. There is no attempt to control how far a particular noodle strays from the strip.

The next example is only for the interested students and will not be assessable. It is an example of a sequence of random variable which converges in probability, but not almost surely.

Example

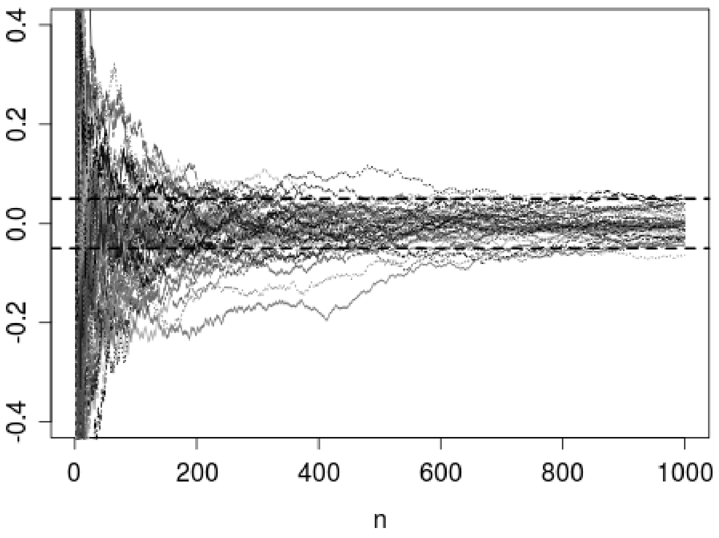


Figure 5.3: Convergence in probability

Consider the iid sequence $X_1, X_2, \ldots$ with

$$\mathbb{P}(X_n = 1) = \frac{1}{n} = 1 - \mathbb{P}(X_n = 0)$$

Show that $X_n \xrightarrow{\mathbb{P}} 0$

This one is easy as:

$$\mathbb{P}(|X_n - 0| > \varepsilon) = \mathbb{P}(X_n = 1) = 1/n \to 0$$

Does $X_n \xrightarrow{\text{a.s.}} 0$?

This one is a bit tricky. Consider the distribution of

$$Y_n = \sup_{k \geqslant n} |X_k - 0| = \sup_{k \geqslant n} X_k$$

$$\begin{aligned}
F_{Y_n}(\varepsilon) = \mathbb{P}(\sup_{k \geqslant n} X_k \leqslant \varepsilon) &= \mathbb{P}(X_n \leqslant \varepsilon, X_{n+1} \leqslant \varepsilon, \ldots) \\
&= \mathbb{P}(X_n \leqslant \varepsilon) \times \mathbb{P}(X_{n+1} \leqslant \varepsilon) \times \cdots \quad \text{(using independence)} \\
&= \lim_{m \uparrow \infty} \prod_{k=n}^{m} \mathbb{P}(X_k \leqslant \varepsilon) \\
&= \lim_{m \uparrow \infty} \prod_{k=n}^{m} \left(1 - \frac{1}{k}\right), \qquad (\varepsilon < 1) \\
&= \lim_{m \uparrow \infty} \frac{n-1}{n} \times \frac{n}{n+1} \times \frac{n+1}{n+2} \times \cdots \times \frac{m-1}{m} \\
&= \lim_{m \uparrow \infty} \frac{n-1}{m} = 0, \qquad \text{for any } n = 1, 2, \ldots .
\end{aligned}$$

It follows that for any $0 < \varepsilon < 1$ and all $n \geqslant 1$

$$\mathbb{P}(\sup_{k \geqslant n} |X_k - 0| > \varepsilon) = 1$$

Thus, it is **not** true that $X_n \xrightarrow{\text{a.s.}} 0$, that is,

$$\lim_{n \uparrow \infty} \mathbb{P}(\sup_{k \geqslant n} |X_k - 0| > \varepsilon) \neq 0$$

Convergence in Distribution

Convergence in probability captures the concept of a sequence of random variables approaching a random variable X. But often we are interested in the probability distribution of the random variable rather than the value that it takes: we are interested in whether the distributions of the X_i converge to the distribution of some random variable X.

convergence in distribution

Let $X_1, X_2, \ldots$ be a sequence of random variables. We say that X_n **converges in distribution** to X if

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$$

for all x where F_X is continuous. A common shorthand is

$$X_n \xrightarrow{\text{d}} X.$$

We say that F_X is the **limiting distribution** of X_n.

This differs subtly (but importantly) from the idea of random variables X_i converging to the random variable X. Convergence in probability is concerned with whether the actual values of the random variables (the x_i) converge. Convergence in distribution, in contrast, is concerned with whether the distributions (the $F_{X_i}(x)$) converge.

Convergence in distribution allows us to make approximate probability statements about X_n, for large n, if we can derive the limiting distribution $F_X(x)$.

Example

Suppose that $\mathbb{P}(X_n = x) = 1/n$ for $x = 1, \ldots, n$. Set $Y_n = X_n/n$.

Show that $Y_n \xrightarrow{\mathrm{d}} Y \sim \mathsf{U}(0,1)$

Here we have for $0 \leqslant y \leqslant 1$

$$F_{Y_n}(y) = \mathbb{P}(Y_n \leqslant y) = \mathbb{P}(X_n \leqslant yn)$$
$$= \frac{\lfloor yn \rfloor}{n} \to y = \mathbb{P}(U \leqslant y) = F_U(y)$$

The last convergence follows from the squeeze principle:

$$y\frac{n-1}{n} \leqslant \frac{\lfloor yn \rfloor}{n} \leqslant y\frac{n+1}{n}$$

In establishing convergence in distribution one can frequently use the following

Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables, each with mgf $M_{X_i}(t)$. Furthermore, suppose that

$$\lim_{n\to\infty} M_{X_n}(t) = M_X(t)$$

If $M_X(t)$ is a moment generating function then there is unique F_X (which gives a random variable X) whose moments are determined by $M_X(t)$ and for all points of continuity $F_X(x)$ we have

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

Example

Let $X_1, X_2, \ldots$ be independent Bernoulli random variables with success probability 1/2, representing the outcomes of the fair coin tosses. Define new random variables $Y_1, Y_2, \ldots$ as

$$Y_n = \sum_{k=1}^{n} X_k \left(\frac{1}{2}\right)^k, \qquad n = 1, 2, \ldots.$$

Show that $Y_n \xrightarrow{\mathrm{d}} Y \sim U(0,1)$

We have

$$\mathbb{E}\mathrm{e}^{-uY_n} = \prod_{k=1}^{n} \mathbb{E}[\mathrm{e}^{-uX_k/2^k}] = 2^{-n}\prod_{k=1}^{n}(1+\mathrm{e}^{-u/2^k})$$

$$(1-\mathrm{e}^{-u/2^n})\prod_{k=1}^{n}(1+\mathrm{e}^{-u/2^k}) = 1-\mathrm{e}^{-u}, \text{ collapsing product!}$$

$$\mathbb{E}\mathrm{e}^{-uY_n} = 2^{-n}\frac{1-\mathrm{e}^{-u}}{1-\mathrm{e}^{-u/2^n}}$$
$$= (1-\mathrm{e}^{-u})\frac{1/2^n}{1-\mathrm{e}^{-u/2^n}}\ .$$

It thus follows that

$$\lim_{n\uparrow\infty} \mathbb{E}\mathrm{e}^{-uY_n} = (1-\mathrm{e}^{-u}) \lim_{n\uparrow\infty} \frac{1/2^n}{1-\mathrm{e}^{-u/2^n}} = \frac{1-\mathrm{e}^{-u}}{u}$$

using L'Hopital's rule $\lim_{n\uparrow\infty} \frac{1/2^n}{1-\mathrm{e}^{-u/2^n}} = \lim_{n\uparrow\infty} \frac{\frac{\mathrm{d}}{\mathrm{d}n}(1/2^n)}{u\mathrm{e}^{-u/2^n}\frac{\mathrm{d}}{\mathrm{d}n}(1/2^n)}$. We recognize $\frac{1-\mathrm{e}^{-u}}{u}$ as the moment generating function of a random variable $U \sim \mathsf{U}[0,1]$, evaluated at $-u$. Hence, $Y_n \xrightarrow{\mathrm{d}} Y \sim U(0,1)$.

Convergence in L^p-norm

Convergence in L^p-mean

The sequence of numerical random variables $X_1, X_2, \ldots$ is said to converge **in L^p-norm** to a numerical random variable X, denoted $X_n \xrightarrow{L^p} X$, if

$$\lim_{n\to\infty} \mathbb{E}|X_n - X|^p = 0 \,,$$

Convergence in L^2-norm is often called **mean square convergence**.

Example

Suppose $X_1, X_2, \ldots$ are independent, each with mean μ and variance $0 < \sigma^2 < \infty$.

Show that $\bar{X}_n \xrightarrow{L^2} \mu$

We have

$$\mathbb{E}[(\bar{X}-\mu)^2] = \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} \to 0 \,.$$

The next example shows that we can have a sequence converging in mean, but not almost surely. Thus, the two types of convergences are quite distinct.

Example

Recall the example in which we have the iid sequence $X_1, X_2, \ldots$ with

$$\mathbb{P}(X_n = 1) = \frac{1}{n} = 1 - \mathbb{P}(X_n = 0)$$

We showed that $X_n \xrightarrow{\text{a.s.}} 0$ is not true. However, we do have $X_n \xrightarrow{L^1} 0$:

$$\mathbb{E}|X_n - 0| = 1 \times \frac{1}{n} + 0 \times (1 - 1/n) = \frac{1}{n} \to 0$$

Complete convergence

For completeness (no pun intended) we will mention one last type of convergence.

Complete Convergence

A sequence of random variables $X_1, X_2, \ldots$ is said to converge **completely** to X, denoted

$$X_n \xrightarrow{\text{cpl.}} X,$$

if

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \varepsilon) < \infty \quad \text{for all } \varepsilon > 0 \ .$$

Example

Suppose $\mathbb{P}(X_n = n^5) = 1/n^2$ and $\mathbb{P}(X_n = 0) = 1 - 1/n^2$. Then, we have

$$\begin{aligned} \sum_n \mathbb{P}(|X_n - 0| > \varepsilon) &= \sum_n \mathbb{P}(X_n = n^5) \\ &= \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} < \infty \end{aligned}$$

Hence, by definition, $X_n \xrightarrow{\text{cpl.}} 0$.

Note that if $\sum_n \mathbb{P}(|X_n - X| > \varepsilon) < \infty$, then by first year Calculus we know that $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$. Thus, complete convergence implies convergence in probability.

It is not obvious, but it also implies almost sure convergence (which implies convergence in probability).

Complete and Almost Sure Convergence

$$X_n \xrightarrow{\text{cpl.}} X \Longrightarrow X_n \xrightarrow{\text{a.s.}} X$$

We give a proof for the interested student.

proof: Recall that $\mathbb{P}(\cup_k A_k) \leqslant \sum_k \mathbb{P}(A_k)$ is one of the defining properties of $\mathbb{P}(\cdot)$.

We can bound the criterion for almost sure convergence as follows:

$$
\begin{aligned}
\mathbb{P}(\sup_{k\geqslant n}|X_k - X| > \varepsilon) &= \mathbb{P}(\cup_{k\geqslant n}\{|X_k - X| > \varepsilon\}) \\
&\leqslant \sum_{k\geqslant n}\mathbb{P}(\{|X_k - X| > \varepsilon\}) \qquad \text{by union property of } \mathbb{P}(\cdot) \\
&\leqslant \underbrace{\sum_{k=1}^{\infty}\mathbb{P}(|X_k - X| > \varepsilon)}_{=c<\infty \text{ from } X_n \xrightarrow{\text{cpl.}} X} - \sum_{k=1}^{n-1}\mathbb{P}(|X_k - X| > \varepsilon) \\
&\leqslant c - \sum_{k=1}^{n-1}\mathbb{P}(|X_k - X| > \varepsilon) \to c - c = 0, \qquad n \uparrow \infty
\end{aligned}
$$

Hence, by definition $X_n \xrightarrow{\text{a.s.}} X$.

Convergence Relationships

The different types of convergence form a big family with close ties to each other. The most general relationships among the various modes of convergence for numerical random variables are shown on the following hierarchical diagram.

$$\begin{tikzcd}
\boxed{X_n \xrightarrow{\text{cpl.}} X} \arrow[r, Rightarrow] & \boxed{X_n \xrightarrow{\text{a.s.}} X} \arrow[d, Rightarrow] & \\
 & \boxed{X_n \xrightarrow{\mathbb{P}} X} \arrow[r, Rightarrow] & \boxed{X_n \xrightarrow{d} X} \\
\boxed{X_n \xrightarrow{L^p} X} \arrow[r, Rightarrow, "p\geqslant q"] & \boxed{X_n \xrightarrow{L^q} X} \arrow[u, Rightarrow] &
\end{tikzcd}$$

We are now going to examine this diagram in some detail. We have already seen that $X_n \xrightarrow{\text{a.s.}} X$ implies $X_n \xrightarrow{\mathbb{P}} X$.

Convergence in probability and distribution

First note that convergence in distribution does not imply convergence in probability (unless additional assumptions are imposed).

Example

Suppose $X_n = 1 - X$, where $X \sim \mathsf{U}(0,1)$. Then,

$$\mathbb{P}(X_n \leqslant x) = \mathbb{P}(1 - X \leqslant x) = \mathbb{P}(X \geqslant 1 - x) = 1 - (1 - x) = x$$

so that $F_{X_n}(x)$ is the cdf of the uniform distribution for all n. Trivially, we have

$$X_n \xrightarrow{\text{d}} X \sim \mathsf{U}(0,1)$$

However,

$$\begin{aligned}\mathbb{P}(|X_n - X| \geqslant \varepsilon) &= \mathbb{P}(|1 - 2X| > \varepsilon)\\ &= 1 - \left(\frac{\varepsilon + 1}{2} - \frac{1-\varepsilon}{2}\right)\\ &= 1 - \varepsilon \not\to 0\end{aligned}$$

Hence, $X_n \overset{\mathbb{P}}{\not\longrightarrow} X$.

We now show that the converse is true.

Convergence in probability implies convergence in distribution

$$X_n \xrightarrow{\mathbb{P}} X \Longrightarrow X_n \xrightarrow{\mathrm{d}} X$$

Before we proceed with a proof, we will take note of the following curious **conjunction fallacy** made famous in psychology.

`Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy.`

`As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.`

`Which is more probable?`

`1. Linda is a bank teller.`

`2. Linda is a bank teller and is active in the feminist movement.`

Most people will choose 2, but they would be wrong. The probability of two events occurring together (in "conjunction") is always less than or equal to the probability of either one occurring alone. Formally,

$$\mathbb{P}(A \cap B) \leqslant \mathbb{P}(A)$$

We now proceed with our proof.

proof: First note that

$$\begin{aligned}F_{X_n}(x) = \mathbb{P}(X_n \leqslant x) &= \mathbb{P}(X_n \leqslant x, |X_n - X| > \varepsilon) + \mathbb{P}(X_n \leqslant x, |X_n - X| \leqslant \varepsilon)\\ &\leqslant \mathbb{P}(|X_n - X| > \varepsilon) + \mathbb{P}(X_n \leqslant x, |X - X_n| \leqslant \varepsilon) \quad \text{by conjunction fallacy}\\ \mathbb{P}(X_n \leqslant x, |X_n - X| \leqslant \varepsilon) &\leqslant \mathbb{P}(X_n \leqslant x, X \leqslant X_n + \varepsilon)\\ &\leqslant \mathbb{P}(X_n \leqslant x, X \leqslant x + \varepsilon) \leqslant \mathbb{P}(X \leqslant x + \varepsilon)\\ \therefore F_{X_n}(x) &\leqslant \mathbb{P}(|X_n - X| > \varepsilon) + \mathbb{P}(X \leqslant x + \varepsilon)\end{aligned}$$

Now, in the arguments above we can switch the roles of X_n and X (there is a symmetry) to deduce the analogous result:

$$F_X(x) \leqslant \mathbb{P}(|X - X_n| > \varepsilon) + \underbrace{\mathbb{P}(X_n \leqslant x + \varepsilon)}_{F_{X_n}(x+\varepsilon)}$$

Therefore, making the switch $x \to x - \varepsilon$ gives

$$F_X(x - \varepsilon) \leqslant \mathbb{P}(|X - X_n| > \varepsilon) + F_{X_n}(x)$$

and putting it all together

$$F_X(x - \varepsilon) - \mathbb{P}(|X - X_n| > \varepsilon) \leqslant F_{X_n}(x) \leqslant \mathbb{P}(|X_n - X| > \varepsilon) + \mathbb{P}(X \leqslant x + \varepsilon)$$

Taking limits on both sides yields for any $\varepsilon > 0$:

$$F_X(x - \varepsilon) \leqslant \lim_{n\uparrow\infty} F_{X_n}(x) \leqslant \mathbb{P}(X \leqslant x + \varepsilon) = F_X(x + \varepsilon)$$

Now if x is a point of continuity, then

$$\lim_{\varepsilon\downarrow 0} F_X(x \pm \varepsilon) = F_X(x)$$

Hence, by taking the limit on both sides as $\varepsilon \downarrow 0$ we deduce by the squeeze principle that

$$\lim_{n\uparrow\infty} F_{X_n}(x) = F_X(x)$$

at points x where $F_X(x)$ is continuous. The last agrees with the definition of convergence in probability. □

Convergence in L^1 mean and convergence in probability

First note that convergence in probability does not imply convergence in L^1 mean (unless additional assumptions are imposed).

Example

Suppose $\mathbb{P}(X_n = n^5) = 1/n^2$ and $\mathbb{P}(X_n = 0) = 1 - 1/n^2$.

Show that $X_n \xrightarrow{\mathbb{P}} 0$.

Does X_n converge to X in L^1 mean?

solution: We have

$$\begin{aligned}\mathbb{P}(|X_n - 0| > \varepsilon) &= \mathbb{P}(X_n > \varepsilon) \\ &= \mathbb{P}(X_n = n^5) \\ &= 1/n^2 \to 0\end{aligned}$$

Hence, by definition $X_n \xrightarrow{\mathbb{P}} 0$. However,

$$\begin{aligned}\mathbb{E}|X_n - 0| &= n^5 \times \frac{1}{n^2} + 0 \times \mathbb{P}(X_n = 0) \\ &= n^3 \not\to 0\end{aligned}$$

So the L_1 mean blows up!

The converse result is however always true.

Convergence in L^1 mean implies convergence in probability

$$X_n \xrightarrow{L^1} X \Longrightarrow X_n \xrightarrow{\mathbb{P}} X$$

proof: We have

$$\begin{aligned}\mathbb{P}(|X_n - X| > \varepsilon) &\leqslant \frac{\mathbb{E}|X_n - X|}{\varepsilon} \quad \text{Chebyshev's inequality} \\ &\leqslant \text{constant} \times \mathbb{E}|X_n - X| \to 0\end{aligned}$$

Hence, $\mathbb{P}(|X_n - X| > \varepsilon)$ is forced to converge to 0.

In assignment 2 you will be asked to prove that

$$\mathbb{E}|X|^u \leqslant (\mathbb{E}|X|^s)^{u/s} \qquad s \geqslant u > 0$$

This implies the following result.

Convergence in L^p mean

For any $p \geqslant q \geqslant 1$, we have

$$X_n \xrightarrow{L^p} X \Longrightarrow X_n \xrightarrow{L^q} X$$

Example

Suppose $X_1, X_2, \ldots$ are independent, each with mean μ and variance $0 < \sigma^2 < \infty$. Then, we know

$$\bar{X}_n \xrightarrow{L^2} \mu\ ,$$

which implies

$$\bar{X}_n \xrightarrow{L^2} \mu \Longrightarrow \bar{X}_n \xrightarrow{L^1} \mu \Longrightarrow \bar{X}_n \xrightarrow{\mathbb{P}} \mu$$

We next give an example showing that in general if $p > q \geqslant 1$, then $X_n \xrightarrow{L^q} X$ does **not** imply $X_n \xrightarrow{L^p} X$.

Example

Assume $p > q \geqslant 1$. Let $\mathbb{P}(X_n = n) = \frac{1}{n^{(p+q)/2}} = 1 - \mathbb{P}(X_n = 0)$, then

$$\mathbb{E}|X_n - 0|^q = n^q / n^{(p+q)/2} = 1/n^{(p-q)/2}$$

and we will get convergence to 0 for $p > q$, but at the same time

$$\mathbb{E}|X_n - 0|^p = 1/n^{(q-p)/2} \to \infty, \qquad n \uparrow \infty$$

As a result, $X_n \xrightarrow{L^q} X$, but $X_n \overset{L^p}{\not\to} X$.

Almost Sure convergence and in L^1 mean

In general, the these two types of convergences are quite distinct as the next example shows.

Example

Consider again the example with $\mathbb{P}(X_n = n^5) = 1/n^2$ and $\mathbb{P}(X_n = 0) = 1 - 1/n^2$. We showed that $X_n \xrightarrow{L^1} 0$ is false, because $X_n = n^5$ can take arbitrarily large values as $n \uparrow \infty$ and this forces the expectation to blow up!

Show that $X_n \xrightarrow{\text{a.s.}} 0$.

This is easy, since

$$\sum_n \mathbb{P}(|X_n - 0| > \varepsilon) = \sum_n \frac{1}{n^2} = \frac{\pi^2}{6} < \infty$$

implies that $X_n \xrightarrow{\text{cpl.}} 0$, which we know in turn implies $X_n \xrightarrow{\text{a.s.}} 0$.

It is easy to remember the relationships between modes of convergences by keeping the diagram at the beginning in your mind's eye.

Converse Results on Modes of Convergence

In this section we explore conditions under which we can reverse the direction of the $\Rightarrow$ in the diagram in the previous section.

Convergence in distribution to a constant

Convergence in distribution to a constant
If c is a constant, then

$$X_n \xrightarrow{\text{d}} c \Longrightarrow X_n \xrightarrow{\mathbb{P}} c$$

The proof is quite simple and all student should be able to follow it.

proof: We are given that

$$\lim_{n\uparrow\infty} F_{X_n}(x) = \lim_{n\uparrow\infty} \mathbb{P}(X_n \leqslant x) = \begin{cases} 1 & x \geqslant c \\ 0 & x < c \end{cases}$$

We now try to bound and squash to zero the criterion for convergence in probability:

$$\begin{aligned} \mathbb{P}(|X_n - c| > \varepsilon) &= \mathbb{P}(X_n > c+\varepsilon) + \mathbb{P}(X_n < c-\varepsilon) \\ &\leqslant 1 - F_{X_n}(c+\varepsilon) + F_{X_n}(c-\varepsilon) \to 1 - 1 + 0 = 0, \quad n\uparrow\infty \end{aligned}$$

Since this is true for any small $\varepsilon > 0$, we have $X_n \xrightarrow{\mathbb{P}} c$.

Convergence in Probability

While convergence in probability does not in general imply convergence in mean, assuming boundedness we have the following converse.

Convergence for Bounded random variables
If $\mathbb{P}(|X_k| \leqslant c) = 1$ for all k, then for any $p \geqslant 1$

$$X_n \xrightarrow{\mathbb{P}} X \Longrightarrow X_n \xrightarrow{L^p} X$$

proof: First note that since $X_n \xrightarrow{\text{d}} X$, we also have that X is bounded in the sense

$$\mathbb{P}(|X| \leqslant c) = \lim_{n\uparrow\infty} \mathbb{P}(|X_n| \leqslant c) = 1$$

We try to bound the criterion for convergence in mean by smuggling an indicator into the expectation:

$$\begin{aligned} \mathbb{E}|X_n - X|^p &= \mathbb{E}[|X_n - X|^p\,\mathrm{I}_{\{|X_n-X|>\varepsilon\}}] + \mathbb{E}[|X_n - X|^p\,\mathrm{I}_{\{|X_n-X|<\varepsilon\}}] \\ &\leqslant \mathbb{E}[|X_n - X|^p\,\mathrm{I}_{\{|X_n-X|>\varepsilon\}}] + \mathbb{E}[\varepsilon^p\,\mathrm{I}_{\{|X_n-X|<\varepsilon\}}] \\ &\leqslant \mathbb{E}[(|X_n| + |X|)^p\,\mathrm{I}_{\{|X_n-X|>\varepsilon\}}] + \varepsilon^p\,\mathbb{P}(|X_n - X| < \varepsilon) \\ &\leqslant (2c)^p \mathbb{E}[\mathrm{I}_{\{|X_n-X|>\varepsilon\}}] + \varepsilon^p\,\mathbb{P}(|X_n - X| < \varepsilon) \\ &\leqslant (2c)^p \mathbb{P}(|X_n - X| > \varepsilon) + \varepsilon^p \to 0 + \varepsilon^p, \quad n\uparrow\infty \end{aligned}$$

Since this is true for any $\varepsilon > 0$, no matter how small, we conclude that $\mathbb{E}|X_n - X|^p \to 0$ as $\varepsilon \downarrow 0$ and $n \uparrow \infty$.

Finally, of interest is when we can go from $X_n \xrightarrow{\mathbb{P}} X$ to $X \xrightarrow{\text{a.s.}} X$. In general, this is a difficult problem, but we can easily show that if $X_n \xrightarrow{\mathbb{P}} X$, then there is a subsequence of $X_1, X_2, X_3, \ldots$, call it $X_{k_1}, X_{k_2}, X_{k_3}, \ldots$ which converges almost surely to X.

Convergence of Subsequence
Suppose that $X_n \xrightarrow{\mathbb{P}} X$, that is, for every $\varepsilon > 0$ and $\delta > 0$, we can find a large enough n such that

$$\mathbb{P}(|X_n - X| > \varepsilon) < \delta$$

Then,

$$X_{k_n} \xrightarrow{\text{a.s.}} X, \quad n \uparrow \infty,$$

where $k_1 < k_2 < k_3 < \cdots$ is selected such that

$$\mathbb{P}(|X_{k_j} - X| > \varepsilon) < \frac{1}{j^2}$$

proof: Since

$$\sum_j \mathbb{P}(|X_{k_j} - X| > \varepsilon) < \sum_j \frac{1}{j^2} = \frac{\pi^2}{6} < \infty,$$

then $X_{k_n} \xrightarrow{\text{cpl.}} X$, which in turn implies $X_{k_n} \xrightarrow{\text{a.s.}} X$.

Almost Sure convergence

In the case of independent random variables almost sure convergence implies complete convergence.

Independent random variables and convergence
If $X_1, X_2, \ldots$ are independent, then

$$X_n \xrightarrow{\text{a.s.}} X \Longrightarrow X_n \xrightarrow{\text{cpl.}} X$$

We give a proof for the interested student only.

proof:

First verify that the global minimum of the function $\mathrm{e}^x - x - 1$ is 0 and hence $\mathrm{e}^x \geqslant x + 1$ or equivalently

$$\mathrm{e}^{-x} \geqslant 1 - x$$

We now argue by contradiction. Assume that $X_n \xrightarrow{\text{a.s.}} X$ and that we do not have complete convergence, that is, assume

$$\sum_n \mathbb{P}(|X_n - X| > \varepsilon) = \infty$$

Then, under this assumption we show that

$$\mathbb{P}(\sup_{k\geqslant n}|X_n - X| > \varepsilon) = \mathbb{P}(\cup_{k\geqslant n}\{|X_n - X| > \varepsilon\}) = 1$$

instead of the expected 0, contradicting $X_n \xrightarrow{\text{a.s.}} X$. We have

$$\begin{aligned}
1 - \mathbb{P}(\cup_{k\geqslant n}\{|X_n - X| > \varepsilon\}) &= \mathbb{P}(\cap_{k\geqslant n}\{|X_n - X| \leqslant \varepsilon\}) \quad \text{de Morgan Law} \\
&= \prod_{k\geqslant n} \mathbb{P}(|X_n - X| \leqslant \varepsilon) \quad \text{independence} \\
&= \prod_{k\geqslant n} (1 - \mathbb{P}(|X_n - X| > \varepsilon)) \\
&\leqslant \prod_{k\geqslant n} \mathrm{e}^{-\mathbb{P}(|X_n - X| > \varepsilon)}, \quad 1 - x \leqslant \mathrm{e}^{-x} \\
&\leqslant \mathrm{e}^{-\sum_{k\geqslant n} \mathbb{P}(|X_n - X| > \varepsilon)} = \mathrm{e}^{-\infty} = 0
\end{aligned}$$

Hence, it is not true that $\mathbb{P}(\sup_{k\geqslant n}|X_n - X| > \varepsilon) \downarrow 0$ contradicting the assumption $X_n \xrightarrow{\text{a.s.}} X$. The only other logical possibility is that

$$\sum_n \mathbb{P}(|X_n - X| > \varepsilon) < \infty,$$

that is, $X_n \xrightarrow{\text{cpl.}} X$. □

In summary, what we have established is the following diagram with reverse arrow directions.

$$\begin{tikzcd}
\boxed{X_n \xrightarrow{\text{cpl.}} X} & \boxed{X_n \xrightarrow{\text{a.s.}} X} \arrow[l, Rightarrow, "\text{indep.}"'] & \\
{} \arrow[u, Rightarrow] & \boxed{X_n \xrightarrow{\mathbb{P}} X} \arrow[u, Rightarrow, "\text{subseq.}"'] \arrow[l, Rightarrow, "\text{subseq.}"'] \arrow[d, Rightarrow, "|X|<c"] & \boxed{X_n \xrightarrow{d} X} \arrow[l, Rightarrow, "X\text{ const.}"'] \\
\boxed{X_n \xrightarrow{L^p} X} & \boxed{X_n \xrightarrow{L^q} X} \arrow[l, Rightarrow, "|X|<c"'] &
\end{tikzcd}$$

Weak Law of Large Numbers

As mentioned in Chapter 1, a fundamental idea in Statistics is to use a sample to represent an unknown distribution, and to use sample characteristics to estimate the corresponding distributional characteristics. The soundness of this idea is verified by the Weak Law of Large Numbers, describing how the sample average converges to the distributional average as the sample size increases.

