

SCS 1307

Probability & Statistics



Complete Reference Notes

Descriptive statistics · Axiomatic probability · Conditional probability
and Bayes' theorem · Independence · Random variables · Binomial,
Poisson and Normal distributions

Compiled from the nine course handouts of
Dr Dilshani Tissera and **Dr Dilanthi Gamage Dona**
Department of Statistics, University of Colombo

*With full worked solutions to every exercise, complete derivations,
and diagrams throughout.*

How to use these notes

These notes cover the entire SCS 1307 syllabus as delivered in Handouts 01–09. They are not a transcript of the slides: every definition has been restated precisely, every result the slides assert has been proved, every blank “Solution” slide has been worked through in full, and diagrams have been added wherever a picture makes the idea clearer than a paragraph.

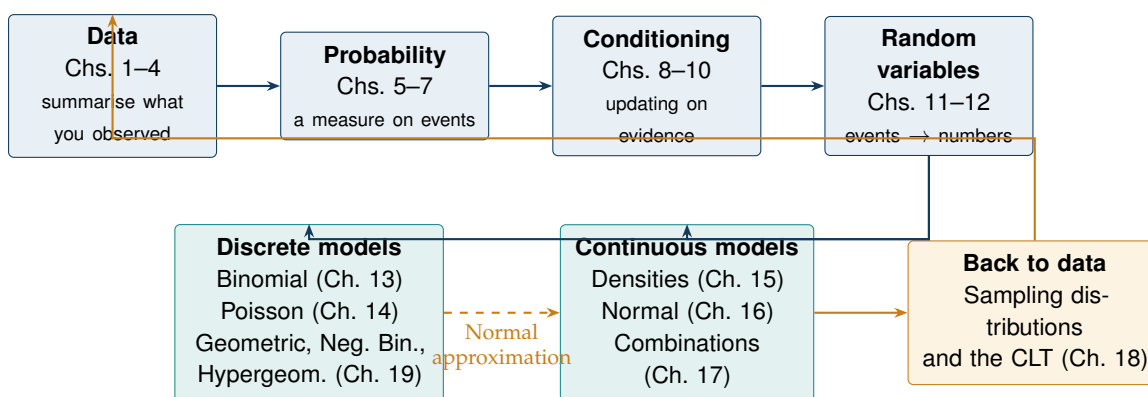
The colour code

Boxes are colour-coded so you can navigate by what you need.

Box	Contains
■ Definition	The precise statement of a term. These are what an exam expects you to reproduce.
■ Theorem / Property	A result that follows from the definitions, together with its proof.
■ Example	A fully worked calculation. Read the reasoning, not just the arithmetic.
■ Exercise	A problem from the handouts, with a complete solution beneath it.
■ Intuition	Why the result is true in plain language, before the algebra.
■ Common pitfall	A mistake that is made every single year.

The shape of the course

The syllabus has a clear logical spine, and it helps to see it before you start. Everything below flows left to right: each box depends only on what precedes it.



Two arrows deserve flagging now. The dashed one is one of the most useful facts in the course: a discrete model (Binomial or Poisson) can be replaced by a Normal one when its parameters get large — that is Chapter 16. The long arrow back to the left is the payoff of the whole subject: Chapter 18 uses everything before it to say how far a sample mean is likely to lie from the population mean, which is the inferential leap that Chapter 1 opened with.

A note on the source handouts

These notes cover Handouts 01–12. (Handout 10 is byte-identical to Handout 09, so the Poisson material appears once, in Chapter 14.)

Four corrections to the handouts are flagged in the text where they occur and collected in Appendix C: two in Handout 02 (Sections 6.5 and 5.2) and two in Handout 12 (the Hypergeometric variance formula in Section 19.5, and the answer to Revision Exercise 3 in Chapter 20). Everything else follows the handouts' notation exactly, including the use of A' for the complement of A and Ω for the outcome space.

Contents

How to use these notes	1
I Foundations of Statistics	8
1 What Statistics Is	9
1.1 The definition, and what it hides	9
1.2 The vocabulary of an investigation	9
1.3 Why we study statistics	10
1.4 The problem that motivates Part I	11
1.5 Learning outcomes for the course	11
2 Types of Variables and Levels of Measurement	13
2.1 Why the type of a variable matters	13
2.2 The two kinds of variable	13
2.3 Discrete versus continuous	14
2.4 Levels of measurement within categorical data	14
2.5 The complete taxonomy	14
2.6 Which summaries are legal at which level	14
3 Measures of Location	16
3.1 What a measure of location does	16
3.2 The mean	16
3.2.1 The weighted mean	17
3.3 The median	17
3.4 Mean versus median	18
3.5 The mode	19
3.6 Resistant measures	19
3.7 Quartiles and percentiles	20
3.7.1 The procedure the course uses for P_k	20
3.8 Worked example: exam marks of 50 students	20
4 Measures of Dispersion	23
4.1 Why a centre alone is not enough	23
4.2 Range and interquartile range	23
4.3 Variance and standard deviation	24
4.3.1 Why we square the deviations	25
4.4 The effect of linear transformations	25
4.5 Worked example: dispersion of the 50 exam marks	26
4.6 Skewness	27

II	Probability	29
5	Sample Spaces and Events	30
5.1	The four objects you need before you can compute anything	30
5.2	Events	31
5.3	Building outcome spaces: the worked cases	32
5.4	Set operations on events, in probability language	32
6	The Axioms of Probability	34
6.1	Probability as a function	34
6.2	The three axioms	35
6.3	Properties derived from the axioms	35
6.4	Assigning probabilities in practice	36
6.4.1	Method 1: symmetry (the classical approach)	37
6.4.2	Method 2: relative frequency	37
6.5	Exercises from Handout 02, worked in full	38
7	Counting Techniques	43
7.1	Why counting deserves a chapter	43
7.2	The multiplication principle	43
7.3	Permutations: order matters	43
7.4	Combinations: order does not matter	44
7.5	Useful identities	45
7.6	Worked counting problems	46
8	Conditional Probability	48
8.1	The idea: shrinking the sample space	48
8.2	The multiplication rule	48
8.3	First examples	49
8.4	Sampling without replacement	50
8.5	A three-coin exercise	51
9	Total Probability and Bayes' Theorem	53
9.1	The two-case version of the Law of Total Probability	53
9.2	The general Law of Total Probability	53
9.3	Worked examples of total probability	54
9.4	Bayes' theorem	56
9.5	The medical test: why Bayes matters	57
9.6	Bayes with more than two causes	58
10	Independence	61
10.1	Definition	61
10.2	Properties and their proofs	61
10.3	Independence versus mutual exclusivity	62
10.4	Independence of more than two events	63
10.5	Worked examples	64
10.6	Pairwise but not mutual independence	66
10.7	Conditional independence	67

III Random Variables and Distributions	70
11 Random Variables and Probability Distributions	71
11.1 Turning outcomes into numbers	71
11.2 First examples	72
11.3 Probability distributions	73
11.4 Functions of random variables	75
11.5 Extended example: the rectangle-cutting machine	75
11.6 The cumulative distribution function	77
12 Expectation and Variance	79
12.1 Why we need summary numbers	79
12.2 Expectation	79
12.2.1 Properties of expectation	80
12.3 Variance	82
12.4 Worked variance problems	83
13 The Binomial Distribution	87
13.1 The binomial situation	87
13.2 Deriving the distribution	87
13.3 Mean and variance	89
13.4 Exercises from Handout 06, worked in full	90
14 The Poisson Distribution	94
14.1 Definition	94
14.2 Expectation and variance	94
14.3 Uses of the Poisson distribution	95
14.4 Worked examples	96
14.5 Scaling the interval	97
14.6 The Normal approximation to the Poisson	98
15 Continuous Random Variables	101
15.1 What is a continuous random variable?	101
15.2 The probability density function	101
15.3 Example: the exponential density	102
15.4 Exercises from Handout 07, worked in full	103
15.5 Expectation and variance	105
15.6 The cumulative distribution function	107
16 The Normal Distribution	109
16.1 Why it matters	109
16.2 The empirical rule	110
16.3 The standard normal distribution	110
16.4 Using the table	112
16.5 The Normal approximation to the Binomial	114
16.5.1 The continuity correction	114
16.6 Choosing the right distribution	115

IV	Sampling and Further Distributions	117
17	Linear Combinations of Normal Variables	118
17.1	Why this chapter exists	118
17.2	Sum and difference of two independent normal variables	118
17.3	Worked examples	119
17.4	Multiples of normal variables	120
17.5	The general linear combination	121
17.6	An extended exercise	122
18	Sampling Distributions and the Central Limit Theorem	124
18.1	Closing the loop	124
18.2	The distribution of the sample mean	124
18.3	The Central Limit Theorem	126
18.4	Two theorems, one picture	128
19	Further Discrete Distributions	130
19.1	Four more models	130
19.2	The Discrete Uniform distribution	130
19.3	The Geometric distribution	131
19.4	The Negative Binomial distribution	133
19.5	The Hypergeometric distribution	134
19.6	How the discrete distributions relate	137
19.7	Summary table	137
20	Revision Exercises	139
20.1	Choosing the right distribution	139
20.2	The exercises	140
20.3	A consolidated worked comparison	143
20.4	Further practice	144
V	Appendices	145
A	Master Formula Sheet	146
	Part I — Descriptive statistics	146
	Part II — Probability	147
	Part III — Random variables	147
	The distributions	148
	Combinations of normal variables and sampling	148
	Normal distribution procedures	149
	Useful constants	149
B	Standard Normal Distribution Table	150
	How to read this table	150
	Table of values	151
	Critical values	151

C	Answers, Notation and Errata	153
	Answer summary	153
	Notation index	156
	Errata in the source handouts	156
	Source handouts	157

PART I

Foundations of Statistics

What Statistics Is

1.1 The definition, and what it hides

The course opens with a one-line definition:

Definition: Statistics

Statistics is the science of **collecting, organising, analysing** and **presenting** data, and finally **drawing conclusions** in the best possible way.

That sentence is easy to nod along to and easy to underestimate. The phrase that carries all the weight is *drawing conclusions*. Collecting and organising data are administrative tasks; analysing it is arithmetic. What makes statistics a distinct discipline is the last step, in which you say something about a group you did *not* measure on the basis of a group you *did*. That leap is never certain, and the whole of probability theory — Chapters 5 to 16 of these notes — exists to quantify exactly how uncertain it is.

Intuition: Why probability is half of a statistics course

Suppose you measure the heights of 200 students and get a mean of 168 cm. The arithmetic is trivial. The hard question is: what does that tell you about the 40,000 students you did not measure? To answer it you need to know how much a sample mean tends to wander from the population mean — and *that* is a probability question. Descriptive statistics (Part I) describes what you saw. Probability (Parts II and III) is the machinery that lets you generalise from it.

1.2 The vocabulary of an investigation

Nearly every sentence in this course uses at least one of the following terms, and they are frequently confused with one another. Learn them as a connected system rather than as isolated definitions.

Definition: The basic terms

Population	The complete collection of individuals or objects that are of interest to the study.
Sample	A subset of the population.
Variable	An observable, measurable or recordable characteristic. It is the basic unit of analysis.
Observation	A value that a variable assumes for a single element.
Data	The set of observations collected for the variable.
Parameter	A numerical summary measure used to describe a characteristic of a <i>population</i> .
Statistic	A numerical summary measure used to describe a characteristic of a <i>sample</i> .

The relationship between these is best seen as a cycle rather than a list. You sample from a population, compute a statistic, and use it to infer a parameter.

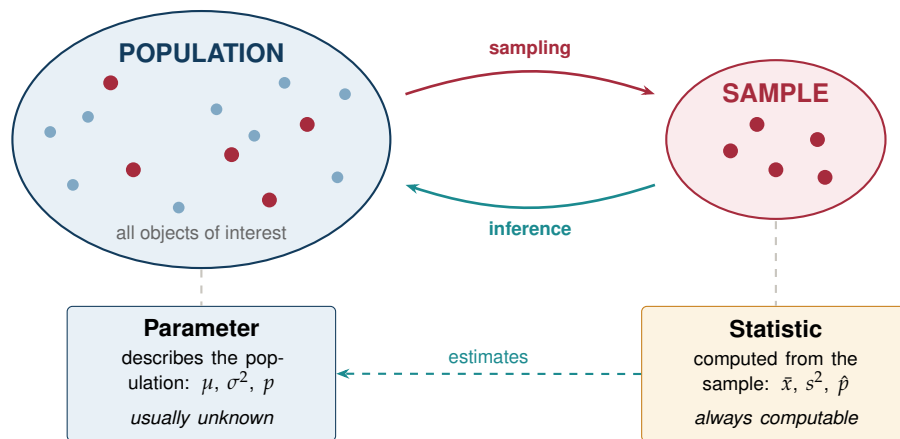


Figure 1.1: The sampling–inference cycle. A *statistic* is a number you can always compute because you hold the sample; a *parameter* is a number you generally cannot compute because you do not hold the whole population. The entire purpose of the subject is the dashed arrow.

Common pitfall: Parameter versus statistic

The distinction is *not* about what the number measures — both may be a mean. It is about *which group* it describes. The mean height of all 40,000 students is a parameter; the mean height of the 200 you measured is a statistic. By convention Greek letters (μ, σ, p, λ) denote parameters and Roman letters ($\bar{x}, s, \hat{p}$) denote statistics.

Example 1.1 (Identifying the terms)

A researcher wants to know the average daily screen time of undergraduates at the University of Colombo. She records the screen time of 350 randomly chosen undergraduates and obtains a mean of 4.7 hours.

Term	In this study
Population	All undergraduates at the University of Colombo
Sample	The 350 undergraduates actually recorded
Variable	Daily screen time (in hours)
Observation	One student's recorded value, e.g. 3.2 hours
Data	The full list of 350 recorded values
Statistic	$\bar{x} = 4.7$ hours
Parameter	μ , the (unknown) mean screen time of <i>all</i> undergraduates

1.3 Why we study statistics

The handout lists seven motivations. They read as a progression from passive to active competence:

1. **To evaluate published data.** The most common use for most graduates. Recognising a misleading average, an unrepresentative sample or an overstated conclusion is a defensive skill.

2. **To collect and summarise data meaningfully.** Deciding *what* to record is a statistical decision made before any arithmetic happens.
3. **To present the findings**, in reports or verbally.
4. **To interpret the statistical results**, which is distinct from producing them. Software will produce a number regardless of whether it means anything.
5. **To employ statistical methods to make decisions.**
6. **To develop new techniques.**
7. **To generalise the decisions** — the inferential leap again.

Statistical software

The handouts name EXCEL, SPSS, MINITAB, SAS, S-Plus, MATLAB, R and Python. In research and industry **R** and **Python** dominate; Excel remains ubiquitous for small datasets and SPSS for the social sciences. None of them will tell you whether your analysis was appropriate — that judgement is what this course is training.

1.4 The problem that motivates Part I

Definition: The summarisation problem

In practice we have a very large quantity of data. It is impossible to look at each observation individually and draw a conclusion. We therefore need ways to **summarise** the data so that its main features are brought out.

This is the agenda of Chapters 2 to 4. A summary must be chosen to match the *kind* of data you hold (Chapter 2), it must say *where* the data sits (Chapter 3), and it must say *how spread out* it is (Chapter 4). Any one of those on its own is misleading.

Intuition: Why one number is never enough

The three datasets $\{2, 5, 6, 8, 9\}$, $\{5, 5, 6, 6, 6\}$ and $\{6, 6, 6, 6, 6\}$ all have median 6. Reported as “the median is 6” they are indistinguishable, yet the first is highly variable and the third is perfectly constant. A measure of *centre* without a measure of *spread* conceals more than it reveals — which is precisely why Chapter 4 exists.

1.5 Learning outcomes for the course

For reference, the six stated learning outcomes and where each is addressed.

	Learning outcome	Covered in
LO1	Calculate probabilities of events and expectations of random variables for elementary problems such as games of chance.	Chs. 6, 12
LO2	Differentiate between dependent and independent events.	Ch. 10
LO3	Explain how events that are independent can be conditionally dependent (and vice versa), and identify real-world examples.	§10.7
LO4	Make a probabilistic inference in a real-world problem using Bayes' theorem to determine the probability of a hypothesis given evidence.	Ch. 9
LO5	Identify cases of the Binomial, Poisson and Normal distributions, compute related probabilities, and apply them to real problems.	Chs. 13, 14, 16
LO6	Apply the tools of probability to solve problems.	Throughout

The course also extends beyond these six outcomes in its later handouts. Chapter 17 covers combinations of normal variables, Chapter 18 the sampling distribution of the mean and the Central Limit Theorem, and Chapter 19 four further discrete distributions — the Uniform, Geometric, Negative Binomial and Hypergeometric. Chapter 20 is a set of revision exercises spanning the whole syllabus.

Types of Variables and Levels of Measurement

2.1 Why the type of a variable matters

Before you compute anything you must know what kind of variable you are holding, because **different statistical methods are appropriate for different types of variable**. Computing the mean of a set of postcodes produces a number, but that number is meaningless. Computing the median of a set of blood groups is not even defined. The taxonomy in this chapter is the gatekeeper for every technique in the rest of the course.

Definition: Variable

A variable is a characteristic of interest that we wish to study. It is observable, measurable or recordable, and it is the basic unit of analysis.

2.2 The two kinds of variable

Definition: Categorical (qualitative) variable

A variable having categories or classifications that are *not numerical in nature*.

Definition: Numerical (quantitative) variable

A variable whose values are *numerical in nature*.

Categorical examples	Numerical examples
Gender (Male, Female)	Weight, height, volume
Social class (High, Medium, Low)	Exam marks
Quality of computers (Good, Bad)	Number of defective parts in a computer
Hair colour	Age of a person
Ethnicity	Blood cholesterol level
Paid income tax last year? (Yes/No)	Amount of rainfall

Common pitfall: Numbers that are not numerical variables

A variable is not numerical merely because it is *recorded* using digits. Student ID numbers, jersey numbers, telephone numbers and coded categories ("1 = Male, 2 = Female") are all categorical. The test is whether arithmetic on the values means anything. The average of two ID numbers is not an ID number of anything; the average of two heights is a height.

2.3 Discrete versus continuous

Numerical variables split further.

Definition: Discrete variable

There is a *gap* between two possible values; the variable takes only countable or finite values.

Examples: the number of customers arriving at a supermarket; the number of defective parts in a computer.

Definition: Continuous variable

Theoretically there is *no gap* between two possible values; the variable takes an uncountable number of values, or any real value in an interval.

Examples: the amount of rainfall; the time taken to complete a computer job.

Intuition: The gap test

Pick any two values the variable can take and ask: is there always another possible value strictly between them? For *number of customers*, there is nothing between 4 and 5, so it is discrete. For *time*, between 4.1 and 4.2 seconds there is 4.15, and between 4.1 and 4.15 there is 4.125, forever — so it is continuous.

Measurement precision does not change this. Time recorded to the nearest second is *conceptually* continuous even though the recorded values happen to be whole numbers; the underlying quantity has no gaps. This distinction matters in Chapter 15, where a continuous variable has $P(X = x) = 0$ for every single value x — something that is never true of a discrete variable.

2.4 Levels of measurement within categorical data

Categorical variables split into two levels according to whether the categories have a meaningful order.

Definition: Nominal and ordinal

A **nominal** variable has categories with *no natural order* — they are names only. A **ordinal** variable has categories that *can be ranked*, although the gaps between ranks are not necessarily equal.

Blood group (O+, A−, B+, ...) is nominal: no group is “more” than another. Social class (High, Medium, Low) is ordinal: the order is real, but you cannot say High − Medium = Medium − Low.

2.5 The complete taxonomy

2.6 Which summaries are legal at which level

This table is the practical payoff of the taxonomy. Refer back to it while reading Chapters 3 and 4.

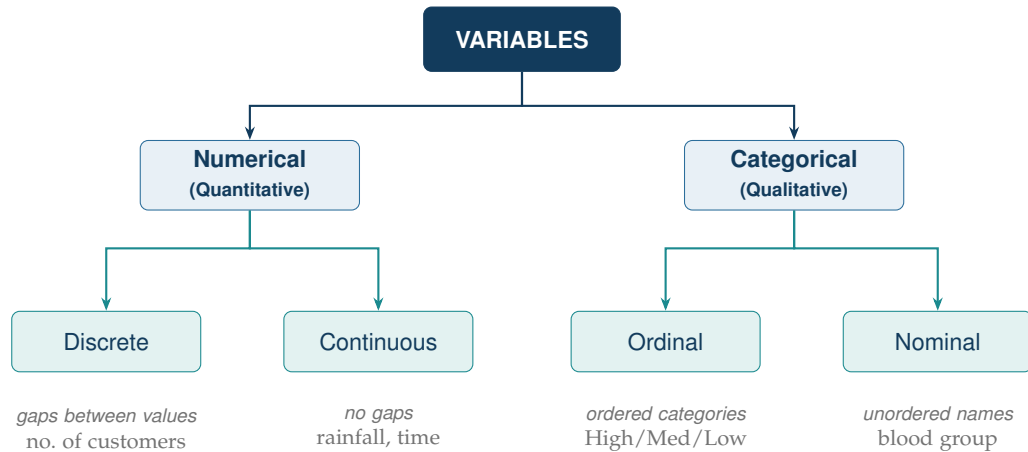


Figure 2.1: The classification of variables. Reading down this tree is the first step of any analysis, because it determines which summaries and which distributions are legitimate.

