

# Contents

## 1. Probability

1.1 Introduction	3
1.2 Sample Spaces	3
1.3 Events	8
1.4 Probability Distributions	15
1.5 Probability Axioms	19
1.6 Conditional Probability	24
1.7 Bayes Theorem	31
1.8 Statistical Independence	35
1.9 Random Variables	38
1.10 Problems	41
1.11 Key Probability Results for Chapter 1	43

## 2. Discrete Probability Distributions

2.1 Introduction	45
2.2 Distribution of Transformed Random Variables	54
2.3 Examples of Discrete Distributions	56
2.4 The Distribution Function, $F_X(x)$	77
2.5 Independent Random Variables	80

## 3. Continuous Random Variables

3.1 Introduction	81
3.2 Examples of Continuous Distributions	91
3.3 Finding the Distribution of $g(X)$	110
3.4 Generating Random Numbers from Continuous Distributions	118

## 4. Multivariate Distributions

4.1 Discrete Bivariate Distributions	123
4.2 Expectation over a Joint Distribution	133
4.3 Covariance and Correlation between Two Random Variables	136
4.4 Conditional Expectation and Conditional Variance	143
4.5 Examples of Discrete Multivariate Distributions	151
4.6 Continuous Joint Distributions	155
4.7 Independence of Continuous Random Variables	165
4.8 Expectation of Jointly Continuous Random Variables	169
4.9 Change of Variable Technique for Continuous Bivariate Distributions	173
4.10 Example of Joint Continuous Distribution: the Bivariate Normal	181

## 5. Moment Generating Functions

5.1 Introduction	185
5.2 Moments	190
5.3 MGFs for Sums of Independent Random Variables	195
5.4 Compound Distributions	197
5.5 Using the MGF to find the Distribution of $g(X)$	199
5.6 Limiting Distributions	200

## 6. Sampling Theory for the Normal Distribution

6.1 Introduction	205
6.2 Distribution Theory	206
6.3 Application to Confidence Intervals and t-Tests	212

# Chapter 1 : Probability

## 1.1 Introduction

*Definition:* A **probability** is a number between 0 and 1 representing how likely it is that an event will occur.

Probabilities can be:

1. *Frequentist (based on frequencies)*

eg.  $\frac{\text{number of times event occurs}}{\text{number of opportunities for event to occur}}$

or

2. *Subjective: probability represents a person's degree of belief that an event will occur,*

eg. *I think there is an 80% chance it will rain today,*  
written as  $\mathbb{P}(\text{rain}) = 0.80$ .

Regardless of how we obtain probabilities, we always combine and manipulate them according to the same rules.

## 1.2 Sample Spaces

*Definition:* A **random experiment** is an experiment whose outcome is *not known until it is observed*.

*Definition:* A **sample space**,  $S$ , is a *set of outcomes of a random experiment*.

Every possible outcome must be listed *once and only once*.

*Definition:* A sample point is an *element of the sample space*.

For example, if the sample space is  $S = \{s_1, s_2, s_3\}$ , then each  $s_i$  is a sample point.

### Examples:

*Experiment:* Toss a coin twice and observe the result.

*Sample space:*  $S = \{HH, HT, TH, TT\}$

An example of a sample point is:  $HT$

*Experiment:* Toss a coin twice and count the number of heads.

*Sample space:*  $S = \{0, 1, 2\}$

*Experiment:* Toss a coin twice and observe whether the two tosses are the same (e.g. HH or TT).

*Sample space:*  $S = \{\text{same, different}\}$

### Types of Sample Space

*Definition:* A sample space is finite if *it has a finite number of elements*.

*Definition:* (Informal definitions) A sample space is discrete if *there are “gaps” between the different elements, or if the elements can be “listed”, even if an infinite list (eg. 1, 2, 3, ...)*.

*(See formal definition later.)*

A sample space is continuous if *there are no gaps between the elements, so the elements cannot be listed (eg. the interval  $[0, 1]$ )*.

## Examples:

$S = \{0, 1, 2, 3\}$  (*discrete and finite*)

$S = \{0, 1, 2, 3, \dots\}$  (*discrete, infinite*)

$S = \{4.5, 4.6, 4.7\}$  (*discrete, finite*)

$S = \{HH, HT, TH, TT\}$  (*discrete, finite*)

$S = \{\text{same, different}\}$  (*discrete, finite*)

$S = [0, 1] = \{\text{all numbers between 0 and 1 inclusive}\}$  (*continuous, infinite*)

**Example:** Sampling with or without replacement.

We have a group of  $N$  people, e.g. students in this class, listed alphabetically.

Let  $x_i$  be the name of student  $i$ , e.g.  $x_3 = \mathbf{Fred}$

*Experiment:* choose one person at random.

*Sample space:*  $S = \{x_1, \dots, x_N\}$  (**discrete, finite**)

*Experiment:* choose two people at random, *without* replacement.

*Sample space:* (two possibilities)

1. **Order matters, so (Fred, Jane) is different from (Jane, Fred).**

$S = \{(x_i, x_j) : i, j = 1, 2, \dots, N, \text{ and } i \neq j\}$ .

2. **Order doesn't matter, so (Fred, Jane) is the same outcome as (Jane, Fred).**

$S = \{(x_i, x_j) : i, j = 1, 2, \dots, N, \text{ and } i < j\}$

*Experiment:* choose two people at random, *with* replacement.

*Sample space:* (two possibilities)

1. **Order matters:**  $S = \{(x_i, x_j) : i, j = 1, 2, \dots, N\}$

2. **Order doesn't matter:**  $S = \{(x_i, x_j) : i, j = 1, 2, \dots, N \text{ and } i \leq j\}$

**Example:** Discrete infinite sample space.

*Experiment:* toss a coin until a Head appears, observe sequence of tosses.

*Sample space:*  $S = \{H, TH, TTH, TTTH, \dots\}$  (discrete, infinite)

$S$  is infinite because there is no number of tails after which a head definitely *must* appear.

*Alternative sample space:* count the number of tails before the first head.

$S = \{0, 1, 2, \dots\}$  (discrete, infinite)

*Question:* is  $S = \{0, 1, 2, (3 \text{ or more})\}$  a possible sample space?

*Answer:* Yes. (Discrete, finite).

*Question:* is  $S = \{1, 2, (3 \text{ or more})\}$  a possible sample space?

*Answer:* No: outcome 0 is omitted.

*Definition:* An infinite sample space is countable if we can index the elements by the natural numbers, 1,2,3,... That is, for every natural number there is a unique element of  $S$ , and for every element of  $S$  there is a unique natural number. (In Mathematical language, there is a bijection from  $\mathbb{N}$  to  $S$ ).

.

In practice, this just means that we can write:  $S = \{s_1, s_2, s_3, \dots\}$ .

Any countable sample space is discrete, because we can list the elements. This gives us our formal definition of a discrete sample space:

*Definition:* A sample space is discrete if it is finite or countable.

## Continuous sample spaces

Any sample space that is not discrete is **continuous**: there are no gaps between the elements. The most common example of a continuous sample space is ***an interval on the real line***.

### ***Example:***

*Experiment:* spin a pointer, and observe the angle  $\theta$  at which it stops.

*Sample space:*

$$S = \{\theta : 0^\circ \leq \theta < 360^\circ\}$$

or  $S = [0^\circ, 360^\circ)$  (continuous).

*Question:* Why not  $\theta = 360^\circ$ ?

*Answer:* **Because  $\theta = 360^\circ$  is the same as  $\theta = 0^\circ$ : outcome must be listed only once.**

*Other possible sample spaces:*

$$S = (0^\circ, 360^\circ] \text{ (continuous)}$$

or  $S = \{[0^\circ, 90^\circ), [90^\circ, 180^\circ), [180^\circ, 360^\circ)\}$  (discrete)

### ***Example:***

*Experiment:* install a light bulb, and observe the time taken before it fails.

*Sample space:*  $S = \{t : t \geq 0\} = [0, \infty)$ .

**Note:** Never write a square bracket after  $\infty$ , because we can never reach  $\infty$ .

**Write  $[0, \infty)$  but not  $[0, \infty]$ .**

### 1.3 Events

*Definition:* An **event** is a subset of the sample space. That is, any collection of outcomes forms an event.

**Example:** Toss a coin twice. Sample space:  $S = \{HH, HT, TH, TT\}$

Let event  $A$  be the event that there is **exactly one head**.

We write:  $A = \text{“exactly one head”}$

Then  $A = \{HT, TH\}$ .

$A$  is a subset of  $S$ , as in the definition. We write  $A \subset S$ .

*Definition:* Event  $A$  **occurs** if we observe an outcome that is a member of the set  $A$ .

**Note:**  $S$  is a subset of itself, so  $S$  is an event. Because  $S$  includes all possible outcomes of the experiment, **event  $S$  occurs every time the experiment is performed**.

The empty set,  $\emptyset = \{\}$ , is also a subset of  $S$ . This is called the **null event**, or **the event with no outcomes**.

**Example:**

*Experiment:* toss coin 3 times.

*Sample space:*  $S = \{HHH, HHT, \dots, TTT\}$

*Event  $A$*  = “no more than one Head” =  $\{HTT, THT, TTH, TTT\}$

*Experiment:* throw 2 dice.

*Sample space:*  $S = \{(1, 1), (1, 2), \dots, (1, 6), (2, 1), (2, 2), \dots, (2, 6), \dots, (6, 6)\}$

*Event  $B$*  = “sum of two faces is 5” =  $\{(1, 4), (2, 3), (3, 2), (4, 1)\}$



## Combining Events

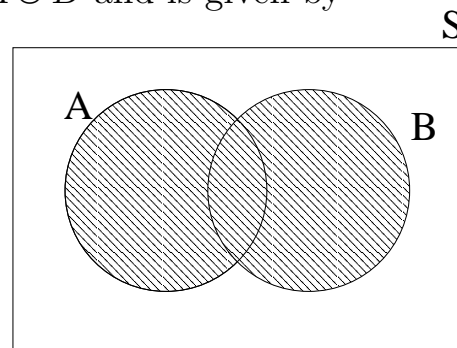
Let  $A$  and  $B$  be events on the same sample space  $S$ : so  $A \subset S$  **and**  $B \subset S$ .

*Definition:* The **union** of events  $A$  and  $B$  is written  $A \cup B$  and is given by

$$A \cup B = \{s : s \in A \text{ or } s \in B \text{ or both}\}$$

Think of  $A \cup B$  as ***A or B or both***.

On a Venn diagram, we show  $A \cup B$  as follows:



**Example:** Spin pointer. Sample space,  $S = [0^\circ, 360^\circ)$

Let event  $A =$  “*acute angle observed*” =  $[0^\circ, 90^\circ)$

Let event  $B =$  “*angle observed is  $> 45^\circ$* ” =  $(45^\circ, 360^\circ)$

Then  $A \cup B = [0^\circ, 90^\circ) \cup (45^\circ, 360^\circ) = [0^\circ, 360^\circ)$ .

**Example:** Pick a person in the class. Sample space,  $S = \{\text{all people in class}\}$

Let event  $A =$  “*person is a male*”

Let event  $B =$  “*person has a cellphone*”

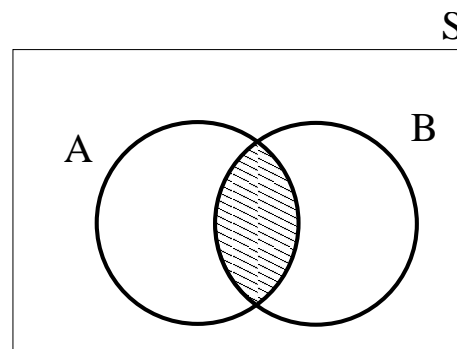
Then event  $A \cup B$  occurs if *the person picked is male, OR has a cellphone, OR both*.

*Definition:* The **intersection** of events  $A$  and  $B$  is written  $A \cap B$  and is given by

$$A \cap B = \{s : s \in A \text{ AND } s \in B\}$$

Think of  $A \cap B$  as “***A and B***”.

On a Venn diagram, we show  $A \cap B$  as follows:

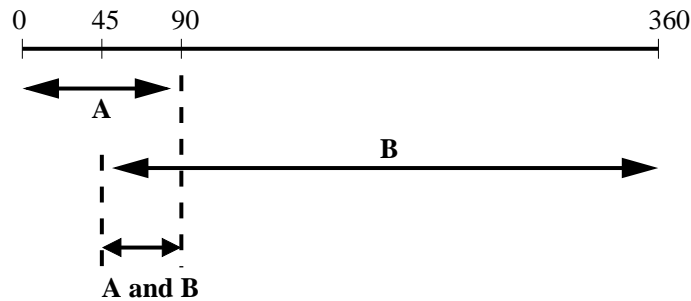


**Example:** Spin pointer. Sample space,  $S = [0^\circ, 360^\circ)$ .

Event  $A = [0^\circ, 90^\circ)$

Event  $B = (45^\circ, 360^\circ)$

Then  $A \cap B = [0^\circ, 90^\circ) \cap (45^\circ, 360^\circ) = (45^\circ, 90^\circ)$  (*angle is acute AND  $> 45^\circ$* )



**Example:** Pick person in class. Sample space,  $S = \{\text{people in class}\}$ .

Event  $A = \text{“person is male”}$

Event  $B = \text{“person has cellphone”}$

Then event  $A \cap B = \{\text{people in class who are male AND have a cellphone}\}$

*Question:* Suppose I pick a female with a cellphone. Which of the following events have occurred?

1)  $A$  No

2)  $B$  Yes

3)  $A \cup B$  Yes

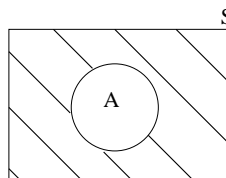
4)  $A \cap B$  No

**Definition:** The complement of event  $A$  is written  $\bar{A}$  and is given by

$$\bar{A} = \{s : s \notin A\}$$

That is,  $\bar{A}$  is the event “not  $A$ ”: *whatever  $A$  was, it didn’t happen.*

Venn diagram: ( $\bar{A}$  shaded)



**Example:** Spin pointer: let  $A = \text{“angle is acute”} = [0^\circ, 90^\circ)$

Then  $\bar{A} = \text{“angle is not acute”} = [90^\circ, 360^\circ)$

**Example:** Pick a person in the class: let  $A = \text{“person is male”}$

Then  $\bar{A} = \text{person is not male} = \{\text{females in class}\}$

*Question:* Let  $A$  = “person is male” and let  $B$  = “person has a cellphone”. Suppose I pick a **male without a cellphone**. Say whether the following events have occurred:

- 1)  $A$     **Yes.**
- 2)  $B$     **No.**
- 3)  $\bar{A}$     **No.**
- 4)  $\bar{B}$     **Yes.**
- 5)  $\bar{A} \cup B = \{\text{females or cellphone owners or both}\}$ . **No.**
- 6)  $A \cap \bar{B} = \{\text{males without cellphones}\}$ . **Yes.**
- 7)  $A \cap B = \{\text{males with cellphones}\}$ . **No.**
- 8)  $\overline{A \cap B} = \text{everything outside } A \cap B$ .  $A \cap B$  did not occur, so  $\overline{A \cap B}$  did occur. **Yes.**

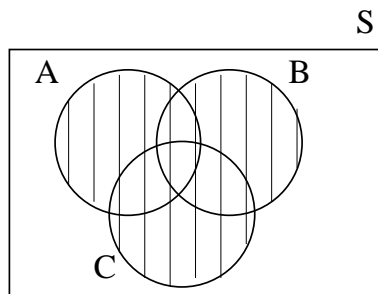
*Question:* What is the event  $\bar{S}$ ?     $\bar{S} = \emptyset$

*Challenge:* can you express  $A \cap B$  using only a  $\cup$  sign? **Answer:**  $A \cap B = \overline{(\bar{A} \cup \bar{B})}$

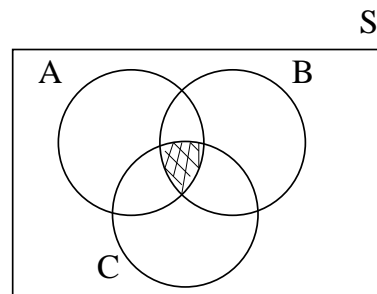
### More than two events

Venn diagrams are generally useful for up to 3 events, although they are not used to provide formal proofs.

**Example:**



(a)  $A \cup B \cup C$



(b)  $A \cap B \cap C$

**Theorem 1.1:** (Not proved here: these are results from Set Theory).

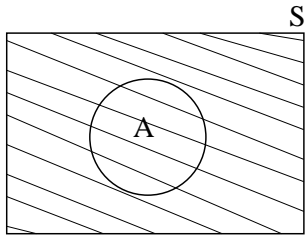
(i)  $\overline{\emptyset} = S$  and  $\overline{S} = \emptyset$

(ii) For any event  $A$ ,  $A \cup \overline{A} = S$  and  $A \cap \overline{A} = \emptyset$

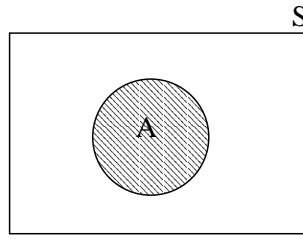
(iii) For any events  $A$  and  $B$ ,  $A \cup B = B \cup A$  and  $A \cap B = B \cap A$  (*Commutative*)

(iv) For any event  $A$ ,

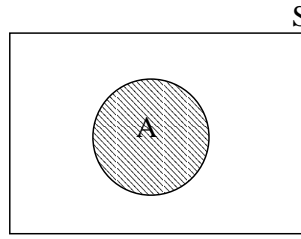
$A \cup S = S:$



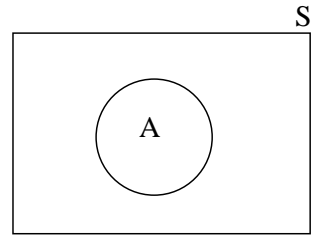
$A \cap S = A:$



$A \cup \emptyset = A:$

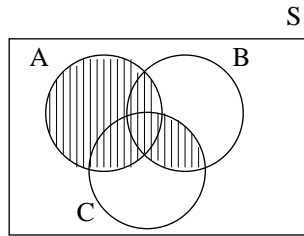


$A \cap \emptyset = \emptyset:$

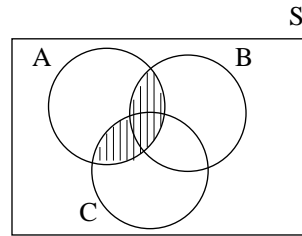


(v) For any  $A$ ,  $B$ , and  $C$ :

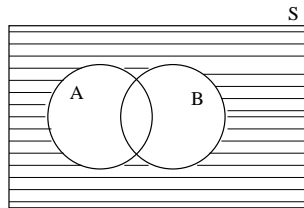
(a)  $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$



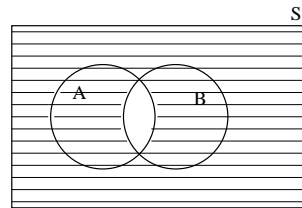
(b)  $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$



(vi) (a)  $\overline{(A \cup B)} = \overline{A} \cap \overline{B}$



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



(vii) Distributive Laws: (extension of (v)):

For events  $A$  and  $B_1, B_2, \dots, B_n$ ,

$$a) A \cup \left( \bigcap_{i=1}^n B_i \right) = \bigcap_{i=1}^n (A \cup B_i)$$

$$ie. A \cup (B_1 \cap B_2 \cap \dots \cap B_n) = (A \cup B_1) \cap (A \cup B_2) \cap \dots \cap (A \cup B_n)$$

$$b) A \cap \left( \bigcup_{i=1}^n B_i \right) = \bigcup_{i=1}^n (A \cap B_i)$$

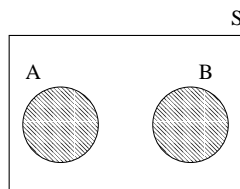
$$ie. A \cap (B_1 \cup B_2 \cup \dots \cup B_n) = (A \cap B_1) \cup (A \cap B_2) \cup \dots \cup (A \cap B_n)$$

**Note:** Often a good way to show that two sets  $A$  and  $B$  are equal is to show that  $A \subseteq B$  and  $B \subseteq A$ ; thus  $A = B$ .

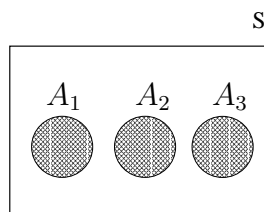
### Fundamental Idea: the Partition

*Definition:* Two events  $A$  and  $B$  are mutually exclusive, or disjoint, if  $A \cap B = \emptyset$ .

*This means events  $A$  and  $B$  cannot happen together. If  $A$  happens, it excludes  $B$  from happening, and vice-versa.*



*Definition:* Any number of events  $A_1, A_2, \dots, A_k$  are mutually exclusive if every pair of the events is mutually exclusive: ie.  $A_i \cap A_j = \emptyset$  for all  $i, j$  with  $i \neq j$ .



*Definition:* A partition of  $S$  is a *collection of mutually exclusive events whose union is  $S$ .*

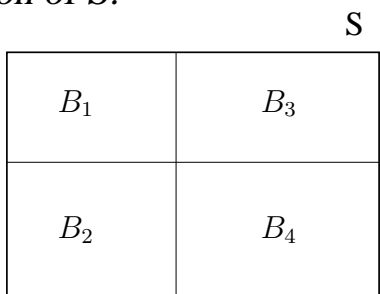
That is, sets  $B_1, B_2, \dots, B_k$  form a partition of  $S$  if

$$B_i \cap B_j = \emptyset \text{ for all } i, j \text{ with } i \neq j,$$

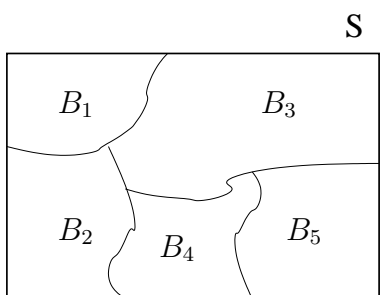
**and**  $\bigcup_{i=1}^k B_i = B_1 \cup B_2 \cup \dots \cup B_k = S.$

*Examples:*

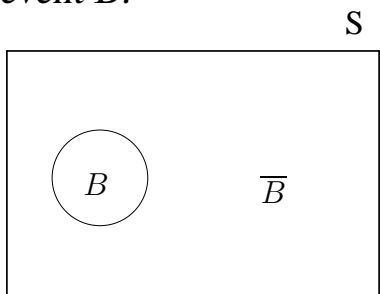
$B_1, B_2, B_3, B_4$  form a partition of  $S$ :



$B_1, \dots, B_5$  partition  $S$ :



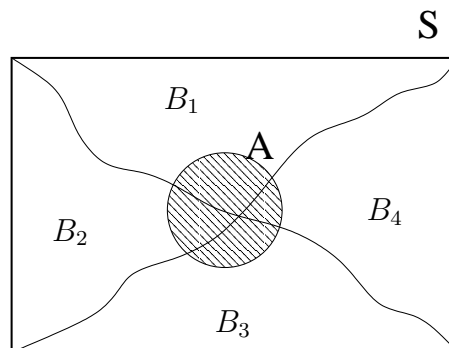
$B$  and  $\overline{B}$  partition  $S$  for any event  $B$ :



## Partitioning an event $A$

*Any set  $A$  can be partitioned: it doesn't have to be  $S$ .*

*If  $B_1, \dots, B_k$  form a partition of  $S$ , then  $(A \cap B_1), \dots, (A \cap B_k)$  form a partition of  $A$ .*



*We will see that this is very useful for finding the probability of event  $A$ .*

## 1.4 Probability Distributions

*Definition:* Let  $S = \{s_1, s_2, \dots\}$  be a discrete sample space.

A **discrete probability distribution** on  $S$  is a set of real numbers  $\{p_1, p_2, \dots\}$  associated with the sample points  $\{s_1, s_2, \dots\}$  such that:

1.  $0 \leq p_i \leq 1$  for all  $i$ ;

2. 
$$\sum_i p_i = 1$$

$p_i$  is called the *probability of the event that the outcome is  $s_i$* .

We write:  $p_i = \mathbb{P}(s_i)$ .

Although there are lots of choices for  $p_1, p_2, \dots$  that are valid (i.e. that fit the definition), we usually aim for  $p_i$  to be a measure of **how likely** outcome  $s_i$  is.

## Probability of an event in a discrete sample space

*Definition:* For a discrete sample space and probability distribution, the probability of an event  $A$  is *the sum of probabilities of the sample points in  $A$ .*

Thus if  $A = \{s_3, s_5, s_{14}\}$   
then  $\mathbb{P}(A) = p_3 + p_5 + p_{14}$

**Notes:**

- i)  $\mathbb{P}(S) = 1$
- ii)  $0 \leq \mathbb{P}(A) \leq 1$  for any event  $A$ .

### Equally likely outcomes

Sometimes, all the outcomes in a discrete finite sample space are equally likely. This makes it easy to calculate probabilities. If:

- i)  $S = \{s_1, \dots, s_k\}$ ;
- ii) each outcome  $s_i$  is equally likely, so  $p_1 = p_2 = \dots = p_k = \frac{1}{k}$ ;
- iii) event  $A = \{s_1, s_2, \dots, s_r\}$  contains  $r$  *possible outcomes*,

then

$$\mathbb{P}(A) = \frac{r}{k} = \frac{\# \text{ outcomes in } A}{\# \text{ outcomes in } S}.$$

**Example:** For a 3-child family, possible outcomes from oldest to youngest are:

$$\begin{aligned} S &= \{GGG, GGB, GBG, GBB, BGG, BGB, BBG, BBB\} \\ &= \{s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8\} \end{aligned}$$

Let  $\{p_1, p_2, \dots, p_8\}$  be a probability distribution on  $S$ . If every baby is equally likely to be a boy or a girl, then *all of the 8 outcomes in  $S$  are equally likely*, so  $p_1 = p_2 = \dots = p_8 = \frac{1}{8}$ .



Let event  $A$  be  $A = \text{“oldest child is a girl”}$ .

Then  $A = \{GGG, GGB, GBG, GBB\}$ .

Event  $A$  contains 4 of the 8 equally likely outcomes, so event  $A$  occurs with probability  $\mathbb{P}(A) = \frac{4}{8} = \frac{1}{2}$ .

### Counting equally likely outcomes

To count the number of equally likely outcomes in an event, we often need to use *permutations* or *combinations*. The number of ways of choosing  $r$  objects from  $n$  distinct objects is:

1) when order matters (so  $(a,b,c)$  is a different choice from  $(b,a,c)$ ):

$$\# \text{permutations} = {}^n P_r = n(n-1)(n-2) \dots (n-r+1) = \frac{n!}{(n-r)!}.$$

( $n$  choices for first object,  $(n-1)$  choices for second, etc.)

2) when order doesn't matter (so  $(a,b,c)$  and  $(b,a,c)$  are the same choice):

$$\# \text{combinations} = {}^n C_r = \binom{n}{r} = \frac{{}^n P_r}{r!} = \frac{n!}{(n-r)!r!}.$$

(because each of the  ${}^n P_r$  permutations is counted  $r!$  times if order doesn't matter).

**Example:** (a) Tom has five elderly great-aunts who live together in a tiny bungalow. They insist on each receiving separate Christmas cards, and threaten to disinherit Tom if he sends two of them the same picture. Tom has Christmas cards with 12 different designs. In how many different ways can he select 5 different designs from the 12 designs available?

*Number of ways of selecting 5 distinct designs from 12 is*

$${}^{12} C_5 = \binom{12}{5} = \frac{12!}{(12-5)!5!} = 792.$$

b) The next Christmas, Tom buys a pack of 40 Christmas cards, featuring 10 different pictures with 4 cards of each picture. He selects 5 cards at random to send to his great-aunts. What is the probability that at least two of the great-aunts receive the same picture?

*Looking for  $\mathbb{P}(\text{at least 2 cards the same}) = \mathbb{P}(A)$  (say).*

*Easiest to find  $\mathbb{P}(\text{all 5 cards are different}) = \mathbb{P}(\bar{A})$ .*

*Number of outcomes in  $\bar{A}$  is*

*(# ways of selecting 5 different designs) =  $40 \times 36 \times 32 \times 28 \times 24$ .*

*(40 choices for first card; 36 for second, because the 4 cards with the first design are excluded; etc.*

*Note that order matters: e.g. we are counting choice 12345 separately from 23154.)*

*Total number of outcomes is*

*(total # ways of selecting 5 cards from 40) =  $40 \times 39 \times 38 \times 37 \times 36$ .*

*(Note: order mattered above, so we need order to matter here too.)*

*So*

$$\mathbb{P}(\bar{A}) = \frac{40 \times 36 \times 32 \times 28 \times 24}{40 \times 39 \times 38 \times 37 \times 36} = 0.392.$$

*Thus*

$$\mathbb{P}(A) = \mathbb{P}(\text{at least 2 cards are the same design}) = 1 - \mathbb{P}(\bar{A}) = 1 - 0.392 = 0.608.$$

## Summary of discrete sample spaces

A discrete sample space can be written  $S = \{s_1, s_2, s_3, \dots\}$ .

$s_1, s_2, \dots$  are called sample points and they describe possible outcomes of the random experiment.

An event  $A$  is a subset of the sample space, i.e. a collection of sample points.

A probability distribution on  $S$  is a set of numbers  $\{p_1, p_2, \dots\}$  such that  $0 \leq p_i \leq 1$  for all  $i$ ,  $\sum_i p_i = 1$ , and each  $p_i$  corresponds to a sample point  $s_i$ . The probability of event  $A$  is  $\mathbb{P}(A) = \sum_{i: s_i \in A} p_i$ .

## Continuous sample spaces

For a continuous sample space  $S$ , we can usually think of both  $S$  and event  $A$  as *intervals on the real line*.

We can no longer sum the probabilities of all outcomes in event  $A$ , because we can't count how many outcomes there are. [E.g. how many numbers are there between 0 and 1? Impossible to say!]

However, if the outcome of the experiment is equally likely to be any point in the interval  $S$ , then we find the probability of event  $A$  as

$$\mathbb{P}(A) = \frac{\text{length of interval } A}{\text{length of interval } S}$$

When the outcome is **not** equally likely to lie anywhere in the interval, we need more advanced methods for determining probabilities. Much of this course is devoted to solving this problem for continuous and discrete sample spaces.

---

### 1.5 Probability Axioms: the three fundamental statements of probability

*Definition:* Let  $S$  be a sample space (continuous or discrete). Let  $\mathbb{P}$  be a function from the set of all events in  $S$  to the real numbers: that is, for every event  $A$ , there is a real number  $\mathbb{P}(A)$ . Then  $\mathbb{P}$  is called a **probability measure** if it satisfies the following axioms:

*Axiom (AI)*  $\mathbb{P}(S) = 1$ .

*Axiom (AII)*  $\mathbb{P}(A) \geq 0$  for all events  $A$ .

*Axiom (AIII)* If  $A_1, A_2, \dots, A_n$  are **mutually exclusive events**, then

$$\mathbb{P}(A_1 \cup A_2 \cup \dots \cup A_n) = \sum_{i=1}^n \mathbb{P}(A_i).$$

Axiom AIII also applies to infinite sequences of mutually exclusive events: if  $S$  is infinite, and  $A_1, A_2, \dots$ , is an infinite series of mutually exclusive events (i.e.  $A_i \cap A_j = \emptyset$  for any  $i \neq j$ ), then

$$\mathbb{P}(A_1 \cup A_2 \cup \dots) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

**Axioms** are statements that can be assumed without proof. Thus, if we are told that  $\mathbb{P}$  is a probability measure, then we can assume it satisfies Axioms AI to AIII. Furthermore, **all properties of**  $\mathbb{P}$  must be derivable using only the three axioms.

The number  $\mathbb{P}(A)$  is called the **probability of event A**.

Axiom AIII is widely used in probability calculations, so it is worth emphasizing; e.g. for the special case  $n = 2$  we have:

*if  $A \cap B = \emptyset$  then  $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$*

**Theorem 1.2:** The probability measure  $\mathbb{P}$  has the following properties.

- (i)  $\mathbb{P}(\emptyset) = 0$ .
- (ii)  $\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A)$  for any event  $A$ .
- (iii)  $\mathbb{P}(A) \leq 1$  for any event  $A$ .
- (iv)  $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap \overline{B})$  for any events  $A, B$ .
- (v)  $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$  for any events  $A, B$ .
- (vi) 

**The Partition Theorem**

If  $B_1, B_2, \dots, B_m$  form a partition of  $S$ , then for any event  $A$ ,

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

(This generalizes part (iv)).

**Proof:**

We must use only the Axioms, and Theorem 1.1.

i)  $A = A \cup \emptyset$ ; and  $A \cap \emptyset = \emptyset$  (mutually exclusive).

So  $\mathbb{P}(A) = \mathbb{P}(A \cup \emptyset) = \mathbb{P}(A) + \mathbb{P}(\emptyset)$  (Axiom AIII)

$\Rightarrow \mathbb{P}(\emptyset) = 0$ .

ii)  $S = A \cup \bar{A}$ ; and  $A \cap \bar{A} = \emptyset$  (mutually exclusive).

So  $\underbrace{1 = \mathbb{P}(S)}_{\text{Axiom AI}} = \mathbb{P}(A \cup \bar{A}) = \mathbb{P}(A) + \mathbb{P}(\bar{A})$ . (Axiom AIII)

iii)  $\mathbb{P}(A) = 1 - \mathbb{P}(\bar{A}) \leq 1$  because  $\mathbb{P}(\bar{A}) \geq 0$  (Axiom AII)

iv) Special case of (vi).

v)

$$\begin{aligned} A \cup B &= (A \cap S) \cup (B \cap S) \quad \text{Thm 1.1(iv)} \\ &= [A \cap (B \cup \bar{B})] \cup [B \cap (A \cup \bar{A})] \quad \text{Thm 1.1(ii)} \\ &= (A \cap B) \cup (A \cap \bar{B}) \cup (B \cap A) \cup (B \cap \bar{A}) \quad \text{Thm 1.1(v)} \\ &= (A \cap \bar{B}) \cup (\bar{A} \cap B) \cup (A \cap B). \end{aligned}$$

These 3 events are mutually exclusive:

eg.  $(A \cap \bar{B}) \cap (A \cap B) = A \cap (\bar{B} \cap B) = A \cap \emptyset = \emptyset$ , etc.

So,  $\mathbb{P}(A \cup B) = \mathbb{P}(A \cap \bar{B}) + \mathbb{P}(\bar{A} \cap B) + \mathbb{P}(A \cap B)$  (Axiom AIII)

$$= \left[ \underbrace{\mathbb{P}(A) - \mathbb{P}(A \cap B)}_{\text{from (iv)}} \right] + \left[ \underbrace{\mathbb{P}(B) - \mathbb{P}(A \cap B)}_{\text{from (iv)}} \right] + \mathbb{P}(A \cap B)$$

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

vi) Suppose  $B_1, \dots, B_m$  are a partition of  $S$ :

then  $B_i \cap B_j = \emptyset$  if  $i \neq j$ , and  $\bigcup_{i=1}^m B_i = S$ .

Thus,  $(A \cap B_i) \cap (A \cap B_j) = A \cap (B_i \cap B_j) = A \cap \emptyset = \emptyset$ , for  $i \neq j$ ,

ie.  $(A \cap B_1), \dots, (A \cap B_m)$  are mutually exclusive also.

$$\begin{aligned} \text{So, } \sum_{i=1}^m \mathbb{P}(A \cap B_i) &= \mathbb{P}\left(\bigcup_{i=1}^m (A \cap B_i)\right) && \text{(Axiom AIII)} \\ &= \mathbb{P}\left(A \cap \bigcup_{i=1}^m B_i\right) && \text{(Thm 1.1 (vii))} \\ &= \mathbb{P}(A \cap S) \\ &= \mathbb{P}(A). && \square \end{aligned}$$

In exercises, quote the Axioms and results from Theorems 1.1 and 1.2 without proof.

**Note:** Part (v) can be extended to three or more events: e.g. for any  $A$ ,  $B$ , and  $C$ ,

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

**Example:** In New Zealand, 52% of drivers are female. The probability of being male and having driven while intoxicated is 15%. In total, 23% of people have driven while intoxicated. 43% of drivers think that the risk of being caught when drink-driving is low. Overall, 50% of drivers have either driven while intoxicated, or believe that there is a low risk of being caught, or both.

*First formulate events:*

let  $F = \text{“female”}$      $M = \overline{F} = \text{“male”}$

let  $D = \text{“has driven while intoxicated”}$

let  $L = \text{“thinks risk of being caught is low”}$

Next write down all the information given:

$$\mathbb{P}(F) = 0.52$$

$$\mathbb{P}(L) = 0.43$$

$$\mathbb{P}(M \cap D) = 0.15$$

$$\mathbb{P}(D \cup L) = 0.50$$

$$\mathbb{P}(D) = 0.23$$

Find the probability that a New Zealand driver:

(a) is male  $\mathbb{P}(M) = \mathbb{P}(\overline{F}) = 1 - \mathbb{P}(F) = 1 - 0.52 = 0.48$ .

(b) is female and has driven while intoxicated

Want  $\mathbb{P}(F \cap D)$ .

$$\text{We know that } \mathbb{P}(F \cap D) + \mathbb{P}(\overline{F} \cap D) = \mathbb{P}(D)$$

$$\text{ie. } \mathbb{P}(F \cap D) + \mathbb{P}(M \cap D) = \mathbb{P}(D)$$

$$\mathbb{P}(F \cap D) + 0.15 = 0.23$$

$$\text{so } \mathbb{P}(F \cap D) = 0.08$$

(c) is male **and/or** has driven while intoxicated

$$\mathbb{P}(M \cup D) = \mathbb{P}(M) + \mathbb{P}(D) - \mathbb{P}(M \cap D)$$

$$= 0.48 + 0.23 - 0.15$$

$$= 0.56$$

(d) has driven while intoxicated, **and** believes that there is a low risk of being caught.

Want  $\mathbb{P}(D \cap L)$  :

$$\text{Now } \mathbb{P}(D \cup L) = \mathbb{P}(D) + \mathbb{P}(L) - \mathbb{P}(D \cap L)$$

$$0.5 = 0.23 + 0.43 - \mathbb{P}(D \cap L)$$

$$\mathbb{P}(D \cap L) = 0.23 + 0.43 - 0.50$$

$$= 0.16$$

(e) has driven while intoxicated, **and** believes that the risk of being caught is **not** low.

Want  $\mathbb{P}(D \cap \overline{L})$  :

$$\text{Now } \mathbb{P}(D \cap L) + \mathbb{P}(D \cap \overline{L}) = \mathbb{P}(D)$$

$$0.16 + \mathbb{P}(D \cap \overline{L}) = 0.23$$

$$\mathbb{P}(D \cap \overline{L}) = 0.07$$

## 1.6 Conditional Probability

Suppose  $A$  and  $B$  are two events on the same sample space. There will often be *dependence* between  $A$  and  $B$ : that is, if we know that  $B$  has occurred, this changes our knowledge of the chance that  $A$  will occur.

**Example:** Toss a die once.

Let event  $A =$  “get a 6”

Let event  $B =$  “get an even number”

If the die is fair, then  $\mathbb{P}(A) = \frac{1}{6}$  and  $\mathbb{P}(B) = \frac{1}{2}$

However, if we *know* that  $B$  has occurred, then there is *an increased* chance that  $A$  has occurred:

$$\mathbb{P}(A \text{ occurs given that } B \text{ has occurred}) = \frac{1}{3} \quad \left( \frac{\text{result 6}}{\text{result 2 or 4 or 6}} \right)$$

**Example:** Probabilities from tables of counts.

The following are the numbers of deaths from heart disease in NZ in 1996.

		Sex		Total
		Male	Female	
	< 45	79	13	92
Age	45 – 64	772	216	988
	65 – 74	1081	499	1580
	74+	1795	2176	3971
	Total	3727	2904	6631

Let event  $A =$  “victim is female”

Let event  $B =$  “victim is <45”



Suppose we choose a person at random from those in the table.

$$\mathbb{P}(A) = \mathbb{P}(\textit{female}) = \frac{\# \textit{female victims}}{\textit{total \# victims}} = \frac{2904}{6631} = 0.44$$

But, if we choose people only from those under 45 years old, then:

$\mathbb{P}(\text{victim is female, given that victim is } < 45)$

$$= \frac{\# \textit{female victims} < 45}{\textit{total \# victims} < 45} = \frac{13}{92} = 0.14.$$

So  $\mathbb{P}(A \textit{ happens, given that } B \textit{ has happened}) = 0.14.$

We write  $\mathbb{P}(A | B) = 0.14.$  We have *conditioned* on event  $B.$

Conditioning on event  $B$  means *restricting attention* to the set for which  $B$  is true.

Think of  $\mathbb{P}(A | B)$  as the chance of getting an  $A$ , from the set of  $B$ 's only.

From above,

$$\begin{aligned} \mathbb{P}(A|B) &= \left( \frac{\text{number of the outcomes in } B \text{ that are also in } A}{\text{total number of outcomes in } B} \right) \\ &= \left( \frac{\# \text{ of outcomes in } A \text{ and } B}{\# \text{ of outcomes in } B} \right) \\ &= \frac{(\# \text{ of outcomes in } A \text{ and } B) / (\# \text{ of outcomes in } S)}{(\# \text{ of outcomes in } B) / (\# \text{ of outcomes in } S)} \\ &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \end{aligned}$$

This is our definition of conditional probability:

*Definition:* Let  $A$  and  $B$  be two events. The conditional probability that event  $A$  occurs, given that event  $B$  has occurred, is written  $\mathbb{P}(A|B)$ .

and is given by

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

Read  $\mathbb{P}(A|B)$  as “probability of  $A$ , given  $B$ ”.

*Note:*  $\mathbb{P}(A|B)$  gives  $\mathbb{P}(A$  and  $B$  , from within the set of  $B$ 's only)  
 $\mathbb{P}(A \cap B)$  gives  $\mathbb{P}(A$  and  $B$  , from the whole sample space).

## Multiplication Rule

For any events  $A$  and  $B$ ,

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

## Proof:

*Immediate from the definitions:*

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

and

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

The Multiplication Rule gives us a new statement of the Partition Theorem:  
 If  $B_1, \dots, B_m$  partition  $S$ , then for any event  $A$ ,

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

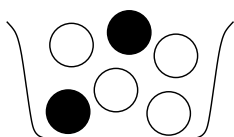
Both formulations of the Partition Theorem are very widely used, but especially the conditional formulation  $\sum_{i=1}^m \mathbb{P}(A|B_i)\mathbb{P}(B_i)$ .

**Example:** Two balls are drawn at random without replacement from a box containing 4 white and 2 red balls.

Find the probability that

- (i) they are both white,
- (ii) the second ball is red.

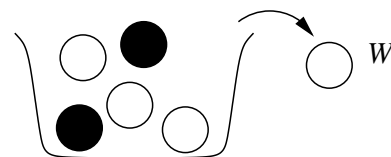
Solution



Let event  $W_i =$  “ $i$ th ball is white” and  $R_i =$  “ $i$ th ball is red”.

i)  $\mathbb{P}(W_1 \cap W_2) = \mathbb{P}(W_2 \cap W_1) = \mathbb{P}(W_2|W_1)\mathbb{P}(W_1)$

Now  $\mathbb{P}(W_1) = \frac{4}{6}$  and  $\mathbb{P}(W_2|W_1) = \frac{3}{5}$ .

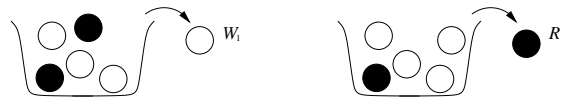


So  $\mathbb{P}(\text{both white}) = \mathbb{P}(W_1 \cap W_2) = \frac{3}{5} \times \frac{4}{6} = \frac{2}{5}$ .

ii) Looking for  $\mathbb{P}(\text{2nd ball is red})$ . We can't find this without conditioning on what happened in the first draw.

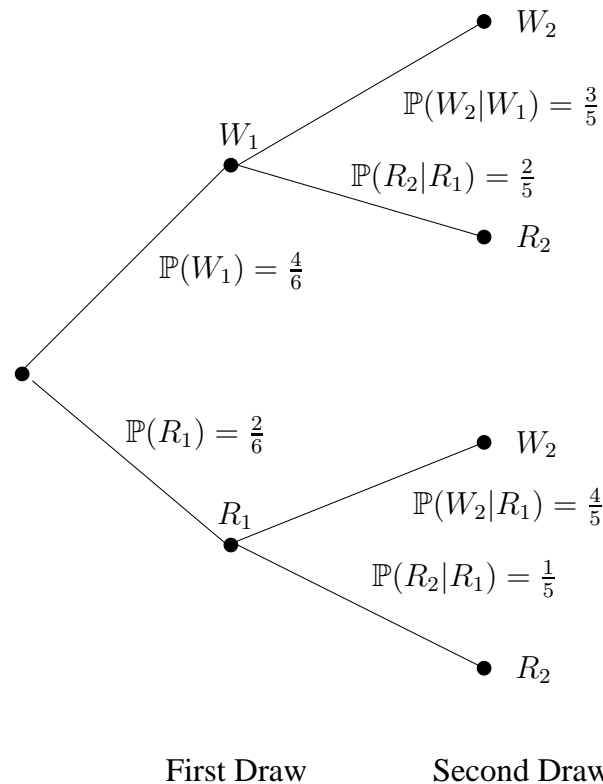
Event "2nd ball is red" is actually event  $\{W_1R_2, R_1R_2\} = (W_1 \cap R_2) \cup (R_1 \cap R_2)$ .

$$\begin{aligned} \text{So } \mathbb{P}(\text{2nd ball is red}) &= \mathbb{P}(W_1 \cap R_2) + \mathbb{P}(R_1 \cap R_2) \quad (\text{mutually exclusive}) \\ &= \mathbb{P}(R_2|W_1)\mathbb{P}(W_1) + \mathbb{P}(R_2|R_1)\mathbb{P}(R_1) \\ &= \frac{2}{5} \times \frac{4}{6} + \frac{1}{5} \times \frac{2}{6} \end{aligned}$$



$$= \frac{1}{3}$$

**Note:** Probability trees are often useful when *events happen in sequence*.



Write conditional probabilities on the branches, and multiply to get probability of an intersection: eg.  $\mathbb{P}(W_1 \cap W_2) = \frac{4}{6} \times \frac{3}{5}$ , or  $\mathbb{P}(R_1 \cap W_2) = \frac{2}{6} \times \frac{4}{5}$ .

# Two separate studies say ...



**So you're better off with AntiCough  
... or are you???**

**Have a look at the figures:**

**Study 1**

	AntiCough	Other Medicine
Given to:	40	80
Cured:	34	64
%Cured:	85% 😊	80% ☹️

**Study 2**

	AntiCough	Other Medicine
Given to:	60	20
Cured:	39	12
%Cured:	65% 😊	60% ☹️

**Combine the studies ... What happens?**

Never believe what you read... This is Simpson's Paradox... Never believe what you read... This is Sim

## Simpson's paradox

## 1.7 Bayes' Theorem: inverting conditional probabilities

---

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

Apply the multiplication rule to each side:

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

Thus, 
$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} \quad \circledast$$

This is the simplest form of Bayes' Theorem, named after Thomas Bayes (c. 1700), English clergyman and founder of Bayesian Statistics.

Bayes' Theorem allows us to “invert” the conditioning, ie. to express  $\mathbb{P}(B|A)$  in terms of  $\mathbb{P}(A|B)$ .

This is very useful. For example, *it might be easy to calculate,*

$$\mathbb{P}(\text{later event}|\text{earlier event}),$$

*but we might only observe the later event and wish to deduce the probability that the earlier event occurred,*

$$\mathbb{P}(\text{earlier event}|\text{later event}).$$

Full statement of Bayes' Theorem:

**Theorem 1.3:** Let  $B_1, B_2, \dots, B_m$  form a partition of  $S$ . Then for any event  $A$ , and for any  $j = 1, \dots, m$ ,

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\sum_{i=1}^m \mathbb{P}(A|B_i)\mathbb{P}(B_i)} \quad (\text{Bayes' Theorem})$$

Proof:

Immediate from  $\circledast$  (put  $B = B_j$ ), and the Partition Rule which gives  $\mathbb{P}(A) = \sum_{i=1}^m \mathbb{P}(A|B_i)\mathbb{P}(B_i)$ .  $\square$

Special case of Bayes' Theorem when  $m = 2$ : use  $B$  and  $\overline{B}$  as the partition of  $S$ :

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

**Example:** The case of the Perfidious Gardener.

Mr Smith owns a *hysterical* rosebush. It will die with probability  $1/2$  if watered, and with probability  $3/4$  if not watered. Worse still, Smith employs a perfidious gardener who will fail to water the rosebush with probability  $2/3$ .

Smith returns from holiday to find the rosebush ... **DEAD!!!**  
What is the probability that the gardener did not water it?

**Solution:**

First step: formulate events

Let :  $D$  = "rosebush dies"

$W$  = "gardener waters rosebush"

$\overline{W}$  = "gardener fails to water rosebush"

Second step: write down all information given

$$\mathbb{P}(D|W) = \frac{1}{2} \quad \mathbb{P}(D|\overline{W}) = \frac{3}{4} \quad \mathbb{P}(\overline{W}) = \frac{2}{3} \quad (\text{so } \mathbb{P}(W) = \frac{1}{3})$$

Third step: write down what we're looking for

$$\mathbb{P}(\overline{W}|D)$$

Fourth step: compare this to what we know

Need to invert the conditioning, so use Bayes' Theorem:

$$\mathbb{P}(\overline{W}|D) = \frac{\mathbb{P}(D|\overline{W})\mathbb{P}(\overline{W})}{\mathbb{P}(D|\overline{W})\mathbb{P}(\overline{W}) + \mathbb{P}(D|W)\mathbb{P}(W)} = \frac{3/4 \times 2/3}{3/4 \times 2/3 + 1/2 \times 1/3} = \frac{3}{4}$$

So the gardener failed to water the rosebush with probability  $\frac{3}{4}$ .



**Example:** The case of the Defective Ketchup Bottle.

Ketchup bottles are produced in 3 different factories, accounting for 50%, 30%, and 20% of the total output respectively. The percentage of defective bottles from the 3 factories is respectively 0.4%, 0.6%, and 1.2%. A statistics lecturer who eats only ketchup finds a defective bottle in her door. What is the probability that it came from Factory 1?