Weak Law of Large Numbers

Suppose $X_1, X_2, \ldots$ are independent, each with mean μ and variance $0 < \sigma^2 < \infty$. If

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \ , \text{ then } \bar{X}_n \xrightarrow{\mathbb{P}} \mu;$$

$$\text{that is, for all } \varepsilon > 0 \ , \ \lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) = 0.$$

Example

There are n independent and identically distributed trials, each with probability p of success. Consider the "sample proportion" $\widehat{p}_n = \frac{X}{n}$, where X is the number of successes in the n trials.

Use the Weak Law of Large Numbers to show that $\widehat{p}_n \xrightarrow{\mathbb{P}} p$.

The following result, known as **Slutsky's Theorem**, is useful for establishing convergence in distribution results.

Slutsky's Theorem

Let $X_1, X_2, \ldots$ be a sequence of random variables that converges in distribution to X, that is,

$$X_n \xrightarrow{\mathrm{d}} X.$$

Let $Y_1, Y_2, \ldots$ be another sequence of random variables that converges in probability to a constant c, that is,

$$Y_n \xrightarrow{\mathbb{P}} c.$$

Then,

1. $X_n + Y_n \xrightarrow{\mathrm{d}} X + c$,
2. $X_n Y_n \xrightarrow{\mathrm{d}} cX$

The proof is omitted in these notes, but may be found in advanced texts such as:

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*, New York: John Wiley & Sons.

Example

Suppose that $X_1, X_2, \ldots$ converges in distribution to $X \sim \mathsf{N}(0,1)$, i.e. $X_n \xrightarrow{\mathrm{d}} \mathsf{N}(0,1)$, and suppose that $nY_n \sim \text{Bin}(n, \frac{1}{2})$.

What are the limiting distributions of $X_n + Y_n$ and X_nY_n?

Answer: First by the Weak Law of Large numbers $Y_n \xrightarrow{\mathbb{P}} \frac{1}{2}$. Therefore, from Slutsky's Theorem,

$$X_n + Y_n \xrightarrow{\mathrm{d}} \mathsf{N}(1/2, 1) \quad \text{and} \quad X_nY_n \xrightarrow{\mathrm{d}} \mathsf{N}(0, 1/4).$$

⌟

Strong Law of Large Numbers

The Weak Law corresponds to convergence in probability, while the Strong Law corresponds to almost sure convergence. Since almost sure convergence implies convergence in probability, it is in this sense that we talk about a Strong and Weak Law. The difference between the Strong and Weak Law is the same as that between a.s. convergence and convergence in probability.

Strong Law of Large numbers

Let $X_1, X_2, \ldots$ be independent with common mean $\mathbb{E}[X] = \mu$ and variance $\mathrm{Var}(X) = \sigma^2 < \infty$, then

$$\bar{X}_n \xrightarrow{\text{a.s.}} \mu$$

proof: We give only a proof for the case $X_k \geqslant 0$ for all k. If from the sequence

$$\{\bar{X}_1, \bar{X}_2, \bar{X}_2, \ldots\}$$

we pick up the subsequence

$$\{\bar{X}_1, \bar{X}_4, \bar{X}_9, \bar{X}_{16}, \ldots\} \equiv \{\bar{X}_{j^2}\}$$

for $j = 1, 2, 3, \ldots$, then

$$\mathbb{P}(|\bar{X}_{j^2} - \mu| > \varepsilon) \leqslant \frac{\mathrm{Var}(\bar{X}_{j^2})}{\varepsilon^2} = \frac{\sigma^2}{j^2\varepsilon^2} < \infty$$

Hence,

$$\sum_{j=1}^{\infty} \mathbb{P}(|\bar{X}_{j^2} - \mu| > \varepsilon) = \frac{\pi^2}{6}\frac{\mathrm{Var}(X)}{\varepsilon^2} < \infty$$

and we can conclude that the sub-sequence of $\{\bar{X}_n\}$, namely, $\{\bar{X}_{j^2},\ j = 1, 2, \ldots\}$ converges almost surely to μ. Next, it is clear that for any arbitrary n, we can always find a k so that $k^2 \leqslant n \leqslant (k+1)^2$. For example, taking $k = \lfloor\sqrt{n}\rfloor$ works. Also note that $k \uparrow \infty$ as $n \uparrow \infty$.

Finally, for $k^2 \leqslant n \leqslant (k+1)^2$ it can be shown (see remark below) that

$$\frac{k^2}{(k+1)^2}\bar{X}_{k^2} \leqslant \bar{X}_n \leqslant \bar{X}_{(k+1)^2}\frac{(k+1)^2}{k^2}\ .$$

Both $\bar{X}_{k^2}$ and $\bar{X}_{(k+1)^2}$ converge almost surely to μ as n (and hence k) go to infinity. Therefore, we conclude that $\bar{X}_n \xrightarrow{\text{a.s.}} \mu$.

Remark 5.1 (Justification of inequality) *Observe that for $X_k \geqslant 0$*

$$\begin{aligned} k^2\bar{X}_{k^2} = X_1 + \cdots + X_{k^2} &\leqslant X_1 + \cdots + X_n \qquad (n \geqslant k^2) \\ &\leqslant n\bar{X}_n \leqslant (k+1)^2\bar{X}_n \qquad ((k+1)^2 \geqslant n) \end{aligned}$$

Rearranging the last gives $\frac{k^2}{(k+1)^2}\bar{X}_{k^2} \leqslant \bar{X}_n$. Similarly,

$$k^2\bar{X}_n \overset{k^2\leqslant n}{\leqslant} n\bar{X}_n = X_1 + \cdots + X_n \overset{n\leqslant(k+1)^2}{\leqslant} X_1 + \cdots + X_{(k+1)^2} = (k+1)^2\bar{X}_{(k+1)^2}$$

Hence, $\bar{X}_n \leqslant \bar{X}_{(k+1)^2}\frac{(k+1)^2}{k^2}$.

Central Limit Theorem

For a general random sample $X_1, \ldots, X_n$ it is often of interest to make probability statements about the sample mean $\bar{X}$. The following theorem provides a pathway to approximating such probabilities when n is large.

Central Limit Theorem

Suppose $X_1, X_2, \ldots$ are independent and identically distributed random variables with common mean $\mu = \mathbb{E}(X_i)$ and common variance $\sigma^2 = \text{Var}(X_i) < \infty$. For each $n \geq 1$ let $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$. Then

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\text{d}} Z$$

where $Z \sim \mathsf{N}(0,1)$. It is common to write

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\text{d}} \mathsf{N}(0,1).$$

Note that

$$\mathbb{E}(\bar{X}_n) = \mu, \quad \text{and} \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

So the Central Limit Theorem states that the limiting distribution of any standardised average of independent random variables is the standard Normal or $\mathsf{N}(0,1)$ distribution.

Note that we made **no assumptions about the common distribution of the $\boldsymbol{X_i}$**. This is an important aspect of this result which makes it particularly useful in practice. In fact, the Central Limit Theorem is often considered the single most important result in statistics, and it forms the basis of most of the statistical inference tools that are used by researchers today.

proof:

Our proof will use the MGF approach, which requires that the MGF function exists. This requirement implies that all moments $\mathbb{E}X^k < \infty$, $k = 1, 2, \dots$ are finite, which is much stronger than the assumption in the theorem of a finite second moment only. Nevertheless, we proceed with this proof, because it is one of the easiest.

We will denote the standardized moments

$$\kappa_k \overset{\text{def}}{=} \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^k\right], \quad k = 1, 2, \dots$$

κ_3 is called the skewness of the distribution of X, and κ_4 is called the kurtosis of the distribution of X. The skewness parameter indicates how symmetric the distribution of X is and the kurtosis indicates how fast the tails of the density decay to zero. Without loss of generality we can assume $\mu = 0$ and $\sigma = 1$. Thus, in our case $\kappa_k = \mathbb{E}[X^k]$. Then, if

$$m_X(t) = \mathbb{E}e^{tX}$$

is the MGF of X, it follows from the iid assumption that

$$m_{\sqrt{n}\bar{X}_n}(t) = \left(m_X\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

Hence,

$$\zeta(t) \overset{\text{def}}{=} \ln m_{\sqrt{n}\bar{X}_n}(t) = n \ln m_X\left(\frac{t}{\sqrt{n}}\right)$$

Now, consider the Taylor expansions

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots, \qquad |x| < 1$$

and

$$\begin{aligned} m_X\left(\frac{t}{\sqrt{n}}\right) &= 1 + \mathbb{E}[X]\frac{t}{\sqrt{n}} + \mathbb{E}[X^2]\frac{t^2}{2!n} + \mathbb{E}[X^3]\frac{t^3}{3!n^{3/2}} + \cdots \\ &= 1 + \frac{t^2}{2n} + \mathbb{E}[X^3]\frac{t^3}{6n^{3/2}} + \mathbb{E}[X^4]\frac{t^4}{4!n^2} + \cdots \\ &= 1 + \underbrace{\frac{t^2}{2n} + \kappa_3\frac{t^3}{6n^{3/2}} + \kappa_4\frac{t^4}{4!n^2} + \cdots}_{x}, \end{aligned}$$

where we choose t such that $|t/\sqrt{n}| < \varepsilon$ with $\varepsilon > 0$ small enough that

$$|x| \leqslant \sum_{k=2}^{\infty} |\kappa_k| \frac{\varepsilon^k}{k!} < 1$$

Now, apply both Taylor expansions to $\zeta(t)$ in a nested fashion to obtain:

$$\begin{aligned}\frac{\zeta(t)}{n} &= \left(\frac{t^2}{2n} + \kappa_3\frac{t^3}{6n^{3/2}} + \kappa_4\frac{t^4}{4!n^2} + \cdots\right) - \\ &\quad - \frac{1}{2}\left(\frac{t^2}{2n} + \kappa_3\frac{t^3}{6n^{3/2}} + \kappa_4\frac{t^4}{4!n^2} + \cdots\right)^2 + \\ &\quad + \frac{1}{3}\left(\frac{t^2}{2n} + \kappa_3\frac{t^3}{6n^{3/2}} + \kappa_4\frac{t^4}{4!n^2} + \cdots\right)^3 - \cdots \\ \text{(collect like powers of } n\text{)} &= \frac{t^2}{2n} + \frac{\kappa_3 t^3}{6n^{3/2}} + \frac{t^4}{n^2}\left(\frac{\kappa_4}{4!} - \frac{1}{8}\right) + \cdots\end{aligned}$$

It follows that

$$\zeta(t) = \frac{t^2}{2} + \frac{\kappa_3 t^3}{6n^{1/2}} + \frac{t^4}{n}\left(\frac{\kappa_4}{4!} - \frac{1}{8}\right) + \cdots$$

or alternatively

$$m_{\sqrt{n}\bar{X}_n}(t) = \mathrm{e}^{\frac{t^2}{2} + \frac{\kappa_3 t^3}{6n^{1/2}} + \frac{t^4}{n}\left(\frac{\kappa_4}{4!} - \frac{1}{8}\right) + \cdots} \to \mathrm{e}^{\frac{t^2}{2}}, \quad n \uparrow \infty$$

Hence, $\sqrt{n}\bar{X}_n \xrightarrow{\mathrm{d}} \mathsf{N}(0,1)$. Note that we have used the fact that

$$\left|\frac{\kappa_3 t^3}{6n^{1/2}} + \frac{t^4}{n}\left(\frac{\kappa_4}{4!} - \frac{1}{8}\right) + \cdots\right| \leqslant \text{const.} \sum_{k=1}^{\infty} \frac{1}{n^{k/2}} = \text{const.}\left(\frac{1}{1 - n^{-1/2}} - 1\right)$$

□

The Central Limit Theorem stated above provides the limiting distribution of

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

However, sometimes probabilities involving related quantities such as the sum $\sum_{i=1}^n X_i$ are required. Since $\sum_{i=1}^n X_i = n\bar{X}$ the Central Limit Theorem also applies to the sum of a sequence of random variables. The following result provides alternative forms of the Central Limit Theorem.

Result

Suppose $X_1, X_2, \ldots$ are independent and identically distributed random variables with common mean $\mu = \mathbb{E}(X_i)$ and common variance $\sigma^2 = \mathrm{Var}(X_i) < \infty$. Then the Central Limit Theorem may also be stated in the following alternative forms:

1. $\sqrt{n}(\bar{X} - \mu) \xrightarrow{\mathrm{d}} \mathsf{N}(0, \sigma^2)$,
2. $\frac{\sum_i X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{\mathrm{d}} \mathsf{N}(0,1)$,
3. $\frac{\sum_i X_i - n\mu}{\sqrt{n}} \xrightarrow{\mathrm{d}} \mathsf{N}(0, \sigma^2)$.

Applications of the Central Limit Theorem

In this section we provide some applications of the Central Limit Theorem (CLT). Being the most important theorem in statistics, its applications are many and varied.

The first application is descriptive. The CLT tells us that the sum of many independent small random variables has an approximate normal distribution. Therefore it is plausible that any real-life random variable, formed by the combined effect of many small independent random influences, will be approximately normally distributed. Thus the CLT provides an explanation for the widespread empirical occurrence of the normal distribution.

The CLT also provides a number of very useful normal approximations to common distributions.

Probability calculations about a sample mean

Say we are interested a random variable X. We take a measurement of this variable on each of n randomly selected subjects, giving us n independently and identically distributed (iid) random variables $X_1, X_2, ..., X_n$. $\bar{X}$ is the average of X from this sample.

Thanks to the central limit theorem, we know that the average of a sample from **any random variable** is approximately normally distributed. So if we know μ and σ for this random variable, we can calculate any probability we like about averages of random variables from this unknown distribution.

Example

It is known that Australians have an average weight of about 68kg, and the variance of this quantity is about 256.

We randomly choose 10 Australians. What is an approximate distribution for the average weight of these people? What is the chance that their average weight exceeds 80kg?

From the central limit theorem,

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow{\text{d}} \mathsf{N}(0,1)$$

so we can say that

$$\bar{X} \overset{\text{appr.}}{\sim} \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right) = \mathsf{N}\left(68, \frac{256}{10}\right) = \mathsf{N}(68, 25.6)$$

So using Chapter 3 methods to calculate normal probabilities,

$$\begin{aligned}\mathbb{P}(\bar{X} > 80) &= \mathbb{P}\left(X > \frac{80-68}{\sqrt{25.6}}\right)\\ &\simeq \mathbb{P}(X > 2.37)\\ &\simeq 0.0089\end{aligned}$$

$\lrcorner$

Normal Approximation to the Binomial Distribution

The Central Limit Theorem also allows us to approximate some common distributions by the normal. An example is the binomial distribution.

Central Limit Theorem for Binomial Distribution
Suppose $X \sim \text{Bin}\ (n, p)$. Then

$$\frac{X - np}{\sqrt{np(1-p)}} \xrightarrow{\text{d}} \mathsf{N}(0,1).$$

Proof:

Let $X_1, \dots, X_n$ be a set of independent Bernoulli random variables with parameter p. Then

$$X = \sum_i X_i$$

From the Central Limit Theorem, as it applies to sums of independent random variables,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X - n\mu}{\sigma\sqrt{n}} \le x\right) = \mathbb{P}(Z \le x)$$

where $Z \sim \mathsf{N}(0,1)$ and $\mu = \mathbb{E}(X_i) = p$ and $\sigma^2 = \text{Var}(X_i) = p(1-p)$. The required result follows immediately.

□

The practical ramifications are that probabilities involving binomial random variables with large n can be approximated by normal probabilities. However a slight adjustment, known as a **continuity correction**, is often used to improve the approximation:

Normal Approximation to Binomial Distribution with Continuity Correction

Suppose $X \sim \text{Bin}\ (n, p)$. Then

$$\mathbb{P}(X \leq x) \simeq \mathbb{P}\left(Z \leq \frac{x - np + \frac{1}{2}}{\sqrt{np(1-p)}}\right)$$

where $Z \sim \mathsf{N}(0, 1)$.

The continuity correction is based on the fact that a discrete random variable is being approximated by a continuous random variable.

Example

Adam tosses 25 piece of toast off a roof, and 10 of them land butter side up.

Is this evidence that toast lands butter side down more often than butter side up? *i.e.* is $\mathbb{P}(X \leqslant 10)$ unusually small?

$X \sim \text{Bin}(25, 0.5)$.

We could answer this question by calculating the exact probability, but this would be time consuming. Instead, we use the fact that

$$\frac{X/n - 1/2}{\sqrt{1/(4n)}} \xrightarrow{\text{d}} Z \sim \mathsf{N}(0, 1)$$

The normal approximation to $\mathbb{P}(X \leqslant 10)$ gives:

$$\mathbb{P}\left(Z < \frac{10 - 12.5}{\sqrt{25/4}}\right) = 0.158655...$$

with the correction

$$\mathbb{P}\left(Z < \frac{10 - 12.5 + 1/2}{\sqrt{25/4}}\right) = 0211855...$$

Compare this with the exact answer, obtained from the binomial distribution:

$$\mathbb{P}(X \leqslant 10) = 0.212178111...$$

How large does n need to be for the normal approximation to the binomial distribution to be reasonable?

Recall that how well the central limit theorem works depends on the skewness of the distribution of X, and its kurtosis. In the case of the Bernouilli and binomial distributions, the skewness and kurtosis are a function of p (and skewness is zero for $p = 0.5$, with kurtosis also very small in this case). This means that how well the normal approximation to the binomial works is a function of p, and it is a better approximation as p approaches 0.5.

A useful "rule of thumb" is that the normal approximation to the binomial will work well when n is large enough that both $np > 5$ and $n(1-p) > 5$.

This rule of thumb means that we don't actually need a very large value of n for this "large sample" approximation to work well – if $p = 0.5$, we only need $n = 10$ for the normal approximation to work well. On the other hand, if $p = 0.005$, we would need a sample size of $n = 1000$...

Normal Approximation to the Poisson Distribution

Result

Suppose $X \sim \text{Poisson}(\lambda)$. Then

$$\lim_{\lambda\to\infty} \mathbb{P}\left(\frac{X-\lambda}{\sqrt{\lambda}} \le x\right) = \mathbb{P}(Z \le x)$$

where $Z \sim \mathsf{N}(0,1)$.

This approximation works increasingly well as λ gets large, and it provides a reasonable approximation to most Poisson probabilities for $\lambda > 5$. Note that because X is discrete, a continuity correction will improve the accuracy of this approximation.

Example

Suppose $X \sim \mathsf{Poi}(100)$. Then

$$\mathbb{P}(X = x) = \mathrm{e}^{-100}\frac{100^x}{x!}, \quad x = 0, 1, 2, \dots \ .$$

Use a normal approximation (with continuity correction) to calculate $\mathbb{P}(80 < X < 120)$.

Now $\frac{X-100}{10}$ is approximately $\mathsf{N}(0,1)$.

$$\begin{aligned}
\therefore \mathbb{P}(80 \le X \le 120) &= \mathbb{P}\left(\frac{80-100}{10} \le \frac{X-100}{10} \le \frac{120-100}{10}\right) \\
&\approx \mathbb{P}(-2 \le Z \le 2) \text{ where } Z \sim \mathsf{N}(0,1) \\
&\approx 0.9488...
\end{aligned}$$

With continuity correction we get:

$$\mathbb{P}\left(\frac{80-100+1/2}{10} \le \frac{X-100}{10} \le \frac{120-100+1/2}{10}\right) \approx 0.9485...$$

The exact answer is

$$\mathbb{P}(80 \le X \le 120) = \sum_{x=80}^{120} \mathrm{e}^{-100}\frac{100^x}{x!} = 0.94912...$$

There is also a normal approximation to the gamma distribution, which works via a similar method (but with no need for a continuity correction).

The Delta Method

The Central Limit Theorem provides a large sample approximation to the distribution of $\bar{X}_n$. But what about other functions of a sequence $X_1, X_2, \ldots$? Some special examples include

1. Functions of $\bar{X}$ such as $(\bar{X}_n)^3$ and $\sin^{-1}(\sqrt{\bar{X}_n})$.
2. Functions defined through a non-linear equation such as the solution in α to

$$\bar{X}_n - \tfrac{\alpha-1}{n}\sum_i \ln(X_i) + \Gamma(\alpha) = 0.$$

This second example is particularly important in statistics, as we will see when we study likelihood-based inference in later chapters.

It turns out that these random variable sequences also converge in distribution to a normal random variable. The general technique for establishing such results has become known as the *delta method.* The reason for this name is a bit mysterious, although it seems to be related to notation (δ) often used in Taylor series expressions.

The Delta Method
Let $Y_1, Y_2, \ldots$ be a sequence of random variables such that

$$\frac{\sqrt{n}(Y_n-\theta)}{\sigma} \xrightarrow{\mathrm{d}} Z \sim \mathsf{N}(0,1).$$

Suppose the function g is differentiable in the neighbourhood of θ and $g'(\theta) \neq 0$. Then,

$$\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{\mathrm{d}} \mathsf{N}(0, \sigma^2\,[g'(\theta)]^2).$$

Equivalently, the Delta Method result can be stated as follows:

If

$$Y_n = \theta + \frac{1}{\sqrt{n}}\sigma Z_n + (\text{terms in } \frac{1}{n} \text{ or smaller})$$

where $Z_n \xrightarrow{\mathrm{d}} \mathsf{N}(0,1)$ then

$$g(Y_n) = g(\theta) + \frac{1}{\sqrt{n}}\sigma g'(\theta) Z_n + (\text{terms in } \frac{1}{n} \text{ or smaller})$$

It is often useful in statistics to use this latter notation, where one expands a statistic into a constant term and terms which vanish at different rates as n increases.

proof:

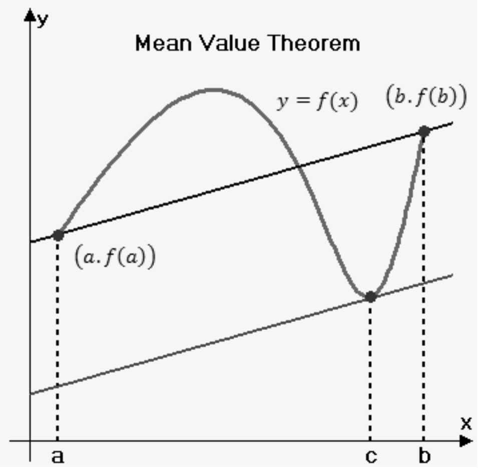


Figure 5.4: Mean Value Theorem

First recall the mean value Theorem from First Year Calculus, which states that if f is continuous in the interval $[a,b]$ and differentiable on (a,b), then there is a number $c \in (a,b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a} \ ,$$

see Figure 5.4. On this figure we can clearly see that there exists a tangent line at c with slope $f'(c)$, which is parallel to the line through the points $(a, f(a))$ and $(b, f(b))$ and with slope $\frac{f(b)-f(a)}{b-a}$.

First, note that

$$\begin{aligned} Y_n - \theta &= \frac{\sigma}{\sqrt{n}} \times \frac{\sqrt{n}(Y_n - \theta)}{\sigma} \\ &= \underbrace{\frac{\sigma}{\sqrt{n}}}_{\xrightarrow{\mathbb{P}} 0} \times \underbrace{\frac{\sqrt{n}(Y_n - \theta)}{\sigma}}_{\xrightarrow{\mathrm{d}} Z \sim \mathsf{N}(0,1)} \xrightarrow{\mathrm{d}} 0 \times Z = 0 \text{ by Slutky's theorem} \end{aligned}$$

In other words, $Y_n - \theta \xrightarrow{\mathrm{d}} 0$, but convergence in distribution to a constant implies convergence in probability. Hence, we have established that

$$Y_n \xrightarrow{\mathbb{P}} \theta \ .$$

Second, by the mean value theorem we can write

$$g'(\vartheta_n) = \frac{g(Y_n) - g(\theta)}{Y_n - \theta} \tag{5.1}$$

for some ϑ_n between θ and Y_n. Note that the mean value result is valid if $\vartheta_n \in (\theta, Y_n)$ or $\vartheta_n \in (Y_n, \theta)$ and this is why we simply say that ϑ_n is between θ and Y_n. Any such random ϑ_n satisfies

$$|\vartheta_n - \theta| \leqslant |Y_n - \theta| \,.$$

From the last inequality it follows that

$$\mathbb{P}(|\vartheta_n - \theta| > \varepsilon) \leqslant \mathbb{P}(|Y_n - \theta| > \varepsilon) \to 0$$

and hence $\vartheta_n \xrightarrow{\mathbb{P}} \theta$ and by the continuity of $g'(\cdot)$ in the neighbourhood of θ

$$g'(\vartheta_n) \xrightarrow{\mathbb{P}} g'(\theta) \,.$$

Then, substituting the expression for $g'(\vartheta)$ in (5.1), we can write

$$\begin{aligned}\frac{\sqrt{n}(g(Y_n) - g(\theta))}{\sigma g'(\theta)} &= \frac{\sqrt{n}(Y_n - \theta)}{\sigma} \times \frac{g'(\vartheta_n)}{g'(\theta)} \\ &= \underbrace{\frac{\sqrt{n}(Y_n - \theta)}{\sigma}}_{\xrightarrow{\mathrm{d}} Z} \times \frac{\overbrace{g'(\vartheta_n)}^{\xrightarrow{\mathbb{P}} g'(\theta)}}{g'(\theta)} \xrightarrow{\mathrm{d}} Z \times 1 \sim \mathsf{N}(0,1) \text{ by Slutsky's Theorem}\end{aligned}$$

□

Yet another way to write the Delta Method result, which is more informal but useful for practical purposes, is as follows:

$$\text{If} \quad Y_n \overset{\text{appr.}}{\sim} \mathsf{N}\left(\theta, \frac{\sigma^2}{n}\right) \quad \text{then} \quad g(Y_n) \overset{\text{appr.}}{\sim} \mathsf{N}\left(g(\theta), [g'(\theta)]^2 \frac{\sigma^2}{n}\right)$$

These two expressions are only valid for finite n, but n must be *large enough* for results when $n \to \infty$ to offer a reasonable approximation to the distribution of $g(Y_n)$. Hence we refer to the above as a *large sample approximation* to the distribution of $g(Y_n)$.

Example

Let $X_1, X_2, \ldots$ be a sequence of independently and identically distributed random variables with mean 2 and variance 7.

Obtain a large sample approximation for the distribution of $(\bar{X}_n)^3$.

Answer: The Central Limit Theorem gives

$$\sqrt{n}(\bar{X}_n - 2) \xrightarrow{\mathrm{d}} \mathsf{N}(0, 7).$$

Application of the Delta Method with $g(x) = x^3$ leads to $g'(x) = 3x^2$ and then

$$\sqrt{n}\{(\bar{X}_n)^3 - 2^3\} \xrightarrow{\mathrm{d}} \mathsf{N}(0, 7 \times (3 \times 2^2)^2).$$

Simplification gives

$$\sqrt{n}\{(\bar{X}_n)^3 - 8\} \xrightarrow{\mathrm{d}} \mathsf{N}(0, 1008).$$

For large n the approximate distribution of $(\bar{X})^3$ is

$\mathsf{N}(8, \frac{1008}{n})$.

Chapter 6

Distributions arising from a Normal Sample

In applications, many data sets consist of *continuous* measurements. Examples include

- Heights in centimetres of a cohort of 137 men.
- Returns of 42 stocks on 18th April, 2011.
- Degree to which the Catecholo-Methyltran gene is differentially expressed between cancerous and normal tissue, based on 16 microarray experiments.