Summary	Nominal	Ordinal	Discrete	Continuous
Mode	✓	✓	✓	✓
Median, quartiles	–	✓	✓	✓
Mean	–	(–) [†]	✓	✓
Range, IQR	–	✓	✓	✓
Variance, standard dev.	–	–	✓	✓

[†] The mean of an ordinal variable is routinely computed in practice (e.g. averaging a 1–5 satisfaction scale) but is formally questionable, because it assumes the gaps between adjacent categories are equal.

The one summary that always works

The **mode** is the only measure of centre defined for *every* type of variable, because it requires only the ability to count occurrences — not to order or to add. That is why the handout stresses it is “the only measure used for both categorical and numerical data”.

Example 2.1 (Classifying a mixed dataset)

A hospital records the following for each patient. Classify each variable.

Variable	Type	Reasoning
Blood group	Categorical, nominal	Named groups, no order
Pain score (1–10)	Categorical, ordinal	Ranked, gaps not equal
Number of prior visits	Numerical, discrete	Countable, gaps between values
Body temperature (°C)	Numerical, continuous	Any real value in a range
Ward number	Categorical, nominal	Digits, but arithmetic meaningless
Length of stay (days, to the nearest day)	Numerical, continuous	Underlying quantity has no gaps

The last two rows are the ones that catch people out. *Ward number* looks numerical and is not; *length of stay* looks discrete because of how it is recorded, but the quantity being measured is time.

Measures of Location

3.1 What a measure of location does

A measure of location answers the question “*whereabouts* is this data?” with a single number. The handout groups them as follows.

Definition: Measures of location

Minimum The first observation of the ordered data set.

Maximum The last observation of the ordered data set.

Measure of centre

A quantity that locates a centre of the data set. The three common ones are the **mean**, the **median** and the **mode**.

Quartiles Quantities that divide the ordered data set into four equal parts.

Percentiles Quantities that divide the ordered data set into one hundred equal parts.

Notice that every one of these except the mean requires the data to be *ordered* first. Sorting is therefore step zero of almost every calculation in this chapter.

3.2 The mean

Definition: Arithmetic mean

The average value of the observations:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}.$$

The mean is the only measure of centre that uses *every* observation’s actual value. That is simultaneously its greatest strength (it wastes no information) and its fatal weakness: it is **very sensitive to extreme values**.

Intuition: The mean as a balance point

Imagine the observations as equal weights placed on a weightless ruler at their values. The mean is the point at which the ruler balances. Move one weight far to the right and the balance point must shift right to compensate — even though every other weight stayed where it was. That is exactly why a single outlier drags the mean.

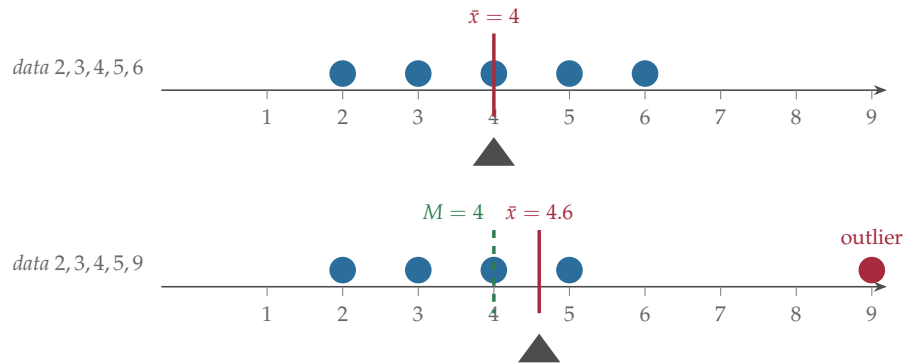


Figure 3.1: Replacing the observation 6 by 9 moves the mean from 4 to 4.6 but leaves the median at 4. The mean is not resistant; the median is.

3.2.1 The weighted mean

Sometimes we wish to give greater emphasis to some values than to others.

Definition: Weighted mean

With weights $w_1, w_2, \dots, w_n$ attached to observations $x_1, x_2, \dots, x_n$,

$$\bar{X}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}.$$

Setting every $w_i = 1$ recovers the ordinary mean, so the arithmetic mean is the special case of equal weights.

Example 3.1 (A weighted mean in practice)

A course is assessed by a quiz (weight 20%), a project (30%) and a final exam (50%). A student scores 80, 65 and 74 respectively. Her overall mark is

$$\bar{X}_w = \frac{20(80) + 30(65) + 50(74)}{20 + 30 + 50} = \frac{1600 + 1950 + 3700}{100} = \frac{7250}{100} = 72.5.$$

The unweighted mean would have been $(80 + 65 + 74)/3 = 73$. The weighted mean is lower because her weakest score carries the heaviest weight.

3.3 The median

Definition: Median M

The value that divides the ordered data set into two equal parts: half the observations are smaller and the other half are larger than M .

Procedure.

1. Arrange all observations in order, smallest to largest.

2. If the number of observations n is

- **odd**, the median is the centre value, the $\left(\frac{n+1}{2}\right)^{\text{th}}$ value;
- **even**, the median is the mean of the two middle values,

$$M = \frac{\left(\frac{n}{2}\right)^{\text{th}} \text{ value} + \left(\frac{n}{2} + 1\right)^{\text{th}} \text{ value}}{2}.$$

Common pitfall: $\frac{n+1}{2}$ is a *position*, not a value

For $n = 51$ the median is the 26th value in the ordered list — it is *not* 26. Students routinely write down the position and stop. Always finish by reading off the observation that sits at that position.

3.4 Mean versus median

This comparison is examined almost every year, so state it carefully.

Mean	Median
The <i>informal</i> midpoint. The data set may or may not be divided into two equal parts.	The <i>formal</i> midpoint. Exactly at the middle; the data set <i>is</i> divided into two equal parts.
Uses every observation's value.	Uses only the position of observations.
Not resistant to skew and outliers.	Resistant to skew and outliers.

The relationship between mean and median tells you the shape

Symmetric distribution	Mean = Median
Left- (negatively) skewed	Mean < Median the long tail pulls the mean down
Right- (positively) skewed	Median < Mean the long tail pulls the mean up

The mean chases the tail. That single sentence is enough to reconstruct all three lines.

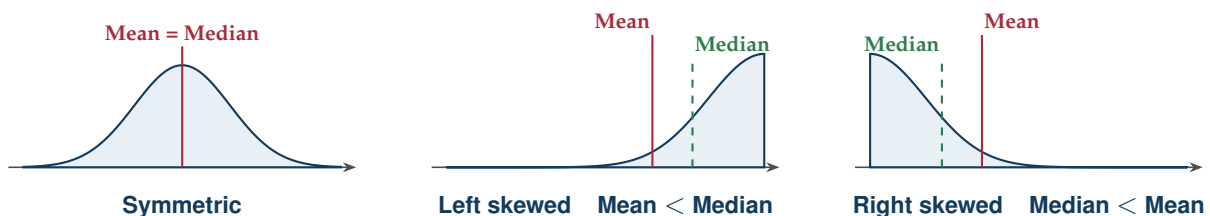


Figure 3.2: The mean is dragged towards the long tail; the median barely moves. In every case the mean sits between the median and the tail.

3.5 The mode

Definition: Mode

The most frequent observation.

Three facts about the mode that the handout emphasises:

- **There may be more than one mode.** In the table below, the values 30, 50 and 60 all occur 10 times, so the data set has three modes.

Value	20	30	40	50	60
Count	5	10	2	10	10

- **There may be no mode at all.** If every value occurs equally often, no value is “most frequent”.

Value	20	30	40	50	60
Count	5	5	5	5	5

- **It is the only measure usable on categorical data.**

Blood group	O+	O−	B+	B−	A+	A−
Count	5	10	12	2	11	8

Here the mode is B+. No mean or median exists.

The mode is used less frequently than the mean or the median for numerical data, because with continuous measurements no two observations may repeat at all.

3.6 Resistant measures

Definition: Resistant measure

A measure is **resistant** if it is not influenced by extreme values — that is, it does not respond strongly to changes in a few observations, no matter how large those changes are.

Which measures resist?

Not resistant	Mean, range, standard deviation, variance
Resistant	Median, mode, interquartile range

The pattern: measures built from *positions* in the ordered list resist; measures built from *values* do not.

3.7 Quartiles and percentiles

Definition: Quartiles

Quartiles divide the ordered data into four equal subsets:

$$Q_1 = \frac{1(n+1)}{4} \text{th value}, \quad Q_2 = \frac{2(n+1)}{4} \text{th value}, \quad Q_3 = \frac{3(n+1)}{4} \text{th value}.$$

Definition: Percentiles

Percentiles divide the ranked data into 100 equal subsets. The k th percentile P_k is the value such that $k\%$ of the data are smaller and $(100 - k)\%$ are larger.

Consequently

$$Q_1 = P_{25}, \quad \text{Median} = P_{50} = Q_2, \quad Q_3 = P_{75}.$$

3.7.1 The procedure the course uses for P_k

Different textbooks compute percentiles slightly differently. The procedure below is the one given in the handout, and it is the one to use in this course.

Recipe: Computing the k th percentile P_k

Step 1 Rank the n data values from smallest to largest.

Step 2 Calculate $\frac{nk}{100}$.

Step 3 If an integer a results: set $d(P_k) = a.5$.
If a number with a fraction results: set $d(P_k) = b$, the next larger integer.

Step 4 If $d(P_k) = a.5$: P_k is the average of the a th and $(a + 1)$ th values.
If $d(P_k) = b$: P_k is the value of the data in the b th position.

3.8 Worked example: exam marks of 50 students

This is the extended example from Handout 01, worked in full.

Example 3.2 (Quartiles and percentiles of 50 exam marks)

The exam marks of 50 students are listed below. Find the first quartile, the 58th percentile and the third quartile.

60	47	82	95	88	72	67	66	68	98
90	77	86	58	64	95	74	72	88	74
77	39	90	63	68	97	70	64	70	70
58	78	89	44	55	85	82	83	72	77
72	86	50	94	92	80	91	75	76	78

Solution.

Step 1 — Rank the data.

Here $n = 50$. Sorted from smallest to largest, with positions shown above the values:

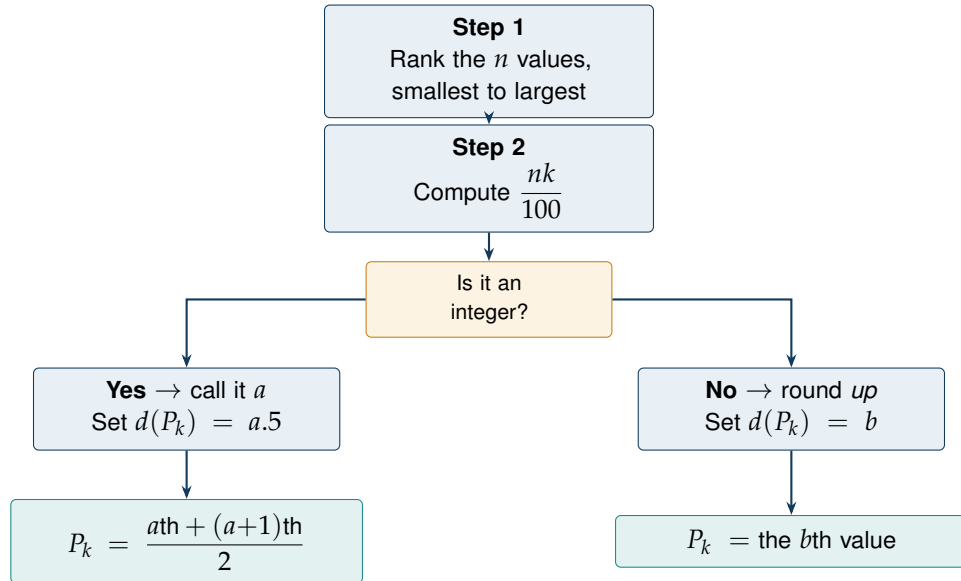


Figure 3.3: The percentile algorithm as a flow chart. Note that the “integer” branch *averages two values* while the “fraction” branch *reads one value*; mixing these up is the most common source of error.

1	2	3	4	5	6	7	8	9	10	11	12	13
39	44	47	50	55	58	58	60	63	64	64	66	67
14	15	16	17	18	19	20	21	22	23	24	25	26
68	68	70	70	70	72	72	72	72	74	74	75	76
27	28	29	30	31	32	33	34	35	36	37	38	39
77	77	77	78	78	80	82	82	83	85	86	86	88
40	41	42	43	44	45	46	47	48	49	50		
88	89	90	90	91	92	94	95	95	97	98		

Finding $Q_1 = P_{25}$.

$$\frac{nk}{100} = \frac{50 \times 25}{100} = 12.5.$$

This has a fraction, so we take the *next larger integer*: $d(Q_1) = 13$. Therefore Q_1 is the 13th value from the lowest, which is

$$Q_1 = 67.$$

Finding P_{58} .

$$\frac{nk}{100} = \frac{50 \times 58}{100} = 29.$$

This *is* an integer, $a = 29$, so $d(P_{58}) = 29.5$ and we average the 29th and 30th values:

$$P_{58} = \frac{77 + 78}{2} = 77.5.$$

Finding $Q_3 = P_{75}$.

$$\frac{nk}{100} = \frac{50 \times 75}{100} = 37.5,$$

a fraction, so $d(Q_3) = 38\dots$ but note the 37th and 38th values are both 86, so whichever convention is used,

$$Q_3 = 86.$$

Sanity check.

We must have $Q_1 \leq P_{50} \leq P_{58} \leq Q_3$. The median is P_{50} : $\frac{50 \times 50}{100} = 25$, an integer, so $M = \frac{75+76}{2} = 75.5$. Indeed $67 \leq 75.5 \leq 77.5 \leq 86$. ✓

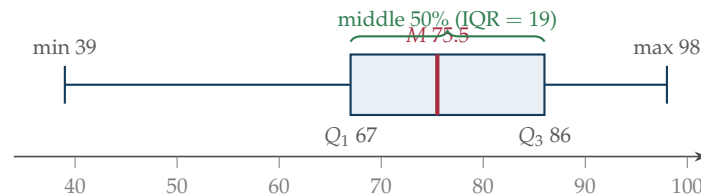
Summary of this dataset

Minimum	39	Median ($Q_2 = P_{50}$)	75.5
Q_1	67	P_{58}	77.5
Q_3	86	Maximum	98
Mean	74.92	Mode	72 (occurs 4 times)

The mean 74.92 sits just below the median 75.5, which suggests a very mild *left* skew — a few low marks (39, 44, 47) pulling the mean down. We quantify this in Section 4.6.

Example 3.3 (Reading a five-number summary)

The five-number summary — minimum, Q_1 , median, Q_3 , maximum — is the standard compact description of a distribution's location, and it is what a box plot draws. For the marks above:



The box holds the middle half of the students; the left whisker is longer than the right, which is the visual signature of the mild left skew noted above.

Measures of Dispersion

4.1 Why a centre alone is not enough

A measure of dispersion (equivalently: spread, or variability) measures **how far each value is away from the measure of centre**. Without one, a measure of centre can be actively misleading. The handout's example makes the point in three lines:

Data	Median	Spread
2, 5, 6, 8, 9	6	$9 - 2 = 7$
5, 5, 6, 6, 6	6	$6 - 5 = 1$
6, 6, 6, 6, 6	6	$6 - 6 = 0$

All three have median 6. Reported by centre alone they are identical; reported with spread they are obviously different data sets.

Definition: Requirements for a measure of variability

A sensible measure of variability should

- be a real number;
- equal **zero** if all the data are identical;
- **increase** as the data become more diverse;
- **never** be less than zero.

These four requirements are worth remembering, because they explain design choices later. In particular, they are why the variance squares the deviations: without squaring, the deviations from the mean always sum to zero (see Section 4.3.1) and the third requirement would fail.

4.2 Range and interquartile range

Definition: Range

$$R = \text{largest value} - \text{smallest value.}$$

- The simplest measure of dispersion.
- Gives an indication of the absolute spread of the distribution.
- Depends *only* on the two extreme values.
- Not very informative, and badly affected by outliers.

Definition: Interquartile range

$$\text{IQR} = Q_3 - Q_1.$$

- The range of the **middle 50%** of the values.
- The upper 25% and lower 25% of observations are discarded, so it is **not affected by outliers**.
- Not centred on the middle unless the distribution is symmetric.
- Often reported alongside the *median* — both are resistant, so they pair naturally.

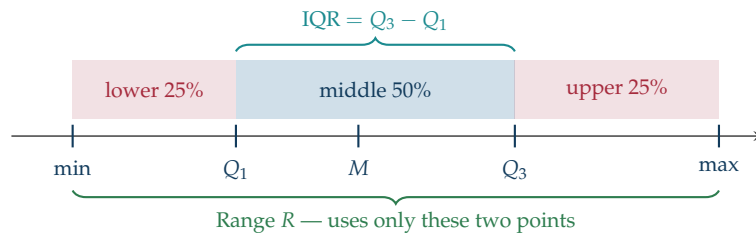


Figure 4.1: Range versus interquartile range. Move the maximum arbitrarily far right and R changes without limit while the IQR does not move at all.

4.3 Variance and standard deviation

Definition: Sample variance

Numerically, the variance equals the average of the squared deviations from the mean. For a sample of size n ,

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

The second form is the **computational formula** and is almost always faster by hand: you accumulate $\sum x_i$ and $\sum x_i^2$ in a single pass through the data, with no need to compute each deviation separately.

Definition: Standard deviation

The positive square root of the variance, $s = \sqrt{s^2}$.

The standard deviation is preferred for *reporting* because it is measured in the same units as the original data. If heights are in centimetres, the variance is in cm^2 — a quantity with no physical meaning — whereas the standard deviation is in centimetres.

4.3.1 Why we square the deviations

Intuition: The deviations always cancel

The obvious measure of spread would be the average deviation from the mean, $\frac{1}{n} \sum (x_i - \bar{x})$. But this is *always exactly zero*:

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0.$$

Positive and negative deviations cancel by construction, so this quantity carries no information at all. Squaring removes the signs before averaging, which is why the variance — and not the mean deviation — became the standard measure.

Theorem: Equivalence of the two variance formulas

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

Proof. Expand the square and use $\sum x_i = n\bar{x}$:

$$\begin{aligned} \sum (x_i - \bar{x})^2 &= \sum (x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\ &= \sum x_i^2 - 2\bar{x} \sum x_i + n\bar{x}^2 \\ &= \sum x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2 \\ &= \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2. \end{aligned} \quad \square$$

A data set that is highly variable will have a larger variance than a data set that is relatively homogeneous.

4.4 The effect of linear transformations

This is a small result with a large payoff: it saves recomputation whenever data are rescaled (Celsius to Fahrenheit, rupees to dollars, marks out of 50 to percentages).

Property: Shifting and scaling

Operation	Effect on the mean	Effect on the standard deviation
Add a constant c to each value	mean is increased by c	standard deviation is unaltered
Multiply each value by a constant k	mean is multiplied by k	standard deviation is multiplied by $ k $

Intuition: Why shifting does not change spread

Sliding every observation five units to the right slides the mean five units to the right as well, so every *deviation* $x_i - \bar{x}$ is unchanged. Since the variance is built entirely out of deviations, it cannot change. Stretching by k , however, stretches every deviation by k , so the variance — built from *squared* deviations — is multiplied by k^2 and the standard deviation by $|k|$.

Example 4.1 (*Applying the transformation rules*)

A set of marks has mean 58 and standard deviation 12. The lecturer rescales every mark by $y = 1.2x + 5$. Find the new mean and standard deviation.

Solution.

$$\bar{y} = 1.2(58) + 5 = 69.6, \quad s_y = 1.2 \times 12 = 14.4.$$

The +5 affects the mean but not the standard deviation. Note the variance becomes $1.2^2 \times 144 = 207.36$.

4.5 Worked example: dispersion of the 50 exam marks**Example 4.2** (*Full dispersion summary*)

Continuing the dataset of Section 3.8, we found $\min = 39$, $Q_1 = 67$, $M = 75.5$, $Q_3 = 86$, $\max = 98$.

Solution.

Range and IQR.

$$R = 98 - 39 = 59, \quad \text{IQR} = 86 - 67 = 19.$$

The range is more than three times the IQR, which already tells you that the extreme marks are far from the bulk of the class.

Mean.

The 50 marks sum to 3746, so

$$\bar{x} = \frac{3746}{50} = 74.92.$$

Variance and standard deviation.

Using the computational formula, $\sum x_i^2 = 290\,524$, so

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} = \frac{290\,524 - 50(74.92)^2}{49} = \frac{290\,524 - 280\,650.32}{49} = \frac{9873.68}{49} \approx 201.50,$$

$$s = \sqrt{201.50} \approx 14.20 \text{ marks.}$$

Interpretation.

