M2S1 Lecture Notes

G. A. Young
http://www2.imperial.ac.uk/∼ayoung

September 2009

Contents

CHAPTER 1

DEFINITIONS, TERMINOLOGY, NOTATION

1.1 EVENTS AND THE SAMPLE SPACE

Definition 1.1.1 An **experiment** is a one-off or repeatable process or procedure for which
(a) there is a well-defined set of *possible* outcomes
(b) the *actual* outcome is not known with certainty.

Definition 1.1.2 A **sample outcome**, ω, is precisely one of the possible outcomes of an experiment.

Definition 1.1.3 The **sample space**, Ω, of an experiment is the set of all possible outcomes.

NOTE : Ω is a set in the mathematical sense, so set theory notation can be used. For example, if the sample outcomes are denoted $\omega_1, ..., \omega_k$, say, then

$$\Omega = \{\omega_1, ..., \omega_k\} = \{\omega_i : i = 1, ..., k\},$$

and $\omega_i \in \Omega$ for $i = 1, ..., k$.

The sample space of an experiment can be

- a FINITE list of sample outcomes, $\{\omega_1, ..., \omega_k\}$
- a (countably) INFINITE list of sample outcomes, $\{\omega_1, \omega_2, ...\}$
- an INTERVAL or REGION of a real space, $\{\omega : \omega \in A \subseteq \mathbb{R}^d\}$

Definition 1.1.4 An **event**, E, is a designated collection of sample outcomes. Event E **occurs** if the actual outcome of the experiment is one of this collection. An event is, therefore, a subset of the sample space Ω.

Special Cases of Events

The event corresponding to the collection of *all* sample outcomes is Ω.

The event corresponding to a collection of *none* of the sample outcomes is denoted $\emptyset$.

i.e. The sets $\emptyset$ and Ω are also events, termed the **impossible** and the **certain** event respectively, and for any event E, $E \subseteq \Omega$.

1.1.1 OPERATIONS IN SET THEORY

Since events are subsets of Ω, set theory operations are used to manipulate events in probability theory. Consider events $E, F \subseteq \Omega$. Then we can reasonably concern ourselves also with events obtained from the three basic set operations:

UNION	$E \cup F$	"E or F or both occur"
INTERSECTION	$E \cap F$	"both E and F occur"
COMPLEMENT	E'	"E does not occur"

Properties of Union/Intersection operators

Consider events $E, F, G \subseteq \Omega$.

COMMUTATIVITY	$E \cup F = F \cup E$ $E \cap F = F \cap E$
ASSOCIATIVITY	$E \cup (F \cup G) = (E \cup F) \cup G$ $E \cap (F \cap G) = (E \cap F) \cap G$
DISTRIBUTIVITY	$E \cup (F \cap G) = (E \cup F) \cap (E \cup G)$ $E \cap (F \cup G) = (E \cap F) \cup (E \cap G)$
DE MORGAN'S LAWS	$(E \cup F)' = E' \cap F'$ $(E \cap F)' = E' \cup F'$

1.1.2 MUTUALLY EXCLUSIVE EVENTS AND PARTITIONS

Definition 1.1.5 Events E and F are **mutually exclusive** if $E \cap F = \emptyset$, that is, if events E and F cannot both occur. If the sets of sample outcomes represented by E and F are **disjoint** (have no common element), then E and F are mutually exclusive.

Definition 1.1.6 Events $E_1, ..., E_k \subseteq \Omega$ form a **partition** of event $F \subseteq \Omega$ if
(a) $E_i \cap E_j = \emptyset$ for $i \neq j$, $i, j = 1, ..., k$
(b) $\bigcup_{i=1}^{k} E_i = F$,
so that each element of the collection of sample outcomes corresponding to event F is in *one and only one* of the collections corresponding to events $E_1, ...E_k$.

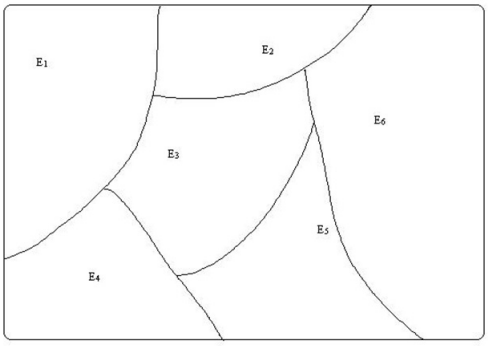

Figure 1.1: Partition of Ω

In Figure 1.1, we have $\Omega = \bigcup_{i=1}^{6} E_i$

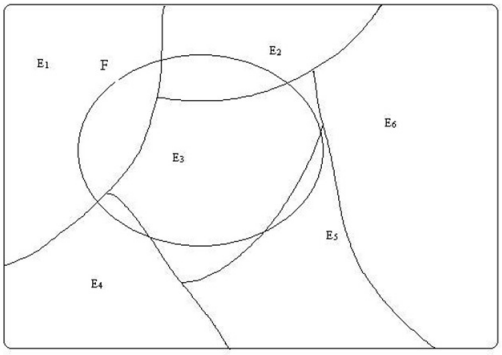

Figure 1.2: Partition of $F \subset \Omega$

In Figure 1.2, we have $F = \bigcup_{i=1}^{6} (F \cap E_i)$, but, for example, $F \cap E_6 = \emptyset$.

1.2 THE σ-FIELD

Events are subsets of Ω, but need all subsets of Ω be events? The answer is negative. But it suffices to think of the collection of events as a subcollection $\mathcal{A}$ of the set of all subsets of Ω. This subcollection should have the following properties:

(a) if $A, B \in \mathcal{A}$ then $A \cup B \in \mathcal{A}$ and $A \cap B \in \mathcal{A}$;

(b) if $A \in \mathcal{A}$ then $A' \in \mathcal{A}$;

(c) $\emptyset \in \mathcal{A}$.

A collection $\mathcal{A}$ of subsets of Ω which satisfies these three conditions is called a **field**. It follows from the properties of a field that if $A_1, A_2, \ldots, A_k \in \mathcal{A}$, then

$$\bigcup_{i=1}^{k} A_i \in \mathcal{A}.$$

So, $\mathcal{A}$ is closed under finite unions and hence under finite intersections also. To see this note that if $A_1, A_2 \in \mathcal{A}$, then

$$A_1', A_2' \in \mathcal{A} \implies A_1' \cup A_2' \in \mathcal{A} \implies \left(A_1' \cup A_2'\right)' \in \mathcal{A} \implies A_1 \cap A_2 \in \mathcal{A}.$$

This is fine when Ω is a finite set, but we require slightly more to deal with the common situation when Ω is infinite. We require the collection of events to be closed under the operation of taking countable unions, not just finite unions.

Definition 1.2.1 A collection $\mathcal{A}$ of subsets of Ω is called a ***σ*−field** if it satisfies the following conditions:

(I) $\emptyset \in \mathcal{A}$;

(II) if $A_1, A_2, \ldots \in \mathcal{A}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$;

(III) if $A \in \mathcal{A}$ then $A' \in \mathcal{A}$.

To recap, with any experiment we may associate a pair $(\Omega, \mathcal{A})$, where Ω is the set of all possible outcomes (or *elementary events*) and $\mathcal{A}$ is a $\sigma-$field of subsets of Ω, which contains all the events in whose occurrences we may be interested. So, from now on, to call a set A an event is equivalent to asserting that A belongs to the $\sigma-$field in question.

1.3 THE PROBABILITY FUNCTION

Definition 1.3.1 For an event $E \subseteq \Omega$, the **probability that E occurs** will be written $P(E)$.

Interpretation: $P(.)$ is a *set-function* that assigns "weight" to collections of possible outcomes of an experiment. There are many ways to think about precisely how this assignment is achieved;

CLASSICAL : "Consider equally likely sample outcomes ..."

FREQUENTIST : "Consider long-run *relative frequencies* ..."

SUBJECTIVE : "Consider personal degree of belief ..."

or merely think of $P(.)$ as a set-function.

Formally, we have the following definition.

Definition 1.3.2 A **probability function** $P(.)$ on $(\Omega, \mathcal{A})$ is a function $P : \mathcal{A} \to [0, 1]$ satisfying:

(a) $P(\emptyset) = 0, \quad P(\Omega) = 1;$

(b) if $A_1, A_2, \ldots$ is a collection of disjoint members of $\mathcal{A}$, so that $A_i \cap A_j = \emptyset$ from all pairs i, j with $i \neq j$, then

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

The triple $(\Omega, \mathcal{A}, P(.))$, consisting of a set Ω, a σ-field $\mathcal{A}$ of subsets of Ω and a probability function $P(.)$ on $(\Omega, \mathcal{A})$ is called a **probability space**.

1.4 PROPERTIES OF P(.): THE AXIOMS OF PROBABILITY

For events $E, F \subseteq \Omega$

1. $P(E') = 1 - P(E)$.
2. If $E \subseteq F$, then $P(E) \leq P(F)$.
3. In general, $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.
4. $P(E \cap F') = P(E) - P(E \cap F)$
5. $P(E \cup F) \leq P(E) + P(F)$.
6. $P(E \cap F) \geq P(E) + P(F) - 1$.

NOTE : The **general addition rule** 3 for probabilities and **Boole's Inequalities** 5 and 6 extend to more than two events. Let $E_1, ..., E_n$ be events in Ω. Then

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i} P(E_i) - \sum_{i<j} P(E_i \cap E_j) + \sum_{i<j<k} P(E_i \cap E_j \cap E_k) - ... + (-1)^n P\left(\bigcap_{i=1}^{n} E_i\right)$$

and

$$P\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} P(E_i).$$

To prove these results, construct the events $F_1 = E_1$ and

$$F_i = E_i \cap \left(\bigcup_{k=1}^{i-1} E_k\right)'$$

for $i = 2, 3, ..., n$. Then $F_1, F_2, ...F_n$ are disjoint, and $\bigcup_{i=1}^{n} E_i = \bigcup_{i=1}^{n} F_i$, so

$$P\left(\bigcup_{i=1}^{n} E_i\right) = P\left(\bigcup_{i=1}^{n} F_i\right) = \sum_{i=1}^{n} P(F_i).$$

Now, by property 4 above

$$P(F_i) = P(E_i) - P\left(E_i \cap \left(\bigcup_{k=1}^{i-1} E_k\right)\right), \qquad i = 2, 3, ..., n,$$

$$= P(E_i) - P\left(\bigcup_{k=1}^{i-1} (E_i \cap E_k)\right)$$

and the result follows by recursive expansion of the second term for $i = 2, 3, ...n$.

NOTE : We will often deal with both probabilities of single events, and also probabilities for intersection events. For convenience, and to reflect connections with distribution theory that will be presented in Chapter 2, we will use the following terminology; for events E and F

$P(E)$ is the **marginal** probability of E

$P(E \cap F)$ is the **joint** probability of E and F

1.5 CONDITIONAL PROBABILITY

Definition 1.5.1 For events $E, F \subseteq \Omega$ the **conditional probability** that F occurs **given** that E occurs is written $\mathrm{P}(F|E)$, and is defined by

$$P(F|E) = \frac{P(E \cap F)}{P(E)},$$

if $\mathrm{P}(E) > 0$.

NOTE: $P(E \cap F) = P(E)P(F|E)$, and in general, for events $E_1, ..., E_k$,

$$P\left(\bigcap_{i=1}^{k} E_i\right) = P(E_1)P(E_2|E_1)P(E_2|E_1 \cap E_2)...P(E_k|E_1 \cap E_2 \cap ... \cap E_{k-1}).$$

This result is known as the CHAIN or MULTIPLICATION RULE.

Definition 1.5.2 Events E and F are **independent** if

$$P(E|F) = P(E), \text{ so that } P(E \cap F) = P(E)P(F).$$

Extension : Events $E_1, ..., E_k$ are independent if, for **every** subset of events of size $l \leq k$, indexed by $\{i_1, ..., i_l\}$, say,

$$P\left(\bigcap_{j=1}^{l} E_{i_j}\right) = \prod_{j=1}^{l} P(E_{i_j}).$$

1.6 THE THEOREM OF TOTAL PROBABILITY

THEOREM

Let $E_1, ..., E_k$ be a (finite) partition of Ω, and let $F \subseteq \Omega$. Then

$$P(F) = \sum_{i=1}^{k} P(F|E_i)P(E_i).$$

PROOF

$E_1, ..., E_k$ form a partition of Ω, and $F \subseteq \Omega$, so

$$\begin{aligned} F &= (F \cap E_1) \cup ... \cup (F \cap E_k) \\ \implies P(F) &= \sum_{i=1}^{k} P(F \cap E_i) = \sum_{i=1}^{k} P(F|E_i)P(E_i), \end{aligned}$$

writing F as a disjoint union and using the definition of a probability function.

Extension: The theorem still holds if $E_1, E_2, ...$ is a (countably) infinite a partition of Ω, and $F \subseteq \Omega$, so that

$$P(F) = \sum_{i=1}^{\infty} P(F \cap E_i) = \sum_{i=1}^{\infty} P(F|E_i)P(E_i),$$

if $\mathrm{P}(E_i) > 0$ for all i.

1.7 BAYES' THEOREM

THEOREM

Suppose $E, F \subseteq \Omega$, with $P(E), P(F) > 0$. Then

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}.$$

PROOF

$$P(E|F)P(F) = P(E \cap F) = P(F|E)P(E), \text{ so } \quad P(E|F)P(F) = P(F|E)P(E).$$

Extension: If $E_1, ..., E_k$ are disjoint, with $P(E_i) > 0$ for $i = 1, ..., k$, and form a partition of $F \subseteq \Omega$, then

$$P(E_i|F) = \frac{P(F|E_i)P(E_i)}{\sum_{j=1}^{k} P(F|E_j)P(E_j)}.$$

NOTE: in general, $P(E|F) \neq P(F|E)$.

1.8 COUNTING TECHNIQUES

Suppose that an experiment has N equally likely sample outcomes. If event E corresponds to a collection of sample outcomes of size $n(E)$, then

$$P(E) = \frac{n(E)}{N},$$

so it is necessary to be able to evaluate $n(E)$ and N in practice.

1.8.1 THE MULTIPLICATION PRINCIPLE

If operations labelled $1, ..., r$ can be carried out in $n_1, ..., n_r$ ways respectively, then there are

$$\prod_{i=1}^{r} n_i = n_1 ... n_r$$

ways of carrying out the r operations in total.

Example 1.1 If each of r trials of an experiment has N possible outcomes, then there are N^r possible sequences of outcomes in total. For example:
(i) If a multiple choice exam has 20 questions, each of which has 5 possible answers, then there are 5^{20} different ways of completing the exam.
(ii) There are 2^m subsets of m elements (as each element is either **in** the subset, or **not in** the subset, which is equivalent to m trials each with two outcomes).

1.8.2 SAMPLING FROM A FINITE POPULATION

Consider a collection of N items, and a sequence of operations labelled $1, ..., r$ such that the ith operation involves **selecting** one of the items remaining after the first $i-1$ operations have been carried out. Let n_i denote the number of ways of carrying out the ith operation, for $i = 1, ..., r$. Then there are two distinct cases;
(a) **Sampling with replacement :** an item is returned to the collection after selection. Then $n_i = N$ for all i, and there are N^r ways of carrying out the r operations.
(b) **Sampling without replacement :** an item is not returned to the collection after selected. Then $n_i = N - i + 1$, and there are $N(N-1)...(N-r+1)$ ways of carrying out the r operations.
e.g. Consider selecting 5 cards from 52. Then

(a) leads to 52^5 possible selections, whereas

(b) leads to 52.51.50.49.48 possible selections.

NOTE : The **order** in which the operations are carried out may be important
e.g. in a raffle with three prizes and 100 tickets, the draw $\{45, 19, 76\}$ is different from $\{19, 76, 45\}$.

NOTE : The items may be **distinct** (unique in the collection), or **indistinct** (of a unique type in the collection, but not unique individually).
e.g. The numbered balls in the National Lottery, or individual playing cards, are **distinct**. However when balls in the lottery are regarded as "WINNING" or "NOT WINNING", or playing cards are regarded in terms of their suit only, they are **indistinct**.

1.8.3 PERMUTATIONS AND COMBINATIONS

Definition 1.8.1 A **permutation** is an *ordered* arrangement of a set of items.
A **combination** is an *unordered* arrangement of a set of items.

RESULT 1 The number of permutations of n distinct items is $n! = n(n-1)...1$.

RESULT 2 The number of permutations of r from n distinct items is

$$P_r^n = \frac{n!}{(n-r)!} = n(n-1)...(n-r+1) \qquad \text{(by the Multiplication Principle)}.$$

If the **order** in which items are selected is not important, then

RESULT 3 The number of combinations of r from n distinct items is the binomial coefficient

$$C_r^n = \binom{n}{r} = \frac{n!}{r!(n-r)!} \qquad (\text{as} \quad P_r^n = r!C_r^n).$$

-recall the **Binomial Theorem**, namely

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

Then the number of subsets of m items can be calculated as follows; for each $0 \leq j \leq m$, choose a subset of j items from m. Then

$$\text{Total number of subsets} = \sum_{j=0}^{m} \binom{m}{j} = (1+1)^m = 2^m.$$

If the items are **indistinct**, but each is of a unique type, say Type I, ..., Type κ say, (the so-called **Urn Model**) then

RESULT 4 The number of distinguishable permutations of n indistinct objects, comprising n_i items of type i for $i = 1, ..., \kappa$ is

$$\frac{n!}{n_1!n_2!...n_\kappa!}.$$

Special Case : if $\kappa = 2$, then the number of distinguishable permutations of the n_1 objects of type I, and $n_2 = n - n_1$ objects of type II is

$$C_{n_2}^n = \frac{n!}{n_1!(n-n_1)!}.$$

RESULT 5 There are C_r^n ways of partitioning n **distinct** items into two "cells", with r in one cell and $n-r$ in the other.