It is common to model data such as these as a *random sample* from the *normal distribution.* Validity of the assumption is often questionable, but because of the central limit theorem, it may be true approximately.

Samples From the Normal Distribution

Definition
Let $X_1, \ldots, X_n$ be a random sample with common distribution $\mathsf{N}(\mu, \sigma^2)$ for some μ and σ^2. Then $X_1, \ldots, X_n$ is a **normal random sample**.

The following results follow immediately:

Result
If $X_1, \ldots, X_n$ is a random sample from the $\mathsf{N}(\mu, \sigma^2)$ distribution then

1. $\sum_i X_i \sim \mathsf{N}(n\mu, n\sigma^2)$,
2. $\bar{X} \sim \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

Note that the Central Limit Theorem from chapter 6 states that a sum or mean of a random sample from any distribution is approximately normal. The above result states that this result is *exact* if we have a normal random sample.

The Chi-Squared Distribution

Definition

If X has density

$$f_X(x) = \frac{e^{-x/2}x^{\nu/2-1}}{2^{\nu/2}\Gamma(\nu/2)}\ , \ x > 0$$

then X has the χ^2 (**chi-squared**) distribution with **degrees of freedom** ν. A common shorthand is

$$X \sim \chi^2_\nu.$$

Note: we pronounce “chi” as “kai”, rhymes with “buy” and “tie”.

Note that the chi-squared distribution is a special case of the Gamma distribution defined in Section 2:

Result

If $X \sim \chi^2_\nu$ then $X \sim \text{Gamma}(\nu/2, 2)$.

This means that results given there can be used to obtain similar results for chi-squared random variables.

Results

If $X \sim \chi^2_\nu$ then

1. $\mathbb{E}(X) = \nu$,
2. $\text{Var}(X) = 2\nu$,
3. $m_X(u) = \left(\frac{1}{1-2u}\right)^{\nu/2}, \quad u < 1/2$.

As the title of the chapter suggests, the χ^2 distribution arises from a normal random sample. The key relation is this result.

Sum of squared iid normals

If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$, then

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

The *t* Distribution

Another distribution that arises naturally when we consider a normal sample is the so called student distribution. This distribution was invented by William Sealy Gosset (June 13, 1876 – October 16, 1937), who was an English statistician. He came up with the distribution while working out how to make better beer at a Guinness brewery. He published the discovery under the pen name Student, because the brewery did not allow its employees to publish anything for fear they will reveal any brewing secrets to their competitors.

Definition

If $T \sim t_\nu$ then

$$f_T(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{(\nu+1)}{2}}, \quad -\infty < t < \infty.$$

A common shorthand is

$$T \sim t_\nu.$$

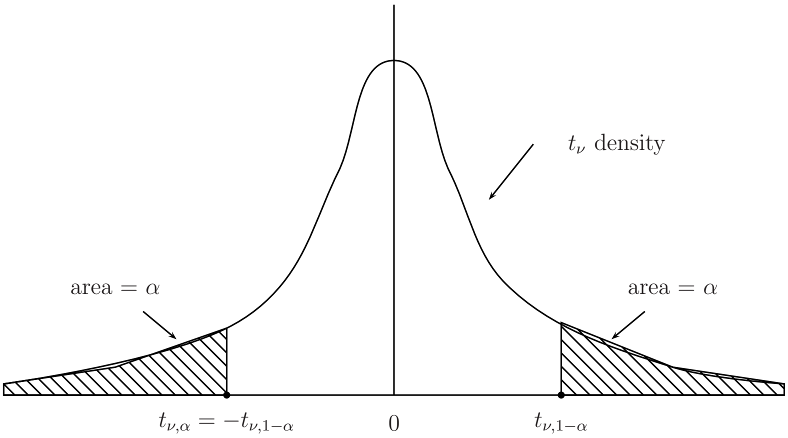
One relation to the normal distribution is the following.

Result

If $T \sim t_\nu$ then as $\nu \to \infty$, T converges to a $\mathsf{N}(0,1)$ random variable.

Proof: $f_T(t) \propto \left(1+\frac{t^2}{\nu}\right)^{-\frac{(\nu+1)}{2}}$ and $\lim_{\nu\to\infty}\left(1+\frac{t^2}{\nu}\right)^{-\nu} = \mathrm{e}^{-t^2}$ so

$$\lim_{\nu\to\infty} f_T(t) \propto \mathrm{e}^{-\frac{t^2}{2}}.$$



If $T \sim t_\nu$, then $\mathbb{P}(T \le t_{\nu,\alpha}) = \alpha$. $t_{\nu,\alpha}$ is the αth quantile of the t_ν distribution.
$t_{\nu,1-\alpha}$ is the $(1-\alpha)$th quantile of the t_ν distribution.
Tables for the t-distribution are in the back of the lecture notes. Again, the tabulated

values are for **right-tailed** probabilities, although two-tailed probabilities can also be read off this table.

The following result is frequently used in applied statistics and is the key reason for the importance of the student's distribution. We will use this result in later chapters.

Result
Let $X_1, \ldots, X_n$ be a random sample from the $\mathsf{N}(\mu, \sigma^2)$ distribution. Then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

The *F* of Fisher distribution

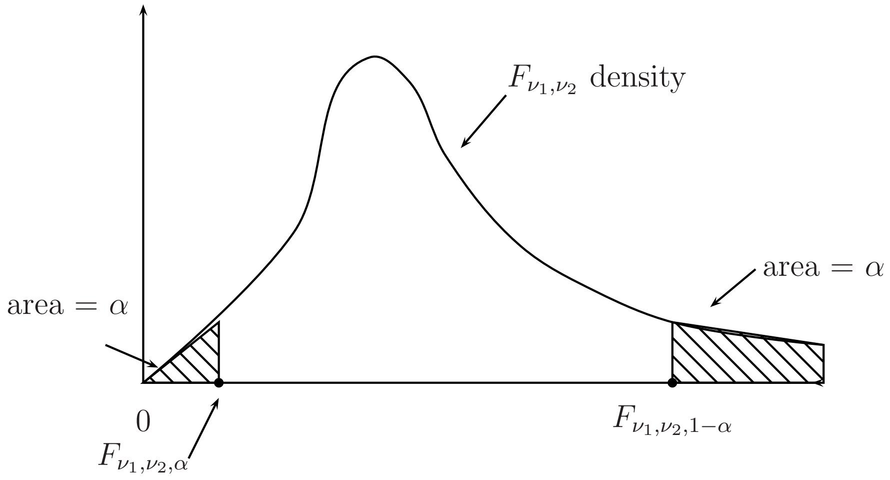
Suppose $X_1, X_2, \ldots, X_{n_X}$ are independent $\mathsf{N}(\mu_X, \sigma_X^2)$ and $Y_1, Y_2, \ldots, Y_{n_Y}$ are independent $\mathsf{N}(\mu_Y, \sigma_Y^2)$ and the samples are independent. When comparing the variances or drawing inferences about σ_X^2/σ_Y^2 we use S_X^2/S_Y^2 (the ratio of the sample variances) and this leads us to the F distribution.

Definition
Suppose $X \sim \chi^2_{\nu_1}$ and $Y \sim \chi^2_{\nu_2}$ and X and Y are independent. Then $F = \frac{X/\nu_1}{Y/\nu_2}$ has the F distribution with degrees of freedom ν_1 and ν_2. We write $F \sim \mathsf{F}_{\nu_1,\nu_2}$.

Result
If $F \sim \mathsf{F}_{\nu_1,\nu_2}$ then F has density function

$$f_F(u) = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} u^{\frac{\nu_1}{2}-1}\left(1 + \frac{\nu_1 u}{\nu_2}\right)^{-\frac{(\nu_1+\nu_2)}{2}}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \qquad , u > 0$$

where $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$.



$F_{\nu_1,\nu_2,\alpha}$ is the αth quantile of the F_{ν_1,ν_2} distribution; i.e., if $U \sim F_{\nu_1,\nu_2}, \mathbb{P}(U \leq F_{\nu_1,\nu_2,\alpha}) = \alpha$. $F_{\nu_1,\nu_2,1-\alpha}$ is the $(1-\alpha)$th quantile of the F_{ν_1,ν_2} distribution; i.e., if $U \sim F_{\nu_1,\nu_2}$, $\mathbb{P}(U \leq F_{\nu_1,\nu_2,1-\alpha}) = 1-\alpha$.

Again, the Fisher distribution arises naturally when we consider statistics of normal samples. The key is this result.

Ratio of sums of squared normal samples

For independent samples

$$X_1, X_2, \ldots, X_{n_X} \overset{\text{iid}}{\sim} \mathsf{N}(\mu_X, \sigma_X^2)$$

$$Y_1, Y_2, \ldots, Y_{n_Y} \overset{\text{iid}}{\sim} \mathsf{N}(\mu_Y, \sigma_Y^2)$$

with sample variances

$$S_X^2 = \frac{1}{n_X - 1}\sum_{i=1}^{n_X}(X_i - \bar{X})^2 = \text{sample variance of the } X\text{'s},$$

$$S_Y^2 = \frac{1}{n_Y - 1}\sum_{i=1}^{n_Y}(Y_i - \bar{Y})^2 = \text{sample variance of the } Y\text{'s}.$$

we have

$$\frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} \sim \underset{1}{\mathsf{F}_{n_X-1,n_Y-1}}.$$

The F distribution was invented/discovered by Sir Ronald Fisher, who made fundamental contributions to both statistics and genetics.

χ^2, t, F and R

The statistics package R can be used to calculate cumulative probabilities and quantiles from the distributions considered in this chapter, and more. The method of use is the same as that described in chapter 2.

Distribution	Random sample	$f_X(x)$	$F_X(x)$	x_p
Normal	rnorm	dnorm	pnorm	qnorm
χ^2	rchisq	dchisq	pchisq	qchisq
t	rt	dt	pt	qt
F	rf	df	pf	qf

To find $\chi^2_{10,0.95}$:

```
> qchisq(0.95,10)
```

To find $\mathbb{P}(t_{15} < 2.602)$:

```
> pt(2.602,15)
```

To find $F_{5,10,0.05}$:

```
> qf(0.05,5,10)
```

To take a random sample of size 20 from the t_1 distribution:

```
> rt(20,1)
```

To plot $f_X(x)$ where $X \sim \chi^2_{10}$ between 0 and 40:

```
> x=0:40
> f=dchisq(x,20)
> plot(x,f,type="l",ylab="f_X(x)", main="The chi-squared (20) distribution")
```

Some Key Results from Chapter 6

Primary Results

Let $X_1, \ldots, X_n$ be a random sample from the $\mathsf{N}(\mu, \sigma^2)$ distribution. Then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

and

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

For independent samples

$$X_1, X_2, \ldots, X_{n_X} \text{ random sample from } \mathsf{N}(\mu_X, \sigma^2_X)$$

$$Y_1, Y_2, \ldots, Y_{n_Y} \text{ random sample from } \mathsf{N}(\mu_Y, \sigma^2_Y)$$

we have

$$\frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} \sim F_{n_X-1,n_Y-1}.$$

Secondary Results

$$Z_1,\ldots,Z_m \sim \mathsf{N}(0,1) \text{ and ind.} \quad \Longrightarrow \quad \textstyle\sum_{i=1}^m Z_i^2 \sim \chi_m^2$$

$$Z \sim \mathsf{N}(0,1), \quad Q \sim \chi_\nu^2, \quad Z, Q \text{ ind.} \quad \Longrightarrow \quad \tfrac{Z}{\sqrt{Q/\nu}} \sim t_\nu$$

$$Q_1 \sim \chi_{\nu_1}^2, \quad Q_2 \sim \chi_{\nu_2}^2, \quad Q_1, Q_2 \text{ ind.} \quad \Longrightarrow \quad \tfrac{Q_1/\nu_1}{Q_2/\nu_2} \sim F_{\nu_1,\nu_2}$$

Part Two —

Basic Statistical Inference

Chapter 7

An Introduction to Statistical Inference

So far, we have dealt only with the situation where we know the distribution of the variable of interest to us, and we know the values of parameters in the distribution. But in practice, we usually do not know the value of the parameters. We would like to use sample data to make **inferences** about the parameters, which is the subject of the next few chapters.

For further reading, consider Hogg *et al* (2005) sections 4.1, 5.1 and 5.4 or Rice (2007) sections 7.1-7.3 (but ignore the finite population correction stuff).

Example

Consider the situation where we would like to know the cadmium concentration in water from a particular dam. We know from Chapter 5 that if we take several water samples and measure cadmium concentration, the average of these samples will be normally distributed and will be centred on μ, the true average cadmium concentration of water in the dam. But what is μ?

In this chapter we will learn how to make inferences about μ based on a sample.

Statistical inference is a very powerful tool – it allows us to make very specific statements about a variable, or about a population, based on just a sample of measurements!

Statistical Models

Given an observed random sample $X_1, \ldots, X_n$ it is common to postulate a **statistical model** for the data. This is a set of density or probability functions f_X that are consistent with the data, and facilitates answering certain questions of interest.

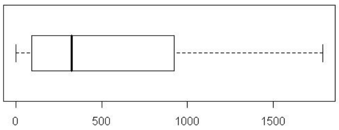
A **parametric model** is a set of f_X's that can be parametrised by a finite number of parameters. We will only deal with parametric models in this course, although the more general non-parametric case is also worth studying – this is considered in MATH3811.

Example

Keating, Glaser and Ketchum (*Technometrics*, 1990) describe data on the lifetimes (hours) of 20 pressure vessels constructed of fibre/epoxy composite materials wrapped around metal liners. The data are:

```
274 28.5 1.7 20.8 871 363 1311 1661 236 828
458 290 54.9 175 1787 970 0.75 1278 776 126
```

The following figure shows a graphical representation of the data:



Given the positive and right-skewed nature of the data a plausible parametric model for the data is:

$$\left\{ f_X(x;\beta) = \tfrac{1}{\beta} \mathrm{e}^{-x/\beta}, \quad x > 0; \quad \beta > 0 \right\}.$$

Since this family of density functions is parameterised by the single parameter $\beta > 0$, this is a parametric model.

We could check how well this parametric model fits the data using a quantile-quantile plot.

A general parametric model with a single parameter θ is

$$\{f_X(x;\theta) : \theta \in \Theta\}.$$

The set $\Theta \subseteq \mathbb{R}$ is the set of possible values of θ and is known as the **parameter space**. If this model is assumed for a random sample $X_1, \ldots, X_n$ then we write

$$X_1, \ldots, X_n \sim f_X(x; \theta), \quad \theta \in \Theta.$$

Note that a model for a random sample induces a probability measure on its members. However, these probabilities depend on the value of θ. This is sometimes indicated using subscripted notation such as: P_θ, E_θ, Var_θ to describe probabilities and expectations according to the model and particular values of θ, although we will minimise the use of this notation.

Estimation

Let $X_1, \ldots, X_n$ be a random sample with model

$$\{f_X(x; \theta) : \theta \in \Theta\}.$$

A fundamental problem in statistics is that of determining a single θ that is "most consistent" with the sample. This is known as the **estimation** problem, sometimes referred to as **point estimation**.

Definition

Suppose

$$X_1, \ldots, X_n \sim f_X(x; \theta), \quad \theta \in \Theta.$$

An **estimator** for θ, denoted by $\widehat{\theta}$, is any real-valued function of $X_1, \ldots, X_n$; i.e.

$$\widehat{\theta} = g(X_1, \ldots, X_n)$$

where the function $g : \mathbb{R}^n \to \mathbb{R}$.

Example

Let

$$p = \text{proportion of UNSW students who watched the Cricket World Cup final.}$$

be a parameter of interest.

Suppose that we survey 8 UNSW students, asking them whether they watched the Cricket World Cup final.

Let $X_1, \ldots, X_8$ be such that

$$X_i = \begin{cases} 1, & \text{if } i\text{th surveyed watched the Cricket World Cup final} \\ 0, & \text{otherwise.} \end{cases}$$

An appropriate model is

$$X_1, \ldots, X_8 \sim f_X(x; p), \; 0 < p < 1.$$

where

$$f_X(x; p) = p^x (1-p)^{1-x} = \begin{cases} p, & x = 1 \\ 1-p, & x = 0. \end{cases}$$

Then the 'natural' estimator for p is

$$\widehat{p} = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8}{8},$$

corresponding to the proportion from the survey that watched the Cricket World Cup final. However, there are many other possible estimators for p; such as

$$\widehat{p}_{\text{alt}} = \frac{X_2 + X_4 + X_6 + X_8}{4},$$

based on every second person in the survey. But even

$$\widehat{p}_{\text{weird}} = \sin(X_1 e^{X_5}) + \pi \coth(X_3/(7X_8 + 12)).$$

satisfies the definition of being an estimator for p! (It's not a very good estimator though...)

The previous definition permits **any** function of the sample to be an estimator for a parameter θ. However, only certain functions have good properties. So how do we identify an estimator of θ that has good properties? This is the subject of the remainder of this section.

Before we start studying properties of estimators, we state a set of facts that are fundamental to statistical inference:

> The estimator $\widehat{\theta}$ is a function of the random variables $X_1, \ldots, X_n$ and is therefore a random variable itself.
>
> It has its own probability function or density function
>
> $$f_{\widehat{\theta}}$$
>
> that depends on θ.
> We use $f_{\widehat{\theta}}$ to study the properties of $\widehat{\theta}$ as an estimator of θ.

This idea will be used repeatedly throughout the rest of these notes.

Bias

The first property of an estimator that we will study is bias. This corresponds to the difference between $\mathbb{E}(\widehat{\theta})$, the "centre of gravity" of $f_{\widehat{\theta}}$, and the target parameter θ:

Definition

Let $\widehat{\theta}$ be an estimator of a parameter θ. The **bias** of $\widehat{\theta}$ is given by

$$\text{bias}(\widehat{\theta}) = \mathbb{E}(\widehat{\theta}) - \theta.$$

If $\text{bias}(\widehat{\theta}) = 0$ then $\widehat{\theta}$ is said to be an **unbiased** estimator of θ.

Bias is a measure of the **systematic error** of an estimator – how far we expect the estimator to be from its true value θ, on average.

Often, we want to use an estimator $\widehat{\theta}$ which is unbiased, or as close to zero bias as possible.

In many practical situations, we can identify an estimator of θ that is unbiased. If we cannot, then we would like an estimator that has as small a bias as possible.

Example

For the Cricket World Cup example

$$\widehat{p} = \frac{Y}{8}$$

where Y is the number of students who watched the Cricket World Cup final, and

$$Y \sim \text{Bin}(8, p)$$

Find $f_{\widehat{p}}(x)$, and $\text{bias}(\widehat{p})$. Compare these to the corresponding results for $\widehat{p}_{\text{alt}} = (X_2 + X_4 + X_6 + X_8)/4$.

Note that Y has probability function

$$f_Y(y) = \binom{8}{y} p^y (1-p)^{8-y}, \quad y = 0, 1, 2, \ldots, 8.$$

Because $\widehat{p} = Y/8$, $Y = 8\widehat{p}$ and using rules for transformation of a discrete random variable:

$$f_{\widehat{p}}(x) = \binom{8}{8x} p^{8x} (1-p)^{8-8x}, \quad x = 0, \tfrac{1}{8}, \tfrac{2}{8} \ldots, 1.$$

Now to find $\text{bias}(\widehat{p}) = \mathbb{E}(\widehat{p} - p)$, we will first find $\mathbb{E}(\widehat{p})$.

Since $\mathbb{E}(Y) = 8p$, $\mathbb{E}(\widehat{p}) = \mathbb{E}(Y/8) = \mathbb{E}(Y)/8 = 8p/8 = p$ so

$$\text{bias}(\widehat{p}) = \mathbb{E}(\widehat{p} - p) = \mathbb{E}(\widehat{p}) - \mathbb{E}(p) = 0$$

This means that $\widehat{p}$ is an unbiased estimator for p.

By similar arguments,

$$f_{\widehat{p}_{\text{alt}}}(x) = \binom{4}{4x} p^{4x}(1-p)^{4-4x}, \quad x = 0, \tfrac{1}{4}, \tfrac{2}{4} \ldots, 1.$$

and $\text{bias}(\widehat{p}_{\text{alt}}) = 0$, so $\widehat{p}_{\text{alt}}$ is also unbiased.

Standard error

The next fundamental property of an estimator is its **standard error:**

Definition
Let $\widehat{\theta}$ be an estimator of a parameter θ. The **standard error** of $\widehat{\theta}$ is the *standard deviation*:

$$\text{se}(\widehat{\theta}) = \sqrt{\text{Var}(\widehat{\theta})},$$

To obtain the *estimated standard error* $\widehat{\text{se}}(\widehat{\theta})$, we first derive $\text{Var}(\widehat{\theta})$, the variance of $\widehat{\theta}$, and then we replace unknown parameters θ by its estimator $\widehat{\theta}$.

Like the bias, the standard error of an estimator is ideally as small as possible. However, unlike the bias the standard error can never be made zero (except in trivial cases).

Example

Consider, again, the Cricket World Cup example.

Find $\text{se}(\widehat{p})$ and $\text{se}(\widehat{p}_{\text{alt}})$. Comment.

From properties of binomial random variables

$$\mathrm{Var}(\widehat{p}) = \mathrm{Var}(Y/8) = (1/8^2)\mathrm{Var}(Y) = \frac{p(1-p)}{8}.$$

Therefore the standard error of $\widehat{p}$ is

$$\mathrm{se}(\widehat{p}) = \sqrt{\frac{p(1-p)}{8}}$$

and the estimated standard error of $\widehat{p}$ is

$$\widehat{\mathrm{se}}(\widehat{p}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{8}}.$$

Similarly,

$$\mathrm{se}(\widehat{p}_{\mathrm{alt}}) = \sqrt{\frac{p(1-p)}{4}} \quad \text{and} \quad \widehat{\mathrm{se}}(\widehat{p}_{\mathrm{alt}}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{4}}.$$

Therefore the standard error of $\widehat{p}_{\mathrm{alt}}$ will always be larger than that of $\widehat{p}$ by a factor of $\sqrt{2} \simeq 1.4$, which suggests that $\widehat{p}$ is a better estimator of p than $\widehat{p}_{\mathrm{alt}}$.

Mean squared error

The bias and standard error of an estimator are fundamental measures of different aspects of the quality of $\widehat{\theta}$, as an estimator of θ: bias is concerned with the systematic error in $\widehat{\theta}$, while the standard error is concerned with its inherent random (or sampling) error. But consider a situation in which we want to choose between two alternative estimators of θ, where one has smaller bias, and the other has smaller standard error. How could we choose between these two estimators?

One approach is to use a combined measure of the quality of $\widehat{\theta}$, which combines the bias and standard error in some way. The **mean squared error** is the most common way of doing this:

Definition

The **mean squared error** of $\widehat{\theta}$ is given by

$$\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}\{(\widehat{\theta} - \theta)^2\}.$$

The following result shows how $\text{MSE}_\theta(\widehat{\theta})$ takes care of both the bias and standard error in $\widehat{\theta}$:

Result

Let $\widehat{\theta}$ be an estimator of a parameter θ. Then

$$\text{MSE}(\widehat{\theta}) = \text{bias}(\widehat{\theta})^2 + \text{Var}(\widehat{\theta}),$$

and the estimated mean squared error is

$$\widehat{\text{MSE}}(\widehat{\theta}) = \text{bias}(\widehat{\theta})^2 + \widehat{\text{se}}(\widehat{\theta})^2.$$

Proof:

$$\begin{aligned}
\text{MSE}(\widehat{\theta}) &= \mathbb{E}\{(\widehat{\theta}-\theta)^2\} \\
&= \mathbb{E}[\{\widehat{\theta}-\mathbb{E}(\widehat{\theta})+\mathbb{E}(\widehat{\theta})-\theta\}^2] \\
&= \mathbb{E}[\{\widehat{\theta}-\mathbb{E}(\widehat{\theta})\}^2]+\mathbb{E}[\{\mathbb{E}(\widehat{\theta})-\theta\}^2]+2E[\{\widehat{\theta}-\mathbb{E}(\widehat{\theta})\}\{\mathbb{E}(\widehat{\theta})-\theta\}] \\
&= \text{Var}(\widehat{\theta})+\mathbb{E}[\{\text{bias}(\widehat{\theta})\}^2]+2\{\mathbb{E}(\widehat{\theta})-\theta\}\mathbb{E}[\{\widehat{\theta}-\mathbb{E}(\widehat{\theta})\}] \\
&= \text{Var}(\widehat{\theta})+\text{bias}(\widehat{\theta})^2+2\{\mathbb{E}(\widehat{\theta})-\theta\}\{\mathbb{E}(\widehat{\theta})-\mathbb{E}(\widehat{\theta})\} \\
&= \text{Var}(\widehat{\theta})+\text{bias}(\widehat{\theta})^2.
\end{aligned}$$

Definition

Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be two estimators of a parameter θ. Then $\widehat{\theta}_1$ is **better than** $\widehat{\theta}_2$ (with respect to MSE) at $\theta_0 \in \Theta$ if

$$\text{MSE}_{\theta_0}(\widehat{\theta}_1) < \text{MSE}_{\theta_0}(\widehat{\theta}_2).$$

If $\widehat{\theta}_1$ is better than $\widehat{\theta}_2$ for all $\theta \in \Theta$ then we say

$$\widehat{\theta}_1 \textbf{ is uniformly better than } \widehat{\theta}_2.$$

Example

For the Cricket World Cup example, find $\text{MSE}(\widehat{p})$ and $\text{MSE}(\widehat{p}_{\text{alt}})$. Is $\widehat{p}$ uniformly better than $\widehat{p}_{\text{alt}}$?

From previous results,

$$\text{MSE}(\widehat{p}) = 0^2 + \frac{p(1-p)}{8} = \frac{p(1-p)}{8}$$

while

$$\text{MSE}(\widehat{p}_{\text{alt}}) = \frac{p(1-p)}{4}.$$

Since

$$\mathrm{MSE}(\widehat{p}) < \mathrm{MSE}(\widehat{p}_{\mathrm{alt}}) \quad \text{for all } 0 < p < 1$$

$\widehat{p}$ is uniformly better than $\widehat{p}_{\mathrm{alt}}$.

(This result makes intuitive sense, since $\widehat{p}$ is based on twice as many responses.)

Common estimated standard error formulas

The standard error expressions for sample means and sample proportions are as follows.

$$\widehat{\mathrm{se}}(\bar{X}) = \frac{\widehat{\sigma}}{\sqrt{n}}$$

where $\widehat{\sigma}$ is the sample (estimated) standard deviation; and

$$\widehat{\mathrm{se}}(\widehat{p}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

Consistency

The **consistency** property of an estimator is concerned with its performance as the amount of data increases. It seems reasonable to demand that $\widehat{\theta} = \widehat{\theta}_n$ gets better as the sample size n grows. Consistency corresponds to $\widehat{\theta}_n$ converging to θ as n becomes larger.

Definition

The estimator $\widehat{\theta}_n$ is **consistent** for θ if

$$\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta.$$

Example

For the Cricket World Cup example, suppose that n people are surveyed and

$$\widehat{p}_n = \frac{Y_n}{n}$$

where

$$Y_n = \text{number of students that watched Cricket World Cup final.}$$

Prove that $\widehat{p}_n$ is consistent.

$Y_n \sim \mathrm{Bin}(n, p)$ and

$$\mathbb{E}(\widehat{p}_n) = p.$$

Application of the Weak Law of Large numbers (to the binary X_i random variables) leads to

$$\widehat{p}_n \xrightarrow{\mathbb{P}} p.$$

Hence $\widehat{p}_n$ is consistent.

In some instances consistency is difficult to check via convergence in probability arguments. The following result reduces the problem to first and second moments of the estimator:

Result

If

$$\lim_{n\to\infty} \mathrm{MSE}(\widehat{\theta}_n) = 0$$

then

$$\widehat{\theta}_n \text{ is consistent for } \theta.$$

Example

For the Cricket World Cup example, find $\mathrm{MSE}(\widehat{p}_n)$. Hence show that $\widehat{p}_n$ is consistent for p.

It is easily shown that

$$\mathrm{MSE}(\widehat{p}_n) = \frac{p(1-p)}{n}$$

so

$$\lim_{n\to\infty} \mathrm{MSE}(\widehat{p}_n) = p(1-p) \lim_{n\to\infty} \frac{1}{n} = 0$$

and therefore $\widehat{p}_n$ is consistent for p.

Asymptotic normality

A final property of an estimator $\widehat{\theta}$ that is of interest is **asymptotic normality**. If this holds then $f_{\widehat{\theta}}$ has an approximate normal distribution for large sample sizes, which facilitates inference about θ.