A standard deviation of about 14 marks around a mean of about 75 says that a typical student sits roughly 14 marks away from the class average. Using the empirical rule anticipated in Chapter 16, roughly two thirds of the class should lie in $\bar{x} \pm s$, that is between about 60.7 and 89.1 — and counting the actual data gives 33 of the 50 students (66%) in that range, against the 68% the rule predicts. That is a close enough match to suggest the marks are roughly bell-shaped.

4.6 Skewness

Definition: Pearson's coefficient of skewness

$$\text{Pearson's coefficient of skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} \approx \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}}.$$

Generally skewness takes a value between -3 and $+3$. The sign is what matters most:

Sign	Shape	Meaning
Positive	Right- (positively) skewed	long tail to the right; mean $>$ median
Zero	Symmetric	mean = median
Negative	Left- (negatively) skewed	long tail to the left; mean $<$ median

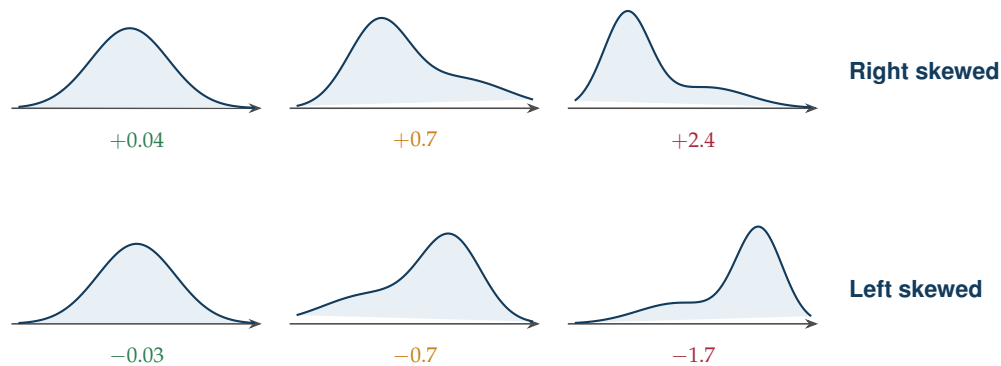


Figure 4.2: Distributions with the skewness coefficients quoted in the handout. The magnitude reflects how pronounced the tail is; the sign says which side the tail is on.

Example 4.3 (Skewness of the exam marks)

Using the values already computed — mean 74.92, median 75.5, mode 72, standard deviation 14.20:

Median version (the reliable one).

$$\frac{3(74.92 - 75.5)}{14.20} = \frac{3(-0.58)}{14.20} \approx -0.12.$$

Mode version.

$$\frac{74.92 - 72}{14.20} = \frac{2.92}{14.20} \approx +0.21.$$

Why they disagree.

The two formulas give opposite *signs* here. This is not an arithmetic error — it is a warning about the mode. The value 72 occurs four times and 70 and 77 each occur three times, so the mode is decided by a margin of a single observation and is extremely unstable. **When the mode is not clearly defined, always prefer the median version.** The correct reading is therefore a very mild left skew, consistent with the long left whisker seen in the box plot of Chapter 3.

Common pitfall: A coefficient near zero does not prove symmetry

Pearson's coefficient compares only the centre and the mode/median. A distribution can be perfectly balanced in that comparison and still be visibly asymmetric in its tails. Always look at a plot as well as a number.

Part I in one table

Question	Not resistant	Resistant	Pairs with
Where is the centre?	Mean $\bar{x}$	Median M , mode	—
How spread out?	Range, s , s^2	IQR	mean $\rightarrow s$; median $\rightarrow$ IQR
What shape?	Pearson's skewness; box plot; histogram		—

Report mean with standard deviation, or median with IQR. Mixing a non-resistant centre with a resistant spread (or vice versa) gives a summary that is internally inconsistent.

PART II

Probability

Sample Spaces and Events

5.1 The four objects you need before you can compute anything

Probability begins with a careful description of *what could happen*. The handout introduces four terms in quick succession; they are best read as a single chain, each one built from the one before.

Definition: The building blocks

Random experiment

An experiment in which there is some *uncertainty* about the outcome that will be observed.

Outcome

The result of an observation of the experiment.

Outcome space

(also called the *universal set* or *sample space*) The set of *all* possible outcomes of the experiment, denoted Ω or S .

Equally likely outcomes

If any outcome of a random experiment is as likely to occur as any other outcome, the outcomes are called equally likely.

Example 5.1 (The standard example)

Roll a die. This is a random experiment because the result is uncertain. The outcomes are 1, 2, 3, 4, 5, 6, and the outcome space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. If the die is properly constructed the six outcomes are equally likely.

Common pitfall: The outcome space is a choice, not a fact

The same physical experiment can be given different outcome spaces depending on what you decide to *observe*. Rolling two dice, you might record the ordered pair, giving $|\Omega| = 36$ equally likely outcomes; or record only the total, giving $\Omega = \{2, 3, \dots, 12\}$ with 11 outcomes that are *not* equally likely. Both are legitimate — but only the first lets you use the counting formula $P(A) = n(A)/n(\Omega)$. **Choose the outcome space in which the outcomes are equally likely whenever you can.** This idea reappears crucially in Exercise 6.3.

5.2 Events

Definition: Event and its relatives

Event A subset of the outcome space.

Equally likely events

If any event is as likely to occur as any other event, they are called equally likely events.

Mutually exclusive events

Two events A and B are mutually exclusive (or *disjoint*) if $A \cap B = \emptyset$.

Exhaustive events Two events A and B are exhaustive if $A \cup B = \Omega$, so that $P(A \cup B) = 1$.

Elementary event An event with just *one* outcome.

A note on the handout's wording

Handout 02 writes the exhaustive condition as " $P(A \cup B) = S$ and $P(A \cup B) = 1$ ". The first of these is a typographical slip: a probability is a number, not a set. The intended statement is

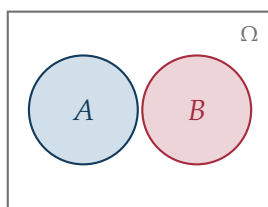
$$A \cup B = \Omega \quad \text{which implies} \quad P(A \cup B) = 1.$$

The set equation is the definition; the probability equation is its consequence.

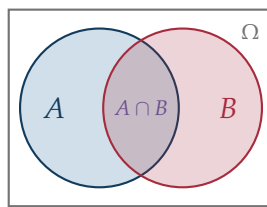
Intuition: Mutually exclusive and exhaustive are different ideas

Mutually exclusive says the events cannot both happen — they do not overlap. *Exhaustive* says at least one of them must happen — together they cover everything. A set of events that is both is called a **partition**, and partitions are precisely what the Law of Total Probability (Chapter 9) needs.

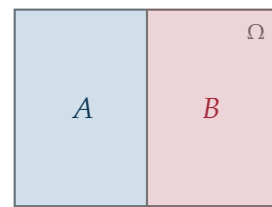
For a single die: $\{1, 2\}$ and $\{3, 4\}$ are mutually exclusive but not exhaustive. $\{1, 2, 3, 4\}$ and $\{3, 4, 5, 6\}$ are exhaustive but not mutually exclusive. $\{1, 2, 3\}$ and $\{4, 5, 6\}$ are both.



Mutually exclusive
 $A \cap B = \emptyset$
but not exhaustive



Exhaustive
 $A \cup B = \Omega$
but not mutually exclusive



Partition
both conditions hold
here $B = A'$

Figure 5.1: Mutually exclusive, exhaustive, and both. Only the third picture supports the Law of Total Probability.

5.3 Building outcome spaces: the worked cases

Most errors in early probability come from writing down the wrong outcome space, so it is worth being slow and explicit here.

Example 5.2 (A drawing pin dropped twice)

A drawing pin is dropped twice and it is observed, each time, whether the pin points upwards (u) or downwards (d). The possible outcomes are:

Outcome	Description
uu	pin points upwards on both occasions
dd	pin points downwards on both occasions
ud	upwards on the first drop, downwards on the second
du	downwards on the first drop, upwards on the second

So $\Omega = \{uu, ud, du, dd\}$, with $|\Omega| = 4$.

Important: ud and du are *different* outcomes. Recording the order matters. Note also that these four outcomes are *not* equally likely — a drawing pin is not symmetric, and the handout's later relative-frequency example finds it lands downwards about 72% of the time.

Example 5.3 (Tossing two and three coins)

Two coins. Writing h for heads and t for tails, and recording the result of each coin in order,

$$\Omega = \{hh, ht, th, tt\}, \quad |\Omega| = 2^2 = 4.$$

Three coins.

$$\Omega = \{hhh, hht, hth, thh, htt, tht, tth, ttt\}, \quad |\Omega| = 2^3 = 8.$$

The three-coin space appears repeatedly in this course — in Chapter 6 (probability), Chapter 8 (conditional probability), Chapter 10 (independence) and Chapter 11 (random variables) — so it is worth memorising in the order above: all-heads first, then the three outcomes with two heads, then the three with one head, then all-tails.

5.4 Set operations on events, in probability language

Because events are sets, the language of set theory carries over directly. The translation table below is used constantly and should become automatic.

Set notation	Probability language	Everyday phrasing
$A \cup B$	union	" A or B (or both) occurs"
$A \cap B$	intersection	" A and B both occur"
A'	complement	" A does not occur"
$A \subseteq B$	A implies B	"if A occurs then B occurs"
$A \cap B = \emptyset$	mutually exclusive	" A and B cannot both occur"
$\emptyset$	impossible event	"nothing happens" — probability 0
Ω	certain event	"something happens" — probability 1

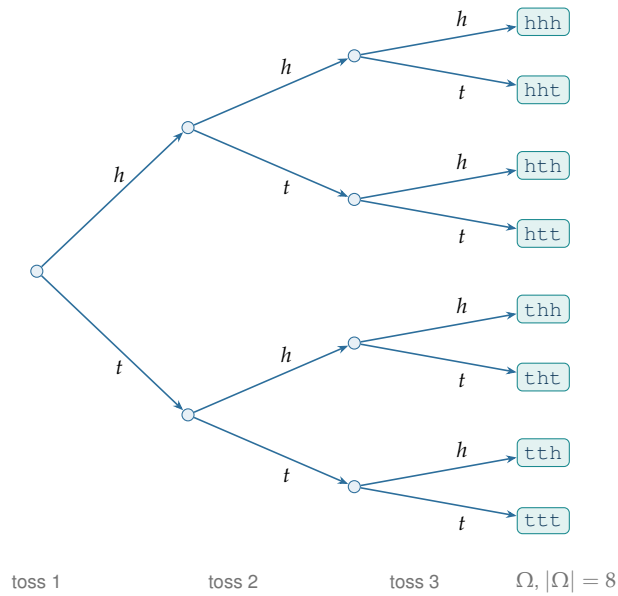


Figure 5.2: Enumerating the outcome space for three coin tosses with a tree. Each level doubles the number of branches, giving $2^3 = 8$ leaves. Every leaf is one outcome; any collection of leaves is an event.

De Morgan's laws — the most useful identities in the course

$$(A \cup B)' = A' \cap B', \quad (A \cap B)' = A' \cup B'.$$

In words: “not (A or B)” means “neither A nor B ”, and “not (A and B)” means “at least one of them fails”.

These convert an awkward “at least one” calculation into an easy “none” calculation. You will use this in Chapters 6, 10, 13 and 14 — for instance $P(\text{at least one head}) = 1 - P(\text{no heads})$.

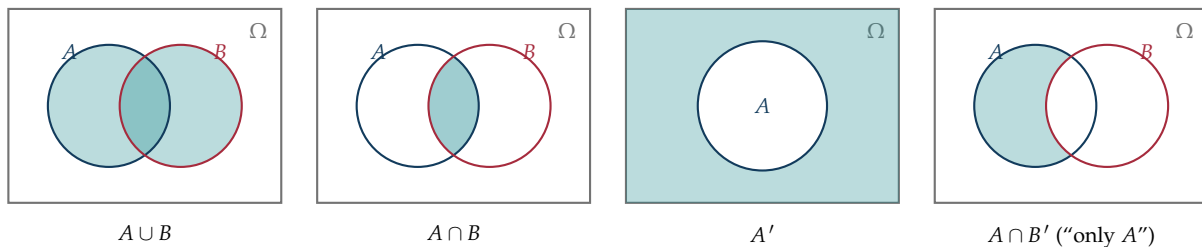


Figure 5.3: The four regions you will need constantly. The last one, “ A but not B ”, is the one students most often forget exists — it is what “only a solar hot-water unit” means in Exercise 6.6.

The Axioms of Probability

6.1 Probability as a function

The handout's framing here is more careful than most introductory treatments, and it repays attention.

Definition: The probability function

Probability is a *function*, which we denote by P . Like all functions, P has a domain, which we denote by $\mathcal{G}$. Since P assigns values to *events*, $\mathcal{G}$ is a *set of events*.

A complete specification of a probability model therefore requires three things:

1. the outcome space Ω corresponding to the random experiment;
2. the domain $\mathcal{G}$ of the probability function;
3. the probability function P itself.

Intuition: Why insist that probability is a function?

Saying “the probability of heads is $\frac{1}{2}$ ” hides the structure. What is really going on is that there is a machine, P , which takes a *set* of outcomes as its input and returns a *number* between 0 and 1 as its output:

$$P : \underbrace{\mathcal{G}}_{\text{sets of outcomes}} \longrightarrow \underbrace{[0, 1]}_{\text{numbers}} .$$

Once you see it this way, questions like “what is $P(A \cup B)$?” stop being mysterious. They are asking what the machine outputs when you feed it a particular set — and the axioms below are the rules the machine must obey.

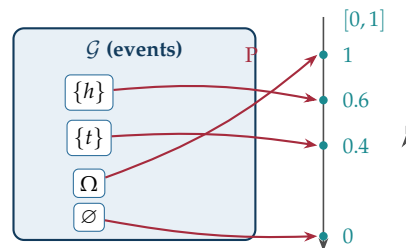


Figure 6.1: The probability function for a biased coin: it maps *sets* of outcomes to *numbers* in $[0, 1]$. Axiom 2 forces $\Omega \mapsto 1$; Property 3 forces $\emptyset \mapsto 0$.

6.2 The three axioms

Definition: Axioms of probability

A probability function P , defined on domain $\mathcal{G}$, assigns to each event $A \in \mathcal{G}$ a unique real number $P(A)$ in a way that satisfies:

Axiom 1 $P(A) \geq 0$ for any $A \in \mathcal{G}$. (non-negativity)

Axiom 2 $P(\Omega) = 1$. (normalisation)

Axiom 3 $P(A \cup B) = P(A) + P(B)$ whenever $A \cap B = \emptyset$, for any $A, B \in \mathcal{G}$. (finite additivity)

Everything else is a theorem

These three statements are the *entire* foundation. Every other formula in Part II — the complement rule, the addition rule, the Law of Total Probability, Bayes' theorem — is derived from them. Nothing else is assumed. That is what makes the theory trustworthy: if you accept these three sentences, you are committed to all the rest.

Note especially the condition on Axiom 3. Additivity holds **only** for mutually exclusive events. Forgetting that restriction is the single most common error in the whole subject.

6.3 Properties derived from the axioms

Property: Complement rule

$$P(A') = 1 - P(A).$$

Proof. Note that $A \cap A' = \emptyset$, so Axiom 3 applies:

$$\begin{aligned} P(A) + P(A') &= P(A \cup A') && \text{by Axiom 3} \\ &= P(\Omega) && \text{since } A \cup A' = \Omega \\ &= 1 && \text{by Axiom 2.} \end{aligned}$$

Hence $P(A') = 1 - P(A)$. □

Property: Addition rule / inclusion–exclusion

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

Proof. Decompose $A \cup B$ into three mutually exclusive pieces:

$$A \cup B = \underbrace{(A \cap B')}_{\text{only } A} \cup \underbrace{(A \cap B)}_{\text{both}} \cup \underbrace{(A' \cap B)}_{\text{only } B},$$

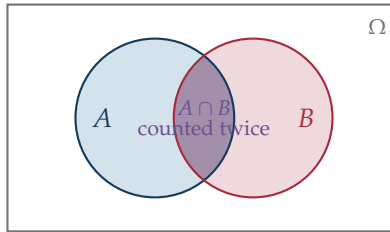
and note that $A = (A \cap B') \cup (A \cap B)$ and $B = (A' \cap B) \cup (A \cap B)$, each a disjoint union. By Axiom 3 applied three times,

$$\begin{aligned} P(A \cup B) &= P(A \cap B') + P(A \cap B) + P(A' \cap B) \\ &= [P(A) - P(A \cap B)] + P(A \cap B) + [P(B) - P(A \cap B)] \\ &= P(A) + P(B) - P(A \cap B). \end{aligned}$$

□

Intuition: Why subtract the overlap

Adding $P(A)$ and $P(B)$ counts every outcome in the overlap *twice* — once as part of A and once as part of B . Subtracting $P(A \cap B)$ removes the double count. If the events are mutually exclusive there is no overlap, $P(A \cap B) = 0$, and the rule collapses back to Axiom 3.



$P(A) + P(B)$ counts the purple region twice, so we subtract $P(A \cap B)$ once.

Figure 6.2: The addition rule, visually.

Property: Further consequences

From the axioms it also follows that

- (a) $P(\emptyset) = 0$;
- (b) $0 \leq P(A) \leq 1$ for every event A ;
- (c) if $A \subseteq B$ then $P(A) \leq P(B)$ (*monotonicity*).

Proof. (a) Take $A = \Omega$ in the complement rule: $P(\emptyset) = P(\Omega') = 1 - P(\Omega) = 1 - 1 = 0$.

(b) Axiom 1 gives $P(A) \geq 0$. Also $P(A') \geq 0$, and by the complement rule $P(A) = 1 - P(A') \leq 1$.

(c) If $A \subseteq B$ then $B = A \cup (B \cap A')$, a union of disjoint sets, so by Axiom 3, $P(B) = P(A) + P(B \cap A') \geq P(A)$ by Axiom 1. □

6.4 Assigning probabilities in practice

The axioms constrain probabilities but do not tell you what they *are*. Two methods are used to assign them.

6.4.1 Method 1: symmetry (the classical approach)

Definition: Symmetry / classical probability

Physical symmetry sometimes lets us assign probability values directly. If the outcome space consists of $n(\Omega)$ **equally likely** outcomes, then for any event A ,

$$P(A) = \frac{n(A)}{n(\Omega)} = \frac{\text{number of outcomes favourable to } A}{\text{total number of outcomes}}.$$

Example 6.1 (A uniform die)

Roll a uniform six-sided die and observe the number of dots. Here $\Omega = \{1, 2, 3, 4, 5, 6\}$ and, provided the die is properly constructed, each of the six elementary events is equally likely, each with probability $\frac{1}{6}$. Hence for $A = \{\text{even}\} = \{2, 4, 6\}$, $P(A) = 3/6 = 1/2$.

Example 6.2 (Tossing a fair coin twice)

$\Omega = \{hh, ht, th, tt\}$, so the probability of each elementary event is $\frac{1}{4}$. Thus $P(\text{exactly one head}) = P(\{ht, th\}) = 2/4 = 1/2$.

Common pitfall: The formula requires equal likelihood

$P(A) = n(A)/n(\Omega)$ is **not** a definition of probability — it is a shortcut that is valid only when every outcome in Ω is equally likely. Applied to $\Omega = \{RR, RB, BB\}$ for two marbles drawn from a bag, it gives $1/3$ for each, which is wrong. See Exercise 6.3.

6.4.2 Method 2: relative frequency

Definition: Relative frequency

Suppose a random experiment can be repeated and A is an event that can be observed. Let $f_r(A)$ denote the number of times A occurs in r repetitions — the *frequency of occurrence* of A in r repetitions. We often assign

$$P(A) \approx \frac{f_r(A)}{r} \quad \text{if } r \text{ is large.}$$

Example 6.3 (The drawing pin)

We drop a drawing pin 1000 times and observe that it comes to rest with the pin pointing downwards 721 times. We therefore assign

$$P(D) = \frac{721}{1000} = 0.721,$$

where D is the event “pin points downwards when dropped”.

This is the method of choice whenever symmetry is unavailable. A drawing pin has no symmetry that could tell us the answer in advance — we must *measure* it.

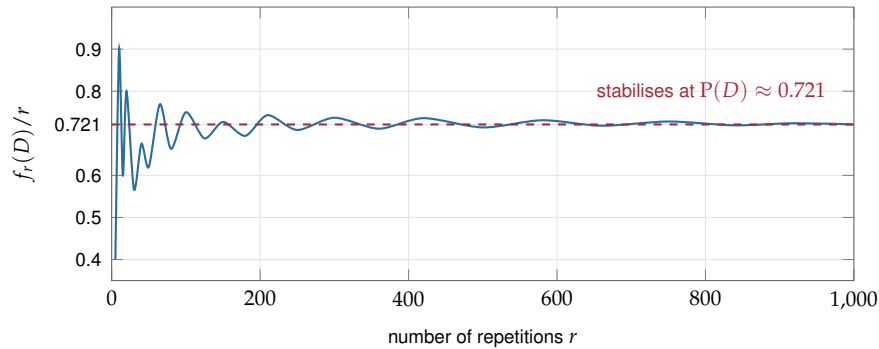


Figure 6.3: Relative frequency in action. Early on the ratio swings wildly; as r grows it settles down. This stabilisation is what justifies calling the limiting value a probability.