1.8.4 PROBABILITY CALCULATIONS

Recall that if an experiment has N equally likely sample outcomes, and event E corresponds to a collection of sample outcomes of size $n(E)$, then

$$P(E) = \frac{n(E)}{N}.$$

Example 1.2 A True/False exam has 20 questions. Let $E =$ "16 answers correct at random". Then

$$P(E) = \frac{\text{Number of ways of getting 16 out of 20 correct}}{\text{Total number of ways of answering 20 questions}} = \frac{\binom{20}{16}}{2^{20}} = 0.0046.$$

Example 1.3 *Sampling without replacement.* Consider an Urn Model with 10 Type I objects and 20 Type II objects, and an experiment involving sampling five objects without replacement. Let E="precisely 2 Type I objects selected" We need to calculate N and $n(E)$ in order to calculate P(E). In this case N is the number of ways of choosing 5 from 30 items, and hence

$$N = \binom{30}{5}.$$

To calculate $n(E)$, we think of E occurring by first choosing 2 Type I objects from 10, and then choosing 3 Type II objects from 20, and hence, by the multiplication rule,

$$n(E) = \binom{10}{2}\binom{20}{3}.$$

Therefore

$$P(E) = \frac{\binom{10}{2}\binom{20}{3}}{\binom{30}{5}} = 0.360.$$

This result can be checked using a conditional probability argument; consider event $F \subseteq E$, where $F =$ "sequence of objects 11222 obtained". Then

$$F = \bigcap_{i=1}^{5} F_{ij}$$

where $F_{ij} =$ "type j object obtained on draw i" $i = 1, ..., 5, j = 1, 2$. Then

$$P(F) = P(F_{11})P(F_{21}|F_{11})...P(F_{52}|F_{11}, F_{21}, F_{32}, F_{42}) = \frac{10}{30}\frac{9}{29}\frac{20}{28}\frac{19}{27}\frac{18}{26}.$$

Now consider event G where $G =$ "sequence of objects 12122 obtained". Then

$$P(G) = \frac{10}{30}\frac{20}{29}\frac{9}{28}\frac{19}{27}\frac{18}{26},$$

i.e. $P(G) = P(F)$. In fact, **any** sequence containing two Type I and three Type II objects has this probability, and there are $\binom{5}{2}$ such sequences. Thus, as all such sequences are mutually exclusive,

$$P(E) = \binom{5}{2}\frac{10}{30}\frac{9}{29}\frac{20}{28}\frac{19}{27}\frac{18}{26} = \frac{\binom{10}{2}\binom{20}{3}}{\binom{30}{5}}$$

as before.

Example 1.4 *Sampling with replacement.* Consider an Urn Model with 10 Type I objects and 20 Type II objects, and an experiment involving sampling five objects with replacement. Let $E =$ "precisely 2 Type I objects selected". Again, we need to calculate N and $n(E)$ in order to calculate P(E). In this case N is the number of ways of choosing 5 from 30 items with replacement, and hence

$$N = 30^5.$$

To calculate $n(E)$, we think of E occurring by first choosing 2 Type I objects from 10, and 3 Type II objects from 20 in any order. Consider such sequences of selection

Sequence	Number of ways
11222	10.10.20.20.20
12122	10.20.10.20.20
.	.

etc., and thus a sequence with 2 Type I objects and 3 Type II objects can be obtained in $10^2 20^3$ ways. As before there are $\binom{5}{2}$ such sequences, and thus

$$P(E) = \frac{\binom{5}{2}10^2 20^3}{30^5} = 0.329.$$

Again, this result can be verified using a conditional probability argument; consider event $F \subseteq E$, where $F =$ "sequence of objects 11222 obtained". Then

$$P(F) = \left(\frac{10}{30}\right)^2 \left(\frac{20}{30}\right)^3$$

as the results of the draws are **independent**. This result is true for any sequence containing two Type I and three Type II objects, and there are $\binom{5}{2}$ such sequences that are mutually exclusive, so

$$P(E) = \binom{5}{2}\left(\frac{10}{30}\right)^2 \left(\frac{20}{30}\right)^3.$$

CHAPTER 2

RANDOM VARIABLES & PROBABILITY DISTRIBUTIONS

This chapter contains the introduction of random variables as a technical device to enable the general specification of probability distributions in one and many dimensions to be made. The key topics and techniques introduced in this chapter include the following:

- **EXPECTATION**
- **TRANSFORMATION**
- **STANDARDIZATION**
- **GENERATING FUNCTIONS**
- **JOINT MODELLING**
- **MARGINALIZATION**
- **MULTIVARIATE TRANSFORMATION**
- **MULTIVARIATE EXPECTATION & COVARIANCE**
- **SUMS OF VARIABLES**

Of key importance is the **moment generating function**, which is a standard device for identification of **probability distributions**. **Transformations** are often used to transform a **random variable** or **statistic** of interest to one of simpler form, whose probability distribution is more convenient to work with. **Standardization** is often a key part of such simplification.

2.1 RANDOM VARIABLES & PROBABILITY MODELS

We are not always interested in an experiment itself, but rather in some consequence of its random outcome. Such consequences, when real valued, may be thought of as functions which map Ω to $\mathbb{R}$, and these functions are called random variables.

Definition 2.1.1 A **random variable** (**r.v.**) X is a function $X : \Omega \to \mathbb{R}$ with the property that $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{A}$ for each $x \in \mathbb{R}$.

The point is that we defined the probability function $P(.)$ on the $\sigma-$field $\mathcal{A}$, so if $A(x) = \{\omega \in \Omega : X(\omega) \leq x\}$, we cannot discuss $P(A(x))$ unless $A(x)$ belongs to $\mathcal{A}$. We generally pay no attention to the technical condition in the definition, and just think of random variables as functions mapping Ω to $\mathbb{R}$.

So, we regard a set $B \subseteq \mathbb{R}$ as an event, associated with event $A \subseteq \Omega$ if

$$A = \{\omega : X(\omega) = x \text{ for some } x \in B\}.$$

A and B are events in **different** spaces, but are equivalent in the sense that

$$P(X \in B) = P(A),$$

where, formally, it is the latter quantity that is defined by the probability function. Attention switches to assigning the probability $P(X \in B)$ for appropriate sets $B \subseteq \mathbb{R}$.

If Ω is a list of discrete elements $\Omega = \{\omega_1, \omega_2, ...\}$, then the definition indicates that the events of interest will be of the form $[X = b]$, or equivalently of the form $[X \leq b]$ for $b \in \mathbb{R}$. For more general sample spaces, we will concentrate on events of the form $[X \leq b]$ for $b \in \mathbb{R}$.

2.2 DISCRETE RANDOM VARIABLES

Definition 2.2.1 A random variable X is **discrete** if the set of all possible values of X (that is, the *range* of the function represented by X), denoted $\mathbb{X}$, is **countable**, that is

$$\mathbb{X} = \{x_1, x_2, ..., x_n\} \quad \text{[FINITE]} \quad \text{or} \quad \mathbb{X} = \{x_1, x_2, ...\} \quad \text{[INFINITE]}.$$

Definition 2.2.2 PROBABILITY MASS FUNCTION

The function f_X defined on $\mathbb{X}$ by

$$f_X(x) = P[X = x], \qquad x \in \mathbb{X}$$

that assigns probability to each $x \in \mathbb{X}$ is the (discrete) **probability mass function**, or **pmf**.

NOTE: For completeness, we define

$$f_X(x) = 0, \qquad x \notin \mathbb{X},$$

so that f_X is defined for all $x \in \mathbb{R}$ Furthermore we will refer to $\mathbb{X}$ as the *support* of random variable X, that is, the set of $x \in \mathbb{R}$ such that $f_X(x) > 0$.

2.2.1 PROPERTIES OF MASS FUNCTION f_X

Elementary properties of the mass function are straightforward to establish using properties of the probability function. A function f_X is a probability mass function for discrete random variable X with range $\mathbb{X}$ of the form $\{x_1, x_2, ...\}$ if and only if

$$\text{(i) } f_X(x_i) \geq 0, \qquad \text{(ii) } \sum f_X(x_i) = 1.$$

These results follow as events $[X = x_1], [X = x_2]$ etc. are **equivalent** to events that partition Ω, that is, $[X = x_i]$ is equivalent to event A_i hence $P[X = x_i] = P(A_i)$, and the two parts of the theorem follow immediately.

Definition 2.2.3 DISCRETE CUMULATIVE DISTRIBUTION FUNCTION

The **cumulative distribution function**, or **cdf**, F_X of a discrete r.v. X is defined by

$$F_X(x) = P[X \leq x], \qquad x \in \mathbb{R}.$$

2.2.2 CONNECTION BETWEEN F_X AND f_X

Let X be a discrete random variable with range $\mathbb{X} = \{x_1, x_2, ...\}$, where $x_1 < x_2 < ...$, and probability mass function f_X and cdf F_X. Then for any real value x, if $x < x_1$, then $F_X(x) = 0$, and for $x \geq x_1$,

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i) \qquad \Longleftrightarrow \qquad f_X(x_i) = F_X(x_i) - F_X(x_{i-1}) \qquad i = 2, 3, ...$$

with, for completeness, $f_X(x_1) = F_X(x_1)$. These relationships follow as events of the form $[X \leq x_i]$ can be represented as countable unions of the events A_i. The first result therefore follows from properties of the probability function. The second result follows immediately.

2.2.3 PROPERTIES OF DISCRETE CDF F_X

(i) In the limiting cases,

$$\lim_{x \to -\infty} F_X(x) = 0, \qquad \lim_{x \to \infty} F_X(x) = 1.$$

(ii) F_X is **continuous from the right** (but not continuous) on $\mathbb{R}$ that is, for $x \in \mathbb{R}$,

$$\lim_{h \to 0^+} F_X(x + h) = F_X(x).$$

(iii) F_X is **non-decreasing**, that is

$$a < b \Longrightarrow F_X(a) \leq F_X(b).$$

(iv) For $a < b$,

$$P[a < X \leq b] = F_X(b) - F_X(a).$$

The key idea is that the functions f_X and/or F_X can be used to describe the **probability distribution** of the random variable X. A graph of the function f_X is non-zero only at the elements of $\mathbb{X}$. A graph of the function F_X is a **step-function** which takes the value zero at minus infinity, the value one at infinity, and is non-decreasing with points of discontinuity at the elements of $\mathbb{X}$.

2.3 CONTINUOUS RANDOM VARIABLES

Definition 2.3.1 A random variable X is <u>continuous</u> if the function F_X defined on $\mathbb{R}$ by

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

for $x \in \mathbb{R}$ is a **continuous** function on $\mathbb{R}$, that is, for $x \in \mathbb{R}$,

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

Definition 2.3.2 CONTINUOUS CUMULATIVE DISTRIBUTION FUNCTION
The <u>cumulative distribution function</u>, or **cdf**, F_X of a continuous r.v. X is defined by

$$F_X(x) = P[X \leq x], \qquad x \in \mathbb{R}.$$

Definition 2.3.3 PROBABILITY DENSITY FUNCTION
A random variable is **absolutely continuous** if the cumulative distribution function F_X can be written

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

for some function f_X, termed the **probability density function**, or **pdf**, of X.

From now on when we speak of a continuous random variable, we will implicitly assume the absolutely continuous case, where a pdf exists.

2.3.1 PROPERTIES OF CONTINUOUS F_X AND f_X

By analogy with the discrete case, let $\mathbb{X}$ be the range of X, so that $\mathbb{X} = \{x : f_X(x) > 0\}$.

(i) The pdf f_X need not exist, but as indicated above, continuous r.v.'s where a pdf f_X cannot be defined in this way will be ignored. The function f_X can be defined piecewise on intervals of $\mathbb{R}$.

(ii) For the cdf of a continuous r.v.,

$$\lim_{x\to-\infty} F_X(x) = 0, \qquad \lim_{x\to\infty} F_X(x) = 1.$$

(iii) Directly from the definition, at values of x where F_X is differentiable,

$$f_X(x) = \frac{d}{dt}\left\{F_X(t)\right\}_{t=x}.$$

(iv) If X is continuous,

$$f_X(x) \neq P[X = x] = \lim_{h\to0^+}[P(X \leq x) - P(X \leq x - h)] = \lim_{h\to0^+}[F_X(x) - F_X(x - h)] = 0.$$

(v) For $a < b$,

$$P[a < X \leq b] = P[a \leq X < b] = P[a \leq X \leq b] = P[a < X < b] = F_X(b) - F_X(a).$$

It follows that a function f_X is a pdf for a continuous random variable X **if and only if**

$$\text{(i) } f_X(x) \geq 0, \qquad \text{(ii)} \int_{-\infty}^{\infty} f_X(x)dx = 1.$$

This result follows direct from definitions and properties of F_X.

Example 2.1 Consider a coin tossing experiment where a fair coin is tossed repeatedly under identical experimental conditions, with the sequence of tosses independent, until a Head is obtained. For this experiment, the sample space, Ω is then the set of sequences $(\{H\}, \{TH\}, \{TTH\}, \{TTTH\}...)$ with associated probabilities 1/2, 1/4, 1/8, 1/16, ...

Define discrete random variable $X : \Omega \longrightarrow \mathbb{R}$, by $X(\omega) = x \Longleftrightarrow$ first H on toss x. Then

$$f_X(x) = P[X = x] = \left(\frac{1}{2}\right)^x, \qquad x = 1, 2, 3, ...$$

and zero otherwise. For $x \geq 1$, let $k(x)$ be the largest integer not greater than x, then

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i) = \sum_{i=1}^{k(x)} f_X(i) = 1 - \left(\frac{1}{2}\right)^{k(x)}$$

and $F_X(x) = 0$ for $x < 1$.

Graphs of the probability mass function (left) and cumulative distribution function (right) are shown in Figure 2.1. Note that the mass function is only non-zero at points that are elements of $\mathbb{X}$, and that the cdf is defined for all real values of x, but is only continuous from the right. F_X is therefore a step-function.

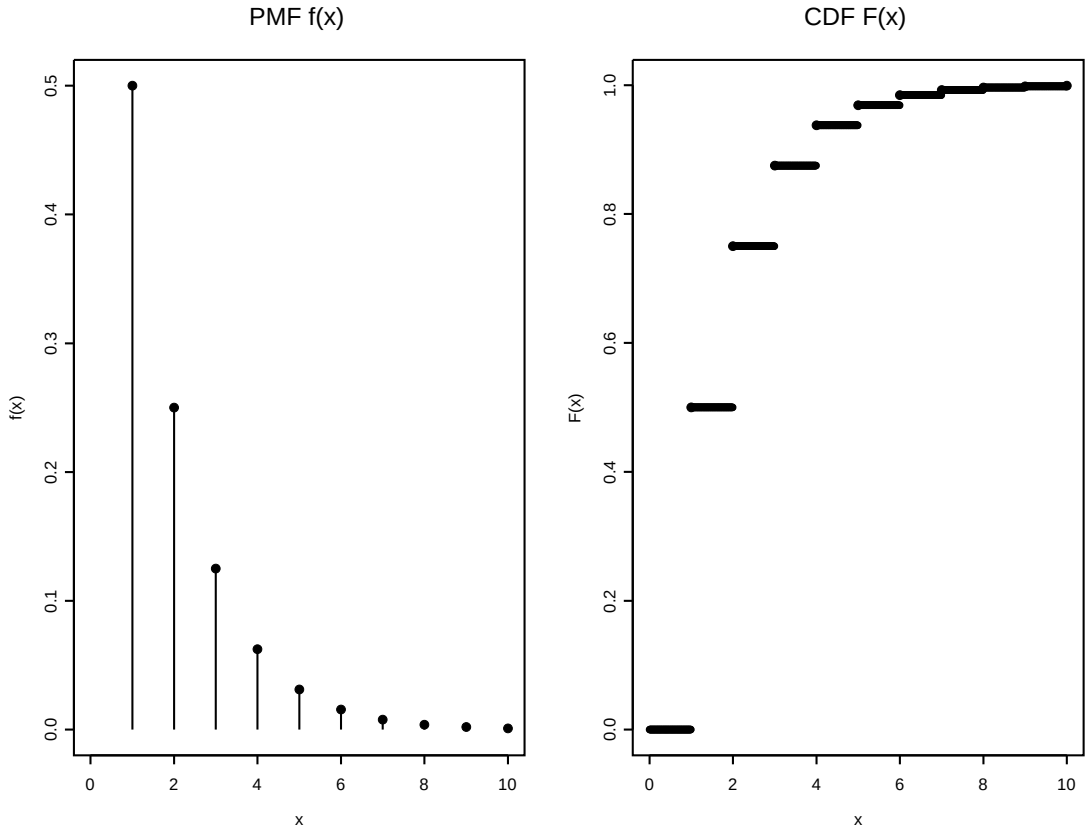

Figure 2.1: PMF $f_X(x) = \left(\frac{1}{2}\right)^x, x = 1, 2, \ldots,$ and CDF $F_X(x) = 1 - \left(\frac{1}{2}\right)^{k(x)}$.

Example 2.2 Consider an experiment to measure the length of time that an electrical component functions before failure. The sample space of outcomes of the experiment, Ω is $\mathbb{R}^+$, and if A_x is the event that the component functions for longer than $x > 0$ time units, suppose that $\mathrm{P}(A_x) = \exp\left\{-x^2\right\}$.

Define continuous random variable $X : \Omega \longrightarrow \mathbb{R}^+$, by $X(\omega) = x \Longleftrightarrow$ component fails at time x. Then, if $x > 0$,

$$F_X(x) = P[X \leq x] = 1 - P(A_x) = 1 - \exp\left\{-x^2\right\}$$

and $F_X(x) = 0$ if $x \leq 0$. Hence if $x > 0$,

$$f_X(x) = \frac{d}{dt}\left\{F_X(t)\right\}_{t=x} = 2x \exp\left\{-x^2\right\},$$

and zero otherwise.

Graphs of the probability density function (left) and cumulative distribution function (right) are shown in Figure 2.2. Note that both the pdf and cdf are defined for all real values of x, and that both are continuous functions.

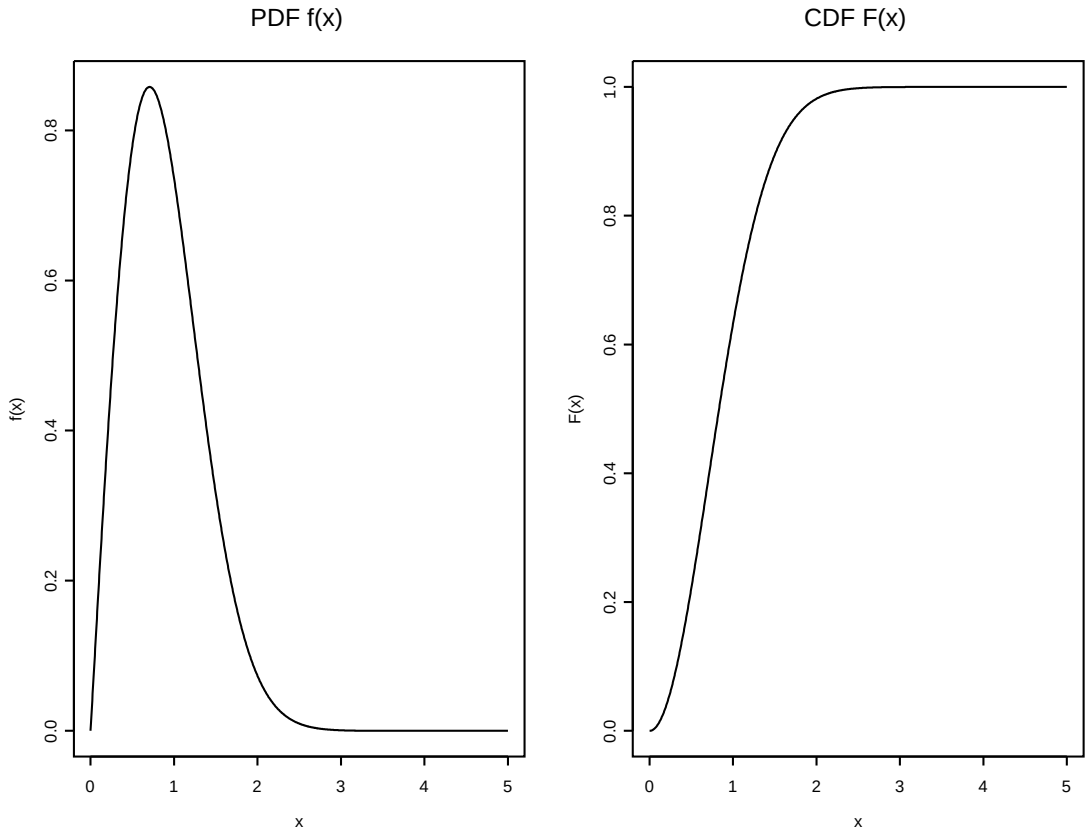

Figure 2.2: PDF $f_X(x) = 2x \exp\{-x^2\}, x > 0$, and CDF $F_X(x) = 1 - \exp\{-x^2\}, x > 0$.