Definition

The estimator $\widehat{\theta}$ is **asymptotically normal** if

$$\frac{\widehat{\theta}-\theta}{\mathrm{se}(\widehat{\theta})} \xrightarrow{\mathrm{d}} \mathsf{N}(0,1).$$

In particular, we know already from the CLT that a sample mean $\widehat{\mu} = \bar{X}$ and a sample proportion $\widehat{p}$ are asymptotically normal.

Example

In the Cricket World Cup example, let $X_1, \ldots, X_n$ be defined by

$$X_i = \begin{cases} 1, & \text{if } i\text{th surveyed student watched Cricket World Cup final} \\ 0, & \text{otherwise.} \end{cases}$$

Assume the X_i are independent. Write $\widehat{p}_n$ in terms of the X_i and hence use convergence results from Chapter 5 to find the asymptotic distribution of $\widehat{p}_n$.

Example

Consider the sample mean $\bar{X}_n$ of n independent random variables with mean μ and variance σ^2.

Show that $\bar{X}_n$ is consistent. Is $\bar{X}_n$ asymptotically normal?

Observed values

Throughout this section we have considered the random sample $X_1, \ldots, X_n$ and the properties of the estimator $\widehat{\theta}$, by treating it as a random variable. This allows us to do theoretical calculations concerning, say, the bias and large sample properties of $\widehat{\theta}$.

In practice, we only take one sample, and observe a single value of $\widehat{\theta}$, known as the **observed value** of $\widehat{\theta}$. This is also sometimes called the **estimate** of θ (as opposed to the **estimator**, which is the random variable we use to obtain the estimate).

Example

Consider the pressure vessel example and let

$$X_i = i\text{th lifetime before the data are observed.}$$

An unbiased estimator for β is

$$\widehat{\beta} = \bar{X} = \frac{1}{n} \sum_i X_i.$$

After the data are observed to be:

```
274 28.5 1.7 20.8 871 363 1311 1661 236 828
458 290 54.9 175 1787 970 0.75 1278 776 126
```

the *observed value* of $\widehat{\beta}$ becomes

$$(274 + 28.5 + \ldots + 126)/20 = 575.53.$$

Notation for observed values

Some statistics texts distinguish between random variables and their observed values via use of lower-case letters. For the previous example this would involve:

$$x_1 = 274, \ x_2 = 28.5, \ldots, x_{20} = 126.$$

and

$$\bar{x} = 575.53.$$

Good notation for the observed value of $\widehat{\beta}$ is a bit trickier since β is already lower-case (in Greek). These notes will not worry too much about making such distinctions. So $\widehat{\beta}$ denotes the random variable $\bar{X}$ before the data are observed. But we will also say $\widehat{\beta} = 575.53$ for the observed value. The meaning of $\widehat{\beta}$ should be clear from the context.

Estimate (standard error) notation

In applied statistics, a common notation when reporting the observed value of an estimator (or estimate) is to add the estimated standard error in parentheses:

estimate (standard error)

Example

Eight students are surveyed and two watched the Cricket World Cup final.

Find $\widehat{p}$ (defined previously) and its standard error, and write you answer in estimate (standard error) notation.

$$2/8 = 0.25.$$

The estimated standard error is

$$\sqrt{0.25(1-0.25)/8} = 0.153.$$

So the estimate (standard error) notation would report the results for p as follows:

$$0.25\ (0.153).$$

Multiparameter models

So far in this chapter we have only considered models with a single parameter. However, it is often the case that a model with more than one parameter is required for the data at hand.

Example

For the pressure vessel data we have previously considered the single parameter model

$$\left\{ f_X(x;\beta) = \tfrac{1}{\beta} e^{-x/\beta}, \quad x > 0; \quad \beta > 0 \right\}.$$

A *two-parameter* model is

$$\left\{ f_X(x;\alpha,\beta) = \tfrac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad x > 0; \quad \alpha,\beta > 0 \right\}.$$

which corresponds to the X_i's having a Gamma(α,β) distribution.

Chapter 6 dealt with the most common two-parameter model

$$X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2).$$

Often inference is concerned with the original parameters: μ and σ for a $\mathsf{N}(\mu,\sigma^2)$ model; α and β for a Gamma(α,β) model. However, sometimes a *transformation* of the parameters is of interest. For example, in the Gamma(α,β) model a parameter of interest is often

$$\tau = \alpha\beta$$

since this corresponds to the mean of the distribution.

Confidence Intervals

An estimator $\widehat{\theta}$ of a parameter θ leads to a single number for inferring the true value of θ. For example, in the Cricket World Cup example if we survey 50 people and 16 watched the Cricket World Cup final then the estimator $\widehat{p}$ has an observed value of 0.32.

However, the number 0.32 alone does not tell us much about the inherent variability in the underlying estimator. Confidence intervals aim to improve this situation with a **range** of plausible values, *e.g.*

p is likely to be in the range 0.19 to 0.45.

Definition

Let $X_1, \ldots, X_n$ be a random sample from a model that includes an unknown parameter θ. Let

$$L = L(X_1, \ldots, X_n) \quad \text{and} \quad U = U(X_1, \ldots, X_n)$$

be statistics (*i.e.* functions of the X_i's) for which

$$\mathbb{P}(L < \theta < U) \geq 1 - \alpha, \quad \text{for all } \theta \in \Theta.$$

Then (L, U) is a $\boldsymbol{1-\alpha}$**, or** $\mathbf{100(1-\boldsymbol{\alpha})\%}$**, confidence interval** for θ.

It is important to note that in the probability statement

$$\mathbb{P}(L < \theta < U) \geq 1 - \alpha$$

the quantity in the middle (θ) is fixed, but the limits (L and U) are random. This is the reverse situation from many probability statements that arise in earlier chapters, such as

$$\mathbb{P}(2 \leq X \leq 7)$$

for a random variable X.

Example

Consider the pressure vessel example

$$X_1, \ldots, X_{20} \sim f_X(x; \beta)$$

where

$$f_X(x; \beta) = \tfrac{1}{\beta}\, e^{-x/\beta}, \quad x > 0; \quad \beta > 0.$$

Then it can be shown (details omitted) that

$$\mathbb{P}\left(0.52\,\bar{X} \leq \beta \leq 1.67\,\bar{X}\right) = 0.99 \quad \text{for all } \beta > 0.$$

Therefore

$$(0.52\,\bar{X}, 1.67\,\bar{X})$$

is a 0.99 or 99% confidence interval for β.

Since the observed value of $\bar{X}$ is $\bar{x} = 575.53$, the observed value of the 99% confidence interval for β is

$$(352, 852).$$

Confidence Intervals for a Normal Random Sample

We now return to the topic of Chapter 6 corresponding to the special case where the random sample can be reasonably modelled as coming from a normal distribution:

$$X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2).$$

Distribution theory results derived there:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \quad \text{and} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

allow for exact confidence interval statements for μ and σ.

Result

Let $X_1, \ldots, X_n$ be a random sample from the $\mathsf{N}(\mu, \sigma^2)$ distribution. Then a $100(1-\alpha)\%$ confidence interval for μ is

$$\left(\bar{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}},\ \bar{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right).$$

Proof: By the definition of $t_{n-1,q}$,

$$\begin{aligned}
1-\alpha &= \mathbb{P}\left(-t_{n-1,1-\alpha/2} < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{n-1,1-\alpha/2}\right) \\
&= \mathbb{P}\left(-t_{n-1,1-\alpha/2}S/\sqrt{n} < \overline{X} - \mu < t_{n-1,1-\alpha/2}S/\sqrt{n}\right) \\
&= \mathbb{P}\left(t_{n-1,1-\alpha/2}S/\sqrt{n} > \mu - \overline{X} > -t_{n-1,1-\alpha/2}S/\sqrt{n}\right) \\
&= \mathbb{P}\left(\overline{X} - \frac{S}{\sqrt{n}}t_{n-1,1-\alpha/2} < \mu < \overline{X} + \frac{S}{\sqrt{n}}t_{n-1,1-\alpha/2}\right)
\end{aligned}$$

The above is a particular powerful result. The above provides us with a method of making potentially quite specific statements about μ, based on a sample: we can estimate a range of values which we are arbitrarily sure will contain μ (such intervals contain μ $100(1-\alpha)\%$ of the time).

This is a particularly useful result in practice, because many interesting research questions can be phrased in terms of means (*e.g.* What is average recovery time for patients given a new treatment? How much weight do people lose, on average, if they exercise for 30 minutes every day?)

Recall from Chapter 5 that it was noted that even when X_i are not normal, t_{n-1} is a reasonable approximation for the distribution of $\frac{\bar{X}-\mu}{S/\sqrt{n}}$, as long as n is large enough for the Central Limit Theorem to come into play. This means that we can use the above method to construct a confidence interval for μ even when data are not normal – this is where the above result becomes really handy in practice!

Confidence Intervals for Two Normal Random Samples

A more common situation in applied statistics is one of *comparison* between two samples. For example, “Is recovery time shorter for patients using a new treatment than for patients on the old treatment?” “Do people lose more weight, on average, if they go on the CSIRO diet than the Atkin’s diet?”)

Suppose the samples are normal:

$$X_1, \dots, X_{n_X} \sim \mathsf{N}(\mu_X, \sigma_X^2)$$

and

$$Y_1, \dots, Y_{n_Y} \sim \mathsf{N}(\mu_Y, \sigma_Y^2)$$

As previously, this assumption is not critical when making inferences about μ, because some robustness to non-normality is inherited from the Central Limit Theorem.

Result

Let

$$X_1, \dots, X_{n_X} \sim \mathsf{N}(\mu_X, \sigma^2)$$

and

$$Y_1, \dots, Y_{n_Y} \sim \mathsf{N}(\mu_Y, \sigma^2)$$

be two independent normal random samples; each with the same variance σ^2. Then a $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is

$$\bar{X} - \bar{Y} \pm t_{n_X+n_Y-2,1-\alpha/2} S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}$$

where

$$S_p^2 = \frac{(n_X-1)S_X^2 + (n_Y-1)S_Y^2}{n_X+n_Y-2},$$

known as the *pooled sample variance.*

The proof is analogous to that for the single sample confidence interval for μ.

Example

Peak expiratory flow is measured for 7 “normal” six-year-old children and 9 six-year-old children with asthma. The data are as follows:

Normal children	asthmatic children
55	53
57	39
80	56
71	54
62	49
56	53
77	45
	37
	44

We would like to know if peak flow is different between normal and asthmatic children, and if so, how different. Use a confidence interval to answer this question.

Graphical inspection of the data shows that the data do not appear to be too far from the normal distribution, for each sample. Let

$$\begin{aligned} \mu_X &= \text{mean peak expiratory flow for normal children} \\ \mu_Y &= \text{mean peak expiratory flow for asthmatic children.} \end{aligned}$$

We will obtain a 95% confidence interval for $\mu_X - \mu_Y$ under the assumption that the variances for each population are equal. The sample sizes, sample means and sample variances are

$$\begin{array}{lll} n_X = 7, & \bar{x} \simeq 65.43, & s_X^2 \simeq 109.62 \\ n_Y = 9, & \bar{y} \simeq 47.78, & s_Y^2 \simeq 47.19 \end{array}$$

The pooled sample variance is

$$s_p^2 \simeq \frac{6 \times 109.62 + 8 \times 47.19}{14} \simeq 73.95.$$

The appropriate t-distribution quantile is $t_{14,0.975} = 2.145$. The confidence interval is then

$$(65.43 - 47.78) \pm 2.15\sqrt{73.95}\sqrt{\frac{1}{7} + \frac{1}{9}} = (8.4, 27.0).$$

In conclusion, we can be 95% confident that the difference in mean peak expiratory flow between the two groups is between 8.4 and 27.0 units.

Confidence Intervals for a Paired Normal Random Sample

If we have a paired normal random sample from (X, Y), *i.e.* two normal random samples that are *dependent*, we can construct a confidence interval for the mean difference by analysing the differences $D = X - Y$ as a single normal random sample, using the methods on page 188.

Example

The number of people exceeding the speed limit, per month, was recorded before and after the installation of speed cameras at four locations (data from the Sydney Morning Herald, 22nd September 2003).

location	Concord West	Edgecliff	Wentworthville	Green Valley
before	5719	7535	6254	2200
after	1786	2228	528	260
difference				

How large was the reduction in number of people speeding per month per camera, on average?

Confidence Intervals for Sample Proportions

Consider a binomial sample $X \sim \mathsf{Bin}(n, p)$, where we are interested in making inferences about p, the probability of a "success".

Recall that we have shown that if $X \sim \mathsf{Bin}(n, p)$, then

$$\frac{\widehat{p} - p}{\sqrt{p(1-p)/n}} \xrightarrow{\text{d}} \mathsf{N}(0, 1).$$

This means that we can use the normal distribution to construct a confidence interval for p, based on a binomial sample.

Result

Let $X \sim \mathsf{Bin}(n, p)$. Then an approximate $100(1-\alpha)\%$ confidence interval for p is

$$\left(\widehat{p} - z_{1-\alpha/2}\,\widehat{\text{se}}(\widehat{p}),\; \widehat{p} + z_{1-\alpha/2}\,\widehat{\text{se}}(\widehat{p})\right).$$

where $\widehat{p} = \frac{X}{n}$ and $\widehat{\text{se}}(\widehat{p}) = \sqrt{\widehat{p}(1-\widehat{p})/n}$.

Note that this is only an "approximate" confidence interval for two reasons: $\widehat{p}$ is only approximately normal, and we are using $\widehat{\text{se}}(\widehat{p})$ in place of $\sqrt{p(1-p)/n}$.

Chapter 8

Methods for parameter estimation and inference

Often key research questions can be phrased in terms of a statistical model, and in particular, in terms of a specific parameter in the statistical model. We would like to be able to make inferences about the parameter of interest, after collecting some data – tools for **parametric inference** are desired. Historically, the majority of statistical work (both theoretical and applied) has been concerned with developing and applying tools for parametric inference. The most important contributions of statistics to scientific research have been in this important area.

Chapter 7 introduced the basic statistical inference concepts of *estimation* and *confidence intervals*. These ideas were worked through in the special cases of means of (approximately) normal samples, and proportions for binomial samples. But what about more general statistical models? How do you come up with a good estimator in such situations? How can you construct a confidence interval for the parameter of interest, in such situations? This chapter provides some tools for answering these important questions.

Throughout this chapter it is useful to keep in mind the distinction between *estimates* and *estimators*. An estimate of a parameter θ is a function $\widehat{\theta} = \widehat{\theta}(x_1, \ldots, x_n)$ of observed values $x_1, \ldots, x_n$, whereas the corresponding estimator is the same function $\widehat{\theta}(X_1, \ldots, X_n)$ of the observable random variables $X_1, \ldots, X_n$. Thus an estimator is a random variable whose properties may be examined and considered before the observation process occurs, whereas an estimate is an actual number, the realized value of the estimator, evaluated after the observations are available.

In deriving estimation formulas it is often easier to work with estimates, but in considering theoretical properties, switching to estimators may be necessary.

For notational convenience, we usually denote the density or probability function f_X simply by f.

For further reading, consider Hogg *et al* (2005) sections 6.1-6.2 or Rice (2007) sections 8.4-8.5 or Chapter 6 of Kroese & Chan (2014).

Method of Moments Estimation

A particularly simple approach to estimation is the *method of moments* suggested by the British polymath Karl Pearson.

Definition

Let $x_1, \ldots, x_n$ be observations from the model

$$f = f(x; \theta_1, \ldots, \theta_k)$$

containing k parameters $\theta = (\theta_1, \ldots, \theta_k)$. Form the system of k equations that equates the moments of f_X with their sample counterparts:

$$\begin{aligned} \mathbb{E}(X) &= \frac{1}{n}\sum_i x_i \\ \mathbb{E}(X^2) &= \frac{1}{n}\sum_i x_i^2 \\ &\vdots \\ \mathbb{E}(X^k) &= \frac{1}{n}\sum_i x_i^k. \end{aligned}$$

Then the **method of moments** estimates are the solutions of these equations in $\theta_1, \ldots, \theta_k$.

Example

Consider a random sample with normal model:

$$X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2).$$

Find the method of moments estimators of μ and σ^2.

The method of moments equations are:

$$\begin{aligned} \mathbb{E}(X) &= \frac{1}{n}\sum_i x_i \\ \mathbb{E}(X^2) &= \frac{1}{n}\sum_i x_i^2 \end{aligned}$$

Figure 8.1: Karl Pearson (27 March 1857 – 27 April 1936) is one of the father of statistics, meteorology, epidemiology, and biometrics. He founded the first department of statistics in the world. We owe the correlation coefficient and the method of moments to Pearson, amongst other things. His books on the philosophy of science influenced the young A. Einstein.

But

$$\mathbb{E}(X) = \mu \quad \text{and} \quad \mathbb{E}(X^2) = \text{Var}(X) + \{\mathbb{E}(X)\}^2 = \sigma^2 + \mu^2$$

which leads to the system of equations:

$$\begin{aligned} \mu &= \bar{x} \\ \sigma^2 + \mu^2 &= \frac{1}{n}\sum_i x_i^2 \end{aligned}$$

Substitution of the first equation into the second leads to

$$\sigma = \sqrt{\frac{1}{n}\sum_i x_i^2 - \bar{x}^2} = \sqrt{\frac{1}{n}\sum_i (x_i - \bar{x})^2}.$$

So the method of moments estimators of μ and σ are:

$$\widehat{\mu} = \bar{X} \quad \text{and} \quad \widehat{\sigma} = \sqrt{\frac{1}{n}\sum_i (X_i - \bar{X})^2}.$$

Consistency of method of moments estimators

Assuming that $\mathrm{Var}(X^k) < \infty$, the Weak Law of Large Numbers states that

$$\frac{1}{n}\sum_i X_i^j \xrightarrow{\mathbb{P}} \mathbb{E}(X^j), \quad 1 \leq j \leq k$$

Hence we can establish that

$$\widehat{\theta}_j \xrightarrow{\mathbb{P}} \theta_j, \quad 1 \leq j \leq k.$$

That is, method of moments leads to consistent estimation of the model parameters.

Methods of moments estimation is useful in practice because it is a simple approach that guarantees us a consistent estimator. However it is not always optimal, in the sense that it does not always provide us with an estimator that has the smallest possible standard errors and mean squared error. There is however a method that is (usually) optimal...

Maximum Likelihood Estimation

In this section we introduce *maximum likelihood estimation*, a procedure that has optimal performance for large samples, *for almost any model*! Hence this method is a very important tool in statistical work. When this estimation method is possible, it is usually as good or better than method of moments estimation.

We will start with the single parameter case. The extension to multi-parameter models will be discussed later in this chapter.

We first define the *likelihood function*:

Definition
Let $x_1, \ldots, x_n$ be observation from the pdf f where

$$f(x) = f(x;\theta)$$

depending on a parameter $\theta \in \Theta$. The **likelihood function** $\mathcal{L}$, a function of θ, is

$$\mathcal{L}(\theta) = f(x_1;\theta)\cdots f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta), \quad \theta \in \Theta,$$

and the **log-likelihood function** of θ is

$$\ell(\theta) = \ln\{\mathcal{L}(\theta)\} = \sum_i \ln\{f(x_i;\theta)\}.$$

Note that the form of the likelihood function as a function of the observations

$x_1, \ldots, x_n$ is the same as the joint density function, but the likelihood function is regarded as a function of θ, for fixed values of $\{x_i\}$.

Example

Let

$$X_i = \begin{cases} 1, & \text{if } i\text{-th surveyed student watched the World Cup Cricket final} \\ 0, & \text{otherwise.} \end{cases}$$

for $i = 1, \ldots, 8$. An appropriate probability function for $\{X_i\}$ is $f(x; p)$, where

$$\begin{aligned} f(x;p) &= \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \end{cases} \\ &= p^x(1-p)^{1-x} \end{aligned}$$

Find the likelihood and log-likelihood functions, ie $\mathcal{L}(p)$ and $\ell(p)$, hence find expressions for $\mathcal{L}(0.3)$ and $\ell(0.3)$.

The likelihood function for p is

$$\mathcal{L}(p) = \prod_{i=1}^{8}\{p^{x_i}(1-p)^{1-x_i}\} = p^{\sum_{i=1}^{8} x_i}(1-p)^{8-\sum_{i=1}^{8} x_i}$$

and the log-likelihood function for p is

$$\ell(p) = \ln\{\mathcal{L}(p)\} = \left(\sum_{i=1}^{8} x_i\right)\ln(p) + \left(8 - \sum_{i=1}^{8} x_i\right)\ln(1-p).$$

Taking $p = 0.3$,

$$\mathcal{L}(0.3) = (0.3)^{\sum_{i=1}^{8} x_i}(0.7)^{8-\sum_{i=1}^{8} x_i}$$

and

$$\ell(0.3) = \left(\sum_{i=1}^{8} x_i\right)\ln(0.3) + \left(8 - \sum_{i=1}^{8} x_i\right)\ln(0.7)$$

Definition

Let $x_1, \ldots, x_n$ be observations from probability/density function f, where

$$f(x) = f(x; \theta)$$

containing the parameter $\theta \in \Theta$. The **maximum likelihood estimate** of θ is the choice

$$\widehat{\theta} = \theta \text{ that maximises } \mathcal{L}(\theta) \text{ over } \theta \in \Theta.$$

Example

In the World Cup Cricket example ($n = 8$) the likelihood function has been shown to be

$$\mathcal{L}(p) = \prod_{i=1}^{8} \{p^{x_i}(1-p)^{1-x_i}\} = p^{\sum_{i=1}^{8} x_i}(1-p)^{8-\sum_{i=1}^{8} x_i} = e^{\left(\sum_{i=1}^{8} x_i\right)\ln(p)+\left(8-\sum_{i=1}^{8} x_i\right)\ln(1-p)}.$$

Find the maximum likelihood estimator of p by finding the value of p that maximises $\mathcal{L}(p)$.

The first derivative of $\mathcal{L}(p)$ with respect to p is then

$$\begin{aligned}\frac{d}{dp}\mathcal{L}(p) &= e^{\left(\sum_{i=1}^{8} x_i\right)\ln(p)+\left(8-\sum_{i=1}^{8} x_i\right)\ln(1-p)}\left[\left(\sum_{i=1}^{8} x_i\right)/p - \left(8 - \sum_{i=1}^{8} x_i\right)/(1-p)\right] \\ &= \mathcal{L}(p)\left(\frac{\sum_{i=1}^{8} x_i}{p} - \frac{8-\sum_{i=1}^{8} x_i}{1-p}\right).\end{aligned}$$

and is zero if and only if

$$\frac{\sum_{i=1}^{8} x_i}{p} - \frac{8-\sum_{i=1}^{8} x_i}{1-p} = 0 \iff p = \frac{\sum_{i=1}^{8} x_i}{8}.$$

Further analysis (see next example) shows that this is the unique maximiser of $\mathcal{L}(p)$ over $0 < p < 1$ so

$$\widehat{p} = \frac{\sum_{i=1}^{8} x_i}{8} = \text{proportion of people in survey that watched the World Cup Cricket final}$$

is the maximum likelihood estimate of p.

Example

Suppose that the observed data are:

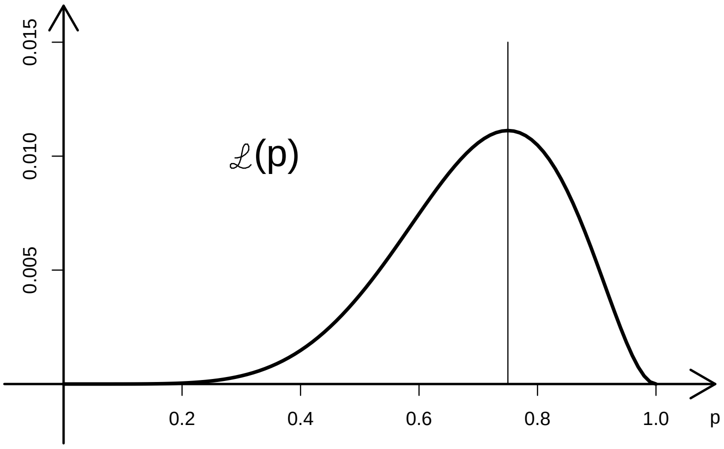
$$x_1 = 0,\ x_2 = 1,\ x_3 = 1,\ x_4 = 1,\ x_5 = 1,\ x_6 = 0,\ x_7 = 1,\ x_8 = 1.$$

Plot the observed likelihood function of p.

The observed value of the likelihood function is

$$\mathcal{L}(p) = p^6(1-p)^2, \quad 0 < p < 1.$$

The following figure is a graphical illustration of the maximum likelihood procedure in this case. Note that $\widehat{p} = 6/8 = 0.75$ is the value of $p \in (0,1)$ that maximises $\mathcal{L}(p)$.



Obtaining maximum likelihood estimators

As the previous definition shows, maximum likelihood estimation boils down to the problem of determining where a function reaches its maximum. The mechanics of determination of the maximum differ depending on the smoothness of $\mathcal{L}(\theta)$.

Smooth likelihood functions

Consider estimation of a general parameter θ. If $\mathcal{L}(\theta)$ is smooth then differential calculus methods can be employed to obtain the maximiser of $\mathcal{L}(\theta)$. However, it is usually simpler to work with the log-likelihood function $\ell(\theta)$. Maximising $\ell(\theta)$ rather than $\mathcal{L}(\theta)$ is justified by:

Result

The point at which $\mathcal{L}(\theta)$ attains its maximum over $\theta \in \Theta$ is also that where

$$\ell(\theta) = \ln\{\mathcal{L}(\theta)\} = \sum_i \ln\{f(x_i;\theta)\}$$

attains its maximum. Therefore, the maximum likelihood estimate of θ is

$$\widehat{\theta} = \theta \text{ that maximises } \ell(\theta) \text{ over } \theta \in \Theta.$$

Example

Re-visit the World Cup Cricket example.

Find the maximum likelihood estimator of p by finding the value of p that maximises $\ell(p)$. Then plot $\ell(p)$ when the observed data are

$$x_1 = 0,\ x_2 = 1,\ x_3 = 1,\ x_4 = 1,\ x_5 = 1,\ x_6 = 0,\ x_7 = 1,\ x_8 = 1$$

$$\ell(p) = \ln\{\mathcal{L}(p)\} = \left(\sum_{i=1}^{8} x_i\right)\ln(p) + \left(8 - \sum_{i=1}^{8} x_i\right)\ln(1-p).$$

The first derivative is

$$\frac{d}{dp}\ell(p) = \frac{\sum_{i=1}^{8} x_i}{p} - \frac{8 - \sum_{i=1}^{8} x_i}{1-p}$$

and is zero if and only if

$$\frac{\sum_{i=1}^{8} x_i}{p} - \frac{8 - \sum_{i=1}^{8} x_i}{1-p} = 0 \iff p = \frac{\sum_{i=1}^{8} x_i}{8},$$

which is the same answer obtained previously using $\mathcal{L}(p)$, but via *simpler calculus*!

Is this the unique maximiser of $\ell(p)$ over $p \in (0,1)$? The second derivative is

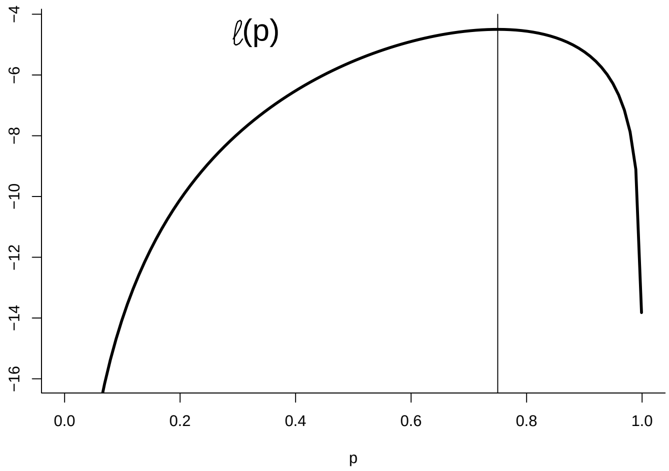
$$\tfrac{d^2}{dp^2}\ell(p) = \tfrac{-\sum_{i=1}^{8} x_i}{p^2} - \tfrac{8-\sum_{i=1}^{8} x_i}{(1-p)^2}$$

which is negative for all $0 < p < 1$ and samples $x_i \in \{0,1\}$, $1 \leq i \leq 8$. Hence $\ell(p)$ is concave (downwards) over $0 < p < 1$ and the point at which $\frac{d}{dp}\ell(p) = 0$ must be a maximum.

