INTRODUCTION TO MATHEMATICAL STATISTICS II

Semester 1, 2004

(Course Code 2002)

Lecturer: Associate Professor Patty Solomon
Statistics, School of Mathematical Sciences
The University of Adelaide

These notes are copies of the overhead transparencies shown in the lectures, and are intended as a guide to this course.

1 INTRODUCTION

1.1 What is Statistics?

A truth: Statistics is an *enabling* discipline. Statisticians have by training the skills of synthesis, empirical investigation, modelling and interpretation which are crucial to application areas such as engineering, finance and bioinformatics.

Statistics is as much the art as the science of collecting, modelling, analysing, and interpeting data.

'*Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.*'
H.G. Wells

1.2 Why do you need this subject?

IMS is about
• building probability models, and
• describing and understanding the properties of those models.

We use models to describe reality. So we want to know:
Do our models fit the observed data or facts?
How do we determine which models are the best for describing the system under study?

It is the presence of *variability* in the real world which underpins the need for probability models which quantify the variation or *uncertainty* in outcomes.

Probability theory is an important field of study in its own right, but we use it primarily as a tool for modelling and analysing data that in some vague sense have a random or chance character.

1.3 Examples

• How should we design a clinical trial to compare a new treatment for leukaemia with standard treatments?

• In the Ash Wednesday bushfires, vast areas of southern Australia were burnt, including many houses. What were the factors which increased or decreased the risk of houses being burnt?

• Algal blooms: data have been collected over many years on the occurrence of algal blooms along the River Murray. What can we say about the conditions which influence the occurrence of these blooms?

• Writing software: a software house supplies computer software under contract. How do we estimate the cost of providing the software, and how can we improve our prediction of software costs over time?

1.4 Some motivating examples

1. Ohm's Law tells us that

$$V = IR$$

where V is the voltage, I is the current, R is the resistance.

This is a *deterministic* model. Suppose 20 electrical engineering students all set up circuits with the same current and resistance, and all measure the voltage.

How many different voltages will be observed?

A better approach is to use a *probabilistic* or *stochastic* model

$$V = IR + \epsilon$$

where ϵ represents random error.

2. DNA sequences

The DNA of an organism consists of very long sequences from an alphabet of four letters called *nucleotides*: *a g c* and *t* for adenine, guanine, cytosine, and thymine. These sequences undergo change within any population over the course of many generations, and random mutations arise and become fixed in the population. Therefore two rather different sequences may well derive from a common ancestor.

Suppose we have two small DNA sequences from two different species, where the arrows indicate paired nucleotides that are the same in both sequences

```
 ↓     ↓        ↓        ↓  ↓  ↓     ↓
 g  g  a  g  a  c  t  g  t  a  g  a  c
 g  a  a  c  g  c  c  c  t  a  g  c  c
 ↓                       ↓  ↓  ↓
 a  g  c  t  a  a  t  g  c  t  a  t  a
 a  c  g  a  g  c  c  c  t  t  a  t  c
```

We wish to gauge whether the two sequences show significant similarity to indicate whether they have a remote common ancestor.

If the sequences were each generated at random, with the four letters *a g c* and *t* having equal probabilities of occurring at any position, then the two sequences should tend to agree at about one quarter of the positions.

The two sequences agree at 11 out of 26 positions.

How unlikely is this outcome if the sequences were generated at random?

Probability theory shows that under the assumption of equal probabilities for *a g c* and *t* at any site, and independence of the nucleotides, the probability of 11 or more matches in a sequence comparison of length 26 is approximately 0.04.

Thus our observation of 11 *matches gives evidence that something other than chance is at work.*

3. Convolution

Convolution forms the basis of the method of *backcalculation* for estimating past HIV infection incidence and predicting future diagnoses of AIDS.

A simple model for AIDS incidence f_A is

$$f_A(a) = \int_y f_X(a-y) f_Y(y) dy$$

where f_X is the *density function* for the time from infection with HIV to development of AIDS, known as the incubation period, and f_Y is the density function for the incidence of HIV infection.

In practice, we observe f_A, assume we know the incubation distribution f_X, and invert the above equation to estimate the past HIV infection f_Y. We can then substitute these estimates of HIV incidence back into the equation and predict future cases of AIDS.

4. Statistical modelling of BSE

BSE has a long and variable incubation period which means that cows showing signs of disease now were infected many years ago. We use these sorts of models to *estimate* the past pattern of infection and to predict future cases of BSE.

A 'simple' model for the *hazard* of infection at time t of horizontal transmission of prions between an infected and susceptible host (i.e. cow) is

$$\int_0^{t-t_0} \beta\Psi(\tau) f(t - t_0 - \tau | t_0) d\tau$$

where β is the age-dependent transmission coefficient, Ψ represents the *expected* infectivity of an individual at time τ since infection, and f is the *density* of hosts born at time t_0 who were infected time τ ago.

2 PROBABILITY

The mathematical theory of probability has been applied to a wide variety of phenomena, for example:

• In genetics as a model for mutations and ensuing natural variability.

• There are highly developed theories that treat noise in electrical devices and communication systems as random processes.

• Many models of atmospheric turbulence use concepts of probability theory.

• Actuarial science, which is used by insurance companies, relies heavily on the tools of probability theory.

• Probability theory is used to study complex systems and improve their reliability, such as in modern commercial or military aircraft.

2.1 Notation and axioms

[WMS, Chapter 2]

Sample space: $\mathcal{S}$ is the set of all possible outcomes.

Event: $A, B, ...$ is a combination of outcomes, and a *subset* of the sample space $\mathcal{S}$.

Probability: is a measure, or function, that tells you the size of the sets.
The probability of an event A is denoted $P(A)$.
It assigns a numerical value to each outcome and event in the sample space, according to specified rules.

Note: a sample space may be discrete (possibly countable) or continuous. WMS (p. 26) refer to 'simple events' rather than 'outcomes'.
The 'sample space' is also referred to as the 'outcome space'.

e.g. The annual rainfall for a given city could take any non-negative value:

$$\mathcal{S} = \{x | x \geq 0, x \in R\}$$

e.g. The number of cars passing a given point on the road in 1 hour could take any non-negative integer:

$$\mathcal{S} = \{x | x = 0, 1, 2, 3, ...\}$$

N.B. Read the '|' as 'given'.

e.g. of an *event*: rainfall less than 600mm in a year:

$$A = \{x | 0 \leq x < 600\}$$

Axioms of probability:

Axiom 1: For any set A, $P(A) \geq 0$.

Axiom 2: $P(\mathcal{S}) = 1$. This is the certain event.

Axiom 3: (Addition Rule.) If $A_1, ..., A_n$ is a set of mutually exclusive events, then

$$P(A_1 \cup A_2 \ldots \cup A_n) = P(A_1) + ... + P(A_n).$$

If we let $A = A_1 \cup A_2 \ldots \cup A_n$, and $A_1, ..., A_n$ are mutually exclusive, i.e. disjoint, then $A_1, ..., A_n$ is said to be a *partition* of A.

[WMS, p.29]

What we mean by a Distribution: For any partition of $\mathcal{S}$, the probability gets 'distributed' onto each member of the partition and it all adds up to 1. In the case of a *countable* sample space $\mathcal{S}$, once we assign probabilities to all the 'outcomes', then we can find the probability of any event we like by summation. (This is easier said than done, as we shall see.)

****

We can also derive a number of results from these basic ones:

Complements: $P(\bar{A}) = 1 - P(A)$.

Differences: If A is contained in B (we write $A \subset B$), then

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

Inclusion-Exclusion:

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

[Check these yourself using set theory or draw the Venn diagrams; WMS p. 22.]

2.2 Equally likely outcomes

Often, we can safely assume that outcomes are equally likely.

Examples: Rolling dice; tossing a fair coin twice.

Why? If we can assume our coin or die is perfect, the answer follows by *symmetry*.

So, for example, the perfect coin is our *model*.

But clearly it is not always true that all outcomes are equally likely.

For example, we cannot assume that all 'rainfalls' in a year are equally likely.

If all outcomes are equally likely in a finite set $\mathcal{S}$, then the probability that event A occurs is:

$$P(A) = \frac{\#(A)}{\#(\mathcal{S})}$$

Chance Odds and the Odds Ratio:

Odds are a useful way of *comparing* probabilities. [Note that odds are not covered in WMS.]

If the outcomes are equally likely, the odds *in favour* of A are

$$\text{Odds}(A) = \frac{\#(A)}{\#(\text{Not}A)}$$

or, more generally,

$$\frac{P(A)}{1 - P(A)}$$

The *log odds* is known as the *logit*.

Example: *Investigating the relationship between Apgar score at birth and measured foetal growth retardation in pregnancy.*

The Apgar score assesses a baby's general state of health at birth on a 0-10 scale. Using ultrasound during pregnancy, growth retardation is assessed as 'symmetric' or 'asymmetric'.

An apgar score of < 7 indicates that the baby is not doing too well. Is symmetric or asymmetric growth indicative of apgar score?

A study of 107 babies who were 'small for dates' (smallest 5% of babies) was conducted.

The data are:

		Symm	Asymm	Tot
Apgar	< 7	2	33	35
score	≥ 7	14	58	72
		16	91	107

What are the odds of an Apgar score < 7?

We can calculate the odds for each group separately:

- odds of <7 if Symmetric: $2/14 = 0.143$
- odds of <7 if Asymmetric: $= 33/58$.

That is, there is a much higher odds of a low Apgar score with asymmetric growth.

The relative odds (or risk) of a low score in the two groups is the ratio of these two odds, and is called the *odds ratio*:

$$(2/14)/(33/58) = 2 \times 58/(14 \times 33) = 0.25.$$

Note that these quantities are *estimated* odds based on a sample.

2.3 Interpretations of probability

Will it rain tomorrow?

Viewed as *either*:

• Limiting relative frequency (i.e. proportion), *or*

• Subjective opinion i.e. a statement which quantifies the speaker's uncertainty about the outcome and is therefore a personal or subjective notion.

Relative frequency forms the basis of *frequentist statistics*. Subjective opinion forms the basis of *Bayesian statistics*. There has been rigorous debate between these two versions. We will discuss these notions in answer to the question 'will it rain tomorrow?'

Opinion or 'subjective probabilities'.

Often referred to as Bayesian statistics after Rev Thomas Bayes (1763) who first developed what we now know as Bayes' Theorem. In essence, the idea is that we start with some idea of what we think the probability is (a *prior probability*) and then, as we collect information, we update our 'subjective' probability on the basis of that information. Bayes was the first to give us a specific formula for doing that 'updating'.

The difficulties with this are:

- How do you determine what is your prior probability/opinion?
- If we are trying to convince others?
- How do you ensure that your subjective probabilities are consistent?

Discussion example: Doctors quoting the probability of survival.

How else might we develop probabilities?

Symmetry

Do not assume symmetry when you shouldn't!

Year	No.births	Propn.boys
1974	3,159,958	0.51333
1975	3,144,198	0.51305
1976	3,167,788	0.51280
1977	3,326,632	0.51281
1978	3,333,279	0.51283
1979	3,494,398	0.51261
1980	3,612,258	0.51287
1981	3,629,238	0.51258

John Arbuthnot(1710):
'it is odds, if a woman be with child, but it shall be a boy, and if you would know the just odds, you must consider the proportion in the Bills that the males bear to females.'

[Ref: Hacking, I. (1975) *The Emergence of Probability*.]

2.4 Conditional probability and independence

[WMS, p. 50]

All probability statements are, to some extent, conditional; consider the fact that $P(A) = P(A|\mathcal{S})$.

Axiom 4: $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

Read this as 'the probability of A given B'.

Interpretation: we are regarding B as the complete space.

Note: $P(A|B)$ is not necessarily the same as $P(B|A)$.

Example: Consider the cards again.

We have 3 cards, each with two sides: one is red on both sides, one is green on both sides, and one is red on one side and green on the other. We can label them (r_1, r_2), (g_1, g_2), (r_3, g_3) where r and g indicate red and green.

If I pick a card, each of the 6 sides are equally likely. If I tell you one side is red, what is the probability that the other side is red?

Exercise: Toss 2 coins. What is the probability of 2 heads?
Given that the first toss gave a head, what is the probability of 2 heads?

Example: Digitalis therapy is often used to treat congestive heart failure. However, it can lead to digitalis toxicity which is difficult to diagnose. To improve the chances of a correct diagnosis, the concentration of digitalis in the blood can be measured (Rice, p.15).

An historical study investigated the relationship between digitalis concentration in the blood and digitalis intoxication in 135 patients.
Notation:
$T+/T-$: high/low blood titre;
$D+/D-$: digitalis toxicity/or not.

		Digitalis toxicity		
		$D+$	$D-$	Total
Titre	$T+$	25	14	39
	$T-$	18	78	96
	Total	43	92	135

Regard the proportions as probabilities. Then

$$P(D+) = 43/135 = 0.3185.$$

We call this the 'prior probability' of digitalis toxicity.

But the conditional probabilities are

$$P(D+|T+) = 25/39 = 0.641$$

$$P(D+|T-) = 18/96 = 0.188$$

Thus, knowing that the high titre is present *doubles* the probability of toxicity.

Note how this is 'evidence' that can be included in the assessment of future patients.

We can of course find the other conditional probabilities:
$P(T+|D+) = 25/43 = 0.581$. This is known as the *sensitivity* of the test.
$P(T-|D-) =$
$P(T+|D-) =$
$P(T-|D+) =$

Technically, of course, these are all 'proportions' and only become probabilities if either (i) we use large enough samples such that the relative frequency is close to the true probability, or (ii) we think of choosing one of these people at random from the population.

In practice, you should assess whether (i) or (ii) is reasonable. If not, interpret the results with caution.

Multiplication Rule:

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

which follows directly from Axiom 4.

It is useful because in practice it is often easier to find $P(A|B)$ or $P(B)$ than the joint probability.

Tree Diagrams can be helpful to depict the Multiplication Rule in action:
The idea is that each branch in the tree represents a possible outcome. The paths to particular events which occur in sequence have the property that the probabilities at the nodes have to sum to 1.

Example: A system has 2 electrical components. The first component has a probability of failure of 10%. If the first component fails, the second fails with probability 20%. If the first works, then second fails with probability 5%.

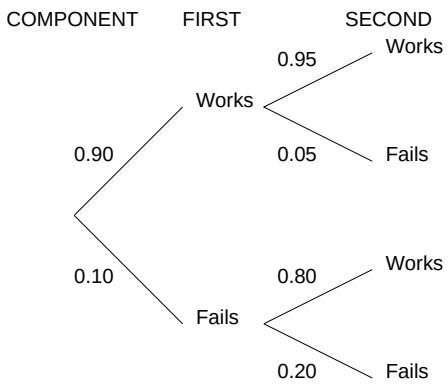

Fig. 2.1: Tree diagram.

Let B be the event that the first component works.
Let A be the event that the second component works.

Find the probability that
- at least one component works:
- exactly one component works:
- the second component works:

Note that there are two ways we can find these probabilities. One is to work out the probabilities along each of the 4 paths and add up the right ones. Alternatively, find each probability separately. We will obtain the solutions in the lectures.

Motivation for Law of Total Probability

$P(A)$ is the probability that the second component works. Clearly it lies between 0.80 and 0.95. (Why?)

If B is the event that the first component works and $\bar{B}$ is that it doesn't, then

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

which is then

$$P(B)P(A|B) + P(\bar{B})P(A|\bar{B}).$$

This is a weighted average of the two conditional probabilities.

Law of total probability

Theorem: If $B_1, ..., B_n$ is a partition of $\mathcal{S}$,

$$P(A) = P(B_1)P(A|B_1) + ... + P(B_n)P(A|B_n).$$

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

This provides a way to *average* conditional probabilities.

How would you represent this in a tree diagram?

Independence

If the probability of the second component working was the same regardless of the first, then

$$P(A|B) = P(A|\bar{B}),$$

and, regardless of the weights, both are equal to $P(A)$. Then B doesn't affect $P(A)$ and we say A and B are independent.

Definition: Events A and B are **independent** if and only if

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

Exercise: Show that $\bar{A}$ and $\bar{B}$ are independent.

2.5 Named Distributions

(i) Bernoulli distribution
Two outcomes, success (S) and failure (F):

Outcome	Failure	Success
X	0	1
Probability	$1-p$	p

p is referred to as a 'parameter' and we often want to estimate it. We write $P(X = 0) = 1-p$ and $P(X = 1) = p$.

The 'numerical' outcome is the *random variable* X. We say 'X has the Bernoulli distribution with parameter p'.

[We can 'draw' such a distribution; called a *probability histogram*.]

(ii) Uniform distribution over a finite set.

Suppose a sample space has a set of n possible outcomes, all equally likely.

Outcome	A_1	A_2	....	A_n
Probability	$1/n$	$1/n$	....	$1/n$

There *may* be a numerical outcome.

It is an important distribution in finite sampling theory.

Examples?

Roll a die. Then $n = 6$ and $P(\text{any outcome}) = 1/6$.

(iii) Empirical distribution (i.e. based on data)

Categorical data:
e.g. smoker/non-smoker/ex-smoker.

For categories $A_1, \ldots, A_m$, we count the number f_i in each category, and give the *proportion* $\widehat{p}_i = f_i/n$ in each.

Note: the (empirical) proportions add to 1.

Measurement data: e.g. time.

Divide the line with breaks at $b_1, ..., b_m$.