Note that here

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \int_{0}^{x} f_X(t)dt$$

as $f_X(x) = 0$ for $x \leq 0$, and also that

$$\int_{-\infty}^{\infty} f_X(x)dx = \int_{0}^{\infty} f_X(x)dx = 1.$$

2.4 EXPECTATIONS AND THEIR PROPERTIES

Definition 2.4.1 For a discrete random variable X with range $\mathbb{X}$ and with probability mass function f_X, the **expectation** or **expected value** of X with respect to f_X is defined by

$$E_{f_X}[X] = \sum_{x=-\infty}^{\infty} x f_X(x) = \sum_{x \in \mathbb{X}} x f_X(x).$$

For a continuous random variable X with range $\mathbb{X}$ and pdf f_X, the **expectation** or **expected value** of X with respect to f_X is defined by

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

NOTE : The sum/integral may not be convergent, and hence the expected value may be infinite. It is important always to check that the integral is finite: a sufficient condition is the *absolute integrability* of the summand/integrand, that is

$$\sum_x |x| \, f_X(x) < \infty \Longrightarrow \sum_x x f_X(x) = E_{f_X}[X] < \infty,$$

or in the continuous case

$$\int_{-\infty}^{\infty} |x| \, f_X(x) dx < \infty \Longrightarrow \int_{-\infty}^{\infty} x f_X(x) dx = E_{f_X}[X] < \infty.$$

Extension : Let g be a real-valued function whose domain includes $\mathbb{X}$. Then

$$E_{f_X}[g(X)] = \begin{cases} \displaystyle\sum_{x \in \mathbb{X}} g(x) f_X(x), & \text{if } X \text{ is discrete,} \\ \displaystyle\int_{\mathbb{X}} g(x) f_X(x) dx, & \text{if } X \text{ is continuous.} \end{cases}$$

PROPERTIES OF EXPECTATIONS

Let X be a random variable with mass function/pdf f_X. Let g and h be real-valued functions whose domains include $\mathbb{X}$, and let a and b be constants. Then

$$E_{f_X}[ag(X) + bh(X)] = aE_{f_X}[g(X)] + bE_{f_X}[h(X)],$$

as (in the continuous case)

$$\begin{aligned} E_{f_X}[ag(X) + bh(X)] &= \int_{\mathbb{X}} [ag(x) + bh(x)] f_X(x) dx \\ &= a \int_{\mathbb{X}} g(x) f_X(x) dx + b \int_{\mathbb{X}} h(x) f_X(x) dx \\ &= aE_{f_X}[g(X)] + bE_{f_X}[h(X)]. \end{aligned}$$

SPECIAL CASES :

(i) For a simple linear function

$$E_{f_X}[aX+b] = aE_{f_X}[X] + b.$$

(ii) Consider $g(x) = (x - \mathrm{E}_{f_X}[X])^2$. Write $\mu = \mathrm{E}_{f_X}[X]$ (a constant that does not depend on x). Then, expanding the integrand

$$\begin{aligned} E_{f_X}[g(X)] &= \int (x-\mu)^2 f_X(x)dx = \int x^2 f_X(x)dx - 2\mu\int x f_X(x)dx + \mu^2\int f_X(x)dx \\ &= \int x^2 f_X(x)dx - 2\mu^2 + \mu^2 = \int x^2 f_X(x)dx - \mu^2 \\ &= E_{f_X}[X^2] - \{E_{f_X}[X]\}^2 . \end{aligned}$$

Then

$$Var_{f_X}[X] = E_{f_X}[X^2] - \{E_{f_X}[X]\}^2$$

is the **variance** of the distribution. Similarly, $\sqrt{Var_{f_X}[X]}$ is the **standard deviation** of the distribution.

(iii) Consider $g(x) = x^k$ for $k = 1, 2, ...$ Then in the continuous case

$$E_{f_X}[g(X)] = E_{f_X}[X^k] = \int_{\mathbb{X}} x^k f_X(x)dx,$$

and $\mathrm{E}_{f_X}[X^k]$ is the kth **moment** of the distribution.

(iv) Consider $g(x) = (x-\mu)^k$ for $k = 1, 2, ...$ Then

$$E_{f_X}[g(X)] = E_{f_X}[(X-\mu)^k] = \int_{\mathbb{X}} (x-\mu)^k f_X(x)dx,$$

and $\mathrm{E}_{f_X}[(X-\mu)^k]$ is the kth **central moment** of the distribution.

(v) Consider $g(x) = aX + b$. Then $Var_{f_X}[aX+b] = a^2 Var_{f_X}[X]$,

$$\begin{aligned} Var_{f_X}[g(X)] &= E_{f_X}[(aX+b-E_{f_X}[aX+b])^2] \\ &= E_{f_X}[(aX+b-aE_{f_X}[X]-b)^2] \\ &= E_{f_X}[(a^2(X-E_{f_X}[X])^2] \\ &= a^2 Var_{f_X}[X]. \end{aligned}$$

2.5 INDICATOR VARIABLES

A particular class of random variables called **indicator variables** are particularly useful. Let A be an event and let $I_A : \Omega \to \mathbb{R}$ be the indicator function of A, so that

$$I_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \in A'. \end{cases}$$

Then I_A is a random variable taking values 1 and 0 with probabilities $P(A)$ and $P(A')$ respectively. Also, I_A has expectation $P(A)$ and variance $P(A)\{1 - P(A)\}$. The usefulness lies in the fact that **any** discrete random variable X can be written as a linear combination of indicator random variables:

$$X = \sum_i a_i I_{A_i},$$

for some collection of events $(A_i, i \geq 1)$ and real numbers $(a_i, i \geq 1)$. Sometimes we can obtain the expectation and variance of a random variable X easily by expressing it in this way, then using knowledge of the expectation and variance of the indicator variables I_{A_i}, rather than by direct calculation.

2.6 TRANSFORMATIONS OF RANDOM VARIABLES

2.6.1 GENERAL TRANSFORMATIONS

Consider a discrete/continuous r.v. X with range $\mathbb{X}$ and probability distribution described by mass/pdf f_X, or cdf F_X. Suppose g is a real-valued function defined on $\mathbb{X}$. Then $Y = g(X)$ is also an r.v. (Y is also a function from Ω to $\mathbb{R}$). Denote the range of Y by $\mathbb{Y}$. For $A \subseteq \mathbb{R}$, the event $[Y \in A]$ is an event in terms of the transformed variable Y. If f_Y is the mass/density function for Y, then

$$P[Y \in A] = \begin{cases} \displaystyle\sum_{y \in A} f_Y(y), & Y \text{ discrete,} \\ \displaystyle\int_A f_Y(y)dy, & Y \text{ continuous.} \end{cases}$$

We wish to derive the probability distribution of random variable Y; in order to do this, we first consider the inverse transformation g^{-1} from $\mathbb{Y}$ to $\mathbb{X}$ defined for set $A \subseteq \mathbb{Y}$ (and for $y \in \mathbb{Y}$) by

$$g^{-1}(A) = \{x \in \mathbb{X} : g(x) \in A\}, \qquad g^{-1}(y) = \{x \in \mathbb{X} : g(x) = y\},$$

that is, $g^{-1}(A)$ is the set of points in $\mathbb{X}$ that map into A, and $g^{-1}(y)$ is the set of points in $\mathbb{X}$ that map to y, under transformation g. By construction, we have

$$P[Y \in A] = P[X \in g^{-1}(A)].$$

Then, for $y \in \mathbb{R}$, we have

$$F_Y(y) = P[Y \leq y] = P[g(X) \leq y] = \begin{cases} \displaystyle\sum_{x \in A_y} f_X(x), & X \text{ discrete,} \\ \displaystyle\int_{A_y} f_X(x)dx, & X \text{ continuous,} \end{cases}$$

where $A_y = \{x \in \mathbb{X} : g(x) \leq y\}$. This result gives the "*first principles*" approach to computing the distribution of the new variable. The approach can be summarized as follows:

- consider the range $\mathbb{Y}$ of the new variable;
- consider the cdf $F_Y(y)$. Step through the argument as follows

$$F_Y(y) \quad = P[Y \leq y] = P[g(X) \leq y] = P[X \in A_y].$$

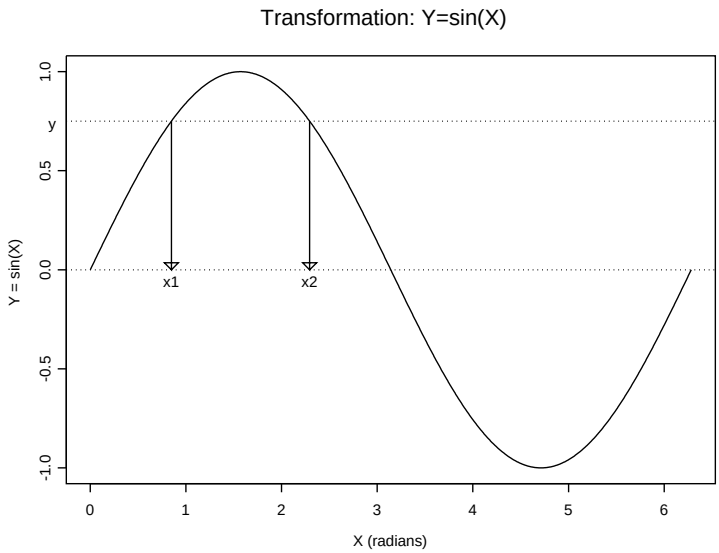

Figure 2.3: Computation of A_y for $Y = \sin X$.

Note that it is usually a good idea to start with the cdf, not the pmf or pdf.

Our main objective is therefore to identify the set

$$A_y = \{x \in \mathbb{X} : g(x) \leq y\}.$$

Example 2.3 Suppose that X is a continuous r.v. with range $\mathbb{X} \equiv (0, 2\pi)$ whose pdf f_X is constant

$$f_X(x) = \frac{1}{2\pi}, \qquad 0 < x < 2\pi,$$

and zero otherwise. This pdf has corresponding continuous cdf

$$F_X(x) = \frac{x}{2\pi}, \qquad 0 < x < 2\pi.$$

Consider the transformed r.v. $Y = \sin X$. Then the range of Y, $\mathbb{Y}$, is $[-1, 1]$, but the transformation is not 1-1. However, from first principles, we have

$$F_Y(y) = P[Y \leq y] = P[\sin X \leq y].$$

Now, by inspection of Figure 2.3, we can easily identify the required set $A_y, y > 0$: it is the union of two disjoint intervals

$$A_y = [0, x_1] \cup [x_2, 2\pi] = \left[0, \sin^{-1} y\right] \cup \left[\pi - \sin^{-1} y, 2\pi\right].$$

Hence

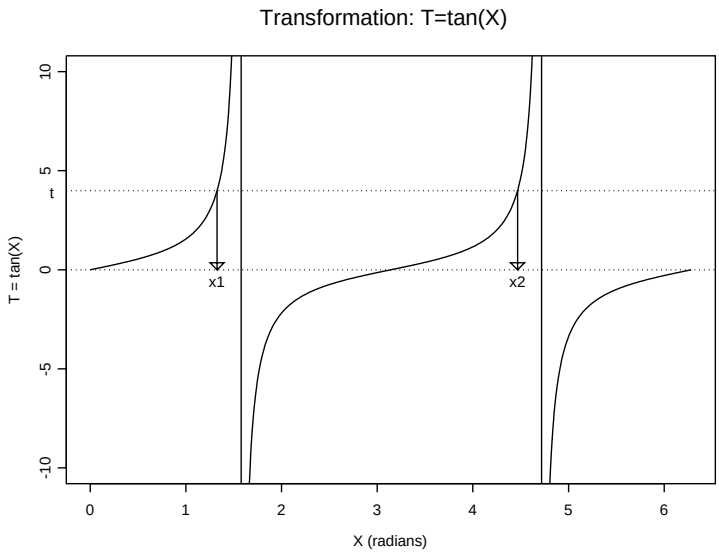

Figure 2.4: Computation of A_y for $T = \tan X$.

$$F_Y(y) = P[\sin X \le y] = P[X \le x_1] + P[X \ge x_2] = \{P[X \le x_1]\} + \{1 - P[X < x_2]\}$$

$$= \left\{\frac{1}{2\pi}\sin^{-1} y\right\} + \left\{1 - \frac{1}{2\pi}\left(\pi - \sin^{-1} y\right)\right\} = \frac{1}{2} + \frac{1}{\pi}\sin^{-1} y,$$

and hence, by differentiation,

$$f_Y(y) = \frac{1}{\pi}\frac{1}{\sqrt{1-y^2}}.$$

[A symmetry argument verifies this for $y < 0$.]

Example 2.4 Consider transformed r.v. $T = \tan X$. Then the range of T, $\mathbb{T}$, is $\mathbb{R}$, but the transformation is not 1-1. However, from first principles, we have, for $t > 0$,

$$F_T(t) = P[T \le t] = P[\tan X \le t].$$

Figure 2.4 helps identify the required set A_t: in this case it is the union of three disjoint intervals

$$A_t = [0, x_1] \cup \left(\frac{\pi}{2}, x_2\right] \cup \left(\frac{3\pi}{2}, 2\pi\right] = \left[0, \tan^{-1} t\right] \cup \left(\frac{\pi}{2}, \pi + \tan^{-1} t\right] \cup \left(\frac{3\pi}{2}, 2\pi\right],$$

(note, for values of $t < 0$, the union will be of only two intervals, but the calculation proceeds identically). Therefore,

$$F_T(t) = P[\tan X \le t] = P[X \le x_1] + P\left[\frac{\pi}{2} < X \le x_2\right] + P\left[\frac{3\pi}{2} < X \le 2\pi\right]$$

$$= \left\{\frac{1}{2\pi}\tan^{-1} t\right\} + \frac{1}{2\pi}\left\{\pi + \tan^{-1} t - \frac{\pi}{2}\right\} + \frac{1}{2\pi}\left\{2\pi - \frac{3\pi}{2}\right\} = \frac{1}{\pi}\tan^{-1} t + \frac{1}{2},$$

and hence, by differentiation,

$$f_T(t) = \frac{1}{\pi}\frac{1}{1+t^2}.$$

2.6.2 1-1 TRANSFORMATIONS

The mapping $g(X)$, a function of X from $\mathbb{X}$, is 1-1 and onto $\mathbb{Y}$ if for each $y \in \mathbb{Y}$, there exists one and only one $x \in \mathbb{X}$ such that $y = g(x)$.

The following theorem gives the distribution for random variable $Y = g(X)$ when g is 1-1.

Theorem 2.6.1 ***THE UNIVARIATE TRANSFORMATION THEOREM***

Let X be a random variable with mass/density function f_X and support $\mathbb{X}$. Let g be a 1-1 function from $\mathbb{X}$ onto $\mathbb{Y}$ with inverse g^{-1}. Then $Y = g(X)$ is a random variable with support $\mathbb{Y}$ and

***Discrete Case** : The mass function of random variable Y is given by*

$$f_Y(y) = f_X(g^{-1}(y)), \quad y \in \mathbb{Y} = \{y | f_Y(y) > 0\},$$

where x is the unique solution of $y = g(x)$ (so that $x = g^{-1}(y)$).

***Continuous Case** : The pdf of random variable Y is given by*

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dt}\left\{g^{-1}(t)\right\}_{t=y} \right|, \qquad y \in \mathbb{Y} = \{y | f_Y(y) > 0\},$$

where $y = g(x)$, provided that the derivative

$$\frac{d}{dt}\left\{g^{-1}(t)\right\}$$

is continuous and non-zero on $\mathbb{Y}$.

Proof. Discrete case : by direct calculation,

$$f_Y(y) = P[Y = y] = P[g(X) = y] = P[X = g^{-1}(y)] = f_X(x)$$

where $x = g^{-1}(y)$, and hence $f_Y(y) > 0 \Longleftrightarrow f_X(x) > 0$.

Continuous case : function g is either (I) a monotonic increasing, or (II) a monotonic decreasing function.

Case (I): If g is increasing, then for $x \in \mathbb{X}$ and $y \in \mathbb{Y}$, we have that

$$g(x) \leq y \Longleftrightarrow x \leq g^{-1}(y).$$

Therefore, for $y \in \mathbb{Y}$,

$$F_Y(y) = P[Y \leq y] = P[g(X) \leq y] = P[X \leq g^{-1}(y)] = F_X(g^{-1}(y))$$

and, by differentiation, because g is monotonic increasing,

$$f_Y(y) = f_X(g^{-1}(y))\frac{d}{dt}\left\{g^{-1}(t)\right\}_{t=y} = f_X(g^{-1}(y))\left|\frac{d}{dt}\left\{g^{-1}(y)\right\}_{t=y}\right|, \quad \text{as} \quad \frac{d}{dt}\left\{g^{-1}(t)\right\} > 0.$$

Case (II): If g is decreasing, then for $x \in \mathbb{X}$ and $y \in \mathbb{Y}$ we have

$$g(x) \leq y \Longleftrightarrow x \geq g^{-1}(y)$$

Therefore, for $y \in \mathbb{Y}$,

$$F_Y(y) = P[Y \leq y] = P[g(X) \leq y] = P[X \geq g^{-1}(y)] = 1 - F_X(g^{-1}(y)),$$

so

$$f_Y(y) = -f_X(g^{-1}(y))\frac{d}{dt}\left\{g^{-1}(y)\right\} = f_X(g^{-1}(y))\left|\frac{d}{dt}\left\{g^{-1}(t)\right\}_{t=y}\right| \quad \text{as} \quad \frac{d}{dt}\left\{g^{-1}(t)\right\} < 0.$$

Definition 2.6.1 Suppose transformation $g : \mathbb{X} \longrightarrow \mathbb{Y}$ is 1-1, and is defined by $g(x) = y$ for $x \in \mathbb{X}$. Then the **Jacobian** of the transformation, denoted $J(y)$, is given by

$$J(y) = \frac{d}{dt}\left\{g^{-1}(t)\right\}_{t=y},$$

that is, the first derivative of g^{-1} evaluated at $y = g(x)$. Note that the inverse transformation $g^{-1} : \mathbb{Y} \longrightarrow \mathbb{X}$ has Jacobian $1/J(x)$.

NOTE :

(i) The Jacobian is precisely the same term that appears as a change of variable term in an integration.