For the case where the observed data are again

$$x_1 = 0,\ x_2 = 1,\ x_3 = 1,\ x_4 = 1,\ x_5 = 1,\ x_6 = 0,\ x_7 = 1,\ x_8 = 1$$

the following figure shows the maximum likelihood estimation procedure via the log-likelihood function $\ell(p)$. As in the case of $\mathcal{L}(p)$ (see previous figure), the maximiser occurs at $\widehat{p} = 6/8 = 0.75$.



Example

Consider an observed sample $x_1, \ldots, x_n$ with common density function f, where

$$f(x;\theta) = 2\theta x \mathrm{e}^{-\theta x^2}, \; x \geq 0; \; \theta > 0.$$

Write down the log-likelihood function for θ. Show that any stationary point on this function maximises $\ell(\theta)$, hence find the maximum likelihood estimator of θ.

Non-smooth likelihood functions*

Not all likelihood functions are differentiable, or even continuous, over $\theta \in \Theta$. In such non-smooth cases calculus methods, alone, cannot be used to locate the maximiser, and it is usually better to work directly with $\mathcal{L}(\theta)$ rather than $\ell(\theta)$. The following notation is useful in non-smooth likelihood situations.

Definition

Let $\mathcal{P}$ be a logical condition. Then the **indicator function** of $\mathcal{P}$, $\mathbb{I}(\mathcal{P})$ is given by

$$\mathbb{I}(\mathcal{P}) = \begin{cases} 1 & \text{if } \mathcal{P} \text{ is true,} \\ 0 & \text{if } \mathcal{P} \text{ is false.} \end{cases}$$

Example

Some examples of use of $\mathbb{I}$ are

$$\begin{aligned}
&\mathbb{I}(\text{Cristina Keneally was the premier of New South Wales on 14th February, 2011}) = 1.\\
&\mathbb{I}(4^2 = 16) = 1,\\
&\mathbb{I}(e^\pi = 17) = 0,\\
&\mathbb{I}(\text{The Earth is bigger than the Moon}) = 1,\\
&\mathbb{I}(\text{The Earth is bigger than the Moon \& The Moon is made of blue cheese}) = 0.
\end{aligned}$$

The $\mathbb{I}$ notation allows one to write density functions in explicit algebraic terms. For example, the Gamma(α, β) density function is usually written as

$$f(x; \alpha, \beta) = \frac{e^{-x/\beta} x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha}, \quad x > 0.$$

However, it can also be written using $\mathbb{I}$ as

$$f(x; \alpha, \beta) = \frac{e^{-x/\beta} x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} \mathbb{I}(x > 0).$$

The following result is useful for deriving maximum likelihood estimators when the likelihood function is non-smooth:

Result

For any two logical conditions $\mathcal{P}$ and $\mathcal{Q}$,

$$\mathbb{I}(\mathcal{P} \cap \mathcal{Q}) = \mathbb{I}(\mathcal{P})\mathbb{I}(\mathcal{Q}).$$

Non-smooth likelihood functions arise when the range of f depends on θ.

Example

Suppose that the observation $x_1, \ldots, x_n$ come from pdf f, where

$$f(x; \theta) = 5(x^4/\theta^5), \quad 0 < x < \theta.$$

Use the $\mathbb{I}$ notation to find an expression for $\mathcal{L}(\theta)$, and simplify.

$$f(x; \theta) = 5(x^4/\theta^5)\mathbb{I}(0 < x < \theta) = 5(x^4/\theta^5)\mathbb{I}(x > 0)\mathbb{I}(\theta > x).$$

The likelihood function is then

$$\begin{aligned}\mathcal{L}(\theta) &= \prod_{i=1}^{n} f(x_i;\theta)\\ &= 5(x_1^4/\theta^5)\mathbb{I}(x_1>0)\mathbb{I}(\theta>x_1)\cdots 5(x_n^4/\theta^5)\mathbb{I}(x_n>0)\mathbb{I}(\theta>x_n)\\ &= 5^n\left(\prod_{i=1}^{n} x_i\right)^4\left\{\prod_{i=1}^{n}\mathbb{I}(x_i>0)\right\}\left\{\prod_{i=1}^{n}\mathbb{I}(\theta>x_i)\right\}\theta^{-5n}\end{aligned}$$

Note that

$$\prod_{i=1}^{n}\mathbb{I}(\theta>x_i)=\mathbb{I}(\theta>x_1,\theta>x_2,\cdots,\theta>x_n)=\mathbb{I}(\theta>\max(x_1,\ldots,x_n)).$$

Also, $\prod_{i=1}^{n}\mathbb{I}(x_i>0)=1$ with probability 1, since $\mathbb{P}(X_i>0)=1$. Hence, the likelihood function is

$$\mathcal{L}(\theta)=5^n\left(\prod_{i=1}^{n} x_i\right)^4\theta^{-5n}\mathbb{I}(\theta>\max(x_1,\ldots,x_n))$$

or, even more digestibly,

$$\mathcal{L}(\theta)=\begin{cases}C_n\theta^{-5n}, & \theta>\max(x_1,\ldots,x_n)\\ 0, & \text{otherwise}\end{cases}$$

where $C_n=5^n\left(\prod_{i=1}^{n}x_i\right)^4$. The accompanying figure shows an example of such an $\mathcal{L}(\theta)$.

Since $C_n\theta^{-5n}$ is clearly decreasing for $\theta>\max(x_1,\ldots,x_n)$ (easily verified via calculus) it is clear that $\mathcal{L}(\theta)$ attains its maximum at $\max(x_1,\ldots,x_n)$. Thus, the maximum likelihood estimator of θ is

$$\widehat{\theta}=\max(X_1,\ldots,X_n).$$

Properties of maximum likelihood estimators

Consistency

Suppose

$$X_1,\ldots,X_n\overset{\text{iid}}{\sim}f(x;\theta^*)$$

for some unknown θ^* in the interior of the set Θ, where Θ is the set of allowable values for θ. Let us have the assumptions

1. The domain of x does not depend on θ^*, that is, all pdfs $f(\cdot;\theta)$ have common support (are nonzero over the same fixed set, independent of θ^*).

2. If $\theta \neq \vartheta$, then
$$f(x;\theta) \neq f(x;\vartheta)\ .$$
In other words, can identify if $\vartheta = \theta$ from the equivalence of the densities $f(x;\vartheta)$ and $f(x;\theta)$.

3. The MLE
$$\widehat{\theta}_n = \operatorname*{argmax}_{\theta\in\Theta} f(\mathbf{X};\theta)\ ,$$
where $f(\mathbf{X};\theta) = \prod_{i=1}^n f(X_i;\theta)$, is unique and lies in the interior of Θ.

Consistency
Under the assumptions above, the maximum likelihood estimator $\widehat{\theta}_n$ of θ^* is consistent; i.e.
$$\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta^*.$$

proof: Define the log-likelihood function based on the n data points
$$\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ln f(X_i;\theta)$$
and note that by the Law of Large Numbers we have
$$\ell_n(\theta) \xrightarrow{\mathbb{P}} \ell(\theta) = \mathbb{E}_{\theta^*} \ln f(X;\theta) \qquad (X \sim f(x;\theta^*))$$
By assumption 3, we have that
$$\ell_n(\widehat{\theta}_n) > \ell_n(\theta)$$
for any $\theta \neq \widehat{\theta}_n$ with equality only when $\theta = \widehat{\theta}_n$.

Similarly,
$$\begin{aligned} \ell(\theta^*) - \ell(\theta) &= \mathbb{E}_{\theta^*} \ln f(X;\theta^*) - \ln f(X;\theta) \\ &= -\mathbb{E}_{\theta^*} \ln \frac{f(X;\theta)}{f(X;\theta^*)} \\ \text{(by Jensen's ineq. and 1)} \quad &\geqslant -\ln \mathbb{E}_{\theta^*} \frac{f(X;\theta)}{f(X;\theta^*)} = -\ln(1) = 0 \end{aligned}$$

Note that $\mathbb{E}_{\theta^*} \ln \frac{f(X;\theta)}{f(X;\theta^*)} = 0$ if and only if $f(X;\theta) = f(X;\theta^*)$. Hence, if $\theta \neq \theta^*$, then by assumption 2 we have $f(X;\theta) \neq f(X;\theta^*)$ and $-\mathbb{E}_{\theta^*} \ln \frac{f(X;\theta)}{f(X;\theta^*)} > 0$. Therefore, if $\theta \neq \theta^*$, then
$$\ell(\theta^*) > \ell(\theta)$$

with equality only if $\theta = \theta^*$. We now observe that

$$\ell_n(\theta^*) - \ell_n(\theta) = \underbrace{\ell_n(\theta^*) - \ell(\theta^*)}_{\xrightarrow{\mathbb{P}} 0} + \underbrace{\ell(\theta) - \ell_n(\theta)}_{\xrightarrow{\mathbb{P}} 0} + \underbrace{\ell(\theta^*) - \ell(\theta)}_{=c \geqslant 0}$$

Hence, for $\theta \neq \theta^*$ we have $c > 0$ and

$$\ell_n(\theta^*) - \ell_n(\theta) \xrightarrow{\mathbb{P}} c > 0$$

This can also be written as

$$\mathbb{P}(\ell_n(\theta^*) > \ell_n(\theta)) \to 1, \qquad \theta \neq \theta^*$$

Now note that by definition of MLE and its assumed uniqueness (assumption 3) we have

$$\{\omega : \widehat{\theta}_n(\omega)\} = \{\omega : \ell_n(\widehat{\theta}_n) \geqslant \ell_n(\theta) \text{ for all } \theta \in \Theta\}$$

Hence, for any $\varepsilon > 0$

$$\begin{aligned} \mathbb{P}(|\widehat{\theta}_n - \theta^*| > \varepsilon) &= \mathbb{P}(|\widehat{\theta}_n - \theta^*| > \varepsilon,\ \ell_n(\widehat{\theta}_n) \geqslant \ell_n(\theta) \text{ for all } \theta \in \Theta) \\ \text{by conjunction fallacy } &\leqslant \mathbb{P}(|\widehat{\theta}_n - \theta^*| > \varepsilon,\ \ell_n(\widehat{\theta}_n) \geqslant \ell_n(\theta^*)) \\ \text{by conjunction fallacy } &\leqslant \mathbb{P}(\ell_n(\theta) \geqslant \ell_n(\theta^*), \text{ for any } \theta \neq \theta^*) \\ &\leqslant 1 - \mathbb{P}(\ell_n(\theta^*) > \ell_n(\theta),\ \theta \neq \theta^*) \to 0 \end{aligned}$$

Hence, by definition $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta^*$.

Equivariance

Maximum likelihood estimators are *equivariant* under functions of the parameter of interest:

Equivariance
Suppose $\widehat{\theta}$ is the maximum likelihood estimator of θ. Then for any function g

$$g(\widehat{\theta}) \text{ is the maximum likelihood estimator of } g(\theta).$$

Example

Let $X_1, \ldots, X_n$ be random variables each with density function f, where

$$f(x;\theta) = 2\theta x e^{-\theta x^2}, \quad x > 0.$$

It has previously been shown that the maximum likelihood estimator of θ is

$$\widehat{\theta} = \frac{n}{\sum_i X_i^2}.$$

Find the maximum likelihood estimators of $\tau = 1/\theta$ and $\omega = \ln(\theta)$.

From the equivariance property of maximum likelihood estimation, the maximum likelihood estimator of $\tau = 1/\theta$ is

$$\widehat{\tau} = \frac{1}{\widehat{\theta}} = \frac{1}{n}\sum_i X_i^2$$

and the maximum likelihood estimator of $\omega = \ln(\theta)$ is

$$\widehat{\omega} = \ln(\widehat{\theta}) = \ln(n) - \ln\left(\sum_i X_i^2\right).$$

Variance and standard error

Let

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$$

be continuous random variables for some unknown $\theta \in \Theta$ and let

$$\ell_n(\theta) = \sum_{i=1}^{n} \ln f(X_i;\theta)$$

be the corresponding log-likelihood function (not normalized by n). In addition to the last three technical assumptions we assume that

4. $f(\cdot;\theta)$ is twice differentiable in θ;

5. $\int f(x;\theta)\mathrm{d}x$ can be twice differentiated under the integral.

To continue we introduce the following quantities.

Definition

The **Fisher score** is defined as

$$S_n(\theta) = \ell_n'(\theta)$$

and the **Fisher information** is defined as

$$I_n(\theta) = -\mathbb{E}_\theta \ell_n''(\theta),$$

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

Figure 8.2: Ronald Fisher: the inventor of the maximum likelihood method and “the greatest biologist since Darwin”. Fisher and Pearson were bitter rivals as each considered their method of estimation and inference superior.

Property of the Fisher score and information

$$\mathbb{E}_\theta S_n(\theta) = 0.$$

$$I_n(\theta) = \mathbb{E}_\theta[\ell_n'(\theta)]^2 = \mathrm{Var}_\theta(S_n(\theta)).$$

proof: We have

$$\begin{aligned} 0 = \frac{\partial}{\partial\theta}(1) &= \frac{\partial}{\partial\theta}\left(\int f(x;\theta)\mathrm{d}x\right) \\ &= \int \frac{\partial}{\partial\theta} f(x;\theta)\mathrm{d}x \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{E}_\theta S_n(\theta) &= n\mathbb{E}_\theta S_1(\theta), \qquad \text{by iid assumption} \\ &= n\int f(x;\theta)\frac{\partial}{\partial\theta}\ln f(x;\theta)\mathrm{d}x \\ &= n\int f(x;\theta)\frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)}\mathrm{d}x \\ &= n\int \frac{\partial}{\partial\theta}f(x;\theta)\mathrm{d}x = 0 \end{aligned}$$

Finally,

$$
\begin{aligned}
I_n(\theta) &= nI_1(\theta) \\
&= -n \int f(x;\theta) \frac{\partial^2}{\partial^2\theta} \ln f(x;\theta) \mathrm{d}x \\
&= -n \int f(x;\theta) \frac{\frac{\partial^2 f}{\partial^2\theta}(x;\theta) f(x;\theta) - \frac{\partial f}{\partial\theta}(x;\theta)\frac{\partial f}{\partial\theta}(x;\theta)}{f^2(x;\theta)} \mathrm{d}x \\
&= -n \int \frac{\partial^2 f}{\partial^2\theta}(x;\theta)\mathrm{d}x + n \int f(x;\theta) \frac{\frac{\partial f}{\partial\theta}(x;\theta)}{f(x;\theta)} \frac{\frac{\partial f}{\partial\theta}(x;\theta)}{f(x;\theta)} \mathrm{d}x \\
&= 0 + n\mathbb{E}_\theta[S_1(\theta)]^2 = \mathrm{Var}_\theta(S_n(\theta))
\end{aligned}
$$

□

It is now possible to show that a maximum likelihood estimate has asymptotic mean square error and standard error that are functions of the **Fisher information**.

Result

Let $X_1, \ldots, X_n$ be random variables with common density function f depending on a parameter θ, and let $\widehat{\theta}_n$ be the maximum likelihood estimator of θ. Then as $n \to \infty$

$$I_n(\theta)\,\mathrm{Var}(\widehat{\theta}_n) \xrightarrow{\mathbb{P}} 1$$

Hence we can say that

$$\mathrm{se}(\widehat{\theta}) \approx \frac{1}{\sqrt{I_n(\widehat{\theta}_n)}}$$

Example

Recall the previous example of a random sample $x_1, \ldots, x_n$ from a common density function f, where

$$f(x;\theta) = 2\theta x e^{-\theta x^2}, \; x \geq 0; \; \theta > 0.$$

Find the Fisher Information for θ, and hence the approximate $\mathrm{se}(\widehat{\theta})$.

From previously, the log-likelihood function, written as a random variable, is

$$\ell(\theta) = n\ln(2) + n\ln(\theta) + \sum_i \ln(X_i) - \theta \sum_i X_i^2.$$

the first and second derivatives of $\ell(\theta)$ are

$$\ell'(\theta) = n\theta^{-1} - \sum_i X_i^2 \qquad \text{and} \qquad \ell''(\theta) = -n\theta^{-2}.$$

Therefore, the Fisher information is

$$I_n(\theta) = -\mathbb{E}(-n\theta^{-2}) = n/\theta^2.$$

Hence the standard error of $\widehat{\theta}$ is approximately

$$\mathrm{se}(\widehat{\theta}) \simeq \frac{1}{\sqrt{I_n(\widehat{\theta})}} = \frac{\widehat{\theta}}{\sqrt{n}}$$

Note that in the above example, $\ell''(\theta)$ has no random variables present so the expected value operation is redundant. This will not be the case in general, as in the example below.

Example

Consider the daily number of hospital admissions due to asthma attacks. This can be considered as a count of rare events, which can be modelled using the Poisson distribution, where

$$f(x;\lambda) = \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}, \quad x \in \{0, 1, 2, ...\}$$

Let $X_1, X_2, \ldots, X_n$ be the observed number of hospital admissions due to asthma attacks on n randomly selected days (which can be considered to be iid).

Find:

1. $\ell(\lambda)$.
2. The maximum likelihood estimator for λ.
3. The Fisher information for λ.
4. Hence approximate $\mathrm{se}(\widehat{\lambda})$ for large n.

Asymptotic normality

In addition to assumptions 1 through 5 we now add the following one

6. Assume the existence of a third derivative of $\ln f(x;\vartheta)$, such that

$$\left|\frac{\partial^3}{\partial^3\vartheta}\ln f(x;\vartheta)\right| \leqslant g(x), \quad \text{for all } \vartheta \in \Theta$$

for some dominating function g with finite expectation $\mathbb{E}_\theta[g(X)] = \mu < \infty$.

Theorem: Asymptotic Normality of Maximum Likelihood Estimators

Under the smoothness assumptions 1 through 6 above,

$$\frac{\widehat{\theta}-\theta}{\sqrt{\mathrm{Var}(\widehat{\theta})}} \xrightarrow{\mathrm{d}} \mathsf{N}(0,1) \quad \text{and} \quad \frac{\widehat{\theta}-\theta}{\mathrm{se}(\widehat{\theta})} \xrightarrow{\mathrm{d}} \mathsf{N}(0,1)$$

where

$$\mathrm{se}(\widehat{\theta}) = \tfrac{1}{\sqrt{I_n(\theta)}}$$

This is one of the most important theorems in statistics!

proof: By the mean value theorem we can write

$$\ell_n'(\widehat{\theta}_n) = \ell_n'(\theta) + \ell_n''(\vartheta_n)(\widehat{\theta}_n - \theta)$$

for some ϑ_n between θ and $\widehat{\theta}_n$. By the consistency of MLE we have $\widehat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ and $\vartheta_n \xrightarrow{\mathbb{P}} \theta$. By the Central Limit Theorem, we have

$$\frac{\ell_n'(\theta) - n \times 0}{\sqrt{nI_1(\theta)}} = \frac{\ell_n'(\theta) - n\mathbb{E}_\theta S_1(\theta)}{\sqrt{nI_1(\theta)}} \xrightarrow{\mathrm{d}} Z \sim \mathsf{N}(0,1)$$

Hence, since $\ell_n'(\widehat{\theta}_n) = 0$ (due to uniqueness and differentiability, the MLE is the critical point of the log-likelihood), we have

$$\begin{aligned}\frac{\sqrt{n}(\widehat{\theta}_n - \theta)}{I_1^{-1/2}(\theta)} &= \frac{\sqrt{n}}{I_1^{-1/2}(\theta)}\frac{-\ell_n'(\theta)}{\ell_n''(\vartheta_n)} \\ &= \frac{I_1(\theta)}{-\ell_n''(\vartheta_n)/n} \times \frac{\ell_n'(\theta)}{\sqrt{nI_1(\theta)}}\end{aligned}$$

It seems plausible that $-\ell_n''(\vartheta_n)/n \xrightarrow{\mathbb{P}} I_1(\theta)$, because $\vartheta_n \xrightarrow{\mathbb{P}} \theta$ and $-\ell_n''(\theta)/n \xrightarrow{\mathbb{P}} I_1(\theta)$ (by the Law of Large Numbers). If we accept this, then the result follows from Slutsky's theorem:

$$\frac{\sqrt{n}(\widehat{\theta}_n - \theta)}{I_1^{-1/2}(\theta)} = \underbrace{\frac{I_1(\theta)}{-\ell_n''(\vartheta_n)/n}}_{\xrightarrow{\mathbb{P}} 1} \times \underbrace{\frac{\ell_n'(\theta)}{\sqrt{nI_1(\theta)}}}_{\xrightarrow{\mathrm{d}} Z} \xrightarrow{\mathrm{d}} Z \sim \mathsf{N}(0,1)\ .$$

For the interested students only, we provide the following.

Remark 8.1 (Convergence of $-\ell_n''(\vartheta_n)/n$) *We can write by the mean value theorem*

$$\ell_n''(\vartheta_n) = \ell_n''(\theta) + (\vartheta_n - \theta)\ell_n'''(\xi_n)$$

for some $\xi_n \xrightarrow{\mathbb{P}} \theta$, which is between θ and ϑ_n. Assume the existence of a third derivative of $\ln f(x;\vartheta)$, such that

$$\left|\frac{\partial^3}{\partial^3\vartheta}\ln f(x;\vartheta)\right| \leqslant g(x), \quad \textit{for all } \vartheta \in \Theta$$

for some dominating function g with finite expectation $\mathbb{E}_\theta[g(X)] = \mu < \infty$. Hence, we can write

$$\left|\frac{\ell_n'''(\xi_n)}{n}\right| \leqslant \frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{\mathbb{P}} \mu < \infty$$

and therefore by Slutsky's theorem and using the fact that $X_n \xrightarrow{\mathrm{d}}$ constant implies $X_n \xrightarrow{\mathbb{P}}$ constant, we have

$$\left|(\vartheta_n - \theta)\ell_n'''(\xi_n)/n\right| = |\vartheta_n - \theta|\left|\frac{\ell_n'''(\xi_n)}{n}\right| \xrightarrow{\mathbb{P}} 0 \times \mu = 0$$

Hence,

$$\frac{-\ell_n''(\vartheta_n)}{n} = \underbrace{\frac{-\ell_n''(\theta)}{n}}_{\xrightarrow{\mathbb{P}} I_1(\theta)} - \underbrace{(\vartheta_n - \theta)\frac{\ell_n'''(\xi_n)}{n}}_{\xrightarrow{\mathbb{P}} 0} \xrightarrow{\mathbb{P}} I_1(\theta)\ .$$

This result is important because it means that maximum likelihood is not only useful for estimation, but is also a method for making inferences about parameters. Because we now know how to find the approximate distribution of any maximum likelihood estimator $\widehat{\theta}$, we can now calculate standard errors and construct confidence intervals for θ using $\widehat{\theta}$ for data from any family of distributions of known form.

Example

Recall the previous example:

$$X_1, \ldots, X_n \sim f, \quad \text{where} \quad f(x;\theta) = 2\theta x \mathrm{e}^{-\theta x^2},\ x \geq 0;\ \theta > 0$$

Find the estimated standard error of $\widehat{\theta}$, and the approximate distribution of $\widehat{\theta}$.

We previously found the Fisher information to be

$$I_n(\theta) = n/\theta^2.$$

Therefore, the approximate variance of $\widehat{\theta}$ is

$$\mathrm{Var}(\widehat{\theta}) \simeq \frac{\theta^2}{n}$$

and the asymptotic standard error of $\widehat{\theta}$ is

$$\widehat{\mathrm{se}}(\widehat{\theta}) = \frac{1}{\sqrt{I_n(\widehat{\theta})}} = \widehat{\theta}/\sqrt{n}.$$

Thus

$$\frac{\widehat{\theta}-\theta}{\theta/\sqrt{n}} \xrightarrow{\text{d}} \mathsf{N}(0,1).$$

This means that we can approximate the distribution of $\widehat{\theta}$ using

$$\widehat{\theta} \overset{\text{appr.}}{\sim} \mathsf{N}(\theta, \theta^2/n)$$

Example

Recall the n measurements of the daily number of hospital admissions due to asthma attacks, $X_1, X_2, \ldots, X_n$, which we will model using the Poisson distribution, *i.e.*

$$f(x;\lambda) = \text{e}^{\lambda}\frac{\lambda^x}{x!}, \quad x \in \{0,1,2,...\}$$

Here, λ has the interpretation of being the average number of hospital admissions due to asthma attacks per day.

Find the approximate distribution of $\widehat{\lambda}$, the maximum likelihood estimator of λ.

We have

$$\ell(\theta) = n\ln(\theta) + \sum_i \ln(X_i) - \theta\sum_i X_i^2 + \text{const.}$$

From here it follows that

$$\dot{\ell}(\theta) = \frac{n}{\theta} - \sum_i X_i^2$$

with solution $\widehat{\theta} = \sum_i X_i^2/n$. This is maximum, because

$$\ddot{\ell}(\theta) = -n/\theta^2 < 0$$

The Delta Method permits extension of the asymptotic normality result to a general smooth function of θ:

Result

Under appropriate regularity conditions, including the existence of two derivatives of $\mathcal{L}(\theta)$, if $\tau = g(\theta)$ and $\widehat{\tau} = g(\widehat{\theta})$, where g is differentiable and $g'(\theta) \neq 0$, then

$$\frac{\widehat{\tau}-\tau}{\sqrt{\text{Var}(\widehat{\tau})}} \xrightarrow{\text{d}} \mathsf{N}(0,1)$$

where

$$\text{Var}(\widehat{\tau}) \simeq \frac{\{g'(\theta)\}^2}{I_n(\theta)}$$

This result follows directly from the delta method result given in Chapter 5.

Example

Recall the example:

$$X_1, \ldots, X_n \sim f, \quad \text{where} \quad f(x;\theta) = 2\theta x \mathrm{e}^{-\theta x^2}, \; x \geq 0; \; \theta > 0$$

Suppose that the parameter of interest is $\omega = \ln(\theta)$. Use maximum likelihood estimation to find an estimator of ω and its approximate distribution.

From previous working, the maximum likelihood estimator of ω is

$$\widehat{\omega} = \ln(n) - \ln\left(\sum_i X_i^2\right).$$

and the variance of $\widehat{\omega}$ is

$$\mathrm{Var}(\widehat{\omega}) \simeq \frac{|1/\theta|^2}{n/\theta^2} = \frac{1}{n}.$$

Thus

$$\frac{\widehat{\omega} - \omega}{1/\sqrt{n}} \xrightarrow{\mathrm{d}} \mathsf{N}(0,1).$$

Example

Recall the n measurements of the daily number of hospital admissions due to asthma attacks, $X_1, X_2, \ldots, X_n$, which we will model using the Poisson distribution, *i.e.*

$$f(x;\lambda) = \mathrm{e}^{\lambda}\frac{\lambda^x}{x!}, \quad x \in \{0, 1, 2, ...\}$$

The parameter $\beta = 1/\lambda$ has the interpretation of being the average time (in days) between hospital admissions due to asthma attacks.

Find the approximate distribution of the maximum likelihood estimator of β.

Asymptotic optimality

In the case of smooth likelihood functions where asymptotic normality results can be derived it is possible to argue that, asymptotically, the maximum likelihood estimator is *optimal* or *best*.

Lower bound on variance

Let

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x; \theta), \quad \theta \in \Theta$$

and suppose that the maximum likelihood estimator $\widehat{\theta}_n$ is asymptotically normal; i.e.

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{\text{d}} \mathsf{N}(0, \{I_1(\theta)\}^{-1}).$$

Let $\tilde{\theta}_n = g(X_1, \ldots, X_n)$ be any other estimator of θ with

$$\mathbb{E}_\theta g(X_1, \ldots, X_n) = \mu_n(\theta)$$

Then,

$$\text{Var}_\theta(\tilde{\theta}_n) \geqslant \frac{(\mu_n'(\theta))^2}{nI_1(\theta)}$$

proof: Assume conditions 1 through 5 and denote the joint pdf of $X_1, \ldots, X_n$ as

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

and note that the log-likelihood can be written in terms of the joint pdf:

$$\ell_n'(\theta) = \sum_{i=1}^n \frac{\partial}{\partial\theta} \ln f(X_i; \theta) = \frac{\partial}{\partial\theta} \ln f(\mathbf{X}; \theta)$$

We have

$$\begin{aligned}
\text{Var}_\theta(\tilde{\theta}_n) I_n(\theta) &= \mathbb{E}_\theta(\tilde{\theta}_n - \mu_n)^2 \mathbb{E}_\theta \left(\ell_n'(\theta)\right)^2 \\
&\geqslant \left(\mathbb{E}_\theta(\tilde{\theta}_n - \mu_n) \frac{\partial}{\partial\theta} \ln f(\mathbf{X}; \theta)\right)^2 \quad \text{using } \mathbb{E}X^2\mathbb{E}Y^2 \geqslant (\mathbb{E}XY)^2 \\
&\geqslant \left(\int f(\mathbf{x}; \theta)(\tilde{\theta}_n - \mu_n) \frac{\frac{\partial}{\partial\theta} f(\mathbf{x}; \theta)}{f(\mathbf{x}; \theta)} \mathrm{d}\mathbf{x}\right)^2 \\
&\geqslant \left(\int \frac{\partial f}{\partial\theta}(\mathbf{x}; \theta)(\tilde{\theta}_n - \mu_n) \mathrm{d}\mathbf{x}\right)^2 \\
&\geqslant \left(\int \frac{\partial f}{\partial\theta}(\mathbf{x}; \theta)\, \tilde{\theta}_n \mathrm{d}\mathbf{x} - \mu_n \int \frac{\partial f}{\partial\theta}(\mathbf{x}; \theta) \mathrm{d}\mathbf{x}\right)^2 \\
\text{via assumption 5} \quad &\geqslant \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \underbrace{\mathbb{E}_\theta[\tilde{\theta}_n]}_{\mu_n} - \mu_n \underbrace{\frac{\mathrm{d}}{\mathrm{d}\theta} \int f(\mathbf{x}; \theta) \mathrm{d}\mathbf{x}}_{\frac{\mathrm{d}}{\mathrm{d}\theta}(1)}\right)^2 = (\mu_n'(\theta))^2
\end{aligned}$$

Therefore, rearranging the inequality

$$\mathrm{Var}_\theta(\tilde{\theta}_n) \geqslant \frac{(\mu_n'(\theta))^2}{I_n(\theta)}$$

□

As a corollary we have the following.