If there are f_i obervations in the interval $(b_i, b_{i+1}]$, the height of the bar is

$$\frac{f_i}{n \times (b_{i+1} - b_i)}.$$

Why?

Explanation: Think of each observation as having an area 1/n. We drop them into their bins from above. If the bin is wider, the observations will not stack up as high. The total area is 1 and the area in the ith bin is proportional to the number falling into that bin. The height must then be the area divided by the width.

2.6 Sequences of events

This simply extends the Multiplication Rule to n events in sequence:
$P(A_1 \cap A_2 \cap \ldots \cap A_n)$

$$= P(A_1)P(A_2|A_1)P\{A_3|(A_1 \cap A_2)\}\ldots$$

Tree diagrams can be extended to have multiple branches at each node and multiple nodes. We may only have some of the information. Independence is not so simple here.

Exercise: *Reliability of two components in parallel and in series.*
Suppose we know the separate probabilities that the two components work are $P(W_1) = 0.9$ and $P(W_2) = 0.8$. Each probability is known as the *reliability*.

Explore the effects of assuming independence of the two components on the overall reliability of the system firstly in series, then in parallel.

Geometric distribution: a discrete waiting time distribution.

Suppose we conduct a sequence of independent Bernoulli trials, where p is the probability of success at each trial. Repeat the trials until we get a success.

What is the probability that we stop at k trials?

Let X be the random variable which takes values equal to the number of trials until we get the first success. Then

X	1	2	3	...
Prob	p	qp	q^2p	...

The *probability function* for the geometric distribution is

$$P(X = k) = q^{k-1}p, \quad k = 1, 2, \ldots.$$

In principle, the sequence of trials can go on indefinitely if a success is never obtained (e.g. tossing a coin and never getting a head). It is called the geometric distribution because its probabilities are terms in a geometric series.

Exercise: Verify that these probabilities add to 1 by showing $\sum_{k=1}^{\infty} q^{k-1}p = 1$.

We will meet this distribution again in Chapter 3.

Example: Gambler's Rule.
If you have a probability $p = 1/N$ of success each time you play a game over and over again, the Gambler's Rule is that you need to play about $2N/3$ games to have a better than 50% chance of at least one win.

How can we show this?

****

The method of solving many of these problems is:
• get a notation,
• think of most problems as sequences of events,
• do a tree diagram,
• break up the answer into bits that can be found relatively easily.

Independence for > 2 events:

Three events A, B, C are independent provided

$$P(B|A) = P(B|\bar{A}) = P(B),$$

and

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

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

It follows that:

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

This is a very strong condition. For n events, there are 2^n possible intersections whose probabilities are all determined by just the probabilities of the n events.

Pairwise independence:
A weaker form of independence, requiring only that A_i and A_j be independent for each pair.

Exercise: Toss 2 coins. Let A be the event that the two coins give the same result.
Then show that the events H_1 (head first time), H_2 (head second time) and A are pairwise independent, but are not fully independent.

2.7 Bayes' Rule

Example: Digitalis revisited.
We know that the probability of a positive titre is $P(T+) = 39/135$.
If we are also given that the person is $D+$, then it is of interest to ask: what is $P(T+|D+)$?

In other words, if we have the *additional* information about the toxicity, how does that change our probabilities?

Bayes' Rule gives a general formula for updating probabilities in the light of new information.

Suppose we have a partition $B_1, ..., B_n$ and we know the probabilities $P(B_i)$ of each.

Now suppose that we find out that the event A has occurred. How do the probabilities of the B_i's change?

In other words, what is $P(B_i|A)$?

How can we visualise this?

We know

$$P(A \cap B_i) = P(B_i|A)P(A) = P(A|B_i)P(B_i).$$

We also know that if $B_1, \ldots, B_n$ partition A, then

$$P(A) = P(B_1)P(A|B_1) + ... + P(B_n)P(A|B_n)$$

by the Law of Total Probability.

From above,

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)},$$

and we are then led to **Bayes' Rule**:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + ... + P(A|B_n)P(B_n)}.$$

Two senses of Bayesian:

1. Frequentist: Bayes' Rule allows the inversion of order in a conditional probability statement.

2. Non-frequentist: the *prior distribution* (here $P(B_i)$) reflects a personal degree of belief which is updated in the light of data (here $P(A|B_i)$) to give a *posterior distribution* for B_i, i.e., $P(B_i|A)$, $i = 1, \ldots, n$.

So now return to our digitalis example. For convenience, the data are given again here:

Recall this study investigated the relationship between digitalis concentration in the blood and digitalis intoxication in 135 patients; the notation is $T+/T-$: high/low blood titre; $D+/D-$: digitalis toxicity/or not.

		Digitalis toxicity		
		$D+$	$D-$	Total
Titre	$T+$	25	14	39
	$T-$	18	78	96
	Total	43	92	135

To keep the notation consistent in this section, let events $T+$ and $T-$ be B_1 and B_2 respectively; note that since there are only two outcomes of positive or negative titre, $B_2 = \bar{B}_1$. The event A is digitalis intoxication $D+$; $\bar{A}$ is $D-$.

Now

$$P(B_1) = 39/135, \quad P(B_1|A) = ?$$

The answer is just

$$\begin{aligned} \frac{P(A|B_1)P(B_1)}{P(A)} &= \frac{(25/39) \times (39/135)}{(43/135)} \\ &= 25/43 \\ &= 0.581, \end{aligned}$$

and the other results follow similarly.

We can see how this works by giving the earlier table with the four individual joint probabilities $P(B_1 \cap A)$, etc:

	A	$\bar{A}$	Total
B_1	.185	.104	.289
B_2	.133	.578	.711
Total	.318	.682	1.000

Convert the (joint) probabilities in the previous table to the conditional probabilities $P(A|B_i)$ by dividing each element by the row total:

	A	$\bar{A}$	Total
B_1	.641	.359	1.000
B_2	.188	.812	1.000

And similarly, convert the joint probabilities to $P(B_i|A)$ by dividing by the column totals:

	A	$\bar{A}$	
B_1	.581	.152	
B_2	.419	.848	
Total	1.000	1.000	

Note how we can move between these tables.

Revisiting odds: [Not in WMS]

If there are just 2 choices or hypotheses, so that $B_1 = B$ and $B_2 = \bar{B}$, then we can consider the *odds* $P(B)/P(\bar{B})$ of B occurring.

If we are then given further information that A has occurred, how do the odds change?

Likelihood:
Consider first a distribution, e.g. the geometric distribution. If we know the value of the parameter p, we can give the probability of each value of X occurring, i.e. $P(X = k|p) = pq^{k-1}$.

Suppose, however, that we have already observed that X takes the value k, a number, but that p is unknown. **The probability is no longer a function of k, but only depends on the unknown p.** In statistics, the latter probability is referred to as a *likelihood*; namely a probability distribution which we consider to be a function of the (unknown) parameter (here p) for a given value of the data k. We write

$$L(p|X = k) = pqk - 1$$

Here, we know that A has occurred (A is the 'data') and we want to assess the odds of the unknown 'parameter', namely whether B or $\bar{B}$ is more likely.

Bayes' Rule shows that:

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

Can you see why?

The ratio

$$\frac{P(A|B)}{P(A|\bar{B})}$$

can be regarded as a 'likelihood ratio'; that is, the relative likelihood of A occurring, given the two different hypotheses.

Thus,
Posterior Odds = Likelihood Ratio x Prior Odds.

Example: Lie detector tests.
B=telling the truth, $\bar{B}$= lying.
A, $\bar{A}$ = lie detector reads positive or negative.

Suppose from historical reliability studies we know:

$$P(A|\bar{B}) = 0.88, \quad P(\bar{A}|B) = 0.86.$$

Say that before the test, $P(B) = 0.99$.

Suppose that an employee tests positive: how does this affect our view of the employee?

What is the probability that the employee is in fact telling the truth (i.e. what is $P(B|A)$)?

We will answer this question in the lecture using Bayes' Rule.

We will also investigate this question in terms of odds:

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

The LHS is the posterior odds that the employee is telling the truth. We want to find this quantity.

3 DISCRETE RANDOM VARIABLES

3.1 Random variables

Now consider outcomes or events which have a numerical value.

In mathematical terms, this is a function:

$$Y : \mathcal{S} \to R$$

which maps each outcome in the sample space $\mathcal{S}$ onto a single point on the real line.

Such a function is called a **random variable.**

So a random variable is simply a variable which takes values according to some probability distribution.

If the set of possible values for Y is countable, the random variable is called **discrete**.

Notation:

$X, Y, Z, \ldots$	random variables	what we might get
$x, y, z, \ldots$	values	what we got

Each time we choose an outcome from $\mathcal{S}$, we get a particular outcome y from the possible values of Y.

Note that random variables are denoted by capital letters.

Values of random variables are denoted by lower case letters.

Examples

Y is the number of people diagnosed with Hepatitis C in South Australia each year; $Y = 0, 1, 2, \ldots$.

Y is the number of farm animals slaughtered during an outbreak of foot and mouth disease; $Y = 0, 1, 2, \ldots$.

Y is the number of heads in two tosses of a fair coin; $Y = 0, 1, 2$.

3.2 Probability distributions

Probability distribution of Y: the probability associated with each possible value of Y,

$$p(y) = P(Y = y), \quad y \in \text{range}(Y).$$

An **event** is a statement about Y, e.g. $Y \leq 3$.

If A is an event,

$$P(A) = P(Y \in A) = \sum_{y \in A} P(Y = y).$$

i.e. the probability of event A is the sum of the probabilities of outcomes that belong to the event.

Notes:
(i) Discrete: the probabilities add to 1, since each outcome maps onto a single y value and takes its probability with it.
(ii) Continuous: $P(Y = y)$ is replaced by the *density* function $f(y)$, where the integral of $f(y)$ over the range of values for Y is 1; i.e. probabilities are given by areas under the curve $f(y)$.

Examples: We have already met the *Bernoulli*, *uniform* (Section 2.5), and *geometric* distributions (Section 2.6).

As with probabilities in general, we have:

- $0 \leq p(y) \leq 1$ for all $y \in range(Y)$.
- $\sum_y p(y) = 1$.

3.3 Expectation

The *expectation*, *expected value* or *population mean* of the random variable Y is

$$E(Y) = \sum_{\text{all } y} yP(Y = y)$$

Notes:
- It is a population parameter, often denoted μ_Y or μ,
- obvious analogy to the sample mean,
- average, weighted by the probability,
- $\sum_y (y - \mu)P(Y = y) = 0$,

indicating that μ is the 'centre of gravity',
- only exists if sum absolutely convergent, i.e.

$$\sum_{\text{all } y} |y|p(y) < \infty$$

Examples:
(i) *Equally likely outcomes.* If Y takes the values $y_1, \ldots, y_n$ with equal probability, then

$$P(Y = y_1) = P(Y = y_2) = \ldots = P(Y = y_n) = \frac{1}{n}$$

and

$$\mu = \sum_y yP(Y = y) = \sum_{i=1}^{n} y_i P(Y = y_i) = \sum_{i=1}^{n} y_i \frac{1}{n} = \bar{y}$$

(ii) *Just two values.*
If Y takes values either a or b with probabilities $(1 - p)$ and p respectively, then

$$\mu = \sum_y yP(Y = y) = a(1 - p) + bp$$

Note that the mean μ shifts between a and b as the probability p moves from 0 to 1.

Functions of a random variable

Suppose we want to find the average kinetic energy of a gas molecule. We know $K = mV^2/2$, and have the distribution of the velocities V. We therefore want the expected value $E(mV^2/2)$.

This leads us to an important result.

Theorem 3.1: If g(y) is any (deterministic) function of Y,

$$E\{g(Y)\} = \sum_{\text{all } y} g(y)P(Y = y) \quad (*)$$

provided the sum is absolutely convergent, i.e.,

$$\sum_y |g(y)|P(Y = y) < \infty.$$

We will typically assume the expectation exists.

We will now prove this Theorem.

Proof: [Not examinable] Any function of a random variable is also a random variable. Let $X = g(Y)$, then by definition

$$E(X) = \mu_X = \sum_x xP(X = x).$$

We need to prove that the rhs is the same as (*) above. Now

$$P(X = x) = \sum_{y:g(y)=x} P(Y = y)$$

by the Addition Rule, where $y : g(y) = x$ is the set of y's mapped onto x by g. So we have, by substituting,

$$\begin{aligned} E(X) &= \sum_x x \left(\sum_{y:g(y)=x} P(Y = y) \right) \\ &= \sum_x \sum_{y:g(y)=x} xP(Y = y) \\ &= \sum_x \sum_{y:g(y)=x} g(y)P(Y = y), \end{aligned}$$

since $x = g(y)$, so that
$E(X) = \sum_y g(y)P(Y = y)$ which is the right-hand-side of (*), as required.
##

This last step follows because the sets $\{y : g(y) = x\}$ are disjoint, and every y belongs to some set.

Note that

$$E[g(Y)] \neq g[E(Y)].$$

Example: Suppose we have a random variable Y with the following probability distribution:

Y	-1	0	1
$P(Y = y)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Let $X = g(Y) = Y^2$. What is $E(X)$?

One easy solution is simply to observe that X takes values

X	0	1
$P(X=x)$	$\frac{1}{2}$	$\frac{1}{2}$

Then

$$E(X) = 0 \times \frac{1}{2} + 1 \times \frac{1}{2} = \frac{1}{2}.$$

Using Theorem 3.1:

$$E(X) = \sum_y g(y)P(Y=y)$$

$$= \sum_y y^2 P(Y=y)$$

$$= (-1)^2 \times \frac{1}{4} + 0^2 \times \frac{1}{2} + 1^2 \times \frac{1}{4} = \frac{1}{2}.$$

The population variance, denoted σ^2, measures the spread of a population:

$$\sigma^2 = \text{Var}(Y) = E[(Y-\mu)^2]$$

$$= \sum_y (y-\mu)^2 P(Y=y)$$

It is known as the *second moment about the mean* or simply the *variance*.

The population standard deviation σ of Y is the square root of the variance.

Notes:
• $E(Y)$ and $\text{Var}(Y)$ can go a long way towards characterising a distribution.
• $\text{Var}(Y) = 0$ if and only if Y has all its probability concentrated at the point $Y = \mu$.
• If $g(Y) = Y^k$, then $E(Y^k)$ is known as the kth *moment* of Y.

Examples:
(i) **Bernoulli distribution**

$$\mu = E(Y) = 1 \times p + 0 \times (1-p) = p.$$

$$\sigma^2 = \text{Var}(Y) = E[(Y-\mu)^2]$$

$$= (1-p)^2 p + (0-p)^2(1-p) = p(1-p).$$

(ii) **Uniform distribution over a finite set**

Let the possible values for Y be $y_1, \ldots, y_n$. We showed earlier that $\mu = E(Y) = \bar{y}$. Then,

$$\sigma^2 = E[(Y-\mu)^2] = (y_1-\mu)^2\frac{1}{n} + \ldots + (y_n-\mu)^2\frac{1}{n}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

(iii) **A special case of the uniform distribution:** We often take $y_i = i$, i.e., each observation is replaced by its rank value. Then we can show in general that

$$\mu = \frac{n+1}{2}$$

and

$$\sigma^2 = \frac{n^2-1}{12}.$$

(*You are asked to find the mean and variance of this uniform distribution in Tutorial 2.*)

These results have important applications to the construction of nonparametric tests based on the ranks of the data.

The geometric distribution

Recall that Y is geometric with probability of success p if

$$P(Y = y) = p(1-p)^{y-1}, \quad y = 1, 2, \ldots$$

If Y has a geometric distribution with probability of success p, then $E(Y) = 1/p$ and $\text{Var}(Y) = (1-p)/p^2$.

We will prove these results in the lecture.

3.4 Expected values of linear functions of random variables

There are several important results, and we will prove each of the following in lectures:

Theorem 3.2:
- $E(c) = c$, for any constant c.
- $E\{cg(Y)\} = cE\{g(Y)\}$, for any constant c.
- $E\{\sum_i c_i g_i(Y)\} = \sum_i c_i E\{g_i(Y)\}$, for any constants c_i.

These results make finding expected values considerably easier.

The proofs are straightforward.

We can use Theorem 3.2 to show that

$$\text{Var}(Y) = E(Y^2) - \mu^2$$

This is an extremely useful result and provides an alternative to finding the variance from first principles.

Examples:
(i) The above result provides an easy proof of the variance for the Bernoulli distribution.

$\sigma^2 = E(Y^2) - p^2$ since we know that $\mu = p$. Now,

$$E(Y^2) = 1^2 \times p + 0^2 \times (1-p) = p.$$

So

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

(ii) Find the variance of the geometric distribution with probability p. (We will work through this in the lecture.)

3.5 Random sampling

In many cases, we sample *items* at random from a *population*. We use this scenario to motivate the *binomial* and *hypergeometric* distributions.

This might be a population of people, e.g. in a city, or parts off a production line.

Consider the case where there are just two outcomes for each item – success or failure.

If the population has size N, of which m are successes, then choosing one item *at random* implies a probability of $p = m/N$ that it will be a success.

What is the probability that the *second* item drawn will be a success? The answer depends on how you do the sampling.

There are two cases:

Sampling with replacement: The item chosen is returned before choosing the next one. Then the probability remains constant at $p = m/N$ for each item drawn.

Taking a random sample in this manner leads to the *binomial* distribution.