(ii) In the Univariate Transformation Theorem, in the continuous case, we take the **modulus** of the Jacobian

(iii) To compute the expectation of $Y = g(X)$, we now have two alternative methods of computation; we either compute the expectation of $g(X)$ with respect to the distribution of X, or compute the distribution of Y, and then its expectation. It is straightforward to demonstrate that the two methods are equivalent, that is

$$E_{f_X}[g(X)] = E_{f_Y}[Y]$$

This result is sometimes known as the *Law of the Unconscious Statistician.*

IMPORTANT NOTE: Note that the apparently appealing "plug-in" approach that sets
$$f_Y(y) = f_X\left(g^{-1}(y)\right)$$
will almost always fail as the Jacobian term **must** be included. For example, if $Y = e^X$ so that $X = \log Y$, then merely setting
$$f_Y(y) = f_X(\log y)$$
is **insufficient**, you **must** have
$$f_Y(y) = f_X(\log y) \times \frac{1}{y}.$$

2.7 GENERATING FUNCTIONS

2.7.1 MOMENT GENERATING FUNCTIONS

Definition 2.7.1 For random variable X with mass/density function f_X, the **moment generating function**, or mgf, of X, M_X, is defined by

$$M_X(t) = E_{f_X}[e^{tX}],$$

if this expectation exists for all values of $t \in (-h, h)$ for some $h > 0$, that is,

$$\text{DISCRETE CASE} \qquad M_X(t) = \sum e^{tx} f_X(x)$$

$$\text{CONTINUOUS CASE} \quad M_X(t) = \int e^{tx} f_X(x) dx$$

where the sum/integral is over $\mathbb{X}$.

NOTE : It can be shown that if X_1 and X_2 are random variables taking values on $\mathbb{X}$ with mass/density functions f_{X_1} and f_{X_2}, and mgfs M_{X_1} and M_{X_2} respectively, then

$$f_{X_1}(x) \equiv f_{X_2}(x), x \in \mathbb{X} \Longleftrightarrow M_{X_1}(t) \equiv M_{X_2}(t),\ t \in (-h, h).$$

Hence there is a **1-1 correspondence between generating functions and distributions**: this provides a key technique for identification of probability distributions.

2.7.2 KEY PROPERTIES OF MGFS

(i) If X is a discrete random variable, the rth derivative of M_X evaluated at t, $M_X^{(r)}(t)$, is given by

$$M_X^{(r)}(t) = \frac{d^r}{ds^r}\left\{M_X(s)\right\}_{s=t} = \frac{d^r}{ds^r}\left\{\sum e^{sx} f_X(x)\right\}_{s=t} = \sum x^r e^{tx} f_X(x)$$

and hence

$$M_X^{(r)}(0) = \sum x^r f_X(x) = E_{f_X}[X^r].$$

If X is a continuous random variable, the rth derivative of M_X is given by

$$M_X^{(r)}(t) = \frac{d^r}{ds^r}\left\{\int e^{sx} f_X(x)dx\right\}_{s=t} = \int x^r e^{tx} f_X(x)dx$$

and hence

$$M_X^{(r)}(0) = \int x^r f_X(x)dx = E_{f_X}[X^r].$$

(ii) If X is a discrete random variable, then

$$\begin{aligned} M_X(t) &= \sum e^{tx} f_X(x) = \sum \left\{\sum_{r=0}^{\infty} \frac{(tx)^r}{r!}\right\} f_X(x) \\ &= 1 + \sum_{r=1}^{\infty} \frac{t^r}{r!}\left\{\sum x^r f_X(x)\right\} = 1 + \sum_{r=1}^{\infty} \frac{t^r}{r!} E_{f_X}[X^r]. \end{aligned}$$

The identical result holds for the continuous case.

(iii) From the general result for expectations of functions of random variables,

$$E_{f_Y}[e^{tY}] \equiv E_{f_X}[e^{t(aX+b)}] \Longrightarrow M_Y(t) = E_{f_X}[e^{t(aX+b)}] = e^{bt} E_{f_X}[e^{atX}] = e^{bt} M_X(at).$$

Therefore, if

$$Y = aX + b, M_Y(t) = e^{bt} M_X(at)$$

Theorem 2.7.1 *Let $X_1, ..., X_k$ be independent random variables with mgfs $M_{X_1}, ..., M_{X_k}$ respectively. Then if the random variable Y is defined by $Y = X_1 + ... + X_k$,*

$$M_Y(t) = \prod_{i=1}^{k} M_{X_i}(t).$$

Proof. For $k = 2$, if X_1 and X_2 are independent, integer-valued, discrete r.v.s, then if $Y = X_1 + X_2$, by the **Theorem of Total Probability**,

$$f_Y(y) = P[Y = y] = \sum_{x_1} P[Y = y | X_1 = x_1] P[X_1 = x_1] = \sum_{x_1} f_{X_2}(y - x_1) f_{X_2}(x_1).$$

Hence

$$\begin{aligned} M_Y(t) &= E_{f_Y}[e^{tY}] = \sum_y e^{ty} f_Y(y) = \sum_y e^{ty} \left\{\sum_{x_1} f_{X_2}(y - x_1) f_{X_2}(x_1)\right\} \\ &= \sum_{x_2} e^{t(x_1+x_2)} \left\{\sum_{x_1} f_{X_2}(x_2) f_{X_2}(x_1)\right\} \qquad \text{(changing variables in the summation, } x_2 = y - x_1) \\ &= \left\{\sum_{x_1} e^{tx_1} f_{X_1}(x_1)\right\} \left\{\sum_{x_2} e^{tx_2} f_{X_2}(x_2)\right\} = M_{X_1}(t) M_{X_2}(t), \end{aligned}$$

and the result follows for general k by recursion.

The result for continuous random variables follows in the obvious way.

Special Case : If $X_1, ..., X_k$ are identically distributed, then $M_{X_i}(t) \equiv M_X(t)$, say, for all i, so

$$M_Y(t) = \prod_{i=1}^{k} M_X(t) = \{M_X(t)\}^k .$$

2.7.3 OTHER GENERATING FUNCTIONS

Definition 2.7.2 For random variable X, with mass/density function f_X, the
factorial moment or **probability generating function**, fmgf or pgf , of X, G_X, is defined by

$$G_X(t) = E_{f_X}[t^X] = E_{f_X}[e^{X \log t}] = M_X(\log t),$$

if this expectation exists for all values of $t \in (1-h, 1+h)$ for some $h > 0$.

Properties :
(i) Using similar techniques to those used for the mgf, it can be shown that

$$\begin{aligned} G_X^{(r)}(t) &= \frac{d^r}{ds^r}\{G_X(s)\}_{s=t} = E_{f_X}\left[X(X-1)...(X-r+1)t^{X-r}\right] \\ \implies & G_X^{(r)}(1) = E_{f_X}[X(X-1)...(X-r+1)], \end{aligned}$$

where $\mathrm{E}_{f_X}[X(X-1)...(X-r+1)]$ is the rth factorial moment.

(ii) For discrete random variables, it can be shown by using a Taylor series expansion of G_X that, for $r = 1, 2, ...,$

$$\frac{G_X^{(r)}(0)}{r!} = P[X = r].$$

Definition 2.7.3 For random variable X with mass/density function f_X, the
cumulant generating function of X, K_X, is defined by

$$K_X(t) = \log\left[M_X(t)\right],$$

for $t \in (-h, h)$ for some $h > 0$.
Moment generating functions provide a very useful technique for identifying distributions, but suffer from the disadvantage that the integrals which define them may not always be finite. Another class of functions which are equally useful and whose finiteness is guaranteed is described next.

Definition 2.7.4 The **characteristic function**, or cf, of X, C_X, is defined by

$$C_X(t) = E_{f_X}\left[e^{itX}\right].$$

By definition

$$C_X(t) = \int_{x\in\mathbb{X}} e^{itx} f_X(x)dx = \int_{x\in\mathbb{X}} [\cos tx + i\sin tx]\, f_X(x)dx$$

$$= \int_{x\in\mathbb{X}} \cos tx f_X(x)dx + i\int_{x\in\mathbb{X}} \sin tx f_X(x)dx$$

$$= E_{f_X}[\cos tX] + iE_{f_X}[\sin tX].$$

We will be concerned primarily with cases where the moment generating function exists, and the use of moment generating functions will be a key tool for identification of distributions.

2.8 JOINT PROBABILITY DISTRIBUTIONS

Suppose X and Y are random variables on the probability space $(\Omega, \mathcal{A}, P(.))$. Their distribution functions F_X and F_Y contain information about their associated probabilities. But how do we describe information about their properties *relative to each other*? We think of X and Y as components of a **random vector** (X, Y) taking values in $\mathbb{R}^2$, rather than as unrelated random variables each taking values in $\mathbb{R}$.

Example 2.5 Toss a coin n times and let $X_i = 0$ or 1, depending on whether the ith toss is a tail or a head. The random vector $X = (X_1, \ldots, X_n)$ describes the whole experiment. The total number of heads is $\sum_{i=1}^n X_i$.

The **joint distribution function** of a random vector $(X_1, \ldots, X_n)$ is $P(X_1 \le x_1, \ldots, X_n \le x_n)$, a function of n real variables $x_1, \ldots, x_n$.

For vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, write $x \le y$ if $x_i \le y_i$ for *each* $i = 1, \ldots, n$.

Definition 2.8.1 The **joint distribution function** of a random vector $X = (X_1, \ldots, X_n)$ on $(\Omega, \mathcal{A}, P(.))$ is given by $F_X : \mathbb{R}^n \longrightarrow [0, 1]$, defined by $F_X(x) = P(X \le x), x \in \mathbb{R}^n$. [Remember, formally, $\{X \le x\}$ means $\{\omega \in \Omega : X(\omega) \le x\}$.]

We will consider, for simplicity, the case $n = 2$, without any loss of generality: the case $n > 2$ is just notationally more cumbersome.

Properties of the joint distribution function.

The joint distribution function $F_{X,Y}$ of the random vector (X, Y) satisfies:

(i)

$$\lim_{x,y\longrightarrow-\infty} F_{X,Y}(x, y) = 0,$$

$$\lim_{x,y\longrightarrow\infty} F_{X,Y}(x, y) = 1.$$

(ii) If $(x_1, y_1) \leq (x_2, y_2)$ then

$$F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2).$$

(iii) $F_{X,Y}$ is continuous from above,

$$F_{X,Y}(x + u, y + v) \longrightarrow F_{X,Y}(x, y),$$

as $u, v \longrightarrow 0^+$.

(iv)

$$\lim_{y \longrightarrow \infty} F_{X,Y}(x, y) = F_X(x) \equiv P(X \leq x),$$
$$\lim_{x \longrightarrow \infty} F_{X,Y}(x, y) = F_Y(y) \equiv P(Y \leq y).$$

F_X and F_Y are the **marginal** distribution functions of the joint distribution $F_{X,Y}$.

Definition 2.8.2 The random variables X and Y on $(\Omega, \mathcal{A}, P(.))$ are (jointly) **discrete** if (X, Y) takes values in a countable subset of $\mathbb{R}^2$ only.

Definition 2.8.3 Discrete variables X and Y are **independent** if the events $\{X = x\}$ and $\{Y = y\}$ are independent *for all* x and y.

Definition 2.8.4 The **joint probability mass function** $f_{X,Y} : \mathbb{R}^2 \longrightarrow [0, 1]$ of X and Y is given by

$$f_{X,Y}(x, y) = P(X = x, Y = y).$$

The marginal pmf of X, $f_X(x)$, is found from:

$$\begin{aligned} f_X(x) &= P(X = x) \\ &= \sum_y P(X = x, Y = y) \\ &= \sum_y f_{X,Y}(x, y). \end{aligned}$$

Similarly for $f_Y(y)$.

The definition of independence can be reformulated as: X and Y are independent iff $f_{X,Y}(x, y) = f_X(x) f_Y(y)$, for all $x, y \in \mathbb{R}$.

More generally, X and Y are independent iff $f_{X,Y}(x, y)$ can be factorized as the product $g(x)h(y)$ of a function of x alone and a function of y alone.

Let $\mathbb{X}$ be the support of X and $\mathbb{Y}$ be the support of Y. Then $Z = (X, Y)$ has support $\mathbb{Z} = \{(x, y) : f_{X,Y}(x, y) > 0\}$. In nice cases $\mathbb{Z} = \mathbb{X} \times \mathbb{Y}$, but we need to be alert to cases with $\mathbb{Z} \subset \mathbb{X} \times \mathbb{Y}$. In general, given a random vector $(X_1, \ldots, X_k)$ we will denote its range or support by $\mathbb{X}^{(k)}$.

Definition 2.8.5 The **conditional distribution function** of Y given $X = x$, $F_{Y|X}(y|x)$ is defined by

$$F_{Y|X}(y|x) = P(Y \leq y|X = x),$$

for any x such that $P(X = x) > 0$. The **conditional probability mass function** of Y given $X = x$, $f_{Y|X}(y|x)$, is defined by

$$f_{Y|X}(y|x) = P(Y = y|X = x),$$

for any x such that $P(x = x) > 0$.

Turning now to the continuous case, we define:

Definition 2.8.6 The random variables X and Y on $(\Omega, \mathcal{A}, P(.))$ are called **jointly continuous** if their joint distribution function can be expressed as

$$F_{X,Y}(x, y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f_{X,Y}(u, v)dvdu,$$

$x, y \in \mathbb{R}$, for some $f_{X,Y} : \mathbb{R}^2 \longrightarrow [0, \infty)$.

Then $f_{X,Y}$ is the **joint probability density function** of X, Y.

If $F_{X,Y}$ is 'sufficiently differentiable' at (x, y) we have

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$$

This is the **usual case**, which we will assume from now on.

Then:

(i)

$$\begin{aligned} P(a \leq X \leq b, c \leq Y \leq d) &= F_{X,Y}(b, d) - F_{X,Y}(a, d) - F_{X,Y}(b, c) + F_{X,Y}(a, c) \\ &= \int_c^d \int_a^b f_{X,Y}(x, y)dxdy. \end{aligned}$$

If B is a 'nice' subset of $\mathbb{R}^2$, such as a union of rectangles,

$$P((X, Y) \in B) = \int\int_B f_{X,Y}(x, y)dxdy.$$

(ii) The marginal distribution functions of X and Y are:

$$F_X(x) = P(X \leq x) = F_{X,Y}(x, \infty),$$

$$F_Y(y) = P(Y \leq y) = F_{X,Y}(\infty, y).$$

Since

$$F_X(x) = \int_{-\infty}^{x} \left\{ \int_{-\infty}^{\infty} f_{X,Y}(u,y)dy \right\} du,$$

we see, differentiating with respect to x, that the marginal pdf of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy.$$

Similarly, the marginal pdf of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx.$$

We **cannot**, as we did in the discrete case, define independence of X and Y in terms of events $\{X = x\}$ and $\{Y = y\}$, as these have **zero** probability and are trivially independent.

So,

Definition 2.8.7 X and Y are **independent** if $\{X \leq x\}$ and $\{Y \leq y\}$ are independent events, for all $x, y \in \mathbb{R}$.

So, X and Y are independent iff

$$F_{X,Y}(x,y) = F_X(x)F_Y(y), \forall x, y \in \mathbb{R},$$

or (equivalently) iff

$$f_{X,Y}(x,y) = f_X(x)f_Y(y),$$

whenever $F_{X,Y}$ is differentiable at (x,y).

(iv) **Definition 2.8.8** The **conditional distribution function** of Y given $X = x$, $F_{Y|X}(y|x)$ or $P(Y \leq y|X = x)$ is defined as

$$F_{Y|X}(y|x) = \int_{v=-\infty}^{y} \frac{f_{X,Y}(x,v)}{f_X(x)} dv,$$

for any x such that $f_X(x) > 0$.

Definition 2.8.9 The **conditional density function** of Y, given $X = x$, $f_{Y|X}(y|x)$, is defined by

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

for any x such that $f_X(x) > 0$.

This is an appropriate point to remark that not all random variables are either continuous or discrete, and not all distribution functions are either absolutely continuous or discrete. Many practical examples exist of distribution functions that are partly discrete and partly continuous.

Example 2.6 We record the delay that a motorist encounters at a one-way traffic stop sign. Let X be the random variable representing the delay the motorist experiences. There is a certain probability that there will be no opposing traffic, so she will be able to proceed without delay. However, if she has to wait, she could (in principle) have to wait for any positive amount of time. The experiment could be described by assuming that X has distribution function $F_X(x) = (1 - pe^{-\lambda x})I_{[0,\infty)}(x)$. This has a jump of $1 - p$ at $x = 0$, but is continuous for $x > 0$: there is a probability $1 - p$ of no wait at all.

We shall see later cases of random vectors, (X, Y) say, where one component is discrete and the other continuous: there is no essential complication in the manipulation of the marginal distributions etc. for such a case.

2.8.1 THE CHAIN RULE FOR RANDOM VARIABLES

As with the chain rule for manipulation of probabilities, there is an explicit relationship between joint, marginal, and conditional mass/density functions. For example, consider three continuous random variables X_1, X_2, X_3, with joint pdf f_{X_1,X_2,X_3}. Then,

$$f_{X_1,X_2,X_3}(x_1, x_2, x_3) = f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) f_{X_3|X_1,X_2}(x_3|x_1, x_2),$$

so that, for example,

$$\begin{aligned} f_{X_1}(x_1) &= \int_{\mathbb{X}_2}\int_{\mathbb{X}_3} f_{X_1,X_2,X_3}(x_1, x_2, x_3)dx_2dx_3 \\ &= \int_{\mathbb{X}_2}\int_{\mathbb{X}_3} f_{X_1|X_2,X_3}(x_1|x_2, x_3) f_{X_2,X_3}(x_2, x_3)dx_2dx_3 \\ &= \int_{\mathbb{X}_2}\int_{\mathbb{X}_3} f_{X_1|X_2,X_3}(x_1|x_2, x_3) f_{X_2|X_3}(x_2|x_3) f_{X_3}(x_3)dx_2dx_3. \end{aligned}$$

Equivalent relationships hold in the discrete case and can be extended to determine the explicit relationship between joint, marginal, and conditional mass/density functions for any number of random variables.

NOTE: the discrete equivalent of this result is a DIRECT consequence of the Theorem of Total Probability; the event $[X_1 = x_1]$ is partitioned into sub-events $[(X_1 = x_1) \cap (X_2 = x_2) \cap (X_3 = x_3)]$ for all possible values of the pair (x_2, x_3).

2.8.2 CONDITIONAL EXPECTATION AND ITERATED EXPECTATION

Consider two discrete/continuous random variables X_1 and X_2 with joint mass function/pdf f_{X_1,X_2}, and the conditional mass function/pdf of X_1 given $X_2 = x_2$, defined in the usual way by

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)}.$$

Then the **conditional expectation** of $g(X_1)$ given $X_2 = x_2$ is defined by

$$E_{f_{X_1|X_2}}[g(X_1)|X_2 = x_2] = \begin{cases} \sum_{x_1 \in \mathbb{X}_1} g(x_1) f_{X_1|X_2}(x_1|x_2), & X_1 \text{ DISCRETE.} \\ \\ \int_{\mathbb{X}_1} g(x_1) f_{X_1|X_2}(x_1|x_2) dx_1, & X_1 \text{ CONTINUOUS,} \end{cases}$$

i.e. the expectation of $g(X_1)$ *with respect to* the conditional density of X_1 given $X_2 = x_2$, (possibly giving a function of x_2). The case $g(x) \equiv x$ is a particular case.