6.5 Exercises from Handout 02, worked in full

Exercise 6.1 (Completing a probability table — one coin)

A random experiment consists of tossing a coin and observing whether Heads or Tails results. The outcome space is $\Omega = \{h, t\}$. Complete the following table in a way consistent with the axioms of probability.

A	$\emptyset$	$\{h\}$	$\{t\}$	$\{h, t\}$
$P(A)$		0.4		

Solution. Work through the axioms in order.

$P(\emptyset)$.

By Property (a), $P(\emptyset) = 0$.

$P(\{h, t\})$.

$\{h, t\} = \Omega$, so by Axiom 2, $P(\Omega) = 1$.

$P(\{t\})$.

$\{h\}$ and $\{t\}$ are mutually exclusive and their union is Ω , so by Axiom 3 and Axiom 2, $P(\{h\}) + P(\{t\}) = 1$, giving $P(\{t\}) = 1 - 0.4 = 0.6$.

A	$\emptyset$	$\{h\}$	$\{t\}$	$\{h, t\}$
$P(A)$	0	0.4	0.6	1

Note the coin here is *biased* — nothing in the axioms requires $P(\{h\}) = \frac{1}{2}$.

Exercise 6.2 (Completing a probability table — one die)

A random experiment consists of rolling a six-sided die and observing the number of dots on the upper face. Complete the table consistently with the axioms.

A	$\emptyset$	$\{1, 2\}$	$\{3\}$	$\{1, 2, 3\}$	$\{4, 5, 6\}$	$\{3, 4, 5, 6\}$	$\{1, 2, 4, 5, 6\}$
$P(A)$				0.5			0.8

Solution.**The two given values.**

$$P(\{1, 2, 3\}) = 0.5 \text{ and } P(\{1, 2, 4, 5, 6\}) = 0.8.$$

$$P(\{4, 5, 6\}).$$

$$\{4, 5, 6\} = \{1, 2, 3\}', \text{ so } P(\{4, 5, 6\}) = 1 - 0.5 = 0.5.$$

$$P(\{3\}).$$

$$\{3\} = \{1, 2, 4, 5, 6\}', \text{ so } P(\{3\}) = 1 - 0.8 = 0.2.$$

$$P(\{1, 2\}).$$

$$\{1, 2\} \text{ and } \{3\} \text{ are disjoint with union } \{1, 2, 3\}, \text{ so by Axiom 3, } P(\{1, 2\}) = 0.5 - 0.2 = 0.3.$$

$$P(\{3, 4, 5, 6\}).$$

$$\{3\} \text{ and } \{4, 5, 6\} \text{ are disjoint, so } P(\{3, 4, 5, 6\}) = 0.2 + 0.5 = 0.7. \text{ (Check: } \{3, 4, 5, 6\} = \{1, 2\}' \text{ gives } 1 - 0.3 = 0.7. \checkmark)$$

A	$\emptyset$	$\{1, 2\}$	$\{3\}$	$\{1, 2, 3\}$	$\{4, 5, 6\}$	$\{3, 4, 5, 6\}$	$\{1, 2, 4, 5, 6\}$
P(A)	0	0.3	0.2	0.5	0.5	0.7	0.8

Again this die is *not* fair: $P(\{3\}) = 0.2 \neq \frac{1}{6}$.

Exercise 6.3 (Two marbles from a bag)

Two marbles are drawn simultaneously from a bag containing three red marbles and two blue marbles, and their colours are observed. Apart from colour the marbles are indistinguishable.

- Specify the outcome space.
- Find the probability of getting 1 red and 1 blue marble.
- Find the probability of getting at least 1 blue marble.

Solution.**(a) The outcome space as observed.**

Since we observe only colours,

$$\Omega = \{RR, RB, BB\}.$$

These three outcomes are not equally likely, so we cannot count in this space.

The extended outcome space.

Label the marbles R_1, R_2, R_3, B_1, B_2 and record which *pair* was drawn. The $\binom{5}{2} = 10$ unordered pairs are:

$$\Omega' = \{R_1R_2, R_1R_3, R_2R_3, R_1B_1, R_2B_1, R_3B_1, R_1B_2, R_2B_2, R_3B_2, B_1B_2\}.$$

Because the marbles are physically identical apart from colour, all 10 pairs *are* equally likely, and now the counting formula applies.

(b) One red and one blue.

The mixed pairs are $R_1B_1, R_2B_1, R_3B_1, R_1B_2, R_2B_2, R_3B_2$ — that is $3 \times 2 = 6$ of them. So

$$P(1 \text{ red and } 1 \text{ blue}) = \frac{6}{10} = \frac{3}{5} = 0.6.$$

(c) At least one blue.

Use the complement: “at least one blue” fails only when both are red, and there are $\binom{3}{2} = 3$ all-red pairs. So

$$P(\text{at least 1 blue}) = 1 - \frac{3}{10} = \frac{7}{10} = 0.7.$$

Common pitfall: A correction to Handout 02

The handout’s answer to part (b) is stated as $5/10$. Counting the mixed pairs in the handout’s own enumerated list gives six of them ($R_1B_1, R_2B_1, R_3B_1, R_1B_2, R_2B_2, R_3B_2$), so the answer should be

$$\frac{6}{10} = \frac{3}{5}.$$

A cross-check confirms this: the three probabilities must sum to 1, and

$$P(RR) = \frac{3}{10}, \quad P(RB) = \frac{6}{10}, \quad P(BB) = \frac{1}{10}, \quad \frac{3+6+1}{10} = 1. \quad \checkmark$$

With $5/10$ the three would sum to $9/10$, violating Axiom 2. The handout’s part (c) answer of $7/10$ is correct and is also consistent with $6/10$, since $6/10 + 1/10 = 7/10$.

Exercise 6.4 (Drawing a seven)

A card is drawn at random from an ordinary pack of 52 playing cards. Find the probability that the card is (a) a seven, (b) not a seven.

Solution. There are four sevens (one per suit), so

$$(a) \quad P(\text{seven}) = \frac{4}{52} = \frac{1}{13} \approx 0.0769.$$

$$(b) \quad P(\text{not a seven}) = 1 - \frac{4}{52} = \frac{48}{52} = \frac{12}{13} \approx 0.923.$$

Part (b) is the complement rule in its simplest form.

Exercise 6.5 (A coin and a die together)

A coin and a die are thrown together. Draw a sample space diagram and find the probability of obtaining (a) a head; (b) a number greater than 4; (c) a head *and* a number greater than 4; (d) a head *or* a number greater than 4.

Solution.

The sample space diagram.

There are $2 \times 6 = 12$ equally likely outcomes.

die→	1	2	3	4	5	6	
h	h1	h2	h3	h4	h5	h6	A: head
t	t1	t2	t3	t4	t5	t6	B: die > 4

(a) A head.

The head row has 6 of the 12 cells: $P(A) = \frac{6}{12} = \frac{1}{2}$.

(b) A number greater than 4.

The columns 5 and 6 contain 4 cells: $P(B) = \frac{4}{12} = \frac{1}{3}$.

(c) A head *and* a number greater than 4.

The purple intersection has 2 cells: $P(A \cap B) = \frac{2}{12} = \frac{1}{6}$.

(d) A head *or* a number greater than 4.

By the addition rule,

$$P(A \cup B) = \frac{6}{12} + \frac{4}{12} - \frac{2}{12} = \frac{8}{12} = \frac{2}{3}.$$

Counting the shaded cells directly also gives 8. ✓

Notice in passing that $P(A \cap B) = \frac{1}{6} = \frac{1}{2} \times \frac{1}{3} = P(A)P(B)$, so A and B are *independent* — as they must be, since the coin cannot influence the die. This is the subject of Chapter 10.

Exercise 6.6 (A survey of country homes)

In a survey of 1500 country homes there were 1200 which had an air-conditioning unit, 900 which had solar hot water, while 700 had both. A country home is selected at random from the same area, and it is observed whether it has an air-conditioning unit and whether it has a solar hot-water unit. Give a suitable outcome space and find the probability of having (a) only solar hot water; (b) a solar hot-water unit *or* an air-conditioning unit.

Solution.**A suitable outcome space.**

Record a pair (has AC?, has solar?), giving four outcomes:

$$\Omega = \{(AC, S), (AC, S'), (AC', S), (AC', S')\}.$$

Let A = “has air conditioning” and S = “has solar hot water”. Then

$$P(A) = \frac{1200}{1500} = 0.8, \quad P(S) = \frac{900}{1500} = 0.6, \quad P(A \cap S) = \frac{700}{1500} = \frac{7}{15} \approx 0.4\bar{6}.$$

A two-way table — the reliable way to do these.

Fill in the “both” cell first, then subtract along rows and columns.

	Solar	No solar	Total
AC	700	500	1200
No AC	200	100	300
Total	900	600	1500

Every interior cell follows from the three given numbers: $1200 - 700 = 500$, $900 - 700 = 200$, and then $1500 - 700 - 500 - 200 = 100$.

(a) Only solar hot water.

This is $A' \cap S$, the “No AC / Solar” cell:

$$P(A' \cap S) = \frac{200}{1500} = \frac{2}{15} \approx 0.133.$$

(b) Solar or air conditioning.

By the addition rule,

$$P(A \cup S) = \frac{1200}{1500} + \frac{900}{1500} - \frac{700}{1500} = \frac{1400}{1500} = \frac{14}{15} \approx 0.933.$$

Equivalently, only 100 of the 1500 homes have neither, so $1 - \frac{100}{1500} = \frac{14}{15}$. ✓

The two-way table is worth the 30 seconds

Whenever a problem gives you “ n have A , m have B , k have both”, draw the 2×2 table before doing anything else. Every question that follows — only A , only B , neither, at least one — is then a single cell or a sum of cells, and it is essentially impossible to double-count.

Exercise 6.7 (Three questions about one card)

A card is drawn at random from a pack of cards. Find the probability that (1) a heart card is drawn; (2) a card which is *not* a face card is drawn; (3) the number on the drawn card is divisible by 3.

Solution.**(1) A heart.**

There are 13 hearts: $P = \frac{13}{52} = \frac{1}{4}$.

(2) Not a face card.

The face cards are Jack, Queen and King in each of the four suits, so there are $3 \times 4 = 12$ of them. Hence

$$P(\text{not a face card}) = 1 - \frac{12}{52} = \frac{40}{52} = \frac{10}{13} \approx 0.769.$$

(3) Number divisible by 3.

Among the numbered cards, those divisible by 3 are the 3s, 6s and 9s — three ranks, four suits each:

$$P = \frac{3 \times 4}{52} = \frac{12}{52} = \frac{3}{13} \approx 0.231.$$

Counting Techniques

7.1 Why counting deserves a chapter

The classical formula $P(A) = n(A)/n(\Omega)$ reduces every equally-likely probability question to a pair of counting problems. When the outcome space is small you can list it, as in Chapter 5. When it is not — and $\binom{52}{5} = 2\,598\,960$ poker hands is not small — you need systematic counting. This chapter supplies the tools, and they are needed again in Chapter 13 where the $\binom{n}{k}$ in the Binomial distribution comes from exactly this reasoning.

7.2 The multiplication principle

Definition: Multiplication principle

If a procedure consists of k stages, and stage i can be performed in n_i ways *regardless of how the earlier stages were performed*, then the whole procedure can be performed in

$$n_1 \times n_2 \times \cdots \times n_k$$

ways.

Example 7.1 (Counting without listing)

Three coins tossed: $2 \times 2 \times 2 = 8$ outcomes, matching the tree of Figure 5.2. A coin and a die: $2 \times 6 = 12$ outcomes, matching the grid in Exercise 6.5. A four-character password using digits with repetition allowed: $10^4 = 10\,000$.

Intuition: The tree picture

The multiplication principle is just the observation that a tree with n_1 branches at the first level, each splitting into n_2 , each into n_3 , has $n_1 n_2 n_3$ leaves. Every counting formula below is a consequence of it.

7.3 Permutations: order matters

Definition: Permutations

The number of ways of arranging r objects chosen from n distinct objects, *where the order of selection matters*, is

$${}^n P_r = \frac{n!}{(n-r)!} = \underbrace{n(n-1)(n-2) \cdots (n-r+1)}_{r \text{ factors}}.$$

In particular ${}^nP_n = n!$ is the number of ways of arranging all n objects.

The formula is the multiplication principle in disguise: there are n choices for the first position, $n - 1$ for the second (one object having been used up), and so on down to $n - r + 1$ for the r th.

Example 7.2 (*Prizes awarded in order*)

In how many ways can gold, silver and bronze medals be awarded among 8 athletes?

$${}^8P_3 = 8 \times 7 \times 6 = 336.$$

Order matters here because “A gold, B silver” differs from “B gold, A silver”.

7.4 Combinations: order does not matter

Definition: Combinations

The number of ways of choosing r objects from n distinct objects *without regard to order* is

$$\binom{n}{r} = {}^nC_r = \frac{n!}{r!(n-r)!}.$$

Theorem: Relationship between permutations and combinations

$${}^nP_r = \binom{n}{r} \times r! \quad \text{equivalently} \quad \binom{n}{r} = \frac{{}^nP_r}{r!}.$$

Proof. Every unordered selection of r objects can be arranged internally in $r!$ different orders, and each of those arrangements is a distinct permutation. Ordered selections therefore outnumber unordered ones by exactly a factor of $r!$. $\square$

How to decide: order or no order?

Ask: *if I swap two of the chosen objects, do I get a different result?*

Situation	Order?	Use
Medals, rankings, seat assignments, PINs	Yes	nP_r
Committees, hands of cards, marbles drawn together	No	$\binom{n}{r}$

The marble problem of Exercise 6.3 says the marbles are drawn “simultaneously”, which is the standard signal that order does *not* matter — hence $\binom{5}{2} = 10$.

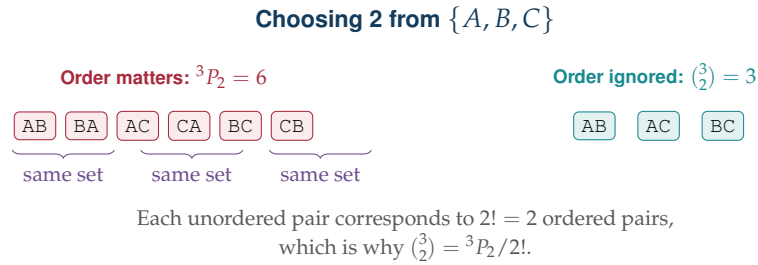


Figure 7.1: Permutations versus combinations. Grouping the 6 ordered pairs into sets of $2! = 2$ leaves 3 unordered selections.

7.5 Useful identities

Property: Properties of binomial coefficients

- (a) $\binom{n}{0} = \binom{n}{n} = 1$;
- (b) $\binom{n}{r} = \binom{n}{n-r}$ (*symmetry*);
- (c) $\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r}$ (*Pascal's rule*);
- (d) $\sum_{r=0}^n \binom{n}{r} = 2^n$.

Proof of (b) and (d), by counting arguments. (b) Choosing which r objects to *take* is the same act as choosing which $n - r$ objects to *leave behind*; the two counts must therefore agree.

(d) The left-hand side counts all subsets of an n -element set, grouped by size. The right-hand side counts them by deciding independently, for each of the n elements, whether it is in or out — 2 choices, n times. $\square$

Why (d) matters in Chapter 13

Identity (d) is the statement that the Binomial probabilities sum to 1 when $p = \frac{1}{2}$:

$$\sum_{k=0}^n \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k} = \frac{1}{2^n} \sum_{k=0}^n \binom{n}{k} = \frac{2^n}{2^n} = 1.$$

The general version, $\sum_k \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1$, is the binomial theorem — and that is precisely where the Binomial distribution gets its name.

7.6 Worked counting problems

Example 7.3 (Marbles, revisited by formula)

The bag of Exercise 6.3 has 3 red and 2 blue marbles, and two are drawn simultaneously.

Event	Count	Probability
Total ways	$\binom{5}{2} = 10$	—
Both red	$\binom{3}{2} = 3$	$3/10$
One red, one blue	$\binom{3}{1}\binom{2}{1} = 3 \times 2 = 6$	$6/10 = 3/5$
Both blue	$\binom{2}{2} = 1$	$1/10$
Sum	$3 + 6 + 1 = 10 \checkmark$	$1 \checkmark$

The pattern $\binom{3}{1}\binom{2}{1}$ — choose the reds, then choose the blues, then multiply — is the multiplication principle applied to combinations, and it generalises immediately to any number of colours.

Example 7.4 (Committee selection)

A committee of 4 is to be chosen from 7 women and 5 men.

- How many committees are possible?
- How many contain exactly 2 women?
- What is the probability a randomly chosen committee has exactly 2 women?

Solution.

$$(a) \quad \binom{12}{4} = \frac{12 \times 11 \times 10 \times 9}{4 \times 3 \times 2 \times 1} = 495.$$

$$(b) \quad \binom{7}{2}\binom{5}{2} = 21 \times 10 = 210.$$

$$(c) \quad P = \frac{210}{495} = \frac{14}{33} \approx 0.424.$$

Example 7.5 (Sampling with and without replacement)

A bag contains 5 red and 7 black counters; two are drawn. Compare the two sampling schemes.

	With replacement	Without replacement
P(both red)	$\frac{5}{12} \cdot \frac{5}{12} = \frac{25}{144} \approx 0.174$	$\frac{5}{12} \cdot \frac{4}{11} = \frac{20}{132} \approx 0.152$
Draws are	independent	dependent
Population changes?	no	yes

The distinction drives everything in Chapters 8–10: *with* replacement gives independent events (Chapter 10), *without* replacement gives conditional probabilities that genuinely change (Chapter 8).

Common pitfall: Do not mix ordered and unordered counts

If you count the numerator with order (permutations) you must count the denominator with order too. Mixing the two is the most common counting error. Both approaches give the same probability when applied consistently: for “both red” without replacement,

$$\text{unordered: } \frac{\binom{5}{2}}{\binom{12}{2}} = \frac{10}{66} = \frac{5}{33}, \quad \text{ordered: } \frac{5 \times 4}{12 \times 11} = \frac{20}{132} = \frac{5}{33} \checkmark$$

Conditional Probability

8.1 The idea: shrinking the sample space

Conditional probability answers the question: *given that I already know B happened, how likely is A?*

Definition: Conditional probability

If A and B are two events with $P(A) \neq 0$ and $P(B) \neq 0$, then the probability of A given that B has already occurred is written $P(A | B)$, read “probability of A , given B ”, and is defined as

$$P(A | B) = \frac{P(A \cap B)}{P(B)} \quad \text{for } P(B) > 0.$$

Similarly,

$$P(B | A) = \frac{P(A \cap B)}{P(A)} \quad \text{for } P(A) > 0.$$

Intuition: Why divide by $P(B)$?

Learning that B occurred eliminates every outcome outside B . The sample space **shrinks** from Ω to B . Within this new, smaller universe, the only part of A that survives is $A \cap B$. But probabilities must still add to 1 over the new universe, and $P(B)$ is generally less than 1. Dividing by $P(B)$ *renormalises*: it rescales the surviving probabilities so that they sum to 1 again. That is the entire content of the formula.

To confirm the renormalisation works, check that $P(B | B) = 1$:

$$P(B | B) = \frac{P(B \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1,$$

which is exactly right — given that B occurred, we are certain B occurred.

8.2 The multiplication rule

Rearranging the definition gives one of the most useful formulas in the course.

Theorem: Multiplication rule

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

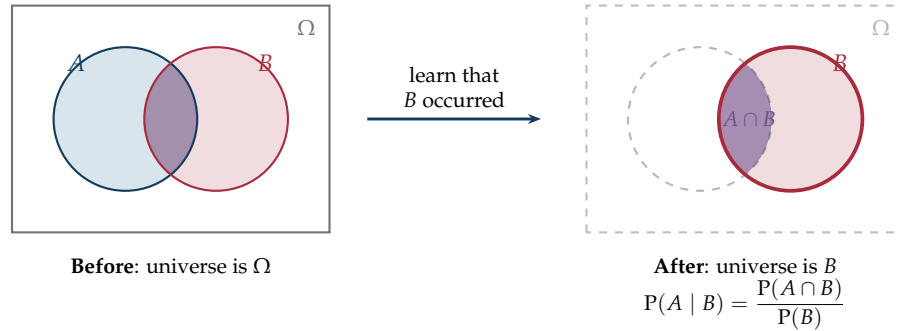


Figure 8.1: Conditioning shrinks the sample space. The purple region is unchanged in absolute terms — what changes is the *denominator* it is measured against.

Read it as a sequence of stages

$P(A \cap B) = P(A) \times P(B | A)$ says: *the chance that both happen is the chance the first happens, times the chance the second happens given the first already did.* That is exactly how you multiply along a branch of a probability tree — trees and this formula are the same idea drawn two ways.

The rule extends to more events by chaining:

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

8.3 First examples

Example 8.1 (Two quick conditionals)

1. Given that a heart is picked at random from a pack of 52 playing cards, find the probability that it is a picture card.
2. When a die is thrown, an odd number occurs. What is the probability that the number is prime?

Solution.