Sampling without replacement: The item chosen is not returned before choosing the next one. Then the probability changes each time, e.g. let S_1 be a success on the first draw, etc, then
$P(S_2|S_1) = (m-1)/(N-1)$
$P(S_2|F_1) = m/(N-1)$

This leads to the *hypergeometric* distribution.

The two distributions are closely related.
Under the right conditions we can approximate the hypergeometric distribution by the binomial distribution.

3.6 Binomial distribution

Suppose n individuals are drawn one-by-one from the population, with replacement between the draws. On each draw, it is assumed that each of the N individuals has the same chance of being chosen and the successive draws are assumed to be independent. Then there are N^n possible sequences of choices. Suppose now that we are interested in the distribution of the number of successes y in the sample. Each draw is an independent trial and $p = m/N$ is the probability of success, so the probability of such a sequence is

$$p \dots p(1-p) \dots (1-p) = p^y(1-p)^{n-y}.$$

There are also

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}$$

ways of getting y successes from n draws.

This gives the binomial probability function of obtaining y successes from n independent trials:

$$P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$$

for $y = 0, 1, \ldots, n$; $0 \leq p \leq 1$. We write $\mathrm{B}(n, p)$.

In summary, the binomial situation is:

• n independent Bernoulli trials,
• at each trial, there can be a failure or a success,
• the probability of failure is $(1 - p)$ and the probability of success is p,
• our random variable Y is the number of successes.

The *tree diagram* nicely describes the probability function for the binomial distribution, and we present it here as an alternative derivation.

Recall that in a tree diagram:

• the probability of any pathway is the product of the (conditional) probabilities along that pathway;

• the probability of reaching any node is the sum of the probabilities of all pathways to that node;

• the sum of all probabilities at the *terminating* nodes is 1.

The tree diagram for the binomial distribution with $n = 4$ is

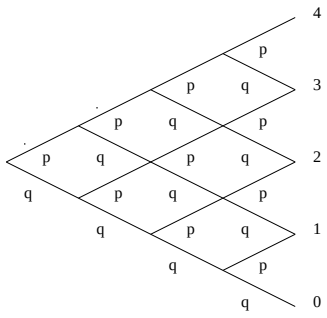

Fig. 3.1: Tree diagram for binomial distribution, $n = 4$.

In general, after n trials, the final nodes at the right hand end have $0, 1, ..., n$ successes.

By comparison, the **geometric** distribution is just followed until the *first* success:

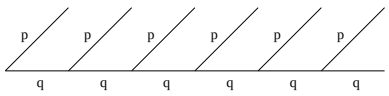

Fig. 3.2: Tree diagram for geometric distribution.

We are only interested in nodes which represented either 0 or 1 success and any number of failures.

Summary of binomial: At any given node, after n trials,

• all pathways have probability $p^y q^{n-y}$, where y is the number of successes,

• the number of paths that lead to that node is the number of ways of ordering y successes among n trials, i.e. $\binom{n}{y}$.

Hence we are led to:
The random variable Y is said to have a *binomial distribution* B(n, p), with n trials and probability of success p if and only if

$$P(Y = y) = \binom{n}{y} p^y q^{n-y}, y = 0, 1, \ldots, n,$$

$0 \leq p \leq 1$.

Note that the binomial probabilities are the terms in a binomial expansion:

$$(p + q)^n = \sum_{y=0}^{n} \binom{n}{y} p^y q^{n-y} = 1.$$

If you are unfamiliar with these ideas you need to work through Section 3.4 in WMS.

Exercises:

(i) Show that $\binom{n}{y} = \binom{n-1}{y} + \binom{n-1}{y-1}$.

(ii) Give a literal explanation of why this formula works.

As an example, consider $n = 20$, $p = 0.5$:

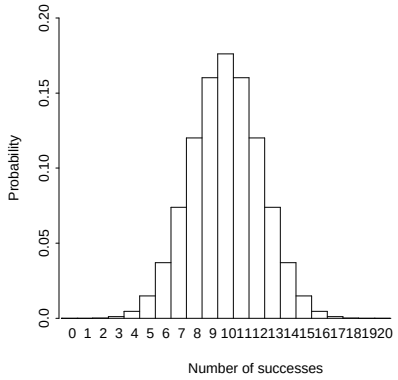

Fig. 3.3: Binomial distribution, n=20, p=0.5

Example: *A simple noise model.* If a single bit (0 or 1) is transmitted over a noisy communications channel, it has probability $p = 0.1$ of being incorrectly transmitted.

Now suppose we use a 'majority decoder'; that is, we send each bit an odd number n of times and we decipher that bit as 0 or 1 according to whichever occurs most often. What is the probability of getting the 'correct' bit for different values of n, say, 3, 5, 7?

Consider $n = 5$, and let Y be the number of bits in error. The probability that the message is received correctly is then the probability of 2 or fewer errors.

Show that $P(Y \leq 2) = 0.9914$.

Mean and variance of binomial distribution

Theorem 3.3: If Y is Bin(n, p), then $E(Y) = np$ and Var$(Y) = np(1-p)$.

Proof:

3.7 Hypergeometric distribution

Here we sample, but the items sampled are not replaced and hence cannot be selected the next time. The probability of selection changes according to what has already been selected.

The number of possible ways of drawing n items, taking order into account, is now

$$N^{(n)} = N(N-1)...(N-n+1),$$

where $n \leq N$.

Note that $N^{(n)} = \binom{N}{n} n!$

How might we get y 'successes' in our sample of size n if there are m 'successes' in the population of N from which we sample without replacement?

Suppose the first y elements in the sample are successes, and the remaining $n-y$ are failures. The probability of this happening is

$$\frac{m}{N}\frac{(m-1)}{(N-1)}\cdots\frac{(m-y+1)}{(N-y+1)}$$
$$\times\frac{(N-m)}{(N-y)}\cdots\frac{(N-m-(n-y)+1)}{(N-y-(n-y)+1)}$$
$$=\frac{m^{(y)}(N-m)^{(n-y)}}{N^{(n)}}$$

But this is just one of the $\binom{n}{y}$ different possible patterns of y successes and $n-y$ failures in an ordered sample of size n, each of which has the same probability. That is

$$P(Y=y)=\binom{n}{y}\frac{m^{(y)}(N-m)^{(n-y)}}{N^{(n)}}=\frac{\binom{m}{y}\binom{N-m}{n-y}}{\binom{N}{n}}.$$

This is the *hypergeometric distribution* for the number of successes Y.

The similarity in the formula for the binomial and hypergeometric distributions can be seen as follows:

Binomial: $P(Y=y)=\binom{n}{y}\frac{m^y(N-m)^{n-y}}{N^n}$
Hypergeometric:

$$P(Y=y)=\binom{n}{y}\frac{m^{(y)}(N-m)^{(n-y)}}{N^{(n)}}$$

where the () around the exponent is as defined earlier.

Note:
(i) the limits on the values Y can take,
(ii) if the *sampling fraction* n/N is low, it is unlikely that you will get the same item sampled again, and the binomial and hypergeometric are very close together.
More formally, $N^{(n)}\approx N^n$ if N is large and n is small relative to N.

In practice, this makes the binomial distribution a useful approximation to the hypergeometric. Hypergeometric probabilities converge to binomial probabilities as N becomes large and m/N is held constant.

Mean and variance

Theorem 3.5: If Y has a hypergeometric distribution, with a sample size n, and m successes in a population of N, then

$$E(Y) = nm/N, \quad \text{var}(Y) = \frac{nm(N-m)(N-n)}{N^2(N-1)}.$$

Proof: Omitted.

Note the similarity with the binomial distribution if we take $p = m/N$.
The factor $(N-n)/(N-1)$ is known as the *finite population correction*.

Here is a comparison of the two distributions. What do you notice, and why?

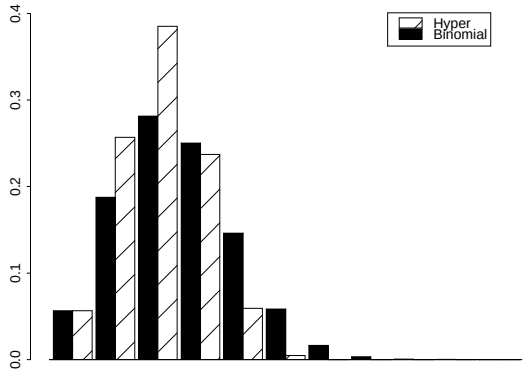

Fig. 3.4: Binomial and Hypergeometric; n=10; m=5, N=20.

Example: A batch of 5000 electrical fuses contains 5% defectives. If a sample of 5 fuses is tested, what is the probability of observing *at least one* defective?

Is this a hypergeometric or binomial situation?

Let Y be the number of defectives observed. It is reasonable to assume that Y is approximately binomial because the batch is large. Then

$$P(Y \geq 1) = 1 - P(Y = 0) = 1 - \binom{5}{0} p^0 q^5$$

$$= 1 - 0.95^5 = 0.226.$$

So even with a small sample, the probability of obtaining at least one defective is still quite high.

N.B. *What assumptions are we making here?*

3.8 Normal distributions

(WMS, p.170)
This is the first *continuous* distribution we have seen. It is included in this Chapter as revision because we need it for the normal approximation to the binomial distribution.

The **normal distribution** N(μ, σ^2) is described by a smooth curve called a *density function* (rather than by a probability histogram):

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/(2\sigma^2)},$$

which has mean μ and standard deviation σ.

Then:
• the total area under the curve is 1;
• the probability of lying within the limits (a, b) is given by the area between vertical lines at $y = a$ and $y = b$. These are obtained by numerical integration (in practice, we use tables or software).

Some examples follow:

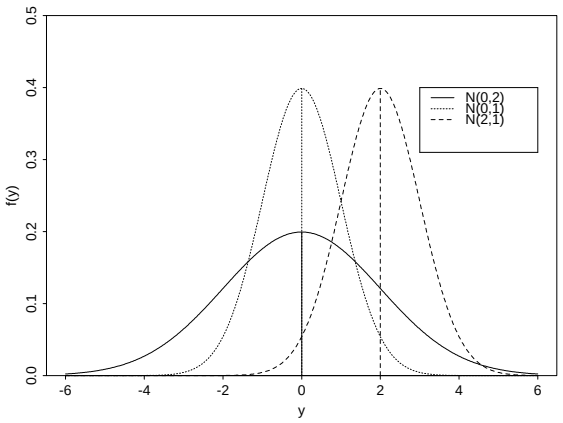

Fig. 3.5: Three normal distributions $N(\mu, \sigma)$.

It is not practical to have a table of probabilities for every pair (μ, σ^2), but happily we only need one table - that for the standard normal distribution.

This is because any random variable $Y \sim N(\mu, \sigma^2)$ can be written as a linear transformation of $Z \sim N(0, 1)$, i.e.,

$$Y = \sigma Z + \mu$$

so that $Z = \frac{Y-\mu}{\sigma}$.

This follows from the general result that if $X \sim N(\mu, \sigma^2)$, then for constants a and b,

$$a + bX \sim N(a + b\mu, b^2\sigma^2).$$

[We will prove this later using moment generating functions.]

The probabilities are determined using the *standard normal distribution* with $\mu = 0$ and $\sigma^2 = 1$ and density function:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

The probability that Y lies between a and b is the probability that the transformed variable $Z = (Y - \mu)/\sigma$ lies between the limits

$$\left(\frac{a-\mu}{\sigma}\right), \left(\frac{b-\mu}{\sigma}\right).$$

If $\Phi(z) = P(Z \leq z)$, then we can tabulate $\Phi(z)$, and the required probability is

$$\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) = \Phi(z_b) - \Phi(z_a).$$

• The tables (e.g. WMS, p.792, Table 4) give the probability of being to the **right** of a given point, i.e.

$$P(Z > z) = 1 - \Phi(z),$$

for values of $z > 0$.

• Probabilities to the **left** are obtained as $P(Z \leq z) = 1 - P(Z > z)$.

• Probabilities for $z < 0$ are obtained by symmetry. For example,

$$P(Z \leq -2) = P(Z > 2)$$

Remember to always draw the picture.

Exercises: Use the tables to convince yourself of the following:

- 68.3% of the time Y will lie within $(\mu - \sigma, \mu + \sigma)$,
- 95.4% of the time Y will lie within $(\mu - 2\sigma, \mu + 2\sigma)$,
- 99.7% of the time Y will lie within $(\mu - 3\sigma, \mu + 3\sigma)$.

e.g.

$$P(\mu - 2\sigma \leq Y \leq \mu + 2\sigma) = P(-2 \leq Z \leq 2)$$

$$= 1 - 2 \times 0.0228 = 0.9544.$$

3.9 Normal approximation to the binomial

Binomial probability calculations quickly become tedious when n is large.

Figure 3.2 demonstrates the 'smoothness' that we get in the probability histogram for large n. Suppose we use a smooth normal curve as an approximation. Which normal curve should we take?

We know that for a binomial, $\mu = np$ and $\sigma = \sqrt{npq}$, so it makes sense to use a normal curve with this mean and standard deviation.

Here, $X \sim B(20, 0.5)$, which gives $\mu = 10$, $\sigma^2 = 5$.

What is the probability that $X = 10$?

Exact calculation using binomial probability function:

$$P(X = 10) = \binom{20}{10}.5^{10}.5^{10} = .1762.$$

Using the normal approximation: since the integral at any single point is always zero, we define a small interval to integrate over. Here, the obvious choice is $(9.5, 10.5)$, and we denote the 'new' random variable by Y. Thus, the normal approximation requires the area between 9.5 and 10.5, i.e.

$$\Phi((10.5 - 10)/\sqrt{5}) - \Phi((9.5 - 10)/\sqrt{5})$$

$$= \Phi(0.5/\sqrt{5}) - \Phi(-0.5/\sqrt{5})$$

$$= 0.5871 - 0.4129 = 0.1742.$$

Continuity Correction

If X is binomial and Y is normal with the same μ and σ, then

$$\begin{aligned} P(X \leq a) &\approx P(Y < a + 0.5) \\ P(X < a) &\approx P(Y < a - 0.5) \\ P(X \geq a) &\approx P(Y > a - 0.5) \\ P(X > a) &\approx P(Y > a + 0.5) \end{aligned}$$

The application of the continuity correction is an important general method.

How good is the approximation?

Excellent when $p = 0.5$, since the binomial distribution is symmetric. It works well if $np \geq 10$ for $p = 0.5$.

It works less well when $p \neq 0.5$, because the binomial is skew, as shown by the following for n=10, p=1/6:

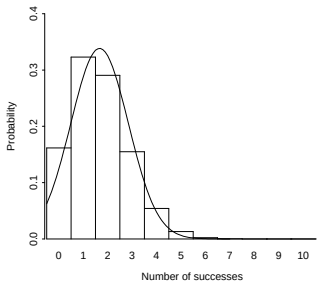

Fig. 3.6: Binomial, n=10, p=1/6

'Rule of thumb': approximation works well if both np and $n(1-p)$ are ≥ 10, i.e. have larger n as p departs from 0.5.

Example 3.3: You sample 100 travel claims by politicians. If the true proportion of claims with errors is 30%, what is the probability that you will see fewer than 20 claims with errors in your sample? Let X be the number of claims with errors.

• First identify the steps you need to take in order to use the normal approximation:

1. Is $np \geq 10$?

2. $\mu = np = 30, \quad \sigma^2 = npq = 21$ for Y.

3. Convert Y to Z, including continuity correction.

• Now do the calculations:

We want $P(X < 20) = P(X \leq 19) =$?

3.10 Poisson distribution

This is probably the most widely used discrete distribution, and arises from random processes. It is named after the French mathematician Simeon Denis Poisson (1781–1840), although it was actually introduced in 1718 by De Moivre.

One of the earliest uses of the Poisson distribution was to model the number of alpha particles emitted from a radioactive source during a given period of time, and is used as a model by insurance companies for freak accidents. It is an important distribution in its own right, but can also be derived as a limiting form of the binomial distribution when n is very large, p is very small and np is still small (roughly < 7).

Motivating example: Consider the BSE/vCJD epidemic and suppose we are investigating whether the disease has reached South Australia.

Let Y be the number of cases of vCJD diagnosed in a year, $Y = 0, 1, \ldots, n$. What is the probability of no cases in 2002, i.e. $P(Y = 0)$?

Consider each person as a Bernoulli trial. Then Y is binomial and

$$P(Y = 0) = \binom{n}{0} p^0 (1-p)^n = (1-p)^n.$$

Let the expected number of cases be $\lambda = np$. Then $p = \lambda/n$, and

$$P(Y = 0) = \left(1 - \frac{\lambda}{n}\right)^n,$$

which converges to $e^{-\lambda}$ as $n \to \infty$. To show this, expand $n\log(1 - \lambda/n)$ as a power series, or use l'Hospital's Rule.

What is the probability of one case?

$$P(Y=1) = \binom{n}{1} p(1-p)^{n-1}$$

$$= n\frac{\lambda}{n}\left(1-\frac{\lambda}{n}\right)^{-1}\left(1-\frac{\lambda}{n}\right)^{n}$$

where the third term in the product tends to 1 as $n \to \infty$, and the fourth term tends to $e^{-\lambda}$ as above. Then

$$P(Y=1) = \lambda e^{-\lambda}.$$

We can repeat this argument for $Y = 2, 3 \ldots$.