**Solution:**

1. *Events:*

let  $F_i$  = “bottle comes from Factory  $i$ ” ( $i=1,2,3$ )

let  $D$  = “bottle is defective”

2. *Information given:*

$$\begin{aligned} \mathbb{P}(F_1) &= 0.5 & \mathbb{P}(F_2) &= 0.3 & \mathbb{P}(F_3) &= 0.2 \\ \mathbb{P}(D|F_1) &= 0.004 & \mathbb{P}(D|F_2) &= 0.006 & \mathbb{P}(D|F_3) &= 0.012 \end{aligned}$$

3. *Looking for:*

$$\mathbb{P}(F_1|D) \quad (\text{so need to invert conditioning}).$$

4. *Bayes Theorem:*

$$\begin{aligned} \mathbb{P}(F_1|D) &= \frac{\mathbb{P}(D|F_1)\mathbb{P}(F_1)}{\mathbb{P}(D|F_1)\mathbb{P}(F_1) + \mathbb{P}(D|F_2)\mathbb{P}(F_2) + \mathbb{P}(D|F_3)\mathbb{P}(F_3)} \\ &= \frac{0.004 \times 0.5}{0.004 \times 0.5 + 0.006 \times 0.3 + 0.012 \times 0.2} \\ &= \frac{0.002}{0.0062} \\ &= 0.322 \end{aligned}$$

## Chains of Events

To find  $\mathbb{P}(A_1 \cap A_2 \cap A_3)$  we can apply the *multiplication rule successively*:

$$\begin{aligned}\mathbb{P}(A_1 \cap A_2 \cap A_3) &= \mathbb{P}(A_3 \cap (A_1 \cap A_2)) \\ &= \mathbb{P}(A_3|A_1 \cap A_2)\mathbb{P}(A_1 \cap A_2) \quad (\text{multiplication rule}) \\ &= \mathbb{P}(A_3|A_1 \cap A_2)\mathbb{P}(A_2|A_1)\mathbb{P}(A_1) \quad (\text{multiplication rule})\end{aligned}$$

Remember as:

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

In general, for  $n$  events  $A_1, A_2, \dots, A_n$ , we have,

$$\mathbb{P}(A_1 \cap A_2 \cap \dots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2 | A_1)\mathbb{P}(A_3 | A_2 \cap A_1) \dots \mathbb{P}(A_n | A_{n-1} \cap \dots \cap A_1)$$

**Example:** A box contains  $w$  white balls and  $r$  red balls. Draw 3 balls without replacement. What is the probability of getting the sequence white, red, white?

Answer:

$$\begin{aligned}\mathbb{P}(W_1 \cap R_2 \cap W_3) &= \mathbb{P}(W_1)\mathbb{P}(R_2|W_1)\mathbb{P}(W_3|R_2 \cap W_1) \\ &= \left(\frac{w}{w+r}\right) \times \left(\frac{r}{w+r-1}\right) \times \left(\frac{w-1}{w+r-2}\right).\end{aligned}$$

---

## 1.8 Statistical Independence

*Two events  $A$  and  $B$  are statistically independent if the occurrence of one does not affect the occurrence of the other.*

*This means  $\mathbb{P}(A|B) = \mathbb{P}(A)$  and  $\mathbb{P}(B|A) = \mathbb{P}(B)$ .*

*Now  $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ ,*

*so if  $\mathbb{P}(A|B) = \mathbb{P}(A)$  then  $\mathbb{P}(A \cap B) = \mathbb{P}(A) \times \mathbb{P}(B)$ .*

We use this as our definition of statistical independence.

*Definition:* Events  $A$  and  $B$  are **statistically independent** if

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

For more than two events, we say:

*Definition:* Events  $A_1, A_2, \dots, A_n$  are **mutually independent** if

$$\mathbb{P}(A_1 \cap A_2 \cap \dots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2) \dots \mathbb{P}(A_n), \text{ AND}$$

*the same multiplication rule holds for every subcollection of the events too.*

*Eg. events  $A_1, A_2, A_3, A_4$  are mutually independent if*

*i)  $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$  for all  $i, j$  with  $i \neq j$ ;  
AND*

*ii)  $\mathbb{P}(A_i \cap A_j \cap A_k) = \mathbb{P}(A_i)\mathbb{P}(A_j)\mathbb{P}(A_k)$  for all  $i, j, k$  that are all different;  
AND*

*iii)  $\mathbb{P}(A_1 \cap A_2 \cap A_3 \cap A_4) = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)\mathbb{P}(A_4)$ .*

**Notes:** 1) If events are *physically* independent, *they will also be statistically independent*.

2) If  $A$  and  $B$  are mutually exclusive, *they are not usually independent*.

“Mutually exclusive” means  $\mathbb{P}(A \cap B) = 0$

“Independent” means  $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ .

**Example:** Toss a fair coin and a fair die together. The coin and die are physically independent.

**Sample space:**  $S = \{H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6\}$   
- all 12 items are equally likely.

Let  $A =$  “heads” and  $B =$  “six”.

Then  $\mathbb{P}(A) = \mathbb{P}(\{H1, H2, H3, H4, H5, H6\}) = \frac{6}{12} = \frac{1}{2}$

$\mathbb{P}(B) = \mathbb{P}(\{H6, T6\}) = \frac{2}{12} = \frac{1}{6}$

Now  $\mathbb{P}(A \cap B) = \mathbb{P}(\text{Heads and } 6) = \mathbb{P}(\{H6\}) = \frac{1}{12}$

But  $\mathbb{P}(A) \times \mathbb{P}(B) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$  also,

So  $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$  and thus  $A$  and  $B$  are statistically independent.

**Example:** A jar contains 4 balls: one red, one white, one blue, and one red, white & blue. Draw one ball at random.

Let

$A =$  “ball has red on it”,  $B =$  “ball has white on it”,  $C =$  “ball has blue on it”.

2 balls satisfy  $A$ , so  $\mathbb{P}(A) = \frac{2}{4} = \frac{1}{2}$ . Likewise,  $\mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2}$ .

Now,  $\mathbb{P}(A \cap B) = \frac{1}{4}$  (one of 4 balls has both red and white on it).

But,  $\mathbb{P}(A) \times \mathbb{P}(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ , so  $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ .

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

So  $A$ ,  $B$  and  $C$  are pairwise independent.

BUT,  $\mathbb{P}(A \cap B \cap C) = \frac{1}{4}$  (one of 4 balls)

while  $\mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} \neq \mathbb{P}(A \cap B \cap C)$ .

So A, B and C are NOT mutually independent, despite being pairwise independent.

**Notes:** 1) If A and B are independent, then

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

Similarly, if A, B, and C are *mutually independent*, then

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

**When independent,**

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

2) If A and B are independent, so are:

(i) A and  $\bar{B}$       (ii)  $\bar{A}$  and B      (iii)  $\bar{A}$  and  $\bar{B}$ .

**Proof of (i):**

$$\begin{aligned}\mathbb{P}(A \cap \bar{B}) &= \mathbb{P}(A) - \mathbb{P}(A \cap B) \text{ (because } B, \bar{B} \text{ partition } S) \\ &= \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) \text{ if } A, B \text{ independent} \\ &= \mathbb{P}(A)(1 - \mathbb{P}(B)) \\ &= \mathbb{P}(A)\mathbb{P}(\bar{B}) \Rightarrow A, \bar{B} \text{ are independent.}\end{aligned}$$

(ii), (iii) exercise.

## 1.9 Random Variables

*Definition:* A random variable (r.v.) is a function from a sample space  $S$  to the real numbers  $\mathbb{R}$ .

We write  $X : S \rightarrow \mathbb{R}$ .

*Example:* Toss a coin 3 times. The sample space is

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

One example of a random variable is  $X : S \rightarrow \mathbb{R}$  such that, for sample point  $s_i$ , we have  $X(s_i) = \# \text{ heads in outcome } s_i$ .

So  $X(HHH) = 3$ ,  $X(THT) = 1$ , etc.

Another example is  $Y : S \rightarrow \mathbb{R}$  such that  $Y(s_i) = \begin{cases} 1 & \text{if 2nd toss is a head} \\ 0 & \text{otherwise} \end{cases}$

Then  $Y(HTH) = 0$ ,  $Y(THH) = 1$ ,  $Y(HHH) = 1$ , etc.

Another example is  $W : S \rightarrow \mathbb{R}$  such that  $W(s_i) = \text{cosine}(\# \text{ tails in } s_i)$ .

Any function is a random variable as long as it is defined on *all* elements of  $S$ , and takes only *real values*.

**Note:** The name ‘random variable’ is misleading, because we are looking at a function on the sample space. This is neither random nor variable.

However, if we observe the outcome of a random experiment, and apply a random variable (i.e. a real-valued function) to it, then we end up with what is essentially a random real number. This helps to explain where the name comes from.

For example, suppose we toss a coin 3 times and observe the outcome. Apply  $X : S \rightarrow \mathbb{R}$ , such that  $X(s_i) = \# \text{ heads in outcome } s_i$ .

The first time we do this, we get outcome *THH* (say)  
so  $X(\text{THH})=2$ .

The second time, we get outcome *HTT*  
so  $X(\text{HTT})=1$ .

and so on.

Thus the random variable produces *random real numbers* as the ‘outcome’ of a random experiment.

### Why do we use random variables?

A random variable allots a number to every outcome in the sample space. This means that totally different sample spaces can be represented on the same numerical scale. Using random variables gives us a way of *describing many different situations at once*.

For example:

*Expt 1:* Let  $X = \#$  heads from 4 tosses of a fair coin.

*Expt 2:* Let  $Y = \#$  boys in a 4-child family.

$X$  and  $Y$  have exactly the same behaviour as random variables, despite being defined upon different sample spaces.

Random variables are the fundamental concept that we need in order to build mathematical models of randomness in the real world.

## Probabilities for random variables

By convention, we use **CAPITAL LETTERS** for random variables (e.g.  $X$ ), and **lower case letters** to represent the values that the random variable takes (e.g.  $x$ ).

For a sample space  $S$  and random variable  $X : S \rightarrow \mathbb{R}$ , and for a real number  $x$ ,

$$\mathbb{P}(X = x) = \mathbb{P}(\text{outcome } s \text{ is such that } X(s) = x) = \mathbb{P}(\{s : X(s) = x\}).$$

**Example:** toss a fair coin 3 times. All outcomes are equally likely:

$$\mathbb{P}(\text{HHH}) = \mathbb{P}(\text{HHT}) = \dots = \mathbb{P}(\text{TTT}) = 1/8.$$

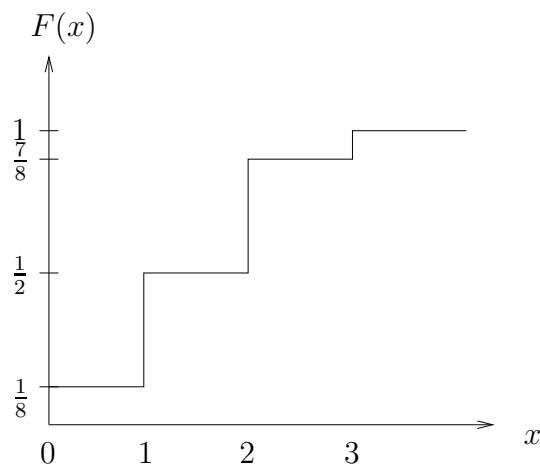
Let  $X : S \rightarrow \mathbb{R}$ , such that  $X(s) = \# \text{ heads in } s$ .

$$\begin{aligned} \text{Then} \quad \mathbb{P}(X = 0) &= \mathbb{P}(\{TTT\}) = 1/8 \\ \mathbb{P}(X = 1) &= \mathbb{P}(\{HTT, THT, TTH\}) = 3/8 \\ \mathbb{P}(X = 2) &= \mathbb{P}(\{HHT, HTH, THH\}) = 3/8 \\ \mathbb{P}(X = 3) &= \mathbb{P}(\{HHH\}) = 1/8 \end{aligned}$$

**Note that**  $\mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \mathbb{P}(X = 2) + \mathbb{P}(X = 3) = 1$ .

**Definition:** The **cumulative distribution function (c.d.f.)** of a r.v.  $X$  is given by  $F_X(x) = \mathbb{P}(X \leq x)$ .

**Example:**  $X(s) = \# \text{heads in } s$ , as above.



$$\begin{aligned} F_X(0) &= \mathbb{P}(X \leq 0) = \frac{1}{8} \\ F_X(1) &= \mathbb{P}(X \leq 1) = \frac{1}{8} + \frac{3}{8} \\ F_X(2) &= \frac{1}{2} + \frac{3}{8} = \frac{7}{8} \\ F_X(3) &= \frac{7}{8} + \frac{1}{8} = 1. \end{aligned}$$

**Questions:** what is  $F_X(-1)$ ? **Ans: 0;**

$F_X(0.5)$ ? **Ans: 1/8;**

$F_X(4)$ ? **Ans: 1.**



## 1.10 Problems

1. In New Zealand, more males are born than females: 51 percent of all babies born are male. However, infant mortality is higher for males than for females: 55 percent of all infant deaths are male. The infant mortality rate is 6.28 per thousand live births.

a) Define events  $M$ ,  $F$ , and  $D$  for ‘baby is male’, ‘baby is female’, and ‘baby dies in infancy’. Express all information given above in terms of these events.

b) What is the probability that a male baby dies in infancy?

c) What is the probability that a female baby dies in infancy?

d) We are interested to know whether the higher death rate for males in their first year balances out the higher birth rate for males: that is, what is the proportion of males among babies that survive infancy? Express this probability in terms of the events  $M$  and  $D$ , and hence find the required probability.

Has the sex ratio been balanced by the higher death rate? Are you surprised?

2. The following figures come from a Western Australian study of hypertension and its connections with weight and alcohol consumption. Alcohol consumption was classed as Low ( $L$ ), Medium ( $M$ ), or High ( $H$ ). A subject’s weight was classed as Average ( $A$ ), or Overweight ( $O$ ). Each subject was diagnosed as suffering from hypertension ( $T$ ), or not suffering from hypertension ( $\bar{T}$ ).

The proportions in the sample are as follows:

i) average weight and with low, medium, and high alcohol consumptions respectively: 0.17, 0.33, 0.16. The probability of hypertension for these three categories is 0.13, 0.23, and 0.31 respectively.

ii) overweight and with low, medium, and high alcohol consumptions respectively: 0.07, 0.17, 0.10. The probability of suffering hypertension for these three categories is 0.27, 0.37, and 0.40 respectively.

a) Express all information given above in terms of the events  $T$ ,  $L$ ,  $M$ ,  $H$ ,  $A$ , and  $O$ .

b) Do the events  $L$ ,  $M$ , and  $H$  form a partition of the sample space? Explain why or why not.

c) Do the events  $A$  and  $O$  form a partition of the sample space? Explain why or why not.

d) Do the events  $A \cap L$ ,  $A \cap M$ , and  $A \cap H$  form a partition of the sample space? Explain why or why not.

e) Find  $\mathbb{P}(A)$  and  $\mathbb{P}(O)$ .

f) Find  $\mathbb{P}(T)$ .

g) Find  $\mathbb{P}(T | L)$  and  $\mathbb{P}(T | H)$ .

h) Find  $\mathbb{P}(T \cap O)$ ,  $\mathbb{P}(T | O)$ , and  $\mathbb{P}(T \cup O)$  for the sample in the study. Describe these events in words.

3. The following probabilities are obtained from weather data for Auckland in February 2002. The weather for each day can be classified as ‘rain’ ( $R$ ), or ‘dry’ ( $D$ ). Given the weather conditions for any specified day, the conditions on the next day are the same with probability  $4/5$ , and different with probability  $1/5$ ; and they do not depend upon the conditions on any previous days. Suppose that on day 1 there is a 10% chance of rain.

a) Formulate events  $R_n$  for ‘rain on day  $n$ ’, and  $D_n$  for ‘dry on day  $n$ ’. Using the information above, state  $\mathbb{P}(R_n | R_{n-1})$ ,  $\mathbb{P}(R_n | D_{n-1})$ ,  $\mathbb{P}(D_n | R_{n-1})$ , and  $\mathbb{P}(D_n | D_{n-1})$ , for  $n > 1$ .

b) Find  $\mathbb{P}(R_1 \cap D_2 \cap R_3 \cap D_4)$ .

c) Let  $r_n = \mathbb{P}(R_n)$ . Find  $r_2$  and  $r_3$ .

d) Show that  $r_n = \frac{3}{5}r_{n-1} + \frac{1}{5}$ .

e) By repeated substitution, show that

$$r_n = \left(\frac{3}{5}\right)^{n-1} \left(r_1 - \frac{1}{2}\right) + \frac{1}{2} = -\left(\frac{2}{5}\right) \left(\frac{3}{5}\right)^{n-1} + \frac{1}{2}.$$

f) What is the probability that it will rain on the last day of February (day 28)? To what extent does this depend upon the conditions on February 1st?

4. An elderly possum retires to live in Albert Park. Having no natural predators, the only threat to its survival comes every day at sunset when the lights to the Sky Tower are turned on. It will survive this experience with probability  $p$ , but with probability  $1 - p$  it will die of shock. This situation continues indefinitely: given that the possum is alive at the start of a day, it survives to the end of the day with probability  $p$  and dies with probability  $1 - p$ .

Assume that the possum is alive at the start of day 1.

a) Let event  $A_2$  be the event that the possum is alive at the end of day 2. Find  $\mathbb{P}(A_2)$ .

b) Let event  $D_4$  be the event that the possum has died by the end of day 4 (i.e. on or before day 4). Find  $\mathbb{P}(D_4)$ .

c) Find  $\mathbb{P}(A_2 | D_4)$ . [Hint: you might find it useful to draw a probability tree.]

## 1.11 Key Probability Results for Chapter 1

---

1. If  $A$  and  $B$  are **mutually exclusive** (i.e.  $A \cap B = \emptyset$ ), then

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

2. Conditional probability:  $\mathbb{P}(A | B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$  for any  $A, B$ .

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

3. For any  $A, B$ , we can write

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

This is a simplified version of Bayes' Theorem. It shows how to 'invert' the conditioning, i.e. how to find  $\mathbb{P}(A | B)$  when you know  $\mathbb{P}(B | A)$ .

4. Bayes' Theorem slightly more generalized:

for any  $A, B$ ,

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

This works because  $A$  and  $\bar{A}$  form a **partition** of the sample space.

5. Complete version of Bayes' Theorem:

If sets  $A_1, \dots, A_m$  form a **partition** of the sample space, i.e. they **do not overlap** (mutually exclusive) and **collectively cover all possible outcomes** (their union is the sample space), then

$$\begin{aligned}\mathbb{P}(A_j | B) &= \frac{\mathbb{P}(B | A_j)\mathbb{P}(A_j)}{\mathbb{P}(B | A_1)\mathbb{P}(A_1) + \dots + \mathbb{P}(B | A_m)\mathbb{P}(A_m)} \\ &= \frac{\mathbb{P}(B | A_j)\mathbb{P}(A_j)}{\sum_{i=1}^m \mathbb{P}(B | A_i)\mathbb{P}(A_i)}.\end{aligned}$$

6. Partition Theorem: if  $A_1, \dots, A_m$  form a **partition** of the sample space, then

$$\mathbb{P}(B) = \mathbb{P}(B \cap A_1) + \mathbb{P}(B \cap A_2) + \dots + \mathbb{P}(B \cap A_m).$$

This can also be written as:

$$\mathbb{P}(B) = \mathbb{P}(B | A_1)\mathbb{P}(A_1) + \mathbb{P}(B | A_2)\mathbb{P}(A_2) + \dots + \mathbb{P}(B | A_m)\mathbb{P}(A_m).$$

These are both very useful formulations.

7. Chains of events:

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

8. Statistical independence:

if  $A$  and  $B$  are **independent**, then

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

and

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

and

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

9. Conditional probability measure:

If  $\mathbb{P}(B) > 0$ , then we can treat  $\mathbb{P}(\cdot | B)$  just like any other probability measure:

e.g. if  $A_1$  and  $A_2$  are mutually exclusive, then  $\mathbb{P}(A_1 \cup A_2 | B) = \mathbb{P}(A_1 | B) + \mathbb{P}(A_2 | B)$  (compare with  $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2)$ );

if  $A_1, \dots, A_m$  partition the sample space, then  $\mathbb{P}(A_1 | B) + \mathbb{P}(A_2 | B) + \dots + \mathbb{P}(A_m | B) = 1$ ; and  $\mathbb{P}(A | B) = 1 - \mathbb{P}(\bar{A} | B)$  for any  $A$ .

(Note: it is **not** generally true that  $\mathbb{P}(A | B) = 1 - \mathbb{P}(A | \bar{B})$ .)

The fact that  $\mathbb{P}(\cdot | B)$  is a probability measure is easily verified by checking that it satisfies the Axioms AI, AII, and AIII.

10. Unions: For **any**  $A, B, C$ ,

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

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

The second expression is obtained by writing  $\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A \cup (B \cup C))$  and applying the first expression to  $A$  and  $(B \cup C)$ , then applying it again to expand  $\mathbb{P}(B \cup C)$ .

# Chapter 2 : Discrete Probability

---

## Distributions

---

### 2.1 Introduction

Recall that a random variable,  $X$ , assigns a real number to every possible outcome of a random experiment. The random variable is discrete if the set of real values it can take is finite or countable, eg.  $\{0,1,2,\dots\}$ .

*Definition:* The probability function,  $f_X(x)$ , for a discrete random variable  $X$ , is given by,

$$\boxed{f_X(x) = \mathbb{P}(X = x),} \quad \text{for all possible outcomes } x \text{ of } X.$$

*Example:* Toss a fair coin once, and let  $X$ =number of heads. Then

$$X = \begin{cases} 0 & \text{with probability } 0.5, \\ 1 & \text{with probability } 0.5. \end{cases}$$

The probability function of  $X$  is given by

$$f_X(x) = \begin{cases} 0.5 & \text{if } x=0 \\ 0.5 & \text{if } x=1 \\ 0 & \text{otherwise} \end{cases}$$

We write (eg.)  $f_X(0) = 0.5$ ,  $f_X(1) = 0.5$ ,  $f_X(7.5) = 0$ , etc.

## Properties of the probability function

i)  $f_X(x) \geq 0$  for all  $x$ ; (probabilities are never negative)

ii)  $\sum_x f_X(x) = 1$ ; (probabilities add to 1 overall)

iii)  $P(X \in A) = \sum_{x \in A} f_X(x)$ ;

eg. if  $X$ =value from one toss of a fair die, then

$$\mathbb{P}(X \in \{1, 2, 3\}) = \mathbb{P}(X = 1) + \mathbb{P}(X = 2) + \mathbb{P}(X = 3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

*Definition:* The expected value, or mean, of a discrete random variable  $X$ , can be written as either  $\mathbb{E}(X)$ , or  $E(X)$ , or  $\mu_X$ , and is given by

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

The expected value is a measure of the centre, or average, of the set of values that  $X$  can take, weighted according to the probability of each value.

*Example:* suppose  $X = \begin{cases} 1 & \text{with probability } 0.9, \\ -1 & \text{with probability } 0.1. \end{cases}$

$X$  takes only the values 1 and  $-1$ . What is the ‘average’ value of  $X$ ?

Using  $\frac{1+(-1)}{2} = 0$  would not be useful, because it ignores the fact that usually  $X = 1$ , and only occasionally is  $X = -1$ .

Instead, think of observing  $X$  many times, say 100 times.

Roughly 90 of these 100 times will have  $X = 1$ .

Roughly 10 of these 100 times will have  $X = -1$

*Take the average of the 100 values: it will be roughly*

$$\frac{90 \times 1 + 10 \times (-1)}{100},$$

*ie.*  $0.9 \times 1 + 0.1 \times (-1) = 0.8$ .

*This is why we take the average as*

$$\mathbb{E}(X) = f_X(1) \times 1 + f_X(-1) \times (-1).$$

*$\mathbb{E}(X)$  is the average (mean) value we would get if we observed  $X$  many times.*

### Expected value of a function of $X$

Let  $X$  be a random variable, and let  $g$  be a (nice) function from  $\mathbb{R} \rightarrow \mathbb{R}$ .

*Then  $g(X)$  is also a random variable.*

*Example:*

$$X = \begin{cases} 3 & \text{with probability } 0.75, \\ 8 & \text{with probability } 0.25. \end{cases}$$

*Let  $g : \mathbb{R} \rightarrow \mathbb{R}$  such that  $g(x) = \sqrt{x}$ .*

*Then*

$$g(X) = \begin{cases} \sqrt{3} & \text{with probability } 0.75, \\ \sqrt{8} & \text{with probability } 0.25. \end{cases}$$

*So the average of  $g(X)$  is:  $\frac{0.75 \times \sqrt{3} + 0.25 \times \sqrt{8}}{1}$ .*

*Definition:* For any function  $g$ , the expected value of  $g(X)$  is given by

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

**Theorem 2.1:** Let  $a$  and  $b$  be constants, and let  $g(x)$ ,  $h(x)$  be functions. Then

i)  $\mathbb{E}[aX + b] = a\mathbb{E}(X) + b$

ii)  $\mathbb{E}[ag(X) + b] = a\mathbb{E}[g(X)] + b$

iii)  $\mathbb{E}[ag(X) + bh(X)] = a\mathbb{E}[g(X)] + b\mathbb{E}[h(X)]$

**Proof:**

Direct from definition of expectation of a function.

*Eg. for (iii),*

$$\begin{aligned}\mathbb{E}[ag(X) + bh(X)] &= \sum_x [ag(x) + bh(x)]f_X(x) \\ &= a \sum_x g(x)f_X(x) + b \sum_x h(x)f_X(x) \\ &= a \mathbb{E}[g(X)] + b \mathbb{E}[h(X)]. \quad \square\end{aligned}$$

**Note:** Part (iii) is related to the important result

$$\mathbb{E}(X_1 + X_2 + \dots + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \dots + \mathbb{E}(X_n), \quad \text{for any } X_1, \dots, X_n.$$



## Variance

*Definition:* The **variance** of a random variable  $X$  is written as either  $\text{Var}(X)$  or  $\sigma_X^2$ , and is given by

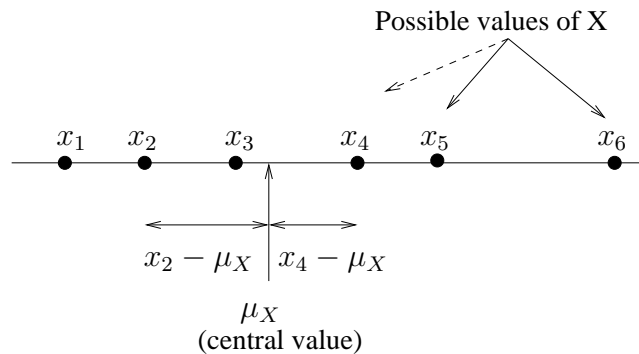
$$\sigma_X^2 = \text{Var}(X) = \mathbb{E} [(X - \mu_X)^2] = \mathbb{E} [(X - \mathbb{E}X)^2].$$

Similarly, the variance of a function of  $X$  is  $\text{Var}(g(X)) = \mathbb{E} \left[ \left( g(X) - \mathbb{E}(g(X)) \right)^2 \right]$ .

**Note:** The variance is the square of the standard deviation of  $X$ , so

$$\text{sd}(X) = \sqrt{\text{Var}(X)} = \sqrt{\sigma_X^2} = \sigma_X.$$

The variance is a measure of how *spread out* are the values that  $X$  can take. It is the *average squared distance between a value of  $X$  and the central (mean) value,  $\mu_X$* .



$$\text{Var}(X) = \underbrace{\mathbb{E}}_{(2)} \left[ \underbrace{(X - \mu_X)^2}_{(1)} \right]$$

- (1) Take distance from observed values of  $X$  to the central point,  $\mu_X$ . Square it to balance positive and negative distances.
- (2) Then take the average over all values  $X$  can take: ie. if we observed  $X$  many times, find what would be the average squared distance between  $X$  and  $\mu_X$ .

**Note:** The mean,  $\mu_X$ , and the variance,  $\sigma_X^2$ , of  $X$  are just *numbers*: there is nothing random or variable about them.

**Example:** Let  $X = \begin{cases} 3 & \text{with probability } 3/4, \\ 8 & \text{with probability } 1/4. \end{cases}$

Then

$$\begin{aligned} \mathbb{E}(X) &= \mu_X = 3 \times \frac{3}{4} + 8 \times \frac{1}{4} = 4.25 \\ \text{Var}(X) &= \sigma_X^2 = \frac{3}{4} \times (3 - 4.25)^2 + \frac{1}{4} \times (8 - 4.25)^2 \\ &= 4.6875 \end{aligned}$$

When we observe  $X$ , we get either 3 or 8: this is random. But  $\mu_X$  is fixed at 4.25, and  $\sigma_X^2$  is fixed at 4.6875, regardless of the outcome of  $X$ .

For a discrete random variable,

$$\text{Var}(X) = \mathbb{E} [(X - \mu_X)^2] = \sum_x (x - \mu_X)^2 f_X(x) = \sum_x (x - \mu_X)^2 \mathbb{P}(X = x).$$

This uses the definition of the expected value of a function of  $X$ :

$$\text{Var}(X) = \mathbb{E}(g(X)) \text{ where } g(X) = (X - \mu_X)^2.$$

**Theorem 2.2:** (important)

$$\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \mathbb{E}(X^2) - \mu_X^2$$

Proof:

$$\begin{aligned} \text{Var}(X) &= \mathbb{E} [(X - \mu_X)^2] \quad \text{by definition} \\ &= \mathbb{E} \left[ \underbrace{X^2}_{\text{r.v.}} - 2 \underbrace{X}_{\text{r.v.}} \underbrace{\mu_X}_{\text{constant}} + \underbrace{\mu_X^2}_{\text{constant}} \right] \\ &= \mathbb{E}(X^2) - 2\mu_X \mathbb{E}(X) + \mu_X^2 \quad \text{by Thm 2.1} \\ &= \mathbb{E}(X^2) - 2\mu_X^2 + \mu_X^2 \\ &= \mathbb{E}(X^2) - \mu_X^2. \quad \square \end{aligned}$$

**Note:**  $\mathbb{E}(X^2) = \sum_x x^2 f_X(x) = \sum_x x^2 \mathbb{P}(X = x)$ . This is not the same as  $(\mathbb{E}X)^2$ :

eg.

$$X = \begin{cases} 3 & \text{with probability } 0.75 \\ 8 & \text{with probability } 0.25, \end{cases}$$

then  $\mu_X = \mathbb{E}(X) = 4.25$ , so  $\mu_X^2 = (\mathbb{E}X)^2 = (4.25)^2 = 18.0625$ .

But  $\mathbb{E}(X^2) = (3^2 \times \frac{3}{4} + 8^2 \times \frac{1}{4}) = 22.75$ .

Thus

$$\mathbb{E}(X^2) \neq (\mathbb{E}X)^2 \text{ in general.}$$

**Theorem 2.3:** If  $a$  and  $b$  are constants and  $g(x)$  is a function, then

i)  $\text{Var}[aX + b] = a^2 \text{Var}(X)$

ii)  $\text{Var}[a g(X) + b] = a^2 \text{Var}[g(X)]$

**Proof:** (part (ii))

$$\begin{aligned} \text{Var}(ag(X) + b) &= \mathbb{E}\left[\{(ag(X) + b) - \mathbb{E}(ag(X) + b)\}^2\right] \\ &= \mathbb{E}\left[\{ag(X) + b - a\mathbb{E}(g(X)) - b\}^2\right] \quad \text{by Thm 2.1} \\ &= \mathbb{E}\left[\{ag(X) - a\mathbb{E}(g(X))\}^2\right] \\ &= \mathbb{E}\left[a^2\{g(X) - \mathbb{E}(g(X))\}^2\right] \\ &= a^2\mathbb{E}\left[\{g(X) - \mathbb{E}(g(X))\}^2\right] \quad \text{by Thm 2.1(i)} \\ &= a^2 \text{Var}[g(X)]. \end{aligned}$$

Part (i) follows by putting  $g(X) = X$ . □

**Note:** These are very different from the corresponding expressions for expectations (Theorem 2.1). Variances are more difficult to manipulate than expectations.

**Example: finding expectation and variance from the probability function**

---

Define  $X$  by the following probability function:

$x$	0	1	2	3
$f_X(x) = \mathbb{P}(X = x)$	$\frac{1}{8}$	$\frac{5}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

Then

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x=0}^3 x f_X(x) = 0 \times \frac{1}{8} + 1 \times \frac{5}{8} + 2 \times \frac{1}{8} + 3 \times \frac{1}{8} \\ &= \frac{10}{8} \\ &= 1.25\end{aligned}$$

$\text{Var}(X)$ : First method, use  $\mathbb{E}[(X - \mu_X)^2]$ :

$$\begin{aligned}\text{Var}(X) &= \sum_{x=0}^3 (x - 1.25)^2 f_X(x) \\ &= (0 - 1.25)^2 \times \frac{1}{8} + (1 - 1.25)^2 \times \frac{5}{8} + (2 - 1.25)^2 \times \frac{1}{8} + (3 - 1.25)^2 \times \frac{1}{8} \\ &= 0.6875\end{aligned}$$

Second method: use  $\mathbb{E}(X^2) - \mu_X^2$ : (usually easier)

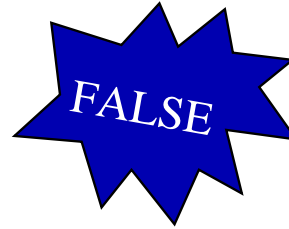
$$\begin{aligned}\mathbb{E}(X^2) &= \sum_{x=0}^3 x^2 f_X(x) = 0^2 \times \frac{1}{8} + 1^2 \times \frac{5}{8} + 2^2 \times \frac{1}{8} + 3^2 \times \frac{1}{8} \\ &= 2.25\end{aligned}$$

So  $\text{Var}(X) = 2.25 - (1.25)^2 = 0.6875$  as before.

# Interlude:



or



??

Guess whether each of the following statements is true or false.

1. Toss a fair coin 10 times. The probability of getting 8 or more heads is less than 1%.
2. Toss a fair coin 200 times. The chance of getting a run of at least 6 heads or 6 tails in a row is less than 10%.
3. Consider a classroom with 30 pupils of age 5, and one teacher of age 50. The probability that the pupils all outlive the teacher is about 90%.
4. Open the Business Herald at the pages giving share prices, or open an atlas at the pages giving country areas or populations. Pick a column of figures.

SHARE	LAST SALE
A Barnett	143
Advantage I	23
AFFCO	18
Air NZ	52
⋮	⋮

The figures are over 5 times more likely to begin with the digit 1 than with the digit 9.

Answers: 1. FALSE it is 5.5%. 2. FALSE it is 97%. 3. FALSE : in NZ the probability is about 50%. 4. TRUE : in fact they are 6.5 times more likely.

## 2.2 Distribution of transformed random variables

Suppose we know the distribution (i.e. *probability function*) of  $X$ .  
How do we find the distribution of  $Y = g(X)$ ?

**Example:** Let  $X$  be as follows:

$x$	-1	0	1
$f_X(x) = \mathbb{P}(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Let  $Y = 2X$  (so  $g(X)=2X$ ).

Clearly, the probability function of  $Y$  is:

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

**Thus**

$$\begin{aligned} f_Y(-2) &= \mathbb{P}(Y = -2) \\ &= \mathbb{P}(2X = -2) \\ &= \mathbb{P}\left(X = -\frac{2}{2}\right) \\ &= \mathbb{P}(X = -1) \\ &= \frac{1}{4} \end{aligned}$$

$$\text{Similarly, } f_Y(0) = \mathbb{P}(Y = 0) = \mathbb{P}\left(X = \frac{0}{2}\right) = \frac{1}{2},$$

$$\text{and } f_Y(2) = \mathbb{P}(Y = 2) = \mathbb{P}\left(X = \frac{2}{2}\right) = \frac{1}{4}.$$

So if  $Y = g(X) = 2X$ , then  $f_Y(y) = f_X\left(\frac{y}{2}\right) = f_X(g^{-1}(y))$ .

This is true in general, as follows.

## General Result:

Suppose  $Y = g(X)$  and  $g$  is *injective (one-to-one)*: that is, there are no two values  $x_1$  and  $x_2$  such that  $g(x_1) = g(x_2)$ . Then the inverse function  $g^{-1}$  is well-defined, and

$$f_Y(y) = f_X(g^{-1}(y))$$

If  $g$  is not injective, there may be more than one value of  $x$  such that  $g(x) = y$ . In this case,

$$f_Y(y) = \sum_{x:g(x)=y} f_X(x)$$

**Example:** let  $X$  be as above:

$$\begin{array}{c|ccc} x & -1 & 0 & 1 \\ \hline f_X(x) & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{array}$$

Let  $Y = X^2$ .

When

$$X = -1, Y = 1 : \text{prob} = \frac{1}{4}$$

$$X = 0, Y = 0 : \text{prob} = \frac{1}{2}$$

$$X = 1, Y = 1 : \text{prob} = \frac{1}{4}$$

$$\therefore \begin{array}{c|cc} y & 0 & 1 \\ \hline f_Y(y) & \frac{1}{2} & \frac{1}{4} + \frac{1}{4} \end{array}$$

In this case, we have:

$$f_Y(0) = \sum_{x:x^2=0} f_X(x) = f_X(0) = \frac{1}{2},$$

$$\text{and } f_Y(1) = \sum_{x:x^2=1} f_X(x) = f_X(-1) + f_X(1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

## 2.3 Examples of Discrete Distributions

Part of the reason for looking at random variables is to be able to describe several different situations all in the same way.

For example, toss a fair coin; let  $X = \begin{cases} 0 & \text{if tail (probability } 1/2), \\ 1 & \text{if head (probability } 1/2), \end{cases}$

or spin a balanced pointer, and let  $Y = \begin{cases} 0 & \text{if } \leq 180^\circ \text{ (probability } 1/2), \\ 1 & \text{if } > 180^\circ \text{ (probability } 1/2). \end{cases}$

The situations are different, but the random variables  $X$  and  $Y$  behave in exactly the same way.

For this reason, we have several ‘standard’ random variables which describe common situations. We work out their properties, and can then apply the results whenever we encounter these situations.

### 1. Binomial distribution

*Definition:* A random experiment is called a set of Bernoulli trials if it consists of several trials such that:

- i) *Each trial has only 2 possible outcomes (usually called “Success” and “Failure”);*
- ii) *The probability of success,  $p$ , remains constant for all trials;*
- iii) *The trials are independent, ie. the event “success in trial  $i$ ” does not depend on the outcome of any other trials.*

*Examples:* 1) Repeated tossing of a fair coin: *each toss is a Bernoulli trial with  $\mathbb{P}(\text{success}) = \mathbb{P}(\text{head}) = \frac{1}{2}$ .*



2) Repeated tossing of a fair die: *success* = “6”, *failure* = “not 6”. Each toss is a *Bernoulli trial* with  $\mathbb{P}(\text{success}) = \frac{1}{6}$ .

*Definition:* The random variable  $Y$  is called a **Bernoulli random variable** if it takes only 2 values, 0 and 1.

The probability function is,

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

That is,

$$\begin{aligned} \mathbb{P}(Y = 1) &= \mathbb{P}(\text{“success”}) = p, \\ \mathbb{P}(Y = 0) &= \mathbb{P}(\text{“failure”}) = 1 - p. \end{aligned}$$

*Definition:* Let  $X$  be the number of successes in  $n$  independent Bernoulli trials each with probability of success =  $p$ . Then  $X$  has the **Binomial distribution** with parameters  $n$  and  $p$ . We write  $X \sim \text{Bin}(n, p)$ , or  $X \sim \text{Binomial}(n, p)$ .

Thus  $X \sim \text{Bin}(n, p)$  if  $X$  is the number of successes out of  $n$  independent trials, each of which has probability  $p$  of success.

## Properties of the Binomial distribution

### i) Probability function

If  $X \sim \text{Binomial}(n, p)$ , then the probability function for  $X$  is

$$f_X(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \quad \text{for } x = 0, 1, \dots, n$$

Explanation: An outcome with  $x$  successes ( $n - x$ ) failures has probability,

$$\underbrace{p^x}_{(1)} \underbrace{(1 - p)^{n-x}}_{(2)}$$

where:

(1) succeeds  $x$  times, each with probability  $p$

(2) fails  $(n - x)$  times, each with probability  $(1 - p)$ .

There are  $\binom{n}{x}$  possible outcomes with  $x$  successes and  $(n - x)$  failures because we must select  $x$  trials to be our “successes”, out of  $n$  trials in total.

Thus,

$$\begin{aligned} \mathbb{P}(\# \text{successes} = x) &= (\# \text{outcomes with } x \text{ successes}) \times (\text{prob. of each such outcome}) \\ &= \binom{n}{x} p^x (1 - p)^{n-x} \end{aligned}$$

**Note:**

$$f_X(x) = 0 \quad \text{if } x \notin \{0, 1, 2, \dots, n\}.$$

Check that  $\sum_{x=0}^n f_X(x) = 1$ :

$$\begin{aligned} \sum_{x=0}^n f_X(x) &= \sum_{x=0}^n \binom{n}{x} p^x (1 - p)^{n-x} = [p + (1 - p)]^n \quad (\text{Binomial Theorem}) \\ &= 1^n = 1 \end{aligned}$$

It is this connection with the Binomial Theorem that gives the Binomial Distribution its name.

## ii) Mean and variance of $\text{Bin}(n, p)$

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

$$\mathbb{E}(X) = \mu_X = np$$

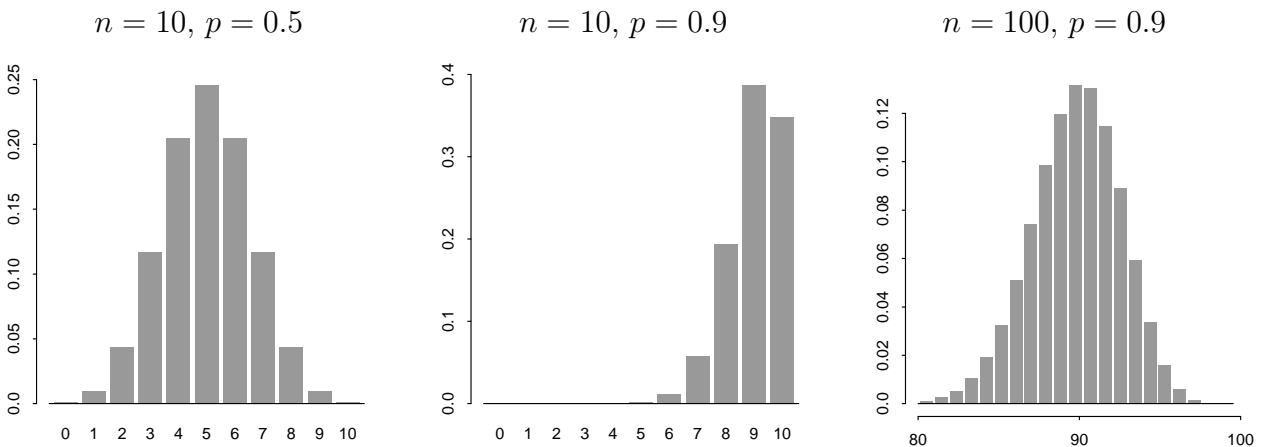
$$\text{Var}(X) = \sigma_X^2 = np(1-p)$$

We often write  $q = 1 - p$ , so  $\text{Var}(X) = npq$ .

## iii) Shape

The shape of the Binomial distribution depends upon the values of  $n$  and  $p$ . For small  $n$ , the distribution is almost symmetrical for values of  $p$  close to 0.5, but highly skewed for values of  $p$  close to 0 or 1. As  $n$  increases, the distribution becomes more and more symmetrical, and there is noticeable skew only if  $p$  is very close to 0 or 1.

The probability functions for various values of  $n$  and  $p$  are shown below.



## iv) Sum of independent Binomial random variables:

If  $X$  and  $Y$  are *independent*, and  $X \sim \text{Binomial}(n, p)$ ,  $Y \sim \text{Binomial}(m, p)$ , then

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

**Proof that  $\mathbb{E}(X) = np$  and  $\text{Var}(X) = np(1 - p)$  for  $X \sim \text{Binomial}(n, p)$**

---

$$\mathbb{E}(X) = \sum_{x=0}^n x f_X(x) = \sum_{x=0}^n x \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=0}^n x \left( \frac{n!}{(n-x)!x!} \right) p^x (1-p)^{n-x}$$

But  $\frac{x}{x!} = \frac{1}{(x-1)!}$  and also the first term  $x f_X(x)$  is 0 when  $x = 0$ .

So, continuing,

$$\mathbb{E}(X) = \sum_{x=1}^n \frac{n!}{(n-x)!(x-1)!} p^x (1-p)^{n-x}$$

Next: make  $n$ 's into  $(n-1)$ 's,  $x$ 's into  $(x-1)$ 's, wherever possible  
eg.

$$\begin{aligned} n-x &= (n-1) - (x-1), & p^x &= p \cdot p^{x-1} \\ n! &= n(n-1)! \text{ etc.} \end{aligned}$$

This gives,

$$\begin{aligned} \mathbb{E}(X) &= \sum_{x=1}^n \frac{n(n-1)!}{[(n-1) - (x-1)]!(x-1)!} p \cdot p^{(x-1)} (1-p)^{(n-1)-(x-1)} \\ &= \underbrace{np}_{\text{what we want}} \underbrace{\sum_{x=1}^n \binom{n-1}{x-1} p^{x-1} (1-p)^{(n-1)-(x-1)}}_{\text{need to show this sum} = 1} \end{aligned}$$

Finally we let  $y = x - 1$  and let  $m = n - 1$ .

When  $x = 1, y = 0$ ; and when  $x = n, y = n - 1 = m$ .

So

$$\begin{aligned} \mathbb{E}(X) &= np \sum_{y=0}^m \binom{m}{y} p^y (1-p)^{m-y} \\ &= np(p + (1-p))^m \quad (\text{Binomial Theorem}) \end{aligned}$$

$$\mathbb{E}(X) = np, \quad \text{as required.}$$

For  $\text{Var}(X)$ , use the same ideas again.

For  $\mathbb{E}(X)$ , we used  $\frac{x}{x!} = \frac{1}{(x-1)!}$ ; so instead of finding  $\mathbb{E}(X^2)$ , it will be easier to find  $\mathbb{E}[X(X-1)] = \mathbb{E}(X^2) - \mathbb{E}(X)$  because then we will be able to cancel  $\frac{x(x-1)}{x!} = \frac{1}{(x-2)!}$ .

Here goes:

$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{x=0}^n x(x-1) \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^n \frac{x(x-1)n(n-1)(n-2)!}{[(n-2)-(x-2)]!(x-2)!x(x-1)} p^2 p^{(x-2)} (1-p)^{(n-2)-(x-2)}\end{aligned}$$

First two terms ( $x=0$  and  $x=1$ ) are 0 due to the  $x(x-1)$  in the numerator.

Thus

$$\begin{aligned}\mathbb{E}[X(X-1)] &= p^2 n(n-1) \sum_{x=2}^n \binom{n-2}{x-2} p^{x-2} (1-p)^{(n-2)-(x-2)} \\ &= n(n-1)p^2 \underbrace{\sum_{y=0}^m \binom{m}{y} p^y (1-p)^{m-y}}_{\text{sum}=1 \text{ by Binomial Theorem}} \quad \text{if } \begin{cases} m = n-2, \\ y = x-2. \end{cases}\end{aligned}$$

So  $\mathbb{E}[X(X-1)] = n(n-1)p^2$ .

$$\begin{aligned}\text{Thus } \text{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \\ &= \mathbb{E}(X^2) - \mathbb{E}(X) + \mathbb{E}(X) - (\mathbb{E}(X))^2 \\ &= \mathbb{E}[X(X-1)] + \mathbb{E}(X) - (\mathbb{E}(X))^2 \\ &= n(n-1)p^2 + np - n^2p^2 \\ &= np(1-p). \quad \square\end{aligned}$$

Note the steps: take out  $x(x-1)$  and replace  $n$  by  $(n-2)$ ,  $x$  by  $(x-2)$  wherever possible.

## 2. Poisson distribution

So far, we have looked at the Binomial distribution, which arises in nature as the number of successes in a sequence of identical, independent Bernoulli trials.

The Poisson distribution is another distribution that arises in nature, through the so-called **Poisson process**. The Poisson process describes a physical situation that is guaranteed to produce a Poisson distribution — just as the number of successes in repeated Bernoulli trials is guaranteed to follow a Binomial distribution. Roughly speaking, the Poisson process counts the *number of events occurring in a fixed time or space, when events occur independently and at a constant average rate*.

The Poisson distribution has one parameter,  $\lambda$ , which in a Poisson process equals the average rate at which events occur.

**Example:** customers arriving at a bank. Suppose that customers arrive at an average rate of 20 per hour, independently of each other.  
*If  $X =$  number of customers to arrive in a 1-hour period, we can use the Poisson distribution with rate  $\lambda = 20$  to model  $X$ .*

We will define the Poisson process formally below.

The Poisson process is a mathematically exact situation that will always result in a Poisson distribution. However, the Poisson distribution is also widely used as a ‘subjective model’ in situations that are not mathematically exact. Statisticians use subjective models when they need to describe the randomness in a situation that has no known mathematical formulation. Essentially, they are suggesting that the shape and variability of the distribution they are interested in is well captured by a Poisson distribution.

The difference between an exact model and a subjective model is important. Exact models, such as the Binomial distribution from Bernoulli trials, or the Poisson distribution from the Poisson process, are quite rare in real life; it is far more common for a subjective model to be required.

**Example:** Let  $X$  be the number of children of a randomly selected NZ woman. There is no mathematical formulation that can describe  $X$  exactly. However, a reasonable subjective model for  $X$  might be  $X \sim \text{Poisson}(\lambda = 2.5)$ .

## Properties of the Poisson distribution

### i) Probability function

If  $X$  has a Poisson distribution with parameter  $\lambda$ , the probability function of  $X$  is

$$f_X(x) = \mathbb{P}(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \text{ for } x = 0, 1, 2, \dots$$

The parameter  $\lambda$  is called the rate of the Poisson distribution. In the bank example above,  $\lambda = 20$  for the rate at which customers arrive. (*20 per hour*)

We write  $X \sim \text{Poisson}(\lambda)$  (eg.  $X \sim \text{Poisson}(20)$ ).

### ii) Mean and variance

*The mean and variance of the  $\text{Poisson}(\lambda)$  distribution are both  $\lambda$ .*

$$\mathbb{E}(X) = \text{Var}(X) = \lambda \text{ when } X \sim \text{Poisson}(\lambda)$$

#### **Notes:**

1. *It makes sense for  $\mathbb{E}(X) = \lambda$ . If events occur at a constant average rate of  $\lambda$  per unit time, then the mean of the number of events to occur in one unit of time should indeed be  $\lambda$ .*

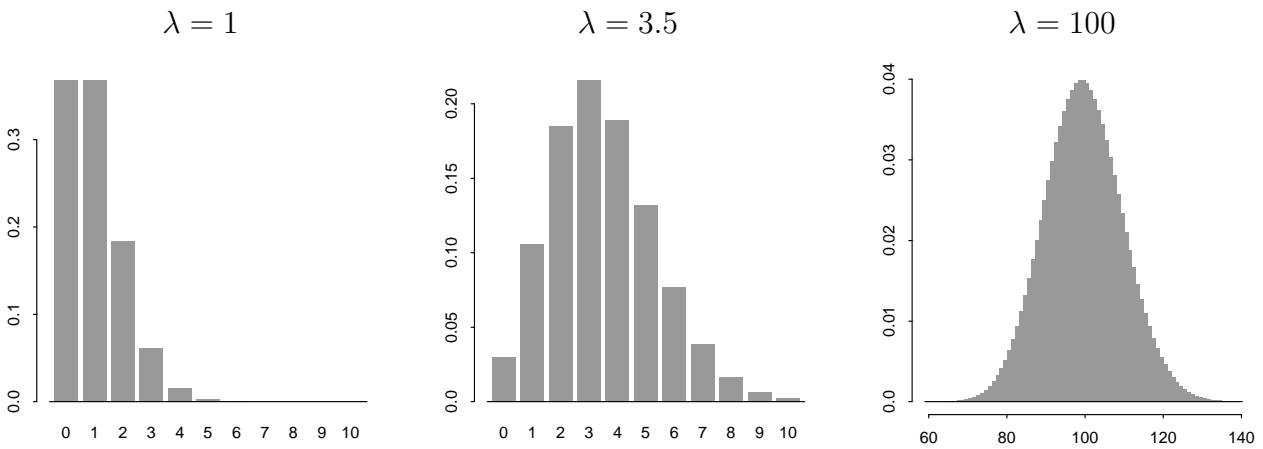
2. *The variance of the Poisson distribution increases with the mean (in fact, variance = mean). This is very often the case in real life: there is more uncertainty associated with larger numbers than with smaller numbers.*

*Despite this, the variance of the Poisson distribution is often too small to describe real-life situations adequately. In real life, the variance of a phenomenon often increases faster than the mean.*

### iii) Shape

The shape of the Poisson distribution depends upon the value of  $\lambda$ . For small  $\lambda$ , the distribution has positive (right) skew. As  $\lambda$  increases, the distribution becomes more and more symmetrical, until for large  $\lambda$  it has the familiar bell-shaped appearance.