1. Picture card given heart.

Let A = “picture card” and B = “heart card”. We require $P(\text{picture card} | \text{heart})$. There are 3 picture cards in hearts (J, Q, K) and 13 hearts, so

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{3/52}{13/52} = \frac{3}{13} \approx 0.231.$$

The 52s cancel: once you know it is a heart, you are simply counting 3 picture cards out of 13 hearts.

2. Prime given odd.

Let A = “prime number” and B = “odd number”. On a die, $A = \{2, 3, 5\}$ and $B = \{1, 3, 5\}$, so $A \cap B = \{3, 5\}$. Hence

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{2/6}{3/6} = \frac{2}{3} \approx 0.667.$$

Note on the number 1.

1 is *not* prime, and 2 is prime but not odd — both facts matter here. Getting either wrong changes the answer.

8.4 Sampling without replacement

Conditional probability is the natural language for drawing objects without putting them back, because each draw changes what is left.

Example 8.2 (*Counters drawn without replacement*)

A bag contains 10 counters, of which 7 are green and 3 are white. A counter is picked at random from the bag and its colour is noted. **The counter is not replaced.** A second counter is then picked out. Find the probability that

- (a) the first counter is green;
- (b) the first counter is green and the second counter is white;
- (c) the counters are of different colours.

Solution. Write G_1 for “first counter is green”, W_2 for “second counter is white”, and so on.

(a) The first counter is green.

$$P(G_1) = \frac{7}{10}.$$

(b) First green and second white.

After a green counter has been removed there are 9 counters left, of which 3 are still white. So $P(W_2 | G_1) = \frac{3}{9} = \frac{1}{3}$, and by the multiplication rule

$$P(W_2 \cap G_1) = P(W_2 | G_1) P(G_1) = \frac{1}{3} \times \frac{7}{10} = \frac{7}{30}.$$

(c) The counters are of different colours.

Different colours means *green then white* **or** *white then green*. These are mutually exclusive, so we add them. We already have $P(W_2 \cap G_1) = \frac{7}{30}$. For the other order,

$$P(W_1) = \frac{3}{10}, \quad P(G_2 | W_1) = \frac{7}{9},$$

$$P(G_2 \cap W_1) = \frac{7}{9} \times \frac{3}{10} = \frac{7}{30}.$$

Therefore

$$P(\text{different colours}) = \frac{7}{30} + \frac{7}{30} = \frac{14}{30} = \frac{7}{15} \approx 0.467.$$

Recipe: Using a probability tree

1. Draw one level of branching per stage of the experiment.
2. Label each branch with a *conditional* probability, given everything on the path so far.

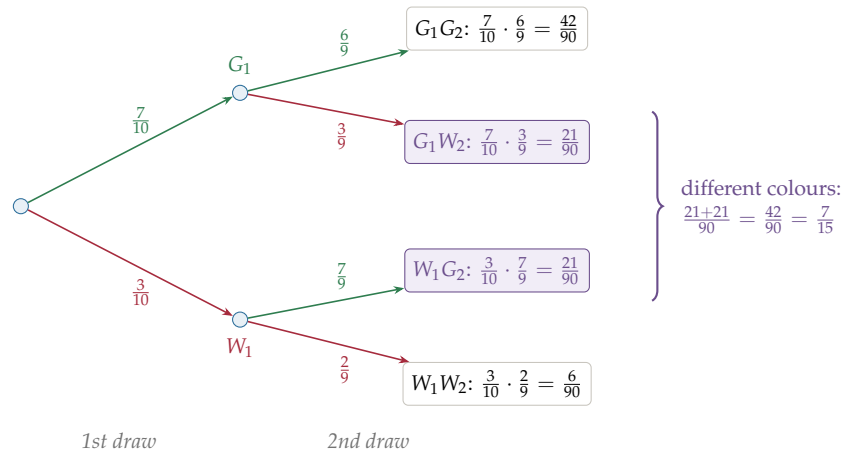


Figure 8.2: A probability tree for two draws without replacement. Note the *second-level* probabilities have denominator 9, not 10 — the bag has changed. **Multiply along a branch; add across branches.** The four leaf probabilities sum to $\frac{42+21+21+6}{90} = 1$. ✓

3. **Multiply** along a path to get the probability of that complete outcome.
4. **Add** the paths that satisfy your event.
5. **Check:** all leaf probabilities must sum to 1.

Step 5 is not optional. It catches almost every arithmetic slip.

8.5 A three-coin exercise

Exercise 8.1 (Conditional probability with three tosses)

A random experiment consists of tossing a fair coin three times and observing the sequence of Heads and Tails. Define

$$A = \text{'first toss is a Head'}, \quad B = \text{'at least two Heads'}.$$

Find (a) $P(A)$; (b) $P(B)$; (c) $P(A \text{ and } B)$; (d) the conditional probability that the first toss is a Head, given that the outcome includes at least two Heads.

Solution.

The outcome space.

$$\Omega = \{hhh, hht, hth, thh, htt, tht, tth, ttt\}, \quad |\Omega| = 8,$$

with all eight outcomes equally likely because the coin is fair.

The two events.

$$\begin{aligned} A &= \{hhh, hht, hth, htt\} && \text{(the four outcomes starting with } h) \\ B &= \{hhh, hht, hth, thh\} && \text{(one with three heads, three with exactly two)} \end{aligned}$$

$$A \cap B = \{hhh, hht, hth\}.$$

(a)–(c).

$$P(A) = \frac{4}{8} = \frac{1}{2}, \quad P(B) = \frac{4}{8} = \frac{1}{2}, \quad P(A \cap B) = \frac{3}{8}.$$

(d) The conditional probability.

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{3/8}{4/8} = \frac{3}{4}.$$

Interpretation — and a preview of Chapter 10.

Unconditionally the first toss is a Head with probability $\frac{1}{2}$. Told that at least two Heads occurred, that rises to $\frac{3}{4}$. The information is relevant, so A and B are **not independent**:

$$P(A \cap B) = \frac{3}{8} \neq \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}.$$

This is worth pausing on. The individual *tosses* are independent of one another, but the *events* A and B are not — because B is defined using all three tosses, including the first. Independence is a property of events, not of the physical mechanism.

Common pitfall: $P(A | B)$ and $P(B | A)$ are different numbers

In the exercise above the two happen to coincide, because $P(A) = P(B)$. That is a coincidence, not a rule. In general

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

which is Bayes' theorem (Chapter 9). Confusing the two directions is the most consequential error in applied probability — see the medical-test example in Section 9.5, where $P(\text{positive} | \text{disease}) = 0.99$ but $P(\text{disease} | \text{positive}) = 0.165$.

Total Probability and Bayes' Theorem

9.1 The two-case version of the Law of Total Probability

Start with the simplest possible partition: an event B and its complement B' .

Theorem: Total probability, two cases

Let A and B be events. Observe that

$$B \cap B' = \emptyset \quad \text{and} \quad B \cup B' = \Omega.$$

Then

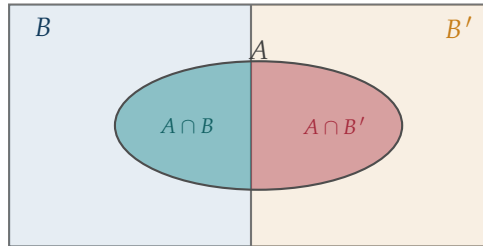
$$A = (A \cap B) \cup (A \cap B'), \quad \text{with} \quad (A \cap B) \cap (A \cap B') = \emptyset,$$

so by Axiom 3,

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

Applying the definition of conditional probability to each term,

$$P(A) = P(A | B) P(B) + P(A | B') P(B').$$



The event A is cut cleanly into two disjoint pieces by the partition $\{B, B'\}$, so
 $P(A) = P(A \cap B) + P(A \cap B') = P(A | B) P(B) + P(A | B') P(B')$.

Figure 9.1: The two-case Law of Total Probability.

9.2 The general Law of Total Probability

Theorem: Law of Total Probability

Suppose $B_1, B_2, B_3, \dots, B_n$ are **mutually exclusive** and **exhaustive** events, and A is an arbitrary event. Then

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

that is,

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_n)P(B_n).$$

Proof. Since $B_1, \dots, B_n$ are mutually exclusive and exhaustive, it follows that $A \cap B_1, A \cap B_2, \dots, A \cap B_n$ are mutually exclusive, and

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

Therefore, with repeated use of Axiom 3,

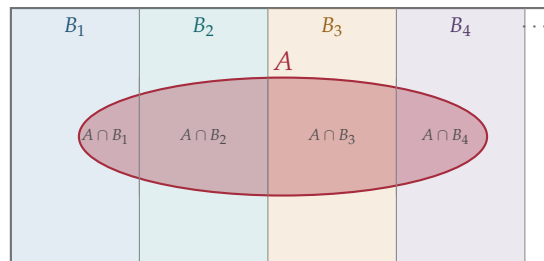
$$P(A) = \sum_{i=1}^n P(A \cap B_i) = \sum_{i=1}^n P(A \mid B_i)P(B_i),$$

the last step being the multiplication rule applied term by term. $\square$

Intuition: A weighted average of conditional probabilities

Because $\sum_i P(B_i) = 1$, the Law of Total Probability says that $P(A)$ is a **weighted average** of the conditional probabilities $P(A \mid B_i)$, weighted by how likely each scenario B_i is.

This gives a free sanity check: $P(A)$ must always lie between the smallest and the largest of the $P(A \mid B_i)$. If your answer falls outside that range you have made an arithmetic error.



$$P(A) = \sum_i P(A \mid B_i)P(B_i) \text{ — the slices of } A \text{ are disjoint and together make all of } A.$$

Figure 9.2: A partition $\{B_1, \dots, B_n\}$ slices any event A into disjoint pieces. Summing the pieces recovers $P(A)$.

9.3 Worked examples of total probability

Example 9.1 (A bag of trick coins)

A bag contains one fair coin, two double-headed coins and three double-tailed coins. A coin is selected at random and tossed. What is the probability that it lands on Heads?

Solution. Define H = “coin falls Heads”, C_F = “fair coin selected”, C_H = “double-headed coin selected”, C_T = “double-tailed coin selected”.

The partition.

$\{C_F, C_H, C_T\}$ is mutually exclusive and exhaustive, with

$$P(C_F) = \frac{1}{6}, \quad P(C_H) = \frac{2}{6}, \quad P(C_T) = \frac{3}{6}.$$

The conditional probabilities.

$$P(H | C_F) = \frac{1}{2}, \quad P(H | C_H) = 1, \quad P(H | C_T) = 0.$$

Apply the law.

$$\begin{aligned} P(H) &= P(H | C_F) P(C_F) + P(H | C_H) P(C_H) + P(H | C_T) P(C_T) \\ &= \frac{1}{2} \times \frac{1}{6} + 1 \times \frac{2}{6} + 0 \times \frac{3}{6} \\ &= \frac{1}{12} + \frac{4}{12} + 0 = \frac{5}{12} \approx 0.417. \end{aligned}$$

Check.

The conditional probabilities range from 0 to 1, and indeed $0 \leq \frac{5}{12} \leq 1$. More usefully: the answer is below $\frac{1}{2}$, which makes sense because double-tailed coins outnumber double-headed ones.

Exercise 9.1 (*Ms. Ferrari's car*)

Ms. Ferrari parks her car in the open air overnight. The probability that it starts on the first attempt in the morning depends on whether or not it rained during the night. If it rained, her car starts first time with probability 0.6. If it did not rain, it starts first attempt with probability 0.9. The probability of rain during the night is 0.4. Find the probability that the car starts on the first attempt.

Solution. Let S = “starts on the first attempt” and R = “it rained during the night”. We are given

$$P(S | R) = 0.6, \quad P(S | R') = 0.9, \quad P(R) = 0.4 \Rightarrow P(R') = 0.6.$$

Then

$$\begin{aligned} P(S) &= P(S \cap R) + P(S \cap R') \\ &= P(S | R) P(R) + P(S | R') P(R') \\ &= 0.6 \times 0.4 + 0.9 \times 0.6 \\ &= 0.24 + 0.54 = \boxed{0.78}. \end{aligned}$$

Check.

$P(S)$ must lie between 0.6 and 0.9, and 0.78 does. It sits closer to 0.9 because dry nights (60%) are more common than wet ones. ✓

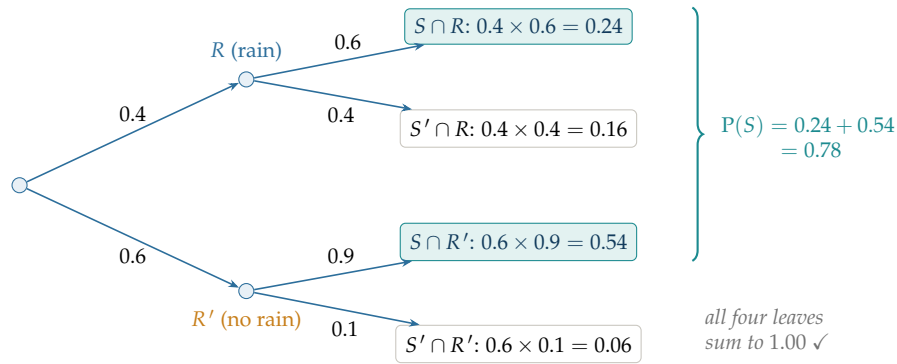


Figure 9.3: The car problem as a tree. The Law of Total Probability is precisely “add the leaves in which S occurs”.

9.4 Bayes' theorem

The Law of Total Probability runs *forwards*: from causes to effects. Bayes' theorem runs *backwards*: from an observed effect to the probability of each possible cause.

Theorem: Bayes' theorem

Suppose $A_1, A_2, A_3, \dots, A_n$ are mutually exclusive and exhaustive events, and B is an arbitrary event with $P(B) > 0$. Then for $i = 1, 2, \dots, n$,

$$P(A_i | B) = \frac{P(B | A_i) P(A_i)}{P(B)} = \frac{P(B | A_i) P(A_i)}{\sum_{j=1}^n P(B | A_j) P(A_j)}.$$

In the two-event case,

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)} \quad \text{for } P(B) > 0.$$

Proof. The proof follows directly from the definition of conditional probability and the Law of Total Probability:

$$P(A_i | B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B | A_i) P(A_i)}{\sum_j P(B | A_j) P(A_j)}.$$

The numerator is the multiplication rule; the denominator is the Law of Total Probability. $\square$

The vocabulary of Bayes

Term	Symbol	Meaning
Prior	$P(A_i)$	belief in cause A_i <i>before</i> seeing evidence
Likelihood	$P(B A_i)$	how well cause A_i explains the evidence B
Evidence	$P(B)$	total probability of the evidence
Posterior	$P(A_i B)$	belief in A_i <i>after</i> seeing the evidence

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}.$$

Intuition: Bayes reverses the arrow

The forward direction — “given the cause, how likely is the effect?” — is usually the one you can measure directly. A laboratory can test its diagnostic kit on known patients and report $P(\text{positive} \mid \text{disease})$. But the question a patient actually asks is the reverse: $P(\text{disease} \mid \text{positive})$. Bayes’ theorem is the bridge, and the crucial insight is that the bridge **depends on the prior**. The same test result means very different things in a high-prevalence population than in a low-prevalence one.

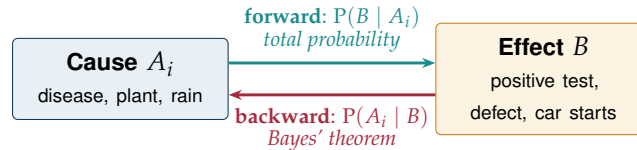


Figure 9.4: Total probability goes forwards; Bayes goes backwards. Almost every Bayes problem gives you the forward numbers and asks for a backward one.

9.5 The medical test: why Bayes matters

Example 9.2 (A laboratory test for a rare disease)

Suppose a laboratory test to detect a certain disease has the following statistics. Let A = “tested person has the disease” and B = “test result is positive”. It is known that

$$P(B \mid A) = 0.99, \quad P(B \mid A') = 0.005,$$

and 0.1% of the population actually has the disease. Find the probability that a person has the disease given that the test result is positive.

Solution.

The prior.

$$P(A) = 0.001, \text{ so } P(A') = 0.999.$$

Apply Bayes’ theorem.

$$\begin{aligned} P(A \mid B) &= \frac{P(B \mid A) P(A)}{P(B \mid A) P(A) + P(B \mid A') P(A')} \\ &= \frac{(0.99)(0.001)}{(0.99)(0.001) + (0.005)(0.999)} \\ &= \frac{0.00099}{0.00099 + 0.004995} = \frac{0.00099}{0.005985} \approx \boxed{0.165}. \end{aligned}$$

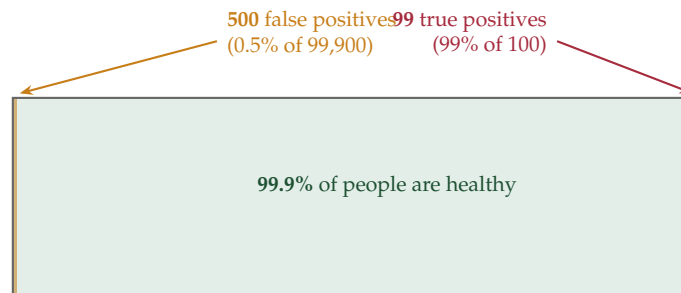
Why this is startling — and correct.

The test is 99% sensitive and has only a 0.5% false-positive rate, yet a positive result means the person has the disease with probability only about **16.5%**. Most positives are *false* positives. The reason is the **base rate**. Imagine 100,000 people:

	Test +	Test –	Total
Has disease	99	1	100
No disease	500	99,400	99,900
Total	599	99,401	100,000

Of the 599 positive tests, only 99 are genuinely ill: $99/599 \approx 0.165$. ✓

There are simply so many more healthy people that even a small false-positive *rate* produces a large false-positive *count*. This is the **base rate fallacy**, and it is why screening programmes for rare conditions always require a confirmatory second test.



Of the $500 + 99 = 599$ positive results, only 99 are true. $P(\text{disease} | +) = \frac{99}{599} \approx 0.165$

Figure 9.5: The base rate fallacy in one picture. The false-positive *rate* is tiny, but it applies to a vastly larger group, so false positives outnumber true positives five to one.

Common pitfall: The prosecutor's fallacy

Asserting that $P(A | B) = P(B | A)$ is so common in legal and medical settings that it has its own name. “The chance of this DNA match if the defendant is innocent is one in a million, therefore he is guilty with probability $999999/1000000$ ” is exactly the error above. The correct conclusion depends entirely on how many other people could have left the sample — the prior.

9.6 Bayes with more than two causes

Example 9.3 (Three manufacturing plants)

A company producing electric relays has three manufacturing plants producing 50, 30 and 20 percent, respectively, of its product. The probabilities that a relay manufactured by these plants is defective are 0.02, 0.05 and 0.01 respectively.

- If a relay is selected at random from the output of the company, what is the probability that it is defective?
- If a relay selected at random is found to be defective, what is the probability that it was manufactured by plant 2?

Solution. Let A_i = “the relay came from plant i ” and D = “the relay is defective”. We are given

$$P(A_1) = 0.50, \quad P(A_2) = 0.30, \quad P(A_3) = 0.20,$$

$$P(D | A_1) = 0.02, \quad P(D | A_2) = 0.05, \quad P(D | A_3) = 0.01.$$

(a) The Law of Total Probability.

$$\begin{aligned} P(D) &= P(D | A_1) P(A_1) + P(D | A_2) P(A_2) + P(D | A_3) P(A_3) \\ &= (0.02)(0.50) + (0.05)(0.30) + (0.01)(0.20) \\ &= 0.010 + 0.015 + 0.002 \\ &= \boxed{0.027}. \end{aligned}$$

So 2.7% of the company’s output is defective. Note this lies between the smallest and largest plant defect rates (0.01 and 0.05), as it must.

(b) Bayes’ theorem.

$$P(A_2 | D) = \frac{P(D | A_2) P(A_2)}{P(D)} = \frac{(0.05)(0.30)}{0.027} = \frac{0.015}{0.027} = \frac{5}{9} \approx \boxed{0.556}.$$

Interpretation.

Plant 2 makes only 30% of the relays, but is responsible for 55.6% of the defective ones — because its defect rate is more than double plant 1’s and five times plant 3’s. The prior of 0.30 has been *revised upward* by the evidence.

The full posterior distribution.

Plant	Prior $P(A_i)$	Likelihood $P(D A_i)$	Joint $P(D \cap A_i)$	Posterior $P(A_i D)$
1	0.50	0.02	0.010	$10/27 \approx 0.370$
2	0.30	0.05	0.015	$15/27 \approx 0.556$
3	0.20	0.01	0.002	$2/27 \approx 0.074$
	1.000	—	0.027	1.000 ✓

The posteriors must sum to 1 — always check this. Plant 1’s share falls from 0.50 to 0.37; plant 3’s collapses from 0.20 to 0.07.

Recipe: Solving any Bayes problem in four columns

Build the table above every time. It removes all the algebra.

- Column 1 — Priors.** List the mutually exclusive, exhaustive causes and their probabilities. They must sum to 1.
- Column 2 — Likelihoods.** For each cause, the probability of the observed evidence given that cause. These need *not* sum to anything.
- Column 3 — Joints.** Multiply columns 1 and 2. Their sum is $P(B)$ — this is the Law of Total Probability.