In general,

$$P(Y=y) = \binom{n}{y} p^y (1-p)^{n-y}$$

$$= \frac{n!}{y!(n-y)!}\left(\frac{\lambda}{n}\right)^{y}\left(1-\frac{\lambda}{n}\right)^{n-y}$$

$$= \frac{\lambda^y}{y!}\frac{n!}{(n-y)!}\frac{1}{n^y}\left(1-\frac{\lambda}{n}\right)^{n}\left(1-\frac{\lambda}{n}\right)^{-y}$$

As $n \to \infty$, $\lambda/n \to 0$,

$$\frac{n!}{(n-y)!n^y} \to 1, \quad \left(1-\frac{\lambda}{n}\right)^{n} \to e^{-\lambda}$$

and

$$\left(1-\frac{\lambda}{n}\right)^{-y} \to 1.$$

Thus we have the Poisson probability function

$$P(Y=y) = \frac{\lambda^y}{y!}e^{-\lambda} \quad y = 0, 1, \ldots$$

Exercise: Show that $\sum_{y=0}^{\infty} P(Y=y) = 1$.

When n is large and p is small, then the binomial distribution Bin(n,p) is well approximated by the Poisson distribution with parameter $\lambda = np$. Here are two examples of Poisson distributions (note that as λ increases the distribution becomes more nearly normal):

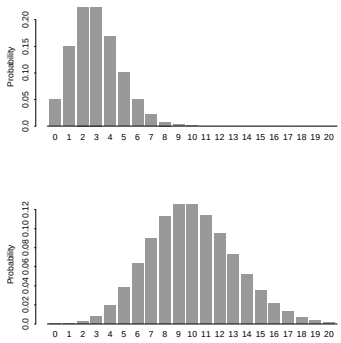

Fig. 3.7: Poisson distributions, $\lambda = 3, 10$.

Mean and variance

Theorem 3.6: If Y has a Poisson distribution with rate λ, then

$$E(Y) = \lambda \quad \text{and} \quad \text{var}(Y) = \lambda.$$

Proof: We know by construction that the expected number of counts per unit time is λ. It is also straightforward to show this directly.

Example 3.4: Death by horsekicks.
von Bortkiewitz collected data on fatalities that resulted from being kicked by a horse for 10 cavalry corps in the Prussian army over a period of 20 years, providing 200 corps-years of data.

The first row in the following table gives the number of deaths per year, Y, ranging from 0 to 4. The second row records how many times that number of deaths was observed. For example, in 65 of the 200 corps-years, there was 1 death. In the third row, the observed numbers are converted to relative frequencies by dividing by 200, and the fourth row gives the Poisson probabilities with 'expected death rate' parameter $\lambda = 0.61$.

No.deaths/year	0	1	2	3	4
No.occurrences	109	65	22	3	1
Empirical dist'n	.545	.325	.110	.015	.005
Fitted Poisson*	.543	.331	.101	.021	.003

* There were 122 deaths in the 200 observations, i.e. a mean of 0.61 per corps per year. The 'fitted' Poisson then has this mean.

Calculating the fitted Poisson probabilities of y deaths in a corps-year:

$$P(Y = 0) = \frac{.61^0}{0!} e^{-.61} = 0.5434$$

$$P(Y = 1) = \frac{.61^1}{1!} e^{-.61} = 0.3314$$

and so on.

Find the probabilities of 2,3,4 deaths as an exercise.

Note that we can find Poisson probabilities recursively:

$$P(Y = y) = \frac{\lambda}{y} P(Y = y - 1).$$

[*Similar relationships hold for other discrete distributions.*]

Exercise: Suppose your lottery ticket has a probability of 0.01 of winning a prize each week. What is the probability that you will win 0,1,2,... prizes during the year if you enter every week?

3.11 Moment generating functions

The kth *moment* of Y about the origin is

$$\mu'_k = E(Y^k) \quad k = 1, 2, \ldots.$$

The kth *moment* of Y about the mean is

$$\mu_k = E\{(Y - \mu)^k\} \quad k = 1, 2, \ldots.$$

The moments about the mean are known as *central moments*.

For example, the *population mean* is μ'_1, and the *variance* is μ_2, the second central moment.

There may be significant difficulties evaluating these, which usually involve summation and integration. We seek an easier way using differentiation.

The **moment generating function** $m(t)$ of Y is defined as $E(e^{tY})$.

The mgf for Y exists if there is some $b > 0$ such that $m(t) < \infty$ for $|t| < b$. That is, $m(t)$ is finite for t in an open interval containing 0.

The name comes from the property that the mgf generates the moments of a distribution.

Theorem 3.7: If the mgf $m(t)$ exists, then for any positive integer k

$$\frac{d^k m(t)}{dt^k}|_{t=0} = m^{(k)}(0) = \mu'_k.$$

Proof: This is an important result and we will prove it in the lectures. It tells us that the kth derivative of the mgf with respect to t, evaluated at $t = 0$, is the kth moment of the distribution.

We will show later that the mgf proves very useful in deriving distributions; e.g. if we know $m(t)$, we can tell what the distribution of Y is.

The mgf

• finds the moments of a distribution by differentiation rather than by summation and integration;

• if the mgf exists, then so do all the moments; and

• if the mgf exists, it is unique for that distribution.

Example 3.5: Find the mgf for the Poisson distribution .

Solution:

$$\begin{aligned} m(t) &= E(e^{tY}) = \sum_y e^{ty} P(Y=y) \\ &= \sum_{y=0}^{\infty} e^{ty} \frac{e^{-\lambda}\lambda^y}{y!} = e^{-\lambda} \sum_{y=0}^{\infty} \frac{(\lambda e^t)^y}{y!} \\ &= e^{-\lambda} e^{\lambda e^t} = e^{\lambda e^t - \lambda} \\ &= e^{\lambda(e^t-1)}. \end{aligned}$$

Now use this result to find the first and second moments of the Poisson distribution.

Example: if $m(t) = \exp\{2.1(e^t - 1)\}$, what is the distribution of Y?

3.12 Bounding probabilities

Tail Sum Formula for the expected value of a random variable:

Consider the simple situation where the possible values of a discrete random variable are $0, 1, \ldots, n$. Then

$$\begin{aligned} E(Y) &= \sum_{i=0}^{n} iP(Y=i) \\ &= P(Y \geq 1) + P(Y \geq 2) + \ldots \\ &= \sum_{j=1}^{n} P(Y \geq j). \end{aligned}$$

Can you see why?

This relationship tells us that the expected value can be written as a sum of tail probabilities. Obviously some bound on these is needed or they will get too big. *Markov's inequality* makes these ideas explicit, and we now motivate the inequality with an example.

Example: Suppose $Y = 0, 1, 2, \ldots$ is a discrete random variable with $E(Y) = 3$.

What is the largest that $P(Y \geq 100)$ *could possibly be*?

Think of balancing the distribution at 3; how could we get as much probability as possible in $[100, \infty]$? Intuitively, we could put some probability at 100 and the rest at 0. The distribution will be balanced at 3 if the probability at 100 is 3/100, since

$$E(Y) = 0 \times P(Y = 0) + 100 \times \frac{3}{100} = 3.$$

This suggests that $P(Y \geq 100)$ can be as large as 3/100 but that it cannot be larger.

We prove this as follows. We know

$$E(Y) = \sum_i iP(Y = i) = 3.$$

The terms with $i \geq 100$ contribute $\sum_{i \geq 100} iP(Y = i)$ to the sum, so that

$$\begin{aligned} 3 &\geq \sum_{i \geq 100} iP(Y = i) \\ &\geq \sum_{i \geq 100} 100P(Y = i) \\ &= 100P(Y \geq 100). \end{aligned}$$

Therefore,

$$P(Y \geq 100) \leq \frac{3}{100}.$$

Markov's Inequality

Recall that $E(Y)$ is the balance point of a distribution. Markov's Inequality puts a bound on how large the tail probability can be; explicitly, it gives the relationship between the tail probabilities and the expected value of a distribution. If $Y \geq 0$, then we cannot go too far out to the right without tipping the 'seesaw'.

We can ask: how much probability can be out beyond a point k? Generalising the argument given above,

$$\begin{aligned} E(Y) &\geq \sum_{i \geq k} iP(Y = i) \\ &\geq \sum_{i \geq k} kP(Y = i) \\ &= kP(Y \geq k). \end{aligned}$$

Hence we obtain **Markov's Inequality**:

$$P(Y \geq k) \leq E(Y)/k.$$

Example 3.6: Suppose the average family income in a region is \$10,000. Find an upper bound for the percentage of families with incomes as large as or over \$50,000.

We are given $E(Y) = 10,000$ and $k = 50,000$. From Markov's Inequality,

$$P(Y \geq 50000) \leq \frac{10000}{50000} = 0.2.$$

That is, at most 20% of families have incomes as large as or over \$50,000, *whatever the shape of the distribution.*

What does it imply if the bound is achieved?

Consider the inequalities that have to be satisfied in the derivation. It implies that there is no probability between 0 and k, and that all the probability at or beyond k is concentrated at k. That is,

$$E(Y) = kP(Y = k) + 0P(Y = 0).$$

For the example, this implies that 80% of families have \$0 income, and 20% have \$50,000 income.

3.13 Tchebyshev's Inequality

Markov's Inequality is quite 'weak', but we now use it to obtain a much better bound on tail probabilities, known as Tchebychev's Inequality. It makes precise the idea that a random variable is unlikely to be more than a few standard deviations away from its mean.

Consider $P(|Y-\mu| \geq k\sigma) = P(\frac{|Y-\mu|}{\sigma} \geq k)$.

If we have a *normal r.v.* Y, these tail probabilities are $P(|Z| \geq k)$, where

k	1	2	3
	.3174	.0456	.0027

(Check these using Table 4.) But what can we say in general about the size of the tail probabilities, *whatever* the shape of the distribution?

Let Y be any discrete random variable, and let

$$W = \frac{(Y-\mu_Y)^2}{\sigma_Y^2}.$$

Then

$$E(W) = \frac{1}{\sigma_Y^2} E[(Y-\mu_Y)^2] = 1.$$

Using Markov's Inequality we can write (dropping the subscript Y)

$$\begin{aligned} P(|Y-\mu| \geq k\sigma) &= P\left(\frac{|Y-\mu|}{\sigma} \geq k\right) \\ &= P(W \geq k^2) \\ &\leq E(W)/k^2 = 1/k^2. \end{aligned}$$

Thus we have **Tchebyshev's Inequality**:

$$P(|Y-\mu| \geq k\sigma) \leq 1/k^2.$$

It states that the probability a random variable differs from its mean by more than k standard deviations is at most $1/k^2$. The significance of the result is that it is true no matter what the shape of the distribution.

Here are the bounds for $k = 1, 2, 3$:

k	1	2	3
$1/k^2$	1.000	.2500	.1111

As this table shows, the bound will be very crude for a distribution that is approximately normal.

Under what conditions does equality occur?

$$P(|Y - \mu| \geq k\sigma) = \frac{1}{k^2} \Rightarrow k^2 P(|Y - \mu| \geq k\sigma) = 1$$

i.e. $k^2 P(W \geq k^2) = 1$. Thus equality is achieved when W has a 2-point distribution with values 0 and k^2, with probabilities $(1 - 1/k^2)$ and $1/k^2$ respectively.

Optional exercise: What values does Y then take, and what are their probabilities?

Note that we can equivalently present Tchebychev's Inequality as a *lower bound*:

$$P(|Y - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}.$$

This is often a more convenient form of the inequality.

4 CONTINUOUS DISTRIBUTIONS

We now assume we have variables which take 'continuous' values. For example:

• survival time of patients following treatment for cancer,

• yield from an agricultural experiment, such as weight or height or protein content,

• time to failure of a piece of equipment,

• consumer price index.

Note: It is necessary for you to revise integrals, integration by parts, etc, from Maths I or IM.

4.1 Cumulative distribution function $F(y)$

The cdf can be used to describe probabiliy distributions for discrete *and* continuous random variables.

If Y takes values on $(-\infty, \infty)$ then the cumulative distribution function $F(y)$ is

$$F(y) = P(Y \leq y).$$

Consider the familiar *discrete* binomial distribution:

Example 4.1. $Y \sim$ Binomial(n, p):

$$F(y) = P(Y \leq y) = \sum_{i=0}^{y} P(Y = i).$$

Example 4.2. $Z \sim N(0, 1)$:

$$F(z) = P(Z \leq z) = \int_{-\infty}^{z} \phi(u)du = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(u^2/2)du$$

The cdf: is usually denoted by upper case letters, e.g., F, Φ.

Properties of the cdf:

• Definition applies to all discrete and continuous variables,

• $F(y)$ is a non-decreasing function of y,

• $\lim_{y\to-\infty} F(y) = 0$,

• $\lim_{y\to\infty} F(y) = 1$,

• $F(y)$ is *right continuous*.

By right continuous, we mean that if you take the limit as $y \to y_0+$ you get $F(y_0)$, but if you take limit as $y \to y_0-$ (i.e. from below) you may not get $F(y_0)$.

4.2 Probability density functions

If Y has a cdf $F(y)$ which is continuous and which is differentiable except at a countable number of points on $-\infty < y < \infty$, then Y is said to be a *continuous random variable*. (We can draw the cdf without lifting the pen off the paper.)

If Y is continuous with cdf $F(y)$, then $f(y)$ defined by

$$f(y) = \frac{dF(y)}{dy} = F'(y)$$

if the derivative exists, and is zero elsewhere, is called the *probability density function* (pdf), or density function of Y.

The pdf is important for describing continuous random variables. Pdf's are usually denoted by lower case letters, such as f.

Properties of $f(y)$:

• $f(y) \geq 0$, (note that the pdf is not a probability function and it can take values greater than 1)

• $f(y)$ is a piece-wise continuous function,

• $\int_{-\infty}^{\infty} f(y)dy = 1$,

• $\int_{-\infty}^{y} f(u)du = F(y) = P(Y \leq y)$, where we note the 'dummy' variable u in the integration,

• there is probability 0 associated with any individual point; only intervals have a probability content,

• $P(a \leq Y \leq b) = \int_a^b f(y)dy = F(b) - F(a)$.

Given that $f()$ is the derivative of $F()$, this is essentially the Fundamental Theorem of Calculus.

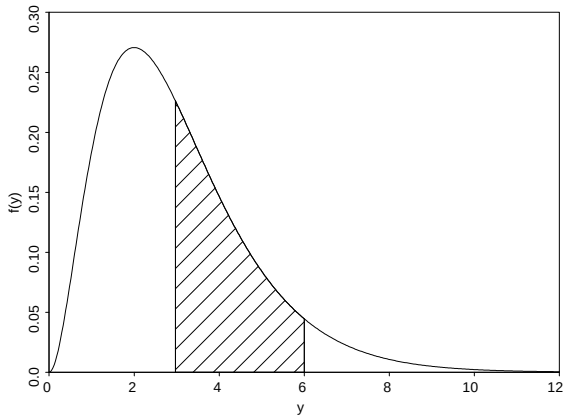

Fig. 4.1: A generic density function $f(y)$

The shaded area is

$$P(3 \leq Y \leq 6) = \int_3^6 f(y)dy = F(6) - F(3).$$

Example 4.3: the standard uniform distribution. Consider both $f(y)$ and $F(y)$ for the *uniform distribution* on $(0,1)$, which has important applications in generating random numbers and simulation; we write $Y \sim U(0,1)$.

$f(y) = 1$ for $0 < y < 1$, and 0 elsewhere. Sketch the density function.

Sketch the cdf:

$$F(y) = \begin{cases} 0 & \text{if } y < 0 \\ y & \text{if } 0 \leq y \leq 1 \\ 1 & \text{if } y > 1 \end{cases}$$

Example 4.4. Suppose $f(y) = cy$ for $0 < y < 2$ and 0 elsewhere, where c is a constant.
What is the value of c?
And what is $P(1 < Y < 2)$?

Quantiles: Suppose the cdf F is strictly increasing on an interval I, 0 to the left of I and 1 to the right of I. Then the inverse function F^{-1} is well-defined.

The pth quantile of the distribution F is defined to be that value y_p of the random variable such that $F(y_p) = p$. Thus $y_p = F^{-1}(p)$.

Special cases:
$p = \frac{1}{2}$

$p = \frac{1}{4}$

$p = \frac{3}{4}$

4.3 Expected values

These are now defined in terms of integrals as

$$E(Y) = \int_{-\infty}^{\infty} y f(y) dy$$

where the expectation is defined provided the integral converges absolutely, i.e.

$$E(|Y|) = \int_{-\infty}^{\infty} |y| f(y) dy < \infty$$

Theorem 4.1: $E\{g(Y)\} = \int_{-\infty}^{\infty} g(y) f(y) dy$.

Proof omitted (it is similar to the discrete case).

Again, this theorem is useful because we don't have to find the density function of $g(Y)$ in order to find its mean and other moments.

Note: The computing formula for the population variance is as before

$$\text{Var}(Y) = E(Y^2) - \{E(Y)\}^2.$$

Theorem 4.2: For any random variable Y, functions $g_i(Y)$ and constants c_i,
- $E(c) = c$,
- $E\{cg(Y)\} = cE\{g(Y)\}$,
- $E\{\sum_i c_i g_i(Y)\} = \sum c_i E\{g_i(Y)\}$.

Proof: This is obtained by straightforward integration, and is left as an exercise. Note that the interchange of summation and integration is possible since the integral, if it exists, is absolutely convergent.

Corollary: $\text{Var}(aY + b) = a^2\text{Var}(Y)$.

Example 4.3 (cont.): U is uniformly distributed on (0,1). Suppose $Y = U^2$. We can find the moments of Y without having to find its distribution.