Cramer-Rao lower bound

If $\tilde{\theta}_n = g(X_1, \ldots, X_n)$ is an unbiased estimator of θ, then

$$\mathrm{Var}_\theta(\tilde{\theta}_n) \geqslant \frac{1}{nI_1(\theta)}$$

Since, for the MLE $\widehat{\theta}$ we have

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{\mathrm{d}} \mathsf{N}(0, I_1^{-1}(\theta)),$$

the maximum likelihood estimator can (under some technical conditions) achieve asymptotically (as $n \uparrow \infty$) the smallest possible variance in estimating θ.

Example

Recall the example:

$$X_1, \ldots, X_n \sim f, \quad \text{where} \quad f(x;\theta) = 2\theta x \mathrm{e}^{-\theta x^2}, \; x \geq 0; \; \theta > 0$$

We have shown previously that if $\widehat{\theta}$ is the maximum likelihood estimator of θ, then

$$\widehat{\theta} = n \Big/ \sum_i X_i^2 \quad \text{and} \quad \mathrm{Var}(\widehat{\theta}) \simeq \tfrac{1}{I_n(\theta)} = \theta^2/n$$

It is known that

$$\mathbb{E}(X) = \frac{1}{2}\sqrt{\frac{\pi}{\theta}} \quad \text{and} \quad \mathrm{Var}(X) = \frac{1}{\theta}\left(1 - \frac{\pi}{4}\right)$$

1. Find the method of moments estimator of θ, $\tilde{\theta}$.

2. Find an expression for the approximate distribution of $\bar{X}$.

3. Hence use the delta method to find the approximate distribution of $\tilde{\theta}$.

4. Compare the asymptotic properties of the estimators $\tilde{\theta}$ and $\widehat{\theta}$. Which is the better estimator?

solution: Rearranging $(\mathbb{E}X)^2 = \frac{1}{4}\frac{\pi}{\theta}$ we obtain the method of moments estimator:

$$\tilde{\theta} = \frac{\pi}{4(\bar{X})^2} = g(\bar{X}),$$

where $g(x) = \pi/(4x^2)$ with $g'(x) = -\frac{\pi}{2x^3}$ and $[g'(x)]^2 = \frac{\pi^2}{4x^6}$. We now derive its asymptotic distribution using the delta method. From the Central Limit Theorem

$$\sqrt{n}(\bar{X} - \sqrt{\pi/(4\theta)}) \xrightarrow{\text{d}} \mathsf{N}\left(0, \frac{1-\pi/4}{\theta}\right)$$

Hence,

$$\sqrt{n}(g(\bar{X}) - g(\sqrt{\pi/(4\theta)})) \xrightarrow{\text{d}} \mathsf{N}\Big(0, \underbrace{\text{Var}(X) \times \frac{4^2\theta^3}{\pi}}_{\theta^2(\frac{16}{\pi}-4)}\Big)$$

Now from the MLE asymptotic theory above we know that

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{\text{d}} \mathsf{N}(0, \theta^2)$$

Thus, we compare θ^2 versus $\theta^2(\frac{16}{\pi} - 4)$. Since

$$\frac{16}{\pi} - 4 = 1.09295817... > 1$$

we conclude that the MLE estimator beats the method of moments estimator, because the method of moments estimator has asymptotic variance roughly 9% higher than the variance of the MLE.

Likelihood-based Confidence Intervals

Because maximum likelihood estimators are asymptotically normal, we can construct confidence intervals easily.

Wald Confidence Intervals

Let $X_1, \ldots, X_n$ be random variables with common density function f, where

$$f(x) = f(x; \theta), \quad \theta \in \Theta$$

and let $\widehat{\theta}$ be the maximum likelihood estimator of θ. Under the regularity conditions for which θ is asymptotically normal

$$\left(\widehat{\theta}_n - z_{1-\alpha/2}\,\text{se}(\widehat{\theta})\,,\, \widehat{\theta} + z_{1-\alpha/2}\,\text{se}(\widehat{\theta})\right)$$

is an approximate $1 - \alpha$ confidence interval for θ for large n, where $\text{se}(\widehat{\theta}) = 1/\sqrt{I_n(\theta)}$.

Example

Recall the example of a sample $x_1, \ldots, x_n$ from the density function f, where

$$f(x;\theta) = 2\theta x \mathrm{e}^{-\theta x^2}, \; x \geq 0; \; \theta > 0.$$

1. Derive a formula for a 95% confidence interval for θ.

Recall that the maximum likelihood estimate and corresponding standard error are:

$$\widehat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i^2} \quad \text{and} \quad \widehat{\mathrm{se}}(\widehat{\theta}) = \frac{\widehat{\theta}}{\sqrt{n}} = \frac{\sqrt{n}}{\sum_{i=1}^{n} x_i^2}.$$

For a 95% confidence interval the appropriate $\mathsf{N}(0,1)$ quantile is

$$z_{0.975} = 1.96.$$

An approximate Wald 95% confidence interval for θ is then:

$$\left(\frac{n}{\sum_{i=1}^{n} x_i^2} - 1.96 \frac{\sqrt{n}}{\sum_{i=1}^{n} x_i^2}, \frac{n}{\sum_{i=1}^{n} x_i^2} + 1.96 \frac{\sqrt{n}}{\sum_{i=1}^{n} x_i^2} \right).$$

2. Use the following data to find an approximate 95% confidence interval for θ, assuming that it is a random sample from a r.v with the density function above.

```
0.366 0.568 0.300 0.115 0.204 0.128 0.277 0.391 0.328 0.451
0.412 0.190 0.207 0.147 0.116 0.326 0.256 0.524 0.217 0.485
0.265 0.375 0.267 0.360 0.250 0.258 0.583 0.413 0.481 0.468
0.406 0.336 0.305 0.321 0.268 0.361 0.632 0.283 0.258 0.466
0.276 0.232 0.133 0.316 0.468 0.496 0.573 0.523 0.256 0.491
0.127 0.054 0.440 0.228 0.249 0.754 0.430 0.111 0.459 0.233
0.257 0.640 0.147 0.273 0.112 0.389 0.126 0.356 0.273 0.296
0.433 0.253 0.234 0.514 0.177 0.221 0.534 0.509 0.510 0.269
0.262 0.625 0.183 0.541 0.705 0.078 0.847 0.149 0.031 0.453
0.299 0.226 0.069 0.211 0.195 0.381 0.317 0.467 0.289 0.593
```

For these data $(x_1, \ldots, x_{100})$ we have $\sum_{i=1}^{100} x_i^2 = 14.018$. So using the formula derived previously, an approximate 95% confidence interval for θ is

$$\left(\frac{100}{14.018} - 1.96 \frac{\sqrt{100}}{14.018}, \frac{100}{14.018} + 1.96 \frac{\sqrt{100}}{14.018} \right) = (5.17, 9.09).$$

Result

Under the same conditions as the previous result, with $\tau = g(\theta)$ and $\widehat{\tau} = g(\widehat{\theta})$,

$$\lim_{n\to\infty} \mathbb{P}\left(\widehat{\tau} - z_{1-\alpha/2}\,\text{se}(\widehat{\tau}) < \tau < \widehat{\tau} + z_{1-\alpha/2}\,\text{se}(\widehat{\tau})\right) = 1-\alpha$$

where $\text{se}(\widehat{\tau}) = |g'(\theta)|/\sqrt{I_n(\theta)}$. Therefore,

$$\left(\,\widehat{\tau} - z_{1-\alpha/2}\,\text{se}(\widehat{\tau})\,,\,\widehat{\tau} + z_{1-\alpha/2}\,\text{se}(\widehat{\tau})\,\right)$$

is an approximate $1-\alpha$ confidence interval for τ for large n.

This result is a confidence interval version of the delta method result.

Multi-parameter Maximum Likelihood Inference

In multi-parameter models such as

$$X_1,\ldots,X_n \sim \mathsf{N}(\mu,\sigma^2)$$

and

$$X_1,\ldots,X_n \sim \text{Gamma}(\alpha,\beta)$$

the maximum likelihood principle still applies. Instead of maximising over a single variable, the maximisation is performed simultaneously over several variables.

Example

Consider the model

$$X_1,\ldots,X_n \sim \mathsf{N}(\mu,\sigma^2), \quad -\infty < \mu < \infty,\ \sigma > 0.$$

Find the maximum likelihood estimators of μ and σ.

The log-likelihood function is

$$\ell(\mu,\sigma) = \ln \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-(x_i-\mu)^2/(2\sigma^2)}\right] = -\tfrac{1}{2\sigma^2}\textstyle\sum_i (x_i-\mu)^2 - \tfrac{n}{2}\ln(2\pi) - n\ln(\sigma).$$

Then,

$$\tfrac{\partial}{\partial\mu}\ell(\mu,\sigma) = \tfrac{1}{\sigma^2}\textstyle\sum_i (x_i-\mu) = \tfrac{n}{\sigma^2}(\bar{x}-\mu) = 0$$

if and only if

$$\mu = \bar{x}$$

and regardless of the value of σ. Also

$$\tfrac{\partial}{\partial\sigma}\ell(\mu,\sigma) = \sigma^{-3}\textstyle\sum_i (x_i-\mu)^2 - n\sigma^{-1} = 0$$

if and only if

$$\sigma = \sqrt{\frac{1}{n}\sum_i (x_i - \mu)^2}$$

The unique stationary point of $\ell(\mu, \sigma)$ is then

$$(\widehat{\mu}, \widehat{\sigma}) = \left(\bar{x}, \sqrt{\frac{1}{n}\sum_i (x_i - \bar{x})^2}\right).$$

Analysis of the second order partial derivatives can be used to show that this is the global maximiser of $\ell(\mu, \sigma)$ over $\mu \in \mathbb{R}$ and $\sigma > 0$. Hence, the maximum likelihood estimators of μ and σ are

$$\widehat{\mu} = \bar{X} \quad \text{and} \quad \widehat{\sigma} = \sqrt{\frac{1}{n}\sum_i (X_i - \bar{X})^2}.$$

In multi-parameter maximum likelihood estimation the extension of Fisher information is as follows:

Definition

Let $\theta = (\theta_1, \ldots, \theta_k)$ be the vector of parameters in a multi-parameter model. The **Fisher information matrix** is given by

$$I_n(\theta) = -\begin{bmatrix} \mathbb{E}(H_{11}) & \mathbb{E}(H_{12}) & \cdots & \mathbb{E}(H_{1k}) \\ \mathbb{E}(H_{21}) & \mathbb{E}(H_{22}) & \cdots & \mathbb{E}(H_{2k}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}(H_{k1}) & \mathbb{E}(H_{k2}) & \cdots & \mathbb{E}(H_{kk}) \end{bmatrix}$$

where

$$H_{ij} = \frac{\partial^2}{\partial\theta_i \partial\theta_j}\ell(\theta).$$

Example

Consider the model

$$X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2), \quad -\infty < \mu < \infty, \; \sigma > 0.$$

Find the Fisher Information matrix for μ and σ.

It was shown previously that the first order partial derivatives are

$$\begin{aligned} \tfrac{\partial}{\partial\mu}\ell(\mu, \sigma) &= \tfrac{n}{\sigma^2}(\bar{x} - \mu) \\ \tfrac{\partial}{\partial\sigma}\ell(\mu, \sigma) &= \sigma^{-3}\sum_i (x_i - \mu)^2 - n\sigma^{-1} \end{aligned}$$

The second order partial derivatives are then

$$\begin{aligned} \frac{\partial^2}{\partial \mu^2}\ell(\mu,\sigma) &= \frac{-n}{\sigma^2} \\ \frac{\partial^2}{\partial \mu \partial \sigma}\ell(\mu,\sigma) &= \frac{-2n}{\sigma^3}(\bar{x}-\mu) \\ \text{and} \quad \frac{\partial^2}{\partial \sigma^2}\ell(\mu,\sigma) &= -3\sigma^{-4}\sum_i (x_i-\mu)^2 + n\sigma^{-2}. \end{aligned}$$

Noting that $\mathbb{E}(X_i)=\mu$ and $\mathbb{E}(X_i-\mu)^2=\sigma^2$ for each i we then get

$$\begin{aligned} \mathbb{E}\left[\frac{\partial^2}{\partial \mu^2}\ell(\mu,\sigma)\right] &= \frac{-n}{\sigma^2,} \\ \mathbb{E}\left[\frac{\partial^2}{\partial \mu \partial \sigma}\ell(\mu,\sigma)\right] &= 0 \\ \text{and} \quad \mathbb{E}\left[\frac{\partial^2}{\partial \sigma^2}\ell(\mu,\sigma)\right] &= -3\sigma^{-4}n\sigma^2 + n\sigma^{-2} = -2n\sigma^{-2}. \end{aligned}$$

The Fisher information matrix is then

$$I_n(\mu,\sigma) = -\begin{bmatrix} -n/\sigma^2 & 0 \\ 0 & -2n\sigma^{-2} \end{bmatrix} = \begin{bmatrix} n/\sigma^2 & 0 \\ 0 & 2n/\sigma^2 \end{bmatrix}$$

Given the Fisher Information matrix, we can find the asymptotic distribution of any function of a vector of maximum likelihood estimators $\widehat{\tau}=g(\widehat{\theta})$. But first we need to define the *gradient vector*, as below.

Definition

Let $\theta=(\theta_1,\ldots,\theta_k)$ be the vector of parameters in a multi-parameter model and $g(\theta)=g(\theta_1,\ldots,\theta_k)$ be a real-valued function. The **gradient vector** of g is given by

$$\nabla g(\theta) = \begin{bmatrix} \frac{\partial g(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial g(\theta)}{\partial \theta_k} \end{bmatrix}$$

Example

Find $\nabla g(\mu,\sigma)$ where $g(\mu,\sigma)=\frac{\mu}{\sigma}$.

$$\nabla g(\mu,\sigma) = \begin{bmatrix} \frac{1}{\sigma} \\ -\frac{\mu}{\sigma^2} \end{bmatrix}$$

Result

Let $\tau = g(\theta)$ be a real-valued function of $\theta = (\theta_1, \ldots, \theta_k)$, with maximum likelihood estimate $\widehat{\theta}$ and $\widehat{\tau} = g(\widehat{\theta})$. Under appropriate regularity conditions, including the existence of all second order partial derivatives of $\mathcal{L}(\theta)$ and first order partial derivatives of g, as $n \to \infty$

$$\frac{\widehat{\tau} - \tau}{\mathrm{se}(\widehat{\tau})} \xrightarrow{\mathrm{d}} \mathsf{N}(0,1)$$

where

$$\mathrm{se}(\widehat{\tau}) = \sqrt{\nabla g(\theta)^\top I_n(\theta)^{-1} \nabla g(\theta)}.$$

This is a multi-parameter extension of the delta method result given on page 213.

Example

Let

$$X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2)$$

Derive a convergence result that gives us the approximate distribution of $\widehat{\tau} = g(\widehat{\mu}, \widehat{\sigma}) = \widehat{\mu}/\widehat{\sigma}$.

Given the invariance property of maximum likelihood estimators, the maximum likelihood estimator of τ is

$$\widehat{\tau} = \frac{\widehat{\mu}}{\widehat{\sigma}} = \frac{\bar{X}}{\sqrt{\frac{1}{n}\sum_i (X_i - \bar{X})^2}}.$$

The Fisher information matrix may be shown to be

$$I_n(\mu, \sigma) = \begin{bmatrix} n/\sigma^2 & 0 \\ 0 & 2n/\sigma^2 \end{bmatrix}$$

which has inverse

$$I_n(\mu, \sigma)^{-1} = \frac{1}{n}\begin{bmatrix} \sigma^2 & 0 \\ 0 & \frac{1}{2}\sigma^2 \end{bmatrix}$$

The gradient vector is

$$\nabla g(\mu, \sigma) = \begin{bmatrix} 1/\sigma \\ -\mu/\sigma^2 \end{bmatrix}.$$

Since

$$\begin{aligned} \nabla g(\mu, \sigma)^\top I_n(\mu, \sigma)^{-1} \nabla g(\mu, \sigma) &= \frac{1}{n}[1/\sigma \quad -\mu/\sigma^2] \begin{bmatrix} \sigma^2 & 0 \\ 0 & \frac{1}{2}\sigma^2 \end{bmatrix} \begin{bmatrix} 1/\sigma \\ -\mu/\sigma^2 \end{bmatrix} \\ &= \frac{1}{n}\left(1 + \frac{\mu^2}{2\sigma^2}\right) \end{aligned}$$

the approximate standard error of $\widehat{\tau}$ is

$$\widehat{\text{se}}(\widehat{\tau}) = \sqrt{\frac{1}{n}\left(1 + \frac{\bar{x}^2}{\frac{2}{n}\sum_i (x_i - \bar{x})^2}\right)}.$$

The asymptotic normality result for $\widehat{\tau}$ is then

$$\frac{\widehat{\tau} - \tau}{\sqrt{\frac{1}{n}\left(1 + \frac{\bar{X}^2}{\frac{2}{n}\sum_i (X_i - \bar{X})^2}\right)}} \xrightarrow{\text{d}} \mathsf{N}(0,1).$$

We can use the asymptotic normality result previously stated to obtain an approximate confidence interval for

$$\tau = g(\theta) = g(\theta_1, \ldots, \theta_k).$$

Result

Under appropriate regularity conditions, including the existence of all second order partial derivatives of $\mathcal{L}(\theta)$, if $\tau = g(\theta)$ where each component of g is differentiable then:

$$\lim_{n\to\infty} \mathbb{P}\left(\widehat{\tau} - z_{1-\alpha/2}\,\text{se}(\widehat{\tau}) < \tau < \widehat{\tau} + z_{1-\alpha/2}\,\text{se}(\widehat{\tau})\right) = 1 - \alpha$$

where

$$\text{se}(\widehat{\tau}) = \sqrt{\nabla g(\theta)^\top I_n(\theta)^{-1} \nabla g(\theta)}.$$

Therefore,

$$\left(\widehat{\tau} - z_{1-\alpha/2}\,\text{se}(\widehat{\tau}_n)\,,\, \widehat{\tau} + z_{1-\alpha/2}\,\text{se}(\widehat{\tau})\right)$$

is an approximate $1 - \alpha$ confidence interval for τ for large n.

Example

Let

$$X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2)$$

Derive a formula for an approximate 99% confidence interval for $\tau = g(\mu, \sigma) = \mu/\sigma$.

As shown previously, the maximum likelihood estimator for τ is

$$\widehat{\tau} = \frac{\bar{X}}{\sqrt{\frac{1}{n}\sum_i (X_i - \bar{X})^2}}$$

and the approximate standard error of $\widehat{\tau}$ is

$$\widehat{\text{se}}(\widehat{\tau}) = \sqrt{\nabla g(\widehat{\theta})^\top I_n(\widehat{\theta})^{-1} \nabla g(\widehat{\theta})} = \sqrt{\frac{1}{n}\left(1 + \frac{\bar{x}^2}{\frac{2}{n}\sum_i (x_i - \bar{x})^2}\right)}.$$

The appropriate quantile from the $\mathsf{N}(0,1)$ distribution is $z_{0.995} = 2.576$.

An approximate 99% confidence interval for μ/σ is then

$$\left(\frac{\bar{X}}{\sqrt{\frac{1}{n}\sum_i (X_i - \bar{X})^2}} - 2.576\sqrt{\frac{1}{n}\left(1 + \frac{\bar{X}^2}{\frac{2}{n}\sum_i (X_i - \bar{X})^2}\right)},\right.$$
$$\left.\frac{\bar{X}}{\sqrt{\frac{1}{n}\sum_i (X_i - \bar{X})^2}} + 2.576\sqrt{\frac{1}{n}\left(1 + \frac{\bar{X}^2}{\frac{2}{n}\sum_i (X_i - \bar{X})^2}\right)}\right).$$

$\lrcorner$

Chapter 9

Hypothesis Testing

Hypothesis testing is the most commonly used formal statistical vehicle for making decisions given some data.

For further reading, consider Hogg *et al* (2005) sections 5.5, 5.6 and 6.3 (likelihood ratio and Wald-type tests only), or Rice (2007) sections 9.1-9.4 and 11.2.1. You may also read through Section 5.3 in Kroese & Chan (2014).

There are many examples of situations in which we are interested in testing a specific hypothesis using sample data. In fact, most science is done using this approach! (And indeed other forms of quantitative research.)

Below are some examples that we will consider in this chapter.

Example

Recall that the Mythbusters were testing whether or not toast lands butter side down more often than butter side up.

In 24 trials, they found that 14 slices of bread landed butter side down.

Is this evidence that toast lands butter-side down more often than butter side up?

Example

Before the installation of new machinery at a chemical plant, the daily yield of fertilizer produced at the plant had a mean $\mu = 880$ tonnes. Some new machinery was installed, and we would like to know if the new machinery is more efficent (*i.e.* if $\mu > 880$).

During the first $n = 50$ days of operation of the new machinery, the yield of fertilizer was recorded and the sample mean obtained was $\bar{x} = 888$ with a standard deviation $s = 21$.

Is there evidence that the new machinery is more efficient?

Example

(Ecology 2005, 86:1057-1060)

Do ravens intentionally fly towards gunshot sounds (to scavenge on the carcass they expect to find)? Crow White addressed this question by going to 12 locations, firing a gun, then counting raven numbers 10 minutes later. He repeated the process at 12 different locations where he didn't fire a gun. Results:

no gunshot	0	0	2	3	5	0	0	0	1	0	1	0
gunshot	2	1	4	1	0	5	0	1	0	3	5	2

Is there evidence that ravens fly towards the location of gunshots?

Stating the hypotheses

A key first step in hypothesis testing is identifying the hypothesis that we would like to test (the null hypothesis), and the alternative hypothesis of interest to us, defined as below.

Definitions

The **null hypothesis**, labelled H_0, is a claim that a parameter of interest to us (θ) takes a particular value (θ_0). Hence H_0 has the form $\theta = \theta_0$ for some pre-specified value θ_0.

The **alternative hypothesis**, labelled H_1, is a more general hypothesis about the parameter of interest to us, which we will accept to be true if the evidence against the null hypothesis is strong enough. The form of H_1 tends to be one of the following:

$$H_1 : \theta \neq \theta_0$$
$$H_1 : \theta > \theta_0$$
$$H_1 : \theta < \theta_0$$

In a hypothesis test, we use our data to test H_0, by measuring how much evidence our data offer against H_0 in favour of H_1.

In later statistics courses you will meet hypothesis tests concerning hypotheses that have a more general form than the above. H_0 does not actually need to have the form $H_0 : \theta = \theta_0$ but instead can just specify some restriction on the range of possible values for θ, *i.e.* $H_0 : \theta \in \omega_0$ for some set $\omega_0 \in \Omega$ where Ω is the parameter space. It is typical for the alternative hypothesis to be more general than the null, *e.g.* $H_1 : \theta \in \overline{\omega_0}$.

Example

Consider the Mythbusters example on page 225.

State the null and alternative hypotheses.

Here we have a sample from a binomial distribution, where the binomial parameter p is the probability that a slice of toast will land butter side down. We are interested in whether there is evidence that the parameter p is larger than 0.5.

$$H_0 : p = 0.5 \qquad H_1 : p > 0.5$$

Example

Consider the machinery example on page 226.

State the null and alternative hypotheses.

In all of the examples of the previous section, there is one hypothesis of primary interest, which involves specifying a **specific value** for a parameter of interest to us.

Can you find the null and alternative hypotheses for all the above situations?

The process of hypothesis testing

A hypothesis test has the following steps.

1. State the null hypothesis (H_0) and the alternative hypothesis (H_1). By convention, the null hypothesis is the more specific of the two hypotheses.

2. Then we use our data to answer the question:

   "How much evidence is there against the null hypothesis?"

   A common way to achieve this is to:

   i) Find a test statistic that measures how "far" our data are from what is expected under the null hypothesis. (You must know the approximate distribution of this test statistic assuming H_0 to be true. This is called the *null distribution.*)

   ii) Calculate a P-value, a probability that measures how much evidence there is against the null hypothesis, for the data we observed. A P-value is defined as the probability of observing a test statistic value as or more unusual than the one we observed, if the null hypothesis were true.

3. Reach a conclusion. A helpful way to think about the conclusion is to return to our original question:

   "How much evidence is there against H_0?"

Example

The Mythbusters example can be used to illustrate the above steps of a hypothesis test.

1. Let

$$p = \mathbb{P}(\text{Toast lands butter side down})$$

Then what we want to do is choose between the following two hypotheses:

$$H_0\colon\ p = \tfrac{1}{2} \qquad \text{versus} \qquad H_1\colon\ p > \tfrac{1}{2}$$

2. We want to answer the question "How much evidence (if any) does our sample (14 of 24 land butter side down) give us against the claim that $p = 0.5$?"

   (a) To answer this question, we will consider $\widehat{p}$, the sample proportion, and in particular we will look at the test statistic

   $$Z = \frac{\widehat{p} - p}{\sqrt{p(1-p)/n}} \xrightarrow{\text{d}} \mathsf{N}(0,1)$$

   Under the null hypothesis,

   $$Z = \frac{\widehat{p} - 0.5}{\sqrt{0.5(1-0.5)/n}} \xrightarrow{\text{d}} \mathsf{N}(0,1)$$

   This statistic measures how far our sample data are with H_0. The further $\widehat{p}$ is from 0.5, the further Z is from 0.

   (b) To find out if $\widehat{p} = \frac{14}{24}$ is unusually large, if $p = 0.5$, we can calculate

   $$\mathbb{P}\left(\widehat{p} \geqslant \frac{14}{24}\right) \simeq \mathbb{P}\left(Z \geqslant \frac{\frac{14}{24} - 0.5}{\sqrt{0.5(1-0.5)/n}}\right) \simeq \mathbb{P}(Z > 0.82) \simeq 0.2071$$

3. So we can say that we would expect p to be at least as large as $\frac{14}{24}$ quite often (22% of the time) due to sample variation alone. Observing an event of probability 0.22 is not particularly surprising, so we conclude that we have no evidence against the claim that $p = 0.5$ – because our data are consistent with this hypothesis.

Interpreting *P*-values

The most common way that a hypothesis test is conducted and reported is via the calculation of a P-value.

Example

When reading scientific research, you will often see comments such as the following:

- "Paired t-tests indicated that there was no change in attitude ($P > 0.05$)"
- "The Yahoo! search activity associated with specific cancers correlated with their estimated incidence (Spearman rank correlation = 0.50, $P = 0.015$)"

- "There was no significant difference ($p > 0.05$) between the first and second responses"

□

So what is a P-value? And how do you interpret a P-value once you've calculated it?

Definition
The P-value of an observed test statistic is

$$\begin{aligned} P\text{-value} &= \mathbb{P}(\text{observing a test statistic as or more ``extreme''} \\ &\qquad \text{than the observed test statistic when } H_0 \text{ is true.}) \end{aligned}$$

The following are some rough guidelines on interpreting P-values. Note though that the interpretation of P-values should depend on the context to some extent, so the below should be considered as a guide only and not as strict rules.