The probability functions for various  $\lambda$  are shown below.



### iv) Sum of independent Poisson random variables

If  $X$  and  $Y$  are independent, and  $X \sim \text{Poisson}(\lambda)$ ,  $Y \sim \text{Poisson}(\mu)$ , then

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



## Proof that $\mathbb{E}(X) = \text{Var}(X) = \lambda$ for $X \sim \text{Poisson}(\lambda)$

For  $X \sim \text{Poisson}(\lambda)$ , the probability function is  $f_X(x) = \frac{\lambda^x}{x!}e^{-\lambda}$  for  $x = 0, 1, 2, \dots$

So

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x f_X(x) = \sum_{x=0}^{\infty} x \left( \frac{\lambda^x}{x!} e^{-\lambda} \right) \\ &= \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} e^{-\lambda} \quad (\text{note that term for } x=0 \text{ is } 0) \\ &= \lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} e^{-\lambda} \quad (\text{writing everything in terms of } x-1) \\ &= \lambda \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} e^{-\lambda} \quad (\text{putting } y = x-1) \\ &= \lambda, \quad \text{because the sum}=1 \text{ (sum of Poisson probabilities).} \end{aligned}$$

So  $\mathbb{E}(X) = \lambda$ , as required.

$$\begin{aligned} \text{For } \text{Var}(X), \text{ we use:} \quad \text{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}X)^2 \\ &= \mathbb{E}[X(X-1)] + \mathbb{E}(X) - (\mathbb{E}X)^2 \\ &= \mathbb{E}[X(X-1)] + \lambda - \lambda^2. \end{aligned}$$

$$\begin{aligned} \text{But } \mathbb{E}[X(X-1)] &= \sum_{x=0}^{\infty} x(x-1) \frac{\lambda^x}{x!} e^{-\lambda} \\ &= \sum_{x=2}^{\infty} \frac{\lambda^x}{(x-2)!} e^{-\lambda} \quad (\text{terms for } x=0 \text{ and } x=1 \text{ are } 0) \\ &= \lambda^2 \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} e^{-\lambda} \quad (\text{writing everything in terms of } x-2) \\ &= \lambda^2 \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} e^{-\lambda} \quad (\text{putting } y = x-2) \\ &= \lambda^2. \end{aligned}$$

So

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[X(X-1)] + \lambda - \lambda^2 \\ &= \lambda^2 + \lambda - \lambda^2 \\ &= \lambda, \quad \text{as required.}\end{aligned}$$

---

## Poisson process with rate $\lambda$

We now define the Poisson process properly. Recall that the Poisson *distribution* is used as a model in a wide range of situations where it is not mathematically exact, but that the Poisson *process* is a single physical situation that *does* give rise to an exact Poisson distribution.

Consider a sequence of events occurring over time (e.g. customers arriving at a bank).

Let  $X_t$  be the *number of events to have occurred by time  $t$ , ie. in the time interval from time 0 to time  $t$ .*

If the events occur according to a Poisson process, the distribution of  $X_t$  can be shown to be Poisson for any  $t > 0$ . In intuitive terms, the conditions for a Poisson process are as follows:

- i) all events are *independent*;
- ii) events occur at a *constant average rate of  $\lambda$* ;
- iii) events *cannot occur simultaneously*.

When these conditions are satisfied, then the number of events to have occurred by time  $t$  has distribution

$$X_t \sim \text{Poisson}(\lambda t) : \text{ so } \mathbb{P}(X_t = x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t} \quad (x = 0, 1, 2, \dots)$$

For a spatial Poisson process,  $X_A = \# \text{ occurrences in area of size } A \sim \text{Poisson}(\lambda A)$ . The mathematical formulation of the Poisson process conditions is as follows.

*Definition:* The random variables  $\{X_t : t > 0\}$  form a **Poisson process with rate  $\lambda$**  if:

i) *events occurring in any time interval are independent of those occurring in any other disjoint time interval;*

ii)

$$\lim_{\delta t \downarrow 0} \left( \frac{\mathbb{P}(\text{exactly one event occurs in time interval}[t, t + \delta t])}{\delta t} \right) = \lambda$$

iii)

$$\lim_{\delta t \downarrow 0} \left( \frac{\mathbb{P}(\text{more than one event occurs in time interval}[t, t + \delta t])}{\delta t} \right) = 0$$

---

### Poisson approximation to the Binomial distribution

Let  $X \sim \text{Binomial}(n, p)$  (so  $X$  is the number of successes out of  $n$  Bernoulli trials, each with probability of success =  $p$ ).

If:

i)  *$n$  is large,*

ii)  *$p$  is small,*

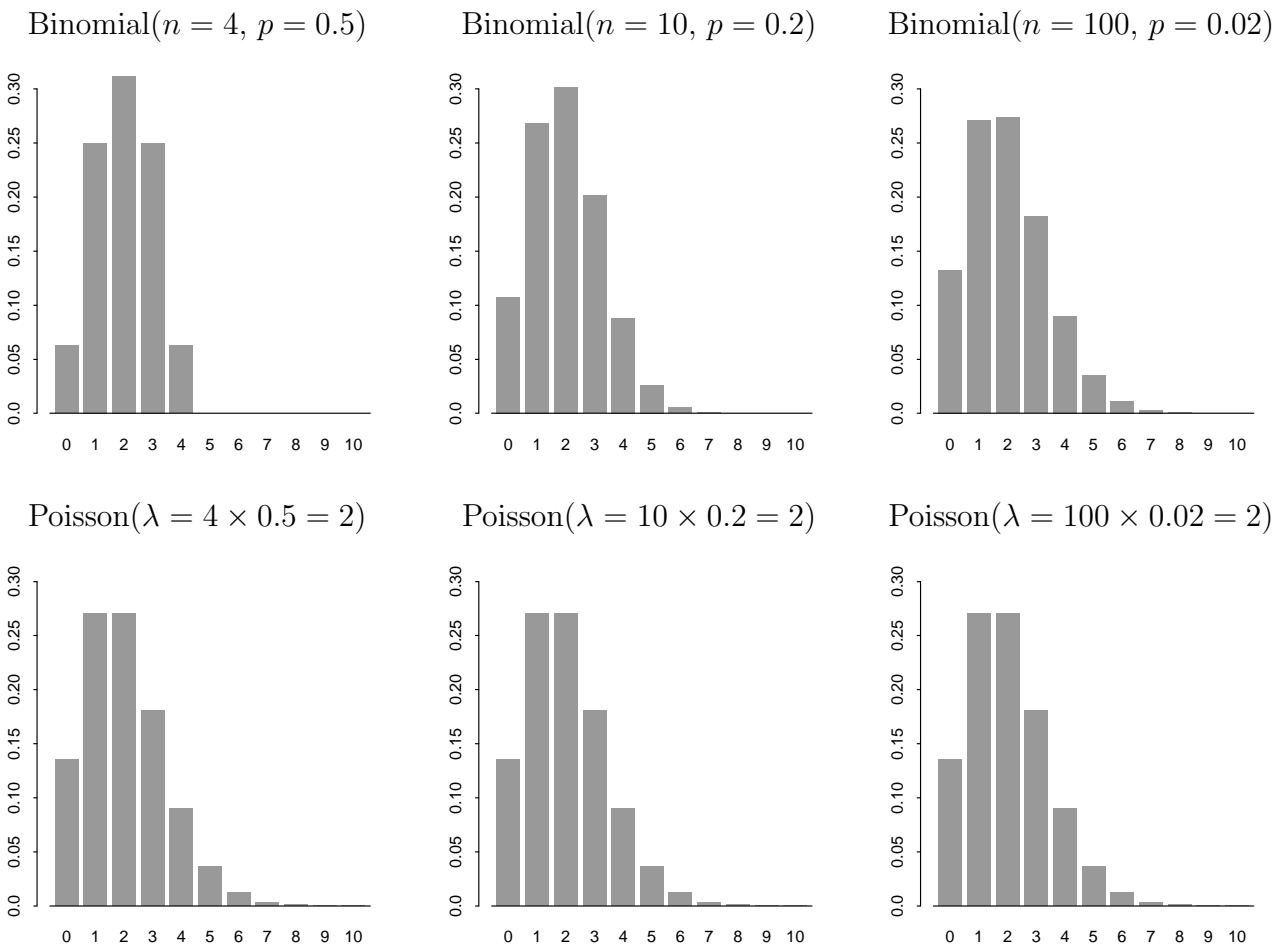
iii)  *$np$  is moderately-sized,*

then

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

So

$\text{Bin}(n, p) \rightarrow \text{Poisson}(\lambda = np)$  when  $n \rightarrow \infty, p \rightarrow 0$  and  $\lambda = np$  is fixed.



The figures show how the probability function of the Binomial( $n, p$ ) distribution looks more like the Poisson( $\lambda = n \times p$ ) distribution as  $n$  becomes large and  $p$  becomes small, although  $np$  is fixed at the value 2.

### Why the approximation works:

The Poisson distribution models the number of events to occur in a fixed time interval, when events occur at a constant average rate. We can imagine splitting the time interval into a large number  $n$  of tiny intervals. In each of the  $n$  tiny intervals, there is a very small probability  $p$  that an event occurs (i.e. that a “success” occurs).

Thus, the number of events to occur in the large time interval, which is Poisson, is also approximately the number of successes in the  $n$  tiny intervals, which is Binomial. The approximation gets better as the number of intervals,  $n$ , becomes *large* and the probability  $p$  becomes *small*.

## Proof of the Poisson approximation to the Binomial

Let  $X \sim \text{Binomial}(n, p)$ , where  $np = \lambda$  (so  $p = \frac{\lambda}{n}$ ).

Then

$$\begin{aligned}\mathbb{P}(X = x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \frac{n(n-1)\dots(n-x+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\ &= \frac{1}{x!} \underbrace{\left(\frac{n}{n}\right) \left(\frac{n-1}{n}\right) \dots \left(\frac{n-x+1}{n}\right)}_{(\rightarrow 1 \text{ as } n \rightarrow \infty)} \lambda^x \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{(\rightarrow e^{-\lambda})} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-x}}_{(\rightarrow 1)}\end{aligned}$$

So as  $n \rightarrow \infty$  and  $p \rightarrow 0$  such that  $np = \lambda$ , we have

$$\mathbb{P}(X = x) \rightarrow \frac{\lambda^x}{x!} e^{-\lambda},$$

which is the probability function for the Poisson( $\lambda$ ) distribution.  $\square$

---

### 3. Geometric distribution

Like the Binomial distribution, the Geometric distribution is defined in terms of a sequence of Bernoulli trials. However, while the Binomial distribution counts the *number of successes out of a fixed number of* Bernoulli trials, the Geometric distribution counts the *number of trials before the first success occurs*.

*Definition:* Let  $X$  be the number of failures that occur before the first success in a sequence of Bernoulli trials with  $\mathbb{P}(\text{success}) = p$ . Then  $X$  has the Geometric distribution with parameter  $p$ . We write  $X \sim \text{Geometric}(p)$ .

# Properties of the Geometric distribution

## i) Probability function

If  $X \sim \text{Geometric}(p)$ , the probability function of  $X$  is

$$f_X(x) = \mathbb{P}(X = x) = (1 - p)^x p \text{ for } x = 0, 1, 2, \dots$$

**Note:**  $\mathbb{P}(X = x) = \underbrace{(1 - p)^x}_{\text{need } x \text{ failures}} \times \underbrace{p}_{\text{final trial must be a success}}$

## ii) Mean and variance

For  $X \sim \text{Geometric}(p)$ ,

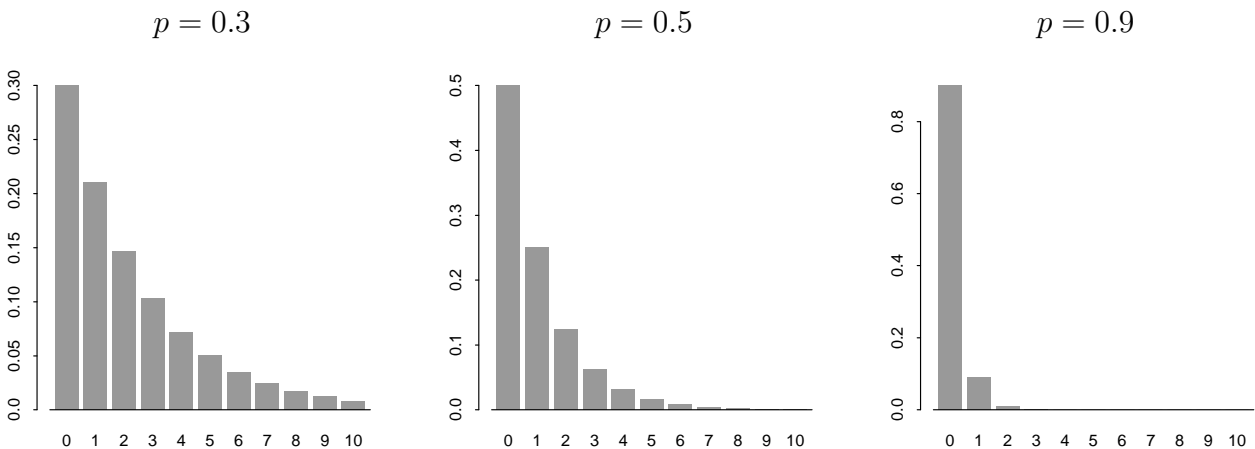
$$\mathbb{E}(X) = \frac{1 - p}{p} = \frac{q}{p}$$

$$\text{Var}(X) = \frac{1 - p}{p^2} = \frac{q}{p^2}$$

## iii) Shape

The shape of the Geometric distribution depends upon the value of  $p$ . For small  $p$ , it is likely that there will be many failures before a success occurs, so the distribution has a long tail. For large  $p$ , a success is likely to occur almost immediately, so the distribution has a short tail. The geometric distribution is always *positively skewed (right skewed)*.

The probability functions for various  $p$  are shown below.



#### iv) Sum of independent Geometric random variables

If  $X_1, \dots, X_k$  are *independent*, and each  $X_i \sim \text{Geometric}(p)$ , then

$$X_1 + \dots + X_k \sim \text{Negative Binomial}(k, p). \quad (\text{see later})$$

Proof that  $\mathbb{E}(X) = \frac{1-p}{p}$  and  $\text{Var}(X) = \frac{1-p}{p^2}$  for  $X \sim \text{Geometric}(p)$

We use the following results:

$$\sum_{x=1}^{\infty} xq^{x-1} = \frac{1}{(1-q)^2} \quad (\text{for } |q| < 1), \quad (1)$$

and

$$\sum_{x=2}^{\infty} x(x-1)q^{x-2} = \frac{2}{(1-q)^3} \quad (\text{for } |q| < 1). \quad (2)$$

#### Proof of (1) and (2):

Consider the infinite sum of a geometric progression:

$$\sum_{x=0}^{\infty} q^x = \frac{1}{1-q} \quad (\text{for } |q| < 1).$$

Differentiate both sides with respect to  $q$ :

$$\begin{aligned} \frac{d}{dq} \left( \sum_{x=0}^{\infty} q^x \right) &= \frac{d}{dq} \left( \frac{1}{1-q} \right) \\ \sum_{x=0}^{\infty} \frac{d}{dq} (q^x) &= \frac{1}{(1-q)^2} \\ \sum_{x=1}^{\infty} xq^{x-1} &= \frac{1}{(1-q)^2}, \quad \text{as stated in (1)}. \end{aligned}$$

Note that the lower limit of the summation becomes  $x = 1$  because the term for  $x = 0$  vanishes.

The proof of (2) is obtained similarly, by differentiating both sides of (1) with respect to  $q$  (Exercise).

Now we can find  $\mathbb{E}(X)$  and  $\text{Var}(X)$ .

$$\begin{aligned}
 \mathbb{E}(X) &= \sum_{x=0}^{\infty} x\mathbb{P}(X = x) \\
 &= \sum_{x=0}^{\infty} xpq^x \quad (\text{where } q = 1 - p) \\
 &= p \sum_{x=1}^{\infty} xq^x \quad (\text{lower limit becomes } x = 1 \text{ because term in } x = 0 \text{ is zero}) \\
 &= pq \sum_{x=1}^{\infty} xq^{x-1} \\
 &= pq \left( \frac{1}{(1-q)^2} \right) \quad (\text{by equation (1)}) \\
 &= pq \left( \frac{1}{p^2} \right) \quad (\text{because } 1 - q = p) \\
 &= \frac{q}{p}, \quad \text{as required.}
 \end{aligned}$$

For  $\text{Var}(X)$ , we use

$$\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \mathbb{E}\{X(X-1)\} + \mathbb{E}(X) - (\mathbb{E}X)^2. \quad (\star)$$

Now

$$\begin{aligned}
 \mathbb{E}\{X(X-1)\} &= \sum_{x=0}^{\infty} x(x-1)\mathbb{P}(X = x) \\
 &= \sum_{x=0}^{\infty} x(x-1)pq^x \quad (\text{where } q = 1 - p) \\
 &= pq^2 \sum_{x=2}^{\infty} x(x-1)q^{x-2} \quad (\text{note that terms below } x = 2 \text{ vanish}) \\
 &= pq^2 \left( \frac{2}{(1-q)^3} \right) \quad (\text{by equation (2)}) \\
 &= \frac{2q^2}{p^2}.
 \end{aligned}$$

Thus by  $(\star)$ ,

$$\text{Var}(X) = \frac{2q^2}{p^2} + \frac{q}{p} - \left( \frac{q}{p} \right)^2 = \frac{q(q+p)}{p^2} = \frac{q}{p^2},$$

as required, because  $q + p = 1$ .



## 4. Negative Binomial distribution

*Definition:* Let  $X$  be the number of failures before the  $k$ 'th success in a sequence of Bernoulli trials, each with  $\mathbb{P}(\text{success}) = p$ . Then  $X \sim \text{Negative Binomial}$  with parameters  $k$  and  $p$ . We write  $X \sim \text{NegBin}(k, p)$ .

### Properties of the Negative Binomial distribution

#### i) Probability function

If  $X \sim \text{NegBin}(k, p)$ , the probability function of  $X$  is

$$f_X(x) = \mathbb{P}(X = x) = \binom{k+x-1}{x} p^k (1-p)^x \quad \text{for } x = 0, 1, 2, \dots$$

*Note:*  $\mathbb{P}(X = x) = \underbrace{\binom{k+x-1}{x}}_{\substack{\text{know that the last trial is a success:} \\ \text{need to choose } (k-1) \text{ other successes} \\ \text{and } x \text{ failures out of } (k-1+x) \text{ trials.}}} \times \overbrace{p^k}^{\text{need } k \text{ successes}} \times \underbrace{(1-p)^x}_{\text{need } x \text{ failures}}$

#### ii) Mean and variance

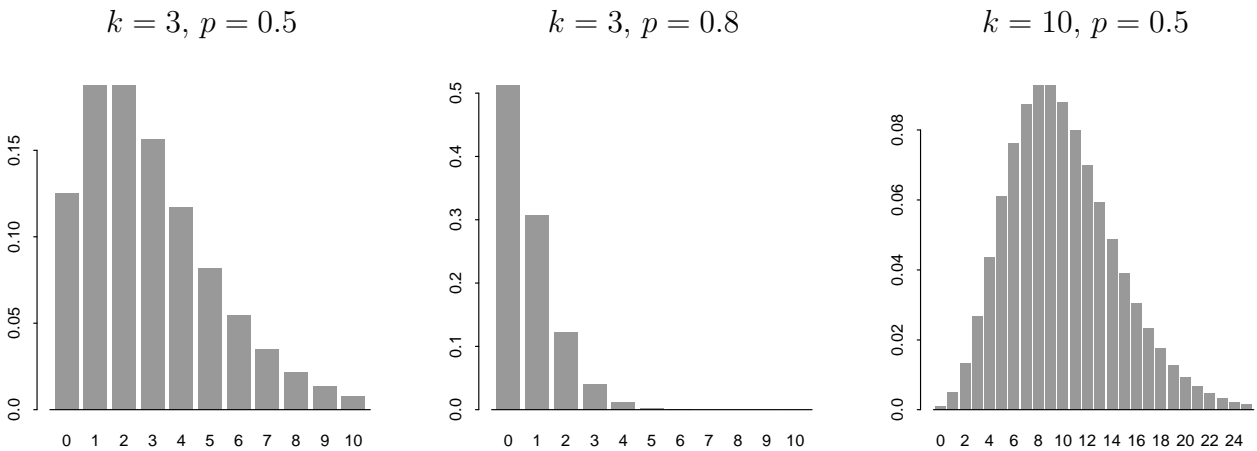
For  $X \sim \text{NegBin}(k, p)$ ,

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

*Proof:* not needed, but note it follows naturally from the result  $X = Y_1 + \dots + Y_k$ , where each  $Y_i \sim \text{Geom}(p)$ .

### iii) Shape

The figure shows the shape of the Negative Binomial distribution for various values of  $k$  and  $p$ .



### iv) Sum of independent Negative Binomial random variables

If  $X$  and  $Y$  are *independent*, and  $X \sim \text{NegBin}(k, p)$ ,  $Y \sim \text{NegBin}(m, p)$ , then

$$X + Y \sim \text{NegBin}(k + m, p).$$

**Note:** (Non-examinable). For the negative binomial distribution,

$$\text{Var}(X) = \frac{\mathbb{E}(X)}{p} > \mathbb{E}(X), \quad \text{because } p < 1.$$

This means that the variance of the negative binomial distribution is always greater than the mean. We can compare this with the Poisson distribution, for which variance is always equal to the mean. The larger variance of the negative binomial distribution makes it a popular choice to use instead of the Poisson distribution in ‘subjective’ modelling situations, because in real life situations there is often high variability.

## 5. Hypergeometric distribution

The hypergeometric distribution is used when we *are sampling without replacement from a finite population.*

*Definition: Suppose we have  $N$  objects, of which  $M$  are “special”:*

*(eg.  $N$  balls in a jar,  $M$  red balls, rest not red.)*

*Draw  $n$  balls without replacement.*

*Let  $X =$  number of the  $n$  balls that are “special”.*

*Then  $X \sim \text{Hypergeometric}(N, M, n)$ .*

### Properties of the Hypergeometric distribution

#### i) Probability function

If  $X \sim \text{Hypergeometric}(N, M, n)$ , the probability function of  $X$  is

$$f_X(x) = \mathbb{P}(X = x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} \text{ for } x = \max(0, n + M - N) \text{ to } x = \min(n, M)$$

*Explanation: there are  $\binom{M}{x}$  ways of choosing  $x$  special objects from the  $M$  special objects available.*

*For each of these ways, there are  $\binom{N-M}{n-x}$  ways of choosing  $(n-x)$  non-special objects from the  $(N-M)$  available.*

*So the total number of ways of choosing  $x$  special objects and  $(n-x)$  non-special objects is  $\binom{M}{x} \binom{N-M}{n-x}$ .*

*Total number of ways of choosing  $n$  objects from  $N$  is  $\binom{N}{n}$ .*

*So*

$$\mathbb{P}(X = x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}$$

**Note:**  $x$  must be at least  $n - (N - M)$  (the number of *special* objects needed to make up a sample of size  $n$  after all  $N - M$  non-special objects have been selected). Similarly,  $x$  cannot be more than  $n$  or  $M$ .

See this more easily by noting that we need  $0 \leq x \leq M$  (# red balls) and  $0 \leq n - x \leq N - M$  (# other balls).

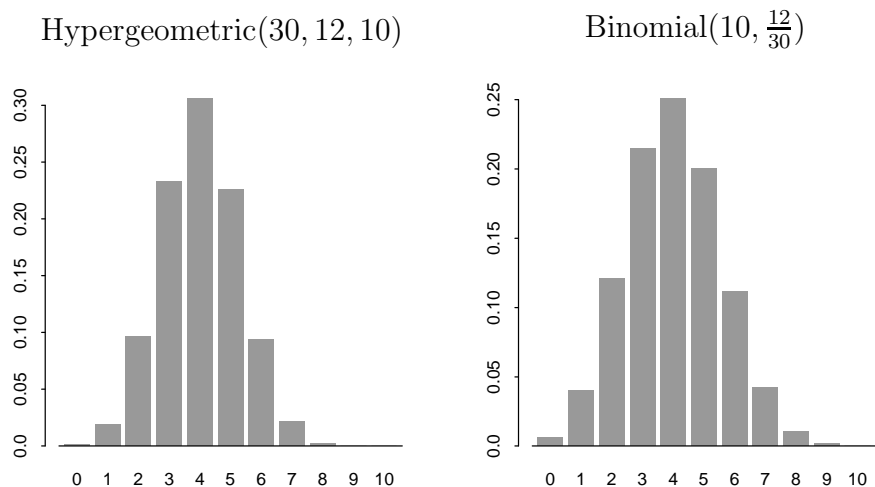
## ii) Mean and variance

For  $X \sim \text{Hypergeometric}(N, M, n)$ ,

$\mathbb{E}(X) = np$ $\text{Var}(X) = np(1 - p) \left( \frac{N-n}{N-1} \right)$	where $p = \frac{M}{N}$ .
---	---------------------------

## iii) Shape

The Hypergeometric distribution is similar to the Binomial distribution when  $n/N$  is small. For  $n/N < 0.1$  we often approximate the Hypergeometric( $N, M, n$ ) distribution by the **Binomial**( $n, p = \frac{M}{N}$ ) *distribution*.



**Note:** The Hypergeometric distribution is used for survey sampling and opinion polls, because *these involve sampling without replacement from a finite population*.

The Binomial distribution is used when the population is sampled with replacement.

As noted above,

$$\text{Hypergeometric}(N, M, n) \rightarrow \text{Binomial}(n, \frac{M}{N}) \text{ as } N \rightarrow \infty.$$

## 2.4 The Distribution Function, $F_X(x)$

We have defined the *probability function*,  $f_X(x)$ , as  $f_X(x) = \mathbb{P}(X = x)$

The *cumulative distribution function*, or just *distribution function*, written as  $F_X(x)$ , provides an alternative way of describing the distribution of  $X$ .

*Definition:* The (cumulative) distribution function (c.d.f.) is

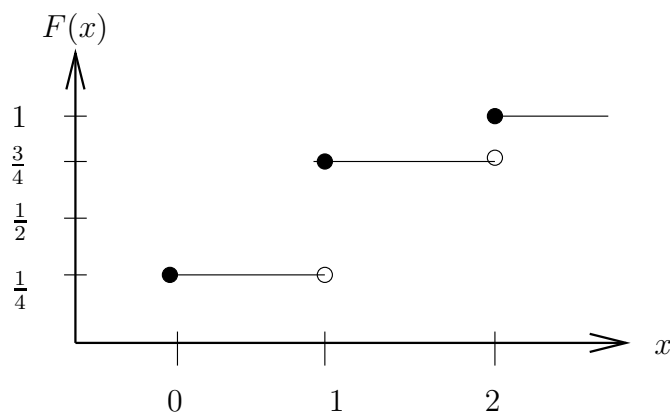
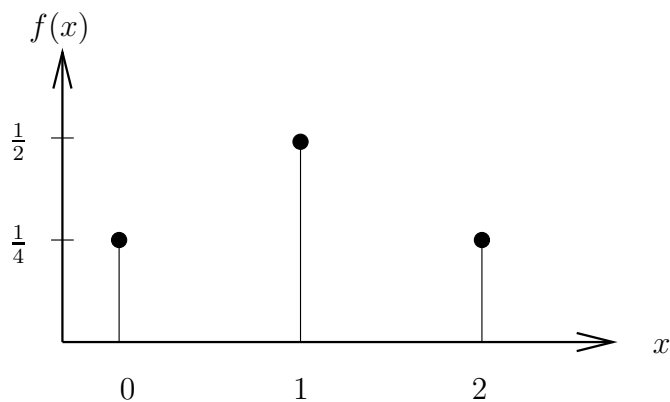
$$F_X(x) = \mathbb{P}(X \leq x) \text{ for } -\infty < x < \infty$$

Either the distribution function,  $F_X(x)$ , or the probability function,  $f_X(x)$ , is sufficient to specify the distribution of  $X$  completely.

**Example:** Let  $X \sim \text{Binomial}(2, \frac{1}{2})$ . 

$x$	0	1	2
$f_X(x) = \mathbb{P}(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

$$\text{So } F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ 0.25 & \text{if } 0 \leq x < 1 \\ 0.25 + 0.5 = 0.75 & \text{if } 1 \leq x < 2 \\ 0.25 + 0.5 + 0.25 = 1 & \text{if } x \geq 2. \end{cases}$$



$F_X(x)$  gives the cumulative probability up to and including point  $x$ .

So

$$F_X(x) = \sum_{y \leq x} f_X(y)$$

Note that  $F_X(x)$  is a step function: it jumps by amount  $f_X(y)$  at every point  $y$  with positive probability.

**Note:** As well as using the probability function to find the distribution function, we can also do the reverse:

$$\begin{aligned} f_X(x) = \mathbb{P}(X = x) &= \mathbb{P}(X \leq x) - \mathbb{P}(X \leq x - 1) \quad (\text{if } X \text{ takes integer values}) \\ &= F_X(x) - F_X(x - 1). \end{aligned}$$

## Properties of the distribution function

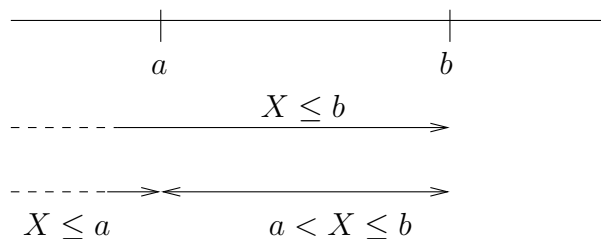
1)  $F(-\infty) = 0, F(+\infty) = 1$  (These are true because values are strictly between  $-\infty$  and  $\infty$ ).

2)  $F_X(x)$  is a non-decreasing function of  $x$ : that is,

$$\text{if } x_1 < x_2, \text{ then } F_X(x_1) \leq F_X(x_2).$$

3)  $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$  if  $b > a$ .

**Proof:**  $\mathbb{P}(X \leq b) = \mathbb{P}(X \leq a) + \mathbb{P}(a < X \leq b)$



$$\text{So } F_X(b) = F_X(a) + \mathbb{P}(a < X \leq b)$$

$$\Rightarrow F_X(b) - F_X(a) = \mathbb{P}(a < X \leq b).$$

4)  $F$  is right-continuous: that is,

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

## 2.5 Independent Random Variables

*Definition:* Random variables  $X$  and  $Y$  are statistically independent if

$$\mathbb{P}(X = x \text{ and } Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) \text{ for any } x \text{ and } y.$$

There are two useful results for independent random variables:

1) *If  $X$  and  $Y$  are independent random variables, then*

$$\mathbb{E}(XY) = (\mathbb{E}X)(\mathbb{E}Y)$$

2) *If  $X$  and  $Y$  are independent random variables, then*

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

These results are not necessarily true if  $X$  and  $Y$  are *not independent*.

Proof of (1) and (2): See Chapter 4.



# Chapter 3: Continuous Random Variables

---

## 3.1 Introduction

A continuous random variable can take values anywhere in some interval of the real line, e.g. in the interval  $[0, 1]$ . Quantities that are commonly modelled with continuous random variables are *time, weight, height, etc.*

Recall that, for a *discrete* random variable  $X$ , the probability function lists all values that  $X$  can take, and gives their probabilities:

$$\text{eg. } \begin{array}{c|ccc} x & 0 & 1 & 2 \\ \hline f_X(x) = \mathbb{P}(X = x) & 0.1 & 0.2 & 0.7 \end{array} \text{ etc.}$$

For a *continuous* random variable  $X$ , it is impossible to list all the values that  $X$  can take. It is also impossible to think of the probability that  $X$  takes any one specific value: e.g. even between the values 0.9999999 and 1.0000001 there are so many values that the probability of each is infinitesimally small. In fact, we write  $\mathbb{P}(X = x) = 0$  for any  $x$ , when  $X$  is continuous.

Thus, for continuous random variables, the probability function is *meaningless*.

Instead, for continuous random variables, we work with intervals:

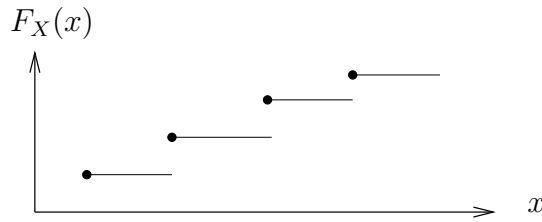
$$\text{eg. } \mathbb{P}(X = 1) = 0, \\ \text{but } \mathbb{P}(0.999 \leq X \leq 1.001) \text{ can be } > 0.$$

To find the probability that  $X$  lies in a given interval, we use the distribution function,  $F_X(x)$ , or its derivative, called the probability density function:  $f_X(x) = \frac{dF_X}{dx}$ .

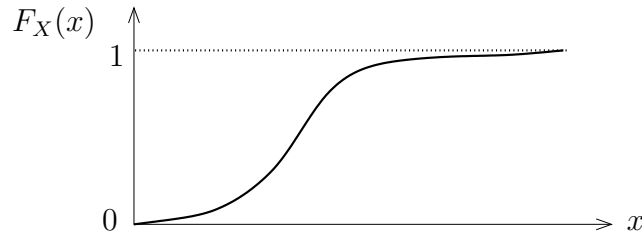
Recall the properties of the distribution function:

- i)  $F(-\infty) = 0, F(+\infty) = 1$ ;
- ii)  $F(x)$  is a *non-decreasing function of  $x$* ;
- iii)  $\mathbb{P}(a < X \leq b) = \mathbb{P}(X \in (a, b]) = F(b) - F(a)$ ;
- iv)  $F$  is *right continuous*.

When  $X$  is a discrete random variable,  $F_X(x)$  is a step function.



When  $X$  is a continuous r.v.,  $F_X(x)$  is a *continuous function*.



Property (iii) of  $F_X$  enables us to use the distribution function to calculate the probability that  $X$  lies in an interval:

$$\mathbb{P}(a < X \leq b) = \mathbb{P}(X \in (a, b]) = F_X(b) - F_X(a)$$

Note that when  $X$  is continuous,  $\mathbb{P}(X = a) = 0$ ,  
so  $\mathbb{P}(a \leq X \leq b) = \mathbb{P}(X = a) + \mathbb{P}(a < X \leq b) = \mathbb{P}(a < X \leq b)$ .  
So  $\mathbb{P}(X \in [a, b]) = \mathbb{P}(X \in (a, b]) = \mathbb{P}(X \in [a, b)) = \mathbb{P}(X \in (a, b))$ .

Thus we can write

$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a)$$

*Endpoints are not important for continuous r.v's (not true for discrete r.v's).*

The distribution function  $F_X(x)$  **characterizes** the random behaviour of  $X$ .

Another tool for characterizing the random behaviour of  $X$  is the **probability density function**,  $f_X(x)$ .

*Definition:* Let  $X$  be a continuous random variable with distribution function  $F_X(x)$ . The **probability density function (p.d.f.)** of  $X$  is defined as

$$f_X(x) = \frac{dF_X}{dx} = F'_X(x).$$

### Use of the probability density function to calculate probabilities

Let  $X$  be a continuous random variable with probability density function  $f_X(x)$ . Then

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(X \in [a, b]) = \int_a^b f_X(x) dx$$

This means that *we can calculate probabilities by integrating the p.d.f.*

#### Proof:

$$\int_a^b f_X(x) dx = \int_a^b \frac{dF_X}{dx} dx = \left[ F_X(x) \right]_a^b = F_X(b) - F_X(a) = \mathbb{P}(a \leq X \leq b).$$

**Note:** When  $X$  is discrete, we use the **probability function**,  $f_X(x) = \mathbb{P}(X = x)$ .

When  $X$  is continuous, we use the **probability density function**,

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx} \mathbb{P}(X \leq x).$$

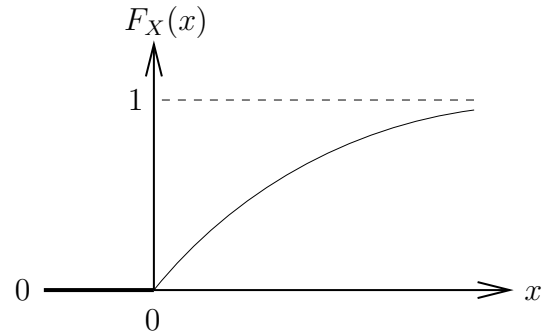
Both discrete and continuous r.v.s have the same definition for the distribution function,  $F_X(x) = \mathbb{P}(X \leq x)$ .

**Example 1:** Let  $F_X(x) = \begin{cases} 1 - e^{-x} & x \geq 0 \\ 0 & x < 0 \end{cases}$

Then  $F_X(-\infty) = 0$ ;

$$F_X(\infty) = 1 - e^{-\infty} = 1 - 0 = 1.$$

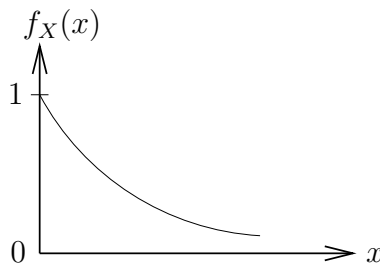
$F_X(x)$  is non-decreasing and continuous:



So  $F_X(x)$  is a **valid distribution function for a continuous r.v.  $X$** .

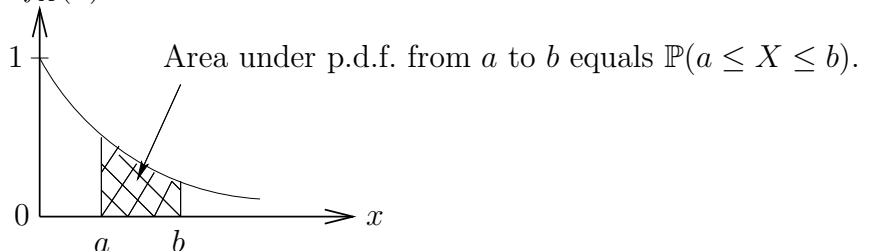
(In fact,  $X$  is said to have an Exponential(1) distribution: see later.)

Probability density function:  $f_X(x) = \frac{d}{dx}(1 - e^{-x}) = e^{-x}$  for  $x \geq 0$ .



We interpret this as follows:  $f_X(x)$

i)

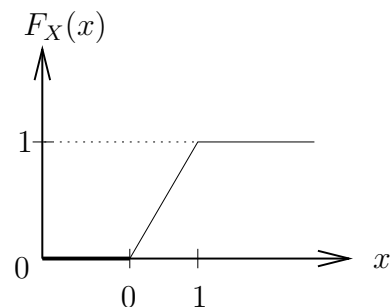


ii)  $X$  is more likely to take values close to 0 (where  $f_X(x)$  is larger), and less likely to take large values (where  $f_X(x)$  is smaller).

However, we can **NOT** say that  $\mathbb{P}(X = 0) = 1$ , even though  $f_X(0) = 1$ . The probability density function is never used in this way.

**Example 2:** Let  $F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } 0 \leq x \leq 1 \\ 1 & \text{for } x > 1. \end{cases}$

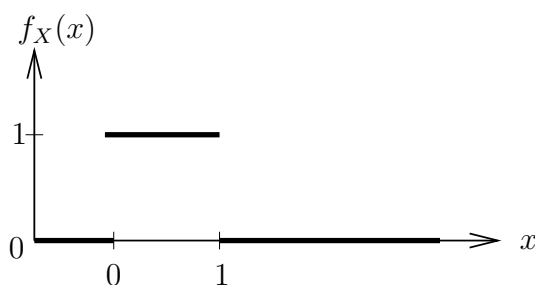
Then  $F_X(-\infty) = 0$ ;  $F_X(\infty) = 1$ ; and  $F_X(x)$  is non-decreasing and continuous:



So  $F_X(x)$  is a valid distribution function for a continuous r.v.  $X$ .

(In fact,  $X$  is said to have a Uniform[0,1] distribution: see later.)

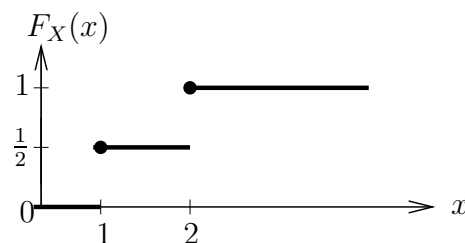
Probability density function :  $\begin{cases} f_X(x) = \frac{dF_X}{dx} = \frac{d}{dx}(x) = 1 & \text{for } 0 \leq x \leq 1, \\ f_X(x) = 0 & \text{when } x < 0 \text{ or } x > 1. \end{cases}$



Interpretation:  $X$  is equally likely to take any value between 0 and 1.

[The p.d.f. gives an intuitive impression of what the distribution looks like.]

**Example 3:** Let  $F_X(x) = \begin{cases} 0 & \text{for } x < 1, \\ 0.5 & \text{for } 1 \leq x < 2, \\ 1 & \text{for } x \geq 2. \end{cases}$



$F_X(x)$  is not continuous, so is not a distribution function for a continuous random variable. It is a distribution function for a discrete random variable with probability function:

$x$	1	2
$f_X(x)$	0.5	0.5

## Properties of the probability density function

If  $f(x)$  is the p.d.f. for a continuous random variable, then

i)  $f(x) \geq 0$  for all  $x$ .

ii)  $\int_{-\infty}^{\infty} f(x) dx = 1$ .

iii) **Distribution function**,  $F(x) = \int_{-\infty}^x f(y) dy$ .

iv)  $\mathbb{P}(a < X \leq b) = \mathbb{P}(a < X < b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a \leq X \leq b) = \int_a^b f(x) dx$ .

### Proof:

i) Because the distribution function  $F(x)$  is non-decreasing, its derivative,  $f(x)$ , is always *non-negative*.

ii) By the Fundamental Theorem of Calculus,  $\int_a^b f(x) dx = F(b) - F(a)$ .

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

[This is saying that the total area under the p.d.f. curve is equal to the total probability that  $X$  takes a value between  $-\infty$  and  $+\infty$ , which is 1.]

iii) By the Fundamental Theorem of Calculus,

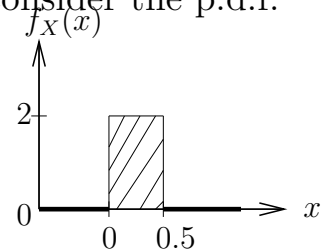
$$\int_{-\infty}^x f(y) dy = F(x) - F(-\infty) = F(x) - 0 = F(x).$$

iv)  $\mathbb{P}(a < X \leq b) = F(b) - F(a) = \int_a^b f(x) dx$ , by previous arguments.

It is more difficult to prove rigorously that  $\mathbb{P}(X = a) = 0$ , in order to show that  $\mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b)$ , etc. This is beyond the scope of this course.  $\square$

**Note:** It is not necessarily true that  $f(x) \leq 1$  for all  $x$ : e.g. consider the p.d.f.

$$f(x) = \begin{cases} 0 & \text{for } x < 0, \\ 2 & \text{for } 0 \leq x \leq 0.5, \\ 0 & \text{for } x > 0.5. \end{cases}$$



This is a valid p.d.f.:  $\int_{-\infty}^{\infty} f(x) dx = \int_0^{0.5} 2 dx = [2x]_0^{0.5} = 1.$

## Expected value of a continuous random variable

*Definition:* The expected value, or expectation, or mean, of a continuous r.v.  $X$  is defined as

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

where  $f_X(x)$  is the probability density function.

Similarly, for any (nice) function  $g(X)$ ,

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

**Note:** Compare these with the definitions for discrete random variables:

$\mathbb{E}(X) = \sum_x x f_X(x)$ ,  $\mathbb{E}(g(X)) = \sum_x g(x) f_X(x)$ , where  $f_X(x)$  is the probability function of  $X$ .

The expectation of a continuous random variable can be manipulated in exactly the same way as that of a discrete random variable:

**Theorem 3.1:** If  $a$  and  $b$  are constants, and  $g(x)$ ,  $h(x)$  are functions, then

i)  $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$

ii)  $\mathbb{E}(ag(X) + b) = a\mathbb{E}(g(X)) + b$

iii)  $\mathbb{E}(ag(X) + bh(X)) = a\mathbb{E}(g(X)) + b\mathbb{E}(h(X)).$

**Proof:** (part (iii): parts (i) and (ii) are special cases).

$$\begin{aligned}\mathbb{E}(ag(X) + bh(X)) &= \int_{-\infty}^{\infty} (ag(x) + bh(x))f_X(x)dx \\ &= a \int_{-\infty}^{\infty} g(x)f_X(x)dx + b \int_{-\infty}^{\infty} h(x)f_X(x)dx \\ &= a\mathbb{E}(g(X)) + b\mathbb{E}(h(X)). \quad \square\end{aligned}$$

Expectation is a linear operator exactly because integration is.

### Variance of a continuous random variable

Variance was defined in Chapter 2 as

$$\text{Var}(X) = \sigma_X^2 = \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}(X^2) - \mu_X^2 = \mathbb{E}(X^2) - (\mathbb{E}X)^2$$

For a continuous random variable, we can either compute the variance using

$$\text{Var}(X) = \mathbb{E}[(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)dx,$$

or

$$\text{Var}(X) = \mathbb{E}(X^2) - \mu_X^2 = \int_{-\infty}^{\infty} x^2 f_X(x)dx - \mu_X^2.$$

The second expression is usually easier (although not always).

The properties of variance for continuous r.v.s are exactly the same as for discrete r.v.s. The proof of the following theorem is exactly the same as that for Theorem 2.3.

**Theorem 3.2:** If  $a$  and  $b$  are constants, and  $g(x)$  is a function, then

- i)  $\text{Var}(aX + b) = a^2\text{Var}(X)$
- ii)  $\text{Var}(ag(X) + b) = a^2\text{Var}(g(X)).$

**Proof:** see Theorem 2.3. □



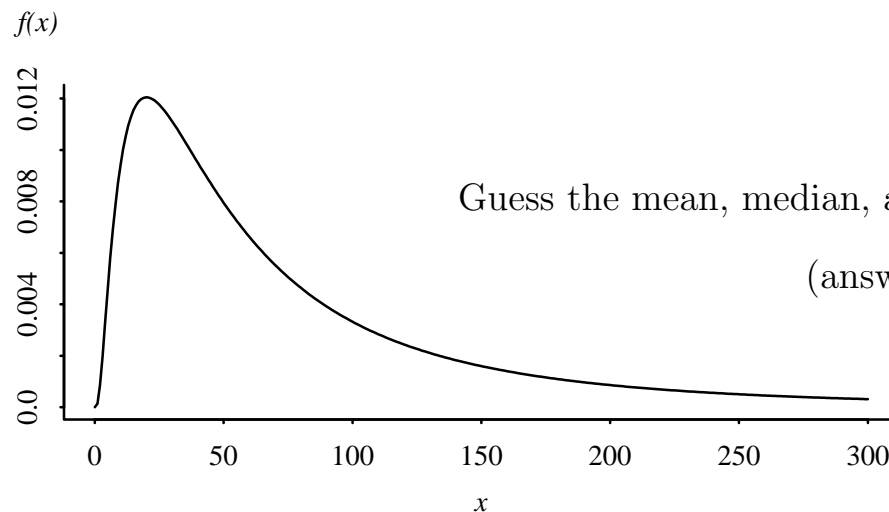
## Interlude: Guess the Mean, Median, and Variance

For any distribution:

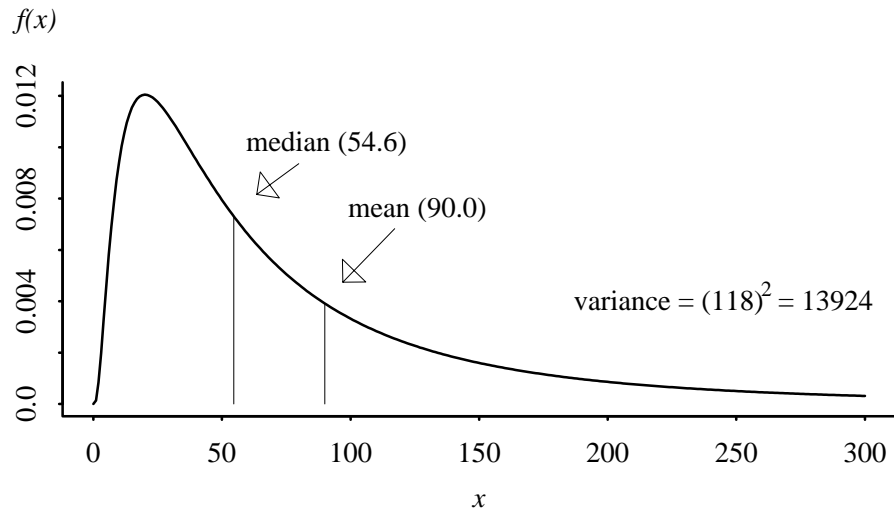
- the **mean** is the **average** that would be obtained if a large number of observations were drawn from the distribution;
- the **median** is the **half-way point** of the distribution: every observation has a 50-50 chance of being above the median or below the median;
- the **variance** is the **average squared distance** of an observation from the mean.

Given the probability density function of a distribution, we should be able to guess roughly the distribution mean, median, and variance . . . but it isn't easy! Have a go at the examples below. As a hint:

- the **mean** is the **balance-point** of the distribution. Imagine that the p.d.f. is made of cardboard and balanced on a rod. The mean is the point where the rod would have to be placed for the cardboard to balance.
- the **median** is the half-way point, so it divides the p.d.f. into two equal areas of 0.5 each.
- the **variance** is the average **squared** distance of observations from the mean; so to get a **rough** guess (not exact), it is easiest to guess an average distance from the mean and square it.



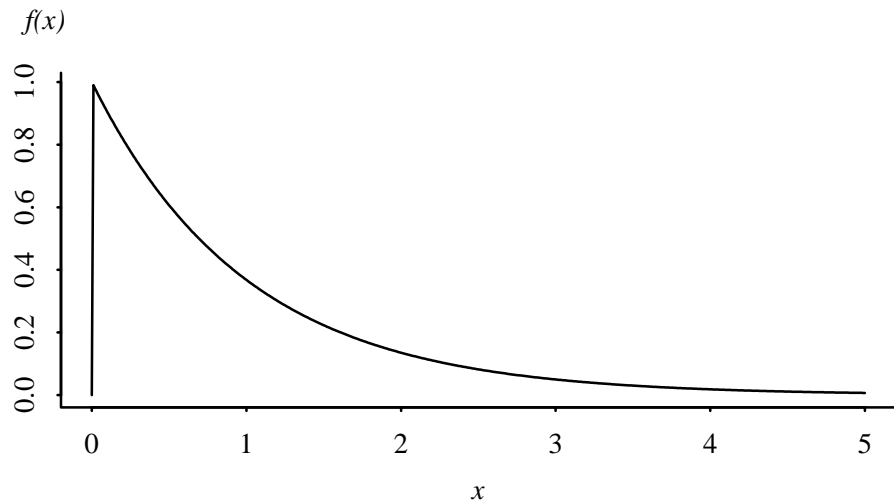
**Answers:**



**Notes:** The mean is larger than the median. This always happens when the distribution has a long right tail (positive skew) like this one.