For example, $E(Y) = E(U^2) = 1/3$, and

$$\begin{aligned}
\text{Var}(Y) &= E(Y^2) - \{E(Y)\}^2 \\
&= E(U^4) - 1/9 \\
&= \int_0^1 u^4 du - 1/9 \\
&= u^5/5|_0^1 = 1/5 - 1/9 \\
&= 4/45 = 0.0889.
\end{aligned}$$

Example 4.4 (cont.): Find the mean and variance of Y when $f(y) = cy, \quad 0 < y < 2$.

Recall $c = 1/2$ and $f(y) = y/2$.

4.4 Uniform distribution on (a, b)

Y has the uniform distribution on the interval (a, b) if

$$f(y) = \begin{cases} 1/(b-a) & \text{if } a < y < b, \\ 0 & \text{otherwise} \end{cases}$$

Note that:

- this fulfils all the criteria for a pdf,
- the cdf $F(y)$ is given by:

$$F(y) = \begin{cases} 0 & y \le a, \\ (y-a)/(b-a) & \text{if } a < y < b, \\ 1 & y \ge b \end{cases}$$

- the probability of being in an interval (x, y) where $a \le x < y \le b$ is

$$P(x < Y < y) = (y-x)/(b-a),$$

i.e. the proportion of the full interval occupied by the interval (x, y).

Properties:

We can rescale Y to the interval $(0, 1)$ by

$$U = (Y-a)/(b-a),$$

where now $0 < U < 1$, and

$$Y = a + (b-a)U.$$

If we generate a random value for U, we can then convert it to a random value for Y.

$$E(U) = \int_0^1 u du = u^2/2|_0^1 = 0.5,$$

so that

$$\begin{aligned} E(Y) &= a + (b-a)E(U) \\ &= a + (b-a)/2 = (a+b)/2. \end{aligned}$$

Exercise: Show $\text{Var}(U) = 1/12$, and that $\text{Var}(Y) = (b-a)^2/12$.

Relationship between the Poisson and uniform distributions.

Suppose that the number of events that occur in a time interval has a Poisson distribution. If it is known that exactly one such event has occurred in the interval $(0, t)$ then the actual time of occurrence is uniformly distributed over this interval.

Example 4.5: The number of defective circuit boards coming off a soldering machine follows a Poisson distribution. During a specific 8-hour day, one defective circuit board was found.

(a) Find the probability that it was produced during the first hour of operation that day.

(b) Find the probability that it was produced during the last hour of operation that day.

4.5 Normal distributions

The random variable Y is normal N(μ, σ^2) if it has the density function

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}.$$

For the particular case N(0,1), we use Z and denote the density function by $\phi(z)$.

There is no simple formula for the cdf, so we have

$$F(y) = \int_{-\infty}^{y} f(w)dw, \quad \Phi(z) = \int_{-\infty}^{z} \phi(u)du,$$

where $\phi(u) = \exp(-u^2/2)/\sqrt{2\pi}$ is the standard normal density function.

Note: We will show later that $F(\infty) = 1$.

Theorem 4.3: The random variable Y as defined above has $E(Y) = \mu$ and $\text{Var}(Y) = \sigma^2$.

Proof: There are two steps here. First we show the result for Z, i.e. when $\mu = 0, \sigma^2 = 1$. Then we extend it to general Y by linear transformation, i.e. $Y = \mu + \sigma Z$.
(i) Moments of $Z \sim N(0, 1)$:

$$\begin{aligned} E(Z) &= \int_{-\infty}^{\infty} z\phi(z)dz \\ &= \int_{-\infty}^{\infty} z\frac{1}{\sqrt{2\pi}}\exp(-z^2/2)dz. \end{aligned}$$

$z\phi(z)$ is an *odd* function of z:
i.e. let $h(z) = z\phi(z)$, then $h(z) = -h(-z)$.

See Figure 4.2.

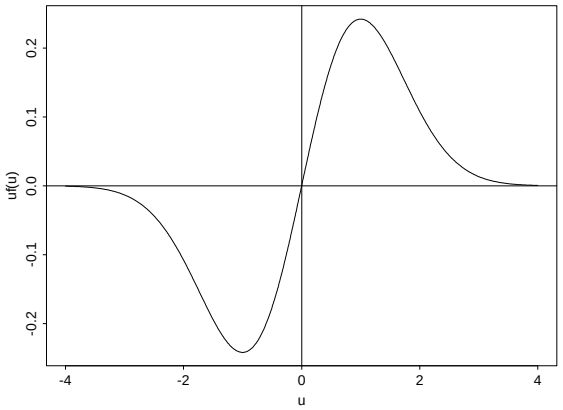

Fig. 4.2: Graph of $z\phi(z)$ versus z.

By symmetry, $E(Z) = 0$.

As an exercise, we will check that the integral is *absolutely* convergent. (Note that absolute convergence holds for almost all the expectations we will consider, and we will usually assume the condition holds unless stated otherwise.)

For $|Z|$, the integrand becomes an *even* function, so the integral is then *double* the area. Analytically:

$$\begin{aligned} E(|Z|) &= \int_{-\infty}^{\infty} |z| \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) dz \\ &= 2\int_{0}^{\infty} z \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) dz. \end{aligned}$$

Recall (Chain Rule, Maths I or IM) that

$$\int f\{g(u)\} g'(u) du = \int f(x) dx.$$

If we let $x = g(z) = z^2/2$, then $g'(z) = z$, and $zdz = dx$, so that $E(|Z|)$ becomes

$$2\int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-x) dx = \frac{2}{\sqrt{2\pi}} (-e^{-x})|_0^{\infty} = \sqrt{\frac{2}{\pi}},$$

which is finite.

Check this as an exercise.

Now, $\text{Var}(Z) = E(Z^2) - E(Z)^2$.
So, how do we get $E(Z^2)$?

Use integration by parts:

(ii) For general Y, write $E(Y-\mu)$ and transform under the integral to $z = (y-\mu)/\sigma$. That is, write

$$E(Y-\mu) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (y-\mu) e^{-(y-\mu)^2/(2\sigma^2)} dy.$$

Let $z = (y-\mu)/\sigma$, so that $dz = dy/\sigma$. Then

$$\begin{aligned} E(Y-\mu) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sigma z e^{-z^2/2} dz \\ &= \sigma E(Z) = 0. \end{aligned}$$

Therefore, $E(Y) = \mu$.

Can you see an easier way to show $E(Y) = \mu$ here?

[Note that in general it is easier to find the moments of normal Y using the mgf.]

Normal probabilities

For a general $Y \sim N(\mu, \sigma^2)$, we can find

$$\begin{aligned} P(Y > a) &= P\{(Y-\mu)/\sigma > (a-\mu)/\sigma\} \\ &= P\{Z > (a-\mu)/\sigma\}, \end{aligned}$$

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

Check: Do the transformation using the integral!

Hence, we only need **one** table for Z.

4.6 Cauchy distribution [Not in WMS]

This is given by the density

$$f(y) = \frac{1}{\pi\{1 + (y-\theta)^2\}} \quad -\infty < y, \theta < \infty,$$

where θ is a location parameter.

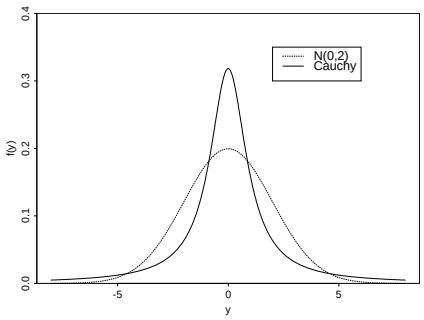

Fig. 4.3: Cauchy distribution with $\theta = 0$ and a normal distribution. Note the similarity between the Cauchy and normal distributions.

Note the long tails for Cauchy: the density decreases so slowly that very large values of Y can exist with substantial probability.

This leads to problems, and the mgf does not exist.

Nevertheless, the Cauchy distribution has a special role in the theory of statistics - it represents an extreme case against which conjectures can be tested. It also turns up in statistical practice when you least expect it. For example, the ratio of two standard normal random variables has a Cauchy distribution.

Here are 20 values from this distribution when $\theta = 0$:

-0.100	-0.667	0.647	-0.674	1.434
-0.439	12.060	0.343	0.842	-0.592
1.815	-2.267	1.204	2.385	0.345
2.044	-0.228	-6.197	8.298	-2.794

If we take $E(Y - \theta)$ here, directly or by symmetry, the areas cancel. However, if we try to establish *absolute* convergence, we find that

$$\begin{aligned} E(|Y-\theta|) &= 2\int_0^\infty u \frac{1}{\pi(1+u^2)} du, \quad u = y - \theta \\ &= \frac{1}{\pi}[\log(1+x)]_0^\infty, \quad x = u^2. \end{aligned}$$

Hence we have to say that $E(Y - \theta)$ does not exist in this case.

Question: How might we estimate θ here? Clearly, $E(\bar{Y})$ for a sample of size n does not exist either!

4.7 Exponential distribution

In many applications, interest centres on the time taken for an event to occur. Examples are:

- survival times for cancer patients,
- time to decay for radioactive atoms,
- time to failure of a piece of equipment.

These are characterised by continuous, non-negative random variables. An example with special properties is the *exponential distribution*, which is used extensively in reliability testing. Generalizations are used in actuarial studies to model human lifetimes as a basis for estimating life-insurance premiums.

The exponential distribution is very useful as a model for right-skewed data.

A non-negative random variable T is *exponential* with mean β if the density function

$$f(t) = (1/\beta)e^{-t/\beta}, \quad t \geq 0, \beta > 0.$$

Note:

- $\int_0^\infty f(t)dt = 1$.

- It is the continuous analogue of the geometric distribution: the exponential distribution models the time to an event.

- Often used as a 'waiting-time' distribution, with mean 'time to event' β.

- cdf is $F(t) = P(T \leq t) = 1 - e^{-t/\beta}$.

- β has units of time.

- $\lambda = 1/\beta$ is called the **rate** and has units of 'per unit time'.

- $S(t) = 1 - F(t) = P(T > t)$ is known as the 'survivor function'. Here, $S(t) = e^{-t/\beta}$. It has an important role in the fields of Survival Analysis and Reliability.

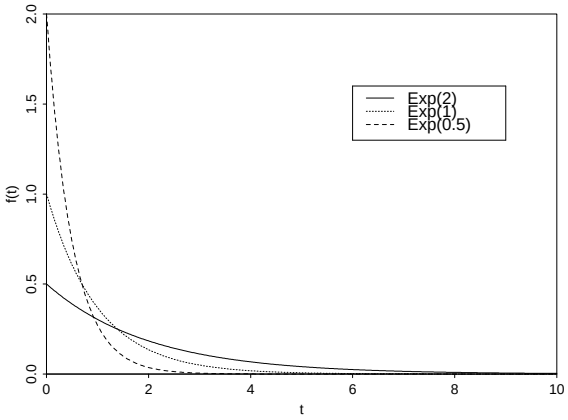

Fig. 4.4: Exponential distributions, Exp(β).

Regarded as 'length of life', lower β represents earlier failure, and fewer survivors.

Suppose we want $P(a < T < b)$. This is equal to

$$\begin{aligned}\int_a^b f(t)dt &= F(b) - F(a) \\ &= (1 - e^{-b/\beta}) - (1 - e^{-a/\beta}) \\ &= e^{-a/\beta} - e^{-b/\beta} \\ &= S(a) - S(b).\end{aligned}$$

Moments:

Using integration by parts,

$$\begin{aligned}E(T) &= \int_0^\infty (1/\beta) t e^{-t/\beta} dt \\ &= \frac{1}{\beta}\int_0^\infty t e^{-t/\beta} dt \\ &= \frac{1}{\beta}\{t(-\beta e^{-t/\beta})|_0^\infty + \int_0^\infty \beta e^{-t/\beta} dt\} \\ &= 0 + \int_0^\infty e^{-t/\beta} dt = -\beta e^{-t/\beta}|_0^\infty \\ &= -\beta(0 - 1) \\ &= \beta.\end{aligned}$$

Exercise: Use the same method to find $E(T^2)$ and hence show that $\text{Var}(T) = \beta^2$.

Note that the mean equals the standard deviation for the exponential distribution.

Memoryless property of exponential distribution

The exponential distribution is the only continuous distribution with this property.

Given that someone has survived to a certain point in time, t_0, what is the probability that they survive *a further* s units?

This is given by the conditional probability

$$\begin{aligned} P\{T > (t_0 + s) | T > t_0\} &= \frac{P(T > (t_0 + s) \cap T > t_0)}{P(T > t_0)} \\ &= \frac{P(T > t_0 + s)}{P(T > t_0)} \\ &= \frac{S(t_0 + s)}{S(t_0)} \\ &= e^{-(t_0+s)/\beta + t_0/\beta} \\ &= e^{-s/\beta} \\ &= P(T > s) = S(s) \end{aligned}$$

which does not depend on t_0.

Thus the probability of surviving a further s units **given that you are alive at** t_0 is the same as having survived s units in the first place. (That is, the probability of surviving a further s units is the same regardless of how long you have already have survived.) This is called the memoryless property of the exponential distribution.

The hazard function $h(t)$:

This is the 'instantaneous failure rate',

$$h(t) = \lim_{\delta \to 0} P(t < T \leq t + \delta | T > t)/\delta$$

i.e. the risk of failing in a short interval $(t, t+\delta]$, given that you are still alive at time t.

Using the same argument as for the memoryless property on the previous slide, we can show that the hazard function is $(1 - e^{-\delta/\beta})/\delta$, which tends to $1/\beta$ as $\delta \to 0$.

It follows that the *hazard function* for an exponential distribution is $h(t) = \lambda = 1/\beta$; it is usually called the *rate* of the exponential distribution and is a **constant**.

Note it is true in general that

$$h(t) = \frac{f(t)}{S(t)}.$$

Thus for the exponential distribution, we have

$$f(t) = h(t)S(t) = \frac{1}{\beta} e^{-t/\beta}.$$

The hazard function has many different names, including the:

- force of mortality in demography
- age-specific failure rate in epidemiology
- conditional failure rate in reliability
- intensity function in stochastic processes
- inverse of Mill's ratio in economics.

Half-life, or median time to failure $t_{\frac{1}{2}}$

Atoms of radioactive isotopes like Carbon 14 or Uranium 235 remain intact up to a random instant of time when they suddenly decay, meaning that they split or turn into some other kind of atom, and emit a pulse of radiation or particles of some kind.

Let T be the random lifetime (time until decay) of such an atom. It is reasonable to assume that the distribution of T must have the memoryless property, and in fact, the *exponential decay* over time of the mass of a radioactive substance has been verified experimentally.

Note that as the number of particles reduces, so will the number decaying.

The 'rate of decay' is often summarised by the *half-life*, the time taken for half the material to decay. This is given by $t_{1/2}$ satisfying

$$\begin{aligned} P(T \leq t_{1/2}) &= F(t_{1/2}) = 0.5 \\ &= \int_0^{t_{1/2}} (1/\beta) e^{-t/\beta} dt \\ &= 1 - e^{-t_{1/2}/\beta}, \end{aligned}$$

so that

$$t_{1/2} = \log_e(2)\beta.$$

Thus the median is smaller than the mean by a factor $\log_e 2 = 0.693$. Why does this happen?

Example 4.6. Strontium 90 is a particularly dangerous component of fallout from nuclear explosions. The substance is toxic, easily absorbed into bones when eaten and has a long half-life of about 28 years.

What is the proportion of atoms that decay in a year?

At the end of the first year, the proportion remaining is $P(Y > 1) = \exp(-1/\beta)$. So first of all find β, then the proportion decayed is $1 - \exp(-1/\beta)$.

Find (i) the mean life of such an atom; (ii) the proportion still remaining after 50 years, or 100 years, and (iii) the number of years after a nuclear explosion before 99% of the Strontium 90 produced by the explosion has decayed.

Example 4.7. Bacteria survival: experimental work has shown that the memoryless property holds here too.

Suppose that under radiation, the half-life of a bacterium is 10 seconds. What is the probability that it will survive 20 seconds?

What is the probability that it will die between 20 and 25 secs?

Example 4.8. Australian AIDS survival in the late 1980's (see handout). This plot is taken from a study of the effects of the widespread introduction of AZT into clinical practice in mid-1987.

• The dotted lines are nonparametric estimates of survival and make no distributional assumptions about the survival times.
• The solid lines assume the survival times are exponentially distributed.

Good agreement implies that the exponential is a good fit.

Conclusions:
• Each curve suggests the 'memoryless' property applies here, i.e., constant hazard of death.
• The differences imply that pre-1987 cases had shorter survival times (e.g. 35% survive one year) than post-1987 case (where 70% survive one year).

Note: How could we plot these to illustrate the exponential decay better?

4.8 Poisson processes

A Poisson process is a model for random events occurring in time, space, etc, at a rate λ. For example, the number of industrial accidents at a certain facility each day.

1. Consider a time interval of length L, and let N_L be the number of events in L. Then N_L has the Poisson distribution with mean λL.

2. Since events are occurring at random, the times between them must be random too. So suppose an event occurs at time t_0, and let T be the random variable representing time to the next event. Then

$$\begin{aligned} P(T > t) &= P(\text{no events in } (t_0, t_0 + t)) \\ &= \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t} \\ &= 1 - F(t) = S(t) \end{aligned}$$

for T an exponential random variable with $\lambda = 1/\beta$.