4. **Column 4 — Posteriors.** Divide each joint by that sum. These must sum to 1.

Independence

10.1 Definition

Definition: Independent events

Events A and B are said to be **independent** if one of the events can occur without being affected by the other. That is, if A is independent of B , then the probability of the occurrence of A is unchanged by the information as to whether or not B has occurred.

Formally, two events A and B are independent if

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

Why the product form is the official definition

The intuitive statement is $P(A | B) = P(A)$: “knowing B tells you nothing about A ”. The product form is preferred as the definition for two reasons.

1. It is **symmetric** in A and B , making it obvious that independence is a mutual relationship rather than a one-way influence.
2. It remains meaningful when $P(B) = 0$, where $P(A | B)$ is undefined.

10.2 Properties and their proofs

Property: Property 1

If $P(A) > 0$ and $P(B) > 0$, then A and B are independent **if and only if**

$$P(A | B) = P(A) \quad \text{and} \quad P(B | A) = P(B).$$

Proof ($\Rightarrow$): independence implies the conditionals are unchanged. Suppose A and B are independent, so $P(A \cap B) = P(A) P(B)$. Using the definition of conditional probability,

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

Substituting independence,

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

The same argument with the roles of A and B exchanged gives $P(B | A) = P(B)$. □

Proof ($\Leftarrow$): $P(A | B) = P(A)$ implies independence. Now suppose $P(A | B) = P(A)$ is true. By the conditional probability formula,

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

Substituting the assumption,

$$\frac{P(A \cap B)}{P(B)} = P(A) \implies P(A \cap B) = P(A)P(B),$$

which means A and B are independent. □

Property: Property 2

A and B independent implies that A and B' are independent.

Proof. Split A using the partition $\{B, B'\}$:

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

Therefore

$$\begin{aligned} P(A \cap B') &= P(A) - P(A \cap B) \\ &= P(A) - P(A)P(B) && \text{(by independence)} \\ &= P(A)[1 - P(B)] \\ &= P(A)P(B'), \end{aligned}$$

so A and B' are independent. □

Intuition: Why the complement version must be true

If knowing that B happened tells you nothing about A , then knowing that B *did not* happen must equally tell you nothing — both are statements about the same piece of information. Repeating the argument shows all four pairs (A, B) , (A, B') , (A', B) , (A', B') are independent together.

10.3 Independence versus mutual exclusivity

These two ideas are routinely confused, and they are almost opposites.

Common pitfall: Mutually exclusive events are *not* independent

If A and B are mutually exclusive with $P(A) > 0$ and $P(B) > 0$, then

$$P(A \cap B) = P(\emptyset) = 0 \quad \text{but} \quad P(A)P(B) > 0,$$

so they are **dependent**. And strongly so: knowing that A occurred tells you with certainty that B did not. Mutual exclusivity is the *maximum* possible dependence, not independence.

	Mutually exclusive	Independent
Defining equation	$P(A \cap B) = 0$	$P(A \cap B) = P(A)P(B)$
Venn diagram	circles do not overlap	circles overlap by exactly the “right” amount — not visible from a picture
Effect of knowing A	B becomes impossible	B ’s probability is unchanged
Addition rule	$P(A \cup B) = P(A) + P(B)$	$P(A \cup B) = P(A) + P(B) - P(A)P(B)$

Independence is invisible in a Venn diagram

You can *see* mutual exclusivity — the circles are separate. You cannot see independence, because it depends on the exact areas, not the shape of the overlap. Independence must always be **checked numerically**, never assumed from a picture.

10.4 Independence of more than two events

Definition: Mutual independence of three events

Three events A , B and C are **mutually independent** if and only if *all four* of the following hold:

$$\begin{aligned} P(A \cap B) &= P(A)P(B), \\ P(A \cap C) &= P(A)P(C), \\ P(B \cap C) &= P(B)P(C), \\ P(A \cap B \cap C) &= P(A)P(B)P(C). \end{aligned}$$

Common pitfall: Pairwise independence is weaker than mutual independence

The first three conditions are called **pairwise independence**. They do *not* imply the fourth. A set of events can be independent in every pair and still fail to be mutually independent — Exercise 10.1 constructs exactly such a case. This is why the definition lists four conditions rather than three.

Property: Consequences of mutual independence

If A , B and C are mutually independent events, then

- (a) A and $B \cap C$ are independent events;
 - (b) A and $B \cup C$ are independent events;
 - (c) A , B and C' are mutually independent events;
 - (d) A , B' and C' are mutually independent events;
- and so on for any combination of complements.

Proof of (a).

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

using mutual independence twice. □

10.5 Worked examples

Example 10.1 (*A die thrown twice*)

A die is thrown twice. Find the probability of obtaining a 4 on the first throw and an odd number on the second throw. Check whether A and B are independent.

Solution. Let A = “4 is obtained on the first throw”, so $P(A) = \frac{1}{6}$. Let B = “an odd number is obtained on the second throw”, so $P(B) = \frac{3}{6} = \frac{1}{2}$.

Physical argument.

The result on the second throw is not affected in any way by the result on the first throw. Therefore A and B are independent events.

Numerical verification.

In the 36-outcome space of ordered pairs, the outcomes in $A \cap B$ are $(4, 1), (4, 3), (4, 5)$ — three of them. So

$$P(A \cap B) = \frac{3}{36} = \frac{1}{12},$$

and indeed

$$P(A)P(B) = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12} = P(A \cap B). \checkmark$$

Example 10.2 (*Counters drawn with replacement*)

A bag contains 5 red counters and 7 black counters. A counter is drawn from the bag, the colour is noted and **the counter is replaced**. A second counter is then drawn. Let A be the event that the first counter is red and B the event that the second counter is black. Are A and B independent? Find $P(A \cap B)$.

Solution. $P(A) = \frac{5}{12}$. Since the first counter is replaced before the second draw is made, the composition of the bag is identical for both draws, so A and B are independent events and $P(B) = \frac{7}{12}$. Hence

$$P(A \cap B) = P(A)P(B) = \left(\frac{5}{12}\right)\left(\frac{7}{12}\right) = \frac{35}{144} \approx 0.243.$$

Contrast with Chapter 8.

Compare this with the green/white counters drawn *without* replacement. The single word “replaced” is what converts a conditional-probability problem into a multiplication of unconditional probabilities. Always look for it.

Example 10.3 (*Recovering $P(B)$ from independence*)

Events A and B are such that $P(A) = \frac{1}{3}$ and $P(A \cap B) = \frac{1}{12}$. If A and B are independent events, find (a) $P(B)$, (b) $P(A \cup B)$.

Solution.

(a).

Since A and B are independent,

$$P(A \cap B) = P(A)P(B) \implies \frac{1}{12} = \frac{1}{3}P(B) \implies P(B) = \frac{3}{12} = \frac{1}{4}.$$

(b).

By the addition rule,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{3} + \frac{1}{4} - \frac{1}{12} = \frac{4 + 3 - 1}{12} = \frac{6}{12} = \frac{1}{2}.$$

Example 10.4 (*A fair die thrown twice: “at least one 4”*)

A fair die is thrown twice. Find the probability that (a) neither throw results in a 4; (b) at least one throw results in a 4.

Solution. Let A = “obtain 4 on the first throw” and B = “obtain 4 on the second throw”. These are independent, and by Property 2 so are A' and B' .

(a) Neither throw is a 4.

$$P(A' \cap B') = P(A')P(B') = \frac{5}{6} \times \frac{5}{6} = \frac{25}{36} \approx 0.694.$$

(b) At least one throw is a 4.

$$P(\text{at least one 4}) = 1 - P(\text{neither is a 4}) = 1 - \frac{25}{36} = \frac{11}{36} \approx 0.306.$$

The “at least one” pattern

For n independent trials each succeeding with probability p ,

$$P(\text{at least one success}) = 1 - (1 - p)^n.$$

This is De Morgan’s law combined with independence, and it is the single most reused formula in the course. Trying to compute “at least one” directly means adding the cases “exactly one”, “exactly two”, ... — always take the complement instead.

Example 10.5 (*Testing independence from conditional probabilities*)

Two events A and B are such that $P(A) = \frac{1}{4}$, $P(A | B) = \frac{1}{2}$, $P(B | A) = \frac{2}{3}$.

(a) Are the events A and B independent?

(b) Find $P(A \cap B)$.

(c) Are A and B mutually exclusive?

(d) Find $P(B)$.

Solution.

(a) Independence.

If A and B were independent we would need $P(A | B) = P(A)$. Since $P(A | B) = \frac{1}{2}$ and $P(A) = \frac{1}{4}$, these differ, so A and B are **not independent**.

(b) The intersection.

By the multiplication rule,

$$P(A \cap B) = P(B | A) P(A) = \frac{2}{3} \times \frac{1}{4} = \frac{1}{6}.$$

(c) Mutually exclusive?

If A and B were mutually exclusive then $P(A \cap B) = 0$. Since $P(A \cap B) = \frac{1}{6} \neq 0$, they are **not mutually exclusive**.

(d) $P(B)$.

Both expressions equal $P(A \cap B)$, so

$$P(A | B) P(B) = P(B | A) P(A) \implies \frac{1}{2} P(B) = \frac{2}{3} \times \frac{1}{4} = \frac{1}{6},$$

$$P(B) = \frac{2}{6} = \frac{1}{3}.$$

Consistency check.

$P(A | B) = P(A \cap B) / P(B) = (1/6) / (1/3) = 1/2$. ✓ Note also that since $P(A | B) = \frac{1}{2} > \frac{1}{4} = P(A)$, the events are *positively* associated: B occurring makes A more likely.

10.6 Pairwise but not mutual independence**Exercise 10.1** (*The classic counterexample*)

Toss a fair coin three times and observe the sequence of Heads and Tails. Define

H_1 = Heads on first toss, H_2 = Heads on second toss, H_3 = Heads on third toss,
 A = all tosses the same.

- (a) Are the events H_1, H_2, H_3 mutually independent?
 (b) Are the events A, H_1 and H_2 mutually independent?

Solution. The outcome space is $\Omega = \{hhh, hht, hth, thh, htt, tht, tth, ttt\}$, all eight outcomes equally likely. Note $A = \{hhh, ttt\}$, so $P(A) = \frac{2}{8} = \frac{1}{4}$, and $P(H_1) = P(H_2) = P(H_3) = \frac{4}{8} = \frac{1}{2}$.

(a) Are H_1, H_2, H_3 mutually independent?

Check all four conditions.

$$P(H_1 \cap H_2) = P(\{hhh, hht\}) = \frac{2}{8} = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(H_1) P(H_2) \checkmark$$

$$P(H_1 \cap H_3) = P(\{hhh, hth\}) = \frac{2}{8} = \frac{1}{4} = P(H_1) P(H_3) \checkmark$$

$$P(H_2 \cap H_3) = P(\{hhh, thh\}) = \frac{2}{8} = \frac{1}{4} = P(H_2) P(H_3) \checkmark$$

$$P(H_1 \cap H_2 \cap H_3) = P(\{hhh\}) = \frac{1}{8} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = P(H_1) P(H_2) P(H_3) \checkmark$$

All four conditions hold, so H_1, H_2, H_3 are **mutually independent** — as expected physically, since the tosses do not influence one another.

(b) Are A, H_1, H_2 mutually independent?

First the three pairwise conditions.

$$\begin{aligned} P(A \cap H_1) &= P(\{hhh\}) = \frac{1}{8}, & P(A)P(H_1) &= \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8} \checkmark \\ P(A \cap H_2) &= P(\{hhh\}) = \frac{1}{8}, & P(A)P(H_2) &= \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8} \checkmark \\ P(H_1 \cap H_2) &= \frac{2}{8} = \frac{1}{4}, & P(H_1)P(H_2) &= \frac{1}{4} \checkmark \end{aligned}$$

So the three events are **pairwise independent**. Now the triple condition:

$$\begin{aligned} P(A \cap H_1 \cap H_2) &= P(\{hhh\}) = \frac{1}{8}, \\ P(A)P(H_1)P(H_2) &= \frac{1}{4} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{16}. \end{aligned}$$

Since $\frac{1}{8} \neq \frac{1}{16}$, the fourth condition **fails**. Therefore A, H_1 and H_2 are pairwise independent but **not mutually independent**.

Why this happens, in words.

Knowing H_1 alone tells you nothing about A : the first toss being a Head still leaves “all the same” with probability $\frac{1}{4}$ (you need two more Heads, probability $\frac{1}{4}$). Same for H_2 alone. But knowing H_1 and H_2 together is different information entirely. Given the first two tosses are both Heads, “all the same” now requires only that the third toss is a Head — probability $\frac{1}{2}$, not $\frac{1}{4}$. Two pieces of individually uninformative evidence have combined into something highly informative.

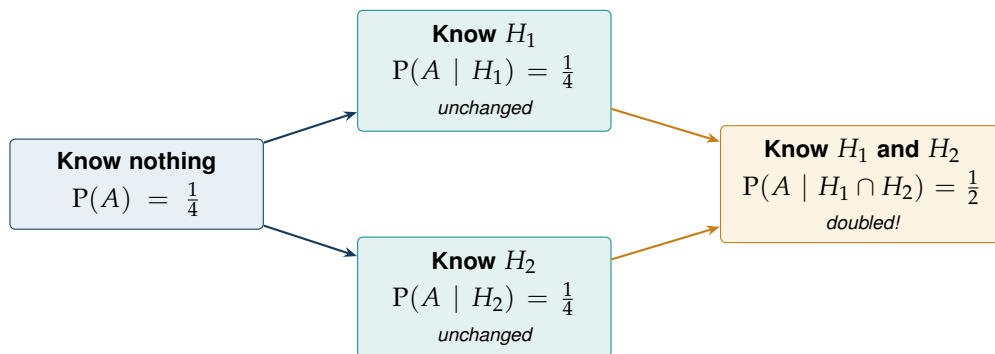


Figure 10.1: Pairwise independence does not compose. Each piece of evidence alone leaves $P(A)$ untouched, but together they double it.

10.7 Conditional independence

Learning outcome LO3 asks you to explain how events that are independent can be *conditionally dependent*, and vice versa. This section addresses it directly.

Definition: Conditional independence

Events A and B are **conditionally independent given C** if

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

The key fact

Conditional independence and unconditional independence are **logically unrelated**. Neither implies the other. All four combinations occur in practice:

	Independent given C	Dependent given C
Independent	the ordinary case	<i>explaining away</i> (Example 10.7)
Dependent	<i>common cause</i> (Example 10.6)	the ordinary case

Example 10.6 (*Dependent, but conditionally independent: a common cause*)

In a primary school, **shoe size** (A) and **reading ability** (B) are strongly positively correlated: children with bigger feet read better. The two are certainly *not* independent.

Now condition on C = the child's **age**. Among children who are all exactly seven years old, shoe size tells you nothing about reading ability. So

$$P(A \cap B) \neq P(A)P(B) \quad \text{but} \quad P(A \cap B \mid C) = P(A \mid C)P(B \mid C).$$

What is going on.

Age is a **common cause** of both. It generates a correlation between effects that have no direct link. Once you hold the common cause fixed, the spurious association disappears. This is the mechanism behind a great many misleading statistics: ice cream sales and drowning deaths (common cause: summer), or the number of fire engines at a fire and the damage caused (common cause: the size of the fire).

Example 10.7 (*Independent, but conditionally dependent: explaining away*)

Your car will not start. Two possible causes: A = “the battery is flat” and B = “the fuel tank is empty”. On any given morning these are **independent** — a flat battery says nothing about how much fuel you have.

Now condition on C = “the car will not start”. Given that the car has failed, suppose you check and find the fuel tank is full. This makes it *more* likely that the battery is flat — because something must explain the failure, and one candidate has been eliminated. So

$$P(A \mid C) < P(A \mid C \cap B'),$$

which means A and B are conditionally *dependent* given C , despite being unconditionally independent.

A numerical version.

Take $P(A) = P(B) = 0.1$, independent, and suppose the car fails to start exactly when at least one fault is present. Then

$$P(C) = 1 - (0.9)(0.9) = 0.19.$$

$$P(A \mid C) = \frac{P(A)}{P(C)} = \frac{0.1}{0.19} \approx 0.526.$$

$$P(A \mid C \cap B') = 1$$

— if the tank is full and the car will not start, the battery *must* be the problem. The probability jumps from 0.526 to 1. Learning about B changed our belief about A , so they are not conditionally independent.

This pattern is called **explaining away**: two independent causes become dependent once you condition on their shared effect.

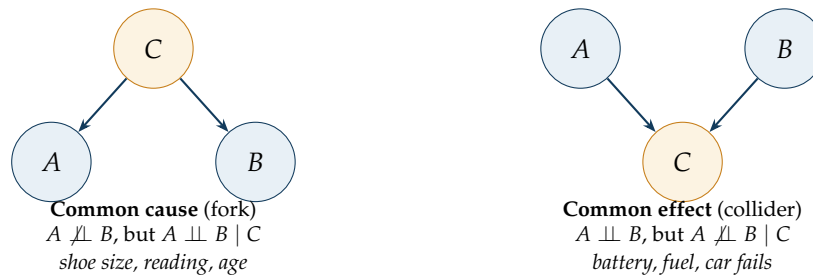


Figure 10.2: The two structures behind LO3. Conditioning on a *cause* destroys dependence; conditioning on an *effect* creates it.

Common pitfall: Never assume independence because it is convenient

Multiplying probabilities is so much easier than the alternative that it is tempting to assume independence by default. But an incorrect independence assumption produces answers that are wrong by orders of magnitude, and it has caused real miscarriages of justice — most famously in cases where the probability of two sudden infant deaths in one family was computed by squaring the probability of one, ignoring shared genetic and environmental causes. Justify independence physically, or verify it numerically.

PART III

Random Variables and Distributions

Random Variables and Probability Distributions

11.1 Turning outcomes into numbers

Many real-world experiments produce uncertain outcomes that are not numbers — “heads”, “defective”, “green”. To analyse them mathematically we convert outcomes into numbers. That conversion is what a random variable is.

Definition: Random variable

Let Ω be the outcome space corresponding to a random experiment. A function

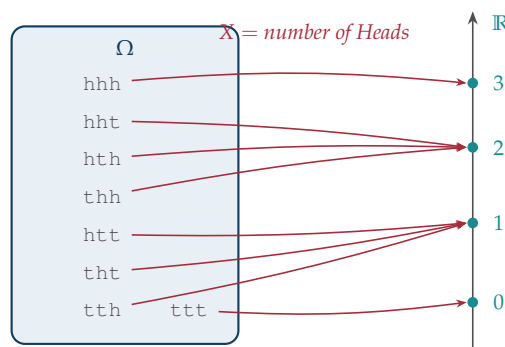
$$X : \Omega \longrightarrow \mathbb{R}$$

is called a **random variable**.

Common pitfall: A random variable is neither random nor a variable

The name is historically unfortunate. X is a **function** — a deterministic rule that assigns a number to each outcome. There is nothing random about the rule itself. The randomness lives entirely in *which outcome ω occurs*; X then reports a number.

Notation reflects this: capital X denotes the function, lower-case x denotes a particular value it might take. “ $X = x$ ” is shorthand for the *event* $\{\omega \in \Omega : X(\omega) = x\}$.



Three outcomes all map to 2: the event $\{X = 2\}$ is $\{hht, hth, thh\}$.

Figure 11.1: A random variable is a function from the outcome space to the real line. Several outcomes may share a value — that is what makes $P(X = 2)$ interesting.

Definition: Discrete random variable

The random variable X is a **discrete random variable** if the set of values taken by the random variable contains only a finite (or countable) number of points. We write Ω_X for this set — the *sample space* of X .

11.2 First examples**Example 11.1** (*Three coin tosses: X , Y and Z*)

A coin is tossed three times and the sequence of Heads and Tails is observed, so

$$\Omega = \{hhh, hht, hth, thh, htt, tht, tth, ttt\}.$$

Let X = number of Heads observed, Y = number of Tails observed, and $Z = XY$. The three functions are specified by the table:

ω	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
$X(\omega)$	3	2	2	2	1	1	1	0
$Y(\omega)$	0	1	1	1	2	2	2	3
$Z(\omega)$	0	2	2	2	2	2	2	0

Therefore

$$\Omega_X = \Omega_Y = \{0, 1, 2, 3\} \quad \text{and} \quad \Omega_Z = \{0, 2\}.$$

Two observations.

First, $Y = 3 - X$ always, so Y carries no information X does not. Second, Ω_Z has only two elements even though Z is built from variables with four — the product $X(3 - X)$ takes the value 0 when $X \in \{0, 3\}$ and the value 2 when $X \in \{1, 2\}$.

Exercise 11.1 (*Tossing until two successive tosses agree*)

A coin is tossed until two successive tosses give the same result. Let X = number of Heads observed and Y = total number of tosses made. Give Ω_X and Ω_Y .

Solution.**Describing Ω .**

The experiment stops the moment a toss matches the one before it. So every outcome is a sequence that *alternates* until the very last toss, which repeats its predecessor:

$$\Omega = \{hh, tt, htt, thh, hthh, thtt, httht, ththh, \dots\}$$

— an infinite set. Listing by length:

Length	Outcomes	X values	Y value
2	hh, tt	2, 0	2
3	htt, thh	1, 2	3
4	hthh, thtt	3, 1	4
5	hthtt, ththh	2, 3	5
6	hththh, ththtt	4, 2	6
$\vdots$	$\vdots$	$\vdots$	$\vdots$

The sample space of Y .