The variance is *huge* . . . but when you look at the numbers along the horizontal axis, it is quite believable that the average squared distance of an observation from the mean is  $118^2$ . Out of interest, the distribution shown is a Lognormal distribution.

**Example 2:** Try the same again with the example below. Answers are written below the graph.



**Answers:** Median = 0.693; Mean=1.0; Variance=1.0.

## 3.2 Examples of Continuous Distributions

---

In Chapter 2 we looked at several examples of discrete distributions. Most of these were mathematically exact distributions arising from fully-specified situations: for example, the Binomial, Geometric, and Negative Binomial distributions (from sequences of Bernoulli trials); the Hypergeometric distribution (sampling without replacement from a finite population); and the Poisson distribution (from the Poisson process). In addition, these distributions can also be used as ‘subjective models’ in other situations that are not mathematically exact. The Poisson distribution and the Negative Binomial distribution are both widely used as subjective models, simply because they have shape and variance properties that could realistically describe many real-world situations.

In the case of continuous distributions, it is quite rare to have mathematically exact situations, and in almost all cases the distributions are used primarily as ‘subjective models’. (Examples of mathematically exact situations are the Exponential distribution from the Poisson process (see later), and the Normal distribution from the Central Limit Theorem, but these are quite unusual.)

To form a subjective model of a situation, we:

- select a probability distribution whose properties could reasonably fit the situation;
- use observed data to estimate the parameters of the probability distribution (e.g. the parameter  $\lambda$  for a Poisson distribution, or  $k$  and  $p$  for the Negative Binomial distribution.)

The result is the ‘best’ set of parameters, *assuming that the model is correct in the first place*. Choosing a good model (probability distribution) is a fundamentally important part of the procedure, but one which is often overlooked in the applied sciences. For example, many scientists automatically assume that their observations follow a Normal distribution (symmetric and bell-shaped), when this is highly inappropriate.

The aim of this section is to introduce some continuous distributions that are widely used in modelling, to show how different distributions provide flexibility in shape and properties. Although the final conclusion obtained from a statistical model is usually the *mean* of the selected distribution, it is the *shape* and *variance* that are most important (and most often forgotten) when selecting a good model.

## 1. Uniform Distribution

$X$  has a Uniform distribution on the interval  $[a, b]$  if  $X$  is equally likely to fall anywhere in the interval  $[a, b]$ .

We write  $X \sim \text{Uniform}[a, b]$ , or  $X \sim U[a, b]$ .

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

### Probability density function, $f_X(x)$

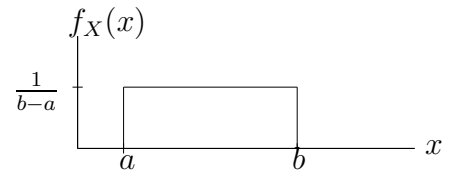
If  $X \sim U[a, b]$ , then

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

(Check area under p.d.f. is 1:

area of rectangle = base  $\times$  height =  $(b - a) \times \frac{1}{b-a} = 1$ .)

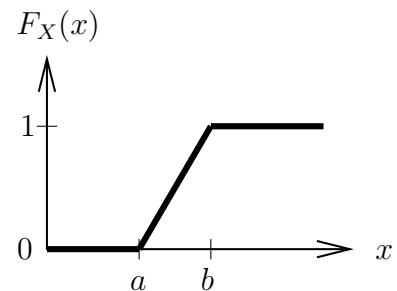
### Distribution function, $F_X(x)$



$$\begin{aligned} F_X(x) &= \int_{-\infty}^x f_Y(y) dy = \int_a^x \frac{1}{b-a} dy \quad \text{if } a \leq x \leq b \\ &= \left[ \frac{y}{b-a} \right]_a^x \\ &= \frac{x-a}{b-a} \quad \text{if } a \leq x \leq b. \end{aligned}$$

Thus

$$F_X(x) = \begin{cases} 0 & \text{if } x < a, \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b, \\ 1 & \text{if } x > b. \end{cases}$$



## Mean and variance:

$$\text{If } X \sim U[a, b], \quad \mathbb{E}(X) = \frac{a+b}{2}, \quad \text{Var}(X) = \frac{(b-a)^2}{12}$$

**Proof:**  $\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_a^b x \left( \frac{1}{b-a} \right) dx = \frac{1}{b-a} \left[ \frac{x^2}{2} \right]_a^b$

$$= \left( \frac{1}{b-a} \right) \cdot \frac{1}{2} (b^2 - a^2)$$
$$= \left( \frac{1}{b-a} \right) \frac{1}{2} (b-a)(b+a)$$
$$= \frac{a+b}{2}.$$

$$\text{Var}(X) = \mathbb{E}[(X - \mu_X)^2] = \int_a^b \frac{(x - \mu_X)^2}{b-a} dx = \frac{1}{b-a} \left[ \frac{(x - \mu_X)^3}{3} \right]_a^b$$
$$= \left( \frac{1}{b-a} \right) \left\{ \frac{(b - \mu_X)^3 - (a - \mu_X)^3}{3} \right\}$$

But  $\mu_X = \mathbb{E}X = \frac{a+b}{2}$ , so  $b - \mu_X = \frac{b-a}{2}$  and  $a - \mu_X = \frac{a-b}{2}$ .  
So,

$$\text{Var}(X) = \left( \frac{1}{b-a} \right) \left\{ \frac{(b-a)^3 - (a-b)^3}{2^3 \times 3} \right\} = \frac{(b-a)^3 + (b-a)^3}{(b-a) \times 24}$$
$$= \frac{(b-a)^2}{12}. \quad \square$$

**Example:** let  $X \sim \text{Uniform}[0, 1]$ . Then

$$f_X(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

$$\mu_X = \mathbb{E}(X) = \frac{0+1}{2} = \frac{1}{2} \quad (\text{half-way through interval } [0, 1]).$$

$$\sigma_X^2 = \text{Var}(X) = \frac{1}{12}(1-0)^2 = \frac{1}{12}.$$

## 2. Exponential Distribution

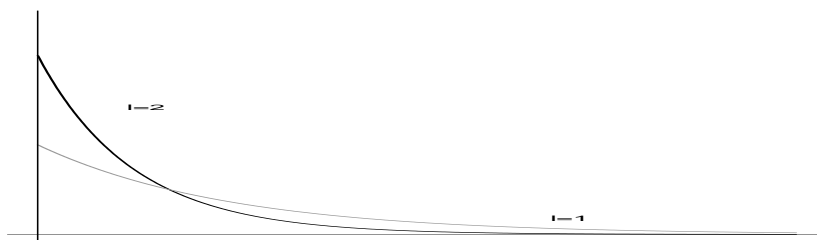
The Exponential distribution has *one parameter*,  $\lambda$ , *which must be positive*.

We write  $X \sim \text{Exponential}(\lambda)$ , or  $X \sim \text{Exp}(\lambda)$ .

### Probability density function, $f_X(x)$

For  $X \sim \text{Exp}(\lambda)$ ,

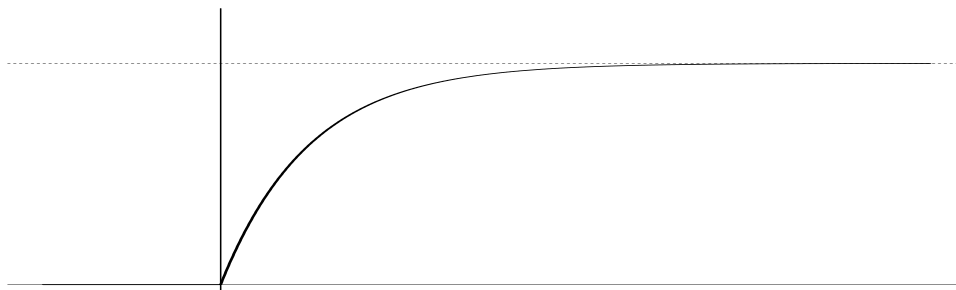
$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0. \end{cases}$$



### Distribution function, $F_X(x)$

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - e^{-\lambda x} & \text{for } x \geq 0. \end{cases}$$

Exercise: check  $F_X(x) = \int_{-\infty}^x f_X(y) dy = 1 - e^{-\lambda x}$ .



## Mean and variance:

For  $X \sim \text{Exp}(\lambda)$ ,

$$\mathbb{E}(X) = \frac{1}{\lambda} \text{ and } \text{Var}(X) = \frac{1}{\lambda^2}.$$

**Proof:**  $\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^{\infty} x \lambda e^{-\lambda x} dx.$

*Integration by parts: recall that  $\int u \frac{dv}{dx} dx = uv - \int v \frac{du}{dx} dx.$*

Let  $u = x$ , so  $\frac{du}{dx} = 1$ , and let  $\frac{dv}{dx} = \lambda e^{-\lambda x}$ , so  $v = -e^{-\lambda x}.$

$$\begin{aligned} \text{Then } \mathbb{E}(X) &= \int_0^{\infty} x \lambda e^{-\lambda x} dx = \int_0^{\infty} u \frac{dv}{dx} dx \\ &= [uv]_0^{\infty} - \int_0^{\infty} v \frac{du}{dx} dx \\ &= [-x e^{-\lambda x}]_0^{\infty} - \int_0^{\infty} (-e^{-\lambda x}) dx \\ &= 0 + \left[ \frac{-1}{\lambda} e^{-\lambda x} \right]_0^{\infty} \\ &= \frac{-1}{\lambda} \times 0 - \left( \frac{-1}{\lambda} \times e^0 \right) \\ \therefore \mathbb{E}(X) &= \frac{1}{\lambda}. \end{aligned}$$

**Variance:**  $\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \mathbb{E}(X^2) - \frac{1}{\lambda^2}.$

Now  $\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx.$

Let  $u = x^2$ , so  $\frac{du}{dx} = 2x$ , and let  $\frac{dv}{dx} = \lambda e^{-\lambda x}$ , so  $v = -e^{-\lambda x}.$

$$\begin{aligned} \text{Then } \mathbb{E}(X^2) &= [uv]_0^{\infty} - \int_0^{\infty} v \frac{du}{dx} dx = [-x^2 e^{-\lambda x}]_0^{\infty} + \int_0^{\infty} 2x e^{-\lambda x} dx \\ &= 0 + \frac{2}{\lambda} \int_0^{\infty} \lambda x e^{-\lambda x} dx \\ &= \frac{2}{\lambda} \times \mathbb{E}(X) = \frac{2}{\lambda^2}. \end{aligned}$$

So

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}X)^2 \\ &= \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 \\ \text{Var}(X) &= \frac{1}{\lambda^2}. \quad \square\end{aligned}$$

### Exponential Distribution arising from the Poisson Process

Suppose that  $\{Y_t : t > 0\}$  forms a Poisson process with rate  $\lambda$ .

[Recall: this means that  $Y_t = \#$  events to have occurred by time  $t$ , and  $Y_t \sim \text{Poisson}(\lambda t)$ .]

Let  $X =$  time we have to wait from time 0 to time of the first event.

What is the distribution of  $X$ ?

To find this, we can calculate the distribution function of  $X$ :

$$\begin{aligned}F_X(x) = \mathbb{P}(X \leq x) &= 1 - \mathbb{P}(X > x) \\ &= 1 - \mathbb{P}(\text{have to wait longer than time } x \text{ before the first event}) \\ &= 1 - \mathbb{P}(\text{there are no events in time from } 0 \text{ to } x) \\ &= 1 - \mathbb{P}(Y_x = 0) \\ &\quad (\text{where } Y_x = \# \text{ events to have occurred by time } x, \text{ and } Y_x \sim \text{Poisson}(\lambda x)) \\ &= 1 - \frac{(\lambda x)^0}{0!} e^{-\lambda x}\end{aligned}$$

$$F_X(x) = 1 - e^{-\lambda x} \quad \text{if } x \geq 0.$$

Clearly,  $F_X(x) = 0$  if  $x < 0$ .

Thus  $F_X(x)$  is the distribution function of the Exponential( $\lambda$ ) distribution, and so  $X \sim \text{Exponential}(\lambda)$ .



*So if  $\{Y_t\}_{t>0}$  is a Poisson process with rate  $\lambda$ , then  $X = (\text{time taken until first event}) \sim \text{Exponential}(\lambda)$ .*

**Note:** 1) We do not have to start at time  $t = 0$ . It can be shown that if

$X = (\text{time taken from time } s \text{ to next subsequent event}),$  for any  $s > 0$ ,  
or

$X = (\text{time taken from } k\text{th event to } (k + 1)\text{th event}),$  for  $k = 1, 2, 3, \dots$ ,  
then  $X \sim \text{Exponential}(\lambda)$ .

Conversely, if the waiting time between events is  $\text{Exponential}(\lambda)$ , then the events form a Poisson process with rate  $\lambda$ .

**Note:** 2) The Poisson process is used to model many situations, e.g. customers arriving at a shop, earthquakes, volcanic eruptions, outbreaks of war or disease, and so on. The exponential distribution can therefore be used to model the waiting time between these events, e.g. time before the next customer arrives, or time before the next earthquake, etc.

## The Memoryless Property of the Exponential Distribution

The Exponential distribution is famous for its property of ‘memorylessness’.

Suppose we have already waited time  $x_0$  for an event.  
How much longer do we have to wait?

*Let  $X \sim \text{Exponential}(\lambda)$ . Then*

$$\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x) = 1 - F_X(x) = e^{-\lambda x}.$$

*Looking for the probability of waiting at least  $x$  more time, given that we have waited  $x_0$  so far.*

$$\mathbb{P}(X > x_0 + x | X > x_0) = \frac{\mathbb{P}(X > x_0 + x)}{\mathbb{P}(X > x_0)} = \frac{e^{-\lambda(x_0+x)}}{e^{-\lambda x_0}} = e^{-\lambda x} \quad \text{if } x \geq 0.$$

But this is equal to  $\mathbb{P}(X > x)$ : so  $\mathbb{P}(X > x_0 + x | X > x_0) = \mathbb{P}(X > x)$ ,

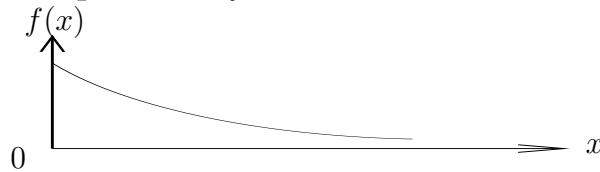
$$\begin{aligned} \text{i.e. } \mathbb{P}(\text{wait at least } x \text{ more time, given we have already waited } x_0 \text{ time}) \\ = \mathbb{P}(\text{wait at least } x \text{ time, starting from } 0). \end{aligned}$$

We say that the Exponential distribution is memoryless: *it forgets the time already waited*.

For example, if bus arrivals follow a memoryless distribution, then even if you have already waited 5 hours for a bus, you still expect to wait the same amount *more* time as you did when you first started.

Similarly, if the lifetime of a lightbulb has a memoryless distribution, then given that the lightbulb has already lasted 2 years, it still has exactly the same lifetime distribution as a new lightbulb.

**Notes:** 1) It is not necessarily desirable for a lifetime distribution to be memoryless. “Old is as good as new”, but put a different way, “new is as bad as old”. A memoryless lightbulb is quite likely to fail almost immediately.



2) The Exponential distribution is the *only* memoryless distribution.

---

### 3. Gamma Distribution

The Gamma distribution has *two parameters,  $k$  and  $\lambda$ , where  $k > 0$  and  $\lambda > 0$* . We write  $X \sim \text{Gamma}(k, \lambda)$ .

#### Probability density function, $f_X(x)$

For  $X \sim \text{Gamma}(k, \lambda)$ ,

$$f_X(x) = \begin{cases} \frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

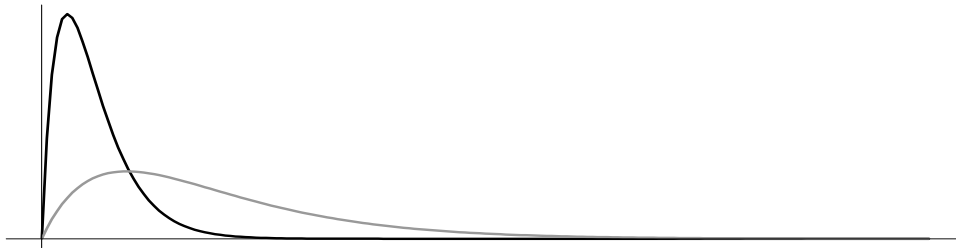
Here,  $\Gamma(k)$ , called the **Gamma function of  $k$** , is simply a constant that ensures  $f_X(x)$  integrates to 1, i.e.  $\int_0^\infty f_X(x)dx = 1$ . (see below).

### Gamma p.d.f.s

$k = 1$



$k = 2$



*Notice: right skew  
(long right tail);  
flexibility in shape  
controlled by the 2  
parameters*

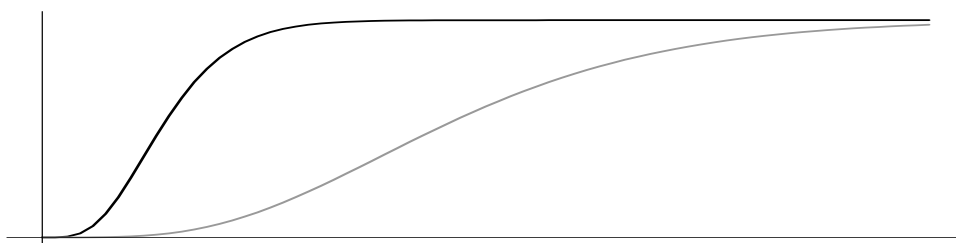
$k = 5$



### Distribution function, $F_X(x)$

There is no closed form for the distribution function of the Gamma distribution. If  $X \sim \text{Gamma}(k, \lambda)$ , then  $F_X(x)$  can only be calculated **by computer**.

$k = 5$



## The Gamma Function, $\Gamma(k)$

Recall that  $\Gamma(k)$  is a constant that is defined to ensure that

$$\int_0^{\infty} f_X(x) dx = \int_0^{\infty} \frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x} dx = 1.$$

*Definition:* For any  $k > 0$ , the Gamma function of  $k$  is defined as

$$\Gamma(k) = \int_0^{\infty} y^{k-1} e^{-y} dy$$

Check that this makes  $\int_0^{\infty} f_X(x) dx = 1$ :

$$\begin{aligned} \int_0^{\infty} f_X(x) dx &= \int_0^{\infty} \frac{1}{\Gamma(k)} \lambda^k x^{k-1} e^{-\lambda x} dx \\ &= \frac{\int_0^{\infty} \lambda (\lambda x)^{k-1} e^{-\lambda x} dx}{\Gamma(k)}. \end{aligned}$$

Change variable: let  $y = \lambda x$ , then  $\frac{dx}{dy} = \frac{1}{\lambda}$ .

Then

$$\begin{aligned} \int_0^{\infty} f_X(x) dx &= \frac{\int_0^{\infty} \lambda y^{k-1} e^{-y} \left(\frac{1}{\lambda}\right) dy}{\Gamma(k)} \\ &= \frac{\int_0^{\infty} y^{k-1} e^{-y} dy}{\int_0^{\infty} y^{k-1} e^{-y} dy} \quad \text{by definition of } \Gamma(k), \\ &= 1. \quad \text{as required.} \quad \square \end{aligned}$$

## Properties of the Gamma function, $\Gamma(k)$

1.  $\Gamma(k) = (k - 1)\Gamma(k - 1)$  for all  $k \geq 1$ .
2. When  $k$  is an integer,  $\Gamma(k) = (k - 1)!$
3.  $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ .

### Proof:

1. 
$$\begin{aligned}\Gamma(k) &= \int_0^{\infty} y^{k-1} e^{-y} dy \\ &= \left[ -y^{k-1} e^{-y} \right]_0^{\infty} + \int_0^{\infty} (k-1)y^{k-2} e^{-y} dy \\ &= 0 + (k-1)\Gamma(k-1).\end{aligned}$$

2. For  $k \in \mathbb{Z}$ , we have:

$$\begin{aligned}\Gamma(k) &= (k-1)\Gamma(k-1) \\ &= (k-1)(k-2)\Gamma(k-2) \\ &= \vdots \\ &= (k-1)(k-2)\dots(3)(2)(1)\Gamma(1).\end{aligned}$$

Now  $\Gamma(1) = \int_0^{\infty} y^0 e^{-y} dy = \left[ -e^{-y} \right]_0^{\infty} = 1$ , so  $\Gamma(k) = (k-1)!$ .

3. Proof not required. □

## Mean and variance of the Gamma distribution:

For  $X \sim \text{Gamma}(k, \lambda)$ ,

$\mathbb{E}(X) = \frac{k}{\lambda}$ and $\text{Var}(X) = \frac{k}{\lambda^2}$
---

Proof that  $\mathbb{E}(X) = \frac{k}{\lambda}$  and  $\text{Var}(X) = \frac{k}{\lambda^2}$

$$\begin{aligned}\mathbb{E}X &= \int_0^{\infty} x f_X(x) dx = \int_0^{\infty} x \cdot \frac{\lambda^k x^{k-1}}{\Gamma(k)} e^{-\lambda x} dx \\ &= \frac{\int_0^{\infty} (\lambda x)^k e^{-\lambda x} dx}{\Gamma(k)} \\ &= \frac{\int_0^{\infty} y^k e^{-y} \left(\frac{1}{\lambda}\right) dy}{\Gamma(k)} \quad (\text{letting } y = \lambda x, \frac{dx}{dy} = \frac{1}{\lambda}) \\ &= \frac{1}{\lambda} \cdot \frac{\Gamma(k+1)}{\Gamma(k)} \\ &= \frac{1}{\lambda} \cdot \frac{k \Gamma(k)}{\Gamma(k)} \quad \text{by Property (1) overleaf,} \\ &= \frac{k}{\lambda}.\end{aligned}$$

$$\begin{aligned}\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 &= \int_0^{\infty} x^2 f_X(x) dx - \frac{k^2}{\lambda^2} \\ &= \int_0^{\infty} \frac{x^2 \lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)} dx - \frac{k^2}{\lambda^2} \\ &= \frac{\int_0^{\infty} \left(\frac{1}{\lambda}\right) (\lambda x)^{k+1} e^{-\lambda x} dx}{\Gamma(k)} - \frac{k^2}{\lambda^2} \\ &= \frac{1}{\lambda^2} \cdot \frac{\int_0^{\infty} y^{k+1} e^{-y} dy}{\Gamma(k)} - \frac{k^2}{\lambda^2} \quad \left[ \text{where } y = \lambda x, \frac{dx}{dy} = \frac{1}{\lambda} \right] \\ &= \frac{1}{\lambda^2} \cdot \frac{\Gamma(k+2)}{\Gamma(k)} - \frac{k^2}{\lambda^2} \\ &= \frac{1}{\lambda^2} \frac{(k+1)k \Gamma(k)}{\Gamma(k)} - \frac{k^2}{\lambda^2} \\ &= \frac{k}{\lambda^2}. \quad \square\end{aligned}$$

**Note:** The  $\text{Gamma}(k, 1)$  distribution is sometimes called the *unscaled Gamma distribution with parameter  $k$* .

and the  $\text{Gamma}(k, \lambda)$  distribution is sometimes called the *scaled Gamma distribution* with parameters  $k$  and  $\lambda$ .

If  $X \sim \text{Gamma}(k, \lambda)$ , it can be shown that  $\lambda X \sim \underline{\text{Gamma}(k, 1)}$ .

## Relationship between the Gamma distribution and the Exponential distribution

The  $\text{Gamma}(k, \lambda)$  distribution arises in nature as the *sum of  $k$  independent Exponential r.v.'s*:

*that is, if  $X_1, \dots, X_k \sim \text{Exponential}(\lambda)$  and are independent then  $X_1 + X_2 + \dots + X_k \sim \text{Gamma}(k, \lambda)$ .*

This is proved later in the course.

**Special Case:** When  $k = 1$ ,

$\text{Gamma}(1, \lambda) = \text{Exponential}(\lambda)$  (the sum of a single Exponential r.v.)

We can see this immediately, as the p.d.f. of  $\text{Gamma}(1, \lambda)$  is

$$f(x) = \frac{\lambda^1}{\Gamma(1)} x^{1-1} e^{-\lambda x} = \lambda e^{-\lambda x}, \text{ which is the same as the pdf of } \text{Exp}(\lambda).$$

## Gamma distribution arising from the Poisson process

Recall that the waiting time between events in a Poisson process with rate  $\lambda$  has the *Exponential( $\lambda$ ) distribution*.

That is, if  $X_i$  = time waited between event  $i - 1$  and event  $i$ , then  $X_i \sim \text{Exp}(\lambda)$ .

Now the time waited from time 0 to the time of the  $k$ th event is

$X_1 + X_2 + \dots + X_k$ , *the sum of  $k$  independent Exponential( $\lambda$ ) r.v's.*

Thus the time waited until the  $k$ th event in a Poisson process with rate  $\lambda$  has the *Gamma( $k, \lambda$ ) distribution.*

[There are some similarities between the Exponential( $\lambda$ ) distribution and the (discrete) Geometric( $p$ ) distribution. Both distributions describe the 'waiting time' before an event. In the same way, the Gamma( $k, \lambda$ ) distribution is similar to the (discrete) Negative Binomial( $k, p$ ) distribution, as they both describe the 'waiting time' before the  $k$ th event.]

### Relationship between the Gamma distribution and the Chi-squared distribution

The Chi-squared distribution with  $\nu$  degrees of freedom,  $\chi_\nu^2$ , is a special case of the Gamma distribution.

$$\chi_\nu^2 = \text{Gamma}(k = \frac{\nu}{2}, \lambda = \frac{1}{2}).$$

So if  $Y \sim \chi_\nu^2$ , then  $\mathbb{E}(Y) = \frac{k}{\lambda} = \nu$ , and  $\text{Var}(Y) = \frac{k}{\lambda^2} = 2\nu$ .

---

### 4. Beta Distribution

The Beta distribution has two parameters,  $\alpha$  and  $\beta$ . We write  $X \sim \text{Beta}(\alpha, \beta)$ .

P.d.f.

$$f(x) = \begin{cases} \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

The function  $B(\alpha, \beta)$  is the *Beta function* and is defined by the integral

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx, \quad \text{for } \alpha > 0, \beta > 0.$$

It can be shown that 
$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

---



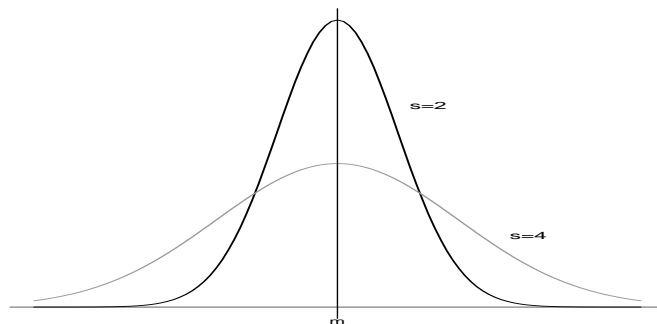
## 5. Normal Distribution

The Normal distribution is the familiar bell-shaped distribution. It has two parameters, *the mean,  $\mu$ , and the variance,  $\sigma^2$ .*

We write  $X \sim \text{Normal}(\mu, \sigma^2)$  or  $X \sim N(\mu, \sigma^2)$ .

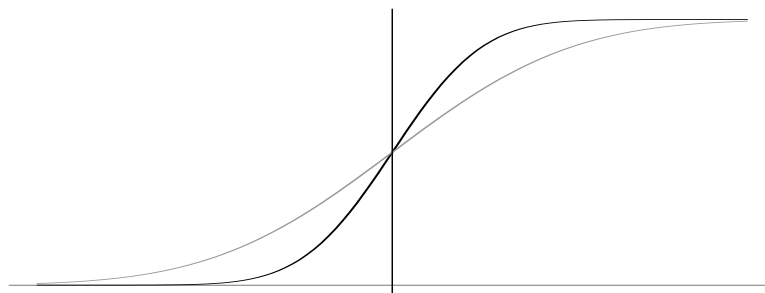
### Probability density function, $f_X(x)$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\{-(x-\mu)^2/2\sigma^2\}} \quad \text{for } -\infty < x < \infty \text{ and } -\infty < \mu < \infty, \sigma^2 > 0.$$



### Distribution function, $F_X(x)$

There is no closed form for the distribution function of the Normal distribution. If  $X \sim \text{Normal}(\mu, \sigma^2)$ , then  $F_X(x)$  can only be calculated *by computer*.



**Note:** To show that  $\int_{-\infty}^{\infty} f_X(x) dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\{-(x-\mu)^2/(2\sigma^2)\}} dx = 1$ ,

the following result is used:

FACT:  $\int_{-\infty}^{\infty} e^{-y^2} dy = \sqrt{\pi}.$

(Proved in Calculus courses.)

### Mean and Variance

For  $X \sim \text{Normal}(\mu, \sigma^2)$ ,  $\mathbb{E}(X) = \mu, \text{Var}(X) = \sigma^2$ .

### Proof:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} dx$$

[Let  $z = \frac{x-\mu}{\sigma}$  : then  $x = \sigma z + \mu$  and  $\frac{dx}{dz} = \sigma$ .]

**Thus** 
$$\mathbb{E}(X) = \int_{-\infty}^{\infty} (\sigma z + \mu) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-z^2/2} \cdot \sigma dz$$

$$= \underbrace{\int_{-\infty}^{\infty} \frac{\sigma z}{\sqrt{2\pi}} \cdot e^{-z^2/2} dz}_{\substack{\text{this is an odd function of } z \\ \text{(i.e. } g(-z) = -g(z)\text{), so it} \\ \text{integrates to 0 over range} \\ -\infty \text{ to } \infty.}} + \mu \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz}_{\text{p.d.f. of } N(0, 1) \text{ integrates to 1.}}$$

$$\begin{aligned} \therefore \mathbb{E}(X) &= 0 + \mu \times 1 \\ &= \mu. \end{aligned}$$

For  $\text{Var}(X)$ ,

$$\begin{aligned}\text{Var}(X) &= E\{(X - \mu)^2\} \\ &= \int_{-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} dx \\ &= \sigma^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} z^2 e^{-z^2/2} dz \quad \left(\text{putting } z = \frac{x - \mu}{\sigma}\right) \\ &= \sigma^2 \left\{ \frac{1}{\sqrt{2\pi}} \left[ -ze^{-z^2/2} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \right\} \quad (\text{integration by parts}) \\ &= \sigma^2 \{0 + 1\} \\ &= \sigma^2. \quad \square\end{aligned}$$

## Linear transformations and Sums of Normal random variables

1. If  $X \sim \text{Normal}(\mu, \sigma^2)$ , then for any constants  $a$  and  $b$ ,

$$\underline{aX + b \sim \text{Normal}(a\mu + b, a^2\sigma^2)}.$$

In particular, if  $Z = \left(\frac{X - \mu}{\sigma}\right)$ , then  $Z \sim \text{Normal}(0, 1)$ .

$\left(\text{Prove this by putting } a = \frac{1}{\sigma} \text{ and } b = -\frac{\mu}{\sigma}.\right)$

$Z \sim \text{Normal}(0, 1)$  is referred to as the *standard Normal random variable*.

Proof: see section 3.3.

2. If  $X_1, X_2, \dots, X_n$  are *independent*, and  $X_i \sim \text{Normal}(\mu_i, \sigma_i^2)$  for  $i = 1, \dots, n$ , then

$$\underline{a_1X_1 + a_2X_2 + \dots + a_nX_n \sim N(a_1\mu_1 + \dots + a_n\mu_n, a_1^2\sigma_1^2 + \dots + a_n^2\sigma_n^2)}$$

Proof: see Chapter 5.

## The Central Limit Theorem (CLT)

The Central Limit Theorem (CLT) is one of the most fundamental results in statistics. In its simplest form, it states that if a large number of independent random variables are drawn from *any* distribution, then the distribution of their sum (or alternatively their average) always converges to the Normal distribution.

### Theorem (The Central Limit Theorem):

Let  $X_1, \dots, X_n$  be independent r.v.'s with mean  $\mu$  and variance  $\sigma^2$ , from ANY distribution.

(eg.  $X_i \sim \text{Bin}(n, p)$  for each  $i$ , so  $\mu = np$  and  $\sigma^2 = np(1 - p)$ ).

Then the sum  $S_n = X_1 + \dots + X_n = \sum_{i=1}^n X_i$  has a distribution that tends to Normal as  $n \rightarrow \infty$ .

$$\mathbb{E}(S_n) = \sum_{i=1}^n \mathbb{E}(X_i) = n\mu$$

$$\begin{aligned} \text{Var}(S_n) &= \text{Var}\left(\sum_{i=1}^n X_i\right) \\ &= \sum_{i=1}^n \text{Var}(X_i) \quad \text{because } X_1, \dots, X_n \text{ are independent (see end of Chapter 2)} \\ &= n\sigma^2 \end{aligned}$$

So  $S_n = X_1 + X_2 + \dots + X_n \rightarrow \text{Normal}(n\mu, n\sigma^2)$  as  $n \rightarrow \infty$ .

Alternatively,  $\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{S_n}{n} \rightarrow \mathbf{N}\left(\mu, \frac{\sigma^2}{n}\right)$  as  $n \rightarrow \infty$ .

A more general form of CLT states that, if  $X_1, \dots, X_n$  are independent, and  $\mathbb{E}(X_i) = \mu_i$ ,  $\text{Var}(X_i) = \sigma_i^2$  (not necessarily all equal), then

$$Z_n = \frac{\sum_{i=1}^n (X_i - \mu_i)}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \rightarrow \text{Normal}(0, 1) \quad \text{as } n \rightarrow \infty.$$

For the present, it is sufficient to remember the principle that *large sums of independent r.v.'s tend towards a Normal distribution, whatever the distribution of the original r.v.'s.*

## Example: Normal approximation to the Binomial

Let  $Y \sim \text{Binomial}(n, p)$ .

We can think of  $Y$  as the *sum of  $n$  Bernoulli random variables*:

$$Y = X_1 + X_2 + \dots + X_n, \text{ where } X_i = \begin{cases} 1 & \text{if trial } i \text{ is a "success" (probability } = p), \\ 0 & \text{otherwise (probability } = 1 - p) \end{cases}$$

So  $Y = X_1 + \dots + X_n$  and each  $X_i$  has  $\mu = \mathbb{E}(X_i) = p$ ,  $\sigma^2 = \text{Var}(X_i) = p(1 - p)$ .

Thus by the CLT,

$$\begin{aligned} Y = X_1 + X_2 + \dots + X_n &\rightarrow \text{Normal}(n\mu, n\sigma^2) \\ &= \text{Normal}(np, np(1 - p)). \end{aligned}$$

Thus,

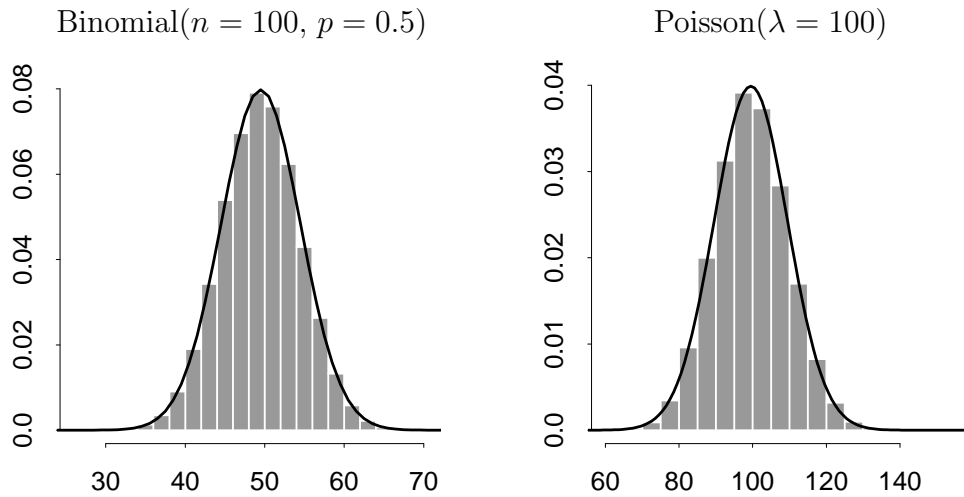
$$\text{Bin}(n, p) \rightarrow \text{Normal}\left(\underbrace{np}_{\text{mean of Bin}(n,p)}, \underbrace{np(1-p)}_{\text{var of Bin}(n,p)}\right) \text{ as } n \rightarrow \infty \text{ with } p \text{ fixed.}$$

The Binomial distribution is therefore well approximated by the Normal distribution when  $n$  is large, for any fixed value of  $p$ .

[Compare this with the Poisson approximation to the Binomial, section 2.3, which had  $\text{Bin}(n, p) \rightarrow \text{Poisson}(n \times p)$  as  $n \rightarrow \infty, p \rightarrow 0$  and  $np$  held fixed.]

The Normal distribution is also a good approximation to the Poisson( $\lambda$ ) distribution when  $\lambda$  is large:

$$\text{Poisson}(\lambda) \rightarrow \text{Normal}(\lambda, \lambda) \text{ when } \lambda \text{ is large.}$$



We will return to the CLT in more detail in Chapter 5 (including a sketch proof).

### 3.3 Finding the distribution of $g(X)$

Suppose we know the distribution of  $X$ . Let  $Y = g(X)$ . Our aim is to find the distribution of  $Y$ .

We look at two techniques:

1. Direct use of the distribution function.
2. Change of Variable technique when  $g(x)$  is a monotone function.

#### 1. Use of the distribution function to find the distribution of $Y = g(X)$

Let  $F_X(x) = \mathbb{P}(X \leq x)$  be the distribution function of  $X$ .

Let  $F_Y(y) = \mathbb{P}(Y \leq y)$  be the distribution function of  $Y = g(X)$ .

Now  $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \in g^{-1}((-\infty, y]))$ ,  
**where**  $g^{-1}((-\infty, y]) = \{x : g(x) \in (-\infty, y]\}$ .

That is, the probability that  $Y \leq y$  is the probability that  $X$  takes a value  $x$  that satisfies  $g(x) \leq y$ . We can use this approach if it is reasonably easy to find the set  $g^{-1}((-\infty, y])$ .

**Example 1:** Let  $X \sim \text{Uniform}(0, 1)$ .

$$\text{Then } f_X(x) = \frac{1}{1-0} = 1 \text{ for } 0 < x < 1,$$

$$F_X(x) = \int_0^x f_X(y) dy = x \text{ for } 0 < x < 1. \quad \circledast$$

Let  $Y = -\log(X)$ . We want to find the distribution of  $Y$ .

**Distribution function,**

$$\begin{aligned} F_Y(y) = \mathbb{P}(Y \leq y) &= \mathbb{P}(-\log(X) \leq y) \\ &= \mathbb{P}(\log(X) \geq -y) \\ &= \mathbb{P}(X \geq e^{-y}) \\ &= 1 - F_X(e^{-y}). \quad \circledast \circledast \end{aligned}$$

If  $y > 0$ , then  $0 < e^{-y} < 1$ , and  $F_X(e^{-y}) = e^{-y}$ . (by  $\circledast$ )

So  $F_Y(y) = 1 - e^{-y}$  if  $y > 0$ . (by  $\circledast \circledast$ )

If  $y \leq 0$ , then  $e^{-y} \geq 1$  so  $F_X(e^{-y}) = 1$ , and  $F_Y(y) = 0$ . (by  $\circledast \circledast$ )

$$\text{Thus } F_Y(y) = \begin{cases} 1 - e^{-y} & \text{if } y > 0, \\ 0 & \text{otherwise,} \end{cases}$$

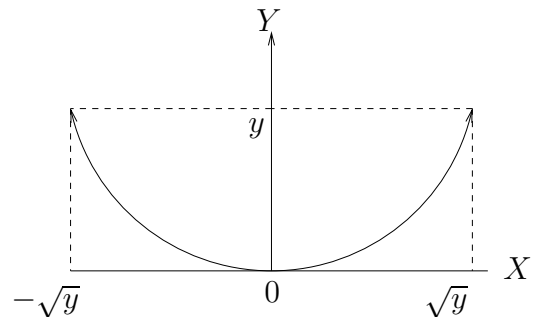
and therefore  $Y = -\log(X) \sim \text{Exponential}(1)$ .

**Example 2:** Let  $X$  have **any** distribution, with distribution function  $F_X(x)$ .

Let  $Y = X^2$ .

Clearly,  $Y \geq 0$ , so  $F_Y(y) = 0$  if  $y < 0$ .

$$\begin{aligned} \text{For } y \geq 0, F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(X^2 \leq y) \\ &= \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$



So

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0, \\ F_X(\sqrt{y}) - F_X(-\sqrt{y}) & \text{if } y \geq 0. \end{cases}$$

So the p.d.f. of  $Y$  is

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y = \frac{d}{dy} (F_X(\sqrt{y})) - \frac{d}{dy} (F_X(-\sqrt{y})) \\ &= \frac{1}{2} y^{-\frac{1}{2}} F'_X(\sqrt{y}) + \frac{1}{2} y^{-\frac{1}{2}} F'_X(-\sqrt{y}) \\ &= \frac{1}{2\sqrt{y}} (f_X(\sqrt{y}) + f_X(-\sqrt{y})) \quad \text{for } y \geq 0. \end{aligned}$$

$$\therefore f_Y(y) = \frac{1}{2\sqrt{y}} (f_X(\sqrt{y}) + f_X(-\sqrt{y})) \quad \text{for } y \geq 0, \text{ whenever } Y = X^2.$$

Special case: let  $X \sim \text{Normal}(0, 1)$ . Then  $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ . By the result above,  $Y = X^2$  has p.d.f.

$$\begin{aligned} f_Y(y) &= \frac{1}{2\sqrt{y}} \cdot \frac{1}{\sqrt{2\pi}} (e^{-y/2} + e^{-y/2}) \\ &= \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2} \quad \text{for } y \geq 0. \end{aligned}$$

But any distribution with p.d.f. of the form  $(\text{constant}) \times (y^{k-1} e^{-\lambda y})$  is **Gamma**( $k, \lambda$ ).

Here,  $k - 1 = -\frac{1}{2}$ , so  $k = \frac{1}{2}$ , and  $\lambda = \frac{1}{2}$ .

So if  $X \sim N(0, 1)$  then  $Y = X^2 \sim \text{Gamma}(k = \frac{1}{2}, \lambda = \frac{1}{2})$ .

But this is the **Chi-Squared** distribution with  $\nu = 1$  degrees of freedom.

So if  $X \sim N(0, 1)$ , then  $Y = X^2 \sim \chi_1^2$ .



## 2. Change of Variable technique for monotone functions

Let  $g(x)$  be a (1-1) function ('one-to-one'), i.e. *for every  $y$  there is a unique  $x$  such that  $g(x) = y$ .*

This means that the inverse function,  $g^{-1}(y)$ , is well-defined as a function for a certain range of  $y$ .

When  $g : \mathbb{R} \rightarrow \mathbb{R}$ , as it is here, then  $g$  can only be (1-1) if it is *monotone* (ie.  *$g$  is an increasing function, or  $g$  is a decreasing function.*

### Change of Variable formula

Let  $g : \mathbb{R} \rightarrow \mathbb{R}$  be a monotone function and let  $Y = g(X)$ . Then *the p.d.f. of  $Y = g(X)$  is*

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

### Easy way to remember

Write  $y = y(x) (= g(x))$   
 $\therefore x = x(y) (= g^{-1}(y))$

Then  $f_Y(y) = f_X(x(y)) \left| \frac{dx}{dy} \right|$

**Proof:** Separate into cases where  $g$  is increasing and where  $g$  is decreasing.

#### i) $g$ increasing

$g$  is increasing if  $u < w \Leftrightarrow g(u) < g(w)$ .  $\otimes$

Note that putting  $u = g^{-1}(x)$ , and  $w = g^{-1}(y)$ , we obtain

$$\begin{aligned} g^{-1}(x) < g^{-1}(y) &\Leftrightarrow g(g^{-1}(x)) < g(g^{-1}(y)) \\ &\Leftrightarrow x < y, \end{aligned}$$

so  $g^{-1}$  is also an increasing function.

Now

$$\begin{aligned} F_Y(y) = \mathbb{P}(Y \leq y) &= \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \leq g^{-1}(y)) \quad \text{put } \begin{cases} u = X, \\ w = g^{-1}(y) \end{cases} \text{ in } \textcircled{*} \text{ to see this.} \\ &= F_X(g^{-1}(y)). \end{aligned}$$

So pdf of  $Y$  is

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) \\ &= \frac{d}{dy} F_X(g^{-1}(y)) \\ &= F'_X(g^{-1}(y)) \frac{d}{dy} (g^{-1}(y)) \quad (\text{Chain Rule}) \\ &= f_X(g^{-1}(y)) \frac{d}{dy} (g^{-1}(y)) \end{aligned}$$

Now  $g$  is increasing, so  $g^{-1}$  is also increasing (by overleaf), so  $\frac{d}{dy}(g^{-1}(y)) > 0$ , and thus  $f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy}(g^{-1}(y)) \right|$  as required.

**ii)  $g$  decreasing,** i.e.  $u > w \iff g(u) < g(w)$ .  $(\star)$

(Putting  $u = g^{-1}(x)$  and  $w = g^{-1}(y)$  gives  $g^{-1}(x) > g^{-1}(y) \iff x < y$ , so  $g^{-1}$  is also decreasing.)

$$\begin{aligned} F_Y(y) = \mathbb{P}(Y \leq y) &= \mathbb{P}(g(X) \leq y) \\ &= \mathbb{P}(X \geq g^{-1}(y)) \quad (\text{put } u = X, w = g^{-1}(y) \text{ in } (\star)) \\ &= 1 - F_X(g^{-1}(y)). \end{aligned}$$

Thus the p.d.f. of  $Y$  is

$$f_Y(y) = \frac{d}{dy} \left( 1 - F_X(g^{-1}(y)) \right) = -f_X(g^{-1}(y)) \frac{d}{dy} (g^{-1}(y)).$$

This time,  $g$  is decreasing, so  $g^{-1}$  is also decreasing, and thus

$$-\frac{d}{dy} (g^{-1}(y)) = \left| \frac{d}{dy} (g^{-1}(y)) \right|.$$

So once again,

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} (g^{-1}(y)) \right|. \quad \square$$

## Working for change of variable questions

- 1) *Show you have checked  $g(x)$  is monotone over the required range.*
- 2) *Write  $y = y(x)$  for  $x$  in  $\langle \text{range of } x \rangle$ .*
- 3) *Thus  $x = x(y)$  for  $y$  in  $\langle \text{range of } y \rangle$ .*
- 4) *Then  $\left| \frac{dx}{dy} \right| = \langle \text{expression involving } y \rangle$ .*
- 5) *So  $f_Y(y) = f_X(x(y)) \left| \frac{dx}{dy} \right|$  by Change of Variable formula,  
= . . . . Quote range of values of  $y$  as part of the FINAL answer.*

**Note:** *There should be no  $x$ 's left in the answer!*

$x(y)$  and  $\left| \frac{dx}{dy} \right|$  are expressions involving  $y$  only.

**Example 1:** Let  $X \sim \text{Uniform}(0, 1)$ , and let  $Y = -\log(X)$ .  
(Same example as before).

- 1)  $y(x) = -\log(x)$  is monotone decreasing, so we can apply the Change of Variable formula.
- 2) Let  $y = y(x) = -\log x$  for  $0 < x < 1$ .
- 3) Then  $x = x(y) = e^{-y}$  for  $-\log(0) > y > -\log(1)$ , ie.  $0 < y < \infty$ .
- 4)  $\left| \frac{dx}{dy} \right| = \left| \frac{d}{dy}(e^{-y}) \right| = |-e^{-y}| = e^{-y}$  for  $0 < y < \infty$ .

5) So

$$\begin{aligned} f_Y(y) &= f_X(x(y)) \left| \frac{dx}{dy} \right| \quad \text{for } 0 < y < \infty \\ &= f_X(e^{-y})e^{-y} \quad \text{for } 0 < y < \infty. \end{aligned}$$

But  $X \sim \text{Uniform}(0, 1)$ , so  $f_X(x) = 1$  for  $0 < x < 1$ ,  
 $\Rightarrow f_X(e^{-y}) = 1$  for  $0 < y < \infty$ .

Thus  $f_Y(y) = f_X(e^{-y})e^{-y} = e^{-y}$  for  $0 < y < \infty$ .

**Note:** In change of variable questions, you lose a mark for:

1. not stating  $g(x)$  is monotone over the required range of  $x$ ;
2. not giving the range of  $y$  for which the result holds, as part of the final answer. (eg.  $f_Y(y) = \dots$  for  $0 < y < \infty$ ).

**Example 2:** Linear transformation of a Normal random variable

Let  $X \sim \text{Normal}(\mu, \sigma^2)$ , and let  $Y = aX + b$ .

1)  $y(x) = ax + b$  is monotone, so we can apply the Change of Variable technique.

2) Let  $y = y(x) = ax + b$  for  $-\infty < x < \infty$ .

3) Then  $x = x(y) = \frac{y-b}{a}$  for  $-\infty < y < \infty$ .

$$4) \left| \frac{dx}{dy} \right| = \left| \frac{1}{a} \right| = \frac{1}{|a|}.$$

5)

$$\begin{aligned} \text{So } f_Y(y) &= f_X(x(y)) \left| \frac{dx}{dy} \right| \\ &= f_X\left(\frac{y-b}{a}\right) \frac{1}{|a|}. \quad \clubsuit \end{aligned}$$

But  $X \sim N(\mu, \sigma^2)$ , so  $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$

$$\begin{aligned} \text{Thus } f_X\left(\frac{y-b}{a}\right) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{y-b}{a}-\mu\right)^2/2\sigma^2} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-(a\mu+b))^2/2a^2\sigma^2} \end{aligned}$$

Returning to ♣,

$$f_Y(y) = f_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{|a|} = \frac{1}{\sqrt{2\pi a^2 \sigma^2}} e^{-(y-(a\mu+b))^2/2a^2\sigma^2} \quad \text{for } -\infty < y < \infty.$$

But this is the pdf of a Normal( $a\mu + b$ ,  $a^2\sigma^2$ ) r.v., so

$$\text{if } X \sim N(\mu, \sigma^2), \text{ then } aX + b \sim N(a\mu + b, a^2\sigma^2)$$

This proves the assertion in Section 3.2.

**Example 3: Proof that if  $X \sim \text{Gamma}(k, \lambda)$ , then  $\lambda X \sim \text{Gamma}(k, 1)$**

Let  $X \sim \text{Gamma}(k, \lambda)$ , and let  $Y = \lambda X$ .

1)  $y(x) = \lambda x$  is monotone increasing (for  $\lambda > 0$ ) so we can apply the Change of Variable technique.

2) Let  $y = y(x) = \lambda x$  for  $0 \leq x < \infty$ .

3) Then  $x = x(y) = \frac{1}{\lambda}y$  for  $0 \leq y < \infty$ .

4)  $\left|\frac{dx}{dy}\right| = \frac{1}{\lambda}$  for  $0 \leq y < \infty$ .