Theorem 2.8.1 ***THE LAW OF ITERATED EXPECTATION***
For two continuous random variables X_1 and X_2 with joint pdf f_{X_1,X_2},

$$E_{f_{X_1}}[g(X_1)] = E_{f_{X_2}}\left[E_{f_{X_1|X_2}}[g(X_1)|X_2 = x_2]\right].$$

Proof

$$\begin{aligned} E_{f_{X_1}}[g(X_1)] &= \int_{\mathbb{X}_1} g(x_1) f_{X_1}(x_1) dx_1 \\ &= \int_{\mathbb{X}_1} g(x_1) \left\{ \int_{\mathbb{X}_2} f_{X_1,X_2}(x_1, x_2) dx_2 \right\} dx_1 \\ &= \int_{\mathbb{X}_1} g(x_1) \left\{ \int_{\mathbb{X}_2} f_{X_1|X_2}(x_1|x_2) f_{X_2}(x_2) dx_2 \right\} dx_1 \\ &= \int_{\mathbb{X}_1} \int_{\mathbb{X}_2} g(x_1) f_{X_1|X_2}(x_1|x_2) f_{X_2}(x_2) dx_2 dx_1 \\ &= \int_{\mathbb{X}_2} \left\{ \int_{\mathbb{X}_1} g(x_1) f_{X_1|X_2}(x_1|x_2) dx_1 \right\} f_{X_2}(x_2) dx_2 \\ &= \int_{\mathbb{X}_2} \left\{ E_{f_{X_1|X_2}}[g(X_1)|X_2 = x_2] \right\} f_{X_2}(x_2) dx_2 \\ &= E_{f_{X_2}}\left[E_{f_{X_1|X_2}}[g(X_1)|X_2 = x_2]\right], \end{aligned}$$

so the expectation of $g(X_1)$ can be calculated by finding the conditional expectation of $g(X_1)$ given $X_2 = x_2$, giving a function of x_2, and then taking the expectation of this function with respect to the marginal density for X_2. Note that this proof only works if the conditional expectation and the marginal expectation are finite. This results extends naturally to k variables.

2.9 MULTIVARIATE TRANSFORMATIONS

Theorem 2.9.1 ***THE MULTIVARIATE TRANSFORMATION THEOREM***
Let $\mathbf{X} = (X_1, ..., X_k)$ be a vector of random variables, with joint mass/density function $f_{X_1,...,X_k}$.

Let $\mathbf{Y} = (Y_1, ..., Y_k)$ *be a vector of random variables defined by* $Y_i = g_i(X_1, ..., X_k)$ *for some functions* $g_i, i = 1, ..., k$*, where the vector function* $\mathbf{g}$ *mapping* $(X_1, ..., X_k)$ *to* $(Y_1, ..., Y_k)$ *is a 1-1 transformation. Then the joint mass/density function of* $(Y_1, ..., Y_k)$ *is given by*

$$\textit{DISCRETE} \qquad f_{Y_1,...,Y_k}(y_1, ..., y_k) = f_{X_1,...,X_k}(x_1, ..., x_k),$$

$$\textit{CONTINUOUS} \quad f_{Y_1,...,Y_k}(y_1, ..., y_k) = f_{X_1,...,X_k}(x_1, ..., x_k)\,|J(y_1, ..., y_k)|\,,$$

where $\mathbf{x} = (x_1, ..., x_k)$ *is the* ***unique*** *solution of the system* $\mathbf{y} = \mathbf{g}(\mathbf{x})$*, so that* $\mathbf{x} = \mathbf{g}^{-1}(\mathbf{y})$*, and where* $J(y_1, ..., y_k)$ *is the* ***Jacobian*** *of the transformation, that is, the determinant of the* $k \times k$ *matrix whose* (i, j)*th element is*

$$\frac{\partial}{\partial t_j} \left\{ {g_i}^{-1}(\mathbf{t}) \right\}_{t_1=y_1,...,t_k=y_k},$$

where g_i^{-1} *is the inverse function uniquely defined by* $X_i = g_i^{-1}(Y_1, ..., Y_k)$*. Note again the* ***modulus****.*

Proof. The discrete case proof follows the univariate case precisely. For the continuous case, consider the **equivalent** events $[\mathbf{X} \in C]$ and $[\mathbf{Y} \in D]$, where D is the image of C under $\mathbf{g}$. Clearly, $P[\mathbf{X} \in C] = P[\mathbf{Y} \in D]$. Now, $P[\mathbf{X} \in C]$ is the k dimensional integral of the joint density $f_{X_1,...,X_k}$ over the set C, and $P[\mathbf{Y} \in D]$ is the k dimensional integral of the joint density $f_{Y_1,...,Y_k}$ over the set D. The result follows by changing variables in the first integral from $\mathbf{x}$ to $\mathbf{y} = \mathbf{g}(\mathbf{x})$, and equating the two integrands.

Note : As for single variable transformations, the ranges of the transformed variables must be considered carefully.

Example 2.7 The multivariate transformation theorem provides a simple proof of the **convolution formula**: if X and Y are independent continuous random variables with pdfs $f_X(x)$ and $f_Y(y)$, then the pdf of $Z = X + Y$ is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(w) f_Y(z - w) dw.$$

Let $W = X$. The Jacobian of the transformation from (X, Y) to (Z, W) is 1. So, the joint pdf of (Z, W) is

$$f_{Z,W}(z, w) = f_{X,Y}(w, z - w) = f_X(w) f_Y(z - w).$$

Then integrate out W to obtain the marginal pdf of Z.

Example 2.8 Consider the case $k = 2$, and suppose that X_1 and X_2 are *independent* continuous random variables with ranges $\mathbb{X}_1 = \mathbb{X}_2 = [0, 1]$ and pdfs given respectively by

$$f_{X_1}(x_1) = 6x_1(1 - x_1), \quad 0 \le x_1 \le 1,$$

$$f_{X_2}(x_2) = 3x_2^2, \qquad 0 \le x_2 \le 1,$$

and zero elsewhere. In order to calculate the pdf of random variable Y_1 defined by

$$Y_1 = X_1 X_2,$$

using the transformation result, consider the additional random variable Y_2, where $Y_2 = X_1$ (note, as X_1 and X_2 take values on $[0,1]$, $X_1 \geq X_1X_2$ so $Y_1 \leq Y_2$).

The transformation $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ is then specified by the two functions

$$g_1(t_1, t_2) = t_1t_2, \qquad g_2(t_1, t_2) = t_1,$$

and the inverse transformation $\mathbf{X} = \mathbf{g}^{-1}(\mathbf{Y})$ (i.e. $\mathbf{X}$ in terms of $\mathbf{Y}$) is

$$X_1 = Y_2, \qquad X_2 = Y_1/Y_2,$$

giving

$$g_1^{-1}(t_1, t_2) = t_2, \qquad g_2^{-1}(t_1, t_2) = t_1/t_2.$$

Hence

$$\frac{\partial}{\partial t_1}\left\{{g_1}^{-1}(t_1, t_2)\right\} = 0, \qquad \frac{\partial}{\partial t_2}\left\{{g_1}^{-1}(t_1, t_2)\right\} = 1,$$

$$\frac{\partial}{\partial t_1}\left\{{g_2}^{-1}(t_1, t_2)\right\} = 1/t_2, \quad \frac{\partial}{\partial t_2}\left\{{g_2}^{-1}(t_1, t_2)\right\} = -t_1/t_2^2,$$

and so the Jacobian $J(y_1, y_2)$ of the transformation is given by

$$\begin{vmatrix} 0 & 1 \\ 1/y_2 & -y_1/y_2^2 \end{vmatrix}$$

so that $J(y_1, y_2) = -1/y_2$. Hence, using the theorem

$$\begin{aligned} f_{Y_1,Y_2}(y_1, y_2) &= f_{X_1,X_2}(y_2, y_1/y_2) \times |J(y_1, y_2)| \\ &= 6y_2(1 - y_2) \times 3(y_1/y_2)^2 \times 1/y_2 \\ &= 18y_1^2(1 - y_2)/y_2^2, \end{aligned}$$

on the set $\mathbb{Y}^{(2)} = \{(y_1, y_2) : 0 \leq y_1 \leq y_2 \leq 1\}$, and zero otherwise. Hence

$$\begin{aligned} f_{Y_1}(y_1) &= \textstyle\int_{y_1}^1 18y_1^2(1 - y_2)/y_2^2 dy_2 \\ &= 18y_1^2\left[-1/y_2 - \log y_2\right]_{y_1}^1 \\ &= 18y_1^2(-1 + 1/y_1 + \log y_1) \\ &= 18y_1(1 - y_1 + y_1 \log y_1), \end{aligned}$$

for $0 \leq y_1 \leq 1$, and zero otherwise.

2.10 MULTIVARIATE EXPECTATIONS AND COVARIANCE

2.10.1 EXPECTATION WITH RESPECT TO JOINT DISTRIBUTIONS

Definition 2.10.1 For random variables $X_1, ..., X_k$ with range $\mathbb{X}^{(k)}$ with mass/density function $f_{X_1,...,X_k}$, the **expectation** of $g(X_1, ..., X_k)$ is defined in the discrete and continuous cases by

$$E_{f_{X_1,...,X_k}}[g(X_1, ..., X_k)] = \begin{cases} \sum_{\mathbb{X}_1} ... \sum_{\mathbb{X}_k} g(x_1, ..., x_k) f_{X_1,...,X_k}(x_1, ..., x_k), \\ \\ \int_{\mathbb{X}_1} ... \int_{\mathbb{X}_k} g(x_1, ..., x_k) f_{X_1,...,X_k}(x_1, ..., x_k) dx_1 ... dx_k. \end{cases}$$

PROPERTIES

(i) Let g and h be real-valued functions and let a and b be constants. Then, if $f_{\mathbf{X}} \equiv f_{X_1,...,X_k}$,

$$E_{f_{\mathbf{X}}}[ag(X_1, ..., X_k) + bh(X_1, ..., X_k)] = aE_{f_{\mathbf{X}}}[g(X_1, ..., X_k)] + bE_{f_{\mathbf{X}}}[h(X_1, ..., X_k)].$$

(ii) Let $X_1, ...X_k$ be **independent** random variables with mass functions/pdfs $f_{X_1}, ..., f_{X_k}$ respectively. Let $g_1, ..., g_k$ be scalar functions of $X_1, ..., X_k$ respectively (that is, g_i is a function of X_i *only* for $i = 1, ..., k$). If $g(X_1, ..., X_k) = g_1(X_1)...g_k(X_k)$, then

$$E_{f_{\mathbf{X}}}[g(X_1, ..., X_k)] = \prod_{i=1}^{k} E_{f_{X_i}}[g_i(X_i)],$$

where $E_{f_{X_i}}[g_i(X_i)]$ is the marginal expectation of $g_i(X_i)$ with respect to f_{X_i}.

(iii) Generally,

$$E_{f_{\mathbf{X}}}[g(X_1)] \equiv E_{f_{X_1}}[g(X_1)],$$

so that the expectation over the **joint** distribution is the same as the expectation over the **marginal** distribution. The proof is an immediate consequence of the fact that the marginal pdf f_{X_1} is obtained by integrating the joint density with respect to $x_2, \ldots, x_k$. So, whevever we wish, it is reasonable to denote the expectation as, say, $E[g(X_1)]$, rather than $E_{f_{X_1}}[g(X_1)]$ or $E_{f_{\mathbf{X}}}[g(X_1)]$: we can 'drop subscripts'.

2.10.2 COVARIANCE AND CORRELATION

Definition 2.10.2 The **covariance** of two random variables X_1 and X_2 is denoted $Cov_{f_{X_1,X_2}}[X_1, X_2]$, and is defined by

$$Cov_{f_{X_1,X_2}}[X_1, X_2] = E_{f_{X_1,X_2}}[(X_1 - \mu_1)(X_2 - \mu_2)] = E_{f_{X_1,X_2}}[X_1X_2] - \mu_1\mu_2,$$

where $\mu_i = E_{f_{X_i}}[X_i]$ is the marginal expectation of X_i, for $i = 1, 2$, and where

$$E_{f_{X_1,X_2}}[X_1X_2] = \int \int x_1x_2 f_{X_1,X_2}(x_1, x_2) dx_1 dx_2,$$

that is, the expectation of function $g(x_1, x_2) = x_1x_2$ with respect to the joint distribution f_{X_1,X_2}.

Definition 2.10.3 The **correlation** of X_1 and X_2 is denoted $Corr_{f_{X_1,X_2}}[X_1, X_2]$, and is defined by

$$Corr_{f_{X_1,X_2}}[X_1, X_2] = \frac{Cov_{f_{X_1,X_2}}[X_1, X_2]}{\sqrt{Var_{f_{X_1}}[X_1]Var_{f_{X_2}}[X_2]}}.$$

If $Cov_{f_{X_1,X_2}}[X_1, X_2] = Corr_{f_{X_1,X_2}}[X_1, X_2] = 0$ then variables X_1 and X_2 are **uncorrelated**.

Note that if random variables X_1 and X_2 are independent then

$$Cov_{f_{X_1,X_2}}[X_1, X_2] = E_{f_{X_1,X_2}}[X_1X_2] - E_{f_{X_1}}[X_1]E_{f_{X_2}}[X_2]$$

$$= E_{f_{X_1}}[X_1]E_{f_{X_2}}[X_2] - E_{f_{X_1}}[X_1]E_{f_{X_2}}[X_2] = 0,$$

and so X_1 and X_2 are also uncorrelated (the converse does **not** hold).

NOTES:
(i) For random variables X_1 and X_2, with (marginal) expectations μ_1 and μ_2 respectively, and (marginal) variances σ_1^2 and σ_2^2 respectively, if random variables Z_1 and Z_2 are defined by

$$Z_1 = (X_1 - \mu_1)/\sigma_1, \quad Z_2 = (X_2 - \mu_2)/\sigma_2,$$

then Z_1 and Z_2 are **standardized** variables. Then $E_{f_{Z_i}}[Z_i] = 0, Var_{f_{Z_i}}[Z_i] = 1$ and

$$Corr_{f_{X_1,X_2}}[X_1, X_2] = Cov_{f_{Z_1,Z_2}}[Z_1, Z_2].$$

(ii) Extension to k variables: covariances can only be calculated for *pairs* of random variables, but if k variables have a joint probability structure it is possible to construct a $k \times k$ *matrix*, $\mathbf{C}$ say, of covariance values, whose (i, j)th element is

$$Cov_{f_{X_i,X_j}}[X_i, X_j],$$

for $i, j = 1, .., k$, that captures the complete covariance structure in the joint distribution. If $i \neq j$, then

$$Cov_{f_{X_j,X_i}}[X_j, X_i] = Cov_{f_{X_i,X_j}}[X_i, X_j],$$

so $\mathbf{C}$ is *symmetric*, and if $i = j$,

$$Cov_{f_{X_i,X_i}}[X_i, X_i] \equiv Var_{f_{X_i}}[X_i].$$

The matrix $\mathbf{C}$ is referred to as the **variance-covariance matrix**.

(iii) If random variable X is defined by $X = a_1X_1 + a_2X_2 + ...a_kX_k$, for random variables $X_1, ..., X_k$ and constants $a_1, ..., a_k$, then

$$E_{f_X}[X] = \sum_{i=1}^{k} a_i E_{f_{X_i}}[X_i],$$

$$\begin{aligned} Var_{f_X}[X] &= \sum_{i=1}^{k} a_i^2 Var_{f_{X_i}}[X_i] + 2\sum_{i=1}^{k}\sum_{j=1}^{i-1} a_i a_j Cov_{f_{X_i,X_j}}[X_i, X_j \\ &= a^T\mathbf{C}a, \quad a = (a_1, \ldots, a_k)^T. \end{aligned}$$

(iv) Combining (i) and (iii) when $k = 2$, and defining standardized variables Z_1 and Z_2,

$$\begin{aligned} 0 \leq Var_{f_{Z_1,Z_2}}[Z_1 \pm Z_2] &= Var_{f_{Z_1}}[Z_1] + Var_{f_{Z_2}}[Z_2] \pm 2Cov_{f_{Z_1,Z_2}}[Z_1, Z_2] \\ &= 1 + 1 \pm 2Corr_{f_{X_1,X_2}}[X_1, X_2] = 2(1 \pm Corr_{f_{X_1,X_2}}[X_1, X_2]) \end{aligned}$$

and hence

$$-1 \leq Corr_{f_{X_1,X_2}}[X_1, X_2] \leq 1.$$

2.10.3 JOINT MOMENT GENERATING FUNCTION

Definition 2.10.4 Let X and Y be jointly distributed. The **joint moment generating function** of X and Y is

$$M_{X,Y}(s,t) = E(e^{sX+tY}).$$

If this exists in a neighbourhood of the origin (0,0), then it has the same attractive properties as the ordinary moment generating function. It determines the joint distribution of X and Y uniquely and it also yields the moments:

$$\frac{\partial^{m+n}}{\partial s^m \partial t^n} M_{X,Y}(s,t)\bigg|_{s=t=0} = E(X^m Y^n).$$

Joint moment generating functions factorize for independent random variables. We have

$$M_{X,Y}(s,t) = M_X(s) M_Y(t)$$

if and only if X and Y are independent.

2.11 ORDER STATISTICS

Order statistics, like sample moments, play a very important role in statistical inference. Let $X_1, \ldots, X_n$ be independent, identically distributed continuous random variables, with cdf F_X and pdf f_X. Then order the X_i: let Y_1 be the smallest of $\{X_1, \ldots, X_n\}$, Y_2 be the second smallest of $\{X_1, \ldots, X_n\}$, ..., Y_n be the largest of $\{X_1, \ldots, X_n\}$. Note that since we assume continuity, the chance of ties is zero. It is customary to use the notation

$$Y_k = X_{(k)},$$

and then $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are known as the **order statistics** of $X_1, \ldots, X_n$. Two key results are:

Result A The order statistics have joint density

$$n! \prod_{i=1}^{n} f_X(y_i), \quad y_1 < y_2 < \ldots < y_n.$$

Result B The order statistic $X_{(k)}$ has density

$$f_{(k)}(y) = k \binom{n}{k} f_X(y) \{1 - F_X(y)\}^{n-k} \{F_X(y)\}^{k-1}.$$

An informal proof of Result A is straightforward, using a symmetry argument based on the independent, identically distributed nature of $X_1, \ldots, X_n$. Result B follows on noting that the event $X_{(k)} \leq y$ occurs if and only if at least k of the X_i lie in $(-\infty, y]$. Recalling the Binomial distribution, this means that $X_{(k)}$ has distribution function

$$F_{(k)}(y) = \sum_{j=k}^{n} \binom{n}{j} \{F_X(y)\}^j \{1 - F_X(y)\}^{n-j}.$$

The pdf follows on differentiating this cdf.