The experiment cannot stop before the second toss, and it can in principle continue forever, so

$$\Omega_Y = \{2, 3, 4, 5, \dots\} = \{n \in \mathbb{N} : n \geq 2\}.$$

The sample space of X .

From the table, X takes the value 0 (from tt), 1 (from htt), 2 (from hh), 3 (from hthh), 4 (from hththh), and so on. Every non-negative integer is attainable:

$$\Omega_X = \{0, 1, 2, 3, \dots\}.$$

Why X can reach every integer but Y cannot reach 1.

$Y \geq 2$ because you need at least two tosses to have a repeated pair. X can be 0 because the outcome tt contains no Heads at all. Both sample spaces are countably infinite, so X and Y are still *discrete* random variables — “discrete” means countable, not finite.

11.3 Probability distributions

Having a random variable is not enough; we need to know how likely each of its values is.

Definition: Probability function

Let X be a random variable with sample space Ω_X . The **probability function** of X is the function P_X defined by

$$P_X(x) = \begin{cases} P(X = x) & \text{if } x \in \Omega_X, \\ 0 & \text{otherwise.} \end{cases}$$

Definition: Probability distribution

By the **probability distribution** of X we mean the way in which the probability mass is distributed over the values in Ω_X . The probability function P_X is one way of specifying the probability distribution of X .

Property: The two defining conditions

Any function P_X is a legitimate probability function if and only if

$$(i) P_X(x) \geq 0 \text{ for all } x, \quad (ii) \sum_{x \in \Omega_X} P_X(x) = 1.$$

Condition (i) is Axiom 1; condition (ii) is Axiom 2 combined with Axiom 3, because the events $\{X = x\}$ for different x are mutually exclusive and together exhaust Ω . **Checking that your probabilities sum to 1 is the single most valuable habit in this chapter.**

Example 11.2 (*A biased coin tossed three times*)

Toss a coin three times and observe the sequence of Heads and Tails, so $\Omega = \{hhh, hht, hth, thh, htt, tht, tth, ttt\}$. Suppose the results on the tosses are independent and any given toss falls Heads with probability $\frac{1}{3}$. Define X = number of Heads and Y = number of Tails.

Solution.**Probabilities of the individual outcomes.**

By independence, an outcome with k Heads and $3 - k$ Tails has probability $(\frac{1}{3})^k (\frac{2}{3})^{3-k}$.

ω	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
$P(\omega)$	$\frac{1}{27}$	$\frac{2}{27}$	$\frac{2}{27}$	$\frac{2}{27}$	$\frac{4}{27}$	$\frac{4}{27}$	$\frac{4}{27}$	$\frac{8}{27}$
$X(\omega)$	3	2	2	2	1	1	1	0
$Y(\omega)$	0	1	1	1	2	2	2	3

Check: $\frac{1+2+2+2+4+4+4+8}{27} = \frac{27}{27} = 1. \checkmark$

The probability function of X .

Collect the outcomes by their X -value:

$$P_X(0) = P(\{ttt\}) = \frac{8}{27} \approx 0.296$$

$$P_X(1) = P(\{htt, tht, tth\}) = 3 \times \frac{4}{27} = \frac{12}{27} \approx 0.444$$

$$P_X(2) = P(\{hht, hth, thh\}) = 3 \times \frac{2}{27} = \frac{6}{27} \approx 0.222$$

$$P_X(3) = P(\{hhh\}) = \frac{1}{27} \approx 0.037$$

Sum: $\frac{8+12+6+1}{27} = 1. \checkmark$

And $P_Y(2)$.

$Y = 2$ exactly when $X = 1$, so $P_Y(2) = \frac{12}{27} = \frac{4}{9}$.

Recognising the pattern.

The coefficients 1, 3, 3, 1 are the binomial coefficients $\binom{3}{k}$, and

$$P_X(k) = \binom{3}{k} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{3-k}.$$

This is the Binomial distribution, derived properly in Chapter 13.

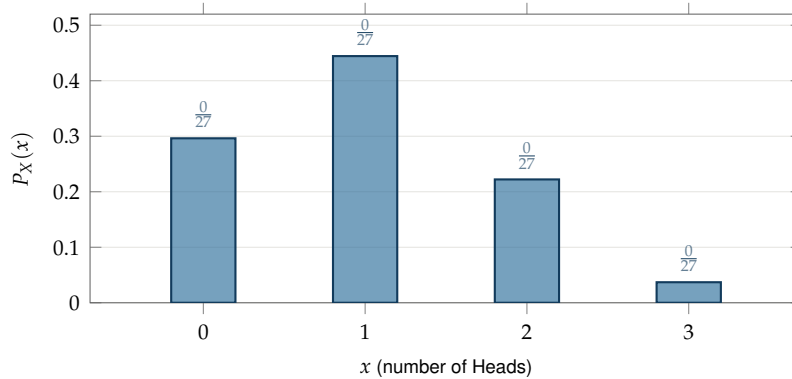


Figure 11.2: The probability distribution of X , the number of Heads in three tosses of a coin biased with $P(\text{head}) = \frac{1}{3}$. The bars sum to 1. The distribution is right-skewed because Heads are unlikely.

11.4 Functions of random variables

It is sometimes useful to define random variables as functions of other random variables. If X is a random variable, then so are $2X + 5$ and X^2 — any function of a random variable is itself a random variable, because a function of a function of ω is still a function of ω .

Recipe: Finding the distribution of a function of X

To find the distribution of $Y = g(X)$:

1. For each value $x \in \Omega_X$, compute $g(x)$.
2. Collect the x values that give the *same* $g(x)$.
3. Add their probabilities.

Step 3 is where the work is: distinct x values often map to the same y value, as with $Z = XY$ in the earlier example.

11.5 Extended example: the rectangle-cutting machine

Example 11.3 (Length, width and perimeter)

A machine is used to cut rectangular shapes out of sheets of metal. Switch A has 2 settings, which select the length L of the rectangle (2 cm or 3 cm). Switch B has 3 settings, which select the width W (1 cm, 2 cm or 3 cm). A mechanic tests the machine by selecting settings for switches A and B and making a rectangle. The length and width are observed. The settings are chosen randomly according to the table:

$\omega = (L, W)$	(2, 1)	(2, 2)	(2, 3)	(3, 1)	(3, 2)	(3, 3)
$P(\omega)$	0.12	0.13	0.25	0.20	0.15	0.15

Find (a) P_L , the distribution of the length; (b) P_W , the distribution of the width; (c) P_X , the distribution of the perimeter $X = 2L + 2W$.

Solution. First check the model: $0.12 + 0.13 + 0.25 + 0.20 + 0.15 + 0.15 = 1.00$. ✓

(a) The distribution of length.

Collect outcomes by their L -value.

$$P_L(2) = P\{(2,1)\} + P\{(2,2)\} + P\{(2,3)\} = 0.12 + 0.13 + 0.25 = 0.50$$

$$P_L(3) = P\{(3,1)\} + P\{(3,2)\} + P\{(3,3)\} = 0.20 + 0.15 + 0.15 = 0.50$$

L	2	3
$P(L)$	0.50	0.50

(b) The distribution of width.

Now collect by W -value:

$$P_W(1) = 0.12 + 0.20 = 0.32$$

$$P_W(2) = 0.13 + 0.15 = 0.28$$

$$P_W(3) = 0.25 + 0.15 = 0.40$$

W	1	2	3
$P(W)$	0.32	0.28	0.40

Sum: $0.32 + 0.28 + 0.40 = 1.00$. ✓

(c) The distribution of the perimeter.

Compute $X = 2L + 2W$ for each outcome:

ω	(2,1)	(2,2)	(2,3)	(3,1)	(3,2)	(3,3)
$X = 2L + 2W$	6	8	10	8	10	12
$P(\omega)$	0.12	0.13	0.25	0.20	0.15	0.15

Note that $X = 8$ arises from *two* different outcomes, (2,2) and (3,1), and $X = 10$ from (2,3) and (3,2). Adding:

$$P_X(6) = 0.12$$

$$P_X(8) = 0.13 + 0.20 = 0.33$$

$$P_X(10) = 0.25 + 0.15 = 0.40$$

$$P_X(12) = 0.15$$

X	6	8	10	12
$P(X)$	0.12	0.33	0.40	0.15

Sum: $0.12 + 0.33 + 0.40 + 0.15 = 1.00$. ✓

A hidden lesson: L and W are *not* independent

It is tempting to assume the two switches are set independently. Check it:

$$P(L = 2, W = 1) = 0.12 \quad \text{but} \quad P_L(2) P_W(1) = 0.50 \times 0.32 = 0.16.$$

Since $0.12 \neq 0.16$, L and W are **dependent**. The mechanic's choice of width is influenced by the length he has selected.

This matters. Had they been independent, you could have found P_X by convolving P_L and P_W . Because they are not, you *must* work from the joint table, exactly as done above. Never assume independence — verify it.

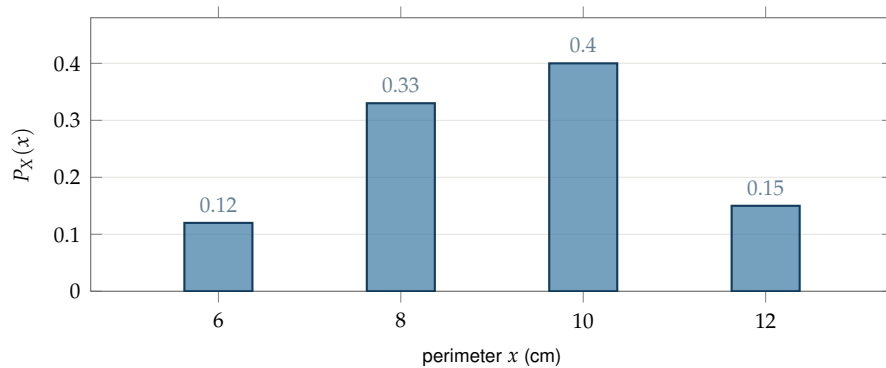


Figure 11.3: The probability distribution of the perimeter. Only even values from 6 to 12 are possible, and 10 cm is the most likely rectangle.

11.6 The cumulative distribution function

Although the handouts work mainly with probability functions, the *cumulative* version is often more convenient — and it is essential in Chapter 15, where continuous variables have no probability function at all.

Definition: Cumulative distribution function

The cumulative distribution function (c.d.f.) of a random variable X is

$$F_X(x) = P(X \leq x) = \sum_{t \leq x} P_X(t).$$

Property: Properties of the c.d.f.

- (a) F_X is non-decreasing;
- (b) $F_X(-\infty) = 0$ and $F_X(+\infty) = 1$;
- (c) $P(a < X \leq b) = F_X(b) - F_X(a)$;
- (d) for a discrete variable, F_X is a **step function**, jumping by $P_X(x)$ at each $x \in \Omega_X$.

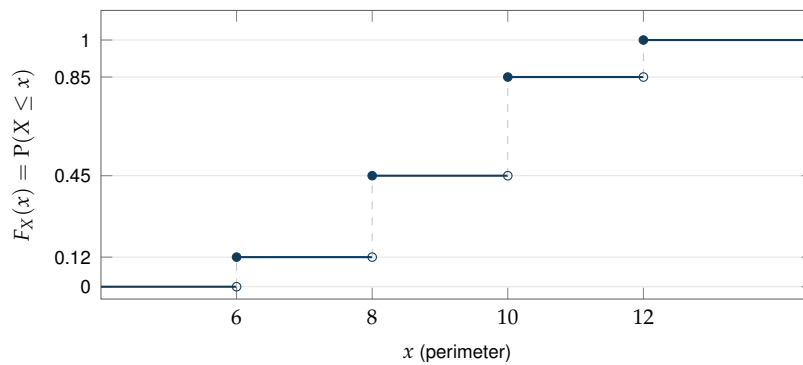


Figure 11.4: The c.d.f. of the perimeter. Filled circles show the value *at* each jump point (F is right-continuous); the size of each jump is exactly $P_X(x)$. For example the jump at $x = 10$ has height $0.85 - 0.45 = 0.40 = P_X(10)$.

Expectation and Variance

12.1 Why we need summary numbers

When the range Ω_X of a random variable contains only a few points it is easy to absorb the entire probability distribution at a glance. But it is often the case that Ω_X is very large. In these cases it is useful to have a measure of *location* and a measure of *spread* for the distribution — exactly the role played by the mean and standard deviation in Part I, but now applied to a theoretical distribution rather than to observed data.

The parallel with Part I

	Data (Part I)	Random variable (here)
Centre	$\bar{x} = \frac{1}{n} \sum x_i$	$E(X) = \sum x P(x)$
Spread	$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$	$\text{Var}(X) = \sum (x - \mu)^2 P(x)$
Status	a <i>statistic</i> (computed)	a <i>parameter</i> (theoretical)

The formulas are the same shape, with *probabilities* playing the role that *relative frequencies* played before. A sample mean estimates $E(X)$; that is the inferential bridge of Chapter 1.

12.2 Expectation

Definition: Expectation

The **expectation** of a random variable X defined on an outcome space Ω is the value $E(X)$ given by

$$E(X) = \sum_{\omega \in \Omega} X(\omega) P(\{\omega\}).$$

This quantity is also referred to as the *mean* of X , the *mean of the probability distribution* of X , or the *expected value* of X . It is usually denoted μ .

Property: Computing over values instead of outcomes

$$E(X) = \sum_{x \in \Omega_X} x P_X(x).$$

This is almost always the more practical formula: group the outcomes by the value of X first, then sum over values. The two agree because each value x collects total probability $P_X(x)$.

Intuition: Expectation is a weighted average

$E(X)$ is the average of the possible values, weighted by how likely each is. Two consequences follow immediately:

- $E(X)$ always lies between $\min \Omega_X$ and $\max \Omega_X$ — a useful sanity check.
- $E(X)$ need not be a value X can actually take. The expected number of Heads in three fair tosses is 1.5, which is impossible on any single trial. “Expected value” means *long-run average*, not “the value to expect”.

Example 12.1 (*Expectation from a probability model*)

Consider the following probability model:

ω	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
$P(\{\omega\})$	0.1	0.2	0.2	0.1	0.3	0.1
$X(\omega)$	3	3	2	1	3	2

Calculate the expected value of X .

Solution.

Method 1 — sum over outcomes.

$$\begin{aligned}
 E(X) &= 3(0.1) + 3(0.2) + 2(0.2) + 1(0.1) + 3(0.3) + 2(0.1) \\
 &= 0.3 + 0.6 + 0.4 + 0.1 + 0.9 + 0.2 \\
 &= \boxed{2.5}.
 \end{aligned}$$

Method 2 — collect by value first.

$$P_X(1) = 0.1, \quad P_X(2) = 0.2 + 0.1 = 0.3, \quad P_X(3) = 0.1 + 0.2 + 0.3 = 0.6.$$

Check: $0.1 + 0.3 + 0.6 = 1.0$. ✓

$$E(X) = 1(0.1) + 2(0.3) + 3(0.6) = 0.1 + 0.6 + 1.8 = 2.5. \checkmark$$

The second method is shorter and less error-prone whenever several outcomes share a value.

Check.

$$1 \leq 2.5 \leq 3. \checkmark$$

12.2.1 Properties of expectation**Property: Linearity**

1. $E(X) = \sum_{x \in \Omega_X} x P(x)$.
2. $E(aX + b) = a E(X) + b$, where a and b are constants.

3. If X and Y are random variables defined on the same outcome space Ω , then $E(X + Y) = E(X) + E(Y)$.

Proof of 2.

$$\begin{aligned} E(aX + b) &= \sum_{\omega} [aX(\omega) + b] P(\{\omega\}) \\ &= a \sum_{\omega} X(\omega) P(\{\omega\}) + b \sum_{\omega} P(\{\omega\}) \\ &= a E(X) + b \cdot 1 = a E(X) + b. \end{aligned}$$

□

Property 3 needs no independence

$E(X + Y) = E(X) + E(Y)$ holds for *any* two random variables on the same space, however strongly they depend on one another. This is remarkable and extremely useful — it is what makes the mean of the Binomial distribution trivial to derive in Chapter 13.

The corresponding statement for variance, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$, **does** require independence. Do not confuse the two.

Common pitfall: $E(g(X)) \neq g(E(X))$ in general

Expectation passes through *linear* functions only. In particular

$$E(X^2) \neq [E(X)]^2.$$

The gap between them is precisely the variance — see Property 1 below. To compute $E(X^2)$ you must square the *values* and then average, not average and then square.

Exercise 12.1 (Computing $E(X)$ and $E(X^2)$)

The random variable X has probability distribution given by

X	-2	-1	0	1	2
$P(X)$	0.3	0.1	0.2	0.1	0.3

(a) Find $E(X)$. (b) Find $E(X^2)$.

Solution. Add a row for X^2 :

X	-2	-1	0	1	2
X^2	4	1	0	1	4
$P(X)$	0.3	0.1	0.2	0.1	0.3

(a) $E(X)$.

$$\begin{aligned} E(X) &= (-2)(0.3) + (-1)(0.1) + (0)(0.2) + (1)(0.1) + (2)(0.3) \\ &= -0.6 - 0.1 + 0 + 0.1 + 0.6 = \boxed{0}. \end{aligned}$$

This is immediate from symmetry: the distribution is symmetric about 0.

(b) $E(X^2)$.

$$\begin{aligned} E(X^2) &= (4)(0.3) + (1)(0.1) + (0)(0.2) + (1)(0.1) + (4)(0.3) \\ &= 1.2 + 0.1 + 0 + 0.1 + 1.2 = \boxed{2.6}. \end{aligned}$$

Note.

Here $[E(X)]^2 = 0$ but $E(X^2) = 2.6$ — a vivid illustration of the pitfall above. The difference $2.6 - 0 = 2.6$ is the variance.

12.3 Variance

Definition: Variance and standard deviation

The mean of the random variable X is $\mu = E(X)$. The **variance** of X is

$$\sigma_X^2 = E[(X - \mu)^2] = \sum_{x \in \Omega_X} (x - \mu)^2 P(x).$$

Then σ_X is called the **standard deviation** of X .

Both the standard deviation and the variance can be used as measures of the spread of a probability distribution; the standard deviation is more commonly quoted because it is measured in the same units as X .

Property: Properties of variance

Property 1 $\text{Var}(X) = E(X^2) - \mu^2$.

Property 2 If c is any constant and X is a random variable, then $\text{Var}(cX) = c^2 \text{Var}(X)$.

Property 3 If X and Y are **independent** random variables, then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \quad \text{and} \quad \text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y).$$

Proof of Property 1 — the computational formula.

$$\begin{aligned} \text{Var}(X) &= E[(X - \mu)^2] \\ &= E[X^2 - 2\mu X + \mu^2] \\ &= E(X^2) - 2\mu E(X) + \mu^2 && \text{(linearity)} \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2. \end{aligned} \quad \square$$

Common pitfall: Why $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$, with a *plus*

This surprises everyone the first time. The reason: variance measures spread, and subtracting a second random quantity makes the result *more* variable, not less — there are now two sources of uncertainty instead of one. Formally, $\text{Var}(-Y) = (-1)^2 \text{Var}(Y) = \text{Var}(Y)$ by Property 2, so the minus sign is squared away. Variances *never* cancel.

Recipe: The standard variance calculation

1. Build a table with rows $x, x^2, P(x)$.
2. Compute $\mu = E(X) = \sum xP(x)$.
3. Compute $E(X^2) = \sum x^2P(x)$.
4. $\text{Var}(X) = E(X^2) - \mu^2$.
5. $\sigma_X = \sqrt{\text{Var}(X)}$.

Always use Property 1 rather than the definition — squaring deviations by hand is slower and invites arithmetic errors. And remember: $\text{Var}(X)$ can never be negative, so if step 4 gives a negative number you have made a mistake.

12.4 Worked variance problems**Exercise 12.2** (*A symmetric gamble*)

Toss an unbiased coin once and win \$ k if it is a Head and lose \$ k if it is a Tail. Find the standard deviation of the profit X , whose distribution is

X	$-k$	k
$P(X)$	0.5	0.5

Solution. Extend the table:

X	$-k$	k
X^2	k^2	k^2
$P(X)$	0.5	0.5

The mean.

$$E(X) = \mu = (-k)(0.5) + (k)(0.5) = 0.$$

The game is **fair**: on average you neither win nor lose.

The variance.

$$E(X^2) = (k^2)(0.5) + (k^2)(0.5) = k^2,$$

$$\text{Var}(X) = E(X^2) - \mu^2 = k^2 - 0 = k^2.$$

The standard deviation.

$$\sigma_X = \sqrt{k^2} = |k|.$$

Interpretation — the point of the exercise.

The expected profit is 0 regardless of k , but the standard deviation is $|k|$. A game with $k = \$1$ and a game with $k = \$10,000$ have *identical* expected values and wildly different risk. This is exactly why a measure of centre alone is never a sufficient description — the same lesson as Section 4.1, now in probability language.