Range of P-value	Conclusion
P-value≥ 0.1	little or no evidence against H_0
$0.01 \leq P$-value < 0.1	some, but inconclusive evidence against H_0
$0.001 \leq P$-value < 0.01	evidence against H_0
P-value < 0.001	strong evidence against H_0

It is common for people to use 0.05 as a "cut-off" between a significant finding ($P < 0.05$) and a non-significant finding ($P > 0.05$) – hence the interpretations in the example quotes on the previous page. Nevertheless, it is helpful to keep in mind that P is continuous and our interpretation of P should reflect this – a P-value of 0.049 (just less than 0.05) is hardly different from a P-value of 0.051 (just larger than 0.05), so there should be little difference in interpretation of these values. In contrast, a P-value of 0.049 offers less evidence against H_0 than a P-value of 0.0001!

Example

Consider again the Mythbusters example.

We calculated that the P-value was about 0.22. (This P-value measured how often you would get further from 0.5 than $\widehat{p} = \frac{14}{24}$.)

Draw a conclusion from this hypothesis test.

This P-value is large – it is in the "little or no evidence against H_0" range. Hence we conclude that there is little or no evidence against the claim that toast lands butter side down just as often as it lands butter side up.

Tests for Normal Samples

In the situation where

$$X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2)$$

it is possible to perform exact hypothesis tests for the parameter μ using the main result from Chapter 6:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \mathsf{t}_{n-1}$$

For testing the hypotheses

$$H_0\colon\ \mu = \mu_0 \qquad \text{versus any of} \qquad \begin{cases} H_1 : \mu < \mu_0 \\ H_1 : \mu \neq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

the appropriate test statistic is

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

which has a t_{n-1} distribution under the null hypothesis. Using this test statistic in a hypothesis test about μ is known as a **t-test** (or a one-sample t-test).

Example

Before the installation of new machinery, the daily yield of fertilizer produced by a chemical plant had a mean $\mu = 880$ tonnes. Some new machinery was installed, and we would like to know if the new machinery is more effcient (*i.e.* if $\mu > 880$).

During the first $n = 50$ days of operation of the new machinery, the yield of fertilizer was recorded. The sample mean was $\bar{x} = 888$ with a standard deviation $s = 21$.

Is there evidence that the new machinery is more efficient? Use a hypothesis test to answer this question, assuming that yield is approximately normal.

Using the hypothesis testing steps given previously:

1. Our null and alternative hypotheses are:

$$H_0 : \mu = 880 \qquad H_1 : \mu > 880$$

2. We estimate μ using the sample mean $\bar{X}$. Now we want to know if our $\bar{X}$ is so far from $\mu = 880$ that it provides evidence against H_0.

   i) We can construct a test statistic based on $\bar{X}$ using

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

   which under the null hypothesis becomes

$$T = \frac{\bar{X} - 880}{S/\sqrt{n}} \sim t_{n-1}$$

   Larger observed values of T suggest more evidence against H_0.

   ii) We now would like to calculate a P-value that measures how unusual it is to get a mean as large as $\bar{x} = 888$, if the true mean were $\mu = 880$:

$$\mathbb{P}\left(T > \frac{\bar{x} - 880}{s/\sqrt{n}}\right) = \mathbb{P}\left(T > \frac{888 - 880}{21/\sqrt{50}}\right) = \mathbb{P}(T > 2.69)$$

   where $T \sim t_{49}$. From tables, $0.0025 < P\text{-value} < 0.005$.

3. This tells us that, if H_0 were true, we would be highly unlikely to observe a T statistic as large as 2.7 (hence a mean yield as high as 888 tonnes) by chance alone. We have strong evidence against the claim that $\mu = 880$.

   Alternatively, you could say that "the mean is significantly different from 880".

One-sided and two-sided tests

The "chemical plant" hypothesis test on page 226 is an example of a *one-sided hypothesis test*, because we are interested in whether $\mu > 880$, *i.e.* we are interested in alternative values of μ on one side of our hypothesised value, $H_0 : \mu = 880$. But there are many situations where we are interested in finding evidence that a parameter lies either side of a hypothesised value. This involves a *two-sided test*.

Definition

A *one-sided* hypothesis test about a parameter θ is either of the form:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0$$

or

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0.$$

A *two-sided* hypothesis test about θ is of the form

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

A two-sided test is sometimes called a two-tailed test, because we calculate the P-value using both tails of the null distribution of the test statistic (instead of just using one tail as in a one-sided test).

Example

A popular brand of yoghurt claims to contain 120 calories per serving. A consumer watchdog group randomly sampled 14 servings of the yoghurt and obtained the following numbers of calories per serving:

160	200	220	230	120	180	140
130	170	190	80	120	100	170

Conduct a hypothesis test to test if the manufacturers' claim is correct.

The hypotheses to be tested are:

H_0: $\mu = 120$ (mean number of calories matches manufacturer's claim)

versus

H_1: $\mu \neq 120$ (mean number of calories differs from manufacturer's claim)

where

$$\mu = \text{mean calorie value of a serving of the yoghurt.}$$

The test statistic we will use is

$$\frac{\bar{X} - 120}{S/\sqrt{14}}$$

which has a t_{13} distribution if the null hypothesis is true. The observed value is

$$\frac{\bar{x} - 120}{s/\sqrt{14}} = \frac{157.8571 - 120}{44.75206/\sqrt{14}} = 3.1651...$$

Then, using t-distribution tables,

$$\begin{aligned} P\text{-value} &= \mathbb{P}_{\mu=120}\left(\left|\frac{\bar{X} - 120}{S/\sqrt{14}}\right| > 3.1651\right) \\ &= 2\,\mathbb{P}(T > 3.1651), \quad T \sim t_{13} \\ &\approx 0.007 \end{aligned}$$

so there is very strong evidence against H_0. The consumer watchdog group should conclude that the manufacturer's calorie claim is not correct.

Rejection regions

Instead of using a P-value to draw a conclusion based on data, an alternative approach that is sometimes used is to base the test on a rejection region (as defined below):

Definition
The **rejection region** is the set of values of the test statistic for which H_0 is rejected in favour of H_1.

The term "rejection region" comes from the fact that often people speak of "rejecting H_0" (if our data provide evidence against H_0) or "retaining H_0" (if our data do not provide evidence against H_0). If out test statistic is in the rejection region we reject H_0, if the test statistic is not in the rejection region we retain H_0.

To determine a rejection region, we first choose a size or *significance level* for the test, which can be defined as the P-value at which we would start rejecting H_0. It should be set to a small number (typically 0.05), and by convention is usually denoted by α.

Once we have determined the desired size of the test, we can then derive the rejection region, as in the below examples.

Example

Recall the fertilizer example – we wanted to test

$$H_0 : \mu = 880 \quad versus \quad H_1 : \mu > 880$$

Our test statistic is

$$T = \frac{\bar{X} - 880}{S/\sqrt{n}} \sim \mathsf{t}_{n-1}$$

and we have a sample of size 50.

1. Find a rejection region for a test of size 0.05
2. Hence test H_0 versus H_1.

1. From tables, if $T \sim t_{49}$, then

   $$\mathbb{P}(T > 1.676) \approx 0.05$$

   and so our rejection region is $T > 1.676$.

   In other words, if our observed value of T is greater than 1.676, we will reject H_0 in favour of H_1. Alternatively, if $T < 1.676$, we retain H_0.

2. Our observed value of T was 2.69 which is in our rejection region $\Longrightarrow$ we reject H_0 and conclude that there is evidence that $\mu > 880$.

Example

Recall the yoghurt example – we wanted to test

$$H_0 : \mu = 120 \quad versus \quad H_1 : \mu \neq 120$$

Our test statistic is

$$T = \frac{\bar{X} - 120}{S/\sqrt{n}} \sim t_{n-1}$$

and we have a sample of size $n = 14$.

Find a rejection region for a test of size 0.05.

We now have two different approaches. One approach involves computing the P-value based on the observed test statistic, and deciding whether or not it is small enough to reject H_0. The other approach is based on setting the significance level α of the test to some number (like 0.05), then working out the range of values of our test statistic for which H_0 should be rejected at this significance level.

We will mainly use the P-value approach because it is more informative, and because it is more commonly used in practice. The significance level approach is however useful in determining important properties of tests, such as their Type I and Type II error.

Type I and Type II error

How could we choose a significance level α? This problem can be answered by considering the possible errors that can made in reaching our decision. These can be categorised into two types of errors, called **Type I error** and **Type II error**.

Definition

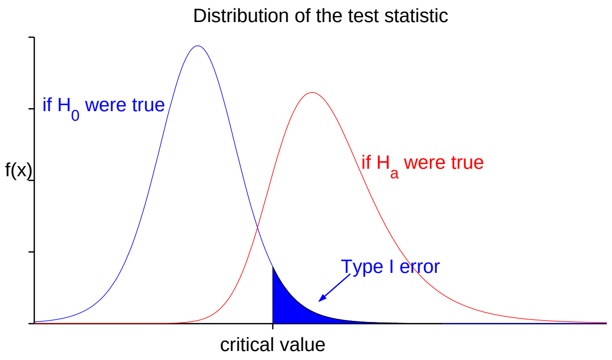
Type I error corresponds to rejection of the null hypothesis when it is really true.

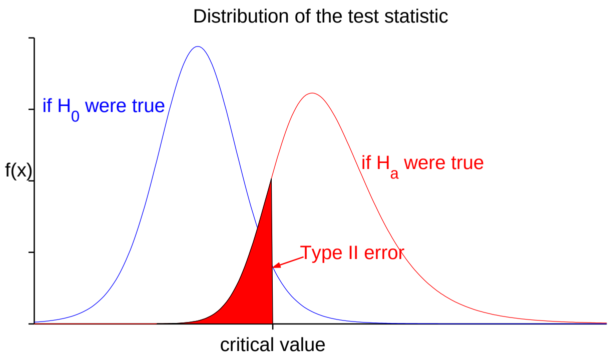
Type II error corresponds to acceptance of the null hypothesis when it is really false.

	reject H_0	accept H_0
H_0 true	Type I error	No error
H_0 false	No error	Type II error

Example

For the "chemical plant" example of page 226, a Type I error would correspond to concluding that the machinery has a positive effect on yield, when in actual fact it doesn't. Type II error corresponds to concluding that the machinery does not have a positive effect on yield, when in actual fact it does.





Clearly we would like to avoid both Type I and Type II errors, however there is always a chance that either one will occur so the idea is to reduce these chances as much as possible. Unfortunately, making the probability of one type of error small has the effect of making the probability of the other large.

In most practical situations, we can readily control Type I error, but Type II error is more difficult to get a handle on.

Definition

The *size* or *significance level* of a test is the probability of committing a Type I error. It is usually denoted by α. Therefore,

$$\begin{aligned} \alpha &= \text{size} \\ &= \text{significance level} \\ &= \mathbb{P}(\text{committing Type I error}) \\ &= \mathbb{P}(\text{reject } H_0 \text{ when } H_0 \text{ is true}) \end{aligned}$$

Type II error is usually quantified through a concept called *power*, which will be discussed in detail in a later section.

A popular choice of α is $\alpha = 0.05$. This corresponds to the following situation:

"We have set up our test in such a way that if we do reject H_0 then there is only a 5% chance that we will wrongfully do so."

There is nothing special about $\alpha = 0.05$. Sometimes it might be better to have $\alpha = 0.1$ or $\alpha = 0.01$, depending on the application.

Example

If we want to test the following hypotheses about a new vaccine:

H_0: vaccine is perfectly safe.

versus

H_1: vaccine has harmful side effects

then it is important to minimise Type II error – we want to detect any harmful side effects that are present. To assist in minimising Type II error, we might be prepared to accept a reasonably high Type I error (such as 0.1 or maybe even 0.25).

Example

Suppose we are testing a potential water source for toxic levels of heavy metals. If the hypotheses are

H_0: the water has toxic levels of heavy metals

versus

H_1: the water is OK to drink

then it is important to minimise type I error – we won't want to make the mistake of saying that toxic water is OK to drink! So we would want to choose a low Type I error, maybe 0.001 say.

Power of a Statistical Test

It is important for a particular hypothesis test procedure to have a small significance level. However, it is also important to realise that this is not the only property that a good hypothesis test should have.

Definition
Broadly speaking, a test procedure with good power properties, usually referred to as a "powerful test", is one that has a good chance of rejecting H_0 when it is not true.

Therefore, a test with high power is able to detect deviation from the null hypothesis with high probability. The formal definition of the power of a test about μ is given by:

Definition
Consider a test about a mean μ. Then

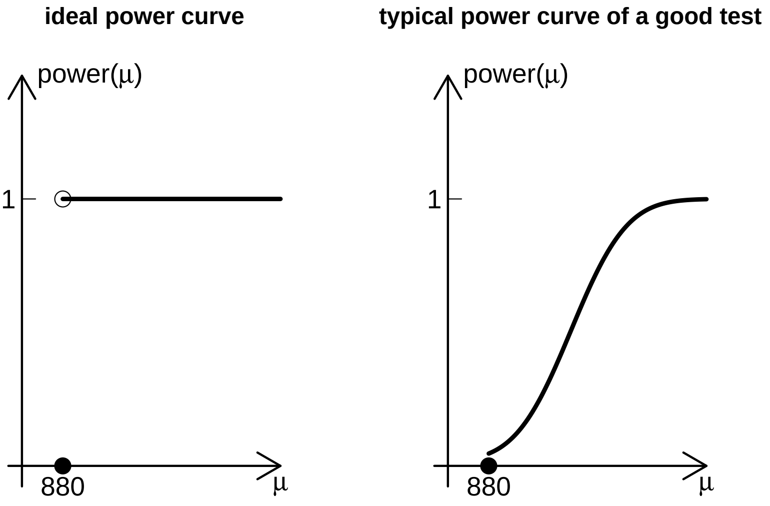
$$\begin{aligned}\text{power}(\mu) &= \mathbb{P}(\text{reject } H_0 \text{ when the true value is } \mu) \\ &= \mathbb{P}_\mu(\text{reject } H_0)\end{aligned}$$

Therefore the power of a test is really a function, or *curve*, that depends on the true value of μ. It is easy to check that

$$\text{power(value of } \mu \text{ under } H_0) = \mathbb{P}(\text{Type I error}) = \alpha$$

But for all other values of μ we want power(μ) to be as close to 1 as possible.

The following graph shows the 'ideal' power curve as well as a typical actual power curve.



ideal power curve

typical power curve of a good test

Also, note that power and Type II error are inversely related:

Result

$$\mathbb{P}(\text{Type II error}) = 1 - \text{power}(\mu)$$

Therefore, high power corresponds to low chance of Type II error.

Power is difficult to calculate by hand unless the test statistic has a normal distribution – hence we will focus on one-sample tests of μ, and treat σ as being known, such that we can use the test statistic

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

Example

Consider again the chemical plant example, where we are testing

$$H_0 : \mu = 880 \quad \text{against} \quad H_1 : \mu > 880$$

using the yield from 50 days of sampling.

Let us suppose that the true value of μ is $\mu = 882$ so H_0 is false (*i.e.* the new machine *has* increased productivity – by 2 tonnes per day).

You may assume that σ is 21 in the following.

What is the (approximate) power of our test when $\mu = 882$ for this test?

Using the test statistic

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

we reject the null hypothesis whenever $\mathbb{P}(Z > 1.645)$.

This corresponds to a sample mean of

$$\bar{X} = \mu + 1.645 \cdot \frac{\sigma}{\sqrt{n}} \simeq 884.9$$

So our rejection region is $\bar{X} > 884.9$.

If the mean is actually 882, what is the chance that our test rejects H_0?

$$\begin{aligned}
\text{power}(882) &= \mathbb{P}(\text{reject } H_0 \text{ if } \mu = 882) \\
&= \mathbb{P}_{\mu=882}\left(\frac{\bar{X} - 882}{21/\sqrt{50}} > \frac{884.9 - 882}{21/\sqrt{50}}\right) \\
&= \mathbb{P}(Z > 0.9764) \\
&= 0.164
\end{aligned}$$

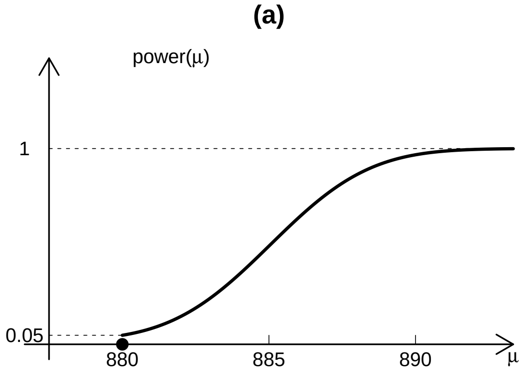
That is, if mean production really was 882 then our test has only a 16% chance of detecting this change and rejecting H_0.

The power seems low in the above scenario – this is because relative to sample variation ($\sigma = 21$), an increase in yield of 2 is not particularly large. If, on the other hand, yield had increased by 6 tonnes to 886, the power could be shown to be 0.64 in this case. And if yield had increased by 10 tonnes to 890, power would be 95%.

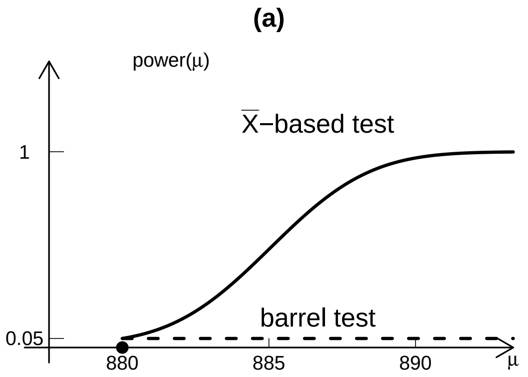
We can do such calculations for general $\mu > 880$ and arrive at a *function* of power values:

$$\begin{aligned}
\text{power}(\mu) &= \mathbb{P}_\mu(\bar{X} > 884.9) \\
&= \mathbb{P}_\mu\left(\frac{\bar{X} - \mu}{21/\sqrt{50}} > \frac{884.9 - \mu}{21/\sqrt{50}}\right) \\
&= \mathbb{P}\left(Z > \frac{884.9 - \mu}{21/\sqrt{50}}\right) \\
&= \mathbb{P}\left(Z < \frac{\mu - 884.9}{21/\sqrt{50}}\right) \\
&= \Phi\left(\frac{\mu - 884.9}{21/\sqrt{50}}\right)
\end{aligned}$$

This leads to the *power curve* shown in the following figure:



It is interesting to graphically compare the power curves from the $\bar{X}$ test and the barrel test for the “chemical plant” example:



Here we see that the $\bar{X}$ test has higher power than the Barrel test for all values of μ, so is clearly superior.

Further notes on power are:

- Power depends on n, σ, and the value of μ under H_1. Power can be increased by increasing the size of the sample, n. If the alternative hypotheses are of the form
$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0$$
then, for a good test, power is larger for higher values of μ and smaller σ.

- In the planning stages of an experiment a question that is often asked is "Does the proposed test have a reasonable chance of detecting a change in population mean?". To answer this question one needs to know how big a change in population mean (μ) that we want to be able to detect, sometimes referred to as a "least significant difference" (LSD). One would also need to know (or have a reasonable estimate of) the standard deviation of data, σ. Then one can decide what value of n is necessary to have high probability of detecting a LSD. This is the most common applied usage of power.

- Power may be used to compare two or more tests for a given significance level. For example, if we had two competing tests with significance level $\alpha = 0.05$ then we would want to use the test that has the higher power. The previous figure shows that the $\bar{X}$ test is clearly superior to the Barrel test.

Some more tests

Hypothesis testing is a tool that can be applied in a range of settings. If the null hypothesis is that some parameter equals a specific value (*e.g.* $p = 0.5$, $\mu = 880$, $\theta = 1$, ...) then all we need to construct a hypothesis test is an estimator of the parameter of interest and an approximate distribution for the estimator, under the null hypothesis.

We have focussed in previous sections on one-sample tests of the mean, *e.g.* the chemical plant and hairdresser examples. In this section we consider some other commonly encountered hypothesis testing situations.

Two-sample test of means

Consider the situation where we have two independent normal random samples:

$$X_1, \ldots, X_{n_X} \sim \mathsf{N}(\mu_X, \sigma^2) \quad \text{and} \quad Y_1, \ldots, Y_{n_Y} \sim \mathsf{N}(\mu_Y, \sigma^2)$$

A hypothesis test that is typically of interest is

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad H_1 : \mu_X \neq \mu_Y$$

An appropriate test statistic is

$$\frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t_{n_X + n_Y - 2} \quad \text{under } H_0$$

where as on page 189,

$$S_p^2 = \frac{1}{n_X + n_Y - 2} \left\{ (n_X - 1) S_X^2 + (n_Y - 1) S_Y^2 \right\}$$

This statistic assumes that the two samples have equal variance – as in the two-sample confidence interval method of page 189.

Example

Do MATH2801 and MATH2901 students spend the same amount at the hairdressers, on average?

A class survey obtained the following results:

$$\begin{array}{lll} n_{2801} = 38 & \bar{x}_{2801} = 28.53 & s_{2801} = 32.89 \\ n_{2901} = 31 & \bar{x}_{2901} = 17.03 & s_{2901} = 19.07 \end{array}$$

Carry out the hypothesis test:

$$H_0 : \mu_{2801} = \mu_{2901} \qquad \text{versus} \qquad H_1 : \mu_{2801} \neq \mu_{2901}$$

Here $s_p^2 = 605.92$ and

$$t = \frac{28.53 - 17.03}{s_p\sqrt{\frac{1}{38} + \frac{1}{31}}} \approx \frac{28.53 - 17.03}{24.61\sqrt{\frac{1}{38} + \frac{1}{31}}} \approx 1.93$$

Hence, the p-value is

$$p = \mathbb{P}(T > |t| \text{ or } T < -|t|) = 2\mathbb{P}(T > 1.93) \approx 0.057$$

where $T \sim \mathsf{t}_{68}$. Therefore, we reject the H_0 hypothesis at the 5% level of significance. There is evidence against the hypothesis that MATH2801 and MATH2901 students spend the same amount on hairdressers.

Wald Tests

So far we have only considered the special cases of normal or binomial data. How about the general situation:

$$X_1, \ldots, X_n \sim f, \quad \text{where} \quad f(x) = f(x;\theta) \ \ ?$$

When the sample size is large the **Wald test** procedure often provides a satisfactory solution:

The Wald Test

Consider the hypotheses

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

and let $\widehat{\theta}$ be an estimator of θ that is asymptotically normal:

$$\frac{\widehat{\theta} - \theta}{\text{se}(\widehat{\theta})} \xrightarrow{\text{d}} \mathsf{N}(0,1).$$

The **Wald test statistic** is

$$W = \frac{\widehat{\theta} - \theta_0}{\widehat{\text{se}}(\widehat{\theta})} \overset{\text{approx.}}{\sim} \mathsf{N}(0,1)$$

Let w be the observed value of W. Then the approximate P-value is given by

$$P\text{-value} \simeq \mathbb{P}(|Z| > |w|) = 2\Phi(-|w|)$$

where $Z \sim \mathsf{N}(0,1)$.

Usually the estimator $\widehat{\theta}$ in the Wald test is the maximum likelihood estimator since, for smooth likelihood situations, this estimator satisfies the asymptotic normality requirement, and a formula for the (approximate) standard error is readily available.

Example

Consider the sample of size $n = 100$

```
0.366 0.568 0.300 0.115 0.204 0.128 0.277 0.391 0.328 0.451
0.412 0.190 0.207 0.147 0.116 0.326 0.256 0.524 0.217 0.485
0.265 0.375 0.267 0.360 0.250 0.258 0.583 0.413 0.481 0.468
0.406 0.336 0.305 0.321 0.268 0.361 0.632 0.283 0.258 0.466
0.276 0.232 0.133 0.316 0.468 0.496 0.573 0.523 0.256 0.491
0.127 0.054 0.440 0.228 0.249 0.754 0.430 0.111 0.459 0.233
0.257 0.640 0.147 0.273 0.112 0.389 0.126 0.356 0.273 0.296
0.433 0.253 0.234 0.514 0.177 0.221 0.534 0.509 0.510 0.269
0.262 0.625 0.183 0.541 0.705 0.078 0.847 0.149 0.031 0.453
0.299 0.226 0.069 0.211 0.195 0.381 0.317 0.467 0.289 0.593
```

which may be considered to be the observed value of a random sample $X_1, \ldots, X_{100}$ with common density function:

$$f(x;\theta) = 2\theta x e^{-\theta x^2}, \;\; x \geq 0; \;\; \theta > 0$$

Use a Wald test to test the hypotheses:

$$H_0 : \theta = 6 \qquad \text{versus} \qquad H_1 : \theta \neq 6$$

In Chapter 8 it was shown that the maximum likelihood estimator is

$$\widehat{\theta} = \frac{100}{\sum_{i=1}^{100} X_i^2}$$

with estimated standard error

$$\widehat{\text{se}}(\widehat{\theta}) = \frac{\widehat{\theta}}{\sqrt{100}}.$$

It may also be shown that $\widehat{\theta}$ is asymptotically normal so the Wald test applies. The Wald test statistic is

$$W = \frac{\widehat{\theta} - 6}{\widehat{\text{se}}(\widehat{\theta})} = \frac{\widehat{\theta} - 6}{\widehat{\theta}/\sqrt{100}} = \frac{\frac{100}{\sum_{i=1}^{100} X_i^2} - 6}{\frac{100}{\sum_{i=1}^{100} X_i^2}/10}.$$

Since $\sum_{i=1}^{100} x_i^2 = 14.081$ the observed value of W is

$$w = \frac{\frac{100}{14.081} - 6}{\frac{100}{14.081}/10} = 1.5514.$$

Then, with $Z \sim \mathsf{N}(0,1)$,

$$\text{p-value} = \mathbb{P}(|Z| > 1.55) = 2\Phi(-1.55) = 0.12.$$

There is little or no evidence against H_0 so we should retain the null hypothesis.

Example

(Ecology 2005, 86:1057-1060)

Do ravens intentionally fly towards gunshot sounds (to scavenge on the carcass they expect to find)? Crow White addressed this question by going to 12 locations, firing a gun, then counting raven numbers 10 minutes later. He repeated the process at 12 different locations where he didn't fire a gun. Results:

no gunshot	0	0	2	3	5	0	0	0	1	0	1	0
gunshot	2	1	4	1	0	5	0	1	0	3	5	2

Is there evidence that ravens fly towards the location of gunshots? Answer this question using an appropriate Wald test.

Here we might proceed by assuming that B_i and A_i is the number of ravens at the i-th location ($i = 1, \dots, 12 = n$) before and after the firing of the gunshots. Note that here $A_1, \dots, A_n$ and $B_1, \dots, B_n$ are pairwise dependent, because the A_i depends on the B_i. We only have independence across the i-th, that is, the pairs

$$(A_1, B_1), \dots, (A_n, B_n)$$

are independent with common mean (μ_a, μ_b). Here, due to the dependence, we cannot use a two sample student test. A crude alternative model is to consider the difference

$$X_i = A_i - B_i$$

Then, $X_1, \dots, X_n$ are independent and we could test

$$H_0 : \ \mu_a = \mu_b \qquad H_1 : \ \mu_a > \mu_b$$

using a one sample student test:

$$T = \frac{\bar{X} - \overbrace{(\mu_a - \mu_b)}^{0 \text{ under} H_0}}{S/\sqrt{n}} \sim \mathsf{t}_{n-1}$$

This gives the p-value

$$2 \times \mathbb{P}(T > \frac{1}{2.73/\sqrt{12}}) \approx 2 \times 0.1153.. \approx 0.23$$

We therefore conclude that there is no evidence that ravens intentionally fly towards gunshots.

Likelihood Ratio Tests

The Wald test described in the previous section is a general testing procedure for the situation where an asymptotically normal estimator is available. An even more general procedure, with good power properties, is the **likelihood ratio test**.