CHAPTER 3

DISCRETE PROBABILITY DISTRIBUTIONS

Definition 3.1.1 DISCRETE UNIFORM DISTRIBUTION

$X \sim Uniform(n)$

$$f_X(x) = \frac{1}{n}, \qquad x \in \mathbb{X} = \{1, 2, ..., n\},$$

and zero otherwise.

Definition 3.1.2 BERNOULLI DISTRIBUTION

$X \sim Bernoulli(\theta)$

$$f_X(x) = \theta^x (1-\theta)^{1-x}, \qquad x \in \mathbb{X} = \{0, 1\},$$

and zero otherwise.

NOTE The Bernoulli distribution is used for modelling when the outcome of an experiment is either a "success" or a 'failure", where the probability of getting a success is equal to θ. Such an experiment is a 'Bernoulli trial'. The mgf is

$$M_X(t) = (1-\theta) + \theta e^t.$$

Definition 3.1.3 BINOMIAL DISTRIBUTION

$X \sim Bin(n, \theta)$

$$f_X(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \qquad x \in \mathbb{X} = \{0, 1, 2, ..., n\}, \qquad n \geq 1, \quad 0 \leq \theta \leq 1.$$

NOTES

1. If $X_1, ..., X_k$ are independent and identically distributed (IID) $Bernoulli(\theta)$ random variables, and $Y = X_1 + ..., X_k$, then by the standard result for mgfs,

$$M_Y(t) = \{M_X(t)\}^k = \left(1 - \theta + \theta e^t\right)^k,$$

so therefore $Y \sim Bin(k, \theta)$ because of the uniqueness of mgfs. Thus the binomial distribution is used to model the total number of successes in a series of independent and identical experiments.

2. Alternatively, consider sampling without replacement from infinite collection, or sampling with replacement from a finite collection of objects, a proportion θ of which are of Type I, and the remainder are of Type II. If X is the number of Type I objects in a sample of n, $X \sim Bin(n, \theta)$.

Definition 3.1.4 POISSON DISTRIBUTION

$X \sim Poisson(\lambda)$

$$f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x \in \mathbb{X} = \{0, 1, 2, ...\}, \qquad \lambda > 0,$$

and zero otherwise

NOTES
1. If $X \sim Bin(n, \theta)$, let $\lambda = n\theta$. Then

$$M_X(t) = \left(1 - \theta + \theta e^t\right)^n = \left(1 + \frac{\lambda(e^t - 1))}{n}\right)^n \longrightarrow \exp\left\{\lambda\left(e^t - 1\right)\right\},$$

as $n \longrightarrow \infty$, which is the mgf of a Poisson random variable. Therefore, the Poisson distribution arises as the **limiting case of the binomial distribution**, when $n \longrightarrow \infty, \theta \longrightarrow 0$ with $n\theta = \lambda$ constant (that is, for "large" n and "small" θ). So, if n is large, θ is small, we can reasonably approximate $Bin(n, \theta)$ by $Poisson(\lambda)$.

2. Suppose that X_1 and X_2 are independent, with $X_1 \sim Poisson(\lambda_1)$, $X_2 \sim Poisson(\lambda_2)$, then if $Y = X_1 + X_2$, using the general mgf result for independent random variables,

$$M_Y(t) = M_{X_1}(t)M_{X_2}(t) = \exp\left\{\lambda_1\left(e^t - 1\right)\right\}\exp\left\{\lambda_2\left(e^t - 1\right)\right\} = \exp\left\{\left(\lambda_1 + \lambda_2\right)\left(e^t - 1\right)\right\}$$

so that $Y \sim Poisson(\lambda_1 + \lambda_2)$. Therefore, the sum of two independent Poisson random variables also has a Poisson distribution. This result can be extended easily; if $X_1, ..., X_k$ are independent random variables with $X_i \sim Poisson(\lambda_i)$ for $i = 1, ..., k$, then

$$Y = \sum_{i=1}^{k} X_i \Longrightarrow Y \sim Poisson\left(\sum_{i=1}^{k}\lambda_i\right).$$

3. **THE POISSON PROCESS** (MATERIAL NOT EXAMINABLE)

The Poisson distribution arises as part of a larger modelling framework. Consider an experiment involving events (such as radioactive emissions) that occur repeatedly and randomly in time. Let $X(t)$ be the random variable representing the number of events that occur in the interval $[0, t]$, so that $X(t)$ takes values $0, 1, 2, ...$ Informally, in a Poisson process we have the following properties. In the interval $(t, t + h)$ there may or may not be events. If h is small, the probability of an event in $(t, t + h)$ is roughly proportional to h; it is not very likely that two or more events occur in a small interval.

Formally, a **Poisson process with intensity** λ is a process $X(t), t \geq 0$, taking values in $\{0, 1, 2, \ldots\}$ such that

(a) $X(0) = 0$ and if $s < t$ then $X(s) \leq X(t)$,

(b)

$$P(X(t + h) = n + m|\ X(t) = n) = \begin{cases} \lambda h + o(h), & m = 1, \\ o(h), & m > 1, \\ 1 - \lambda h + o(h), & m = 0. \end{cases}$$

(c) If $s < t$, the number of events $X(t) - X(s)$ in $(s, t]$ is independent of the number of events in $[0, s]$.

Here [definition]

$$\lim_{h \longrightarrow 0} \frac{o(h)}{h} = 0.$$

Then, if

$$P_n(t) = P[X(t) = n] = P[n \text{ events occur in } [0, t]]$$

it can be shown that

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

(that is, the random variable corresponding to the number of events that occurs in the interval $[0, t]$ has a Poisson distribution with parameter λt.)

Definition 3.1.5 GEOMETRIC DISTRIBUTION

$X \sim Geometric(\theta)$

$$f_X(x) = (1-\theta)^{x-1}\theta, \qquad x \in \mathbb{X} = \{1, 2, ...\}, \qquad 0 \leq \theta \leq 1.$$

NOTES

1. The cdf is available analytically as

$$F_X(x) = 1 - (1-\theta)^x, \qquad x = 1, 2, 3, ...$$

2. If $X \sim Geometric(\theta)$, then for $x, j \geq 1$,

$$P[X = x+j | X > j] = \frac{P[X = x+j, X > j]}{P[X > j]} = \frac{P[X = x+j]}{P[X > j]} = \frac{(1-\theta)^{x+j-1}\theta}{(1-\theta)^j} = (1-\theta)^{x-1}\theta = P[X = x].$$

So P$[X = x + j | X > j]$ =P$[X = x]$. This property is **unique** (among discrete distributions) to the geometric distribution, and is called the **lack of memory** property.

3. Alternative representations are sometimes useful:

$$f_X(x) = \phi^{x-1}(1-\phi), \qquad x = 1, 2, 3... \text{ (that is, } \phi = 1 - \theta),$$

$$f_X(x) = \phi^x(1-\phi), \qquad x = 0, 1, 2, \ldots.$$

4. The geometric distribution is used to model the number, X, of independent, identical Bernoulli trials until the first success is obtained. It is a discrete **waiting time** distribution.

Definition 3.1.6 NEGATIVE BINOMIAL DISTRIBUTION

$X \sim NegBin(n, \theta)$

$$f_X(x) = \binom{x-1}{n-1}\theta^n(1-\theta)^{x-n}, \qquad x \in \mathbb{X} = \{n, n+1, ...\}, \qquad n \geq 1, 0 \leq \theta \leq 1.$$

NOTES
1. If $X \sim Bin(n,\theta)$, $Y \sim NegBin(r,\theta)$, then for $r \leq n$, $P[X \geq r] = P[Y \leq n]$.

2. The Negative Binomial distribution is used to model the number, X, of independent, identical Bernoulli trials needed to obtain exactly n successes. [The number of trials up to and including the nth success].

3. **Alternative representation**: let Y be the number of **failures** in a sequence of independent, identical Bernoulli trials that contains exactly n successes. Then $Y = X - n$, and hence

$$f_Y(y) = \binom{n+y-1}{n-1}\theta^n(1-\theta)^y, \qquad y \in \{0,1,...\}.$$

4. If $X_i \sim Geometric(\theta)$, for $i = 1,...n$, are i.i.d. random variables, and $Y = X_1 + ... + X_n$, then $Y \sim NegBin(n,\theta)$ (result immediately follows using mgfs).

5. If $X \sim NegBin(n,\theta)$, let $n(1-\theta)/\theta = \lambda$ and $Y = X - n$. Then

$$M_Y(t) = e^{-nt}M_X(t) = \left\{\frac{\theta}{1-e^t(1-\theta)}\right\}^n = \left\{\frac{1}{1-\frac{\lambda}{n}(e^t-1)}\right\}^n \longrightarrow \exp\left\{\lambda(e^t-1)\right\},$$

as $n \longrightarrow \infty$, hence the alternate form of the negative binomial distribution tends to the Poisson distribution as $n \longrightarrow \infty$ with $n(1-\theta)/\theta = \lambda$ constant.

Definition 3.1.7 HYPERGEOMETRIC DISTRIBUTION
$X \sim HypGeom(N,R,n)$ for $N \geq R \geq n$

$$f_X(x) = \frac{\binom{N-R}{n-x}\binom{R}{x}}{\binom{N}{n}}, \qquad x \in \mathbb{X} = \{\max(0,n-N+R),...,\min(n,R)\},$$

and zero otherwise.

NOTES
1. The hypergeometric distribution is used as a model for experiments involving sampling without replacement from a finite population. Specifically, consider a finite population of size N, consisting of R items of Type I and $N-R$ of Type II: take a sample of size n **without replacement**, and let X be the number of Type I objects on the sample. The mass function for the hypergeometric distribution can be obtained by using combinatorics/counting techniques. However the form of the mass function does not lend itself readily to calculation of moments etc..

2. As $N, R \longrightarrow \infty$ with $R/N = \theta$(constant), then

$$P[X = x] \longrightarrow \binom{n}{x}\theta^x(1-\theta)^{n-x},$$

so the distribution tends to a Binomial distribution.

CHAPTER 4

CONTINUOUS PROBABILITY DISTRIBUTIONS

Definition 4.1.1 CONTINUOUS UNIFORM DISTRIBUTION

$X \sim Uniform(a, b)$

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

NOTES

1. The cdf is

$$F_X(x) = \frac{x-a}{b-a}, \qquad a \leq x \leq b.$$

2. The case $a = 0$ and $b = 1$ gives the **Standard uniform**.

Definition 4.1.2 EXPONENTIAL DISTRIBUTION

$X \sim Exp(\lambda)$

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

NOTES

1. The cdf is

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

2. **An alternative representation** uses $\theta = 1/\lambda$ as the parameter of the distribution. This is sometimes used because the expectation and variance of the Exponential distribution are

$$E_{f_X}[X] = \frac{1}{\lambda} = \theta, \qquad Var_{f_X}[X] = \frac{1}{\lambda^2}.$$

3. If $X \sim Exp(\lambda)$, then, for all $x, t > 0$,

$$P[X > x + t | X > t] = \frac{P[X > x + t, X > t]}{P[X > t]} = \frac{P[X > x + t]}{P[X > t]} = \frac{e^{-\lambda(x+t)}}{e^{-\lambda t}} = e^{-\lambda x} = P[X > x].$$

Thus, for all $x, t > 0$, $P[X > x + t | X > t] = P[X > x]$ - this is known as the **Lack of Memory Property**, and is **unique** to the exponential distribution amongst continuous distributions.

4. Suppose that $X(t)$ is a Poisson process with rate parameter $\lambda > 0$, so that

$$P[X(t) = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!}.$$

Let $X_1, ..., X_n$ be random variables defined by $X_1 =$ "time that first event occurs", and, for $i = 2, ..., n$, $X_i =$ "time interval between occurrence of $(i-1)$st and ith events". Then $X_1, ..., X_n$

are IID because of the assumptions underlying the Poisson process. So consider the distribution of X_1; in particular, consider the probability $\mathrm{P}[X_1 > x]$ for $x > 0$. The event $[X_1 > x]$ is equivalent to the event "No events occur in the interval (0,x]", which has probability $e^{-\lambda x}$. But

$$F_{X_1}(x) = P[X_1 \leq x] = 1 - P[X_1 > x] = 1 - e^{-\lambda x} \Longrightarrow X_1 \sim Exp(\lambda).$$

5. The exponential distribution is used to model failure times in continuous time. It is a continuous waiting time distribution, the continuous analogue of the geometric distribution.

6. If $X \sim Uniform(0, 1)$, and

$$Y = -\frac{1}{\lambda}\log(1 - X),$$

then $Y \sim Exp(\lambda)$.

7. If $X \sim Exp(\lambda)$, and

$$Y = X^{1/\alpha},$$

for $\alpha > 0$, then Y has a (two-parameter) **Weibull** distribution, and

$$f_Y(y) = \alpha\lambda y^{\alpha-1}e^{-\lambda y^\alpha}, \qquad y > 0.$$

Definition 4.1.3 GAMMA DISTRIBUTION

$X \sim Ga(\alpha, \beta)$

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}, \qquad x > 0, \qquad \alpha, \beta > 0,$$

where, for any real number $\alpha > 0$, the **gamma function**, $\Gamma(.)$ is defined by

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

NOTES

1. If $X_1 \sim Ga(\alpha_1, \beta), X_2 \sim Ga(\alpha_2, \beta)$ are independent random variables, and

$$Y = X_1 + X_2$$

then $Y \sim Ga(\alpha_1 + \alpha_2, \beta)$ (directly from properties of mgfs).

2. $Ga(1, \beta) \equiv Exp(\beta)$.

3. If $X_1, ..., X_n \sim Exp(\lambda)$ are independent random variables, and

$$Y = X_1 + ... + X_n$$

then $Y \sim Ga(n, \lambda)$ (directly from 1. and 2.).

4. For $\alpha > 0$, integrating by parts, we have that

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

and hence if $\alpha = 1, 2, ...$, then $\Gamma(\alpha) = (\alpha - 1)!$. A useful fact is that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

5. **Special Case** : If $\alpha = 1, 2, ...$ the $Ga(\alpha/2, 1/2)$ distribution is also known as the **chi-squared distribution** with α degrees of freedom, denoted by χ^2_α.

6. If $X_1 \sim \chi^2_{n_1}$ and $X_2 \sim \chi^2_{n_2}$ are independent chi-squared random variables with n_1 and n_2 degrees of freedom respectively, then random variable F defined as the ratio

$$F = \frac{X_1/n_1}{X_2/n_2}$$

has an **F-distribution** with (n_1, n_2) degrees of freedom.

Definition 4.1.4 BETA DISTRIBUTION

$X \sim Be(\alpha, \beta)$

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 < x < 1, \qquad \alpha, \beta > 0.$$

NOTES
1. If $\alpha = \beta = 1$, $Be(\alpha, \beta) \equiv Uniform(0, 1)$

2. If $X_1 \sim Ga(\alpha_1, \beta)$, $X_2 \sim Ga(\alpha_2, \beta)$ are independent random variables, and

$$Y = \frac{X_1}{X_1 + X_2}$$

then $Y \sim Be(\alpha_1, \alpha_2)$ (using standard multivariate transformation techniques).

3. Suppose that random variables X and Y have a joint probability distribution such that the conditional distribution of X, given $Y = y$ for $0 < y < 1$, is binomial, $Bin(n, y)$, and the marginal distribution of Y is beta, $Be(\alpha, \beta)$, so that

$$\begin{aligned} f_{X|Y}(x|y) &= \binom{n}{x} y^x (1-y)^{n-x}, \qquad x = 0, 1, ..., n, \\ f_Y(y) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \qquad 0 < y < 1. \end{aligned}$$

Then the marginal distribution of X is given by

$$\begin{aligned} f_X(x) &= \int_0^1 f_{X|Y}(x|y) f_Y(y) dy \\ &= \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(n+\alpha+\beta)}, \qquad x = 0, 1, 2...n. \end{aligned}$$

Note that this provides an example of a joint distribution of continuous Y and discrete X.

Definition 4.1.5 NORMAL DISTRIBUTION

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

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

NOTES

1. Special Case : If $\mu = 0$, $\sigma^2 = 1$, then X has a **standard** or **unit** normal distribution. Usually, the pdf of the standard normal is written $\phi(x)$, and the cdf is written $\Phi(x)$.

2. If $X \sim N(0,1)$, and
$$Y = \sigma X + \mu$$
then $Y \sim N(\mu, \sigma^2)$. Re-expressing this result, if $X \sim N(\mu, \sigma^2)$, and $Y = (X-\mu)/\sigma$, then $Y \sim N(0,1)$ (using transformation or mgf techniques).

3. **The Central Limit Theorem** Suppose $X_1, ..., X_n$ are IID random variables with *some* mgf M_X, with $\mathrm{E}_{f_X}[X_i] = \mu$ and $Var_{f_X}[X_i] = \sigma^2$ that is, the mgf and the expectation and variance of the X_i's are specified, but the pdf is not. Let the standardized random variable Z_n be defined by
$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}}$$
and let Z_n have mgf M_{Z_n}. Then, as $n \longrightarrow \infty$,
$$M_{Z_n}(t) \longrightarrow \exp\left\{t^2/2\right\},$$
irrespective of the distribution of the X_i's, that is, the distribution of Z_n tends to a standard normal distribution as n tends to infinity. This theorem will be proved and explained in Chapter 6.

4. If $X \sim N(0,1)$, and $Y = X^2$, then $Y \sim \chi^2_1$, so that the square of a unit normal random variable has a **chi-squared distribution** with 1 degree of freedom.

5. If $X \sim N(0,1)$, and $Y \sim N(0,1)$ are independent random variables, and Z is defined by $Z = X/Y$, then Z has a **Cauchy distribution**
$$f_Z(z) = \frac{1}{\pi}\frac{1}{1+z^2}, \qquad z \in \mathbb{R}.$$

6. If $X \sim N(0,1)$, and $Y \sim Ga(n/2, 1/2)$ for $n = 1, 2, ...$ (so that $Y \sim \chi^2_n$), are independent random variables, and T is defined by
$$T = \frac{X}{\sqrt{Y/n}}$$
then T has a **Student-t distribution with n degrees of freedom**, $T \sim St(n)$,
$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{1}{n\pi}\right)^{1/2}\left\{1+\frac{t^2}{n}\right\}^{-(n+1)/2}, \qquad t \in \mathbb{R}.$$

Taking limiting cases of the Student-t distribution

$$n \longrightarrow \infty : St(n) \longrightarrow N(0,1), \qquad n \longrightarrow 1 : St(n) \longrightarrow Cauchy.$$

7. If $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are independent and a, b are constants, then $T = aX_1 + bX_2 \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$.