5) Thus  $f_Y(y) = f_X(x(y)) \left|\frac{dx}{dy}\right| = f_X\left(\frac{1}{\lambda}y\right) \cdot \frac{1}{\lambda}$  for  $0 \leq y < \infty$ . ♠

Now  $X \sim \text{Gamma}(k, \lambda)$ , so  $f_X(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}$  (for  $0 \leq x < \infty$ ),

$$\text{hence } f_X\left(\frac{1}{\lambda}y\right) = \lambda^k \cdot \frac{y^{k-1}}{\lambda^{k-1}} \cdot \frac{e^{-\lambda \cdot y/\lambda}}{\Gamma(k)} = \frac{\lambda y^{k-1} e^{-\lambda \cdot y/\lambda}}{\Gamma(k)}.$$

Returning to ♠,

$$f_Y(y) = f_X\left(\frac{1}{\lambda}y\right) \cdot \left(\frac{1}{\lambda}\right) = \frac{\lambda y^{k-1} e^{-y}}{\Gamma(k)} \cdot \frac{1}{\lambda} \text{ for } 0 \leq y < \infty$$

$$\therefore f_Y(y) = \frac{y^{k-1} e^{-y}}{\Gamma(k)} \text{ for } 0 \leq y < \infty.$$

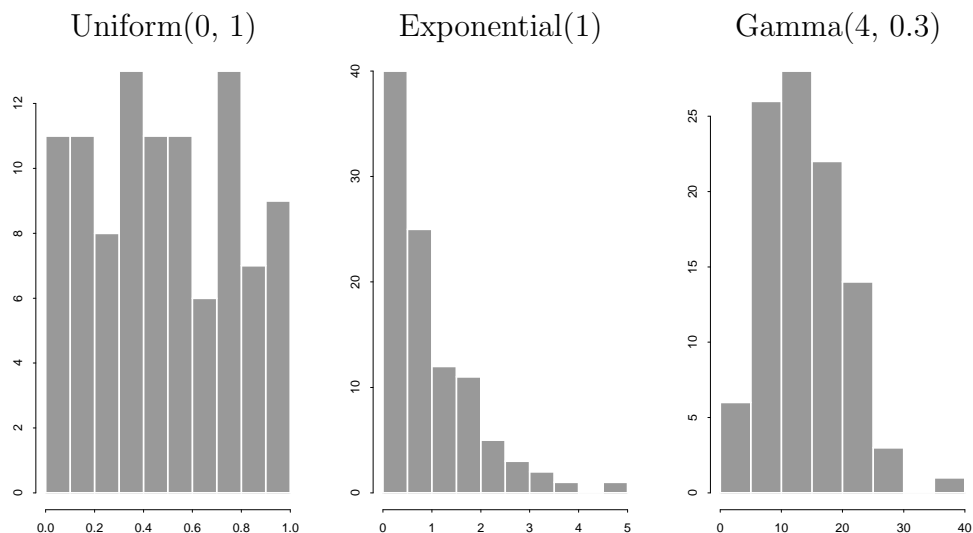
This is the pdf of  $\text{Gamma}(k, 1)$ ,

so if  $X \sim \text{Gamma}(k, \lambda)$ , then  $Y = \lambda X \sim \text{Gamma}(k, 1)$ , as claimed in §3.2.

### 3.4 Generating random numbers from continuous probability distributions

It is quite straightforward to generate random (or pseudo-random) numbers from a  $\text{Uniform}(0, 1)$  distribution: for example, most calculators have a random number generator (button marked RAN or RND or similar).

What if we want to generate a sample of random numbers from a different distribution, e.g. Exponential or Normal?



Histograms show samples of size 100 from the distributions indicated.

The following result is often helpful.

**Theorem 3.3:** Let  $F$  be a distribution function. Suppose  $F$  is **strictly** increasing on some interval  $(a, b)$  with  $F(a) = 0$ ,  $F(b) = 1$ , and  $a, b \in \mathbb{R}$ . Then  $F^{-1}(u)$  is a well-defined function for  $0 < u < 1$ .

Now let  $U \sim \text{Uniform}(0, 1)$  and let  $Y = F^{-1}(U)$ .

Then  $Y$  is a random variable with distribution function  $F$ .

**Proof:**

If  $U \sim \text{Uniform}(0, 1)$ , then  $F_U(u) = u$  for  $0 < u < 1$ ,  
ie.  $\mathbb{P}(U \leq u) = u$  for  $0 < u < 1$ .

Let  $Y = F^{-1}(U)$ . We want to show that the distribution function of  $Y$  is  $F$ ,  
ie. that  $\mathbb{P}(Y \leq y) = F(y)$ .

**LHS:**

$$\begin{aligned}\mathbb{P}(Y \leq y) &= \mathbb{P}(F^{-1}(U) \leq y) \\ &= \mathbb{P}(U \leq F(y)) \\ &= F(y) \text{ by } \textcircled{*}, \text{ because } \mathbb{P}(U \leq u) = u \text{ for any } u \in (0, 1).\end{aligned}$$

So  $Y$  has distribution function  $F$ , as required. □

This is quite a powerful result:

if  $U \sim U(0, 1)$ , then  $Y = F^{-1}(U)$  has distribution function  $F$ .

The Theorem tells us that to generate a sample  $y_1, y_2, \dots, y_n$  from a distribution with distribution function  $F$ , simply:

- i) generate  $u_1, u_2, \dots, u_n$  as random numbers from the  $U(0, 1)$  distribution (eg. using a calculator);
- ii) find the function  $F^{-1}$ , and compute  $y_1 = F^{-1}(u_1), \dots, y_n = F^{-1}(u_n)$ . The sample  $y_1, y_2, \dots, y_n$  are then a sample from the required distribution.

**Example:** The following random numbers are drawn from the Uniform(0, 1) distribution:

$$0.98 \quad 0.77 \quad 0.38 \quad 0.66 \quad 0.24$$

Use these numbers to find a sample of size 5 from the Exponential(3) distribution.

For the Exponential(3) distribution,  $F(y) = 1 - e^{-3y}$  ( $y \geq 0$ ).

Write  $u = F(y) = 1 - e^{-3y}$ .

Then

$$\begin{aligned} 1 - u &= e^{-3y}, \\ -\log(1 - u) &= 3y, \\ y &= -\frac{1}{3} \log(1 - u). \end{aligned}$$

So the inverse function is  $F^{-1}(u) = -\frac{1}{3} \log(1 - u)$ .

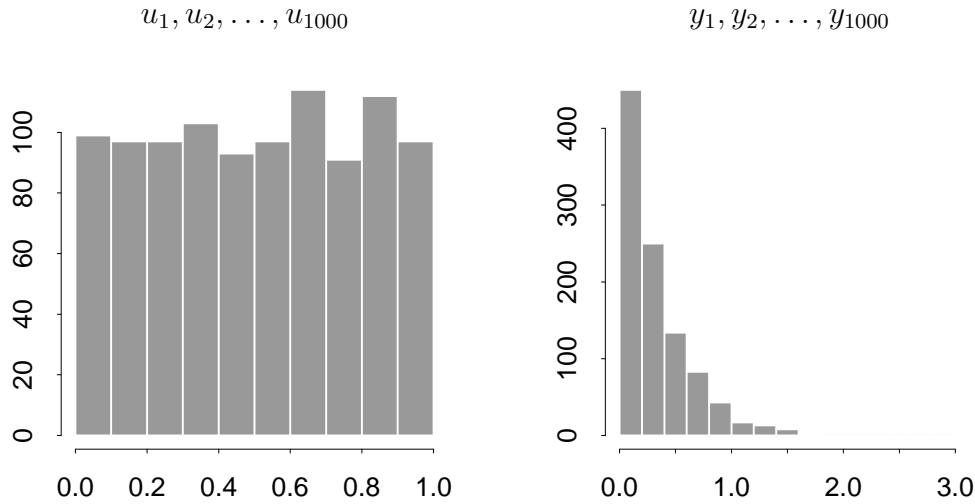
Given numbers  $u_1 = 0.98, u_2 = 0.77, \dots, u_5 = 0.24$ , we can construct  $y_1 = -\frac{1}{3} \log(1 - u_1), \dots, y_5 = -\frac{1}{3} \log(1 - u_5)$  as a sample from the Exponential(3) distribution.

The required sample is

$$y_1 = 1.304 \quad y_2 = 0.490 \quad y_3 = 0.159 \quad y_4 = 0.360 \quad y_5 = 0.091.$$



The figure shows a histogram of 1000 random numbers  $u_1, \dots, u_{1000}$  generated from the  $\text{Uniform}(0, 1)$  distribution, and the same 1000 numbers transformed using  $y_i = -\frac{1}{3} \log(1 - u_i)$ . The distribution of  $y_1, \dots, y_{1000}$  has the characteristic Exponential shape, as it should.



**Corollary :** The Theorem stated that if  $U \sim \text{Uniform}(0, 1)$ , then  $Y = F^{-1}(U)$  has distribution function  $F$ . An alternative statement is:

*if  $Y$  is a r.v. with strictly increasing distribution function  $F_Y$ , then  $F_Y(Y) \sim \text{Uniform}(0, 1)$ .*

Proof:

Let  $X = F_Y(Y)$ .

$$\begin{aligned}
 F_X(x) = \mathbb{P}(X \leq x) &= \mathbb{P}(F_Y(Y) \leq x) \\
 &= \mathbb{P}(Y \leq F_Y^{-1}(x)) \text{ because } F_Y \text{ is strictly increasing} \\
 &= F_Y(F_Y^{-1}(x)) \text{ by definition of } F_Y
 \end{aligned}$$

$$\therefore F_X(x) = x, \text{ for } 0 < x < 1.$$

But  $F_X(x) = x$  is the distribution function of the  $\text{Uniform}(0, 1)$  distribution, so we must have  $X = F_Y(Y) \sim U(0, 1)$ . □

## The Hazard Function (non-examinable)

If  $X$  is a random variable representing the lifetime of some object (e.g. a human), then the *hazard function* for  $X$  is defined as

$$\text{hazard function, } h(x) = \frac{f_X(x)}{1 - F_X(x)}.$$

The hazard function may be thought of as the instantaneous death rate at age  $x$ , i.e.:

$$\mathbb{P}(\text{dies in interval } (x, x + \delta x) \mid \text{has survived until age } x) = h(x)\delta x.$$

**Example:** If  $X \sim \text{Exponential}(\lambda)$ , then  $h(x) = \frac{\lambda e^{-\lambda x}}{1 - (1 - e^{-\lambda x})} = \lambda = \text{constant}$ .

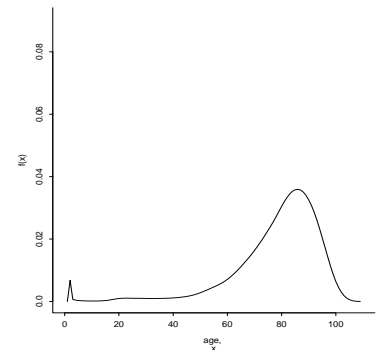
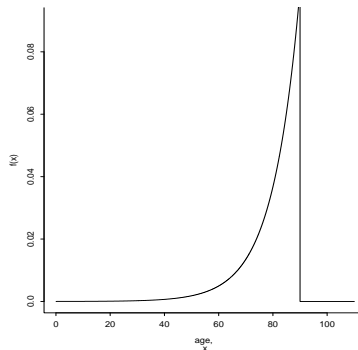
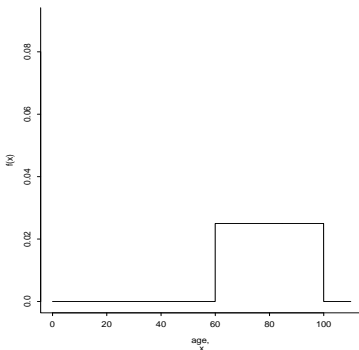
So the Exponential distribution describes the lifetime of an object that does not age: its death rate is constant ( $\lambda$ ) at all ages.

---

## Endnote . . . which lifetime distribution?

You have been given the choice of three distributions for your lifetime:

1. Uniform(60, 100):      2. 90 – Exponential ( $\frac{1}{10}$ ):      3. True NZ distribution



## Which are you going to choose . . . ?

# Chapter 4: Multivariate Distributions

---

## 4.1 Discrete Bivariate Distributions

---

Suppose  $X$  and  $Y$  are *discrete* random variables. If there is *dependence* between  $X$  and  $Y$ , we might be interested in *their joint behaviour*.

*Definition:* The **joint probability function**,  $f_{X,Y}(x, y)$ , of  $X$  and  $Y$  is given by

$$f_{X,Y}(x, y) = \mathbb{P}(X = x \text{ and } Y = y)$$

We often write

$$f(x, y) = \mathbb{P}(X = x, Y = y)$$

We can also write

$$f_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} = \mathbf{x})$$

where  $\mathbf{X} = (X, Y)$  is a vector of random variables  $X$  and  $Y$ , and  $\mathbf{x} = (x, y)$  is a vector of observations:  $X = x, Y = y$ .

$f_{X,Y}(x, y)$  is called a bivariate probability function, because it involves two random variables,  $X$  and  $Y$ . (*Bivariate = two variables*).

### Properties of the joint probability function

---

- i)  $f(x, y) \geq 0$  for all  $x$  and  $y$ ;
- ii)  $\sum_x \sum_y f(x, y) = 1$ .

**Example:** A milkman delivers bottles of milk and boxes of eggs to a house. He gets a daily note to say how many milk bottles and egg boxes are required.

Let  $X =$  *number of egg boxes*.

and  $Y =$  *number of milk bottles*.

Suppose the joint probability function of  $X$  and  $Y$  is as follows:

$f_{X,Y}(x, y)$		$y$ (milk bottles)				Total
		0	1	2	3	
$x$	0	0.05	0.05	0.10	0	0.20
(egg	1	0.05	0.10	0.25	0.10	0.50
boxes)	2	0	0.15	0.10	0.05	0.30
Total		0.10	0.30	0.45	0.15	1

We interpret this as follows:

$$\mathbb{P}(X = 0, Y = 0) = f_{X,Y}(0, 0) = 0.05 \text{ (no eggs, no milk)}$$

$$\mathbb{P}(X = 2, Y = 1) = f_{X,Y}(2, 1) = 0.15 \text{ (2 eggs, 1 milk)}$$

We will use this example in the following definitions.

## Bivariate Distribution Function

*Definition:* Let  $X$  and  $Y$  be discrete random variables. The **bivariate distribution function** is  $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$ .

It is given by,

$$\begin{aligned} F_{X,Y}(x, y) &= \sum_{\{x^*: x^* \leq x\}} \sum_{\{y^*: y^* \leq y\}} \mathbb{P}(X = x^*, Y = y^*) \\ &= \sum_{x^* \leq x} \sum_{y^* \leq y} f_{X,Y}(x^*, y^*). \end{aligned}$$

**Example:** In the milkman example,

$$\begin{aligned} F_{X,Y}(1, 2) &= \mathbb{P}(\# \text{ egg boxes} \leq 1 \text{ and } \# \text{ milk bottles} \leq 2) \\ &= \mathbb{P}(X \leq 1, Y \leq 2). \end{aligned}$$

$f_{X,Y}(x, y)$		$y$				Total
		0	1	2	3	
$x$	0	0.05	0.05	0.10	0	0.20
	1	0.05	0.10	0.25	0.10	0.50
	2	0	0.15	0.10	0.05	0.30
Total		0.10	0.30	0.45	0.15	1

We sum all the entries that satisfy  $x \leq 1$  and  $y \leq 2$ :

$$\begin{aligned} F_{X,Y}(1, 2) &= 0.05 + 0.05 + 0.10 + 0.05 + 0.10 + 0.25 \\ &= 0.6 \end{aligned}$$

### Marginal probability functions

Given a joint probability function  $f_{X,Y}(x, y)$ , we can find the *individual probability functions of  $X$  and  $Y$ ,  $f_X(x)$  and  $f_Y(y)$* .

These are called the marginal probability functions.

*Definition:* Let  $X$  be a discrete random variable. The marginal probability function of  $X$  is given by  $f_X(x) = \mathbb{P}(X = x)$ .

*The marginal probability function is exactly the same as the univariate probability function for  $X$  that we defined in Chapter 2. The term “marginal” is usually used when there is the possibility of confusion with a joint probability function.*

## Finding the marginal probability functions using $f_{X,Y}(x, y)$

Consider the milkman example again:

$f_{X,Y}(x, y)$	$y$				Total	
	0	1	2	3		
$x$	0	0.05	0.05	0.10	0	0.20
	1	0.05	0.10	0.25	0.10	0.50
	2	0	0.15	0.10	0.05	0.30
Total	0.10	0.30	0.45	0.15	1	

The overall probability that  $X = 0$  is the *sum of all table entries that have  $x = 0$ : that is, the row total for  $x = 0$ , 0.20.*

Similarly, the probability that  $Y = 2$  is the column total for  $y = 2$ , 0.45.

The marginal probabilities are therefore obtained by looking in the margins of the table.

*In fact, we are implicitly using the Partition Theorem:*

$$\begin{aligned}\mathbb{P}(X = 0) &= \mathbb{P}(X = 0, Y = 0) + \mathbb{P}(X = 0, Y = 1) \\ &\quad + \mathbb{P}(X = 0, Y = 2) + \mathbb{P}(X = 0, Y = 3) \quad (\text{row total})\end{aligned}$$

- because events  $\{Y = 0\}$ ,  $\{Y = 1\}$ ,  $\{Y = 2\}$  and  $\{Y = 3\}$  form a partition of the sample space  $\Omega = \{(x, y) : x = 0, 1, 2; y = 0, 1, 2, 3\}$ .

*Similarly,*

$$\mathbb{P}(Y = 2) = \mathbb{P}(X = 0, Y = 2) + \mathbb{P}(X = 1, Y = 2) + \mathbb{P}(X = 2, Y = 2) \quad (\text{column total})$$

- because events  $\{X = 0\}$ ,  $\{X = 1\}$  and  $\{X = 2\}$  form a partition of  $\Omega$ .

In general, the marginal probability functions are given by:

$$\mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y)$$

ie.  $f_X(x) = \sum_y f_{X,Y}(x, y)$  (marginal probability function of  $X$ )

Similarly,  $f_Y(y) = \sum_x f_{X,Y}(x, y)$  (marginal probability function of  $Y$ )

**Example:** In the milkman example,  
Marginal probability function of  $X$  is

$x$	$0$	$1$	$2$
$f_X(x)$	$0.20$	$0.50$	$0.30$

Marginal probability function of  $Y$  is

$y$	$0$	$1$	$2$	$3$
$f_Y(y)$	$0.10$	$0.30$	$0.45$	$0.15$

Note: sum=1 in each case.

### Conditional probability functions

*Definition:* Let  $X$  and  $Y$  be discrete random variables. The conditional probability function of  $X$ , given that  $Y$  takes the value  $y$ , is:

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)} \quad \text{if } f_Y(y) > 0.$$

Similarly,

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

**Note:** The conditional probability function  $f_{X|Y}(x|y)$  is a function of  $x$ . *Usually,  $y$  is a single fixed number, eg.  $f_{X|Y}(x|Y = 5)$  or  $f_{X|Y}(x|Y = -6)$ .  $x$  is variable and ranges over all values that  $X$  can take.*

**Example:** For the milkman, the conditional probability function of  $X$ , given that  $Y = 2$ , is:

$x$	0	1	2
$f_{X Y}(x 2)$	$\frac{0.1}{0.45} = \frac{2}{9}$	$\frac{0.25}{0.45} = \frac{5}{9}$	$\frac{0.1}{0.45} = \frac{2}{9}$

Thus  $x$  ranges from 0 to 2, while  $y$  stays fixed at 2.

Note that the sum=1:  $\sum_x f_{X|Y}(x|y) = 1$ .

**Exercise:** Show that the conditional probability function of  $Y$ , given that  $X = 0$ , is:

$y$	0	1	2	3
$f_{Y X}(y 0)$	0.25	0.25	0.5	0

**Example:** An insect lays eggs on a leaf. Let  $X$  be the number of eggs the insect lays, and suppose that  $X \sim \text{Poisson}(\lambda)$ .

Suppose also that every egg laid survives to maturity with probability  $p$ , independently of other eggs.

Let  $Y$  be the number of eggs surviving to maturity.

- a) Find the joint probability function of  $X$  and  $Y$ ,  $f_{X,Y}(x, y)$ .
- b) Find the marginal probability function of  $Y$ ,  $f_Y(y)$ .  
Hence name the distribution of  $Y$ .

### Solution

a) We are told that, given a fixed number  $x$  of eggs, they survive to maturity independently with probability  $p$ .

Thus, given  $x$  eggs to start with,

$Y = (\# \text{ surviving to maturity}) \sim \text{Binomial}(x, p)$ .



This is therefore the conditional distribution of  $Y$  given that  $X = x$ .

$Y|(X = x) \sim \text{Binomial}(x, p)$  ie.  $Y|X \sim \text{Binomial}(X, p)$ .

$$\begin{aligned} \text{So } f_{Y|X}(y|x) &= \mathbb{P}(Y = y|X = x) = \binom{x}{y} p^y (1-p)^{x-y} \quad (1) \\ &\text{(for } y = 0, 1, \dots, x\text{).} \end{aligned}$$

$$\begin{aligned} \text{Looking for } f_{X,Y}(x, y) &= \mathbb{P}(X = x, Y = y) \\ &= \mathbb{P}(Y = y|X = x)\mathbb{P}(X = x) \\ &= f_{Y|X}(y|x)f_X(x). \quad (2) \end{aligned}$$

We know that  $X \sim \text{Poisson}(\lambda)$ , so  $f_X(x) = \frac{\lambda^x}{x!}e^{-\lambda}$ .

Thus from (1) and (2),

$$\begin{aligned} f_{X,Y}(x, y) &= \binom{x}{y} p^y (1-p)^{x-y} \cdot \frac{\lambda^x}{x!} e^{-\lambda} \\ &= \frac{x!}{(x-y)!y!} p^y (1-p)^{x-y} \cdot \frac{\lambda^x}{x!} e^{-\lambda} \\ f_{X,Y}(x, y) &= \frac{p^y (1-p)^{x-y} \lambda^x e^{-\lambda}}{(x-y)!y!} \quad \text{for } x = 0, 1, 2, \dots \text{ and } y = 0, 1, 2, \dots, x \end{aligned}$$

b) *Marginal probability function of Y:*

$$\begin{aligned}
 f_Y(y) &= \sum_{x=y}^{\infty} f_{X,Y}(x,y) \quad (\text{note that } \underbrace{x}_{\# \text{ eggs}} \geq \underbrace{y}_{\# \text{ surviving eggs}} : f_{X,Y}(x,y) = 0 \text{ if } x < y.) \\
 &= \sum_{x=y}^{\infty} \frac{p^y(1-p)^{x-y}\lambda^x e^{-\lambda}}{(x-y)!y!} \\
 &= \frac{p^y}{y!} e^{-\lambda} \sum_{x=y}^{\infty} \frac{(1-p)^{x-y}\lambda^x}{(x-y)!} \quad (\text{taking all terms not involving } x \text{ out of the sum}) \\
 &= \frac{p^y}{y!} e^{-\lambda} \sum_{m=0}^{\infty} \frac{(1-p)^m \lambda^{m+y}}{m!} \quad (\text{where } m=x-y) \\
 &= \frac{(\lambda p)^y}{y!} e^{-\lambda} \sum_{m=0}^{\infty} \frac{\{\lambda(1-p)\}^m}{m!} \\
 &= \frac{(\lambda p)^y}{y!} e^{-\lambda} \cdot e^{\lambda(1-p)} \\
 f_Y(y) &= \frac{(\lambda p)^y}{y!} e^{-\lambda p} \quad \text{for } y = 0, 1, 2, \dots
 \end{aligned}$$

*But this is a Poisson probability, with parameter  $(\lambda p)$ .  
So the marginal distribution of Y is  $Y \sim \text{Poisson}(\lambda p)$ .*

This is a general result:

Let  $X$  = number of objects. Suppose that  $X \sim \text{Poisson}(\lambda)$ .

For each object, let  $\mathbb{P}(\text{object is 'special'}) = p$ , and let all objects be independent.

Let  $Y$  = number of *special* objects. Then  $Y \sim \text{Poisson}(\lambda p)$ .

However, note that  $X$  and  $Y$  are not independent.

**Note:** Conditional probability functions have all the usual properties of probability functions:

ie.  $f_{X|Y}(x|y) \geq 0$  for all  $x, y$

$$\begin{aligned} \text{and} \quad \sum_x f_{X|Y}(x|y) &= \sum_x \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{1}{f_Y(y)} \sum_x f_{X,Y}(x,y) \\ &= \frac{1}{f_Y(y)} \cdot f_Y(y) \\ \therefore \sum_x f_{X|Y}(x|y) &= 1 \end{aligned}$$

### Independence of discrete random variables

*Definition:* Let  $X$  and  $Y$  be discrete random variables.  $X$  and  $Y$  are statistically independent if and only if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \text{ for all } x, y.$$

**Notes:**

1. Compare with the definition of statistical independence for **events**  $A$  and  $B$ :

$A$  and  $B$  are independent if and only if  $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ .

2. Check that the definition of independence ensures that

$$\mathbb{P}(X = x|Y = y) = \mathbb{P}(X = x) \text{ for all } x, y. \quad (\text{Exercise}).$$

**Theorem 4.1:** Discrete random variables  $X$  and  $Y$  are independent if and only if  $f_{X,Y}(x,y)$  can be written as the product of a function of  $x$  only and a function of  $y$  only: that is, if and only if there exist functions  $g$  and  $h$  such that,

$$f_{X,Y}(x,y) = g(x)h(y) \text{ for ALL } x, y.$$

If  $f_{X,Y}(x,y) = g(x)h(y)$ , then the marginal probability functions are

$$f_X(x) = \frac{g(x)}{\sum_u g(u)}, \quad f_Y(y) = \frac{h(y)}{\sum_u h(u)}$$

**Proof:**

We need to show that:

$$(i) \ X \text{ and } Y \text{ are independent} \quad \Rightarrow \quad f_{X,Y}(x, y) = g(x)h(y);$$

$$(ii) \ f_{X,Y}(x, y) = g(x)h(y) \quad \Rightarrow \quad X \text{ and } Y \text{ are independent.}$$

**Proof of (i):**

By the definition of independence,

$$X \text{ and } Y \text{ are independent} \quad \Rightarrow \quad f_{X,Y}(x, y) = f_X(x)f_Y(y) \text{ for all } x \text{ and } y.$$

$$(i) \text{ follows by putting } g(x) = f_X(x) \text{ and } h(y) = f_Y(y).$$

**Proof of (ii):**

Suppose that  $f_{X,Y}(x, y) = g(x)h(y)$  for some functions  $g$  and  $h$  and for all  $x, y$ .

Now the marginal probability function of  $X$  is given by

$$f_X(x) = \sum_y f_{X,Y}(x, y) = \sum_y g(x)h(y) = g(x) \sum_y h(y) = g(x)H, \quad (a)$$

say, where  $H = \sum_y h(y)$ .

Similarly,

$$f_Y(y) = \sum_x f_{X,Y}(x, y) = \sum_x g(x)h(y) = h(y) \sum_x g(x) = h(y)G, \quad (b)$$

where  $G = \sum_x g(x)$ .

Results (a) and (b) give  $g(x) = \frac{f_X(x)}{H}$  and  $h(y) = \frac{f_Y(y)}{G}$ , so

$$f_{X,Y}(x, y) = g(x)h(y) = \frac{f_X(x)}{H} \frac{f_Y(y)}{G} = \frac{f_X(x)f_Y(y)}{GH}.$$

This shows that the joint probability function  $f_{X,Y}(x, y)$  is *proportional* to  $f_X(x)f_Y(y)$ , which is close to the result that we need for demonstrating independence. We must now show that  $GH = 1$ , so that  $f_{X,Y}(x, y)$  is *equal* to  $f_X(x)f_Y(y)$ .

By definition of  $G$  and  $H$ ,

$$GH = \sum_x g(x) \sum_y h(y) = \sum_x \sum_y g(x)h(y) = \sum_x \sum_y f_{X,Y}(x, y) = 1,$$

because  $f_{X,Y}(x, y)$  is the joint probability function so it must sum to 1.

Thus

$$f_{X,Y}(x, y) = \frac{f_X(x)f_Y(y)}{GH} = \frac{f_X(x)f_Y(y)}{1} = f_X(x)f_Y(y) \quad \text{for all } x, y,$$

and so  $X$  and  $Y$  are independent.

Further, because  $GH = 1$ , we have  $H = \frac{1}{G}$  and  $G = \frac{1}{H}$ . So from (a) and (b), the marginals are:

$$f_X(x) = g(x)H = \frac{g(x)}{G} = \frac{g(x)}{\sum_u g(u)},$$

and

$$f_Y(y) = h(y)G = \frac{h(y)}{H} = \frac{h(y)}{\sum_u h(u)},$$

as required. □

## 4.2 Expectation over a joint distribution

Recall from Chapter 2 that if  $X$  is a univariate discrete random variable, then

$$\mathbb{E}(g(X)) = \sum_x g(x)f_X(x).$$

How do we calculate expectations over a joint distribution? For example, if  $X_1$  and  $X_2$  are jointly distributed discrete random variables, what is  $\mathbb{E}(X_1/\sqrt{X_2})$ ?

*Definition:* Suppose that  $\mathbf{X} = (X_1, X_2, \dots, X_k)$  is a  $k$ -variate discrete random variable: that is, each  $X_i$  is a **univariate** discrete random variable. Let the function  $g : \mathbb{R}^k \rightarrow \mathbb{R}$  be a **scalar** function on  $\mathbb{R}^k$  (i.e.  $g$  takes **scalar** values in  $\mathbb{R}$ ). Then the expectation of  $g(\mathbf{X})$  is

$$\mathbb{E}(g(\mathbf{X})) = \sum_{\mathbf{x}} g(\mathbf{x})f_{\mathbf{X}}(\mathbf{x})$$

**Note:** This is a scalar sum, not a vector sum.

**Example:** If  $\mathbf{X} = (X, Y)$  is a discrete bivariate random variable, and  $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ , then

$$\mathbb{E}(g(X, Y)) = \sum_x \sum_y g(x, y) f_{X,Y}(x, y).$$

For example, 
$$\mathbb{E}\left(\frac{X}{\sqrt{Y}}\right) = \sum_x \sum_y \frac{x}{\sqrt{y}} f_{X,Y}(x, y).$$

### Properties of expectation

i) If  $\mathbf{X} = (X_1, X_2, \dots, X_k)$  is a  $k$ -variate discrete random variable, then for any constants  $a$  and  $b$  and any functions  $g$  and  $h$ ,

$$\mathbb{E}(ag(\mathbf{X}) + bh(\mathbf{X})) = a\mathbb{E}(g(\mathbf{X})) + b\mathbb{E}(h(\mathbf{X})).$$

Proof:

$$\begin{aligned} \mathbb{E}(ag(\mathbf{X}) + bh(\mathbf{X})) &= \sum_{\mathbf{x}} (ag(\mathbf{x}) + bh(\mathbf{x})) f_{\mathbf{X}}(\mathbf{x}) \\ &= a \sum_{\mathbf{x}} g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) + b \sum_{\mathbf{x}} h(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) \\ &= a\mathbb{E}(g(\mathbf{X})) + b\mathbb{E}(h(\mathbf{X})) \end{aligned}$$

ii) For **any** discrete random variables  $X$  and  $Y$ ,

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

Consequently, for any discrete random variables  $X_1, \dots, X_k$ ,

$$\mathbb{E}(X_1 + X_2 + \dots + X_k) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \dots + \mathbb{E}(X_k).$$

Note that we do **not** require  $X_1, \dots, X_k$  to be independent.

Proof:

$$\begin{aligned}\mathbb{E}(X + Y) &= \sum_x \sum_y (x + y) f_{X,Y}(x, y) \\ &= \sum_x x \sum_y f_{X,Y}(x, y) + \sum_y y \sum_x f_{X,Y}(x, y) \\ &= \sum_x x f_X(x) + \sum_y y f_Y(y) \\ &= \mathbb{E}(X) + \mathbb{E}(Y)\end{aligned}$$

iii) If  $X$  and  $Y$  are *independent* discrete random variables, and  $g, h$  are functions, then

$$\mathbb{E}(XY) = (\mathbb{E}X)(\mathbb{E}Y)$$

and

$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X)) \cdot \mathbb{E}(h(Y))$$

Note that this result *DOES* require  $X$  and  $Y$  to be *INDEPENDENT*.

Proof:

$$\begin{aligned}\mathbb{E}(XY) &= \sum_x \sum_y xy f_{X,Y}(x, y) \\ &= \sum_x \sum_y xy f_X(x) f_Y(y) \quad \text{if } X, Y \text{ independent} \\ &\quad \text{(so } f_{X,Y}(x, y) = f_X(x) f_Y(y)\text{)} \\ &= \left( \sum_x x f_X(x) \right) \left( \sum_y y f_Y(y) \right) \\ &= (\mathbb{E}X)(\mathbb{E}Y).\end{aligned}$$

Proof for  $\mathbb{E}(g(X)h(Y))$  similar. This proves the assertion made in section 2.5.

**Example:** (milkman example).

$X$  = number of egg boxes.

$Y$  = number of milk bottles.

		$y$				Total
		0	1	2	3	
$x$	0	0.05	0.05	0.10	0	0.20
	1	0.05	0.10	0.25	0.10	0.50
	2	0	0.15	0.10	0.05	0.30
Total		0.10	0.30	0.45	0.15	1

Suppose the milkman wants to know the expected number of *items* (egg boxes plus milk bottles) he is to deliver to a house.

Total number of items =  $X + Y$ .

Working directly from the definition of expectation,

$$\begin{aligned}\mathbb{E}(X + Y) &= \sum_x \sum_y (x + y) f_{X,Y}(x, y) \\ &= (0 + 0) \times 0.05 + (0 + 1) \times 0.05 + (0 + 2) \times 0.10 + (0 + 3) \times 0 \\ &\quad + (1 + 0) \times 0.05 + (1 + 1) \times 0.10 + \dots + (2 + 3) \times 0.05 \\ &= 2.75\end{aligned}$$

**Exercise:** Verify that  $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$  in this example.

---

### 4.3 Covariance and correlation between two random variables

---

Recall that the *variance* of a random variable  $X$  is  $\text{Var}(X) = \mathbb{E}((X - \mu_X)^2)$ .

When we have two random variables,  $X$  and  $Y$ , we often wish to quantify the relationship between them. One tool for doing this is the *covariance*. The covariance measures the linear association between  $X$  and  $Y$ .

*Definition:* The covariance between random variables  $X$  and  $Y$  is given by

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \quad \text{where } \mu_X = \mathbb{E}(X), \mu_Y = \mathbb{E}(Y)$$

Immediate from the definition is the alternative result:

$$\text{cov}(X, Y) = \mathbb{E}(XY) - \mu_X \mu_Y = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$



## Proof:

$$\begin{aligned}\text{cov}(X, Y) &= \mathbb{E}((X - \mu_X)(Y - \mu_Y)) \\ &= \mathbb{E}(XY - \mu_Y X - \mu_X Y + \mu_X \mu_Y) \\ &= \mathbb{E}(XY) - \mu_Y \mathbb{E}(X) - \mu_X \mathbb{E}(Y) + \mu_X \mu_Y \\ &= \mathbb{E}(XY) - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y \\ &= \mathbb{E}(XY) - \mu_X \mu_Y.\end{aligned}$$

## Intuitive explanation of covariance

If  $\text{cov}(X, Y)$  is *positive*, then  $\mathbb{E}((X - \mu_X)(Y - \mu_Y)) > 0$ . This means that  $(X - \mu_X)$  and  $(Y - \mu_Y)$  will *tend on the whole* to be either *both positive* or *both negative* (so that their product is positive on average).

Thus a positive covariance tends to suggest a positive association between  $X$  and  $Y$ : if  $X$  is *larger* than average ( $X - \mu_X > 0$ ), then  $Y$  will often be *larger* than average too ( $Y - \mu_Y > 0$ ). Similarly, if  $X$  is *smaller* than average, then  $Y$  will often be *smaller* than average too.

By contrast, if  $\text{cov}(X, Y) < 0$ , then when  $X - \mu_X > 0$  we will often have  $Y - \mu_Y < 0$ , and vice versa. This indicates a *negative* association between  $X$  and  $Y$ .

**Notes:** 1.  $X$  and  $Y$  are both random, but they might have no, some, or complete dependence on each other.

2. It is usually easiest to calculate  $\text{cov}(X, Y)$  using the formula

$$\text{cov}(X, Y) = \mathbb{E}(XY) - \mu_X \mu_Y. \text{ Recall that } \mathbb{E}(XY) = \sum_x \sum_y xy f_{X,Y}(x, y).$$

## Covariance of independent random variables

When  $X$  and  $Y$  are *independent*, then  $\text{cov}(X, Y) = 0$ .

The converse is *NOT TRUE*:

$X, Y$  independent  $\Rightarrow$   $\text{cov}(X, Y) = 0$ , but  $\text{cov}(X, Y) = 0 \not\Rightarrow X, Y$  independent.

## Proof:

When  $X$  and  $Y$  are independent,  $\mathbb{E}(XY) = (\mathbb{E}X)(\mathbb{E}Y)$  from page 13. Thus  $\text{cov}(X, Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y) = 0$ .

Conversely, consider the following joint distribution:

$$(X, Y) = \begin{cases} (1, 0) & \text{with probability } 1/4 \\ (0, 1) & \text{with probability } 1/4 \\ (-1, 0) & \text{with probability } 1/4 \\ (0, -1) & \text{with probability } 1/4 \end{cases}$$

Now  $\mathbb{E}(X) = \sum_x x\mathbb{P}(X = x) = 1 \times \frac{1}{4} + 0 \times \frac{1}{4} + (-1) \times \frac{1}{4} + 0 \times \frac{1}{4} = 0$ .

Similarly,  $\mathbb{E}(Y) = 0$ .

Also,  $\mathbb{E}(XY) = 0$  because  $XY = 0$  with probability 1.

So  $\text{cov}(X, Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y) = 0$ .

However,  $\mathbb{P}(X = 0 \text{ and } Y = 0) = 0$ , but  $\mathbb{P}(X = 0)\mathbb{P}(Y = 0) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ .

So  $\mathbb{P}(X = 0, Y = 0) \neq \mathbb{P}(X = 0)\mathbb{P}(Y = 0)$ , so  $X$  and  $Y$  are not independent.

Intuitively,  $\text{cov}(X, Y) = 0$  when  $X$  and  $Y$  are independent because whether  $X$  is above average or below average has no effect on the value of  $Y$ .

## Using the covariance to find the variance of a sum

The covariance is particularly useful for finding  $\text{Var}(X + Y)$ .

**Theorem 4.2:** For any random variables  $X$  and  $Y$ , and constants  $a, b$ :

i)  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{cov}(X, Y)$ .    **LEARN!**

ii)  $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{cov}(X, Y)$ .

iii)  $\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{cov}(X, Y)$ .    **LEARN!**

iv) For constants  $a_1, \dots, a_n$ ,

$$\text{Var} \left( \sum_{i=1}^n a_i X_i \right) = \sum_{i=1}^n a_i^2 \text{Var}(X_i) + 2 \sum_{i=1}^n \sum_{j>i} a_i a_j \text{cov}(X_i, X_j).$$

**Memory Aid:** Remember these results by thinking of  $(X + Y)^2$ ,  $(X - Y)^2$ , and  $(aX + bY)^2$ . Whenever you see an  $X^2$  or  $Y^2$ , replace by  $\text{Var}(X)$  and  $\text{Var}(Y)$ . Whenever you see  $XY$ , replace by  $\text{cov}(X, Y)$ .

eg.  $(aX + bY)^2 = a^2 X^2 + b^2 Y^2 + 2abXY$

$\rightsquigarrow \text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{cov}(X, Y)$ .

**Note:** When  $X$  and  $Y$  are independent,  $\text{cov}(X, Y) = 0$ . Thus

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \text{ when } X, Y \text{ independent.}$$

*This proves the assertion in section 2.5.*

### Proof of Theorem 4.2:

*Sufficient to prove (iii). (i) and (ii) follow directly and (iv) with some extra work.*

$$\begin{aligned} \text{iii) } \text{Var}(aX + bY) &= \mathbb{E} \left\{ aX + bY - \mathbb{E}(aX + bY) \right\}^2 \\ &= \mathbb{E} \left\{ aX + bY - a \underbrace{\mu_X}_{\mathbb{E}X} - b \underbrace{\mu_Y}_{\mathbb{E}Y} \right\}^2 \\ &= \mathbb{E} \{ a(X - \mu_X) + b(Y - \mu_Y) \}^2 \\ &= \mathbb{E} \{ a^2(X - \mu_X)^2 + b^2(Y - \mu_Y)^2 + 2ab(X - \mu_X)(Y - \mu_Y) \} \\ &= a^2 \mathbb{E}(X - \mu_X)^2 + b^2 \mathbb{E}(Y - \mu_Y)^2 + 2ab \mathbb{E}\{(X - \mu_X)(Y - \mu_Y)\} \\ &= a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{cov}(X, Y). \end{aligned}$$

## Correlation between two random variables

Let  $X$  and  $Y$  be random variables. The *correlation* between  $X$  and  $Y$  is closely related to the covariance, but it is scaled to be a number between -1 and 1.

*Definition:* The **correlation** between  $X$  and  $Y$  (also called the **correlation coefficient**, or  $\rho_{XY}$ ) is given by

$$\text{corr}(X, Y) = \rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\text{cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

$$\left( \sigma_X = \sqrt{\text{Var}(X)}, \sigma_Y = \sqrt{\text{Var}(Y)} \right).$$

The correlation measures **linear association between  $X$  and  $Y$** .

**Theorem 4.3:** The correlation coefficient  $\rho_{XY}$  has the following properties:

- i)  $-1 \leq \rho_{XY} \leq 1$ .
- ii)  $\rho_{XY}^2 = 1 \iff Y = aX + b$  for some constants  $a$  and  $b$ , where  $a > 0$  if  $\rho_{XY} = 1$ ,  $a < 0$  if  $\rho_{XY} = -1$ .
- iii) If  $X$  and  $Y$  are independent, then  $\rho_{XY} = 0$ . However, if  $\rho_{XY} = 0$  it is NOT necessarily true that  $X$  and  $Y$  are independent.  
( $\rho_{XY} = 0$  is necessary but not sufficient for independence.)

**Proof:**

- i) Let  $Z = Y - aX$ , where  $a$  is any constant.

Now for any random variable  $Z$ , we know that  $\text{Var}(Z) \geq 0$ . Thus

$$\text{Var}(Z) = \text{Var}(Y - aX) = \text{Var}(Y) + a^2\text{Var}(X) - 2a \text{cov}(X, Y) \geq 0.$$

Rearrange this to form a quadratic in  $a$ :

$$\text{Var}(Z) = a^2\sigma_X^2 - 2a \text{cov}(X, Y) + \sigma_Y^2 \geq 0.$$

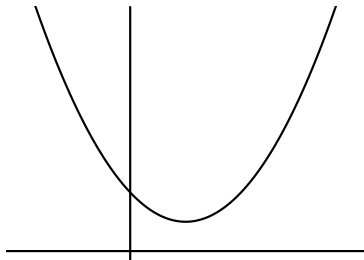
Divide by  $\sigma_X\sigma_Y$  :

$$\frac{\text{Var}(Z)}{\sigma_X\sigma_Y} = \left(\frac{\sigma_X}{\sigma_Y}\right) a^2 - (2\rho_{XY}) a + \frac{\sigma_Y}{\sigma_X} \geq 0. \quad (*)$$

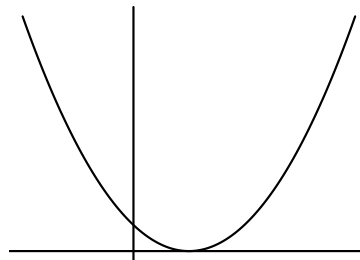
Equation (\*) holds for *all* values of  $a$ .

Now if a quadratic in  $a$  is *always*  $\geq 0$ , the quadratic never crosses the  $a$ -axis; so it must have either no real roots or exactly one real root. Recall that the number of roots of a quadratic is determined by the *discriminant*: for a quadratic in  $x$ , the standard equation is  $ax^2 + bx + c = 0$  and the discriminant is  $b^2 - 4ac$ . The equation has no real roots if and only if the discriminant is  $< 0$ . The equation has exactly one real root if and only if the discriminant equals 0.

**No real roots : discriminant  $< 0$**



**One real root : discriminant = 0**



Expression (\*) therefore indicates that the discriminant of the quadratic in (\*) must be  $\leq 0$ , so that the quadratic has either no real roots or exactly one real root. This gives

$$\begin{aligned} \text{discriminant} &= (2\rho_{XY})^2 - 4 \left(\frac{\sigma_X}{\sigma_Y}\right) \left(\frac{\sigma_Y}{\sigma_X}\right) \leq 0 \\ &4\rho_{XY}^2 - 4 \leq 0 \\ &\rho_{XY}^2 \leq 1. \end{aligned}$$

Thus

$$-1 \leq \rho_{XY} \leq 1, \quad \text{as required.}$$

- ii) We must show both that  $\rho_{XY}^2 = 1 \Rightarrow Y = aX + b$ ,  
and that  $Y = aX + b \Rightarrow \rho_{XY}^2 = 1$ .

Suppose that  $\rho_{XY}^2 = 1$ , so that  $\rho_{XY} = \pm 1$ . Recall that  $Z = Y - aX$ , and that (from (\*) overleaf),

$$\frac{\text{Var}(Z)}{\sigma_X \sigma_Y} = \left( \frac{\sigma_X}{\sigma_Y} \right) a^2 - (2\rho_{XY}) a + \frac{\sigma_Y}{\sigma_X}. \quad (*)$$

We can solve this equation to see if there are any values of  $a$  that make  $\text{Var}(Z) = 0$ . If so, then  $Z = Y - aX$  must be constant at these values of  $a$ .

Solving the quadratic (\*), we find that  $\text{Var}(Z) = 0$  implies that

$$\begin{aligned} a &= \frac{2\rho_{XY} \pm \sqrt{4\rho_{XY}^2 - 4}}{2\sigma_X/\sigma_Y} \\ &= \frac{2\rho_{XY} \pm \sqrt{4 \times 1 - 4}}{2\sigma_X/\sigma_Y} \quad (\text{because } \rho_{XY}^2 = 1) \\ &= \frac{\sigma_Y}{\sigma_X} \rho_{XY}. \end{aligned}$$

Thus, when  $\rho_{XY} = \pm 1$  and  $a = \left( \frac{\sigma_Y}{\sigma_X} \right) \rho_{XY}$ , then  $\text{Var}(Z) = \text{Var}(Y - aX) = 0$ , so  $Y - aX$  is constant and thus  $Y = aX + b$  for some constant  $b$ , as required.

Conversely, suppose that  $Y = aX + b$ . Then  $\text{Var}(Y) = \sigma_Y^2 = a^2 \sigma_X^2$ . Also,

$$\begin{aligned} \text{cov}(X, Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}\left(X(aX + b)\right) - \mathbb{E}(X)\mathbb{E}(aX + b) \\ &= a\mathbb{E}(X^2) + b\mathbb{E}(X) - \mathbb{E}(X)\left(a\mathbb{E}(X) + b\right) \\ &= a\left(\mathbb{E}(X^2) - (\mathbb{E}X)^2\right) + b\left(\mathbb{E}X - \mathbb{E}X\right) \\ &= a\left(\text{Var}(X)\right) \\ &= a\sigma_X^2. \end{aligned}$$

Thus

$$\rho_{XY}^2 = \frac{\text{cov}(X, Y)^2}{\sigma_X^2 \sigma_Y^2} = \frac{a^2 \sigma_X^4}{\sigma_X^2 (a^2 \sigma_X^2)} = 1,$$

as required.

Thus  $\rho_{XY}^2 = 1 \iff Y = aX + b$  for some constants  $a$  and  $b$ , as required.

iii) We showed on page 137 that

$$X, Y \text{ independent} \Rightarrow \text{cov}(X, Y) = 0,$$

but

$$\text{cov}(X, Y) = 0 \not\Rightarrow X, Y \text{ independent.}$$

But  $\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$ , so it follows that

$$X, Y \text{ independent} \Rightarrow \rho_{XY} = 0, \text{ but } \rho_{XY} = 0 \not\Rightarrow X, Y \text{ independent.}$$

---

## 4.4 Conditional Expectation and Conditional Variance

---

Suppose that we fix  $Y$  at the value  $y$ . We have seen that we can find the conditional distribution of  $X$  given that  $Y = y$ : for example  $X | (Y = y)$  has probability function  $f_{X|Y}(x|y)$ .

We can also find the expectation and variance of  $X$  with respect to this conditional distribution. That is, if we know that the value of  $Y$  is  $y$ , then we can find the mean value of  $X$  *given that*  $Y$  takes the value  $y$ , and also the variance of  $X$  given that  $Y = y$ .

*Definition:* Let  $X$  and  $Y$  be discrete random variables. The **conditional expectation of  $X$ , given that  $Y = y$** , is

$$\mu_{X|Y=y} = \mathbb{E}(X|Y = y) = \sum_x x f_{X|Y}(x|y).$$

$\mathbb{E}(X | Y = y)$  is the *mean value of  $X$ , when  $Y$  is fixed at  $y$* .

### Conditional expectation as a random variable

The unconditional expectation of  $X$ ,  $\mathbb{E}(X)$ , is just *a number*:  
eg.  $\mathbb{E}X = 2$  or  $\mathbb{E}X = 5.8$  (*in case you need examples of numbers*).

The conditional expectation,  $\mathbb{E}(X | Y = y)$ , is *a number depending on  $y$* :  
eg. *usually*  $\mathbb{E}(X|Y = 2)$  *will be different from*  $\mathbb{E}(X|Y = 3)$ .

We can therefore regard  $\mathbb{E}(X | Y = y)$  as a function of  $y$ , say  $\mathbb{E}(X|Y=y) = h(y)$ .

To evaluate this function,  $h(y) = \mathbb{E}(X | Y = y)$ , we:

- i) *fix  $Y$  at the chosen value  $y$ ;*
- ii) *evaluate the expectation of  $X$  when  $Y$  is fixed at this value.*

However, we could also evaluate the function at a *random value* of  $Y$ :

- i) *observe a random value of  $Y$ ;*
- ii) *fix  $Y$  at that observed random value;*
- iii) *evaluate  $\mathbb{E}(X|Y = \text{observed random value})$ .*

We obtain a random variable:  $\mathbb{E}(X|Y) = h(Y)$ .

*The randomness comes from the randomness in  $Y$ , not in  $X$ .*

*Conditional expectation,  $\mathbb{E}(X|Y)$ , is a random variable with randomness inherited from  $Y$ , not  $X$ .*

**Example:**

Suppose  $Y = \begin{cases} 1 & \text{with probability } 1/8 \\ 2 & \text{with probability } 7/8 \end{cases}$

and  $X|Y = \begin{cases} 2Y & \text{with probability } 3/4 \\ 3Y & \text{with probability } 1/4 \end{cases}$

Then  $X|(Y = 1) = \begin{cases} 2 & \text{with probability } 3/4 \\ 3 & \text{with probability } 1/4 \end{cases}$

$$\text{so, } \mathbb{E}(X|Y = 1) = 2 \times \frac{3}{4} + 3 \times \frac{1}{4} = \frac{9}{4}.$$

Then  $X|(Y = 2) = \begin{cases} 4 & \text{with probability } 3/4 \\ 6 & \text{with probability } 1/4 \end{cases}$

$$\text{so, } \mathbb{E}(X|Y = 2) = 4 \times \frac{3}{4} + 6 \times \frac{1}{4} = \frac{18}{4}.$$



$$\text{Thus } \mathbb{E}(X|Y = y) = \begin{cases} 9/4 & \text{if } y = 1 \\ 18/4 & \text{if } y = 2, \end{cases}$$

so it is a number depending on  $y$ .

$$\text{Now } \mathbb{E}(X|Y) = \begin{cases} 9/4 & \text{if } Y = 1 \text{ (probability } 1/8) \\ 18/4 & \text{if } Y = 2 \text{ (probability } 7/8) \end{cases}$$

$$\text{So } \mathbb{E}(X|Y) = \begin{cases} 9/4 & \text{with probability } 1/8 \\ 18/4 & \text{with probability } 7/8 \end{cases}$$

ie.  $\mathbb{E}(X|Y)$  is a random variable, but the randomness is inherited from  $Y$ , not from  $X$ .