Repeating this argument for the time to the next event, etc, we can show that the times between events in a Poisson process are independent and identically exponentially distributed.

Thus a **Poisson process** may be described in two alternative ways:

• The number of events N_L in an interval of length L follows a Poisson distribution with rate λL, and the number of events in non-overlapping intervals are independent, or

• The waiting time T to the first success follows an exponential distribution of rate λ, and the waiting times between each success and the next are independent, with the same exponential distribution.

4.9 Gamma distribution

This is probably the most widely used two-parameter distribution for a non-negative continuous random variable. It is extremely important as a distribution for right-skewed data.

Let T be the sum of r independent exponential random variables $W_1, W_2, \ldots, W_r$, each with mean β. Then $T = W_1 + \ldots + W_r$, the time to the rth event/success, has the **gamma density function** with parameters r, β:

$$f(t) = \frac{1}{\beta^r \Gamma(r)} t^{r-1} \exp(-t/\beta),$$

which exists for all $r, \beta > 0$; $\quad t \geq 0$.

r is called a *shape* parameter, β is called a *scale* parameter, and $\Gamma(r)$ is the gamma function.

What distribution do we get when $r = 1$?

(We will work through most of the following calculations in the lectures.)

• Check that $\int_0^\infty f(t)dt = 1$.

• We can also show that

$$E(T) = r\beta, \quad \text{Var}(T) = r\beta^2.$$

• Note that r need not be integer, in which case

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

and this is the gamma function.

Exercise: Show that $\Gamma(r+1) = r\Gamma(r)$; this is the recursion formula for the gamma function. Also show that for integer r, $\Gamma(r+1) = r!$.

• Note that $\int_c^d f(t)dt$ does not in general have an analytic form and we need to use software or tables of the incomplete gamma function.

Example. The gamma distribution is often used as a model for the AIDS incubation period, which is the time from infection with HIV to a diagnosis of AIDS.

An early study of people infected via blood transfusions estimated that $r = 2$ and $\beta = 7.143$ (in years).

What is the mean time to AIDS (in years) under the gamma model?

What is the probability of remaining AIDS-free 2, 10 or 15 years following infection?

A useful device for integer r, based on an interesting relationship between the gamma and Poisson distributions.

Tail probabilities for the gamma can be quite difficult to obtain, involving a complicated integral. However, we can often get to the solution using properties of the Poisson distribution and Poisson process.

If a Poisson process has rate λ, the following two statements are equivalent:

• the rth event occurs before time $T = t$, where T has a Gamma$(r, 1/\lambda)$ distribution, and

• there are at least $R = r$ events in the interval $(0, t]$, where R is Poisson with mean λt.

Thus the probabilities of both events are the same, i.e. $P(T < t) = P(R \geq r)$, so that

$$\int_0^t \frac{\lambda^r}{\Gamma(r)} t^{r-1} e^{-\lambda t} dt = \sum_{i=r}^{\infty} \frac{(\lambda t)^i}{i!} \exp(-\lambda t).$$

[This can be shown using integration by parts.]

Example 4.9. Let $T_2 = W_1 + W_2$, i.e., the time to the second event in a Poisson process of rate λ.

What is the probability that $T_2 < 10$ when $\lambda = 0.5$?

Here, the number of events in the first 10 units of time is R which has a Poisson distribution with mean $10\lambda = 5$. Thus,

$$P(T_2 < 10) = P\{R \geq 2\}.$$

T_2 is gamma $(2, 1/.5)$. So

$$\begin{aligned} P(T_2 < 10) &= P(R \geq 2) \\ &= 1 - P(R < 2) \\ &= 1 - P(R = 0) - P(R = 1) \\ &= 1 - e^{-5} - 5e^{-5} \\ &= 0.9596. \end{aligned}$$

The Chi-square distribution. This is an important special case of the gamma distribution.

Suppose $Z \sim N(0,1)$ and $Y = Z^2$. What is the density of Y?

It is easiest to find it using the *cdf method*:

$$\begin{aligned} F(y) = P(Y \le y) &= P(Z^2 \le y) \\ &= P(|Z| \le \sqrt{y}) = P(-\sqrt{y} \le Z \le \sqrt{y}) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} \phi(z)dz \\ &= 2\int_0^{\sqrt{y}} \phi(z)dz \end{aligned}$$

since $\phi(z)$ is an even function. Then, since $f(y) = F'(y)$,

$$\begin{aligned} f(y) &= 2\phi(\sqrt{y})\frac{1}{2}y^{-1/2} \\ &= \frac{1}{\sqrt{2\pi y}}e^{-y/2}, \quad y > 0. \end{aligned}$$

This is the density function for the chi-square distribution with 1 degree of freedom, denoted $Y \sim \chi^2_1$.

This is also a gamma distribution with $r = 1/2$ and $\beta = 2$.

Exercise: Show that $P(Y < 3.84) = 0.95$.

4.10 Moment generating functions

As with discrete distributions, the **moment generating function** $m(t)$ of Y is defined as

$$E(e^{tY}) = \int e^{ty} f(y) dy.$$

The mgf for Y exists if there is some $b > 0$ such that $m(t) < \infty$ for $|t| < b$.

Example 4.10. The mgf for an exponential distribution with random variable Y is given by

$$\begin{aligned} E(e^{tY}) &= \int_0^\infty e^{ty} \frac{1}{\beta} \exp^{-y/\beta} dy \\ &= \frac{1}{\beta} \int_0^\infty \exp\{(t - 1/\beta)y\} dy \\ &= \frac{1}{\beta(t - 1/\beta)} \exp\{(t - 1/\beta)y\}|_0^\infty = \frac{1}{1 - \beta t} \end{aligned}$$

provided $t < 1/\beta$.

Under what conditions is this integration valid?

An important example: Show that the mgf for the standard normal distribution is

$$m(t) = e^{t^2/2}.$$

Check that $E(Z) = 0$ and $\text{Var}(Z) = 1$.

The same results for the mgf hold as for the discrete case: $m^{(k)}(t)$ evaluated at $t = 0$ gives the moments μ'_k about the origin.

Exercise: Find the mgf for a gamma distribution with parameters (r, β).

By differentiating with respect to t, find the mean and variance of the gamma distribution.

We will show later that the mgf proves very useful, for example:

• If we know $m_Y(t)$, we can often tell what the distribution of Y is (this is the uniqueness property).

• If $T = Y_1 + \ldots + Y_n$ is the sum of n independent random variables, then the mgf of T is just the *product* of the mgf's of the n random variables.

Exercise: Use the above two results to show that if Y_1 and Y_2 are independent Poissons with means λ_1 and λ_2, then $X = Y_1 + Y_2$ is also Poisson, but with mean $\lambda_1 + \lambda_2$.

Sums of Poisson random variables are also Poisson.

This applies also for continuous random variables. The proof is analogous to the discrete case.

Theorem 4.4: If Y is a random variable with finite mean μ and finite variance σ^2, then for any k,

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

Proof:

$$\begin{aligned}\sigma^2 &\geq \int_{-\infty}^{\mu-k\sigma}(y-\mu)^2 f(y)dy + \int_{\mu+k\sigma}^{\infty}(y-\mu)^2 f(y)dy \\ &\geq \int_{-\infty}^{\mu-k\sigma} k^2\sigma^2 f(y)dy + \int_{\mu+k\sigma}^{\infty} k^2\sigma^2 f(y)dy \\ &= k^2\sigma^2 P(|Y-\mu| > k\sigma).\end{aligned}$$

Hence result.

Note that if Var$(Y) = 0$, then $P(X = \mu) = 1$.

5 MULTIVARIATE PROBABILITY DISTRIBUTIONS

5.1 Bivariate distributions

We are interested in how the random variables X, Y, ... behave together.

The event $(X, Y) = (x, y)$ is the *intersection* of the events $X = x$ and $Y = y$.

Examples:

• In ecological studies, counts (modelled as random variables) of several species are often made. One species is often the prey of another, and clearly the number of predators will be related to the number of prey.

• A model for the joint distribution of age and length of fish populations can be used to estimate the age distribution from the length distribution; the age distribution is relevant to the setting of reasonable harvesting policies.

• The joint probability distribution of the x, y and z components of wind velocity can be measured experimentally in studies of atmospheric turbulence.

• The joint distribution of factors such as cholesterol, blood pressure and age is important in studies for determining an individual's risk of heart attack.

• Interest may centre on the joint distribution of quality of life and time since a diagnosis of HIV/AIDS.

Consider first the discrete case.

Suppose the outcomes in a sample space S are indexed by **two** random variables (X, Y).

Then each outcome (x, y) has an associated probability $P(X = x, Y = y)$.

Definition 5.1: If X and Y are discrete random variables, then the (joint) **probability distribution** of (X, Y) is defined by

$$p(x, y) = P(X = x, Y = y).$$

Theorem 5.1: $p(x, y) \geq 0$ and $\sum p(x, y) = 1$.

Example 5.1: Roll two dice.
(i) X on Die 1, Y on Die 2. Then

$$P(X = x, Y = y) = P(X = x)P(Y = y),$$

by independence, and hence if the dice are fair each outcome (x, y) has probability 1/36.

(ii) Let W be the sum and U be the product of the two numbers showing. Then the events $\{W = w\}$ and $\{U = u\}$ are **not** independent and the joint probabilities $P(W = w, U = u)$ are more complex.
(Can you see why?)

We will discuss this example in the lectures.

Example 5.1 (ii) Enumerate the 36 outcomes and calculate the sum W and the product U for each:

W		2	3	4	5	6	7	8	9	10	11	12
U	1	1										
	2		2									
	3			2								
	4			1	2							
	5					2						
	6				2		2					
	8					2						
	9					1						
	10						2					
	12						2	2				
	15							2				
	16							1				
	18								2			
	20								2			
	24									2		
	25									1		
	30										2	
	36											1

Note that $\sum_{x,y} p(x,y) = 1$.

The joint behaviour of two random variables X and Y is determined by the *cumulative distribution function*. As for the univariate case, the cdf is defined for both discrete and continuous random variables.

Definition 5.2: If X and Y are any random variables, then the (joint) **cumulative distribution distribution** (cdf) of (X,Y) is defined by

$$F(x,y) = P(X \leq x, Y \leq y)$$

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

The cdf gives the probability that the point (X,Y) belongs to a semi-infinite rectangle in the plane.

For two discrete random variables, $F(x, y)$ has the form

$$F(x_1, y_1) = \sum_{x=-\infty}^{x_1} \sum_{y=-\infty}^{y_1} p(x, y)$$

Example 5.1 (cont.): In the table on the previous slide, the cdf can be thought of as summing the probabilities in the top left hand rectangle from any nominated point (w, u).

Thus, for example, $F(6, 9)$ is

$$\begin{aligned} P(W \leq 6, U \leq 9) &= p(1,1) + p(3,2) + \ldots + p(6,9) \\ &= 15/36. \end{aligned}$$

The bivariate continuous case: we now integrate over regions in the plane to get probabilities, which are volumes rather than areas.

Definition 5.3: Suppose X and Y are each continuous random variables, and suppose there exists a function f such that

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) dv du.$$

Then X and Y are said to be **jointly continuous random variables**. f is the joint probability density function; it is a piecewise continuous function of two variables.

- For any region R,

$$P\{(X, Y) \in R\} = \int_R f(x, y) dx dy.$$

- Probabilities are given by **volumes**.

The volume *under* $f(x,y)$ *over* the small rectangle $dxdy$ is approximately $f(x,y)dxdy$, i.e.,

$$P(x \leq X \leq x+dx, y \leq Y \leq y+dy) \approx f(x,y)dxdy.$$

To obtain the volume for a whole region R, sum all these little volumes.

Theorem 5.2: If X and Y have joint density $f(x,y)$, then $f(x,y) \geq 0$ for all x, y, and

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1.$$

Note:

• If F is continuous in x and y, then

$$f(x,y) = \frac{\partial^2}{\partial x \partial y}F(x,y).$$

This is essentially the Fundamental Theorem of Multivariable Calculus.

Properties of the cdf:

- $F(x, -\infty) = F(-\infty, y) = F(-\infty, -\infty) = 0$,
- $F(\infty, \infty) = 1$,
- for $x_1 \geq x_0$ and $y_1 \geq y_0$,

$$P(x_0 \leq X \leq x_1, y_0 \leq Y \leq y_1)$$

$$= F(x_1, y_1) - F(x_1, y_0) - F(x_0, y_1) + F(x_0, y_0) \geq 0.$$

Example 5.2: Consider (X, Y) jointly uniformly distributed on the unit square, so that $0 < x < 1$, $0 < y < 1$.

Then $f(x, y) = \frac{1}{\text{area}} = 1$.

$P(X < 0.5) = F(0.5, 1) =$

$F(x, y) = P(X \leq x, Y \leq y) =$

Example 5.3: Consider the joint distribution of (X, Y) defined by the density function

$$f(x, y) = c, \quad 0 \leq x \leq 1; \quad 0 \leq y \leq x.$$

Find c, and hence find $P(X < 0.5) = F(0.5, x)$.

Always draw the region of integration first.

Double integrals: see handout on multivariable integrals.

• Probabilities and expectations here are double integrals.

• It is wise to always **draw** the region in the (X, Y) plane, and shade in the region of integration.

• Often the region is expressed in terms of the values of x, and then values of y which vary according to x.

• It is then necessary to integrate out y at a given value of x and then integrate out x.

• If it makes the integration easier, reverse the order of integration so that y has limits independent of x, but the limits on x depend on y. Then integrate out x first.

The multivariate case

Both the discrete and continuous cases generalise to n random variables in an obvious way.

5.2 Marginal distributions

Suppose X and Y have a known joint distribution.

Definition 5.4: If X and Y are discrete, the **marginal distributions** of X and Y are defined by

$$p_1(x) = \sum_y p(x, y), \quad p_2(y) = \sum_x p(x, y).$$

(ii) If X and Y are jointly continuous, the **marginal density functions** are

$$f_1(x) = \int_y f(x, y)dy, \quad f_2(y) = \int_x f(x, y)dx,$$

where the integrals are over the whole real line.

Example 5.4: Toss a fair coin 3 times. Then the sample space $\mathcal{S}$ is {HHH, TTT, HHT, TTH, HTH, THT, HTT, THH} and each outcome has probability 1/8.

Let X be the number of heads in the first 2 tosses. What values can X take?

Let Y be the number of heads in the second 2 tosses. What values can Y take?

What is the joint distribution of (X, Y)?

We will work through this example in the lectures, and will find the appropriate marginal distributions.

Example 5.5: Consider the joint density function

$$f(x, y) = \lambda^2 e^{-\lambda y} \quad 0 \leq x \leq y < \infty$$

for a constant $\lambda > 0$.

Sketch the region of integration.

Let $\lambda = 1$ so that $f(x, y) = e^{-y}$ for $0 \leq x \leq y < \infty$.

Find the marginal density function of X, and hence establish that $f(x, y)$ is a valid density function by showing that the total volume under the surface is 1.

Find the marginal density function of Y. (Do you recognise these density functions?)

We will work through this example in the lectures.

Exercise: Example 5.1 (continued): (ii) The joint distribution of W and U is given in the table on Slide 5.5. The marginal distribution of W, for example, is obtained by summing **down** each column to get:

W	2	3	4	5	6	7	8	9	10	11	12
$36p_w$	1	2	3	4	5	6	5	4	3	2	1

Exercise: Example 5.3 (cont.): (X, Y) defined as

$$f(x, y) = 2, \quad 0 \leq x \leq 1, \quad 0 \leq y \leq x.$$

Find the marginal density functions $f_1(x)$ and $f_2(y)$.

5.3 Conditional distributions

Recall that $P(A|B) = P(A \cap B)/P(B)$. For events $\{X = x\}$ and $\{Y = y\}$, the same applies. Hence for discrete distributions we have:

Definition 5.5: If X and Y are jointly discrete with joint probability distribution $p(x, y)$, and marginal probability distributions $p_1(x)$ and $p_2(y)$, the **conditional distribution of Y given** $X = x$ is

$$p_2(y|x) = P(Y = y|X = x) = \frac{p(x, y)}{p_1(x)},$$

defined for all values of X such that $p_1(x) > 0$.

Similarly, the conditional distribution of X given $Y = y$ is

$$p_1(x|y) = P(X = x|Y = y) = \frac{p(x, y)}{p_2(y)},$$

defined for all values of Y such that $p_2(y) > 0$.

Example 5.4 continued: toss 3 coins.

Find $P(Y = 1|X = 1)$

Find $P(Y = 0|X = 1)$

Find $P(Y = 2|X = 1)$

This is the conditional distribution of Y given $X = 1$.

Exercise: example 5.1 (cont.): Use this definition to get the conditional distribution of U given $W = 7$, say.

W		2	3	4	5	6	7	8	9	10	11	12
U	1	1										
	2		2									
	3			2								
	4			1	2							
	5					2						
	6				2		2					
	8					2						
	9					1						
	10						2					
	12						2	2				
	15							2				
	16							1				
	18								2			
	20								2			
	24									2		
	25									1		
	30										2	
	36											1
$36p_w$		1	2	3	4	5	6	5	4	3	2	1

Still on bivariate discrete distributions:

Note that we can write the joint probability

$$P(X = x, Y = y) = p(x, y) = p_1(x)p_2(y|x),$$

or equivalently, as

$$p(x, y) = p_2(y)p_1(x|y). \quad (*)$$

These formulae are useful for finding the joint distribution of X and Y when, say, the marginal distribution of Y *and* the conditional distribution of X given $Y = y$ are known, but the joint distribution is not known.