CHAPTER 5

MULTIVARIATE PROBABILITY DISTRIBUTIONS

For purely notational reasons, it is convenient in this chapter to consider a random vector X as a **column vector**, $X = (X_1, \ldots, X_k)^T$, say

5.1 THE MULTINOMIAL DISTRIBUTION

The multinomial distribution is a multivariate generalization of the binomial distribution. Recall that the binomial distribution arose from an infinite Urn model with two types of objects being sampled with replacement. Suppose that the proportion of "Type 1" objects in the urn is θ (so $0 \leq \theta \leq 1$) and hence the proportion of "Type 2" objects in the urn is $1-\theta$. Suppose that n objects are sampled, and X is the random variable corresponding to the number of "Type 1" objects in the sample. Then $X \sim Bin(n, \theta)$, and

$$f_X(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \qquad x \in \{0, 1, 2, ..., n\}.$$

Now consider a generalization; suppose that the Urn contains $k+1$ types of objects ($k = 1, 2, ...$), with θ_i being the proportion of Type i objects, for $i = 1, ..., k+1$. Let X_i be the random variable corresponding to the number of type i objects in a sample of size n, for $i = 1, ..., k$. Then the joint distribution of vector $X = (X_1, ..., X_k)^T$ is given by

$$f_{X_1,...,X_k}(x_1, ..., x_k) = \frac{n!}{x_1!...x_k!x_{k+1}!}\theta_1^{x_1}....\theta_k^{x_k}\theta_{k+1}^{x_{k+1}} = \frac{n!}{x_1!...x_k!x_{k+1}!}\prod_{i=1}^{k+1}\theta_i^{x_i},$$

where $0 \leq \theta_i \leq 1$ for all i, and $\theta_1 + ... + \theta_k + \theta_{k+1} = 1$, and where x_{k+1} is defined by $x_{k+1} = n - (x_1 + ... + x_k)$. This is the mass function for the multinomial distribution which reduces to the binomial if $k = 1$. It can also be shown that the marginal distribution of X_i is $\text{Bin}(n, \theta_i)$.

5.2 THE DIRICHLET DISTRIBUTION

The Dirichlet distribution is a multivariate generalization of the beta distribution. Recall that the beta distribution arose as follows; suppose that V_1 and V_2 are independent Gamma random variables with $V_1 \sim Ga(\alpha_1, \beta), V_2 \sim Ga(\alpha_2, \beta)$. Then if X is defined by $X = V_1/(V_1 + V_2)$, we have that $X \sim Be(\alpha_1, \alpha_2)$. Now consider a generalization; suppose that $V_1, ..., V_{k+1}$ are independent Gamma random variables with $V_i \sim Ga(\alpha_i, \beta)$, for $i = 1, ..., k+1$. Define

$$X_i = \frac{V_i}{V_1 + ... + V_{k+1}}$$

for $i = 1, ..., k$. Then the joint distribution of vector $X = (X_1, ..., X_k)^T$ is given by density

$$f_{X_1,...,X_k}(x_1, ..., x_k) = \frac{\Gamma(\alpha)}{\Gamma(\alpha_1)...\Gamma(\alpha_k)\Gamma(\alpha_{k+1})}x_1^{\alpha_1-1}...x_k^{\alpha_k-1}{x_{k+1}}^{\alpha_{k+1}-1},$$

for $0 \leq x_i \leq 1$ for all i such that $x_1 + ... + x_k + x_{k+1} = 1$, where $\alpha = \alpha_1 + ... + \alpha_{k+1}$ and where x_{k+1} is defined by $x_{k+1} = 1 - (x_1 + ... + x_k)$. This is the density function which reduces to the beta distribution if $k = 1$. It can also be shown that the marginal distribution of X_i is $Beta(\alpha_i, \alpha)$.

5.3 THE MULTIVARIATE NORMAL DISTRIBUTION

The random vector $X = (X_1, \ldots, X_k)^T$ has a **multivariate normal distribution** if the joint pdf is of the form:

$$f_X(x_1, ..., x_k) = \left(\frac{1}{2\pi}\right)^{k/2} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\}.$$

Here $\mathbf{x}$ is the (column) vector of length k formed by $x_1, \ldots, x_k$, $\boldsymbol{\mu}$ is a (column) vector of length k, Σ is a $k \times k$ symmetric, positive definite matrix [$\Sigma = \Sigma^T$, $x^T \Sigma x > 0$ for all $x \neq 0$], and $|\Sigma|$ denotes the determinant of Σ.

We write $X \sim N_k(\boldsymbol{\mu}, \Sigma)$.

Properties

1. $E[X] = \boldsymbol{\mu}$: $\boldsymbol{\mu}$ is the mean vector of X. If $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^T$, we have $E[X_i] = \mu_i$. Further, Σ is the variance-covariance matrix of X, $\Sigma = [\sigma_{ij}]$, where $\sigma_{ij} = cov[X_i, X_j]$.

2. Since Σ is symmetric and positive definite, there exists a matrix $\Sigma^{1/2}$ [the 'square root of Σ'] such that: (i) $\Sigma^{1/2}$ is symmetric; (ii) $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$; (iii) $\Sigma^{1/2}\Sigma^{-1/2} = \Sigma^{-1/2}\Sigma^{1/2} = I$, the $k \times k$ identity matrix, with $\Sigma^{-1/2} = (\Sigma^{1/2})^{-1}$.

Then, if $Z \sim N_k(0, I)$, so that $Z_1, \ldots Z_k$ are IID $N(0, 1)$, and $X = \boldsymbol{\mu} + \Sigma^{1/2} Z$, then $X \sim N_k(\boldsymbol{\mu}, \Sigma)$. Conversely, if $X \sim N_k(\boldsymbol{\mu}, \Sigma)$, then $\Sigma^{-1/2}(X - \boldsymbol{\mu}) \sim N_k(0, I)$.

3. A useful result is the following: if $X \sim N_k(\boldsymbol{\mu}, \Sigma)$ and D is a $m \times k$ matrix of rank $m \leq k$, then $Y \equiv DX \sim N_m(D\boldsymbol{\mu}, D\Sigma D^T)$. A special case is where $X \sim N_k(0, I)$ and D is a $k \times k$ matrix of full rank k, so that D is invertible: then $Y = DX \sim N_k(0, DD^T)$.

4. Suppose we partition X as

$$X = \begin{pmatrix} X_a \\ X_b \end{pmatrix},$$

with

$$X_a = \begin{pmatrix} X_1 \\ \vdots \\ X_m \end{pmatrix}, X_b = \begin{pmatrix} X_{m+1} \\ \vdots \\ X_k \end{pmatrix}.$$

We can similarly partition $\boldsymbol{\mu}$ and Σ:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \ \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}.$$

Then if $X \sim N_k(\boldsymbol{\mu}, \Sigma)$ we have:

(I) The **marginal distribution** of X_a is $N_m(\mu_a, \Sigma_{aa})$.

(II) The **conditional distribution** of X_b, given $X_a = x_a$ is

$$X_b|X_a = x_a \sim N_{k-m}(\mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(x_a - \mu_a), \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}).$$

(III) If $a = (a_1, \ldots, a_k)^T$, then

$$a^T X \equiv \sum_{i=1}^{k} a_i X_i \sim N(a^T \boldsymbol{\mu}, a^T \Sigma a).$$

(IV) $V = (X - \boldsymbol{\mu})^T \Sigma^{-1} (X - \boldsymbol{\mu}) \sim \chi_k^2$.

CHAPTER 6

PROBABILITY RESULTS & LIMIT THEOREMS

6.1 BOUNDS ON PROBABILITIES BASED ON MOMENTS

Theorem 6.1.1 *If X is a random variable, then for non-negative function h, and $c > 0$,*

$$P\left[h(X) \geq c\right] \leq \frac{E_{f_X}\left[h(X)\right]}{c}.$$

Proof. *(continuous case) : Suppose that X has density function f_X which is positive for $x \in \mathbb{X}$. Let $\mathcal{A} = \{x \in \mathbb{X} : h(x) \geq c\} \subseteq \mathbb{X}$. Then, as $h(x) \geq c$ on $\mathcal{A}$,*

$$\begin{aligned} E_{f_X}\left[h(X)\right] &= \int h(x) f_X(x) dx = \int_{\mathcal{A}} h(x) f_X(x) dx + \int_{\mathcal{A}'} h(x) f_X(x) dx \\ &\geq \int_{\mathcal{A}} h(x) f_X(x) dx \geq \int_{\mathcal{A}} c f_X(x) dx = cP\left[X \in \mathcal{A}\right] = cP\left[h(X) \geq c\right]. \end{aligned}$$

SPECIAL CASE I - THE MARKOV INEQUALITY : If $h(x) = |x|^r$ for $r > 0$, so

$$P\left[|X|^r \geq c\right] \leq \frac{1}{c} E_{f_X}\left[|X|^r\right].$$

SPECIAL CASE II - THE CHEBYCHEV INEQUALITY: Suppose that X is a random variable with expectation μ and variance σ^2. Then taking $h(x) = (x-\mu)^2$ and $c = k^2\sigma^2$, for $k > 0$, gives

$$P\left[|X - \mu| \geq k\sigma\right] \leq 1/k^2,$$

and setting $\epsilon = k\sigma$ gives

$$P\left[|X - \mu| \geq \epsilon\right] \leq \sigma^2/\epsilon^2, \qquad P\left[|X - \mu| < \epsilon\right] \geq 1 - \sigma^2/\epsilon^2.$$

Theorem 6.1.2 ***JENSEN'S INEQUALITY***

Suppose that X is a random variable, and function g is ***convex*** *so that $\frac{d^2}{dt^2}\left\{g(t)\right\}_{t=x} = g''(x) > 0$, $\forall x$, with Taylor expansion around expectation μ of the form*

$$g(x) = g(\mu) + (x-\mu)g'(\mu) + \frac{1}{2}(x-\mu)^2 g''(x_0), \tag{6.1}$$

for some x_0 such that $x < x_0 < \mu$. Then

$$E_{f_X}\left[g(X)\right] \geq g(E_{f_X}\left[X\right]).$$

Proof. Taking expectations in (6.1), and noting that $E_{f_X}\left[(X-\mu)\right] = 0$, $E_{f_X}\left[(X-\mu)^2\right] = \sigma^2$, $g''(x_0) \geq 0$, we have that

$$E_{f_X}\left[g(X)\right] = g(\mu) + \left(0 \times g'(\mu)\right) + \frac{1}{2}\left(\sigma^2 \times g''(x_0)\right) \geq g(\mu) = g(E_{f_X}\left[X\right]),$$

as $\sigma^2, g''(x_0) > 0$.

6.2 THE CENTRAL LIMIT THEOREM

Theorem 6.2.1 *Suppose $X_1, ..., X_n$ are i.i.d. random variables with mgf M_X, with*

$$E_{f_X}[X_i] = \mu, \qquad Var_{f_X}[X_i] = \sigma^2,$$

both finite. Let the random variable Z_n be defined by

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}},$$

and let Z_n have mgf M_{Z_n}. Then, as $n \longrightarrow \infty$,

$$M_{Z_n}(t) \longrightarrow \exp\left\{\frac{t^2}{2}\right\},$$

***irrespective** of the form of M_X.*

Proof. First, let $Y_i = (X_i - \mu)/\sigma$ for $i = 1, ..., n$. Then $Y_1, ..., Y_n$ are i.i.d. with mgf M_Y say, and by the elementary properties of expectation,

$$E_{f_Y}[Y_i] = 0, \qquad Var_{f_Y}[Y_i] = 1,$$

for each i. Using the power series expansion result for mgfs, we have that

$$\begin{aligned} M_Y(t) &= 1 + tE_{f_Y}[Y] + \frac{t^2}{2!}E_{f_Y}[Y^2] + \frac{t^3}{3!}E_{f_Y}[Y^3] + \frac{t^4}{4!}E_{f_Y}[Y^4] + ... \\ &= 1 + \frac{t^2}{2!} + \frac{t^3}{3!}E_{f_Y}[Y^3] + \frac{t^4}{4!}E_{f_Y}[Y^4] + ... \end{aligned}$$

Now, the random variable Z_n can be rewritten

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

and thus, again by a standard mgf result, as $Y_1, ..., Y_n$ are independent, we have that

$$M_{Z_n}(t) = \prod_{i=1}^{n} \left\{M_Y(t/\sqrt{n})\right\} = \left\{1 + \frac{t^2}{2n} + \frac{t^3}{6n^{3/2}}E_{f_Y}[Y^3] + \frac{t^4}{6n^2}E_{f_Y}[Y^4]...\right\}^n.$$

In the limit as $n \longrightarrow \infty$, it follows that

$$\left\{1 + \frac{t^2}{2n} + \frac{t^3}{6n^{3/2}}E_{f_Y}[Y^3] + ...\right\}^n \approx \left\{1 + \frac{t^2}{2n}\right\}^n$$

for large n. Thus, as $n \longrightarrow \infty$, using the properties of the exponential function,

$$M_{Z_n}(t) \longrightarrow \exp\left\{\frac{t^2}{2}\right\}.$$

INTERPRETATION: Sums of independent and identically distributed random variables have a limiting distribution that is Normal, irrespective of the distribution of the variables.

6.3 MODES OF STOCHASTIC CONVERGENCE

6.3.1 CONVERGENCE IN DISTRIBUTION

Definition 6.3.1 Consider a sequence $\{X_n\}, n = 1, 2, \ldots$, of random variables and a corresponding sequence of cdfs, $F_{X_1}, F_{X_2}, \ldots$ so that for $n = 1, 2, \ldots$, $F_{X_n}(x) = \mathrm{P}[X_n \leq x]$. Suppose that there exists a cdf, F_X, such that **for all x at which F_X is continuous**,

$$\lim_{n \longrightarrow \infty} F_{X_n}(x) = F_X(x).$$

Then the sequence $\{X_n\}$ **converges in distribution** to the random variable X with cdf F_X. This is denoted

$$X_n \xrightarrow{d} X,$$

and F_X is the limiting distribution.

Convergence of a sequence of mgfs also indicates convergence in distribution. That is, if for all t at which $M_X(t)$ is defined, if as $n \longrightarrow \infty$, we have

$$M_{X_n}(t) \longrightarrow M_X(t)$$

then $X_n \xrightarrow{d} X$.

Definition 6.3.2 The sequence $\{X_n\}$ of random variables converges in distribution to the constant c if the limiting distribution of X_n is **degenerate at** c, that is,

$$X_n \xrightarrow{d} X$$

and $\mathrm{P}[X = c] = 1$, so that

$$F_X(x) = \begin{cases} 0, & x < c, \\ 1, & x \geq c. \end{cases}$$

This special type of convergence in distribution occurs when the limiting distribution is discrete, with the probability mass function only being non-zero at a single value. That is, if the limiting random variable is X, then

$$f_X(x) = 1, \qquad x = c, \text{ and zero otherwise.}$$

Theorem 6.3.1 *The sequence of random variables $\{X_n\}$* ***converges in distribution*** *to c if and only if, for all $\epsilon > 0$,*

$$\lim_{n \longrightarrow \infty} P\left[|X_n - c| < \epsilon\right] = 1.$$

This theorem indicates that convergence in distribution to a constant c occurs if and only if the probability becomes increasingly concentrated around c as $n \longrightarrow \infty$.

6.3.2 CONVERGENCE IN PROBABILITY

Definition 6.3.3 CONVERGENCE IN PROBABILITY TO A CONSTANT
The sequence of random variables $\{X_n\}$ converges in probability to the constant c, denoted

$$X_n \xrightarrow{P} c$$

if for all $\epsilon > 0$

$$\lim_{n\longrightarrow\infty} P\left[|X_n - c| < \epsilon\right] = 1, \text{ or, equivalently, } \lim_{n\longrightarrow\infty} P\left[|X_n - c| \geq \epsilon\right] = 0,$$

that is, if the limiting distribution of X_n is **degenerate at** c.

Interpretation. Convergence in probability to a constant is precisely equivalent to convergence in distribution to a constant.

A very useful result is Slutsky's Theorem which states that if $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} c$, where c is a finite constant, then: (i) $X_n + Y_n \xrightarrow{d} X + c$, (ii) $X_n Y_n \xrightarrow{d} cX$, (iii) $X_n/Y_n \xrightarrow{d} X/c$, if $c \neq 0$.

Theorem 6.3.2 ***WEAK LAW OF LARGE NUMBERS***
Suppose that $\{X_n\}$ is a sequence of i.i.d. random variables with expectation μ and variance σ^2. Let Y_n be defined by

$$Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then, for all $\epsilon > 0$,

$$\lim_{n\longrightarrow\infty} P\left[|Y_n - \mu| < \epsilon\right] = 1,$$

that is, $Y_n \xrightarrow{P} \mu$, and thus the mean of $X_1, \ldots, X_n$ converges in probability to μ.

Proof. Using the properties of expectation, it can be shown that Y_n has expectation μ and variance σ^2/n, and hence by the Chebychev Inequality,

$$P\left[|Y_n - \mu| \geq \epsilon\right] \leq \frac{\sigma^2}{n\epsilon^2} \longrightarrow 0, \qquad \text{as } n \longrightarrow \infty$$

for all $\epsilon > 0$. Hence

$$P\left[|Y_n - \mu| < \epsilon\right] \longrightarrow 1, \qquad \text{as } n \longrightarrow \infty$$

and $Y_n \xrightarrow{P} \mu$.

Definition 6.3.4 CONVERGENCE TO A RANDOM VARIABLE
The sequence of random variables $\{X_n\}$ converges in probability to the random variable X, denoted $X_n \xrightarrow{P} X$, if, for all $\epsilon > 0$,

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

Theorem 6.3.3 *For the sequence* $\{X_n\}$ *of random variables,*

$$X_n \xrightarrow{P} X \Longrightarrow X_n \xrightarrow{d} X,$$

so convergence in probability to a random variable implies convergence in distribution.

CHAPTER 7

STATISTICAL ANALYSIS

7.1 STATISTICAL SUMMARIES

Definition 7.1.1 A collection of i.i.d. random variables $X_1, ..., X_n$ each of which has distribution defined by cdf F_X (or mass/density function f_X) is a **random sample** of size n from F_X (or f_X).

Definition 7.1.2 A function, T, of a random sample, $X_1, ..., X_n$, that is, $T = t(X_1, ..., X_n)$ that depends only on $X_1, ..., X_n$ is a **statistic**. A statistic is a random variable. For example, the sample mean

$$\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

is a statistic. But a statistic T need not necessarily be constructed from a random sample (the random variables need not be i.i.d.), but that is the case encountered most often. In many circumstances it is necessary to consider statistics constructed from a collection of independent, but not identically distributed, random variables.

7.2 SAMPLING DISTRIBUTIONS

Definition 7.2.1 If $X_1, ..., X_n$ is a random sample from F_X, say, and $T = t(X_1, ..., X_n)$ is a statistic, then F_T (or f_T), the cdf (or mass/density function) of random variable T, is the **sampling distribution** of T. This notion extends immediately to the case of a statistic T constructed from a general collection of random variables $X_1, ..., X_n$.