The Likelihood Ratio Test (Single Parameter Case)
Consider the hypotheses

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

The **likelihood ratio test statistic** is

$$\Lambda(\widehat{\theta}_n) = 2 \ln \left(\frac{\mathcal{L}(\widehat{\theta}_n)}{\mathcal{L}(\theta_0)} \right) = 2\{\ell_n(\widehat{\theta}_n) - \ell_n(\theta_0)\}.$$

Under H_0 and certain regularity conditions

$$\Lambda(\widehat{\theta}_n) \xrightarrow{\mathrm{d}} \chi_1^2.$$

We will frequently drop the n's from the notation for convenience. Let λ be the observed value of Λ. Then the approximate P-value is given by

$$P\text{-value} \simeq \mathbb{P}_{\theta_0}(\Lambda > \lambda) = \mathbb{P}(Q > \lambda) = 2\Phi(-\sqrt{\lambda})$$

where $Q \sim \chi_1^2$.

proof: Assume the H_0 hypothesis is true and that the data $X_1, \ldots, X_n$ are iid random variables from the 'true' $f(x; \theta_0)$. Then, recall that under some regularity conditions

$$\sqrt{nI_1(\theta_0)}(\theta_0 - \widehat{\theta}_n) \xrightarrow{\mathrm{d}} Z \sim \mathsf{N}(0, 1)$$

By Taylor's expansion around $\widehat{\theta}_n$ and the Mean Value Theorem we have

$$\ell_n(\theta_0) = \ell_n(\widehat{\theta}_n) + \ell_n'(\widehat{\theta}_n)(\theta_0 - \widehat{\theta}_n) + \frac{1}{2}(\theta_0 - \widehat{\theta}_n)^2 \ell_n''(\vartheta_n)$$

for some ϑ_n between $\widehat{\theta}_n$ and θ_0. Since $\ell_n'(\widehat{\theta}_n) = 0$, we obtain after rearrangement

$$
\begin{aligned}
2(\ell_n(\widehat{\theta}_n) - \ell_n(\theta_0)) &= -\ell_n''(\vartheta_n)(\theta_0 - \widehat{\theta}_n)^2 \\
&= \frac{-\ell_n''(\vartheta_n)}{nI_1(\theta_0)} \left(\sqrt{nI_1(\theta_0)}(\theta_0 - \widehat{\theta}_n)\right)^2 \\
&= \underbrace{\frac{-\ell_n''(\vartheta_n)}{n}}_{\xrightarrow{\mathbb{P}} I_1(\theta_0)} I_1^{-1}(\theta) \Big(\underbrace{\sqrt{nI_1(\theta_0)}(\theta_0 - \widehat{\theta}_n)}_{\xrightarrow{\mathrm{d}} Z} \Big)^2 \xrightarrow{\mathrm{d}} 1 \times (Z^2) \sim \chi_1^2
\end{aligned}
$$

where we used the result from Remark 8.1 and Slutsky's theorem. □

Remark 9.1 (Relation between Wald and Likelihood Ratio Test Statistics) *A Wald statistic uses the horizontal axis, for θ, to construct a test statistic – we take $\widehat{\theta}$ and compare it to θ_0, to see if it $\widehat{\theta}$ is significantly far from θ_0.*

In constrast, a likelihood ratio statistic uses the vertical axis, for $\ell(\theta)$, to construct a test statistic – we take the maximised log-likelihood, $\ell(\widehat{\theta})$, and compare it to the log-likelihood under the null hypothesis, $\ell(\theta_0)$, to see if $\ell(\theta_0)$ is significantly far from the maximum.

Note that if $W = (\widehat{\theta}_n - \theta_0)/se(\widehat{\theta}_n)$ is the Wald statistic, then

$$
\Lambda = 2(\ell_n(\widehat{\theta}_n) - \ell_n(\theta_0)) = \frac{-\ell_n''(\vartheta_n)}{n}\, se^2(\widehat{\theta}_n)\, W^2 \xrightarrow{\mathbb{P}} W^2
$$

Thus, the Wald and likelihood ratio tests are asymptotically equivalent when the null hypothesis is true – and in large samples, they typically return similar test statistics hence similar conclusions. However, when the null hypothesis is not true, these tests can have quite different properties, especially in small samples.

Example

Consider, one last time, the sample of size $n = 100$

```
0.366 0.568 0.300 0.115 0.204 0.128 0.277 0.391 0.328 0.451
0.412 0.190 0.207 0.147 0.116 0.326 0.256 0.524 0.217 0.485
0.265 0.375 0.267 0.360 0.250 0.258 0.583 0.413 0.481 0.468
0.406 0.336 0.305 0.321 0.268 0.361 0.632 0.283 0.258 0.466
0.276 0.232 0.133 0.316 0.468 0.496 0.573 0.523 0.256 0.491
0.127 0.054 0.440 0.228 0.249 0.754 0.430 0.111 0.459 0.233
0.257 0.640 0.147 0.273 0.112 0.389 0.126 0.356 0.273 0.296
0.433 0.253 0.234 0.514 0.177 0.221 0.534 0.509 0.510 0.269
0.262 0.625 0.183 0.541 0.705 0.078 0.847 0.149 0.031 0.453
0.299 0.226 0.069 0.211 0.195 0.381 0.317 0.467 0.289 0.593
```

which may be considered to be the observed value of a random sample $X_1, \ldots, X_{100}$ with common density function f given by:

$$f(x;\theta) = 2\theta x e^{-\theta x^2}, \; x \geq 0; \; \theta > 0.$$

Use a likelihood ratio test to test the hypotheses:

$$H_0 : \theta = 6 \qquad \text{versus} \qquad H_1 : \theta \neq 6$$

First, note that the likelihood ratio statistic is

$$\lambda = 2\ln\left(\frac{\mathcal{L}(\widehat{\theta})}{\mathcal{L}(6)}\right) = 2\{\ell(\widehat{\theta}) - \ell(6)\}$$

where

$$\ell(\theta) = \sum_{i=1}^{100} \ln\{f(X_i;\theta)\} = 100\ln(2) + 100\ln(\theta) + \sum_{i=1}^{100}\ln(X_i) - \theta\sum_{i=1}^{100} X_i^2.$$

As shown before, the maximum likelihood estimator is

$$\widehat{\theta} = \frac{100}{\sum_{i=1}^{100} X_i^2}$$

so

$$\begin{aligned} \ell(\widehat{\theta}) &= 100\ln(2) + 100\ln\left(\frac{100}{\sum_{i=1}^{100} X_i^2}\right) + \sum_{i=1}^{100}\ln(X_i) - \left(\frac{100}{\sum_{i=1}^{100} X_i^2}\right)\sum_{i=1}^{100} X_i^2 \\ &= 100\{\ln(2) + \ln(100)\} - 100\ln\left(\sum_{i=1}^{100} X_i^2\right) + \sum_{i=1}^{100}\ln(X_i) - 100. \end{aligned}$$

Also,

$$\ell(6) = 100\ln(2) + 100\ln(6) + \sum_{i=1}^{100}\ln(X_i) - 6\sum_{i=1}^{100} X_i^2$$

and so the likelihood ratio statistic is

$$\Lambda = 2\left[100\{\ln(100/6) - 1\} - 100\ln\left(\sum_{i=1}^{100} X_i^2\right) + 6\sum_{i=1}^{100} X_i^2\right].$$

Since $\sum_{i=1}^{100} x_i^2 = 14.081$ the observed value of Λ is

$$\lambda = 2\left[100\{\ln(100/6) - 1\} - 100\ln(14.081) + 6 \times 14.081\right] = 2.689.$$

Then

$$\begin{aligned} \text{p-value} &= \mathbb{P}_{\theta=6}(\Lambda > \lambda) \\ &= \mathbb{P}(Q > 2.689), \quad Q \sim \chi_1^2 \\ &= \mathbb{P}(Z^2 > 2.689), \quad Z \sim \mathsf{N}(0,1) \\ &= 2\mathbb{P}(Z \leq -\sqrt{2.689}) \\ &= 2\Phi(-1.64) = 0.10 \end{aligned}$$

which is close to the p-value of 0.12 obtained via the Wald test in the previous section. The conclusion remains that there is little or no evidence against H_0 and that H_0 should be retained.

Multiparameter Extension of the Likelihood Ratio Test

The likelihood ratio test procedure can be extended to hypothesis tests involving several parameters simultaneously. Such situations arise, for example, in the important branch of statistics known as *regression.*

Consider a model with parameter vector $\boldsymbol{\theta}$ and corresponding parameter space Θ. A general class of hypotheses is:

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \notin \Theta_0$$

where Θ_0 is a subset of Θ. Then the likelihood ratio statistic is

$$\Lambda = 2 \ln \left(\frac{\sup_{\theta \in \Theta} \mathcal{L}(\boldsymbol{\theta})}{\sup_{\theta \in \Theta_0} \mathcal{L}(\boldsymbol{\theta})} \right).$$

Under H_0 and regularity conditions on $\mathcal{L}$,

$$\Lambda \xrightarrow{\mathrm{d}} \chi^2_d$$

where d is the dimension of Θ minus the dimension of Θ_0.

We will now show how the multivariate version motivates the one- and two-sample student hypothesis tests.

Derivation of One Sample t-test

Suppose we have the model

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$$

with both μ, σ^2 unknown. We wish to test

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0$$

Then, if $\boldsymbol{\theta} = (\mu; \sigma^2)$, the likelihood is

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2} \sigma^n} \mathrm{e}^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2}$$

and the ratio test statistic is

$$\Lambda = 2 \ln \frac{\max_{\mu; \sigma^2} \mathcal{L}(\mu; \sigma^2)}{\max_{\mu = \mu_0; \sigma^2} \mathcal{L}(\mu; \sigma^2)} = 2 \ln \frac{\mathcal{L}(\widehat{\mu}; \widehat{\sigma}^2)}{\mathcal{L}(\mu_0; \widehat{\sigma}^2_0)},$$

where

$$\begin{aligned}
\widehat{\mu} &= \bar{X}\\
\widehat{\sigma}^2 &= \frac{1}{n}\sum_i (X_i - \bar{X})^2\\
\widehat{\mu}_0 &= \mu_0\\
\widehat{\sigma}_0^2 &= \frac{1}{n}\sum_i (X_i - \mu_0)^2
\end{aligned}$$

are the MLE of $(\mu; \sigma^2)$ restricted to $\{(\mu; \sigma^2) : \mu \in \mathbb{R}; \sigma > 0\}$ and $\Theta_0 = \{(\mu; \sigma^2) : \mu = \mu_0; \sigma > 0\}$, respectively. We now simplify the likelihood ratio to obtain:

$$\begin{aligned}
\Lambda(\boldsymbol{\theta}) &= 2\ln\frac{\widehat{\sigma}_0^n \mathrm{e}^{-\frac{1}{2\widehat{\sigma}^2}\sum_{i=1}^n (X_i - \widehat{\mu})^2}}{\widehat{\sigma}^n \mathrm{e}^{-\frac{1}{2\widehat{\sigma}_0^2}\sum_{i=1}^n (X_i - \mu_0)^2}}\\
&= n\ln\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}^2}\\
&= n\ln\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\\
&= n\ln\frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (\bar{X} - \mu_0)^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\\
&= n\ln\left(1 + n\frac{(\bar{X} - \mu_0)^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)\\
&= n\ln\left(1 + \frac{n}{n-1}\frac{(\bar{X} - \mu_0)^2}{\sum_{i=1}^n (X_i - \bar{X})^2/(n-1)}\right)\\
&= n\ln\left(1 + \frac{1}{n-1}\frac{(\bar{X} - \mu_0)^2}{S^2/n}\right)\\
&= n\ln\left(1 + \frac{1}{n-1}T^2\right),
\end{aligned}$$

where

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

is the student statistic. It follows that the likelihood ratio test statistic is a monotonic function of T^2. If T^2 is too large, then Λ is too large and we reject H_0. Thus, a p-value for Λ can be related to a p-value for T as follows:

$$\mathbb{P}(\Lambda \geqslant \lambda) = \mathbb{P}(T^2 \geqslant (n-1)(\mathrm{e}^{\lambda/n} - 1)) = 2\mathbb{P}(T > t),$$

where $t = \sqrt{(n-1)(\mathrm{e}^{\lambda/n} - 1)}$, Hence, instead of dealing with Λ, we could perform this two sided hypothesis test using $T \sim \mathsf{t}_{n-1}$ only. This is the motivation for the one-sample student test. The test has optimal asymptotic power properties similar to those of the MLE. Thus, again the maximum likelihood principle gives us a method for deriving not only optimal estimators, but optimal test statistics for a given hypothesis.

Derivation of Two Sample t-test

Here the model is

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu_X; \sigma^2)$$

is independent of

$$X_1, \ldots, X_m \overset{\text{iid}}{\sim} \mathsf{N}(\mu_Y; \sigma^2)$$

and we wish to test:

$$H_0 : \quad \mu_X = \mu_Y = \mu \quad \text{versus} \quad H_1 : \mu_X \neq \mu_Y$$

where $\mu, \mu_X, \mu_Y, \sigma$ are all unknown. We now derive the two sample test statistic

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathsf{t}_{m+n-2},$$

where

$$S_p^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^m (Y_i - \bar{Y})^2}{m+n-2}$$

using the likelihood ratio test. The likelihood function of the joint data is

$$\mathcal{L}(\mu_X, \mu_Y, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{m+n}{2}}} \mathrm{e}^{-\frac{\sum_{i=1}^n (X_i-\mu_X)^2 + \sum_{i=1}^m (Y_i-\mu_Y)^2}{2\sigma^2}}$$

Hence, the log-likelihood is

$$\ell(\boldsymbol{\theta}) = -\frac{\sum_{i=1}^n (X_i - \mu_X)^2 + \sum_{i=1}^m (Y_i - \mu_Y)^2}{2\sigma^2} - \frac{m+n}{2}\ln(2\pi\sigma^2)$$

and the likelihood ratio can be written as

$$\Lambda(\boldsymbol{\theta}) = 2(\ell(\widehat{\mu}_X, \widehat{\mu}_Y, \widehat{\sigma}^2) - \ell(\widehat{\mu}, \widehat{\mu}, \widetilde{\sigma}^2))$$

where

$$\begin{aligned}
\widehat{\mu}_X &= \bar{X} \\
\widehat{\mu}_Y &= \bar{Y} \\
\widehat{\sigma}^2 &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^m (Y_i - \bar{Y})^2}{m+n}
\end{aligned}$$

is the MLE of $\boldsymbol{\theta}$ without any restrictions and

$$\begin{aligned}
\widehat{\mu} &= \frac{n\bar{X} + m\bar{Y}}{n+m} = \frac{n}{m+n}\bar{X} + \frac{n}{m+n}\bar{Y} \\
\widetilde{\sigma}^2 &= \frac{\sum_{i=1}^n (X_i - \widehat{\mu})^2 + \sum_{i=1}^m (Y_i - \widehat{\mu})^2}{m+n}
\end{aligned}$$

is the MLE of $\boldsymbol{\theta}$ with the restriction $\mu_X = \mu_Y = \mu$. We now simplify the likelihood ratio to obtain:

$$\begin{aligned}
\Lambda(\boldsymbol{\theta}) &= 2(\ell(\widehat{\mu}_X, \widehat{\mu}_Y, \widehat{\sigma}^2) - \ell(\widehat{\mu}, \widehat{\mu}, \widetilde{\sigma}^2)) \\
&= -(m+n)\ln\frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^m (Y_i - \bar{Y})^2}{\sum_{i=1}^n (X_i - \widehat{\mu})^2 + \sum_{i=1}^m (Y_i - \widehat{\mu})^2}
\end{aligned}$$

To simplify this further note that

$$\sum_{i=1}^{n}(X_i - \widehat{\mu})^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + \frac{m^2 n}{(m+n)^2}(\bar{X} - \bar{Y})^2$$

and similarly

$$\sum_{i=1}^{m}(Y_i - \widehat{\mu})^2 = \sum_{i=1}^{m}(Y_i - \bar{Y})^2 + \frac{mn^2}{(m+n)^2}(\bar{Y} - \bar{X})^2$$

Hence, the bottom of the fraction in ln() above becomes:

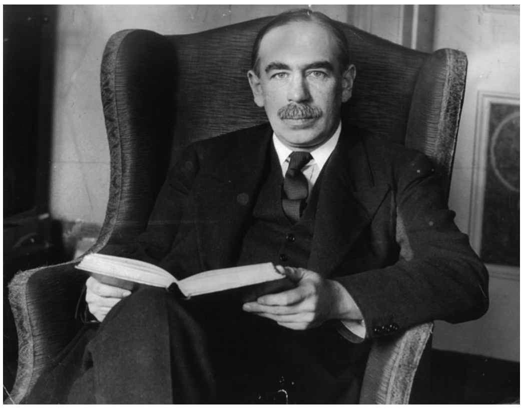
$$\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2 + \frac{mn}{m+n}(\bar{Y} - \bar{X})^2$$

Therefore,

$$\begin{aligned}
\Lambda &= -(m+n)\ln\left(\frac{(m+n-2)S_p^2}{(m+n-2)S_p^2 + \frac{mn}{m+n}(\bar{Y} - \bar{X})^2}\right) \\
&= (m+n)\ln\left(\frac{m+n-2+\frac{(\bar{X}-\bar{Y})^2}{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)}}{m+n-2}\right) \\
&= (m+n)\ln\left(1 + \frac{T^2}{m+n-2}\right)
\end{aligned}$$

Again, Λ is a monotonic function of T^2 and therefore instead of dealing with the distribution of Λ when computing p-values, we could equivalently work with the two-sample student statistic T.

Figure 9.1: John Maynard Keynes: the most famous and influential economist of the 20-th century and author of *A Treatise on Probability* (1921).



That's it – good luck in your exams and remember that all models in Statistics are wrong, but that some are useful or as Keynes put it "It is better to be roughly right than precisely wrong". You will consider statistical modeling and probability theory in much more detail in your third year courses.

A Brief Introduction to R

This set of exercises introduces you to the statistics package R. R is a free and widely used statistics program that you can download from `http://cran.r-project.org/`. R is the statistics package that was used to construct most of the graphs in your lecture notes. Some of your assignment questions will require computer-aided data analysis, for which you are encouraged to use R.

By the end of this set of exercises you will know how to:

- Load R in Windows.
- Enter data into R, manually or by importing it.
- Construct numerical and graphical summaries on R.
- Save R graphs as jpg files, and save your work before exiting R.

Starting R

Log onto a maths computer under the Windows environment. You can also use R under Linux, by typing R at the command line, and you are welcome to explore this option yourself. These notes however are written for the Windows environment, and the methods are different for Linux (in particular, no drag-down menus are available, and the saving of graphs is a little trickier).

Go to **Start**...**Programs**...**R**...**R2.4.1**.

Notice that R opens in the window called “R Console”. R is a command-line program, which means that you need to type commands into the R Console line-by-line after the red prompt “>” at the bottom of the window.

If at any time the prompt disappears or changes into a “+”, you’ve probably made a mistake. In this situation, if you press the escape key `Esc` you will be back in business.

Entering data into R

Entering your data into R manually

If your dataset is small, you can enter it in manually. For example, to enter the following dataset and to store it as the object dat:

2.3 2.5 6.5 7.4 1.1 3.9

type at the prompt (>)

```
> dat = c(2.3, 2.5, 6.5, 7.4, 1.1, 3.9)
```

then press return.

You can type dat and press return to check that R has recorded your data correctly – it should return to you the values you entered.

Now that R has your data stored in it, you can do some calculations to summarise the variable dat (such as finding the mean, median, etc). We will do this a little later.

Importing data into R

When entering data into a computer, people typically use a spreadsheet such as Excel, or some other program. We can take data from these other programs and import it into R for analysis. We will learn how to do this for a dataset that has been stored in tab-delimited format.

Download the file 1041_07.txt from My eLearning and open this file. This file contains MATH1041 student responses to a survey that was administered at the start of session 2, 2007. (If you are curious about what questions were asked in the survey, you can find them in the file 1041_07Evaluation.pdf.) The 1041_07.txt dataset is stored in a tab-delimited file, which means that the Tab key has been used to distinguish between columns of the dataset. There are many columns in the dataset, and each has a label in the first row identifying which variable it is (*e.g.* travel.time). Notice that the column labels have no spaces in them – this is important because R will have problems reading the dataset if there are spaces.

To import this dataset into R:

- First save 1041_07.txt in your personal drive. This drive is labelled using your student number.
- Now change R's working directory to the directory on your personal drive that contains 1041_07.txt, by going to "**File...Change dir...Browse...**" and navigating to the directory on your personal drive that contains 1041_07.txt.

The reason for doing this is to tell R where it will be able to find the file 1041_07.txt. Whenever you ask R to load a file, you need to also tell R which directory to find the file in, and this is a simple way of doing that.

- To load the file 1041_07.txt and to store it under the name survey, type:
```
> survey=read.table("1041_07.txt",header=T)
```
(read.table is the R command that reads data in from a table that is stored in a text file, and header=T tells R that the first row of the file is a "header row" which contains the names for each column.)

To check that the data are in R, ask R to show you the first 10 rows of the dataset, which you can do using the command survey[1:10,]

Numerical summaries on R

Numerical summaries of a quantitative variable

We will now illustrate how to calculate numerical summaries of quantitative variables using the dataset dat which you entered earlier. It's pretty easy! Just use the following commands...

- To calculate the mean of dat, type the command mean(dat) at the prompt, then press return.

- To calculate the standard deviation of dat, type the command sd(dat)

- To calculate the median of dat, type the command median(dat)

- To calculate a five number summary, use the command summary(dat). Note that this actually returns a six number summary – it throws in the mean as well.

- To calculate the 25th percentile (otherwise known as the first quartile), use the command quantile(dat,0.25)

All of these commands will also work for quantitative variables in the survey data, which you have stored in R as survey, although you should specify which variable from this dataset you want to use when typing a command. For example, to access the travel-time variable, you will need to refer to survey$travel.time in your command, not just travel.time. So to calculate a five-number summary of travel time, use the command summary(survey$travel.time).

Numerical summary of a categorical variable

Use the `summary` command to obtain a table of frequencies of a categorical variable. For example, type `summary(survey$gender)`

Removing NA's

In some cases, a MATH1041 student attempted the survey but did not complete all questions. This means that there are some places in the dataset with no response, which have been labelled as `NA`. You might have noticed that `R` tells you how many NA values there are for a given variable, when you use the `summary` command. The `travel.time` variable, for example, contains 12 NA values.

A problem with having NA's in a quantitative variable is that when you try to calculate some function of this variable, such as the mean, it will return `NA`. For example, see what you get when you type `mean(survey$travel.time)`. To overcome this problem, there is an option in most functions called `na.rm` (which stands for "remove NA's"). To calculate the average travel time, ignoring the NA values, use the command `mean(survey$travel.time, na.rm=T)`.

Graphical summaries on R

Graphical summary of one quantitative variable

To graph the `travel.time` variable from the `survey` dataset, type

- `boxplot(survey$travel.time)` for a boxplot.
- `hist(survey$travel.time)` for a histogram.

Graphical summary of one categorical variable

To construct a bar chart of the `gender` variable from the `survey` dataset, use the command `plot(survey$gender)`

Graphical summary comparing several samples of a quantitative variable

To construct comparative boxplots for the hair cut costs of different genders, type any of the following:

- `plot(survey$hair.cost~survey$gender)`
- `boxplot(survey$hair.cost~survey$gender)`

- `plot(hair.cost~gender, data=survey)`

Note that most functions have an optional `data=` argument so you don't have to use the "`survey$`" form every time you want to use a variable from within the dataset `survey`.

The "~" stands for "against", that is, `plot(hair.cost~gender, data=survey)` means "plot `hair.cost` against `gender`, for the dataset `survey`."

Saving your work

Saving graphs as jpeg files

To save a graph as a **jpeg** image file, click on it (to make the graph window active), then go to **File...Save As... jpeg... 75% quality...** and type a suitable filename. Notice that it is possible to save your graphs in other formats too, such as `PDF`.

Saving your commands

To keep a record, or a **History**, of every command you have typed during an `R` session, go to **File...Save History...** and choose a name for the file. A text file will be created, which will list every single command you have typed at the command line in your current `R` session. You could then rerun your analyses later by opening this file and copying and pasting the commands into `R` (all at once, if you wanted to).

This function is really handy in statistical analysis, because sometimes when analysing data you find a mistake in the original dataset, or in your method of analysis, so you need to go "back to the drawing board" and repeat your whole analysis. But it's so easy to repeat your analysis on `R` – just copy and paste all commands from your history file!

Saving your workspace

To save all objects currently in your workspace (if you have completed all the above steps, your workspace currently consists of `dat` and `survey`) go to **File...Save Workspace...** and suggest a filename. If you ever want to continue analyses from the point you are currently at, you can simply find the workspace file on My Computer and double-click on it (provided that it has been saved with the `.RData` file extension).

Getting help on R

There are a few different ways to find help on R, if you want to explore it further...

- Click on the R Console (so that it is the active window) and you will find a drag-down **Help** menu. You can browse through **html help** or an introductory manual available at **Manuals (In PDF)... An Introduction to R**.

- To search for help on a particular technique, use the `help.search` command, *e.g.* type `help.search("histogram")` at the command line to find out what functions are available for constructing histograms.

- If you know what function you want to find out more about, use the `help` command. *e.g.* to find out more about the `hist` command, type `help(hist)` at the command line.

- Buy a book that introduces you to R, or download a free book from the Internet. There are a lot of free books around, see suggestions at `http://cran.r-project.org/` under "Documentation – Contributed".

Other stuff

There's so much you can do on R. One particular example is doing arithmetic calculations – you can use R like a calculator.

```
> 3*log(2.4)-3
> choose(5,2) * 0.4

94         3

* (1-0.4)

94         2
```

You can also store the results of previous commands, which is important for doing more complex calculations or for retrieving results later on.

```
> mn.trav = mean(survey$travel.time, na.rm=T)
> sd.trav = sd(survey$travel.time, na.rm=T)
> mn.trav + 3*sd.trav
> x = 0:5
> p = choose(5,x) * 0.4

94         x

* (1-0.4)

94         (
5-x)
> plot(x,p)
```

Exiting R

Make the R Console the active window (by clicking on it) then go to **File... Exit**.

Statistical Tables

UNIVERSITY OF NEW SOUTH WALES • SCHOOL OF MATHEMATICS AND STATISTICS

Statistical Tables

t distribution critical values

Key: Table entry for p and C is the critical value t^* with probability p lying to its right
and probability C lying between $-t^*$ and t^*.

	Upper tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
	.50	.60	.70	0.80	.90	.95	.96	.98	.99	.995	.998	.999
	Probability C											

Standard normal probabilities

Key: Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

χ^2 distribution critical values

Key: Table entry for p is the critical value with probability p lying to its right.

	Upper tail probability p									
df	.995	.99	.975	.95	.90	.10	.05	.025	.01	.005
1	0.000039	0.00016	0.00098	0.0039	0.0158	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.61	5.99	7.38	9.21	10.60
3	0.0717	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33	116.32
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81	140.17

F distribution critical values

Key: p=Upper tail probability p, $\mathrm{df_n}$=degrees of freedom in numerator, $\mathrm{df_d}$=degrees of freedom in denominator, $*$ Multiply by 10, † Multiply by 100.

	$\mathrm{df_n}$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
$\mathrm{df_d}$	p																			
1	.05	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
	.025	648	800	864	900	922	937	948	957	963	969	977	986	993	997	1001	1006	1010	1014	1018
	.01	405*	500*	540*	563*	576*	586*	593*	598*	602*	606*	611*	616*	621*	624*	626*	629*	631*	634*	637*
	.005	162†	200†	216†	225†	231†	234†	237†	239†	241†	242†	244†	246†	248†	249†	250†	251†	253†	254†	255†
2	.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
	.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
	.005	199	199	199	199	199	199	199	199	199	199	199	199	199	200	200	200	200	200	200
3	.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
	.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
	.005	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43.69	43.39	43.08	42.78	42.62	42.47	42.31	42.15	41.99	41.83
4	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
	.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
	.005	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97	20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	19.32
5	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
	.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
	.005	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	13.38	13.15	12.90	12.78	12.66	12.53	12.40	12.27	12.14
6	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
	.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
	.005	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	8.88
7	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.41	4.36	4.31	4.25	4.20	4.14
	.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
	.005	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	8.18	7.97	7.75	7.64	7.53	7.42	7.31	7.19	7.08
8	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
	.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
	.005	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	5.95
9	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
	.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
	.005	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30	5.19
10	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
	.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
	.005	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.66	5.47	5.27	5.17	5.07	4.97	4.86	4.75	4.64
12	.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
	.005	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.91	4.72	4.53	4.43	4.33	4.23	4.12	4.01	3.90
15	.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
	.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
	.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
	.005	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.25	4.07	3.88	3.79	3.69	3.58	3.48	3.37	3.26
20	.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
	.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
	.005	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	2.69
24	.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
	.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
	.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
	.005	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55	2.43
30	.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
	.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
	.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
	.005	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.18	3.01	2.82	2.73	2.63	2.52	2.42	2.30	2.18
40	.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
	.025	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
	.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
	.005	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	2.95	2.78	2.60	2.50	2.40	2.30	2.18	2.06	1.93
60	.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
	.025	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
	.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
	.005	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.74	2.57	2.39	2.29	2.19	2.08	1.96	1.83	1.69
120	.05	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
	.025	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
	.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
	.005	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.54	2.37	2.19	2.09	1.98	1.87	1.75	1.61	1.43
∞	.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00
	.025	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00
	.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00
	.005	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.36	2.19	2.00	1.90	1.79	1.67	1.53	1.36	1.00