We can take this a step further by marginalising $(*)$ over Y to obtain the marginal distribution of X. Can you see how to do this?)

This is the Law of Total Probability.

Conditional density functions

Care is needed in the continuous case because the event $X = x$ has probability 0.

By analogy with the discrete case, we use the intuitive definitions:

$$f_2(y|x) = \frac{f(x, y)}{f_1(x)}, \quad f_1(x) > 0$$

and

$$f_1(x|y) = \frac{f(x, y)}{f_2(y)}, \quad f_2(y) > 0.$$

We can visualise, for example, $f_2(y|x)$ as the profile of a slice through the joint density $f(x, y)$ with x held constant, normalised to have unit area.

Example 5.5 continued:

$$f(x,y) = e^{-y} \quad 0 \le x \le y < \infty.$$

What is the conditional density of Y given $X = x$?

What is the conditional density of X given $Y = y$?

Example 5.3 continued: (X, Y) defined as

$$f(x,y) = 2, \quad 0 \le x \le 1, \quad 0 \le y \le x.$$

Find $f_2(y|x)$ and $f_1(x|y)$.

Law of Total Probability for the continuous case:

As for the discrete case, the joint density can be expressed in terms of the marginal and conditional densities. For example,

$$f(x,y) = f_2(y|x) f_1(x).$$

Then integrating both sides over x gives the marginal distribution of y as

$$f_2(y) = \int_{-\infty}^{\infty} f_2(y|x) f_1(x) dx.$$

5.4 Independence

Definitions for independence of events are:
- $P(A \cap B) = P(A)P(B)$,
- $P(B|A) = P(B)$,
- $P(A|B) = P(A)$.

For independence of **random variables**, we need this to be true for all events such that A is an event concerning X and B is an event concerning Y.

Definition 5.6: Let X and Y have cdfs $F_1(x)$ and $F_2(y)$, and joint cdf $F(x, y)$. Then X and Y are **independent** if and only if

$$F(x, y) = F_1(x)F_2(y),$$

for every pair of real numbers (x, y).

That is,

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$$

for all (x, y).

Theorem 5.3: (i) If X and Y are discrete, they are independent if and only if

$$p(x, y) = p_1(x)p_2(y),$$

for all real pairs (x, y).
(ii) If X and Y are continuous, they are independent if and only if

$$f(x, y) = f_1(x)f_2(y),$$

for all real pairs (x, y).

Proof omitted.

Note that the ranges of $f_1(x)$ and $f_2(y)$ cannot depend on y or x, respectively. So we cannot have independence unless $f(x, y)$ has ranges which are independent of x and y.

Examples: For Example 5.1(ii) and Example 5.3, the random variables are not independent. Show this for Example 5.3.

Example 5.6: Suppose X and Y are independent exponential random variables with means β and γ. Then

$$\begin{aligned} f(x,y) &= f_1(x)f_2(y) \\ &= \frac{1}{\beta\gamma}\exp(-\frac{1}{\beta}x - \frac{1}{\gamma}y). \end{aligned}$$

Where the limits of integration are fixed (e.g. at 0 or ∞), there is a simple way to show independence [WMS p.236]:

Theorem 5.4: Suppose X and Y have a joint density $f(x,y)$ positive if and only if $a \le x \le b$ and $c \le y \le d$. Then X and Y are independent if and only if

$$f(x,y) = g(x)h(y),$$

where $g()$ is a nonnegative function of x only and $h()$ is a nonnegative function of y only.

So when the conditions of the theorem are satisfied, we don't need to derive the marginal densities in order to show independence.

Example 5.2 (cont.): If $f(x,y) = 1$ for $0 \le x, y \le 1$, then X and Y are independent.

Example 5.3 (cont.): If $f(x,y) = 2$ for $0 \le y \le x \le 1$, then this theorem cannot be applied. However, we can state that X and Y are not independent. Why?

Note that for X and Y independent random variables,

$$f_1(x|y) = f_1(x)$$

and

$$f_2(y|x) = f_2(y),$$

i.e. the conditional density functions reduce to the marginal density functions.

Example 5.7: extreme values and order statistics.

Suppose a system has n components connected in parallel, so that the system fails only if all the components fail. Suppose also that the lifetimes of the components $T_1, \ldots, T_n$ are independent, identically distributed exponential random variables with mean parameter β.

Let U be the random variable representing the length of time the system operates; this is the *maximum* of the T_i.
Find the density function of $U = \max(T_1, \ldots, T_n)$.

Example 5.8: convolution.

Let X and Y be discrete random variables with joint probability $p(x, y)$.

Let $Z = X + Y$, and find $p(z)$.

Note that $Z = z$ whenever $x + y = z$, i.e., when $X = x$, $Y = z - x$. Then $p(z)$ is the sum over all x of these joint probabilities, i.e.,

$$p(z) = \sum_{x=-\infty}^{\infty} p(x, z - x).$$

If X, Y are independent, then

$$p(x, y) = p_1(x)p_2(y)$$

and

$$p(z) = \sum_{x=-\infty}^{\infty} p_1(x)p_2(z - x).$$

This sum is the *convolution* of the sequences p_1, p_2.

5.5 Expected values

Let X and Y be discrete random variables with joint probability function

$$p(x, y) = P(X = x, Y = y).$$

Let $g(X, Y)$ be a function of X and Y. Then the expected value of $g(X, Y)$ is

$$E\{g(X, Y)\} = \sum_y \sum_x g(x, y)p(x, y).$$

If X and Y are continuous random variables with joint density function $f(x, y)$ then

$$E\{g(X, Y)\} = \int_y \int_x g(x, y)f(x, y)dxdy.$$

Example 5.4 continued: Toss coin 3 times.

Let $g(X,Y) = XY$. Find $E(XY)$.

Example 5.2 continued: (X,Y) uniform on unit square.

Suppose we are interested in $Z = g(X,Y) = X^2Y$. What is its expectation?

The above results generalise to k random variables.

Let $\mathbf{Y} = (Y_1, \ldots, Y_k)'$ be a vector of random variables.

Definition 5.7: For any function $Z = g(\mathbf{Y})$, the **expected value** of Z is defined as
(i) for a discrete set of random variables

$$E\{g(\mathbf{Y})\} = \sum \ldots \sum g(\mathbf{y})p(\mathbf{y}),$$

(ii) for a continuous set of random variables

$$E\{g(\mathbf{Y})\} = \int \ldots \int g(\mathbf{y})f(\mathbf{y})dy_k \ldots dy_1.$$

If there are k random variables, but $Z = g()$ is a function of only some of them, then we can use **either** the full density function **or** the marginal density for any subset that includes those involved in $g()$.

Example 5.3 continued: If we want $E(X)$, we can use the joint density $f(x,y)$ or the marginal density $f_1(x)$.

5.6 Combining expectations

These follow as in the univariate case.

Theorem 5.5: For any random variables $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, functions $g_i(\mathbf{Y})$ and constants c_i,
- $E(c) = c$,
- $E\{cg_i(\mathbf{Y})\} = cE\{g_i(\mathbf{Y})\}$,
- $E\{\sum_i c_i g_i(\mathbf{Y})\} = \sum c_i E\{g_i(\mathbf{Y})\}$.

Example 5.3 continued: We know that $E(X) = 2/3, E(Y) = 1/3$.
So what is $E(X - Y)$?

Group testing:

Suppose that a large number, n, of blood samples are to be screened for a rare disease. If each sample is assayed individually, n tests will be required. On the other hand, if each sample is divided in half, and one of the halves is pooled with some of the other halves, the *pooled* blood can be tested. The idea is that if the pooled blood tests negative, then no further testing of the samples in the pool is required. If however the pooled blood tests positive, each reserved half-sample can then be tested individually.

Suppose the n samples are first grouped into m subgroups of k samples in each group, i.e. $n = mk$. Each subgroup is then tested: if a subgroup tests positive, each individual in the subgroup is tested. Let p be the probability of a negative test on any individual, and let X_i be the number of tests run on the ith subgroup.

If N is the total number of tests run, find the expected value of N.

Theorem 5.6: If $Y_1, \ldots, Y_n$ are independent random variables, and the n functions $g_i(Y_i)$ are each a function of just one Y_i,

$$E\{\prod g_i(Y_i)\} = \prod E\{g_i(Y_i)\},$$

provided the expectations exist.

Proof:

Corollary: In particular, if Y_1 and Y_2 are independent, then

$$E(Y_1 Y_2) = E(Y_1)E(Y_2).$$

This is a very useful result.

We can now prove the following important result:

If X and Y are independent random variables with moment generating functions $m_X(t)$ and $m_Y(t)$, and $Z = X + Y$, then

$$m_Z(t) = m_X(t)m_Y(t)$$

on the common interval where both mgfs exist.

Proof:

By induction, this result can be extended to sums of several independent random variables.

Example 5.9:

If X follows a gamma distribution with parameters (r, β), and Y follows a gamma distribution with parameters (s, β), then the mgf of $X + Y$ is

$$\left(\frac{1}{1-\beta t}\right)^r \left(\frac{1}{1-\beta t}\right)^s = \left(\frac{1}{1-\beta t}\right)^{r+s}$$

which is also gamma with parameters $(r+s, \beta)$.

Note that this example is atypical. For example, if the scale parameters are different, we don't get a gamma distribution.

5.7 Covariance

We have already defined the variance of a random variable as a measure of its variability:

$$\text{Var}(Y) = E\{(Y-\mu)^2\},$$

where $\mu = E(Y)$.

The *covariance* of two random variables is a measure of their linear dependence or association, or joint variability.

Definition 5.8: The **covariance** of two random variables X and Y is defined as

$$\text{Cov}(X, Y) = E\{(X-\mu_X)(Y-\mu_Y)\},$$

where $\mu_X = E(X)$ and $\mu_Y = E(Y)$. This is the average value of the product of the deviation of X from its mean and the deviation of Y from its mean.

A snag is that covariance depends on the scale of measurement, which makes it hard to assess what is 'big' and what is 'small'. So we standardise it:

Definition 5.9: The **correlation** ρ of two random variables X and Y is defined as

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

This is a *dimensionless* measure of the association between two random variables.

Theorem 5.7: The correlation of any two random variables satisfies $|\rho| \leq 1$, with equality if and only if there is a linear relationship between the two.

Theorem 5.8:
(i) $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$.
(ii) If X and Y are independent, then $\text{Cov}(X, Y) = 0$.

Proof:

Note: The converse of (ii) is not true in general, and zero covariance does not imply independence unless X and Y are jointly normally distributed.

Example 5.3 (cont.): $f(x,y) = 2, 0 \leq y \leq x \leq 1$. Find the variances of X and Y and their covariance.

Example 5.10: Suppose X is uniform on $(-1,1)$. Let $Y = X^2$. Are X and Y independent? Find Cov(X,Y).

5.8 Linear combinations

This section is about finding the mean and variance of linear combinations of random variables, not necessarily independent.

Let $U = X + Y$. Then

$$\begin{aligned} \text{Var}(U) &= \text{Var}(X+Y) \\ &= E[\{(X+Y) - (\mu_X + \mu_Y)\}^2] \\ &= \ldots \\ &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y). \end{aligned}$$

This result generalises to more than two random variables, and to more general linear combinations. For example,

$$\text{Var}(aX+bY) = a^2\text{Var}(X)+b^2\text{Var}(Y)+2ab\text{Cov}(X,Y)$$

WMS p.257 works through the more general cases in detail. Here we give some further key results:

Let $U = \sum_{i=1}^{n} a_i Y_i$, a linear combination. Then

$$\text{Var}(U) = \sum_{i=1}^{n} a_i^2 \text{Var}(Y_i) + 2 \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \text{Cov}(X, Y)$$

where the double sum is over all pairs (i, j) for which $i < j$.

Now let $V = \sum_{j=1}^{m} b_j X_j$, a linear combination of a different set of random variables. Then the covariance between the two linear combinations is

$$\text{Cov}(U, V) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \text{Cov}(Y_i, X_j).$$

If X and Y are independent, then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

And this result generalises to more than two random variables: if Y_i, $i = 1, \ldots, n$ are independent, then

$$\text{Var}\left(\sum_{i=1}^{n} Y_i \right) = \sum_{i=1}^{n} \text{Var}(Y_i).$$

Example 5.11: For any independent random variables Y_i, $i = 1, \ldots, n$, with common mean μ and variance σ^2, it follows that

$$E(\bar{Y}) = \mu, \quad \text{Var}(\bar{Y}) = \frac{\sigma^2}{n}.$$

Example 5.12: Consider $Y \sim B(n, p)$.

Now let X_i be the indicator function defined by

$$X_i = \begin{cases} 1 & \text{if success at trial} \\ 0 & \text{if failure at trial } i \end{cases}$$

That is, X_i, $i = 1, \ldots, n$, are independent Bernoulli trials.

It is easily shown that $E(X_i) = p$, $\text{Var}(X) = p(1-p)$, and that $Y = \sum_{i=1}^{n} X_i$, where the X_i are independent. It follows that $E(Y) = np$, $\text{Var}(Y) = np(1-p)$.

5.9 Multinomial distribution

This generalises the binomial distribution to cases where there are more than the two (success, failure) categories. As with the binomial, we can define it as:

• There are n independent, identical, trials.
• The outcome of each trial falls into one of k classes or cells.
• At each trial, there is a probability p_i of falling into the ith class or cell, where $\sum_{i=1}^{k} p_i = 1$.
• The random variables are the numbers $Y_1, \ldots, Y_k$ falling into each of the k classes; note that $\sum Y_i = n$.

l *We will motivate the formal definition by a classification problem.*

A simple classification problem:

Suppose we have n randomly sampled individuals, and we want to classify each according to one of three blood types characterised by erythrocyte antigen. The three blood phenotypes are M, MN, N with probabilities p_1, p_2, p_3, where $\sum_{i=1}^{3} p_i = 1$.

Suppose we observe that y_1 individuals fall into class 1, y_2 into class 2, and y_3 into class 3, where $\sum_{i=1}^{3} y_i = n$, and y_i is the observed value of the random variable Y_i, representing the number of individuals who fall into class i.

The probability of observing such an outcome is

$$p_1^{y_1} p_2^{y_2} p_3^{y_3}.$$

How many ways can this occur?

There are

$$\binom{n}{y_1}\binom{n-y_1}{y_2}\binom{n-y_1-y_2}{y_3} = \frac{n!}{y_1!y_2!y_3!}$$

ways of obtaining the above probability.

This quantity is called the *multinomial coefficient*.

Thus,

$$P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) = \frac{n!}{y_1!y_2!y_3!} p_1^{y_1} p_2^{y_2} p_3^{y_3}.$$

This is called the *trinomial distribution* and generalises to the *multinomial distribution* for k classes.

Definition 5.10: If $p_1, \ldots, p_k$ are each $>$ 0 and $\sum p_i = 1$, then the random variables $Y_1, \ldots, Y_k$ have a **multinomial** distribution Mn$(n; p_1, \ldots, p_k)$ if the joint distribution is

$$p(y_1, \ldots, y_k) = \frac{n!}{y_1! \ldots y_k!} p_1^{y_1} \ldots p_k^{y_k},$$

where $\sum y_i = n$.

Example 5.13: Suppose a Poisson process has a mean of 2 events per hour. In a period of 3 hours, suppose we observe 5 events. What is the probability of at least one in each hour?

Some key results:

We can show that the marginal distribution of Y_1 is binomial (n, p_1). Similarly for Y_2 and Y_3.

Also, Cov$(Y_1, Y_2) = -np_1p_2$.

Note that the covariance is negative. This is because if there is a large number of outcomes in class 1, this would force the number of outcomes in class 2 to be small, and vice versa.

Theorem 6.3: If $Y = (Y_1, \ldots, Y_n)$ are independent normal variables with means $\mu = (\mu_1, \ldots, \mu_n)$ and variances $\sigma_1^2, \ldots, \sigma_n^2$, and if $Z_i = (Y_i - \mu_i)/\sigma_i$, then $W = \sum Z_i^2$ has a χ^2 distribution with n degrees of freedom.

Proof: We have already seen that $n = 1$ gives a χ^2 with 1 d.f., and that this is Gamma$(1/2, 2)$. It has an mgf of $(1 - 2t)^{-1/2}$.

Then

$$m_W(t) = \prod_{i=1}^{n} (1 - 2t)^{-1/2} = (1 - 2t)^{-n/2},$$

which is the mgf of a Gamma$(n/2, 2)$ distribution.

It follows that the density of W is given by

$$f_W(w) = \frac{1}{2^{n/2}\Gamma(n/2)} w^{n/2-1} e^{-w/2}.$$

Theorem 6.4: If $Y_1, \ldots, Y_n$ are independent $N(\mu, \sigma^2)$, then
- $\bar{Y} \sim N(\mu, \sigma^2/n)$,
- $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, and
- these two are independent.

Proof: Not given, but uses the mgf method.

6 FUNCTIONS OF RANDOM VARIABLES

6.1 The three methods

In many cases, we form a *statistic* $W = g(Y)$ based on a random sample $Y = (Y_1, \ldots, Y_n)'$ of size n.

We then need the distribution of W.

In this section, we discuss *three* methods for doing this:

1. Distribution functions
- $F(w) = P(W \leq w)$,
- we find the probability of lying in the region $(y_1, \ldots, y_n)$ defined by $W \leq w$.