EXAMPLE: If $X_1, ..., X_n$ are independent random variables, with $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, ..., n$, and $a_1, ..., a_n$ are constants, consider the distribution of random variable Y defined by

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

Using standard mgf results, the distribution of Y is derived to be normal with parameters

$$\mu_Y = \sum_{i=1}^{n} a_i \mu_i, \qquad \sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2.$$

Now consider the special case of this result when $X_1, ..., X_n$ are i.i.d. with $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2$, and where $a_i = 1/n$ for $i = 1, ..., n$. Then

$$Y = \sum_{i=1}^{n} \frac{1}{n} X_i = \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Definition 7.2.2 For a random sample $X_1, ..., X_n$ from a probability distribution, then the **sample variance**, s^2, is the statistic defined by

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

Theorem 7.2.1 ***SAMPLING DISTRIBUTION FOR NORMAL SAMPLES*** *If $X_1, ..., X_n$ is a random sample from a normal distribution, say $X_i \sim N(\mu, \sigma^2)$, then:*

(a) $\bar{X}$ is independent of $\left\{X_i - \overline{X}, i = 1, ..., n\right\}$;
(b) $\bar{X}$ and s^2 are independent random variables;
(c) The random variable

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

*has a **chi-squared distribution** with $n-1$ degrees of freedom.*

Proof. (**NOT EXAMINABLE**) See M2S2 and

http://stats.ma.ic.ac.uk/~ayoung/public_html/m2s1/NormalFactsheet.PDF

Theorem 7.2.2 *Suppose that $X_1, ..., X_n$ is a random sample from a normal distribution, say $X_i \sim N(\mu, \sigma^2)$. Then the random variable*

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

has a Student-t distribution with $n-1$ degrees of freedom.

Proof. Consider the random variables

$$\begin{aligned} Z &= \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0,1), \\ V &= \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}, \end{aligned}$$

and

$$T = \frac{Z}{\sqrt{\frac{V}{n-1}}},$$

and use the properties of the normal distribution and related random variables (**NOTE 6, following Definition 4.1.5).** Also, see **EXERCISES 5, Q4 (b).**

7.3 HYPOTHESIS TESTING

7.3.1 TESTING FOR NORMAL SAMPLES - THE Z-TEST

We concentrate initially on random data samples that we can assume to have a normal distribution, and utilize the Theorem from the previous section. We will look at two situations, namely **one sample** and **two sample** experiments. So, we suppose that $X_1, ..., X_n \sim N\left(\mu, \sigma^2\right)$ (one sample) and $X_1, ..., X_n \sim N(\mu_X, \sigma_X^2)$, $Y_1, ..., Y_n \sim N(\mu_Y, \sigma_Y^2)$ (two sample): in the latter case we assume also independence of the two samples.

- **ONE SAMPLE** Possible tests of interest are: $\mu = \mu_0, \sigma = \sigma_0$ for some specified constants μ_0 and σ_0.
- **TWO SAMPLE** Possible tests of interest are: $\mu_X = \mu_Y, \sigma_X = \sigma_Y$.

Recall from Theorem 7.2.1 that, if $X_1, ..., X_n \sim N(\mu, \sigma^2)$ are the i.i.d. outcome random variables of n experimental trials, then

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{and} \qquad \frac{(n-1)\, s^2}{\sigma^2} \sim \chi^2_{n-1},$$

with $\bar{X}$ and s^2 statistically independent. Suppose we want to test the **hypothesis** that $\mu = \mu_0$, for some specified constant μ_0, (for example, $\mu_0 = 20.0$) is a plausible model. More specifically, we want to test

$$\begin{array}{lcll} H_0 & : & \mu = \mu_0, & \text{the } \mathbf{NULL} \text{ hypothesis, against} \\ H_1 & : & \mu \neq \mu_0, & \text{the } \mathbf{ALTERNATIVE} \text{ hypothesis.} \end{array}$$

So, we want to test whether H_0 is true, or whether H_1 is true. In the case of a Normal sample, the distribution of $\bar{X}$ is Normal, and

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Longrightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N\left(0, 1\right),$$

where Z is a **random variable**. Now, when we have observed the data sample, we can calculate $\overline{x}$, and therefore we have a way of testing whether $\mu = \mu_0$ is a plausible model; we calculate $\overline{x}$ from $x_1, ..., x_n$, and then calculate

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}.$$

If H_0 is **true**, and $\mu = \mu_0$, then the **observed** z should be an observation from an $N(0, 1)$ distribution (as $Z \sim N(0, 1)$), that is, it should be near zero with high probability. In fact, z should lie between -1.96 and 1.96 with probability $1 - \alpha = 0.95$, say, as

$$P[-1.96 \leq Z < 1.96] = \Phi(1.96) - \Phi(-1.96) = 0.975 - 0.025 = 0.95.$$

If we observe z to be outside of this range, then there is evidence that H_0 is **not true**.
So, basically, if we observe an extreme value of z, either H_0 is true, but we have observed a rare event, or we prefer to disbelieve H_0 and conclude that the data contains evidence against H_0. Notice the asymmetry between H_0 and H_1. The null hypothesis is 'conservative', reflecting perhaps a current state of belief, and we are testing whether the data is consistent with that hypothesis,

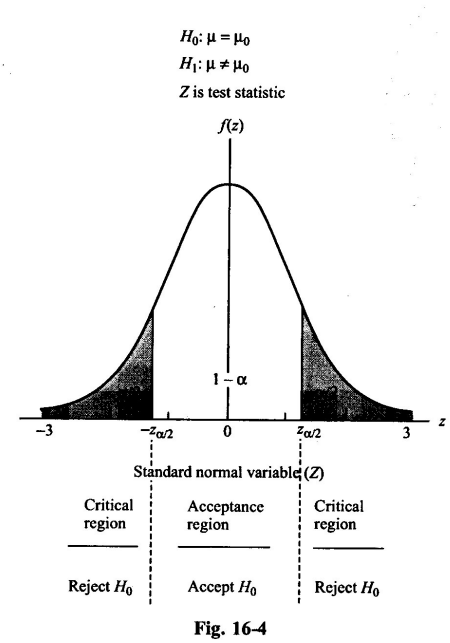

Figure 7.1: CRITICAL REGION IN A Z-TEST (taken from Schaum's ELEMENTS OF STATISTICS II, *Bernstein & Bernstein.*

only rejecting H_0 in favour of the alternative H_1 if evidence is clear i.e. when the data represent a rare event under H_0.
As an alternative approach, we could calculate the probability p of observing a z value that is **more extreme** than the z we did observe; this probability is given by

$$p = \begin{cases} 2\Phi(z), & z < 0, \\ 2(1 - \Phi(z)), & z \geq 0. \end{cases}$$

If this p is very small, say $p \leq \alpha = 0.05$, then again there is evidence that H_0 is **not true**. This approach is called **significance testing**.
In summary, we need to assess whether z is a **surprising** observation from an $N(0, 1)$ distribution - if it is, then we **reject** H_0. Figure 6.1 depicts the "critical region" in a Z-test.

7.3.2 HYPOTHESIS TESTING TERMINOLOGY

There are five crucial components to a hypothesis test, namely:

- the **TEST STATISTIC**;
- the **NULL DISTRIBUTION** of the statistic;
- the **SIGNIFICANCE LEVEL** of the test, usually denoted by α;
- the **P-VALUE**, denoted p;
- **CRITICAL VALUE(S)** of the test.

In the Normal example given above, we have that:

- z is the **test statistic**;
- The distribution of random variable Z if H_0 is true is the **null distribution**;
- $\alpha = 0.05$ is the **significance level** of the test (choosing $\alpha = 0.01$ gives a "stronger" test);
- p is the **p-value** of the test statistic under the null distribution;
- the solution C_R of $\Phi(C_R) = 1 - \alpha/2$ gives the **critical values** of the test $\pm C_R$. These critical values define the boundary of a **critical region**: if the value z is in the critical region we reject H_0.

7.3.3 THE t-TEST

In practice, we will often want to test hypotheses about μ when σ is unknown. We cannot perform the Z-test, as this requires knowledge of σ to calculate the z statistic. Recall that we know the sampling distributions of $\bar{X}$ and s^2, and that the two estimators are statistically independent. Now, from the properties of the Normal distribution, if we have independent random variables $Z \sim N(0,1)$ and $Y \sim \chi^2_\nu$, then we know that random variable T defined by

$$T = \frac{Z}{\sqrt{Y/\nu}}$$

has a Student-t distribution with ν degrees of freedom. Using this result, and recalling the sampling distributions of $\bar{X}$ and s^2, we see that

$$T = \frac{\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)s^2/\sigma^2}{(n-1)}}} = \frac{(\bar{X} - \mu)}{s/\sqrt{n}} \sim t_{n-1} :$$

T has a Student-t distribution with $n-1$ degrees of freedom, denoted $St(n-1)$, which does not depend on σ^2. Thus we can repeat the procedure used in the σ known case, but use the sampling distribution of T rather than that of Z to assess whether the value of the test statistic is "surprising" or not. Specifically, we calculate

$$t = \frac{(\overline{x} - \mu)}{s/\sqrt{n}}$$

and find the critical values for a $\alpha = 0.05$ test by finding the ordinates corresponding to the 0.025 and 0.975 percentiles of a Student-t distribution, $St(n-1)$ (rather than a $N(0,1)$) distribution.

7.3.4 TEST FOR σ

The Z-test and t-test are both tests for the parameter μ. To perform a test about σ, say

$$\begin{aligned} H_0 &: \sigma = \sigma_0, \\ H_1 &: \sigma \neq \sigma_0, \end{aligned}$$

we construct a test based on the estimate of variance, s^2. In particular, we saw from Theorem 7.2.1 that the random variable Q, defined by

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

if the data have an $N(\mu, \sigma^2)$ distribution. Hence if we define test statistic value q by

$$q = \frac{(n-1)s^2}{\sigma_0}$$

then we can compare q with the critical values derived from a χ^2_{n-1} distribution; we look for the 0.025 and 0.975 quantiles - note that the chi-squared distribution is not symmetric, so we need two distinct critical values.

7.3.5 TWO SAMPLE TESTS

It is straightforward to extend the ideas from the previous sections to two sample situations where we wish to compare the distributions underlying two data samples. Typically, we consider sample one, $x_1, ..., x_{n_X}$, from a $N(\mu_X, \sigma^2_X)$ distribution, and sample two, $y_1, ..., y_{n_Y}$, independently from a $N(\mu_Y, \sigma^2_Y)$ distribution, and test the equality of the parameters in the two models. Suppose that the sample mean and sample variance for samples one and two are denoted $(\overline{x}, s^2_X)$ and $(\overline{y}, s^2_Y)$ respectively.

1. First, consider the hypothesis testing problem defined by

$$H_0 : \mu_X = \mu_Y,$$
$$H_1 : \mu_X \neq \mu_Y,$$

when $\sigma_X = \sigma_Y = \sigma$ is known, so the two samples come from normal distributions with the same, known, variance. Now, from the sampling distributions theorem we have, under H_0

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma^2}{n_X}\right), \qquad \bar{Y} \sim N\left(\mu_Y, \frac{\sigma^2}{n_Y}\right), \Longrightarrow \bar{X} - \bar{Y} \sim N\left(0, \frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y}\right),$$

since $\bar{X}$ and $\bar{Y}$ are independent. Hence by the properties of normal random variables

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim N(0, 1),$$

if H_0 is true, giving us a test statistic z defined by

$$z = \frac{\overline{x} - \overline{y}}{\sigma\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}},$$

which we can compare with the standard normal distribution. If z is a surprising observation from $N(0, 1)$, and lies in the critical region, then we reject H_0. This procedure is the **Two Sample Z-Test**.

2. If we can assume that $\sigma_X = \sigma_Y$, but the common value, σ, say, is unknown, we parallel the one sample t-test by replacing σ by an estimate in the two sample Z-test. First, we obtain an estimate of σ by "pooling" the two samples; our estimate is the **pooled estimate**, s_P^2, defined by
$$s_P^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2},$$
which we then use to form the test statistic t defined by
$$t = \frac{\overline{x} - \overline{y}}{s_P\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}.$$
It can be shown that if H_0 is true then t should be an observation from a Student-t distribution with $n_X + n_Y - 2$ degrees of freedom. Hence we can derive the critical values from the tables of the Student-t distribution.

3. If $\sigma_X \neq \sigma_Y$, but both parameters are known, we can use a similar approach to the one above to derive a test statistic z defined by
$$z = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}},$$
which has an $N(0, 1)$ distribution if H_0 is true.

4. If $\sigma_X \neq \sigma_Y$, but both parameters are unknown, we can use a similar approach to the one above to derive a test statistic t defined by
$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}.$$
The distribution of this statistic when H_0 is true is not analytically available, but can be adequately approximated by a *Student* (m) distribution, where
$$m = \frac{(w_X + w_Y)^2}{\left(\frac{w_X^2}{n_X - 1} + \frac{w_Y^2}{n_Y - 1}\right)},$$
with
$$w_X = \frac{s_X^2}{n_X}, \qquad w_Y = \frac{s_Y^2}{n_Y}.$$

Clearly, the choice of test we use depends on whether $\sigma_X = \sigma_Y$ or not. We may test this hypothesis formally; to test
$$H_0 : \sigma_X = \sigma_Y,$$
$$H_1 : \sigma_X \neq \sigma_Y,$$
we compute the test statistic
$$q = \frac{s_X^2}{s_Y^2},$$

which has as null distribution the F distribution with $(n_X - 1, n_Y - 1)$ degrees of freedom. This distribution can be denoted $F(n_X - 1, n_Y - 1)$, and its quantiles are tabulated. Hence we can look up the 0.025 and 0.975 quantiles of this distribution (the F distribution is not symmetric), and hence define the critical region. Informally, if the test statistic value q is very small or very large, then it is a surprising observation from the F distribution and hence we reject the hypothesis of equal variances.

HYPOTHESIS TESTING SUMMARY In general, to test a hypothesis H_0, we consider a statistic calculated from the sample data. We derive mathematically the probability distribution of the statistic, considered as a random variable, when the hypothesis H_0 is true, and compare the actual observed value of the statistic computed from the data sample with the hypothetical probability distribution. We ask the question "Is the value a likely observation from this probability distribution "? If the answer is "No", then reject the hypothesis, otherwise accept it.

7.4 ESTIMATION

Definition 7.4.1 Let $X_1, ..., X_n$ be a random sample from a distribution with mass/density function f_X that depends on a (possibly vector) parameter θ. Then $f_{X_1}(x_1) = f_X(x_1; \theta)$, so that

$$f_{X_1,...,X_k}(x_1, ..., x_k) = \prod_{i=1}^{k} f_X(x_i; \theta).$$

A statistic $T = t(X_1, ..., X_n)$ that is used to represent or estimate a function $\tau(\theta)$ of θ based on an observed sample of the random variables $x_1, ..., x_n$ is an **estimator**, and $t = t(x_1, ..., x_n)$ is an **estimate**, $\hat{\tau}(\theta)$, of $\tau(\theta)$.

7.4.1 ESTIMATION TECHNIQUES I: METHOD OF MOMENTS

Suppose that $X_1, ..., X_n$ is a random sample from a probability distribution with mass/density function f_X that depends on a vector parameter θ of dimension k, and suppose that a sample $x_1, ..., x_n$ has been observed. Let the rth moment of f_X be denoted μ_r, and let the rth sample moment, denoted m_r be defined for $r = 1, 2, \ldots$ by

$$m_r = \frac{1}{n} \sum_{i=1}^{n} x_i^r.$$

Then m_r is an **estimate** of μ_r, and

$$M_j = \frac{1}{n} \sum_{i=1}^{n} X_i^r$$

is an **estimator** of μ_r.

PROCEDURE : The method of moments technique of estimation involves matching the theoretical moments $\mu_r \equiv \mu_r(\theta)$ to the sample moments $m_r, r = 1, 2, \ldots, l$, for suitable l, and solving for θ. In most situations taking $l = k$, the dimension of θ, suffices: we obtain k equations in the k elements of vector θ which may be solved simultaneously to find the parameter estimates. We may, however, need $l > k$. Intuitively, and recalling the **Weak Law of Large Numbers**, it

is reasonable to suppose that there is a close relationship between the theoretical properties of a probability distribution and estimates derived from a large sample. For example, we know that, for large n, the sample mean converges in probability to the theoretical expectation.

7.4.2 ESTIMATION TECHNIQUES II: MAXIMUM LIKELIHOOD

Definition 7.4.2 Let random variables $X_1, ..., X_n$ have joint mass or density function, denoted $f_{X_1,...,X_k}$, that depends on a vector parameter $\theta = (\theta_1, ..., \theta_k)$. Then the joint/mass density function considered as a function of θ for the (fixed) observed values $x_1, ..., x_n$ of the variables is the **likelihood function**, $L(\theta)$:

$$L(\theta) = f_{X_1,...,X_n}(x_1, ..., x_n; \theta).$$

If $X_1, ..., X_n$ represents a random sample from joint/mass density function f_X

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i; \theta).$$

Definition 7.4.3 Let $L(\theta)$ be the likelihood function derived from the joint/mass density function of random variables $X_1, ..., X_n$, where $\theta \in \Theta \subseteq \mathbb{R}^k$, say, and Θ is termed the **parameter space**. Then for a fixed set of observed values $x_1, ..., x_n$ of the variables, the estimate of θ termed the **maximum likelihood estimate** (MLE) of θ, $\widehat{\theta}$, is defined by

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} L(\theta).$$

That is, the maximum likelihood estimate is the value of θ for which $L(\theta)$ is maximized in the parameter space Θ.
DISCUSSION : The method of estimation involves finding the value of θ for which $L(\theta)$ is maximized. This is generally done by setting the first partial derivatives of $L(\theta)$ with respect to θ_j equal to zero, for $j = 1, ..., k$, and solving the resulting k simultaneous equations. But we must be alert to cases where the likelihood function $L(\theta)$ is not differentiable, or where the maximum occurs on the boundary of Θ! Typically, it is easier to obtain the MLE by maximising the (natural) logarithm of $L(\theta)$: we maximise $l(\theta) = \log L(\theta)$, the **log-likelihood**.

THE FOUR STEP ESTIMATION PROCEDURE: Suppose a sample $x_1, ..., x_n$ has been obtained from a probability model specified by mass or density function $f_X(x; \theta)$ depending on parameter(s) θ lying in parameter space Θ. The maximum likelihood estimate is produced as follows;

1. Write down the likelihood function, $L(\theta)$.

2. Take the natural log of the likelihood, collect terms involving θ.

3. Find the value of $\theta \in \Theta$, $\widehat{\theta}$, for which $\log L(\theta)$ is maximized, for example by differentiation. Note that, if parameter space Θ is a bounded interval, then the maximum likelihood estimate may lie on the boundary of Θ. If the parameter is a k vector, the maximization involves evaluation of partial derivatives.

4. Check that the estimate $\widehat{\theta}$ obtained in STEP 3 truly corresponds to a maximum in the (log) likelihood function by inspecting the second derivative of $\log L(\theta)$ with respect to θ. In the single parameter case, if the second derivative of the log -likelihood is negative at $\theta = \widehat{\theta}$, then $\widehat{\theta}$ is confirmed as the MLE of θ (other techniques may be used to verify that the likelihood is maximized at $\hat{\theta}$).