The conditional variance is found in a similar manner:

*Definition:* Let  $X$  and  $Y$  be random variables. The conditional variance of  $X$ , given  $Y$ , is given by

$$\text{Var}(X|Y) = \mathbb{E}(X^2|Y) - [\mathbb{E}(X|Y)]^2 = \mathbb{E}\{(X - \mu_{X|Y})^2|Y\}$$

As with expectation,  $\text{Var}(X|Y = y)$  is a number depending on  $y$  (a function of  $y$ ), while  $\text{Var}(X|Y)$  is a random variable with randomness inherited from  $Y$ .

Conditional expectation is an extremely useful tool for finding the **unconditional** expectation of  $X$  (see Theorem 4.4 below). Just like the Partition Theorem, it is useful because it is often easier to specify conditional probabilities than to specify overall probabilities.

**Theorem 4.4: Formulae for conditional expectation and variance.**

If all the expectations below are finite, then for ANY random variables  $X$  and  $Y$ , we have:

i)  $\mathbb{E}(X) = \mathbb{E}_Y[\mathbb{E}(X|Y)]$       *Formula for Conditional Expectation: LEARN!*

*Note that we can pick any r.v.  $Y$ , to make the expectation as easy as we can.*

ii)  $\mathbb{E}(g(X)) = \mathbb{E}_Y[\mathbb{E}(g(X)|Y)]$  for any function  $g$ .

iii)  $\text{Var}(X) = \mathbb{E}_Y[\text{Var}(X|Y)] + \text{Var}_Y[\mathbb{E}(X|Y)]$

*Formula for Conditional Variance: LEARN!*

**Notes:** 1.  $\mathbb{E}_Y$  and  $\text{Var}_Y$  denote expectation over  $Y$  and variance over  $Y$ .

*ie. the expectation or variance is computed over the randomness due to the r.v.  $Y$  (see example above).*

2. The same formulae hold for discrete and continuous random variables.

*(See Theorem 4.7.)*

*Proof later: first some examples.*

## Examples of conditional expectation

### *Example 1:*

Let  $Y \sim \text{Geometric}(p)$ : so  $\mathbb{E}(Y) = \frac{1-p}{p}$ .

Let  $(X|Y) \sim \text{Poisson}(\lambda Y)$ : so  $\mathbb{E}(X|Y) = \text{Var}(X|Y) = \lambda Y$ .

$$\begin{aligned}\text{Then } \mathbb{E}(X) &= \mathbb{E}_Y[\mathbb{E}(X|Y)] \\ &= \mathbb{E}_Y(\lambda Y) \\ &= \lambda \mathbb{E}_Y(Y)\end{aligned}$$

$$\therefore \mathbb{E}(X) = \frac{\lambda(1-p)}{p}$$

### *Example 2: Sum of a random number of random variables*

Let  $N \sim \text{Poisson}(\lambda)$ , and consider the sum  $X_1 + X_2 + \dots + X_N$ , where each  $X_i \sim \text{NegBin}(k, p)$ , and  $X_1, X_2, \dots$  are independent of  $N$ .

This is a sum of a random number ( $N$ ) of random variables ( $X_1, X_2, \dots$ ).

$$\begin{aligned}\text{Then } \mathbb{E} \left\{ \sum_{i=1}^N X_i \right\} &= \mathbb{E}_N \left\{ \mathbb{E} \left( \sum_{i=1}^N X_i \mid N \right) \right\} \\ &= \mathbb{E}_N \left\{ N \times \frac{k(1-p)}{p} \right\} \quad NX_i\text{'s added together, each with mean } \frac{k(1-p)}{p}. \\ &= \frac{k(1-p)}{p} \mathbb{E}(N) \\ \mathbb{E} \left\{ \sum_{i=1}^N X_i \right\} &= \frac{k(1-p)}{p} \lambda \quad \text{because } N \sim \text{Poisson}(\lambda), \text{ so } \mathbb{E}N = \lambda.\end{aligned}$$

## General result: sum of a random number of random variables

---

If  $X_1, X_2, \dots$  each have the same mean  $\mu$ , and if  $N$  is independent of  $X_1, X_2, \dots$  then

$$\mathbb{E} \left\{ \sum_{i=1}^N X_i \right\} = (\mathbb{E}N) \times \mu.$$

*Example 3: Insect eggs on a leaf again: see example on page 128.*

Recall  $X = \# \text{ eggs laid on a leaf} \sim \text{Poisson}(\lambda)$ , so  $\mathbb{E}X = \text{Var}(X) = \lambda$   
and  $Y = \# \text{ eggs surviving to maturity}$   
and  $(Y|X) \sim \text{Binomial}(X, p)$ : so  $\mathbb{E}(Y|X) = Xp$ ,  $\text{Var}(Y|X) = Xp(1 - p)$ .

$$\begin{aligned} \text{Then } \mathbb{E}(Y) &= \mathbb{E}_X[\mathbb{E}(Y|X)] \\ &= \mathbb{E}_X(Xp) \\ &= p\mathbb{E}_X(X) \\ \mathbb{E}(Y) &= p\lambda \end{aligned}$$

$$\begin{aligned} \text{Var}(Y) &= \mathbb{E}_X(\text{Var}(Y|X)) + \text{Var}_X(\mathbb{E}(Y|X)) \\ &= \mathbb{E}_X(Xp(1 - p)) + \text{Var}_X(Xp) \\ &= p(1 - p)\mathbb{E}(X) + p^2 \text{Var}(X) \\ &= p(1 - p)\lambda + p^2\lambda \\ \text{Var}(Y) &= p\lambda \end{aligned}$$

So  $\mathbb{E}(Y) = \text{Var}(Y) = p\lambda$ , and this supports the earlier finding that  $Y \sim \text{Poisson}(p\lambda)$ .

Example 3 cont...

Calculating the covariance and correlation using conditional expectation:

$\text{cov}(X, Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y)$ . We know that  $\mathbb{E}X = \lambda$ ,  $\mathbb{E}Y = \lambda p$ .

To find  $\mathbb{E}(XY)$ , once again use the formula for conditional expectation:

$$\mathbb{E}(XY) = \mathbb{E}_X[\mathbb{E}(XY|X)] = \mathbb{E}_X[X\mathbb{E}(Y|X)],$$

because, conditional on  $X$ , we can take  $X$  outside the expectation like a constant.

But  $\mathbb{E}(Y|X) = Xp$ , so  $\mathbb{E}(XY) = \mathbb{E}_X(X \times Xp) = p\mathbb{E}(X^2)$ .

*Trick for calculating  $\mathbb{E}(X^2)$ : use*

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}X)^2 \\ \Rightarrow \mathbb{E}(X^2) &= \text{Var}(X) + (\mathbb{E}X)^2\end{aligned}$$

Thus  $\mathbb{E}(XY) = p\mathbb{E}(X^2) = p(\text{Var}(X) + (\mathbb{E}X)^2) = p(\lambda + \lambda^2)$ .

$$\begin{aligned}\text{So } \text{cov}(X, Y) &= \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y) \\ &= p(\lambda + \lambda^2) - \lambda \times p\lambda \\ \text{cov}(X, Y) &= p\lambda\end{aligned}$$

$$\begin{aligned}\text{Finally, } \text{corr}(X, Y) = \rho_{XY} &= \frac{\text{cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} \\ &= \frac{p\lambda}{\sqrt{\lambda \times p\lambda}} \\ &= \frac{p}{\sqrt{p}} \\ \Rightarrow \text{corr}(X, Y) &= \sqrt{p}.\end{aligned}$$

### Proof of Theorem 4.4:

(i) is a special case of (ii).

(ii) Wish to show that  $\mathbb{E}(g(X)) = \mathbb{E}_Y\left(\mathbb{E}(g(X) | Y)\right)$ , for any function  $g$ .

*Begin at RHS:*

$$\begin{aligned}\mathbb{E}_Y\left[\mathbb{E}(g(X)|Y)\right] &= \mathbb{E}_Y\left[\sum_x g(x)\mathbb{P}(X = x|Y)\right] \\ &= \sum_y \left[\sum_x g(x)\mathbb{P}(X = x|Y = y)\right] \mathbb{P}(Y = y) \\ &= \sum_y \sum_x g(x)\mathbb{P}(X = x|Y = y)\mathbb{P}(Y = y) \\ &= \sum_x g(x) \sum_y \mathbb{P}(X = x|Y = y)\mathbb{P}(Y = y) \\ &= \sum_x g(x)\mathbb{P}(X = x) \quad (\text{partition rule}) \\ &= \mathbb{E}(g(X)).\end{aligned}$$

(iii) *Wish to prove*  $\text{Var}(X) = \mathbb{E}_Y[\text{Var}(X|Y)] + \text{Var}_Y[\mathbb{E}(X|Y)]$

*Begin at RHS:*

$$\begin{aligned}\mathbb{E}_Y[\text{Var}(X|Y)] + \text{Var}_Y[\mathbb{E}(X|Y)] \\ &= \mathbb{E}_Y\left\{\mathbb{E}(X^2|Y) - (\mathbb{E}(X|Y))^2\right\} + \underbrace{\mathbb{E}_Y\left\{[\mathbb{E}(X|Y)]^2\right\}}_{\text{by definitions}} - \underbrace{\left[\mathbb{E}_Y(\mathbb{E}(X|Y))\right]^2}_{\mathbb{E}(X) \text{ by part (i)}} \\ &= \underbrace{\mathbb{E}_Y\{\mathbb{E}(X^2|Y)\}}_{\mathbb{E}(X^2) \text{ by part (i)}} - \mathbb{E}_Y\{[\mathbb{E}(X|Y)]^2\} + \mathbb{E}_Y\{[\mathbb{E}(X|Y)]^2\} - (\mathbb{E}X)^2\end{aligned}$$

*giving*

$$\begin{aligned}RHS &= \mathbb{E}_Y[\text{Var}(X|Y)] + \text{Var}_Y[\mathbb{E}(X|Y)] \\&= \mathbb{E}(X^2) - (\mathbb{E}X)^2 \\&= \text{Var}(X) \quad \text{as required.}\end{aligned}$$

---

## 4.5 Examples of discrete multivariate distributions

### 1. Multinomial distribution

Recall the Binomial distribution from Chapter 2:

- $n$  independent trials;
- 2 possible outcomes per trial;
- $\mathbb{P}(\text{success}) = \text{constant} = p$ ;
- $X =$  number of successes. Then  $X \sim \text{Binomial}(n, p)$   
and  $\mathbb{P}(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$ .

Now consider the following situation:

- $n$  independent trials;
- $k$  possible outcomes per trial;
- $\mathbb{P}(\text{outcome } i) = p_i$  (constant) where  $\sum_{i=1}^k p_i = 1$ .
- $\mathbf{X} = (X_1, \dots, X_k)$ , where  $X_i =$  # trials with outcome  $i$ . Then  $\mathbf{X} = (X_1, \dots, X_k)$  has a **Multinomial distribution** with parameters  $n =$  # trials,  $p_1, \dots, p_k$ .

We write:

$$\mathbf{X} \sim \text{Multinomial}(n; p_1, \dots, p_k).$$

**Example:** Throwing paper darts in lectures. Each dart:

- hits the lecturer with probability 0.2;
- hits another student with probability 0.5;
- *self-destructs* with probability 0.3.

Throw 7 darts.

Let  $\mathbf{X} = (X_1, X_2, X_3) = (\# \text{ hit lecturer}, \# \text{ hit another student}, \# \text{ self destructed})$

Then  $\mathbf{X} \sim \text{Multinomial}(7; 0.2, 0.5, 0.3)$ .

## Probability function for Multinomial distribution

$$f_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$$

for  $x_i = 0, \dots, n \forall_i$ , and  $\sum_{i=1}^k x_i = n$ ,

and where  $p_i \geq 0 \forall_i$ ,  $\sum_{i=1}^k p_i = 1$ .

**Notes:**

1)  $\sum_{x_1} \sum_{x_2} \dots \sum_{x_k} f(x_1, \dots, x_k) = (p_1 + \dots + p_k)^n = 1^n = 1$ .

2) The marginal distributions are  $X_i \sim \text{Binomial}(n, p_i)$ , because we can reduce the situation to 2 outcomes: “*i*” and “not *i*”.



3) Similarly,  $X_i + X_j \sim \text{Binomial}(n, p_i + p_j)$  if  $i \neq j$ .

4) Because  $X_i \sim \text{Binomial}(n, p_i)$ , we have  $\mathbb{E}(X_i) = np_i$ ,  $\text{Var}(X_i) = np_i(1 - p_i)$ .

**Example:** Blood types in New Zealand have the following frequencies:

	A	B	AB	O	Total
Maori	0.51	0.04	0.02	0.43	1
Non-Maori	0.40	0.10	0.03	0.47	1

Given a random sample of size 30 from each population, find the probability that:  $\#(A) = 10$      $\#(B) = 4$      $\#(AB) = 1$      $\#(O) = 15$ .

Solution: For Maori population,  $\mathbf{X} \sim \text{Multinomial}(30; 0.51, 0.04, 0.02, 0.43)$

$$\begin{aligned} f_{\mathbf{X}}(10, 4, 1, 15) &= \frac{30!}{10!4!1!15!} (0.51)^{10} (0.04)^4 (0.02)^1 (0.43)^{15} \\ &= 0.00045 \end{aligned}$$

For non-Maori population,  $\mathbf{X} \sim \text{Multinomial}(30; 0.40, 0.10, 0.03, 0.47)$

$$\begin{aligned} f_{\mathbf{X}}(10, 4, 1, 15) &= \frac{30!}{10!4!1!15!} (0.40)^{10} (0.10)^4 (0.03)^1 (0.47)^{15} \\ &= 0.0088 \end{aligned}$$

### Covariance and correlation of $X_i, X_j$

If  $\mathbf{X} \sim \text{Multinomial}(n; p_1, p_2, \dots, p_k)$ , then

$$\text{cov}(X_i, X_j) = -np_i p_j \qquad \text{corr}(X_i, X_j) = -\sqrt{\frac{p_i p_j}{(1 - p_i)(1 - p_j)}}$$

**Note:** Negative correlation makes sense: the more outcomes that fall into category  $i$ , the fewer there are to fall into category  $j$ .

## 2. Multivariate Hypergeometric distribution

Recall the Hypergeometric distribution from Chapter 2:

- $N$  balls in a jar;
- 2 colours:  $M$  balls black,  $(N - M)$  balls white;
- Sample  $n$  balls *without replacement*;
- $X$  = number of black balls in the sample of size  $n$ .

Then  $X \sim \text{Hypergeometric}(N, M, n)$ .  $\mathbb{P}(X = x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}$ .

The multivariate hypergeometric distribution is similar, but there are balls of  $k$  different colours instead of just 2 *different colours*.

### Multivariate hypergeometric distribution:

- $N$  balls in a jar;
- $k$  colours:  $M_i$  balls with colour  $i$ , where  $\sum_{i=1}^k M_i = N$ .
- Sample  $n$  balls without replacement.
- Let  $\mathbf{X} = (X_1, X_2, \dots, X_k)$  where  $X_i = \#$ balls of colour  $i$  in sample of size  $n$ .

Then  $\mathbf{X} \sim \text{Multivariate Hypergeometric}(N; M_1, \dots, M_k; n)$ .

### Probability function:

$$\mathbb{P}(X_1 = x_1, \dots, X_k = x_k) = \frac{\prod_{i=1}^k \binom{M_i}{x_i}}{\binom{N}{n}} \quad \text{for } x_i = 0, \dots, M_i \quad \forall i.$$

### Marginal distributions:

The marginal distribution of  $X_i$  is Hypergeometric  $(N, M_i, n)$ .

## 4.6 Continuous joint distributions

The random vector  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  has a continuous joint distribution if  $X_1, X_2, \dots, X_n$  are each continuous random variables, and they interact ‘nicely’. We define this formally below.

### Joint distribution functions and probability density functions

*Definition:* Let  $\mathbf{X} = (X_1, X_2, \dots, X_k)$  be a random vector. The **joint distribution function** of  $\mathbf{X}$  is

$$F_{\mathbf{X}}(\mathbf{x}) = F_{\mathbf{X}}(x_1, \dots, x_k) = \mathbb{P}(X_1 \leq x_1, \dots, X_k \leq x_k).$$

*Definition:* Let  $\mathbf{X} = (X_1, X_2, \dots, X_k)$  be a random vector with joint distribution function  $F(x_1, x_2, \dots, x_k) = \mathbb{P}(X_1 \leq x_1, \dots, X_k \leq x_k)$ . Then  $\mathbf{X}$  has a **continuous joint distribution** if  $F$  is continuous, and if the partial derivative  $\frac{\partial^k F}{\partial x_1 \dots \partial x_k}$  exists, except possibly on a  $(k - 1)$ -dimensional subset of  $\mathbb{R}^k$ .

*Definition:* Let  $\mathbf{X} = (X_1, X_2, \dots, X_k)$  have a continuous joint distribution. The **joint probability density function** of  $\mathbf{X}$ , or simply **joint density** of  $\mathbf{X}$ , is given by

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}}(x_1, \dots, x_k) = \frac{\partial^k F(x_1, \dots, x_k)}{\partial x_1 \dots \partial x_k} \quad (\text{partial derivative})$$

The joint density is used to find probabilities by integration. In the univariate case, for a set  $A \subseteq \mathbb{R}$  (i.e.  $A = (a, b)$  for some  $a$  and  $b$ ), we have

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

In the multivariate case, we have for  $A \subseteq \mathbb{R}^k$ ,

$$\mathbb{P}(\mathbf{X} \in A) = \int \int \dots \int_A f_{\mathbf{X}}(x_1, \dots, x_k) dx_k \dots dx_2 dx_1.$$

## Properties of the joint density function

i)  $f_{\mathbf{X}}(x_1, \dots, x_k) \geq 0$  for all  $x_1, \dots, x_k$ .

ii)  $\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, \dots, x_k) dx_k \dots dx_1 = 1$  (total probability=1)

iii)

$$F_{\mathbf{X}}(x_1, \dots, x_k) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_k} f_{\mathbf{X}}(y_1, \dots, y_k) dy_k \dots dy_1$$

(immediate from definitions)

iv) For any reasonable region  $A \subseteq \mathbb{R}^k$ ,

$$\mathbb{P}((X_1, \dots, X_k) \in A) = \int_A f_{\mathbf{X}}(x_1, \dots, x_k) dx_k \dots dx_1$$

Practical use of the joint density.

Conversely, the *conditions* required for  $f(x_1, \dots, x_k)$  to be a valid joint density are:

i)  $f(x_1, \dots, x_k) \geq 0$  for all  $x_1, \dots, x_k$ .

ii)  $\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, \dots, x_k) dx_k \dots dx_1 = \int_{\mathbb{R}^k} f(x_1, \dots, x_k) dx_k \dots dx_1 = 1$

**Example 1:** Let  $\mathbf{X} \in \mathbb{R}^2$  have joint density  $f(x, y) = \begin{cases} 1 & (0 \leq x \leq 1, 0 \leq y \leq 1), \\ 0 & \text{otherwise.} \end{cases}$

a) Show that  $f(x, y)$  is a valid joint density.

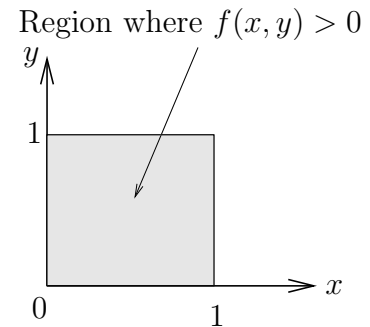
b) Find the joint distribution function,  $F(x, y)$ .

c) Find  $\mathbb{P}(X + Y \leq 1)$ .

a) i)  $f(x, y) \geq 0 \forall x, y$  by definition.

ii) Check  $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dy dx = 1$ :

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dy dx &= \int_0^1 \int_0^1 1 dy dx \\ &= \int_0^1 [y]_0^1 dx \\ &= \int_0^1 1 dx \\ &= [x]_0^1 = 1. \end{aligned}$$



So  $f(x, y)$  is a valid p.d.f. by (i) and (ii).

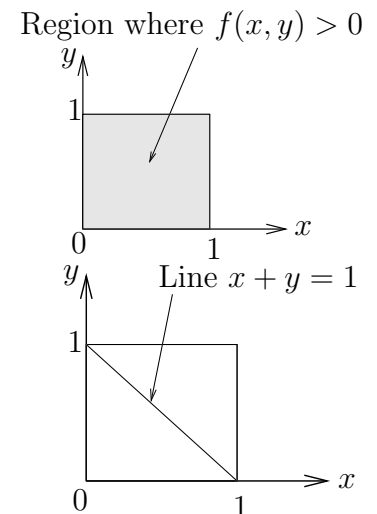
$$\begin{aligned} \text{b)} \quad F(x, y) &= \int_{-\infty}^x \int_{-\infty}^y f(u, v) dv du \\ &= \int_0^x \int_0^y 1 dv du \quad \text{for} \quad \begin{cases} 0 \leq x \leq 1, \\ 0 \leq y \leq 1 \end{cases} \\ &= \int_0^x y du \\ &= y [u]_0^x \end{aligned}$$

$$F(x, y) = xy \quad \text{for} \quad \begin{cases} 0 \leq x \leq 1, \\ 0 \leq y \leq 1. \end{cases}$$

c) To find  $\mathbb{P}(X+Y \leq 1)$ , we need to do a double integration of the joint density over the correct region. Follow the following steps:

1) Draw the area where  $f_{X,Y}(x, y) > 0$ :  
this shows where we have  
to restrict our attention.

2) We need to find the region where **BOTH**  
 $f_{X,Y}(x, y) > 0$  AND  $x + y \leq 1$ . Draw on the diagram  
the boundary line for  $x + y \leq 1$ : ie. the line  $x + y = 1$ .



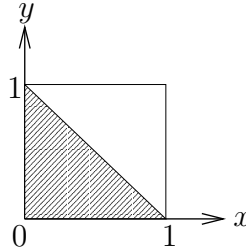
3) Work out *which side of the line corresponds to  $x + y \leq 1$* .

(Note: this requires care! People often make mistakes at this stage because it seems easy.)

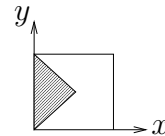
If in doubt, pick a point on one side of the line and test the condition:

e.g.  $(x, y) = (0, 0) \Rightarrow x + y < 1$ , so we want the area below the line.

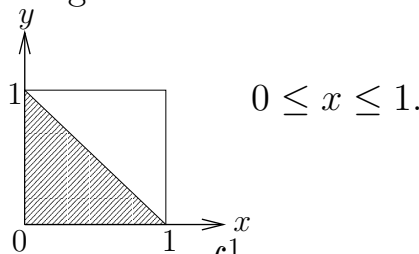
Shade this area.



4) We need to find the limits of integration that match this area. *Select one variable,  $x$  or  $y$ , to “lead”*:  $x$  is often more natural, but it can be easier to use  $y$  instead if the area follows the  $y$ -axis but not the  $x$ -axis, eg.



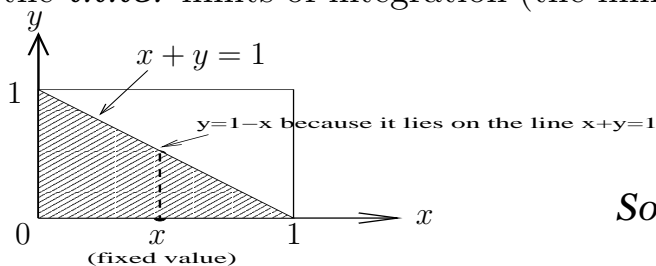
Here we select  $x$ . First find the range of values of  $x$  that lie inside the shaded area:



This gives the *outer* limits of integration:  $\int_{x=0}^1$

To find the inner limits of integration, we need to see how  $y$  varies for *any given value of  $x$* .

Fix a typical value of  $x$ , and mark it on the diagram. Find the *range of values of  $y$  that lie in the shaded area, for this fixed value of  $x$* . This gives the *inner* limits of integration (the limits for  $y$ ).



So as  $x$  ranges from 0 to 1,

$y$  ranges from 0 to  $1 - x$ .

So the limits of integration are  $\int_{x=0}^1 \int_{y=0}^{1-x}$ .

5) Perform the integration using the limits just obtained:

$$\begin{aligned}
 \mathbb{P}(X + Y \leq 1) &= \int_{x=0}^1 \int_{y=0}^{1-x} f_{X,Y}(x, y) dy dx \\
 &= \int_{x=0}^1 \int_{y=0}^{1-x} 1 dy dx \\
 &= \int_{x=0}^1 \left[ y \right]_{y=0}^{1-x} dx \\
 &= \int_{x=0}^1 (1 - x) dx \\
 &= \left[ x - \frac{x^2}{2} \right]_0^1 \\
 \mathbb{P}(X + Y \leq 1) &= \frac{1}{2}
 \end{aligned}$$

**Example 2:** Suppose  $(X, Y)$  have joint density  $f_{X,Y}(x, y) = \begin{cases} e^{-x-y} & (x, y \geq 0), \\ 0 & \text{otherwise.} \end{cases}$

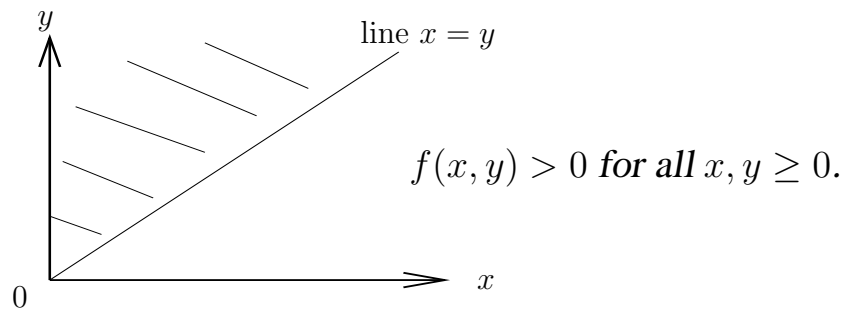
a) Find  $F_{X,Y}(x, y)$ .

b) Find  $\mathbb{P}(X \leq Y)$ .

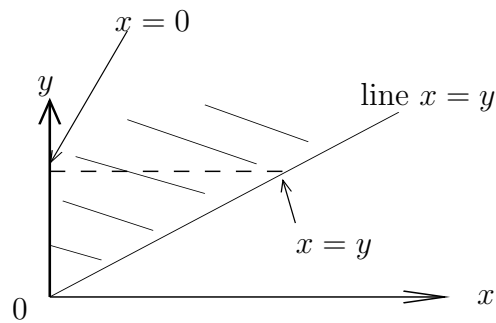
$$\begin{aligned}
 \text{a) } F_{X,Y}(x, y) &= \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(u, v) dv du = \int_0^x \int_0^y e^{-u-v} dv du \text{ for } x, y \geq 0 \\
 &= \int_0^x \int_0^y e^{-u} e^{-v} dv du \\
 &= \int_0^x e^{-u} \left\{ \int_0^y e^{-v} dv \right\} du \\
 &= \int_0^x e^{-u} \left[ -e^{-v} \right]_0^y du \\
 &= \int_0^x e^{-u} (1 - e^{-y}) du \\
 &= (1 - e^{-y}) \left[ -e^{-u} \right]_0^x
 \end{aligned}$$

$$F_{X,Y}(x, y) = (1 - e^{-y})(1 - e^{-x}) \text{ for } x, y \geq 0.$$

b) Looking for  $\mathbb{P}(X \leq Y)$ : boundary line  $x = y$ .



Select variable  $y$  to “lead”, because area follows the  $y$ -axis.



$y$  ranges from 0 to  $\infty$ :  $\int_{y=0}^{\infty}$ .

For fixed  $y$ ,  $x$  ranges from 0 to  $y$ :  $\int_{y=0}^{\infty} \int_{x=0}^y$ .

$$\begin{aligned}
 \text{So } \mathbb{P}(X \leq Y) &= \int_{y=0}^{\infty} \int_{x=0}^y f(x, y) dx dy \\
 &= \int_{y=0}^{\infty} \int_{x=0}^y e^{-x-y} dx dy \\
 &= \int_{y=0}^{\infty} e^{-y} \left[ -e^{-x} \right]_{x=0}^y dy \\
 &= \int_{y=0}^{\infty} e^{-y} (1 - e^{-y}) dy \\
 &= \int_{y=0}^{\infty} (e^{-y} - e^{-2y}) dy \\
 &= \left[ -e^{-y} + \frac{1}{2}e^{-2y} \right]_0^{\infty} \\
 &= e^0 - \frac{1}{2}e^0 \\
 \mathbb{P}(X \leq Y) &= \frac{1}{2}.
 \end{aligned}$$

Makes sense by the symmetry of  $X$  and  $Y$ .



## Marginal and Conditional Densities

All the concepts for discrete random variables have continuous analogues, although the ideas can be less intuitive in the continuous case.

For simplicity, we restrict attention to bivariate random vectors:  $\mathbf{X} = (X, Y)$ .

*Definition:* Suppose that  $\mathbf{X} = (X, Y)$  has a continuous joint distribution with joint density  $f(x, y)$ . The marginal p.d.f. of  $X$ , or the marginal density of  $X$ , is

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

Similarly, the marginal density of  $Y$  is

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

**Note:** Compare with the discrete case:  $f_X(x) = \sum_y f(x, y)$ .

*To get from discrete to continuous, replace probability functions by pdfs, and replace  $\sum$ 's with  $\int$ 's. The idea is the same: eliminate all but the required argument through summing / integration.*

*Definition:* If  $\mathbf{X} = (X, Y)$  has a continuous joint distribution, then the conditional density of  $X$  given  $Y$  is defined as

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \quad \text{as long as } f_Y(y) > 0.$$

### Justifications:

To justify results for continuous random variables, we generally use the distribution function. The distribution function gives us probabilities, which we understand how to manipulate. Working directly with the probability density function is harder, because it is difficult to conceptualize how it should behave.

## Justification of the marginal density:

Consider the distribution function of  $X$ :

$$\begin{aligned} F_X(x) = \mathbb{P}(X \leq x) &= \mathbb{P}(X \leq x \text{ and } -\infty < Y < \infty) \\ &= F_{X,Y}(x, \infty) && \text{(by definition of } F_{X,Y}\text{)} \\ &= \int_{-\infty}^x \int_{-\infty}^{\infty} f_{X,Y}(u, y) dy du \\ &= \int_{-\infty}^x g(u) du, \quad \text{say, where } g(u) = \int_{-\infty}^{\infty} f_{X,Y}(u, y) dy. \end{aligned}$$

Thus  $X$  has marginal density

$$f_X(x) = F'_X(x) = g(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$

Similarly,  $Y$  has marginal density  $f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$ .

## Justification of the conditional density:

The exact meaning of the conditional density  $f_{X|Y}(x|y)$  is hard to understand. We cannot work with conditional probabilities of the form  $\mathbb{P}(X \leq x | Y = y)$ , because the event  $\{Y = y\}$  has probability zero so we cannot condition on it. Instead, we must resort to limiting arguments.

Define the limiting conditional distribution function as

$$\begin{aligned} F_{X|Y}(x|y) &= \lim_{h \rightarrow 0} \mathbb{P}(X \leq x | y - h \leq Y \leq y + h) \\ &= \lim_{h \rightarrow 0} \left\{ \frac{\int_{-\infty}^x \int_{y-h}^{y+h} f_{X,Y}(u, v) dv du}{\int_{y-h}^{y+h} f_Y(v) dv} \right\} \\ &= \lim_{h \rightarrow 0} \left\{ \frac{\int_{-\infty}^x 2h f_{X,Y}(u, y) du}{2h f_Y(y)} \right\}, \end{aligned}$$

because  $\int_{y-h}^{y+h} f_{X,Y}(u, v) dv \rightarrow 2h f(u, y)$  and  $\int_{y-h}^{y+h} f_Y(v) dv \rightarrow 2h f_Y(y)$  as  $h \rightarrow 0$ .

So  $F_{X,Y}(x|y) = \frac{\int_{-\infty}^x f_{X,Y}(u, y) du}{f_Y(y)}$ . Taking the derivative to find the conditional p.d.f., we obtain

$$f_{X|Y}(x|y) = \frac{d}{dx} \left( F_{X|Y}(x|y) \right) = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \text{ as long as } f_Y(y) > 0.$$

**Notes:** 1. These justifications are not rigorous proofs. For a full treatment, Measure Theory is needed.

2. When calculating marginal and conditional densities, great attention must be paid to the **limits of integration**, just as for calculating probabilities.

**Example 1:** Let  $(X, Y)$  have joint density  $f_{X,Y}(x, y) = \begin{cases} \lambda^2 e^{-\lambda y} & (0 \leq x \leq y), \\ 0 & \text{otherwise.} \end{cases}$

a) Find the marginal p.d.f. of  $X$ ,  $f_X(x)$ .

b) Find the marginal p.d.f. of  $Y$ ,  $f_Y(y)$ .

c) Find the conditional density,  $f_{X|Y}(x|y)$ .

**Solution**

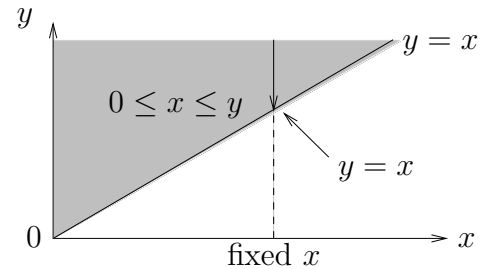
a)  $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy$

$= \int_x^{\infty} \lambda^2 e^{-\lambda y} dy$  because  $f_{X,Y}(x, y) = 0$  if  $y < x$ .

$= \left[ -\frac{\lambda^2}{\lambda} e^{-\lambda y} \right]_x^{\infty}$

$= (-\lambda e^{-\infty} + \lambda e^{-\lambda x})$

$f_X(x) = \lambda e^{-\lambda x}$  for  $x \geq 0$ . So  $X \sim \text{Exponential}(\lambda)$ .

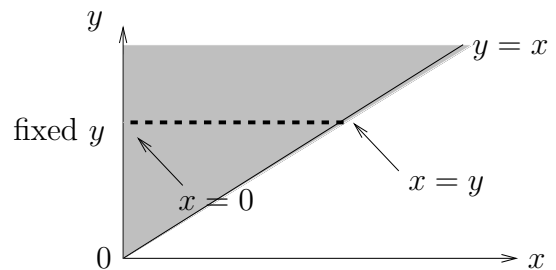


b)  $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx$

$= \int_0^y \lambda^2 e^{-\lambda y} dx$  because  $f_{X,Y}(x, y) = 0$  if  $x > y$ .

$= \lambda^2 e^{-\lambda y} \left[ x \right]_0^y$

$f_Y(y) = \lambda^2 y e^{-\lambda y}$  for  $y \geq 0$ .

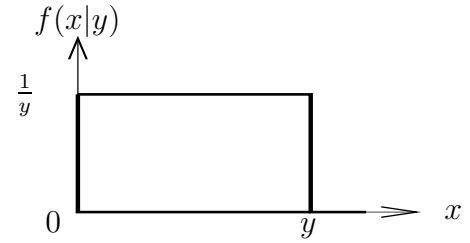


So  $Y \sim \text{Gamma}(k = 2, \lambda)$ .

$$c) f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\lambda^2 e^{-\lambda y}}{\lambda^2 y e^{-\lambda y}} \quad \text{for } 0 \leq x \leq y.$$

$$f_{X|Y}(x|y) = \frac{1}{y} \quad \text{for } 0 \leq x \leq y.$$

Thus  $(X|Y) \sim \text{Uniform}[0, Y]$ .



**Example 2:** Let  $(X, Y)$  have joint density

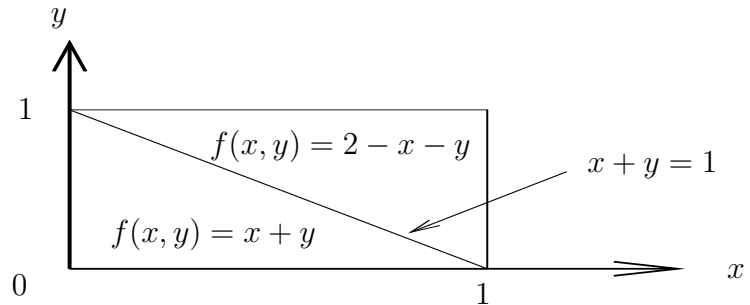
$$f_{X,Y}(x,y) = \begin{cases} x+y & \text{for } 0 \leq x \leq 1; \quad 0 \leq y \leq 1; \quad x+y \leq 1; \\ 2-x-y & \text{for } 0 \leq x \leq 1; \quad 0 \leq y \leq 1; \quad x+y > 1; \\ 0 & \text{otherwise.} \end{cases}$$

a) Find the marginal p.d.f. of  $X$ ,  $f_X(x)$ .

b) Find the marginal p.d.f. of  $Y$ ,  $f_Y(y)$ .

c) Find the conditional density,  $f_{X|Y}(x|y)$ .

**Solution** First draw picture:



$$a) f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy \quad \text{where } f_{X,Y}(x,y) = \begin{cases} 0 & \text{for } y < 0 \\ 0 & \text{for } y > 1 \\ x+y & \text{for } 0 \leq y \leq 1-x \\ 2-x-y & \text{for } 1-x < y \leq 1 \end{cases}$$

$$= \int_0^{1-x} (x+y) dy + \int_{1-x}^1 (2-x-y) dy$$

$$= \left[ xy + \frac{y^2}{2} \right]_0^{1-x} + \left[ 2y - xy - \frac{y^2}{2} \right]_{1-x}^1$$

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

$$f_X(x) = -x^2 + x + \frac{1}{2} \quad \text{for } 0 \leq x \leq 1.$$

b) By symmetry,  $X$  and  $Y$  have the same marginal distribution, because  $x$  and  $y$  are treated identically in  $f_{X,Y}(x, y)$ .

Thus  $f_Y(y) = -y^2 + y + \frac{1}{2}$  ( $0 \leq y \leq 1$ ).

$$c) \quad f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \begin{cases} \frac{x+y}{-y^2 + y + \frac{1}{2}} & \text{if } x+y \leq 1 \\ \frac{2-x-y}{-y^2 + y + \frac{1}{2}} & \text{if } x+y > 1. \end{cases} \quad \text{for } x, y \in [0, 1].$$


---

## 4.7 Independence of continuous random variables

---

Recall that **discrete** random variables  $X$  and  $Y$  are statistically independent if and only if

$$\mathbb{P}(X = x, Y = y) = f_{X,Y}(x, y) = f_X(x)f_Y(y) = \mathbb{P}(X = x)\mathbb{P}(Y = y).$$

The analogous definition holds for continuous random variables.

*Definition:* Let  $(X, Y)$  be jointly continuous random variables with joint density  $f_{X,Y}$  and marginal densities  $f_X$  and  $f_Y$ . Then  $X$  and  $Y$  are **statistically independent** if and only if

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \quad \text{for all } x, y.$$

**Theorem 4.5:**  $X$  and  $Y$  are statistically independent if and only if

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \quad \text{for all } x, y, \text{ where } F \text{ denotes distribution function.}$$

**Proof:**

First suppose that  $F_{X,Y}(x, y) = F_X(x)F_Y(y)$ . Then

$$\begin{aligned} f_{X,Y}(x, y) &= \frac{\partial^2}{\partial x \partial y} \left( F_{X,Y}(x, y) \right) = \frac{\partial^2}{\partial x \partial y} \left( F_X(x)F_Y(y) \right) = \left( \frac{\partial F_X(x)}{\partial x} \right) \left( \frac{\partial F_Y(y)}{\partial y} \right) \\ &= f_X(x)f_Y(y). \end{aligned}$$

Thus  $X$  and  $Y$  are statistically independent, by definition.

Conversely, suppose that  $X$  and  $Y$  are statistically independent.

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

$$\begin{aligned} F_{X,Y}(x, y) &= \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(u, v) dv du = \int_{-\infty}^x \int_{-\infty}^y f_X(u)f_Y(v) dv du \\ &= \left( \int_{-\infty}^x f_X(u) du \right) \left( \int_{-\infty}^y f_Y(v) dv \right) \\ &= F_X(x) F_Y(y), \end{aligned}$$

as required. □

---

**Theorem 4.6:**

*Continuous random variables  $X$  and  $Y$  are independent if and only if their joint density  $f_{X,Y}(x, y)$  can be written as a product  $f_{X,Y}(x, y) = g(x)h(y)$  for some functions  $g$  and  $h$ , and for ALL  $x, y \in \mathbb{R}$ .*

*If  $f_{X,Y}(x, y) = g(x)h(y)$ , then the marginal densities are*

$$f_X(x) = \frac{g(x)}{\int_{-\infty}^{\infty} g(u) du}, \quad f_Y(y) = \frac{h(y)}{\int_{-\infty}^{\infty} h(u) du}.$$

**Proof:**

As for discrete case (Theorem 4.1), but with sums  $\sum_x \sum_y$  replaced by integrals  $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty}$ . □

## Using the joint density to determine whether $X$ and $Y$ are independent

---

The following is a common exam question:

$$\text{Let } f_{X,Y}(x, y) = \begin{cases} \dots\dots & \text{for } (x, y) \in \text{region } A, \\ 0 & \text{otherwise.} \end{cases}$$

Are  $X$  and  $Y$  independent?

### Solution:

We use Theorem 4.6, but to do this we need to find a single expression for  $f_{X,Y}(x, y)$  that holds for **all**  $(x, y) \in \mathbb{R}^2$ .

Define the **indicator function**:  $I\{(x, y) \in A\} = \begin{cases} 1 & \text{for } (x, y) \in A \\ 0 & \text{otherwise.} \end{cases}$

*Solve the question by seeing if  $f_{X,Y}(x, y)I\{(x, y) \in A\}$  can factorize into  $g(x)h(y)$ .*

*Sometimes, it is possible to factorize  $I\{(x, y) \in A\} = I\{x \in A_x\}I\{y \in A_y\}$ ,*

*eg.  $I\{0 \leq x \leq 1, 0 \leq y \leq 1\} = I\{0 \leq x \leq 1\}I\{0 \leq y \leq 1\}$ .*

*Other times we cannot factorize  $I\{(x, y) \in A\}$ ,*

*eg.  $I\{0 \leq x \leq y \leq 1\}$  cannot be factorized.*

---

**Example 1:** Let  $f_{X,Y}(x, y) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1; 0 \leq y \leq 1; \\ 0 & \text{otherwise.} \end{cases}$

*Then*

$$\begin{aligned} f_{X,Y}(x, y) &= 1 \times I\{0 \leq x \leq 1\} \times I\{0 \leq y \leq 1\} \\ &= g(x) \times h(y) \text{ for } x, y \in \mathbb{R}. \end{aligned}$$

So  $X$  and  $Y$  are independent.

**Example 2:** Let  $f_{X,Y}(x, y) = \begin{cases} e^{-x-y} & \text{for } x, y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$

Then

$$\begin{aligned} f_{X,Y}(x, y) &= e^{-x-y} \times I\{x \geq 0\} \times I\{y \geq 0\} \\ &= (e^{-x}I\{x \geq 0\})(e^{-y}I\{y \geq 0\}) \\ &= g(x) \times h(y) \text{ for } x, y \in \mathbb{R}. \end{aligned}$$

So  $X$  and  $Y$  are independent.

**Example 3:** Let  $f_{X,Y}(x, y) = \begin{cases} \lambda^2 e^{-\lambda y} & \text{for } 0 \leq x \leq y, \\ 0 & \text{otherwise.} \end{cases}$

Then  $f_{X,Y}(x, y) = \lambda^2 e^{-\lambda y} \underbrace{I\{0 \leq x \leq y\}}_{\text{does not factorize}}$

So  $X$  and  $Y$  are NOT independent.

**Example 4:** Let  $(X, Y)$  have joint density

$$f_{X,Y}(x, y) = \begin{cases} x + y & \text{for } 0 \leq x \leq 1; \quad 0 \leq y \leq 1; \quad x + y \leq 1; \\ 2 - x - y & \text{for } 0 \leq x \leq 1; \quad 0 \leq y \leq 1; \quad x + y > 1; \\ 0 & \text{otherwise.} \end{cases}$$

This gives,

$$\begin{aligned} f_{X,Y}(x, y) &= \left\{ (x + y)I\{x + y \leq 1\} + (2 - x - y)I\{x + y > 1\} \right\} \\ &\quad \times I\{0 \leq x \leq 1\} \times I\{0 \leq y \leq 1\} \end{aligned}$$

Thus  $X$  and  $Y$  are NOT independent.



## 4.8 Expectation of jointly continuous random variables

---

*Definition:* Let  $X_1, X_2, \dots, X_k$  be jointly continuous random variables with joint density  $f(x_1, \dots, x_k)$ . Let  $g : \mathbb{R}^k \rightarrow \mathbb{R}$  be a (nice enough) function. Then

$$\mathbb{E}(g(X_1, \dots, X_k)) = \int_{x_1=-\infty}^{\infty} \dots \int_{x_k=-\infty}^{\infty} g(x_1, \dots, x_k) f(x_1, \dots, x_k) dx_k \dots dx_1$$

*Example:* for two variables,

$$\mathbb{E}(g(X, Y)) = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} g(x, y) f(x, y) dy dx.$$

### Properties of expectation for continuous random variables

---

All properties of expectation are exactly the same for continuous random variables as they are for discrete random variables. For proofs of the statements below, see the proofs for the discrete case on page 134, and replace sums  $\sum_x$  with integrals  $\int_{-\infty}^{\infty}$  where necessary.

i) If  $\mathbf{X} = (X_1, X_2, \dots, X_k)$  is a  $k$ -variate continuous random variable, then for any constants  $a$  and  $b$  and any functions  $g$  and  $h$ , ( $g : \mathbb{R}^k \rightarrow \mathbb{R}, h : \mathbb{R}^k \rightarrow \mathbb{R}$ ),

$$\mathbb{E}(ag(\mathbf{X}) + bh(\mathbf{X})) = a\mathbb{E}(g(\mathbf{X})) + b\mathbb{E}(h(\mathbf{X})).$$

ii) For **any** continuous random variables  $X$  and  $Y$ ,

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

Consequently, for any continuous random variables  $X_1, \dots, X_k$ ,

$$\mathbb{E}(X_1 + \dots + X_k) = \mathbb{E}(X_1) + \dots + \mathbb{E}(X_k).$$

Note that we do **not** require  $X_1, \dots, X_k$  to be independent.

iii) If  $X$  and  $Y$  are **independent**, and  $g, h$  are functions, ( $g, h : \mathbb{R}^k \rightarrow \mathbb{R}$ ), **then**

$$\mathbb{E}(XY) = (\mathbb{E}X)(\mathbb{E}Y)$$

**and**

$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$$

Note that this result **DOES** require  $X$  and  $Y$  to be **INDEPENDENT**.

## Covariance of continuous random variables:

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y) \text{ as before.}$$

$$\text{Note: } \mathbb{E}(XY) = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} xy f_{X,Y}(x, y) dy dx.$$

## Correlation:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \text{ as before.}$$

## Conditional expectation:

$$\mathbb{E}(X|Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx.$$

$$\text{Similarly, } \mathbb{E}(g(X)|Y = y) = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y) dx.$$

Recall that  $\mathbb{E}(g(X)|Y = y)$  is a function of  $y$  (a number depending on  $y$ ), while  $\mathbb{E}(g(X)|Y)$  is a random variable, with randomness inherited from  $Y$  (not  $X$ ).

## Theorem 4.7: Formulae for conditional expectation and variance.

(Exactly the same as for the discrete case, Theorem 4.4.)

If all expectations below are finite, then for ANY random variables  $X$  and  $Y$ :

$$\text{i) } \mathbb{E}X = \mathbb{E}_Y \left\{ \mathbb{E}(X|Y) \right\}.$$

$$\text{ii) } \mathbb{E}(g(X)) = \mathbb{E}_Y \left\{ \mathbb{E}(g(X)|Y) \right\}.$$

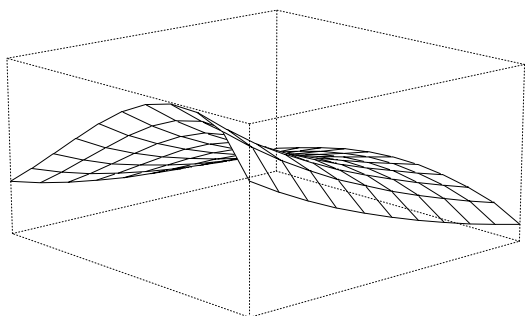
$$\text{iii) } \text{Var}(X) = \mathbb{E}_Y \left( \text{Var}(X|Y) \right) + \text{Var}_Y \left( \mathbb{E}(X|Y) \right).$$

## Proof:

Exactly as for Theorem 4.4, with sums  $\sum$  replaced by integrals  $\int$ . □

## Interlude: What *is* a bivariate density?

To build a mental picture of a bivariate density, we must think in *3 dimensions*.



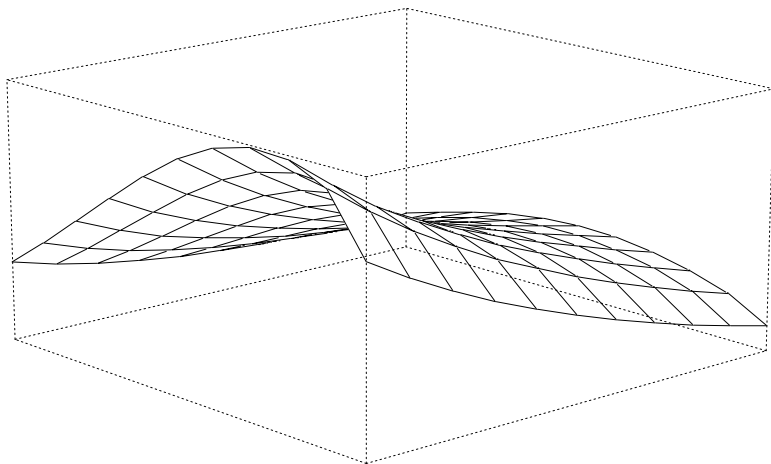
The joint density  $f(x, y)$  is a *surface*.

The height of the surface at point  $(x, y)$  tells you “*how likely*” point  $(x, y)$  is, *compared with other points*. *The higher the surface at point  $(x, y)$ , the more likely it is.*

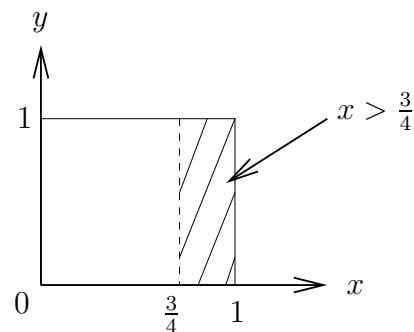
Probabilities are given by *volumes underneath the surface*.