2. Transformations
- we transform to include W, then
- integrate out the other random variables.

3. Moment generating functions
- defined as $E(e^{tY})$ for general t,
- there is in general a 1-1 correspondence between probability distributions and moment generating functions.

Example 6.1: Consider X and Y each uniform on (0, 1). What is the distribution of $W = X + Y$?

$$f(x, y) = 1, \quad 0 \leq x, y \leq 1.$$

$F_W(w) = P(X + Y \leq w) =$

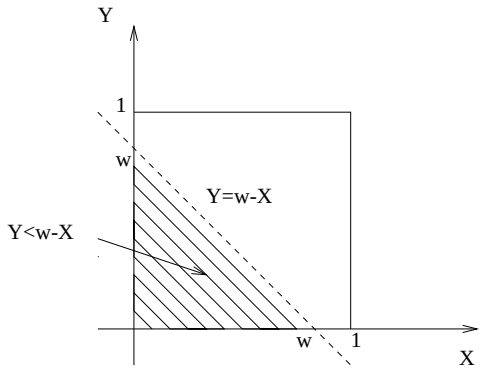

Fig. 6.1: Bivariate uniform distribution.

Example 6.2: Consider the exponential distribution $f(y) = \exp(-y)$,
for which $F(y) =$

Consider the transformation $W = -\log(Y)$. Then $P(W \leq w) =$

Then the density of W is:

So how can we generate exponentials with mean β?

Example 6.3: Suppose X and Y are each $N(0,1)$. Let $W = Y/X$. Use the cdf method to determine the distribution of W.

$P(W \leq w) = P(Y/X \leq w) =$

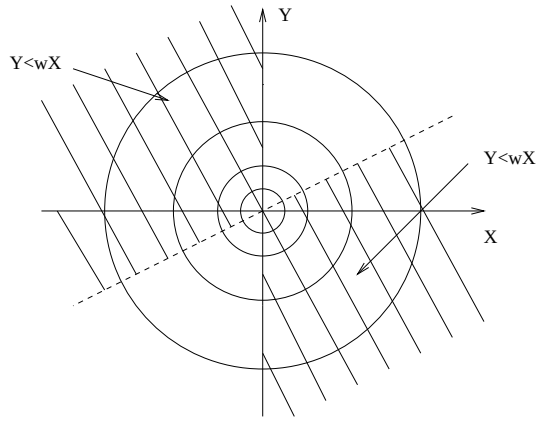

Fig. 6.2: Bivariate normal distribution.

Example 6.4: Maximums and minimums. This method is particularly useful for finding the distribution of the maximum or minimum of n independent random variables.

Suppose $(Y_1, \ldots, Y_n)$ are independent uniforms on (0,1). What is the distribution of $Y_{(n)}$, the largest order statistic?

$P(Y_{(n)} \leq y) =$

Exercise: Find the distribution of the minimum in the same way.

Summary of method:

- Find the region $W = w$ in $(y_1, \ldots, y_n)$,
- find the region defined by $W \leq w$,
- find $F_W(w) = P(W \leq w)$ by integrating out $f(y_1, \ldots, y_n)$ over the region,
- find $f_W(w)$ by differentiating $F_W(w)$.

6.3 Transformations

We are used to 'changing variables' to evaluate integrals.

However, we often need the distribution of W where $W = g(Y)$. It turns out to be essentially the same thing.

Consider first 'linear functions'.

If $f_Y(y)$ is uniform on (0,1), and $W = aY + b$, then we know that to keep the area equal to 1, we need to rescale the vertical axis, and that for $a > 0$

$$f_W(w) = \begin{cases} 1/|a| & \text{if } b < w < a + b, \\ 0 & \text{otherwise} \end{cases}$$

Reason:

• Recall how the vertical axis in histograms was **relative frequency per unit length on the x-axis**, this being necessary to get areas which sum to 1.

• Probability density gives the **probability per unit length** in the same way.

• Hence, to map **areas** onto **areas**, we need to watch how the horizontal axis gets compressed or expanded, and then do the **reverse** to the vertical axis.

• It follows that we require

$$f_W(w)|dw| = f_Y(y)|dy|,$$

and hence that

$$f_W(w) = f_Y(y)|dy/dw|.$$

Example 6.4: Distribution of $W = Y^2$, when Y is uniform.

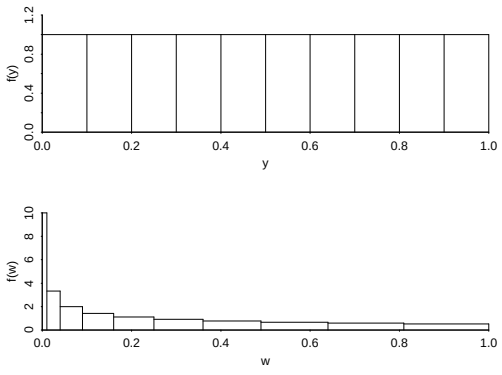

Fig. 6.3: Transforming the uniform by $W = Y^2$.

1:1 Differentiable functions

Suppose we have a density $f_Y(y)$ defined on (a, b).

Consider the transformation, or 'change of variable' to $W = g(Y)$. Then W is defined on the interval $(g(a), g(b))$.

The method is:

- invert the transformation as $Y = h(W)$,
- we require $f_W(w)|dw| = f_Y(y)|dy|$,
- form the Jacobian $|dy/dw| = |dh(w)/dw|$,
- the density function $f_W(w)$ for W is then

$$f_W(w) = f_Y(y)|dy/dw|,$$

where both y and dy/dw are expressed solely in terms of w, i.e.

$$f_W(w) = f_Y\{h(w)\}|dh(w)/dw|.$$

Note: the Jacobian is always $|d\text{old}/d\text{new}|$.

Example 6.4 (cont.): $W = Y^2$, when Y is uniform.

In this case, $g(y) = y^2$ is 1:1 and differentiable over the interval in question, i.e. $(0, 1)$.

Here $h(w) = \sqrt{w}$, and the Jacobian is

$$dh(w)/dw = 1/(2\sqrt{y}).$$

We know that $f_Y(y) = 1$ over the interval. Thus we have

$$f_W(w) = f_Y(y).|dh(w)/dw| = 1/(2\sqrt{y}),$$

which is also defined on $(0, 1)$.

Example 6.5: Log of uniform.

Notes:
• This is used as a way of generating random exponential distributions.
• As an exercise, what happens if we let $W = -\beta \log(Y)$?

Many:1 transformations

A function $w = g(y)$ may be differentiable but have several cases where different values of y lead to the same value of w.

Note: you will see this when you express y as a function of w.

In this case,

$$f_W(w) = \sum_{y:g(y)=w} \frac{f_Y(y)}{|dy/dw|}.$$

Example 6.6: $W = Y^2$, in general.

Here $y = -\sqrt{w}$ and $y = +\sqrt{w}$ both give the same value of w, so the density of W is

$$f_W(y) = \{f_Y(\sqrt{w}) + f_Y(\sqrt{w})\}/(2\sqrt{w}).$$

Chi-square distribution

Suppose $Y \sim N(0,1)$, and $W = Y^2$. What is the density of W?

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2}, \quad y > 0.$$

This is the chi-square distribution with 1 degree of freedom, denoted $Y \sim \chi^2_1$.

This is a gamma distribution with $r = 1/2; \beta = 2$.

Exercise: Show that $P(W < 3.84) = 0.95$.

Summary of method

- Consider the transformation $W = g(Y)$.
- Invert the function and express Y in terms of W as $Y = h(W)$.
- This will identify if $g(Y)$ is monotone.
- If $W = g(Y)$ is monotone,

$$f_W(w) = f_Y(y)|dy/dw|,$$

where we replace y by $h(w)$ in both $f_Y(y)$ and $|dy/dw|$.

- And if $W = g(Y)$ is a many:1 function,

$$f_W(w) = \sum_{y:w=g(y)} f_Y(y)|dy/dw|,$$

where again we replace y by $h(w)$.

6.4 Moment generating functions

The **moment generating function** $m(t)$ of Y is defined as $E(e^{ty})$.

The mgf for Y exists if there is some $b > 0$ such that $m(t) < \infty$ for $|t| < b$.

Theorem 6.1: If for X and Y the moment generating functions $m_X(t)$ and $m_Y(t)$ exist, and if $m_X(t) = m_Y(t)$ for all values of t, then X and Y have the same disribution.

Proof: Not given.

Example 6.7: Suppose that Y is normal with mean μ, variance σ^2. Show that $W = (Y-\mu)/\sigma$ is $N(0,1)$.

First, the mgf of a standard normal is:

$$\begin{aligned} E\{e^{tZ}\} &= \frac{1}{\sqrt{2\pi}}\int \exp(tz - z^2/2)dz \\ &= \frac{1}{\sqrt{2\pi}}\int \exp\{-(z-t)^2/2 + t^2/2\}dz \\ &= \exp(t^2/2). \end{aligned}$$

In the same way, if $Y \sim N(\mu, \sigma^2)$, its mgf is:

$$m_Y(t) = E\{\exp(tY)\} = \exp(\mu t + t^2\sigma^2/2).$$

Now

$$\begin{aligned} m_W(t) = E(e^{tW}) &= E\{\exp(tY/\sigma - \mu t/\sigma)\} \\ &= \exp(-\mu t/\sigma) m_Y(t/\sigma) \\ &= \end{aligned}$$

Example 6.8: Sums of independent Poissons random variables.
The moment generating function of a Poisson, mean λ, is

$$m(t) = E\{e^{tY}\} = \exp\{\lambda(e^t - 1)\}.$$

If $Y_1, \ldots, Y_n$ are n independent Poissons with means $\lambda_1, \ldots, \lambda_n$, then the mgf of their sum $W = Y_1 + \ldots + Y_n$ is:

$m_W(t) =$

Theorem 6.2: If $Y = (Y_1, \ldots, Y_n)$ are independent normal variables with means $\mu = (\mu_1, \ldots, \mu_n)$ and variances $\sigma_1^2, \ldots, \sigma_n^2$, and if $W = a'Y = a_i Y_1 + \ldots a_n Y_n$, then W is normal with
• mean $\mu_W = \sum a_i \mu_i = a'\mu$ and
• variance $\sigma_W^2 = \sum a_i \sigma_i^2 = a'\Sigma a$,
where $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$.

Proof: We already know (Theorem 5.9) that the mean and variance are correct.

$$E\{\exp(tW)\} \;=\; E\{\exp(t \sum a_i Y_i)\}$$

7 Appendix of additional topics (not examinable)

7.1 Bivariate normal distribution

We start with the 'standardised bivariate normal with a correlation ρ'. The density is given by:

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)}\right),$$

where $-\infty < x, y < \infty$.

This is denoted by $(X,Y) \sim N_2(\rho)$.

Figure 7.1 shows the densities for these for $\rho = 0.7, 0.9$.

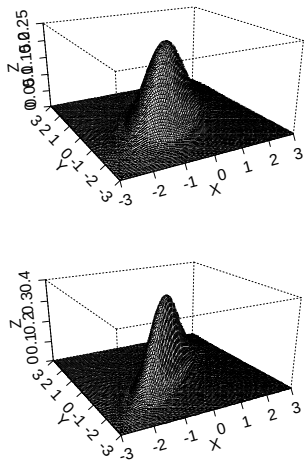

Fig. 7.1: Bivariate normals, with $\rho = 0.7, 0.9$.

Marginal density of X:

• We have to integrate out y.

• Take the exponent and make the Y part of it look like a normal density by completing the square:

$$\begin{aligned}(x^2+y^2-2\rho xy) &= x^2+(y-\rho x)^2-\rho^2x^2\\ &= (1-\rho^2)x^2+(y-\rho x)^2\end{aligned}$$

• The integral over y now looks like a normal with mean ρx and variance $(1-\rho^2)$, so we get:

$$\begin{aligned}f_X(x) &= \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\times\\ &\quad \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi(1-\rho^2)}}\exp\left(-\frac{(y-\rho x)^2}{2(1-\rho^2)}\right)dy\\ &= \frac{1}{\sqrt{2\pi}}e^{-x^2/2}.\end{aligned}$$

• Hence the marginal density of X is just N(0,1).

Conditional density:

Here again we use the ratio:

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

Use the version of $f(x,y)$ where we completed the square, and divide by $f_X(x)$:

$$f_{Y|X}(y|X=x)=\frac{1}{\sqrt{2\pi(1-\rho^2)}}\exp\left\{-\frac{(y-\rho x)^2}{2(1-\rho^2)}\right\}.$$

What does this say?

• If we don't know X, then Y is just N(0,1).

• If we are *told* the value of X is x, then Y is still normal, but has mean ρx and variance $(1-\rho^2)$.

• Knowing $X = x$ helps us *predict* the value of Y, with a variance *smaller* than the unconditional variance.

• The closer ρ is to 1, the better we can predict the value of Y, given $X = x$.

General bivariate normal

If X, Y have means μ_X, μ_Y and standard deviations σ_X, σ_Y, and if

$$\{(X-\mu_X)/\sigma_X, (Y-\mu_Y)/\sigma_Y\}$$

is $N_2(\rho)$, then (X, Y) is bivariate normal, or

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left\{ \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \right\}$$

Exercise: Show that, in this case, the distribution of Y given $X = x$ is

$$N\left(\mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(x-\mu_X), \sigma_Y^2(1-\rho^2)\right).$$

Independence in bivariate normal

In the standardised bivariate normal with $\rho = 0$,

$$f(x,y) = \frac{1}{\sqrt{2\pi}} \exp\left\{-(x^2+y^2)/2\right\} = \phi(x)\phi(y).$$

It follows by Theorem 5.4 that X and Y are independent.

This holds in general for bivariate normal; this is one case where zero covariance implies independence.

Note for Statistics level III subjects: These results generalise to the multivariate case, where $E(Y) = \mu$ and $\mathrm{Var}(Y) = \Sigma$ contains the variances and covariances. The density is then:

$$f(y) = \frac{1}{2\pi^{n/2}\det(\Sigma)} \exp\left\{-(y-\mu)'\Sigma^{-1}(y-\mu)/2\right\}.$$

Origin of 'regression'

Francis Galton looked at heights of fathers (X) and their sons (Y). Each is marginally normal with the same σ and a correlation of about $\rho = 0.6$.

Given the father's height (i.e. $X = x$), the son's height Y is predicted to be

$$E(Y|X = x) = \mu_y + \rho(x - \mu_x),$$

which is a line of slope ρ as shown in Fig. 5.2, *not* a line of slope 1.

Galton called this 'regression towards the mean'.

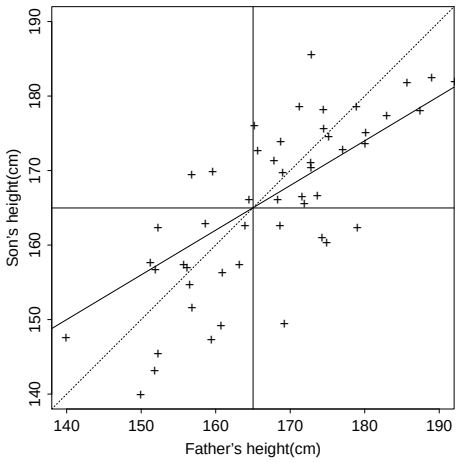

Fig. 7.2: Heights of fathers and sons.

7.2 Conditional expectation

Definition: For two random variables (X, Y), the **conditional expectation** of $g(Y)$ given $X = x$ is defined as

$$E\{g(Y)|X = x\} = \int_y g(y) f_{Y|X}(y|x) dy$$
$$= h(x), \text{say.}$$

Example 5.3 (cont.): We have $f(y|x) = 1/x, 0 < y < x$. Then $E(Y|X = x) =$

Now, the expectation depends on X and can be regarded in two ways:
• for given $X = x$, it specifies a value, or
• since X is a random variable, then $h(X) = E\{g(Y)|X = x\}$ can be thought of as a random variable with its own distribution, mean, etc.

Theorem: For random variables X and Y,

$$E\{g(Y)\} = E_X[E\{g(Y)|X = x\}],$$

where the **inside** expectation is over $Y|X = x$ and the **outer** one over the marginal distribution of X.

Proof: If we denote $h(X) = E\{g(Y)|X = x\}$,

$$\begin{aligned} E_X[h(X)] &= \int_x h(x) f_X(x) dx \\ &= \int_x \int_y g(y) f_{Y|X}(y|x) f_X(x) dy dx \\ &= \int \int g(y) f(x, y) dy dx \\ &= E\{g(Y)\}. \end{aligned}$$

Exercise: Show that if (X, Y) in $N_2(\rho)$ then $E(XY) = \rho$.

We use the previous result:

$$\begin{aligned} E(XY) &= E_X[E_{Y|X}\{XY\}] \\ &= \end{aligned}$$

Theorem: For random variables X and Y,

$$\begin{aligned}\text{Var}\{g(Y)\} &= \text{Var}[E\{g(Y)|X=x\}] \\ &\quad +E[\text{Var}\{g(Y)|X=x\}],\end{aligned}$$

where the **inside** expectation is over $Y|X=x$ and the **outer** one over the marginal distribution of X.

Proof: The previous result holds with $g(Y)$ equal to both Y and Y^2.

Example: Suppose we observe events occurring as a Poisson process, where each event is a Bernoulli trial. If we observe the process for a fixed period of time, the number of events N is Poisson with mean λ. We obtain a value n. Then, given the value of $N=n$, the number of successes Y is Bin(n,p).

What is the mean and variance of Y?