Total probability =  $\iint f(x, y) dy dx = 1$  means that *the total volume underneath the surface is 1*.

To calculate (say)  $\mathbb{P}(X > \frac{3}{4})$ , calculate the *volume underneath the surface corresponding to the region  $x > \frac{3}{4}$* :  $\int_{x=3/4}^1 \int_{y=0}^1 f(x, y) dy dx$ .

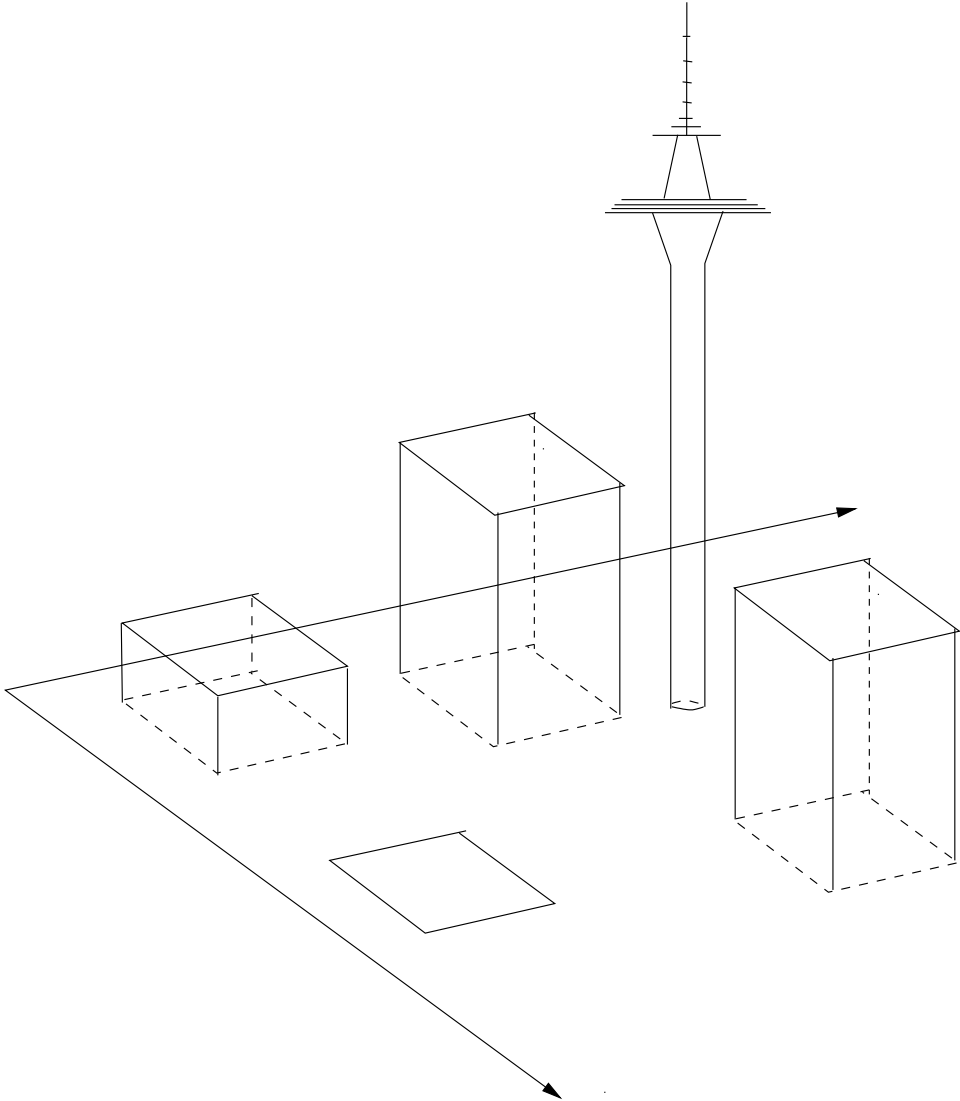


This is *not* the same as the area of the shaded region  $x > \frac{3}{4}$ .



**Example: Lost Dog**

Your dog is lost somewhere in Central Auckland and it knows how to climb stairs and use the lift. Where do you spend most time looking for it?



## 4.9 Change of Variable Technique for Continuous Bivariate Distributions

---

Recall that if  $X$  is a **univariate** random variable, and  $Y = g(X)$  where  $g : \mathbb{R} \rightarrow \mathbb{R}$  is a  $(1 - 1)$  function, then the p.d.f. of  $Y$  is

$$f_Y(y) = f_X(x(y)) \left| \frac{dx}{dy} \right|.$$

Now suppose we have  $\mathbf{X} = (X, Y)$ : a random **vector** in  $\mathbb{R}^2$ .

Suppose  $\mathbf{U} = (U, V) = (g_1(X, Y), g_2(X, Y)) = \mathbf{g}(\mathbf{X})$ .

If  $\mathbf{g}(\mathbf{X}) = \mathbf{U}$  is **smooth and**  $(1 - 1)$  over some region, then the inverse function  $\mathbf{g}^{-1}(\mathbf{U}) = \mathbf{X}$  exists, and we can apply the **change of variable technique in 2 dimensions** to find the joint density of  $\mathbf{U} = (U, V)$ .

*Definition:* **Jacobian**.

Let  $\mathbf{x} = (x, y)$ .

Let  $\mathbf{u} = (u, v) = (u(x, y), v(x, y))$  be a  $(1 - 1)$  transformation of  $\mathbf{x}$  over some region.

We can write  $\mathbf{x} = (x(u, v), y(u, v))$  for the **inverse** transformation.

Define the matrix  $J = \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix}$ .

Then the **Jacobian** of the transformation is

$$\text{Jacobian} = \det J = \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix}$$

### Theorem 4.8: Change of Variable formula in 2 dimensions.

If  $U = (U, V)$  is a smooth, (1-1) function of  $\mathbf{X} = (X, Y)$  over some region, then the joint density of  $\mathbf{U} = (U, V)$  is given by

$f_{\mathbf{U}}(\mathbf{u}) = f_{\mathbf{X}}(\mathbf{x}(\mathbf{u})) |\det J|$  where  $J$  is the Jacobian of the transformation,

or in other words,

$$f_{U,V}(u, v) = f_{X,Y}(x(u, v), y(u, v)) \left| \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} \right|$$

**Proof:** Not required.

**Important Note:** Applying the Change of Variable formula is easy: the difficult part is working out the correct region in the  $(u, v)$  plane.

### Examples: how to work out the region

1. Write down the equations of all lines bounding the region in the  $(x, y)$  plane. Rewrite each equation in terms of  $u$  and  $v$ . Sketch the resulting lines in the  $(u, v)$  plane.
2. A mathematical description of the region is needed: it is not enough just to shade it on a diagram.

*First try to use the mathematical description of the  $(x, y)$  region,*

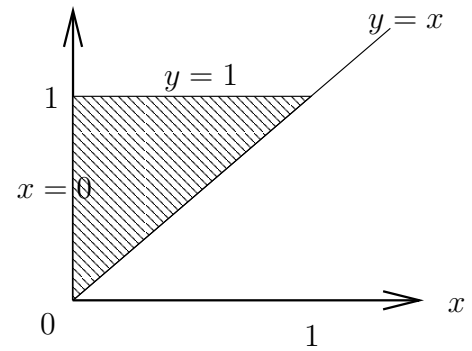
*eg.  $0 < x < y < 1$ .*

*Translate it directly in terms of  $u$  and  $v$  and see if it gives a neat mathematical expression.*

*If not, work from scratch, using your sketch.*

**Example 1:** Suppose the  $(x, y)$  region is  $0 < x < y < 1$ . Let  $u = x$ ,  $v = \log(y)$ .

1) Sketch region in  $(x, y)$  plane.



2) Invert the transformation:  $u = x \Rightarrow x = u$ ,  $v = \log y \Rightarrow y = e^v$ .

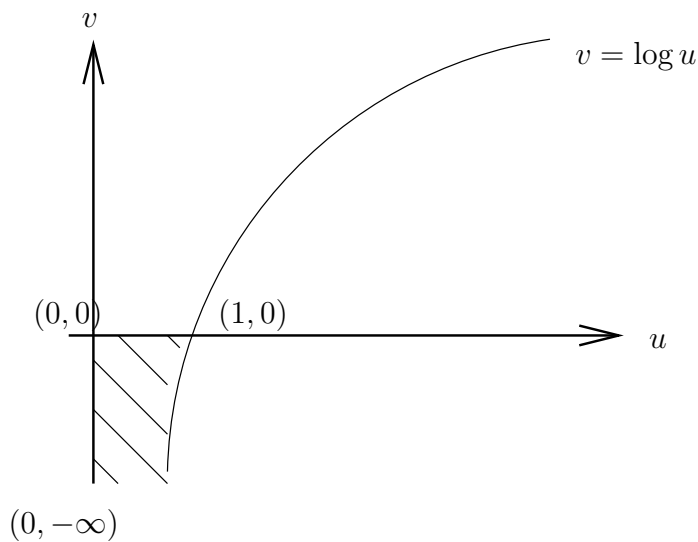
3) Rewrite the equations of all bounding lines:

$$y = 1 \Rightarrow e^v = 1 \Rightarrow v = \log 1 = 0.$$

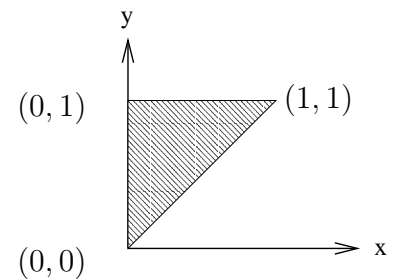
$$x = 0 \Rightarrow u = 0.$$

$$y = x \Rightarrow e^v = u \Rightarrow v = \log u.$$

4) Sketch new region in  $(u, v)$  plane:



Transform points to decide which area to shade:



Use boundary points, or one single inside point is enough.

5) Look for mathematical description:

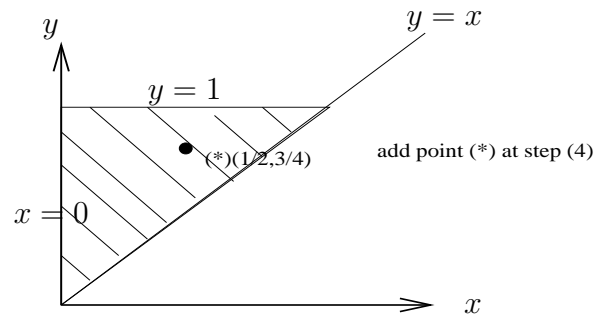
first try  $0 < x < y < 1 \Rightarrow 0 < u < e^v < 1$

→ gives 
$$\begin{array}{l} 0 < u < 1 \\ \log u < v < 0. \end{array}$$

**Example 2:** Suppose the  $(x, y)$  region is  $0 < x < y < 1$  again. Let

$$u = \frac{x+y}{2}, \quad v = \frac{x-y}{2}.$$

1) Sketch region in  $(x, y)$  plane:



2) Invert transformation:  $u = \frac{x+y}{2} \Rightarrow x = u+v$

$$v = \frac{x-y}{2} \Rightarrow y = u-v$$

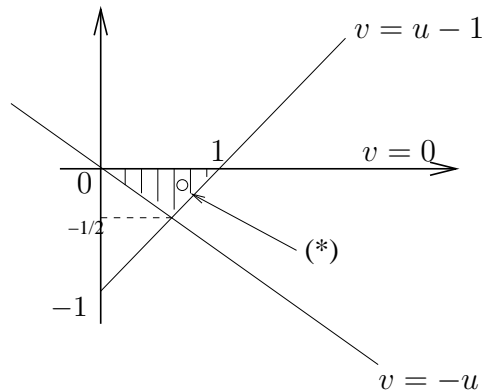
3) Rewrite equations of bounding lines:

$$y = 1 \Rightarrow u - v = 1 \Rightarrow v = u - 1.$$

$$x = 0 \Rightarrow u + v = 0 \Rightarrow v = -u.$$

$$y = x \Rightarrow u + v = u - v \Rightarrow 2v = 0 \Rightarrow v = 0.$$

4) Sketch in  $(u, v)$  plane:



5) Look for mathematical description:

first try  $0 < x < y < 1 \Rightarrow 0 < u + v < u - v < 1$ : too complicated for easy understanding.

Instead, look directly at sketch:

$$\begin{aligned} 0 < u < 1 \\ \max(-u, u-1) < v < 0. \end{aligned}$$

Alternative:

$$\begin{aligned} -1/2 < v < 0 \\ -v < u < v + 1. \end{aligned}$$



## Examples of the Change of Variable technique

**Example 1:** Let  $(X, Y)$  have joint density  $f(x, y) = \begin{cases} 1 & (0 < x < 1, 0 < y < 1), \\ 0 & \text{otherwise.} \end{cases}$

a) Find the joint density of  $\mathbf{U} = (U, V) = (X + Y, Y)$ .

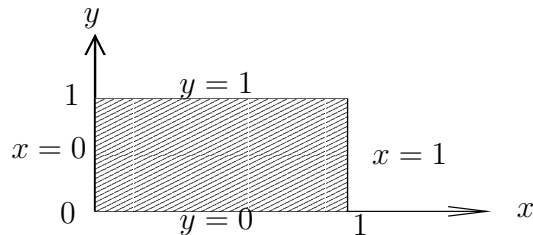
b) Use your answer to (a) to find the marginal p.d.f. of  $U = X + Y$ .

### Solution:

a) Let 
$$\begin{aligned} u &= u(x, y) = x + y \\ v &= v(x, y) = y \quad \text{for } 0 < x < 1, 0 < y < 1. \end{aligned}$$

This is a  $(1 - 1)$  transformation. (Must state this).

Sketch:

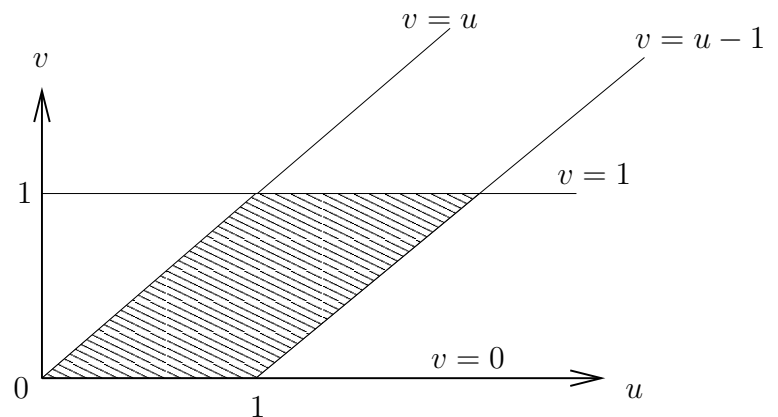


Invert transformation: 
$$\begin{aligned} x &= x(u, v) = u - v \\ y &= y(u, v) = v. \end{aligned}$$

New Sketch:

Line equations:

$$\begin{aligned} y = 1 &\Rightarrow v = 1 \\ y = 0 &\Rightarrow v = 0 \\ x = 1 &\Rightarrow u - v = 1, v = u - 1 \\ x = 0 &\Rightarrow u - v = 0, v = u. \end{aligned}$$



Region:

$\begin{aligned} 0 &< v < 1 \\ v &< u < v + 1 \end{aligned}$
--

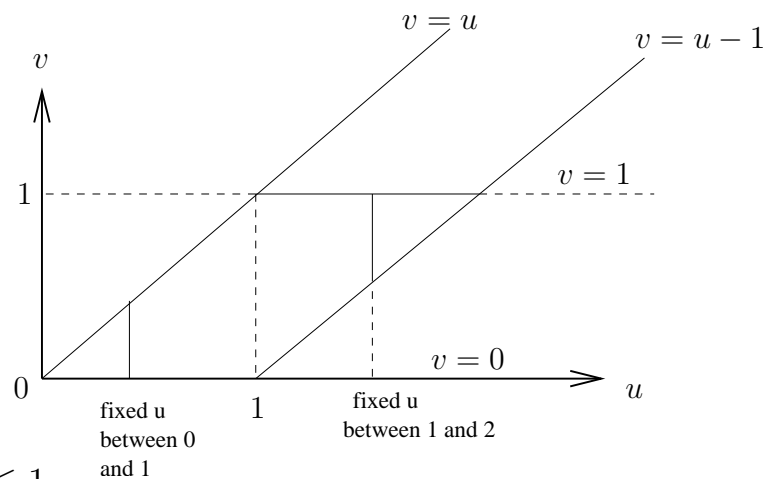
Now change variable:

$$|\det J| = \left| \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} \right| = \left| \det \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \right| = 1.$$

$$\begin{aligned} \text{So } f_{U,V}(u, v) &= f_{X,Y}(x(u, v), y(u, v)) |\det J| \\ &= f_{X,Y}(u - v, v) \times 1 \\ &= 1 \times 1 \end{aligned}$$

$$f_{U,V}(u, v) = 1 \text{ for } \begin{cases} 0 < v < 1 \\ v < u < v + 1 \end{cases}$$

b)



$$\begin{aligned} f_U(u) &= \int_{-\infty}^{\infty} f_{U,V}(u, v) dv \\ &= \begin{cases} \int_{v=0}^u 1 dv & \text{if } 0 < u \leq 1 \\ \int_{v=u-1}^1 1 dv & \text{if } 1 < u < 2 \end{cases} \end{aligned}$$

$$= \begin{cases} [v]_{v=0}^u & \text{for } 0 < u \leq 1 \\ [v]_{v=u-1}^1 & \text{for } 1 < u < 2 \end{cases}$$

$$f_U(u) = \begin{cases} u & \text{for } 0 < u \leq 1, \\ 1 - (u - 1) = 2 - u & \text{for } 1 < u < 2. \end{cases}$$

**Example 2:** Let  $(X, Y)$  have joint density  $f(x, y) = \begin{cases} 4xy & (0 < x < 1, 0 < y < 1), \\ 0 & \text{otherwise.} \end{cases}$

Find the p.d.f. of  $U = \frac{X}{Y}$ .

**Solution:** Three steps.

i) Let  $U = \frac{X}{Y}$  and choose a suitable  $V$ .

ii) Find the joint pdf of  $(U, V)$ .

iii) Integrate out  $V$  to give the marginal  $f_U(u)$ .

**Step (i)**

If we put  $V = Y$ , then we can uniquely recover  $X$  and  $Y$  from  $U = \frac{X}{Y}$  and  $V = Y$ .

So let  $u = u(x, y) = \frac{x}{y}$ ,  $v = v(x, y) = y$  for  $\begin{cases} 0 < x < 1 \\ 0 < y < 1. \end{cases}$

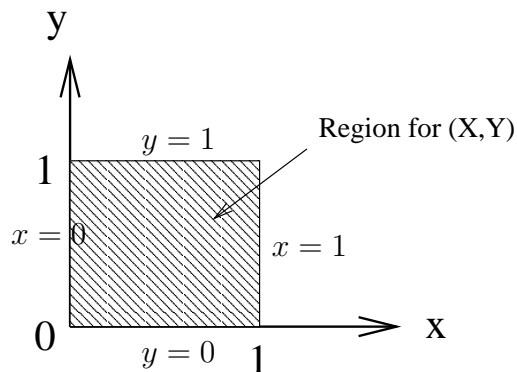
Invert:

$$x = x(u, v) = uv$$

$$y = y(u, v) = v.$$

**Step (ii)**

First find the region for  $(U, V)$ :



**Lines:**

$$x = 0 \Rightarrow uv = 0 \Rightarrow u = 0 \text{ or } v = 0$$

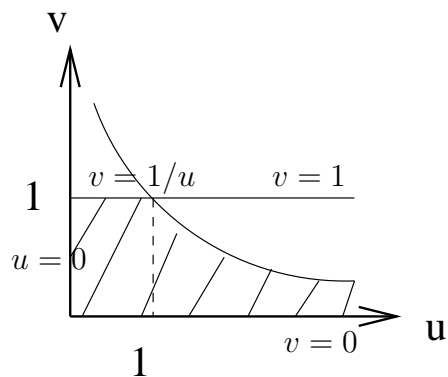
$$x = 1 \Rightarrow uv = 1 \Rightarrow v = \frac{1}{u}$$

$$y = 0 \Rightarrow v = 0$$

$$y = 1 \Rightarrow v = 1$$

**Shaded Region:**

$$\begin{aligned} 0 < v < 1 \\ 0 < u < \frac{1}{v} \end{aligned}$$



**Jacobian:**  $|\det J| = \left| \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} \right| = \left| \det \begin{pmatrix} v & u \\ 0 & 1 \end{pmatrix} \right| = v$

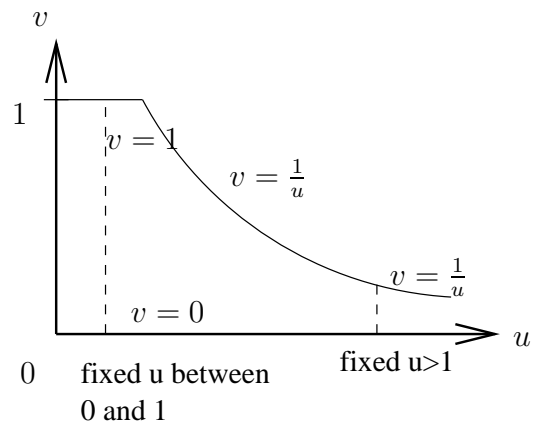
**Change of Variable:** the transformation is (1 - 1) so we can apply the technique.

$$\begin{aligned} f_{U,V}(u, v) &= f_{X,Y}(x(u, v), y(u, v)) |\det J| \\ &= 4(uv)(v) \times v \\ f_{U,V}(u, v) &= 4uv^3 \text{ for } \begin{cases} 0 < v < 1 \\ 0 < u < \frac{1}{v} \end{cases} \quad (\text{joint pdf of } U, V.) \end{aligned}$$

**Step (iii)**

Need marginal pdf of  $U = \frac{X}{Y}$  :

$$\begin{aligned} f_U(u) &= \int_{-\infty}^{\infty} f_{U,V}(u, v) dv \\ &= \begin{cases} \int_{v=0}^1 4uv^3 dv & \text{for } 0 < u \leq 1 \\ \int_{v=0}^{1/u} 4uv^3 dv & \text{for } 1 < u < \infty \end{cases} \\ &= \begin{cases} u \left[ v^4 \right]_{v=0}^1 = u & \text{for } 0 < u \leq 1 \\ u \left[ v^4 \right]_{v=0}^{1/u} = \frac{1}{u^3} & \text{for } 1 < u < \infty \end{cases} \end{aligned}$$



$$f_U(u) = \begin{cases} u & \text{for } 0 < u \leq 1 \\ \frac{1}{u^3} & \text{for } 1 < u < \infty \\ 0 & \text{otherwise.} \end{cases}$$

## 4.10 Example of joint continuous distribution: the Bivariate Normal

---

The bivariate Normal distribution arises as an interaction between two univariate Normal random variables.

*Definition:*  $X$  and  $Y$  have a **bivariate Normal distribution** if their joint density is:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{\sigma_X^2\sigma_Y^2(1-\rho^2)}} \exp \left\{ \frac{-1}{2(1-\rho^2)} \left[ \left( \frac{x-\mu_X}{\sigma_X} \right)^2 - 2\rho \left( \frac{x-\mu_X}{\sigma_X} \right) \left( \frac{y-\mu_Y}{\sigma_Y} \right) + \left( \frac{y-\mu_Y}{\sigma_Y} \right)^2 \right] \right\}$$

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

$$\text{Here, } \left. \begin{array}{l} -\infty < \mu_X, \mu_Y < \infty \\ 0 < \sigma_X, \sigma_Y < \infty \\ -1 < \rho < 1 \end{array} \right\} \text{ five parameters required.}$$

### Properties:

If  $(X, Y)$  has a Bivariate Normal distribution, then:

i) *The marginals are univariate Normal:*

$$\begin{aligned} X &\sim N(\mu_X, \sigma_X^2) \\ Y &\sim N(\mu_Y, \sigma_Y^2) \end{aligned}$$

ii) *Parameter  $\rho$  is the correlation between  $X$  and  $Y$ .*

iii) *Any linear combination  $Z = aX + bY$  is univariate Normal.*  
*[Proof left till Chapter 5.]*

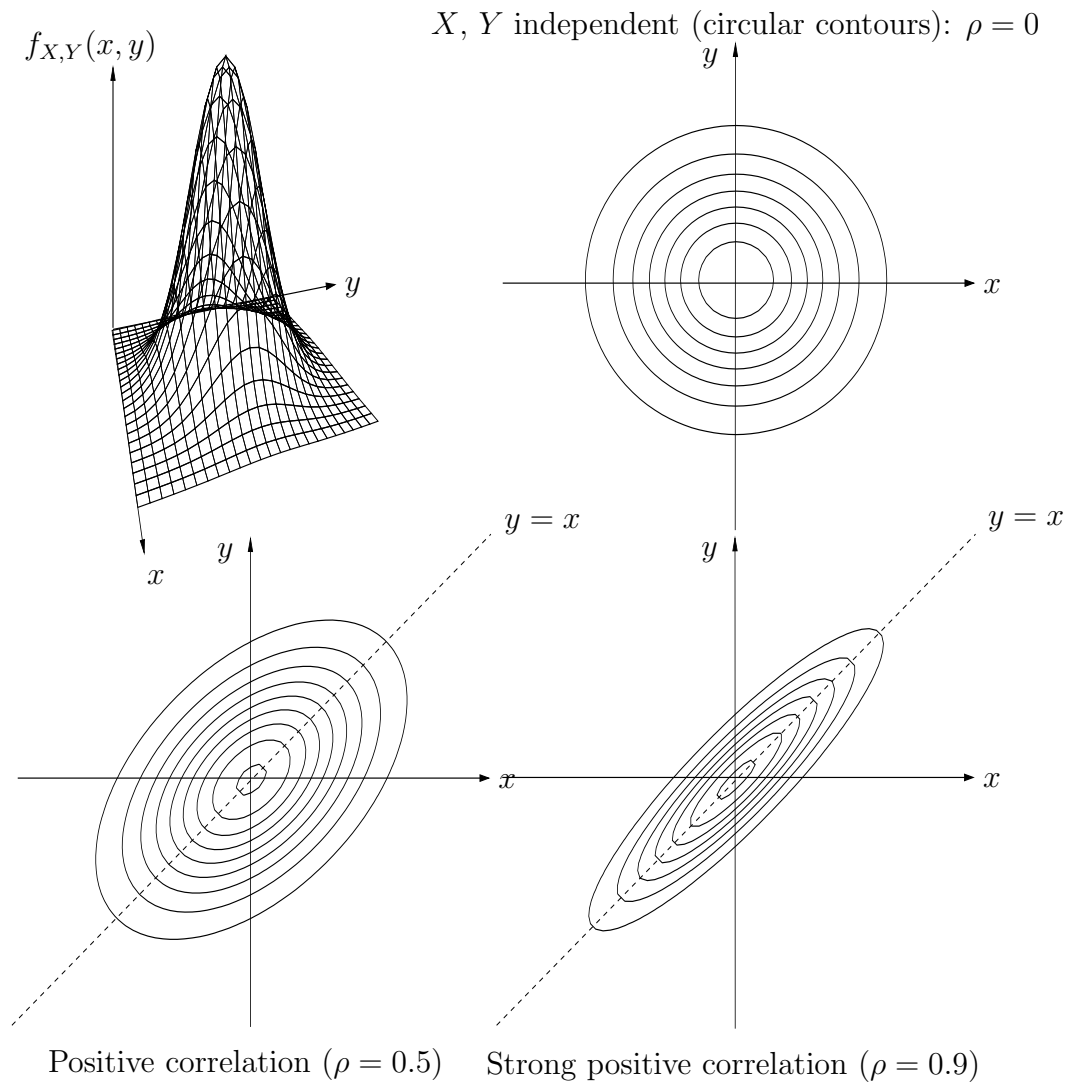
iv)  $X$  and  $Y$  are independent if and only if  $\rho = 0$ .

**Note: this is different from usual:**

usually  $X, Y$  independent  $\Rightarrow \rho_{XY} = 0$  but  $\rho_{XY} = 0 \not\Rightarrow X, Y$  independent.

However, if  $(X, Y) \sim \text{Bivariate Normal}$  then  $X, Y$  independent  $\Leftrightarrow \rho_{XY} = 0$ .

v) The graph of  $f_{X,Y}(x, y)$  is like a ‘mountain’ centred on  $(\mu_X, \mu_Y)$ . If  $X$  and  $Y$  are independent ( $\rho = 0$ ), with equal variance, the mountain has a circular cross-section. As  $|\rho| \rightarrow 1$ , the cross-section becomes elliptical and eventually almost a straight line.



vi) *The conditional distribution of  $X$  given  $Y = y$  is univariate Normal:*

$$X|(Y = y) \sim N \left( \underbrace{\mu_X + \rho \frac{\sigma_X}{\sigma_Y}(y - \mu_Y)}_{\text{mean}}, \underbrace{\sigma_X^2(1 - \rho^2)}_{\text{variance}} \right).$$

**Note:**  $\mathbb{E}(X|Y = y) = \mu_X + \rho \frac{\sigma_X}{\sigma_Y}(y - \mu_Y)$  is a linear function of  $y$ : this is called the regression of  $X$  upon  $Y$ .

**Proof:** (i), (ii), (iv), (vi)

(i) The marginal density of  $X$  is

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x, y) dy \\ &= \int_{-\infty}^{\infty} \left( \frac{\exp \left\{ -\frac{1}{2(1-\rho^2)} \left[ \left( \frac{x-\mu_X}{\sigma_X} \right)^2 - 2\rho \left( \frac{x-\mu_X}{\sigma_X} \right) \left( \frac{y-\mu_Y}{\sigma_Y} \right) + \left( \frac{y-\mu_Y}{\sigma_Y} \right)^2 \right] \right\}}{2\pi\sigma_X\sigma_Y\sqrt{(1-\rho^2)}} \right) dy \\ &= \int_{-\infty}^{\infty} \left( \frac{\exp \left\{ -\frac{1}{2(1-\rho^2)} \left[ (1-\rho^2) \left( \frac{x-\mu_X}{\sigma_X} \right)^2 + \left( \rho \left( \frac{x-\mu_X}{\sigma_X} \right) - \left( \frac{y-\mu_Y}{\sigma_Y} \right) \right)^2 \right] \right\}}{\sqrt{2\pi\sigma_X^2}\sqrt{2\pi\sigma_Y^2(1-\rho^2)}} \right) dy \end{aligned}$$

Put  $z = \rho \left( \frac{x-\mu_X}{\sigma_X} \right) - \left( \frac{y-\mu_Y}{\sigma_Y} \right)$ . The integral becomes:

$$\begin{aligned} f_X(x) &= \left( \frac{\exp \left\{ -\frac{1}{2} \left( \frac{x-\mu_X}{\sigma_X} \right)^2 \right\}}{\sqrt{2\pi\sigma_X^2}} \right) \times \int_{-\infty}^{\infty} \left( \frac{\exp \left\{ -\frac{z^2}{2(1-\rho^2)} \right\}}{\sqrt{2\pi(1-\rho^2)}} \right) dz \\ &= \left( \frac{\exp \left\{ -\frac{1}{2} \left( \frac{x-\mu_X}{\sigma_X} \right)^2 \right\}}{\sqrt{2\pi\sigma_X^2}} \right) \times 1. \end{aligned}$$

(The integral is the integral of the p.d.f. of a Normal( $\mu = 0, \sigma^2 = (1 - \rho^2)$ ) random variable, so it is unity.)

By examining the form of the marginal p.d.f.  $f_X(x)$ , we see that

$$X \sim \text{Normal}(\mu_X, \sigma_X^2).$$

By symmetry, the marginal distribution of  $Y$  is Normal( $\mu_Y, \sigma_Y^2$ ).

(ii) Method of proof: integrate the bivariate Normal p.d.f. to obtain

$$\text{cov}(X, Y) = \mathbb{E}\left((X - \mu_X)(Y - \mu_Y)\right) = \rho \sigma_X \sigma_Y.$$

The result  $\text{corr}(X, Y) = \rho$  follows.

(iv) We know that  $X, Y$  independent  $\Rightarrow \rho = 0$ , as always.

Suppose conversely that  $\rho = 0$ . The bivariate density  $f_{X,Y}(x, y)$  factorizes into an expression  $g(x)h(y)$ , so  $X$  and  $Y$  are independent by Theorem 4.6.

Thus

$$X \text{ and } Y \text{ are independent} \iff \rho = 0.$$

(vi) The conditional density of  $X$  given  $Y = y$  is

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f(x, y)}{f_Y(y)} \\ &= \frac{\frac{1}{2\pi\sqrt{(1-\rho^2)\sigma_X^2\sigma_Y^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}}{\frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}} \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)\sigma_X^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X}\right) - \rho\left(\frac{y-\mu_Y}{\sigma_Y}\right)\right]^2\right\} \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)\sigma_X^2}} \exp\left\{-\frac{1}{2\sigma_X^2(1-\rho^2)} \left[x - \left(\mu_X + \frac{\rho\sigma_X}{\sigma_Y}(y-\mu_Y)\right)\right]^2\right\}, \end{aligned}$$

which is the density of the Normal distribution with mean  $\mu_X + \frac{\rho\sigma_X}{\sigma_Y}(y - \mu_Y)$  and variance  $\sigma_X^2(1 - \rho^2)$ .  $\square$



# Chapter 5: Moment Generating Functions

---

## 5.1 Introduction

---

Recall that the distribution function,  $F_X(x)$ , and the probability function or probability density function,  $f_X(x)$ , both **characterize** the distribution of a random variable  $X$ : that is, specifying either  $F_X(x)$  or  $f_X(x)$  **uniquely defines the whole distribution**.

A third characterization of a distribution is the **moment generating function**,  $M_X(t)$ .

*Definition:* The **moment generating function (m.g.f.)** of a random variable  $X$

is the function  $M_X(t) = \mathbb{E}(e^{Xt})$  provided this exists in some interval containing  $t = 0$ .

### Reference List

(derivations later)

Distribution of $X$	M.G.F.	Special cases	M.G.F.
Normal( $\mu, \sigma^2$ )	$e^{\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)}$	Normal(0, 1)	$e^{\left(\frac{1}{2}t^2\right)}$
Uniform( $a, b$ )	$\frac{e^{bt} - e^{at}}{t(b - a)}$	Uniform(0, 1)	$\frac{e^t - 1}{t}$
Binomial( $n, p$ )	$(pe^t + q)^n$		
Poisson( $\lambda$ )	$e^{\lambda(e^t - 1)}$		
Gamma( $k, \lambda$ )	$\left(1 - \frac{t}{\lambda}\right)^{-k}$	Chisquare( $\nu$ )	$(1 - 2t)^{-\frac{\nu}{2}}$
		Exponential( $\lambda$ )	$\left(1 - \frac{t}{\lambda}\right)^{-1}$
NegBin( $k, p$ )	$\frac{p^k}{(1 - qe^t)^k}$		

---

**Note:** The moment generating function is written  $M_X(t)$  and is a **function of  $t$** :

for example,  $M_X(2) = \mathbb{E}(e^{2X})$ : a fixed number;

$M_X(3) = \mathbb{E}(e^{3X})$ : a fixed number, different from  $M_X(2)$ .

$M_X(t)$  simply describes how  $\mathbb{E}(e^{tX})$  changes with the value of  $t$ . (Why this is useful will become clear later on.)

Note that  $t$  is **not** random;  $X$  is the only random quantity.  $\mathbb{E}(e^{tX})$  is a fixed number giving the mean of  $e^{tX}$  if  $X$  were observed many times.

### Calculating the moment generating function

1. When  $X$  is discrete,

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

2. When  $X$  is continuous,

$$M_X(t) = \mathbb{E}(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx \quad (dx \text{ not } dt)$$

**Theorem 5.1:** Let  $X$  be any random variable with m.g.f.  $M_X(t)$ . Let  $Y = aX + b$  where  $a$  and  $b$  are constants. Then the m.g.f. of  $Y$  is

$$M_Y(t) = e^{bt} M_X(at).$$

**Proof:**

$$M_Y(t) = \mathbb{E}(e^{Yt}) = \mathbb{E}(e^{(aX+b)t}) = \mathbb{E}(\underbrace{e^{bt}}_{\text{constant}} e^{(at)X}) = e^{bt} \mathbb{E}(e^{(at)X}) = e^{bt} M_X(at). \quad \square$$

### Derivations of m.g.f.s for selected distributions

#### 1. Binomial distribution

Let  $X \sim \text{Binomial}(n, p)$ , so  $f_X(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x q^{n-x}$ .

$$\begin{aligned} M_X(t) &= \sum_{x=0}^n e^{xt} \binom{n}{x} p^x q^{n-x} \\ &= \sum_{x=0}^n \binom{n}{x} (pe^t)^x q^{n-x} \\ &= (pe^t + q)^n \quad \text{by the Binomial Theorem: true for all } t. \end{aligned}$$

Thus  $M_X(t) = (pe^t + q)^n$  for all  $t \in \mathbb{R}$ .

## 2. Poisson distribution

Let  $X \sim \text{Poisson}(\lambda)$ , so  $f_X(x) = \mathbb{P}(X = x) = \frac{\lambda^x}{x!}e^{-\lambda}$ .

$$\begin{aligned}M_X(t) &= \sum_{x=0}^{\infty} e^{xt} f_X(x) = \sum_{x=0}^{\infty} e^{xt} \frac{\lambda^x}{x!} e^{-\lambda} \\&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \quad (\text{Note: sum = series expansion of } e^{(\lambda e^t)}) \\&= e^{-\lambda} e^{(\lambda e^t)}\end{aligned}$$

$$M_X(t) = e^{\lambda(e^t - 1)} \text{ for all } t \in \mathbb{R}.$$

## 3. Normal(0, 1) distribution

Let  $X \sim \text{Normal}(0, 1)$ , so  $f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ .

$$\begin{aligned}M_X(t) = \mathbb{E}(e^{Xt}) &= \int_{-\infty}^{\infty} e^{xt} f_X(x) dx \quad (\text{Note: integrate } \int dx, \text{ NOT } \int dt) \\&= \int_{-\infty}^{\infty} e^{xt} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^2 - 2tx + t^2 - t^2)} dx \\&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2} e^{\frac{1}{2}t^2} dx \\&= e^{\frac{1}{2}t^2} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2} dx}_{\text{integral of } N(t, 1) \text{ pdf} = 1}\end{aligned}$$

$$M_X(t) = e^{\frac{1}{2}t^2} \text{ for } t \in \mathbb{R}.$$

#### 4. Normal( $\mu, \sigma^2$ ) distribution

Use Theorem 5.1: if  $Y \sim N(\mu, \sigma^2)$ , then  $\frac{Y-\mu}{\sigma} \sim N(0, 1)$ , so we can write  $Y = \sigma X + \mu$ , where  $X \sim N(0, 1)$ .

$$\begin{aligned} \text{Thus by Thm 5.1, } M_Y(t) &= e^{\mu t} M_X(\sigma t) \\ &= e^{\mu t} e^{\frac{1}{2}(\sigma t)^2} \\ M_Y(t) &= e^{\mu t + \frac{1}{2}\sigma^2 t^2} \text{ for all } t \in \mathbb{R}. \end{aligned}$$

#### 5. Gamma( $k, \lambda$ ) distribution

Let  $X \sim \text{Gamma}(k, \lambda)$ , so  $f_X(x) = \frac{1}{\Gamma(k)} \lambda^k x^{k-1} e^{-\lambda x}$  for  $x > 0$ .

$$\begin{aligned} M_X(t) = \mathbb{E}(e^{Xt}) &= \int_0^\infty e^{xt} f_X(x) dx \\ &= \int_0^\infty e^{xt} \frac{1}{\Gamma(k)} \lambda^k x^{k-1} e^{-\lambda x} dx \\ &= \int_0^\infty \frac{1}{\Gamma(k)} \lambda^k x^{k-1} e^{-(\lambda-t)x} dx. \end{aligned}$$

**Important:** we need  $t < \lambda$  for this integral to be finite. Assume  $t < \lambda$ , so that  $(\lambda - t)$  is a positive number.

$$\begin{aligned} \text{Then } M_X(t) &= \int_0^\infty \frac{1}{\Gamma(k)} (\lambda - t)^k x^{k-1} e^{-(\lambda-t)x} \frac{\lambda^k}{(\lambda - t)^k} dx \text{ for } t < \lambda \\ &= \frac{\lambda^k}{(\lambda - t)^k} \int_0^\infty \frac{1}{\Gamma(k)} (\lambda - t)^k x^{k-1} e^{-(\lambda-t)x} dx \text{ for } t < \lambda \\ &= \left( \frac{\lambda}{\lambda - t} \right)^k \text{ for } t < \lambda \\ M_X(t) &= \frac{1}{(1 - t/\lambda)^k} \text{ for } t < \lambda. \end{aligned}$$

## 6. Chi-square distribution $\chi_\nu^2$

Recall that  $\chi_\nu^2 = \text{Gamma}(\frac{\nu}{2}, \frac{1}{2})$ . Put  $k = \frac{\nu}{2}, \lambda = \frac{1}{2}$  above to get

$$M_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-k} = (1 - 2t)^{-\nu/2} \text{ when } X \sim \chi_\nu^2. \text{ Valid for } t < \frac{1}{2}.$$

## 7. Exponential distribution

Recall that  $\text{Exponential}(\lambda) = \text{Gamma}(k = 1, \lambda)$ . Put  $k = 1$  above to get

$$M_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-1} \text{ when } X \sim \text{Exponential}(\lambda), \text{ for } t < \lambda.$$

## Uniqueness of the moment generating function

**Theorem 5.2:** Let  $X$  be a random variable. If the m.g.f. of  $X$ ,  $M_X(t)$ , exists for all  $t$  with  $|t| < t_0$  for some  $t_0 > 0$ , then the whole distribution of  $X$  is uniquely determined by  $M_X(t)$ .

**Proof:** beyond the scope of this course.

The Theorem tells us that, if the m.g.f.  $M_X(t)$  exists for  $t$  in some interval containing 0, then

*the m.g.f. uniquely determines the distribution of  $X$ .*

Thus, if we can recognize the m.g.f. of an unknown random variable  $X$  as one of the functions on the reference list on page 185, then we have established what the distribution of  $X$  is.

## Why is the moment generating function useful?

The moment generating function is a powerful tool for solving problems that are difficult to solve using distribution functions and p.d.f.s or probability functions. Examples are: (i) calculating moments; (ii) finding the distribution of a sum of independent random variables; (iii) finding the distribution of a compound random variable; (iv) finding the distribution of a function of  $X$ ; (v) finding a limiting distribution. We will look at these in turn.

## 5.2 Moments

*Definition:* Let  $X$  be a random variable and let  $r$  be a positive integer.

The  $r$ 'th moment of  $X$  (about the origin) is  $\mathbb{E}(X^r)$ .

The  $r$ 'th central moment of  $X$  ( $r$ 'th moment of  $X$  about the mean of  $X$ ) is  $\mathbb{E}\{(X - \mu_X)^r\}$ .

*Examples:*  $\mathbb{E}(X) = \mu_X$  is the first moment of  $X$ .

$\text{Var}(X) = \mathbb{E}((X - \mu_X)^2)$  is the second central moment of  $X$ .

### Using the power series expansion of the m.g.f. to calculate moments

The moment generating function gets its name because it gives us a quick way of calculating the moments of  $X$ , using the power series expansion of  $e^{tX}$ .

$$\begin{aligned} \text{Consider } M_X(t) &= \mathbb{E}(e^{tX}) \\ &= \mathbb{E}\left\{1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \dots\right\} \\ &\quad \text{using power series expansion of } e^{tX} \\ &= 1 + t \underbrace{\mathbb{E}(X)}_{\text{1st moment}} + \frac{t^2}{2!} \underbrace{\mathbb{E}(X^2)}_{\text{2nd moment}} + \frac{t^3}{3!} \underbrace{\mathbb{E}(X^3)}_{\text{3rd moment}} + \dots \end{aligned}$$

We can recover the moments by *differentiating the power series and evaluating at  $t = 0$* .

1st moment  $\frac{d}{dt}(M_X(t)) = \mathbb{E}(X) + t\mathbb{E}(X^2) + \frac{t^2}{2!}\mathbb{E}(X^3) + \dots$

So  $\frac{d}{dt}(M_X(t))\Big|_{t=0} = M'_X(0) = \mathbb{E}(X)$  : *the 1st moment.*

2nd moment  $\frac{d^2}{dt^2}(M_X(t)) = \mathbb{E}(X^2) + t\mathbb{E}(X^3) + \frac{t^2}{2!}\mathbb{E}(X^4) + \dots$

So  $\frac{d^2}{dt^2}(M_X(t))\Big|_{t=0} = M''_X(0) = \mathbb{E}(X^2)$  : *the 2nd moment.*

## General expression

$$\begin{aligned}\mathbb{E}(X) &= M'_X(0) = \left. \frac{d}{dt} M_X(t) \right|_{t=0} \\ \mathbb{E}(X^2) &= M''_X(0) = \left. \frac{d^2}{dt^2} M_X(t) \right|_{t=0} \\ \mathbb{E}(X^r) &= M_X^{(r)}(0) = \left. \frac{d^r}{dt^r} M_X(t) \right|_{t=0}\end{aligned}$$

This can be a ***much*** quicker way of calculating the mean and variance than the traditional integrations or summations.

Compare the following examples with the effort required in chapters 2 and 3.

### 1. Binomial distribution mean and variance

$X \sim \text{Binomial}(n, p)$ , so MGF is  $M_X(t) = (pe^t + q)^n$ .

$$M'_X(t) = n(pe^t + q)^{n-1}pe^t \quad (\text{Note: } \frac{d}{dt}, \text{ not } \frac{d}{dx})$$

$$M''_X(t) = n(n-1)(pe^t + q)^{n-2}(pe^t)^2 + n(pe^t + q)^{n-1}pe^t \quad (\text{don't bother to simplify})$$

So

$$\begin{aligned}\mathbb{E}(X) &= M'_X(0) = n(pe^0 + q)^{n-1}pe^0 = n(p + q)^{n-1}p = np \quad (\text{because } p + q = 1) \\ \mathbb{E}(X^2) &= M''_X(0) = n(n-1)(p + q)^{n-2}p^2 + n(p + q)^{n-1}p \\ &= n(n-1)p^2 + np\end{aligned}$$

$$\text{So } \text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = n(n-1)p^2 + np - n^2p^2 = np(1-p).$$

### 2. Poisson distribution mean and variance

$X \sim \text{Poisson}(\lambda)$ , so MGF is  $M_X(t) = e^{\lambda(e^t-1)}$ .

$$M'_X(t) = \lambda e^t e^{\lambda(e^t-1)} = \lambda e^{t+\lambda e^t-\lambda}$$

$$M''_X(t) = \lambda(1 + \lambda e^t) e^{t+\lambda e^t-\lambda}$$

$$\begin{aligned}\text{So } \mathbb{E}(X) &= M'_X(0) = \lambda e^{0+\lambda e^0-\lambda} = \lambda \\ \mathbb{E}(X^2) &= M''_X(0) = \lambda(1 + \lambda e^0)e^{0+\lambda e^0-\lambda} = \lambda(1 + \lambda)\end{aligned}$$

$$\text{So } \text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \lambda(1 + \lambda) - \lambda^2 = \lambda$$

### 3. Normal( $\mu, \sigma^2$ ) distribution mean and variance

$X \sim N(\mu, \sigma^2)$ , so MGF is  $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

$$M'_X(t) = (\mu + \sigma^2 t)e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

$$M''_X(t) = (\mu + \sigma^2 t)^2 e^{\mu t + \frac{1}{2}\sigma^2 t^2} + \sigma^2 e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

$$\begin{aligned}\text{So } \mathbb{E}(X) &= M'_X(0) = (\mu + 0)e^0 = \mu. \\ \mathbb{E}(X^2) &= M''_X(0) = (\mu + 0)^2 e^0 + \sigma^2 e^0 = \mu^2 + \sigma^2;\end{aligned}$$

$$\text{So } \text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \mu^2 + \sigma^2 - \mu^2 = \sigma^2.$$

### 4. Gamma( $k, \lambda$ ) distribution mean and variance

$X \sim \text{Gamma}(k, \lambda)$ , so MGF is  $M_X(t) = (1 - \frac{t}{\lambda})^{-k}$  ( $t < \lambda$ )

The MGF is defined at  $t = 0$ , so we can proceed.

$$M'_X(t) = -k \left(1 - \frac{t}{\lambda}\right)^{-k-1} \left(\frac{-1}{\lambda}\right) = \frac{k}{\lambda} \left(1 - \frac{t}{\lambda}\right)^{-k-1}$$

$$M''_X(t) = \frac{k}{\lambda} (-k-1) \left(1 - \frac{t}{\lambda}\right)^{-k-2} \left(\frac{-1}{\lambda}\right) = \frac{k}{\lambda^2} (k+1) \left(1 - \frac{t}{\lambda}\right)^{-k-2}$$

$$\text{So } \mathbb{E}(X) = M'_X(0) = \frac{k}{\lambda}.$$

$$\mathbb{E}(X^2) = M''_X(0) = \frac{k}{\lambda^2} (k+1);$$

$$\text{So } \text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \frac{k(k+1)}{\lambda^2} - \frac{k^2}{\lambda^2} = \frac{k}{\lambda^2}.$$



## Skewness and Kurtosis

The mean,  $\mu = \mathbb{E}(X)$ , of a distribution measures its *location (centre)*. and the variance,  $\sigma^2 = \text{Var}(X) = \mathbb{E}\{(X - \mu)^2\}$ , measures its *spread*.

Two other commonly used measures of distributional shape are *skewness and kurtosis*.

$$\gamma_1 = \mathbb{E}\left\{\left(\frac{X-\mu}{\sigma}\right)^3\right\}.$$

*Definition:* For any r.v.  $X$ , the skewness of  $X$  is

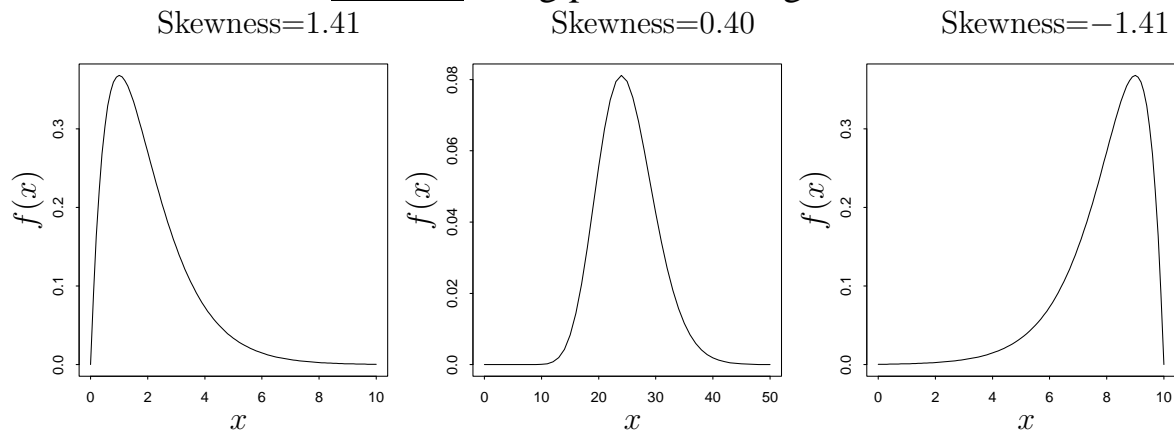
### How does skewness measure shape?

If the distribution of  $X$  is symmetric about the mean  $\mu$ , then skewness=0.

Proof: If  $X$  is symmetric, then  $f_X(\mu - y) = f_X(\mu + y)$  for all  $y$ .

$$\begin{aligned} \text{Then } \gamma_1 &= \frac{1}{\sigma^3} \mathbb{E}\{(X - \mu)^3\} = \frac{1}{\sigma^3} \int_{-\infty}^{\infty} (x - \mu)^3 f_X(x) dx \\ &= \frac{1}{\sigma^3} \int_{-\infty}^{\infty} y^3 f_X(\mu + y) dy \quad (\text{putting } y = x - \mu.) \\ \text{Split integral: } \gamma_1 &= \frac{1}{\sigma^3} \left\{ \int_{-\infty}^0 y^3 f_X(\mu + y) dy + \int_0^{\infty} y^3 f_X(\mu + y) dy \right\} \\ &= \frac{1}{\sigma^3} \left\{ \int_0^{\infty} -v^3 f_X(\mu - v) dv + \int_0^{\infty} y^3 f_X(\mu + y) dy \right\} \quad (v = -y) \\ &= \frac{1}{\sigma^3} \left\{ - \int_0^{\infty} v^3 \underbrace{f_X(\mu + v)}_{\text{by symmetry}} dv + \int_0^{\infty} y^3 f_X(\mu + y) dy \right\} = 0. \end{aligned}$$

If the distribution of  $X$  is not symmetric, then skewness  $\neq 0$  and we say that the distribution is skewed: *long positive or negative tail*.



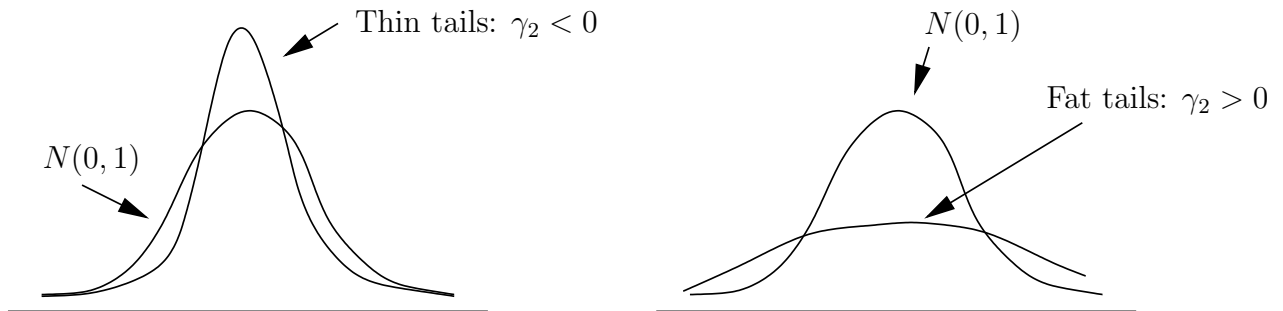
*Definition:* The **kurtosis** of  $X$  is

$$\gamma_2 = \mathbb{E} \left\{ \left( \frac{X - \mu}{\sigma} \right)^4 \right\} - 3.$$

Kurtosis measures the heavy-tailedness of  $X$ , relative to the Normal distribution. (The  $-3$  in the formula makes  $\gamma_2 = 0$  for any  $\text{Normal}(\mu, \sigma^2)$  distribution.)

When the kurtosis is negative ( $\gamma_2 < 0$ ), the tails are ‘thin’ relative to the Normal distribution.

When the kurtosis is positive ( $\gamma_2 > 0$ ), the tails are ‘fat’ relative to the Normal distribution.



**Notes:**

1. *Distributions with the first few moments equal are similar in shape. (Same mean, same variance, same skewness, etc.  $\Rightarrow$  similar shape).*
2. Not all distributions possess finite moments: for example, the Cauchy distribution has  $\mathbb{E}(X) = \infty$  and  $\mathbb{E}(X^2) = \infty$ , although any observation from the Cauchy distribution is of course finite.
3. Central moments,  $\mathbb{E} \{(X - \mu)^r\}$ , can always be expressed in terms of moments about the origin,  $\mathbb{E}(X), \mathbb{E}(X^2), \dots, \mathbb{E}(X^r)$ . Simply expand  $(X - \mu)^r$  and take expectations.

Similarly,  $\mathbb{E}(X^r)$  can be expressed in terms of  $\mathbb{E}(X - \mu), \mathbb{E} \{(X - \mu)^2\}, \dots, \mathbb{E} \{(X - \mu)^r\}$ , by writing  $\mathbb{E}(X^r) = \mathbb{E} \{(X - \mu + \mu)^r\}$  and expanding.

An example is the well-known variance equivalence:

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

### 5.3 Moment generating functions for sums of independent r.v.s

---

The moment generating function turns a sum into a product:

$$\mathbb{E} \left( e^{(X_1+X_2)t} \right) = \mathbb{E} \left( e^{X_1t} e^{X_2t} \right) .$$

This makes it especially useful for finding the distribution of  $(a_1X_1 + \dots + a_nX_n)$ .

**Theorem 5.3:** Suppose that  $X_1, \dots, X_n$  are *independent* random variables, and let  $Y = a_1X_1 + \dots + a_nX_n$  for constants  $a_1, \dots, a_n$ . Then

$$M_Y(t) = \prod_{i=1}^n M_{X_i}(a_it)$$

**Proof:**

$$\begin{aligned} M_Y(t) &= \mathbb{E}(e^{(a_1X_1 + \dots + a_nX_n)t}) \\ &= \mathbb{E}(e^{X_1(a_1t)} e^{X_2(a_2t)} \dots e^{X_n(a_nt)}) \\ &= \mathbb{E}(e^{X_1(a_1t)}) \mathbb{E}(e^{X_2(a_2t)}) \dots \mathbb{E}(e^{X_n(a_nt)}) \quad (\text{because } X_1, \dots, X_n \text{ are independent}) \\ &= \prod_{i=1}^n M_{X_i}(a_it). \quad \text{as required.} \end{aligned}$$

### Sums and means of independent, identically distributed random variables

---

Let  $X_1, \dots, X_n$  be independent and *identically distributed*, with common moment generating function  $M_X(t)$ . Theorem 5.3 gives the following results about the m.g.f.s of the sum,  $S_n$ , and the mean,  $\bar{X}_n$ :

**Sum:** If  $S = X_1 + \dots + X_n$ , then  $M_S(t) = \{M_X(t)\}^n$ .

**Mean:** If  $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$ , then  $M_{\bar{X}}(t) = \{M_X(\frac{t}{n})\}^n$ .

**Examples:** The following examples of Theorem 5.3 are all important results.

1. Sum of independent Poisson random variables is Poisson.
2. Sum of independent Normal random variables is Normal.
3. Sum of independent Chi-square random variables is Chi-square.

**Example 1:** Let  $X_1, \dots, X_n$  be independent with  $X_i \sim \text{Poisson}(\lambda_i)$  ( $i = 1, \dots, n$ ). Then  $X_1 + \dots + X_n \sim \text{Poisson}(\lambda_1 + \dots + \lambda_n)$ .

**Proof:**

$$\begin{aligned} \text{Let } Y = \sum_{i=1}^n X_i \text{ then } M_Y(t) &= \prod_{i=1}^n M_{X_i}(t) \\ &= \prod_{i=1}^n e^{\lambda_i(e^t-1)} \\ &= e^{\sum_{i=1}^n \lambda_i(e^t-1)} = e^{(\sum_{i=1}^n \lambda_i)(e^t-1)}, \end{aligned}$$

which is the m.g.f. of the Poisson  $\left(\sum_{i=1}^n \lambda_i\right)$  distribution. So  $Y \sim \text{Poisson}\left(\sum_{i=1}^n \lambda_i\right)$ .

**Example 2:** Let  $X_1, \dots, X_n$  be independent with  $X_i \sim \text{Normal}(\mu_i, \sigma_i^2)$  ( $i = 1, \dots, n$ ). Then  $a_1X_1 + \dots + a_nX_n \sim N(a_1\mu_1 + \dots + a_n\mu_n, a_1^2\sigma_1^2 + \dots + a_n^2\sigma_n^2)$ .

**Proof:**

$$\begin{aligned} \text{Let } Y = \sum_{i=1}^n a_i X_i \text{ then } M_Y(t) &= \prod_{i=1}^n M_{X_i}(a_i t) = \prod_{i=1}^n e^{\mu_i(a_i t) + \frac{1}{2}\sigma_i^2(a_i t)^2} \\ &= e^{(\sum_{i=1}^n \mu_i a_i)t + \frac{1}{2}(\sum_{i=1}^n \sigma_i^2 a_i^2)t^2}, \end{aligned}$$

which is the m.g.f. of the Normal  $\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right)$  distribution.

**Example 3:** Let  $X_1, \dots, X_n$  be independent with  $X_i \sim \text{Chisquare}(\nu_i) = \chi_{\nu_i}^2$  ( $i = 1, \dots, n$ ). Then  $X_1 + \dots + X_n \sim \chi_{\sum_{i=1}^n \nu_i}^2 = \text{Chisquare}(\sum_{i=1}^n \nu_i)$ .

**Proof:**

$$\begin{aligned} \text{Let } Y = \sum_{i=1}^n X_i \text{ then } M_Y(t) &= \prod_{i=1}^n M_{X_i}(t) \\ &= \prod_{i=1}^n (1 - 2t)^{-\nu_i/2} \\ &= (1 - 2t)^{-\frac{1}{2}\sum_{i=1}^n \nu_i}, \end{aligned}$$

which is the m.g.f. of the Chisquare  $\left(\sum_{i=1}^n \nu_i\right)$  distribution.

## 5.4 Compound distributions

A random variable has a compound distribution if it is defined in terms of two or more other distributions:

*e.g.*  $X \sim \text{Poisson}(\lambda)$ ,  $(Y|X) \sim \text{Binomial}(X, p)$ ;  
Then  $Y$  has a compound distribution.

To find the m.g.f. of a compound random variable, use the formula for conditional expectation:

$$M_Y(t) = \mathbb{E}(e^{Yt}) = \mathbb{E}_X\{\mathbb{E}(e^{Yt}|X)\}$$

**Useful Tip:** For questions of this sort, we often need to find  $\mathbb{E}(a^X)$  for constant  $a$ .  
Use

$$\mathbb{E}(a^X) = \mathbb{E}(e^{\log(a^X)}) = \mathbb{E}(e^{X \log a}) = M_X(\log a).$$

**Example 1:** (insect eggs on a leaf again: see example in Chapter 4).

Let  $X \sim \text{Poisson}(\lambda)$ , so  $M_X(t) = \mathbb{E}(e^{Xt}) = e^{\lambda(e^t-1)}$ .

Let  $(Y|X) \sim \text{Binomial}(X, p)$ , so  $\mathbb{E}(e^{Yt}|X) = (pe^t + q)^X$

(this is Binomial m.g.f. replacing “ $n$ ” by “ $X$ ” in usual formula, and where  $q = 1 - p$ ).

$$\begin{aligned} \text{So } M_Y(t) = \mathbb{E}(e^{Yt}) &= \mathbb{E}_X\{\mathbb{E}(e^{Yt}|X)\} \\ &= \mathbb{E}_X\{(pe^t + q)^X\} \\ &= \mathbb{E}_X\{e^{X \log(pe^t + q)}\} \quad (\text{using tip above}) \\ &= M_X(\log(pe^t + q)). \end{aligned}$$

But  $X \sim \text{Poisson}(\lambda)$ , so  $M_X(t) = e^{\lambda(e^t-1)}$ .

$$\begin{aligned}
 \text{So } M_Y(t) &= M_X(\log(pe^t + q)) \\
 &= e^{\lambda(e^{\log(pe^t+q)}-1)} \\
 &= e^{\lambda(pe^t+q-1)} \\
 &= e^{\lambda(pe^t+1-p-1)} \quad (\text{because } q = 1 - p) \\
 M_Y(t) &= e^{\lambda p(e^t-1)}
 \end{aligned}$$

This is the m.g.f. of the  $\text{Poisson}(\lambda p)$  distribution, so  $Y \sim \underline{\text{Poisson}(\lambda p)}$  as also derived in Chapter 4.

**Example 2:** Sum of a random number of random variables.

Suppose that  $N$  has m.g.f.  $M_N(t) = \mathbb{E}(e^{Nt})$ , and let  $X_1, X_2, \dots$  be independent of each other and of  $N$ , with common m.g.f.  $M_X(t)$ .

Let  $Y = X_1 + \dots + X_N$  (sum of a random number of random variables).

$$\begin{aligned}
 \text{Then } M_Y(t) &= \mathbb{E}(e^{Yt}) \\
 &= \mathbb{E}_N\{\mathbb{E}(e^{Yt}|N)\} \\
 &= \mathbb{E}_N\{\mathbb{E}(e^{(X_1+\dots+X_N)t}|N)\} \\
 &= \mathbb{E}_N\{\mathbb{E}(e^{X_1t}e^{X_2t} \dots e^{X_Nt}|N)\} \\
 &= \mathbb{E}_N\{\mathbb{E}(e^{X_1t})\mathbb{E}(e^{X_2t}) \dots \mathbb{E}(e^{X_Nt})\} \quad (\text{because } X_1, \dots, X_N \text{ are} \\
 &\hspace{15em} \text{independent of each other and of } N) \\
 &= \mathbb{E}_N\{(M_X(t))^N\}
 \end{aligned}$$

$$\begin{aligned}
 \text{So } M_Y(t) &= \mathbb{E}_N\{(M_X(t))^N\} = \mathbb{E}_N(e^{N \log M_X(t)}) \\
 &\Rightarrow M_Y(t) = M_N(\log M_X(t)).
 \end{aligned}$$

**Example:** if  $X_i \sim \text{Poisson}(\lambda)$  for all  $i$ , and if  $N \sim \text{Poisson}(\mu)$ , then  $Y = X_1 + \dots + X_N$  has the compound Poisson distribution:

$$\underline{M_Y(t) = M_N(\log M_X(t)) = e^{\mu(e^{\lambda(e^t-1)}-1)}}$$

## 5.5 Using the m.g.f. to find the distribution of $g(X)$

Let  $X$  be a random variable, and let  $Y = g(X)$ . Usually (for monotone transformations) we find the distribution of  $Y$  by using the change of variable technique to convert the p.d.f.  $f_X(x)$  into the p.d.f.  $f_Y(y)$ . However, we can also use the moment generating function.

**Example:** Let  $X \sim \text{Normal}(0, 1)$ , and let  $Y = X^2$ . (**Note:** this transformation is *not* monotone over the range of  $X$ .) In Section 3.3 we worked with the distribution function to show that  $Y \sim \text{Chisquare}(1) = \chi_1^2$ . Here we use the m.g.f. instead.

*The m.g.f. of  $Y$  is:*

$$\begin{aligned} M_Y(t) &= \mathbb{E}(e^{Yt}) \\ &= \mathbb{E}(e^{X^2t}) \\ &= \int_{-\infty}^{\infty} e^{x^2t} f_X(x) dx \\ &= \int_{-\infty}^{\infty} e^{x^2t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2(1-2t)} dx && \text{(need } 1 - 2t > 0, \text{ i.e. } t < \frac{1}{2}, \text{ for integral} \\ &&& \text{to be finite)} \\ &= \sqrt{(1-2t)^{-1}} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-2t)^{-1}}} e^{-\frac{x^2}{2(1-2t)^{-1}}} dx}_{\text{p.d.f. of } N(0, (1-2t)^{-1}) \text{ integrates to 1}} \\ &= (1-2t)^{-1/2}. \end{aligned}$$

So  $M_Y(t) = (1 - 2t)^{-1/2}$  for  $t < \frac{1}{2}$ , and this is the m.g.f. of a  $\chi_1^2$  random variable.

So  $Y = X^2 \sim \text{Chisquare}(1)$  as expected.

## 5.6 Limiting distributions

*Definition:* Let  $X_1, \dots, X_n$  be a sequence of random variables such that the r.v.  $X_i$  has distribution function  $F_i(x)$  for each  $i$ . Then the sequence  $X_1, \dots, X_n$  **converges in distribution** to the random variable  $X$ , with distribution function  $F(x)$ , if  $\lim_{n \rightarrow \infty} F_n(x) = F(x)$  for all  $x$  at which  $F(x)$  is continuous.

We write  $X_n \xrightarrow{D} X$  ( $X_n$  converges in distribution to  $X$ )  
or (same thing)  $X_n \xrightarrow{W} X$  ( $X_n$  converges weakly to  $X$ ).

We can therefore use the distribution of  $X$  to gain approximate probabilities for  $X_n$ , if  $n$  is large enough.

$$\mathbb{P}(a < X_n \leq b) = F_n(b) - F_n(a) \simeq F(b) - F(a) \text{ for large } n.$$

This is useful when  $F(x)$  is easier to calculate than  $F_n(x)$ : for example, many complicated distributions converge to the Normal distribution (Central Limit Theorem), for which  $F(x)$  can be calculated by computer.

Moment generating functions are useful for finding the limiting distribution  $F$ .

**Theorem 5.4:** Suppose that  $X_1, X_2, \dots$  is a sequence of random variables with m.g.f.s  $M_{X_1}(t), M_{X_2}(t), \dots$  all defined for  $|t| < t_0$  (for some  $t_0 > 0$ ).

If  $M_{X_n}(t) \rightarrow M_X(t)$  for all  $|t| < t_0$  and for some r.v.  $X$ , then  $X_n \xrightarrow{D} X$ .

**Proof:** beyond the scope of this course.



## Practical use of Theorem 5.4:

If we can prove that  $M_{X_n}(t) \rightarrow M_X(t)$  as  $n \rightarrow \infty$ ,  
or (often easier) that  $\log M_{X_n}(t) \rightarrow \log M_X(t)$  as  $n \rightarrow \infty$ ,  
then we have proved that  $X_n \xrightarrow{D} X$  as  $n \rightarrow \infty$ .

## Theorem 5.5: The Central Limit Theorem

Let  $X_1, \dots, X_n$  be independent, identically distributed r.v.'s with m.g.f  $M_X(t)$  defined for all  $|t| < t_0$  (where  $t_0 > 0$ ).

Let  $\mathbb{E}(X_i) = \mu$  and  $\text{Var}(X_i) = \sigma^2$  for all  $i$ ,  
and let  $S_n = X_1 + X_2 + \dots + X_n$  be the sum of the first  $n$   $X_i$ 's.

$$\text{Let } Z_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{\text{Var}(S_n)}}.$$

Then  $Z_n \xrightarrow{D} Z$  as  $n \rightarrow \infty$ , where  $Z \sim N(0, 1)$ .

That is, 
$$\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{D} N(0, 1) \text{ as } n \rightarrow \infty.$$

**Proof:** (non-examinable)

i) Standardize  $X_1, \dots, X_n$  to have mean 0 and variance 1:

$$\text{Let } Y_i = \frac{X_i - \mu}{\sigma}.$$

Then  $Y_1, \dots, Y_n$  are independent and identically distributed, with  $\mathbb{E}(Y_i) = 0$ ,  $\text{Var}(Y_i) = 1$ , for all  $i$ .

$$\text{Also, } \sum_{i=1}^n Y_i = \frac{S_n - n\mu}{\sigma} = \sqrt{n} \left( \frac{S_n - n\mu}{\sqrt{n\sigma^2}} \right).$$

ii) Find  $M_Y(t)$ , the m.g.f. of  $Y_1, \dots, Y_n$ :

Any m.g.f. satisfies  $M_Y(t) = 1 + t\mathbb{E}(Y) + \frac{t^2}{2}\mathbb{E}(Y^2) + \frac{t^3}{3!}\mathbb{E}(Y^3) + \dots$

Here,  $\mathbb{E}(Y_i) = 0$ ,  $\mathbb{E}(Y_i^2) = \text{Var}(Y_i) + (\mathbb{E}Y_i)^2 = 1$ .

So

$$M_Y(t) = 1 + (t \times 0) + \left(\frac{t^2}{2} \times 1\right) + \underbrace{O(t^3)}_{\text{terms in } t^3 \text{ and above}}$$
$$M_Y(t) = 1 + \frac{t^2}{2} + O(t^3).$$

iii) Find the m.g.f. of  $Z_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}$  in terms of  $M_Y(t)$ :

---

We have  $Z_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \sum_{i=1}^n \frac{Y_i}{\sqrt{n}}$ , so by Theorem 5.3,

$$\begin{aligned} M_{Z_n}(t) &= \prod_{i=1}^n \left( M_Y \left( \frac{t}{\sqrt{n}} \right) \right) \\ &= \left[ M_Y \left( \frac{t}{\sqrt{n}} \right) \right]^n \\ &= \left\{ 1 + \frac{t^2}{2n} + O \left( \frac{t}{\sqrt{n}} \right)^3 \right\}^n. \end{aligned}$$

iv) Take logs:

$$\begin{aligned} \log(M_{Z_n}(t)) &= n \log \left\{ 1 + \frac{t^2}{2n} + O \left( \frac{t}{\sqrt{n}} \right)^3 \right\} \\ &= n \left\{ \left( \frac{t^2}{2n} + O \left( \frac{t}{\sqrt{n}} \right)^3 \right) - \frac{1}{2} \left( \frac{t^2}{2n} + O \left( \frac{t}{\sqrt{n}} \right)^3 \right)^2 + \dots \right\} \\ &= \frac{t^2}{2} + (\text{terms that } \rightarrow 0 \text{ as } n \rightarrow \infty). \end{aligned}$$

So  $\log M_{Z_n}(t) \rightarrow \frac{t^2}{2}$  as  $n \rightarrow \infty$ .

Thus  $M_{Z_n}(t) \rightarrow e^{t^2/2} = M_Z(t)$  as  $n \rightarrow \infty$ , where  $Z \sim N(0, 1)$ .  $\square$

### Notes:

1. This is a remarkable theorem, because the limit holds for **any** distribution of  $X_1, \dots, X_n$ .
2. The condition that  $M_X(t)$  exists is stronger than necessary: it is actually sufficient that  $\text{Var}(X)$  is finite. Still more versions of the Central Limit Theorem relax the conditions that  $X_1, \dots, X_n$  are independent and have the same distribution.
3. The **speed** of convergence of  $\frac{S_n - n\mu}{\sqrt{n\sigma^2}}$  to the Normal(0, 1) distribution **does** depend upon the distribution of  $X$ : distributions with large skewness and kurtosis converge more slowly than symmetric Normal-like distributions.

### Using the Central Limit Theorem to find the distribution of the mean, $\bar{X}$

Let  $\bar{X} = \frac{X_1 + \dots + X_n}{n} = \frac{S_n}{n}$ . Note that  $\mathbb{E}(\bar{X}) = \mu$ ,  $\text{Var}(\bar{X}) = \left(\frac{1}{n^2}\right) n\sigma^2 = \frac{\sigma^2}{n}$ .

Then  $\frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{n(\bar{X} - \mu)}{\sqrt{n\sigma^2}} = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} = \frac{\bar{X} - \mathbb{E}(\bar{X})}{\sqrt{\text{Var}(\bar{X})}}$ .

So the CLT also states that  $\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \xrightarrow{D} N(0, 1)$  as  $n \rightarrow \infty$ , ie.  $\bar{X} \xrightarrow{D} N(\mu, \sigma^2/n)$ .

The essential point to remember about the Central Limit Theorem is that large sums or sample means of independent random variables converge to a Normal distribution. With some distributions the CLT applies for as few as  $n = 4$  observations, while other distributions require larger  $n$ . Generally speaking, it is safe to assume that the Central Limit Theorem provides a good approximation whenever  $n \geq 30$ .

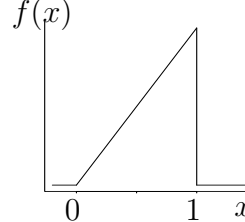
### Central Limit Theorem in action : simulation studies

The following simulation study illustrates the Central Limit Theorem, making use of several of the techniques learnt in STATS 210. Check all the working in the examples below.

**Example 1:** Triangular distribution:  $f_X(x) = 2x$  for  $0 < x < 1$ .

We find that  $\mathbb{E}(X) = \mu = \frac{2}{3}$ ,  $\text{Var}(X) = \sigma^2 = \frac{1}{18}$ .

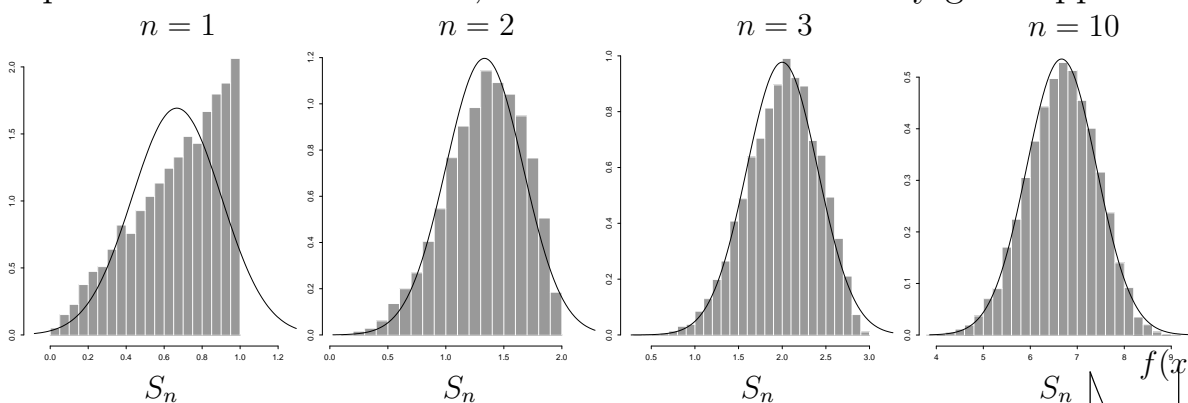
The distribution function is  $F_X(x) = x^2$  for  $0 < x < 1$ , with  $F_X(x) = 0$  for  $x \leq 0$ ,  $F_X(x) = 1$  for  $x \geq 1$ .



The inverse distribution function is therefore  $F_X^{-1}(u) = \sqrt{u}$ , for  $0 < u < 1$ .

We can generate samples of size  $n$  from this distribution using the method of §3.4: generate  $U_1, \dots, U_n \sim \text{Uniform}(0, 1)$  and let  $X_i = \sqrt{U_i}$  for  $i = 1, \dots, n$ .

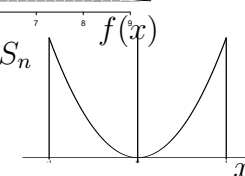
The graph shows histograms of 10 000 values of  $S_n = X_1 + \dots + X_n$  for  $n = 1, 2, 3$ , and 10. The Normal p.d.f.  $N(n\mu, n\sigma^2) = N(\frac{2}{3}n, \frac{1}{18}n)$  is superimposed across the top. Even for  $n$  as low as 10, the Normal curve is a very good approximation.



**Example 2:** U-shaped distribution:  $f_X(x) = \frac{3}{2}x^2$  for  $-1 < x < 1$ .

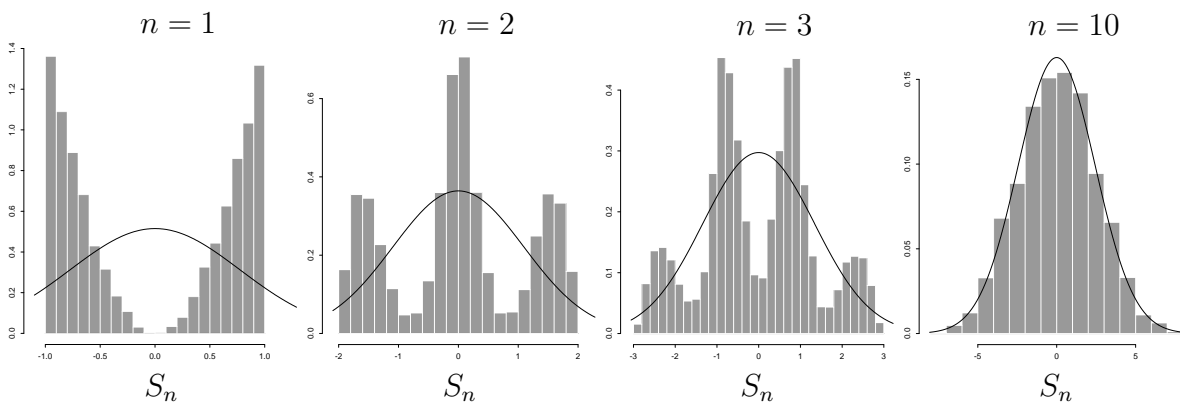
We find that  $\mathbb{E}(X) = \mu = 0$ ,  $\text{Var}(X) = \sigma^2 = \frac{3}{5}$ .

$F_X(x) = \frac{1}{2}(x^3 + 1)$  for  $-1 < x < 1$ , so  $F_X^{-1}(u) = (2u - 1)^{1/3}$ , for  $0 < u < 1$ .



We generate samples  $X_1, \dots, X_n$  using  $X_i = (2U_i - 1)^{1/3}$  for  $i = 1, \dots, n$ .

Even with this highly non-Normal distribution for  $X$ , the Normal curve provides a good approximation to  $S_n = X_1 + \dots + X_n$  for  $n$  as small as 10.



# Chapter 6: Sampling Theory for the

---

## Normal Distribution

---

### 6.1 Introduction

---

The aim in this chapter is to establish the theory behind the  $t$ -tests and  $t$ -based confidence intervals described in Stage I courses. These tests are designed for **Normal distributions**: you might remember from Stage I that we only use  $t$ -tests and  $t$ -based confidence intervals when we are satisfied that plots of the data show no evidence of severe non-Normality.

We need to establish the following results.

Let  $X_1, \dots, X_n$  be *independent* and *identically distributed* such that each  $X_i \sim \text{Normal}(\mu, \sigma^2)$ . Then:

1. The sample mean,  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  has distribution  $\bar{X} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$ .
2. The sample variance,  $S_X^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$  satisfies

$$\left(\frac{n-1}{\sigma^2}\right) S_X^2 \sim \text{Chisquare}(n-1).$$

3. The random variables  $\bar{X}$  and  $S_X^2$  are independent.
4. The  $t$ -ratio,

$$T = \frac{\bar{X} - \mu}{\sqrt{(S_X^2/n)}} = \frac{\bar{X} - \mu}{\text{se}(\bar{X})},$$

has a distribution called the Student's  $t$ -distribution, with p.d.f. to be determined.

The reason for needing to use the  $t$ -distribution is that *we are interested in the unknown mean,  $\mu$ , but not in the unknown variance,  $\sigma^2$ . The  $t$ -ratio involves  $\mu$ , but not  $\sigma^2$ , so it eliminates the nuisance parameter  $\sigma^2$ .*

## 6.2 Distribution Theory

**Theorem 6.1:** Let  $X_1, \dots, X_n$  be independent, with  $X_i \sim \text{Normal}(\mu, \sigma^2)$  for all  $i$ .

Then

(a)  $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma^2}{n})$

(b)  $\sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 \sim \text{Chisquare}(n)$ .

**Proof:**

a)

$\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$ . Find the m.g.f. of  $\bar{X}$ :

$$\begin{aligned} \text{By Theorem 5.3, } M_{\bar{X}}(t) &= \left\{ M_X \left( \frac{t}{n} \right) \right\}^n \\ &= \left\{ e^{\left( \mu \frac{t}{n} + \frac{1}{2} \sigma^2 \frac{t^2}{n^2} \right)} \right\}^n \\ &= e^{\left\{ \mu t + \frac{1}{2} \left( \frac{\sigma^2}{n} \right) t^2 \right\}}, \end{aligned}$$

which is the m.g.f. of the  $\text{Normal}(\mu, \frac{\sigma^2}{n})$  distribution. So,  $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma^2}{n})$ .

b)

Let  $Z_i = \frac{X_i - \mu}{\sigma}$ : then  $Z_i \sim \text{Normal}(0, 1)$ .

By Example in Section 5.5, this means that  $Z_i^2 \sim \text{Chisquare}(1)$ .

Now by Example 3, Section 5.3, the sum of independent  $\text{Chisquare}(\nu_i)$  r.v's has distribution  $\text{Chisquare}(\sum \nu_i)$ .

$$\text{Thus } \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^n Z_i^2 \sim \text{Chisquare}(\sum_{i=1}^n 1) = \chi_n^2 \quad (\text{Chisquare}(n)).$$

**Note:** If  $X_1, \dots, X_n$  are **not** Normal, then (a) still holds approximately for large  $n$ , by the Central Limit Theorem. The approximation is less good for (b): Normality is more important.

### Drawing inference about the unknown mean, $\mu$

Usually,  $\mu$  and  $\sigma^2$  are unknown: in real life, we observe  $X_1, \dots, X_n$  and use them to make inferences (statements) about the mean,  $\mu$ .

$\sigma^2$  is usually a nuisance parameter: it is unknown, but not of primary interest.

We aim to find a quantity with a **known distribution**, that **does not depend on  $\sigma^2$** , so that we can concentrate on drawing inference about the mean,  $\mu$ .

Consider the following.

**Lemma :** Let  $X_1, \dots, X_n$  be independent, with  $X_i \sim \text{Normal}(\mu, \sigma^2)$  for all  $i$ .

Define the **vector of residuals**,

$$\mathbf{R} = \left( (X_1 - \bar{X}), (X_2 - \bar{X}), \dots, (X_n - \bar{X}) \right).$$

Then  $\bar{X}$  and  $\mathbf{R}$  are independent.

**Proof:** (sketch)

- Find the multivariate moment generating function of the vector  $(\bar{X}, \mathbf{R})$ :

$$\begin{aligned} M(t_0, t_1, \dots, t_n) &= \mathbb{E} \left( e^{t_0 \bar{X} + t_1 R_1 + \dots + t_n R_n} \right) \\ &= \mathbb{E} \left( e^{t_0 \bar{X} + t_1 (X_1 - \bar{X}) + \dots + t_n (X_n - \bar{X})} \right). \end{aligned}$$

- Show that  $M(t_0, t_1, \dots, t_n)$  factorizes as  $a(t_0)b(t_1, \dots, t_n)$ . There is a theorem that states that random variables are independent if and only if their multivariate moment generating functions factorize in this way.  $\square$

**Theorem 6.2:** Let  $X_1, \dots, X_n$  be independent, with  $X_i \sim \text{Normal}(\mu, \sigma^2)$  for all  $i$ .

Let  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  be the sample mean,

and let  $S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  be the sample variance.

Then

(a)  $\bar{X}$  and  $S_X^2$  are independent.

(b)  $\left(\frac{n-1}{\sigma^2}\right) S_X^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \text{Chisquare}(n-1)$ .

**Proof:**

(a) Direct from the Lemma: if  $\bar{X}$  and  $\mathbf{R}$  are independent, then  $\bar{X}$  and  $S_X^2$  must also be independent, because  $S_X^2$  is a function of  $\mathbf{R}$ .

(b) Let  $U = \frac{n(\bar{X} - \mu)^2}{\sigma^2}$ , and let  $V = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}$ .

We wish to prove that  $V \sim \text{Chisquare}(n-1)$ .

**Method:**

- (i) show that  $U$  and  $V$  are independent;
- (ii) find the MGF of  $U + V$ ,  $M_{U+V}(t)$ ;
- (iii) by independence,  $M_{U+V}(t) = M_U(t)M_V(t)$ ;
- (iv) hence, knowing  $M_U(t)$ , find  $M_V(t)$ .

(i)  $U$  is a function of  $\bar{X}$  only, and  $V$  is a function of  $\mathbf{R} = ((X_1 - \bar{X}), \dots, (X_n - \bar{X}))$  only. Thus  $U$  and  $V$  are independent by the Lemma.

(ii)  $U + V = \frac{n(\bar{X} - \mu)^2}{\sigma^2} + \underbrace{\sum_{i=1}^n \frac{\{(X_i - \mu) - (\bar{X} - \mu)\}^2}{\sigma^2}}_{\text{expand and use the fact that } \sum_{i=1}^n (X_i - \mu) = n(\bar{X} - \mu)}$

$\rightarrow$  gives  $U + V = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$ .



But we have already found the distribution of  $\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$ : by Theorem 6.1, we have  $\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \sim \text{Chisquare}(n)$ .

Thus  $U + V \sim \text{Chisquare}(n)$ , so  $M_{U+V}(t) = (1 - 2t)^{-n/2}$ .  $\circledast$

(iii) By independence,  $M_{U+V}(t) = (1 - 2t)^{-n/2} = M_U(t)M_V(t)$ .  $\circledast\circledast$

Now  $U = \frac{(\bar{X} - \mu)^2}{(\sigma^2/n)}$ , and by Theorem 6.1(a),  $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma^2}{n})$ , so  $U \sim \text{Chisquare}(1)$  (square of a  $\text{Normal}(0, 1)$  r.v.).

So  $M_U(t) = (1 - 2t)^{-1/2}$ .

(iv) Thus  $\circledast$  and  $\circledast\circledast$  give:

$$\begin{aligned} M_V(t) &= \frac{M_{U+V}(t)}{M_U(t)} = \frac{(1 - 2t)^{-n/2}}{(1 - 2t)^{-1/2}} \\ &= (1 - 2t)^{-(n-1)/2} \end{aligned}$$

So  $V \sim \text{Chisquare}(n - 1)$  as required.

---

We are now able to eliminate the nuisance parameter  $\sigma^2$ .

We have  $\bar{X} \sim \text{Normal}(\mu, \frac{\sigma^2}{n}) \Rightarrow Z = \frac{(\bar{X} - \mu)\sqrt{n}}{\sigma} \sim \text{Normal}(0, 1)$ . Thm 6.1(a)

Also,  $V = \left(\frac{n-1}{\sigma^2}\right) S_X^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \text{Chisquare}(n-1)$ . Thm 6.2(b)

Also,  $V$  and  $Z$  are independent. Thm 6.2(a)

Consider the quantity

$$T = \frac{Z}{\sqrt{\frac{V}{n-1}}} = \frac{\left(\frac{(\bar{X} - \mu)\sqrt{n}}{\sigma}\right)}{\sqrt{\frac{(n-1) S_X^2}{\sigma^2} \frac{1}{n-1}}} = \frac{(\bar{X} - \mu)}{\sqrt{\frac{S_X^2}{n}}}$$

$$T = \frac{(\bar{X} - \mu)}{\sqrt{\frac{S_x^2}{n}}} \text{ does not depend upon } \sigma^2.$$

Furthermore, the distribution of  $T$  is quite easy to find.

**Theorem 6.3:** Let  $Z \sim \text{Normal}(0, 1)$ , and let  $V \sim \text{Chisquare}(r)$ , and suppose that  $Z$  and  $V$  are independent.

Let  $T = \frac{Z}{\sqrt{\frac{V}{r}}}$ . Then  $T$  has p.d.f.

$$f_T(t) = \left( \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi} \Gamma\left(\frac{r}{2}\right)} \right) \left( 1 + \frac{t^2}{r} \right)^{-(r+1)/2} \quad \text{for } -\infty < t < \infty.$$

This is defined as the *Student's  $t$ -distribution with  $r$  degrees of freedom*:

$$T \sim t_r \quad \text{or} \quad T \sim \text{Student}(df = r).$$

**Proof:** (sketch)

Use the bivariate change of variable technique:

- find the joint density of  $Z$  and  $V$  by independence:

$$f_{Z, V}(z, v) = f_Z(z) f_V(v).$$

- define two new random variables:  $T = \frac{Z}{\sqrt{\frac{V}{r}}}$ , and  $U = V$ .
- Use the bivariate change of variable technique to find  $f_{T, U}(t, u)$ .
- Find the marginal p.d.f. of  $T$ ,

$$f_T(t) = \int_0^{\infty} f_{T, U}(t, u) du.$$

**Proof:** (detailed)

Let  $Z \sim \text{Normal}(0, 1)$  and let  $V \sim \chi_r^2$  (i.e.  $V \sim \text{Chisquare}(r)$ ), and let  $Z$  and  $V$  be independent. The joint density of  $Z$  and  $V$  is

$$f(z, v) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \frac{1}{2^{\frac{r}{2}} \Gamma(\frac{r}{2})} v^{\frac{r}{2}-1} e^{-\frac{v}{2}}, \quad -\infty < z < \infty, \quad v \geq 0.$$

Now let  $T = \sqrt{r} Z / \sqrt{V}$ , and  $U = V$ . The transformation is monotone. Inverting, we obtain  $Z = \sqrt{U} T / \sqrt{r}$  and  $V = U$ , so

$$|\det J| = \left| \det \begin{pmatrix} \frac{\partial z}{\partial t} & \frac{\partial z}{\partial u} \\ \frac{\partial v}{\partial t} & \frac{\partial v}{\partial u} \end{pmatrix} \right| = \left| \det \begin{pmatrix} \frac{\sqrt{u}}{r} & \frac{t}{2\sqrt{ru}} \\ 0 & 1 \end{pmatrix} \right| = \sqrt{\frac{u}{r}}.$$

Thus  $T$  and  $U$  have joint p.d.f.

$$\begin{aligned} f_{T,U}(t, u) &= \frac{1}{\sqrt{2\pi} 2^{\frac{r}{2}} \Gamma(\frac{r}{2})} e^{-\frac{ut^2}{2r}} u^{\frac{r}{2}-1} e^{-\frac{u}{2}} \frac{\sqrt{u}}{\sqrt{r}}, \quad -\infty < t < \infty, \quad u \geq 0 \\ &= \frac{u^{\frac{r-1}{2}} e^{-\frac{u}{2}(1+\frac{t^2}{r})}}{\sqrt{2\pi r} 2^{\frac{r}{2}} \Gamma(\frac{r}{2})}, \quad -\infty < t < \infty, \quad u \geq 0. \end{aligned}$$

Thus  $T$  has marginal p.d.f.

$$f_T(t) = \frac{1}{\sqrt{2\pi r} 2^{\frac{r}{2}} \Gamma(\frac{r}{2})} \int_0^\infty u^{\frac{r+1}{2}-1} e^{-\frac{u}{2}(1+\frac{t^2}{r})} du.$$

The integral is proportional to the integral of a Gamma p.d.f. with  $k = (\frac{r+1}{2})$  and  $\lambda = \frac{1}{2} \left(1 + \frac{t^2}{r}\right)$ . In general, for a Gamma( $k, \lambda$ ) integral we have

$$\int_0^\infty \frac{1}{\Gamma(k)} \lambda^k u^{k-1} e^{-\lambda u} du = 1, \quad \Rightarrow \quad \int_0^\infty u^{k-1} e^{-\lambda u} du = \frac{\Gamma(k)}{\lambda^k}.$$

Substituting  $k = (\frac{r+1}{2})$  and  $\lambda = \frac{1}{2} \left(1 + \frac{t^2}{r}\right)$  gives

$$\begin{aligned} f_T(t) &= \left( \frac{1}{\sqrt{2\pi r} 2^{\frac{r}{2}} \Gamma(\frac{r}{2})} \right) \left( \frac{\Gamma(\frac{r+1}{2})}{\left\{ \frac{1}{2} \left(1 + \frac{t^2}{r}\right) \right\}^{\frac{r+1}{2}}} \right) \\ &= \left( \frac{\Gamma(\frac{r+1}{2})}{\sqrt{r\pi} \Gamma(\frac{r}{2})} \right) \left( 1 + \frac{t^2}{r} \right)^{-(r+1)/2} \quad \text{for } -\infty < t < \infty. \quad \square \end{aligned}$$

The results above together prove the following theorem.

**Theorem 6.4:** Let  $X_1, \dots, X_n$  be independent, with  $X_i \sim \text{Normal}(\mu, \sigma^2)$  for all  $i$ .

Then 
$$Z = \frac{(\bar{X} - \mu)\sqrt{n}}{\sigma} \sim \text{Normal}(0, 1);$$

$$V = \left(\frac{n-1}{\sigma^2}\right) S_X^2 = \left(\frac{n-1}{\sigma^2}\right) \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \sim \text{Chisquare}(n-1);$$

$Z$  and  $V$  are independent;

and 
$$T = \frac{Z}{\sqrt{V/(n-1)}} = \frac{\bar{X} - \mu}{\sqrt{(S_X^2/n)}} \sim \text{Student}(\text{df} = n-1). \quad \square$$

---

### 6.3 Application to confidence intervals and $t$ -tests

---

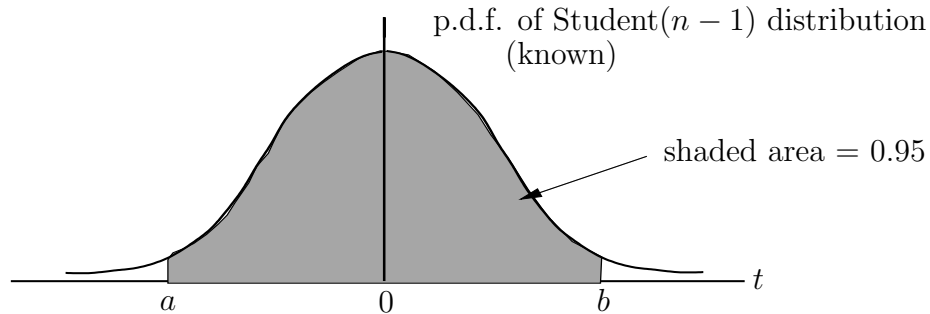
We have discovered that we can derive a quantity  $T = \frac{\bar{X} - \mu}{\sqrt{(S_X^2/n)}}$  with a known distribution, where  $T$  depends upon the unknown mean  $\mu$  but **not** on the unknown variance  $\sigma^2$  (the nuisance parameter). This means that, even without any knowledge of  $\sigma^2$ , we can predict how  $T$  should behave and draw conclusions about the unknown mean,  $\mu$ . Two examples are:

1. Because  $T$  is related to  $(\bar{X} - \mu)$ , we can predict **how far away** the observed sample mean,  $\bar{X}$ , should lie from the true mean,  $\mu$ , and therefore construct an **interval** in which  $\mu$  is likely to lie (a **confidence interval for  $\mu$** ).
2. We can **guess** (hypothesize) a value of  $\mu$  and **test** whether it is plausible. For example, if the true value of  $\mu$  is 5, then the true  $T$ -statistic is  $T = \frac{\bar{X} - 5}{\sqrt{(S_X^2/n)}}$ , and its distribution is known to be the Student(df= $n-1$ ) distribution. However, if the true value of  $\mu$  is **not** 5, then the quantity  $\frac{\bar{X} - 5}{\sqrt{(S_X^2/n)}}$  will have a **different** (unknown) distribution. Therefore, we can look at the value of  $\frac{\bar{X} - 5}{\sqrt{(S_X^2/n)}}$ , to see whether it is consistent with the Student(df= $n-1$ ) distribution. If it is not, we have to conclude that the true value of  $\mu$  is probably **not** 5.

## 1. Confidence intervals for the mean, $\mu$

We have  $T = \frac{\bar{X} - \mu}{\sqrt{S_X^2/n}} \sim \text{Student}(df = n - 1)$ . Because the Student( $df = n - 1$ ) distribution is known, we are able to find points  $a$  and  $b$  such that

$$\mathbb{P}\left(a < \frac{\bar{X} - \mu}{\sqrt{S_X^2/n}} < b\right) = 0.95.$$



Usually, we choose  $a = -b$ , so:

$$\begin{aligned} & \mathbb{P}\left(-b < \frac{\bar{X} - \mu}{\sqrt{S_X^2/n}} < b\right) = 0.95 \\ \Rightarrow & \mathbb{P}\left\{\left(\bar{X} - b\sqrt{\frac{S_X^2}{n}}\right) < \mu < \left(\bar{X} + b\sqrt{\frac{S_X^2}{n}}\right)\right\} = 0.95. \end{aligned}$$

Thus, with 95% probability, the interval  $\left(\bar{X} - b\sqrt{\frac{S_X^2}{n}}, \bar{X} + b\sqrt{\frac{S_X^2}{n}}\right)$  encloses the unknown value  $\mu$ . This is called a

95% confidence interval for  $\mu$ .

**Note:**  $\bar{X}$  and  $S_X^2$  are observed from the data: they are *random*.

$b$  is calculated from the  $t$ -distribution, and it is *not random*. It is the unique value that satisfies  $\mathbb{P}(-b < T < b) = 0.95$  where  $T \sim \text{Student}(df = n - 1)$ .

$\mu$  is unknown, but fixed (*not random*).

The confidence interval is random because of  $\bar{X}$  and  $S_X^2$ . It contains  $\mu$  with probability 0.95, but this is a probability statement about  $\bar{X}$  and  $S_X^2$ , *not* about  $\mu$  (which is fixed).

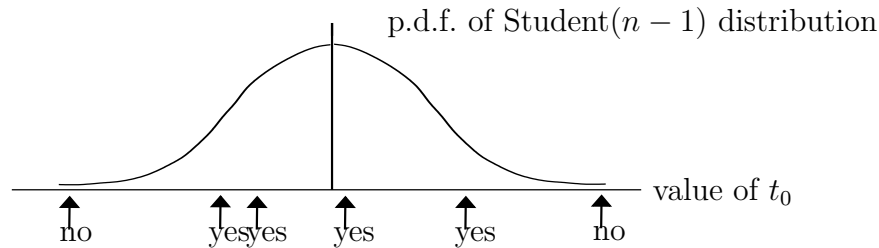
## 2. Hypothesis tests

Let  $H_0 : \mu = \mu_0$  be the null (favoured) hypothesis. ( $\mu_0$  here is a specified number, for example  $\mu_0 = 5$  or  $\mu_0 = 0$ .)

If  $H_0$  is **true**, then  $\mu = \mu_0$ , so  $T_0 = \frac{\bar{X} - \mu_0}{\sqrt{S_X^2/n}} \sim \text{Student}(df = n - 1)$ .

Testing  $H_0$  : Calculate the value of  $t_0 = \frac{\bar{x} - \mu_0}{\sqrt{(s_x^2/n)}}$ .

Does it **look** as if it came from the Student( $df = n - 1$ ) distribution?

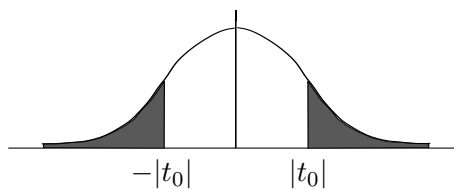


If **yes**, accept that  $H_0$  is possibly true.

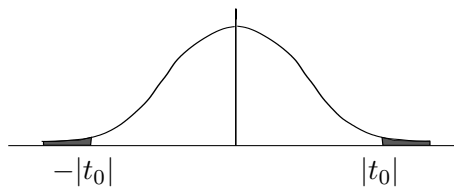
If **no**, we have evidence against  $H_0$ .

We summarize evidence by the  $p$ -value:

$$p = \mathbb{P}(T \geq |t_0| \text{ if } T \sim \text{Student}(n - 1)) = \text{total shaded area.}$$



Large  $p$ -value  $\Rightarrow$  no evidence against  $H_0$ .  
 $t_0$  is a reasonable observation from the Student( $n - 1$ ) distribution.



Small  $p$ -value  $\Rightarrow$  evidence against  $H_0$ .  
 $t_0$  is a very unusual observation from the Student( $n - 1$ ) distribution.

### Note: Non-Normal populations

The procedures above ( $t$ -tests and  $t$ -based confidence intervals) are often applied when  $X_1, \dots, X_n$  are **not** drawn from a Normal distribution. This is acceptable in large samples if the distribution of  $X_1, \dots, X_n$  is reasonably symmetric. However, the procedures are not valid for highly skewed distributions.