Exercise 12.3 (A game with a die)

Suppose that a game is played with a single die assumed fair. In this game a player wins \$20 if a 2 turns up, \$40 if a 4 turns up, and loses \$30 if a 6 turns up; he neither wins nor loses if any other face turns up. Find the expected sum of money won.

Solution.**The distribution of the winnings X .**

Faces 1, 3 and 5 all give $X = 0$, so they combine:

Face	2	4	6	1, 3, 5
X (\$)	20	40	-30	0
$P(X)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{3}{6}$

Check: $\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{3}{6} = 1$. ✓

The expected winnings.

$$\begin{aligned} E(X) &= 20 \left(\frac{1}{6}\right) + 40 \left(\frac{1}{6}\right) + (-30) \left(\frac{1}{6}\right) + 0 \left(\frac{3}{6}\right) \\ &= \frac{20 + 40 - 30}{6} = \frac{30}{6} = \boxed{\$5}. \end{aligned}$$

Interpretation.

The player expects to win \$5 per game in the long run — so this game *favours the player*. A casino offering it would charge an entry fee of more than \$5 per play to make it profitable.

Bonus: how risky is it?

$$\begin{aligned} E(X^2) &= \frac{400 + 1600 + 900 + 0}{6} = \frac{2900}{6} \approx 483.33, \\ \text{Var}(X) &= 483.33 - 5^2 = 458.33, \quad \sigma_X = \sqrt{458.33} \approx \$21.41. \end{aligned}$$

The expected gain of \$5 comes with a standard deviation of over \$21, so individual games swing widely. Only over many plays does the \$5 average emerge — which is the practical meaning of “expected value”.

Example 12.2 (Expectation and variance of the perimeter)

Return to the rectangle machine of Chapter 11, where the perimeter X had distribution

X	6	8	10	12
X^2	36	64	100	144
$P(X)$	0.12	0.33	0.40	0.15

Solution.

$$E(X) = 6(0.12) + 8(0.33) + 10(0.40) + 12(0.15) = 0.72 + 2.64 + 4.00 + 1.80 = 9.16 \text{ cm.}$$

$$E(X^2) = 36(0.12) + 64(0.33) + 100(0.40) + 144(0.15) = 4.32 + 21.12 + 40.00 + 21.60 = 87.04.$$

$$\text{Var}(X) = 87.04 - (9.16)^2 = 87.04 - 83.9056 = 3.1344, \quad \sigma_X = \sqrt{3.1344} \approx 1.77 \text{ cm.}$$

Cross-check using linearity.

Since $X = 2L + 2W$, linearity of expectation gives

$$E(X) = 2E(L) + 2E(W).$$

From the marginal distributions,

$$E(L) = 2(0.5) + 3(0.5) = 2.5,$$

$$E(W) = 1(0.32) + 2(0.28) + 3(0.40) = 0.32 + 0.56 + 1.20 = 2.08.$$

$$2(2.5) + 2(2.08) = 5 + 4.16 = 9.16. \checkmark$$

Why the same shortcut fails for the variance.

Linearity of expectation held even though L and W are dependent. For the variance we would need $\text{Var}(2L + 2W) = 4\text{Var}(L) + 4\text{Var}(W)$, which requires independence — and we showed in Chapter 11 that L and W are *dependent*. Indeed $\text{Var}(L) = 0.25$ and $\text{Var}(W) \approx 0.7136$, so $4(0.25) + 4(0.7136) = 3.854 \neq 3.1344$. The dependence has reduced the variability of the perimeter.

Chapter 12 summary

Quantity	Formula
Expectation	$\mu = E(X) = \sum_x x P(x)$
Expectation of a function	$E[g(X)] = \sum_x g(x)P(x)$
Variance (definition)	$\sigma^2 = E[(X - \mu)^2]$
Variance (computational)	$\sigma^2 = E(X^2) - \mu^2$
Standard deviation	$\sigma = \sqrt{\text{Var}(X)}$
Linear transformation	$E(aX + b) = aE(X) + b; \quad \text{Var}(aX + b) = a^2 \text{Var}(X)$
Sums (always)	$E(X + Y) = E(X) + E(Y)$
Sums (independent only)	$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$

Note that b disappears from the variance — shifting a distribution does not change its spread,

exactly as in Chapter 4.

The Binomial Distribution

13.1 The binomial situation

Consider an experiment which has two possible outcomes, one of which is considered a **success** and the other a **failure**. A *binomial situation* arises when n **independent** trials of the experiment are conducted. For example:

- Toss a coin 6 times, considering obtaining a head on a single trial as the success.
- Roll a die 10 times, considering obtaining an even number as the “success”.

The four conditions — check all of them

A binomial model applies if and only if:

1. there is a **fixed** number n of trials, decided in advance;
2. each trial has exactly **two** outcomes, success or failure;
3. the trials are **independent**;
4. the probability of success p is **constant** across trials.

Sampling *with* replacement satisfies (3) and (4); sampling *without* replacement violates both, because the population changes. The mnemonic is **B-I-N-S**: **B**inary outcomes, **I**ndependent trials, **N** fixed, **S**ame probability.

13.2 Deriving the distribution

Definition: Binomial distribution

Let p be the probability that an event will happen in any single trial, so that the probability it fails to happen in any single trial is $1 - p$. The probability that the event will happen exactly k times in n independent trials is

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, 2, \dots, n.$$

When the random variable X denotes the number of successes in n independent trials, this function is called the **Binomial distribution**, since for $k = 0, 1, \dots, n$ it corresponds to the successive terms in the binomial expansion. We write $X \sim \text{Bin}(n, p)$.

Theorem: Why the formula has this form

The probability function above follows from independence and counting.

Proof. Fix one particular sequence of n trials containing exactly k successes, for instance

$$\underbrace{SS \cdots S}_k \underbrace{FF \cdots F}_{n-k}.$$

Because the trials are independent, the probability of that specific sequence is the product of the individual probabilities:

$$\underbrace{p \times p \times \cdots \times p}_k \times \underbrace{(1-p) \times \cdots \times (1-p)}_{n-k} = p^k (1-p)^{n-k}.$$

Crucially, *every* sequence with k successes has this same probability, because multiplication does not care about order. The number of such sequences is the number of ways to choose which k of the n positions are successes, namely $\binom{n}{k}$. Since these sequences are mutually exclusive, Axiom 3 lets us add:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

□

Intuition: The three ingredients

$$P(X = k) = \underbrace{\binom{n}{k}}_{\text{how many arrangements}} \times \underbrace{p^k}_{\substack{k \text{ successes} \\ \text{each } p}} \times \underbrace{(1-p)^{n-k}}_{\substack{n-k \text{ failures} \\ \text{each } 1-p}}$$

If you can reconstruct these three factors you can reconstruct the formula without memorising it.

Property: The probabilities sum to 1

$$\sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = [p + (1-p)]^n = 1^n = 1,$$

by the binomial theorem. This is the origin of the distribution's name.

Example 13.1 (The introductory example)

The probability of getting exactly 2 heads in 6 tosses of a fair coin.

Solution. Here $X \sim \text{Bin}(6, \frac{1}{2})$ and

$$P(X = 2) = \binom{6}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^4 = 15 \times \frac{1}{4} \times \frac{1}{16} = \frac{15}{64} \approx 0.234.$$

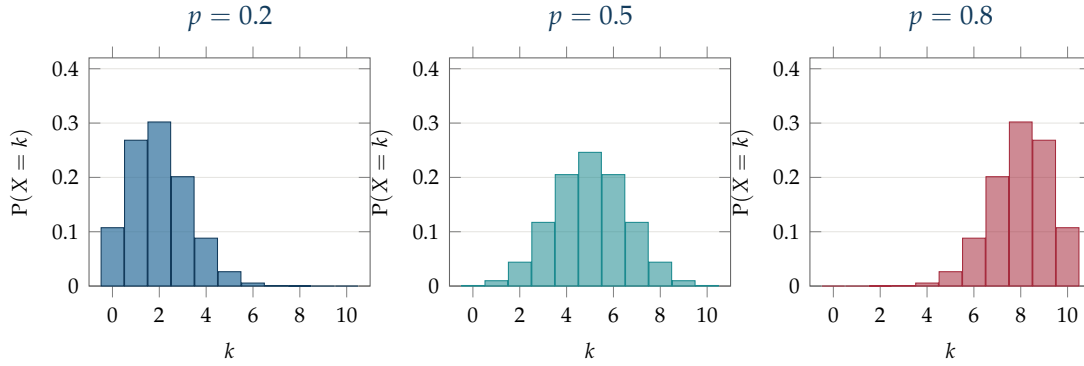


Figure 13.1: $\text{Bin}(10, p)$ for three values of p . The distribution is symmetric only when $p = 0.5$; it is right-skewed for small p and left-skewed for large p . The peak sits near np .

13.3 Mean and variance

Property: Some properties of the Binomial Distribution

If $X \sim \text{Bin}(n, p)$ and $q = 1 - p$, then

$$E(X) = np, \quad \text{Var}(X) = npq = np(1 - p), \quad \text{SD}(X) = \sqrt{np(1 - p)}.$$

Proof. Write $X = X_1 + X_2 + \cdots + X_n$, where $X_i = 1$ if trial i is a success and $X_i = 0$ otherwise. Each X_i has

$$E(X_i) = 1(p) + 0(q) = p, \quad E(X_i^2) = 1^2(p) + 0^2(q) = p,$$

$$\text{Var}(X_i) = E(X_i^2) - [E(X_i)]^2 = p - p^2 = p(1 - p) = pq.$$

By linearity of expectation (which needs no independence),

$$E(X) = \sum_{i=1}^n E(X_i) = np.$$

Because the trials are *independent*, variances add:

$$\text{Var}(X) = \sum_{i=1}^n \text{Var}(X_i) = npq. \quad \square$$

Intuition: Why $E(X) = np$ is obvious in hindsight

If you toss a coin 100 times and each toss is a head with probability $\frac{1}{2}$, of course you expect 50 heads. The formula np is just “number of trials $\times$ chance per trial”. The variance is less obvious, but note $pq = p(1 - p)$ is largest at $p = 0.5$ and shrinks to 0 as p approaches 0 or 1 — an almost-certain outcome has almost no variability.

13.4 Exercises from Handout 06, worked in full

Exercise 13.1 (Tossing a fair coin three times)

Find the probability that in tossing a fair coin three times there will appear (a) 3 heads; (b) 2 tails and 1 head; (c) at least 1 head; (d) not more than 1 tail.

Solution. Let X be the number of heads, $X \sim \text{Bin}(3, \frac{1}{2})$. Since $p = q = \frac{1}{2}$, every term carries a factor $(\frac{1}{2})^3 = \frac{1}{8}$, so $P(X = k) = \binom{3}{k}/8$.

(a) 3 heads.

$$P(X = 3) = \binom{3}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^0 = \frac{1}{8} = 0.125.$$

(b) 2 tails and 1 head.

Let Y be the number of tails; $Y \sim \text{Bin}(3, \frac{1}{2})$ too. Then

$$P(Y = 2) = \binom{3}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^1 = \frac{3}{8} = 0.375.$$

Equivalently, 2 tails and 1 head means $X = 1$, and $P(X = 1) = \binom{3}{1}/8 = 3/8$. ✓

(c) At least 1 head.

Use the complement — “at least one” almost always should.

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \binom{3}{0} \left(\frac{1}{2}\right)^3 = 1 - \frac{1}{8} = \frac{7}{8} = 0.875.$$

(d) Not more than 1 tail.

“Not more than 1 tail” means 0 or 1 tail, i.e. 3 heads or 2 heads:

$$P(X \geq 2) = P(X = 2) + P(X = 3) = \frac{\binom{3}{2}}{8} + \frac{\binom{3}{3}}{8} = \frac{3}{8} + \frac{1}{8} = \frac{4}{8} = \frac{1}{2}.$$

Full distribution, for reference.

k	0	1	2	3	Total
$P(X = k)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1 ✓

Exercise 13.2 (Five tosses of a fair die)

Find the probability that in 5 tosses of a fair die a 3 appears (a) twice; (b) at most once; (c) at least two times.

Solution. Let X be the number of 3s. A “success” is rolling a 3, so $p = \frac{1}{6}$, $q = \frac{5}{6}$, and $X \sim \text{Bin}(5, \frac{1}{6})$.

(a) Exactly twice.

$$\begin{aligned} P(X = 2) &= \binom{5}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3 \\ &= 10 \times \frac{1}{36} \times \frac{125}{216} \\ &= \frac{1250}{7776} = \frac{625}{3888} \approx 0.161. \end{aligned}$$

(b) At most once.

$$P(X \leq 1) = P(X = 0) + P(X = 1):$$

$$\begin{aligned} P(X = 0) &= \binom{5}{0} \left(\frac{5}{6}\right)^5 = \frac{3125}{7776}, \\ P(X = 1) &= \binom{5}{1} \left(\frac{1}{6}\right) \left(\frac{5}{6}\right)^4 = 5 \times \frac{1}{6} \times \frac{625}{1296} = \frac{3125}{7776}, \\ P(X \leq 1) &= \frac{3125 + 3125}{7776} = \frac{6250}{7776} = \frac{3125}{3888} \approx 0.804. \end{aligned}$$

(The two terms being equal is a coincidence of $n = 5$, $p = \frac{1}{6}$.)

(c) At least two times.

$$P(X \geq 2) = 1 - P(X \leq 1) = 1 - \frac{3125}{3888} = \frac{763}{3888} \approx 0.196.$$

Note how part (b) was set up precisely to make part (c) a one-line complement.

Exercise 13.3 (A family of four children)

Find the probability that in a family of 4 children there will be (a) at least 1 boy; (b) at least 1 boy and at least 1 girl. Assume the probability of a male birth is $\frac{1}{2}$.

Solution. Let X be the number of boys, $X \sim \text{Bin}(4, \frac{1}{2})$, so $P(X = k) = \binom{4}{k}/16$.

(a) At least 1 boy.

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{1}{16} = \frac{15}{16} = 0.9375.$$

(b) At least 1 boy and at least 1 girl.

This fails only in the two extreme cases: all boys ($X = 4$) or all girls ($X = 0$). These are mutually exclusive, so

$$P(\text{at least 1 of each}) = 1 - P(X = 0) - P(X = 4) = 1 - \frac{1}{16} - \frac{1}{16} = \frac{14}{16} = \frac{7}{8} = 0.875.$$

Common error.

Do not compute $P(X \geq 1) \times P(4 - X \geq 1)$. The events “at least one boy” and “at least one girl”

are *not* independent — they concern the same four children. Taking the complement of the two extremes is the correct route.

Example 13.2 (*Expected number of heads in 100 tosses*)

In 100 tosses of a fair coin, what is the expected number of Heads?

Solution. $X \sim \text{Bin}(100, \frac{1}{2})$, so

$$E(X) = np = 100 \times \frac{1}{2} = 50.$$

For completeness, $\text{Var}(X) = npq = 100 \times \frac{1}{2} \times \frac{1}{2} = 25$ and $\text{SD}(X) = 5$. So a typical result is 50 ± 5 heads, and by the empirical rule of Chapter 16 roughly 95% of the time you would see between 40 and 60 heads.

Exercise 13.4 (*Defective bolts*)

If the probability of a defective bolt is 0.1, find (a) the mean and (b) the standard deviation for the number of defective bolts in a total of 400 bolts.

Solution. Let X be the number of defective bolts, $X \sim \text{Bin}(400, 0.1)$.

(a) **The mean.**

$$E(X) = np = 400 \times 0.1 = 40 \text{ defective bolts.}$$

(b) **The standard deviation.**

$$\text{Var}(X) = npq = 400 \times 0.1 \times 0.9 = 36,$$

$$\text{SD}(X) = \sqrt{36} = 6 \text{ bolts.}$$

Interpretation.

Typically about 40 ± 6 bolts in a batch of 400 will be defective. Since $np = 40 > 5$ and $nq = 360 > 5$, this Binomial is well approximated by $N(40, 36)$ — see Section 16.5.

Exercise 13.5 (*Recovering n and p from the mean and variance*)

The random variable X has a binomial distribution with mean 5.76 and standard deviation 1.92. Find $P(X = 6)$.

Solution. This is a simultaneous-equations problem in disguise.

Set up the equations.

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

$$np = 5.76, \quad \sqrt{np(1-p)} = 1.92 \implies np(1-p) = 1.92^2 = 3.6864.$$

Divide to eliminate n .

$$\frac{np(1-p)}{np} = \frac{3.6864}{5.76} \implies 1-p = 0.64 \implies p = 0.36.$$

Back-substitute for n .

$$n = \frac{5.76}{p} = \frac{5.76}{0.36} = 16.$$

So $X \sim \text{Bin}(16, 0.36)$. Check: $np = 5.76$ ✓ and $npq = 16(0.36)(0.64) = 3.6864$ ✓, and $n = 16$ is a whole number as it must be.

Compute $P(X = 6)$.

$$P(X = 6) = \binom{16}{6} (0.36)^6 (0.64)^{10}.$$

With $\binom{16}{6} = 8008$, $(0.36)^6 \approx 2.1768 \times 10^{-3}$ and $(0.64)^{10} \approx 1.1529 \times 10^{-2}$,

$$P(X = 6) \approx 8008 \times 2.1768 \times 10^{-3} \times 1.1529 \times 10^{-2} \approx \boxed{0.201}.$$

Sanity check.

The mean is 5.76, so $X = 6$ should be near the peak of the distribution — a probability around 0.2 is exactly what one expects for the modal value of a distribution with standard deviation 1.92.

Recognising a Binomial problem

Phrases that signal $\text{Bin}(n, p)$:

Wording	Translation
"exactly k "	$P(X = k)$
"at least k "	$P(X \geq k) = 1 - P(X \leq k - 1)$
"at most k ", "not more than k "	$P(X \leq k)$
"more than k "	$P(X \geq k + 1)$
"fewer than k "	$P(X \leq k - 1)$
"none"	$P(X = 0) = q^n$
"all"	$P(X = n) = p^n$

Whenever the phrase contains "at least one", reach for $1 - P(X = 0)$ first.

The Poisson Distribution

14.1 Definition

Definition: Poisson distribution

A discrete random variable X having probability function of the form

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, 3, \dots \text{ to infinity,}$$

where λ can take any positive value, is said to follow the **Poisson distribution**. We write $X \sim \text{Poisson}(\lambda)$.

Note that unlike the Binomial, the Poisson has **no upper limit**: X can in principle be any non-negative integer. There is one parameter, λ .

Property: The probabilities sum to 1

$$\sum_{x=0}^{\infty} \frac{e^{-\lambda} \lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1,$$

using the Maclaurin series $e^{\lambda} = \sum_{x \geq 0} \lambda^x / x!$. This is where the factor $e^{-\lambda}$ comes from — it is exactly the normalising constant needed to make the probabilities add to one.

14.2 Expectation and variance

Property: Mean and variance

If $X \sim \text{Poisson}(\lambda)$ then it can be shown that

$$E(X) = \lambda \quad \text{and} \quad \text{Var}(X) = \lambda.$$

Proof. For the mean, note the $x = 0$ term contributes nothing:

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x \frac{e^{-\lambda} \lambda^x}{x!} = \sum_{x=1}^{\infty} \frac{e^{-\lambda} \lambda^x}{(x-1)!} \\ &= \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda e^{-\lambda} e^{\lambda} = \lambda. \end{aligned}$$

For the variance, first compute $E[X(X-1)]$, where now the $x=0$ and $x=1$ terms vanish:

$$E[X(X-1)] = \sum_{x=2}^{\infty} \frac{e^{-\lambda} \lambda^x}{(x-2)!} = \lambda^2 e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda^2.$$

Hence $E(X^2) = E[X(X-1)] + E(X) = \lambda^2 + \lambda$, and

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

□

The Poisson signature

Mean = variance is the fingerprint of the Poisson distribution. If a dataset of counts has a sample mean and sample variance close to each other, a Poisson model is plausible; if the variance is much larger than the mean (called *overdispersion*), it is not.

Contrast with the Binomial, where $\text{Var} = np(1-p) < np = E$ — the variance is always *smaller* than the mean.

14.3 Uses of the Poisson distribution

Definition: When to use it

If an event is randomly scattered in time or space and has mean number of occurrences λ in a given interval of time or space, and if X is the random variable “the number of occurrences in the given interval”, then $X \sim \text{Poisson}(\lambda)$.

Examples given in the handout — the number of:

- car accidents on a particular stretch of road in one day;
- accidents in a factory in one week;
- telephone calls made to a switchboard in a given minute;
- insurance claims made to a company in a given time.

Intuition: The Poisson as a limit of the Binomial

Divide an hour into $n = 3600$ one-second slots. Suppose a call can arrive in each slot with small probability p , independently. Then the number of calls in the hour is $\text{Bin}(n, p)$ with mean $np = \lambda$.

Now let the slots get finer: $n \rightarrow \infty$ and $p \rightarrow 0$ with $np = \lambda$ held fixed. In the limit,

$$\binom{n}{k} p^k (1-p)^{n-k} \rightarrow \frac{e^{-\lambda} \lambda^k}{k!}.$$

So the Poisson is what the Binomial becomes when you have a very large number of opportunities, each individually very unlikely. This explains both the shape of the formula and the kinds of situation it models: **rare events, many chances**.

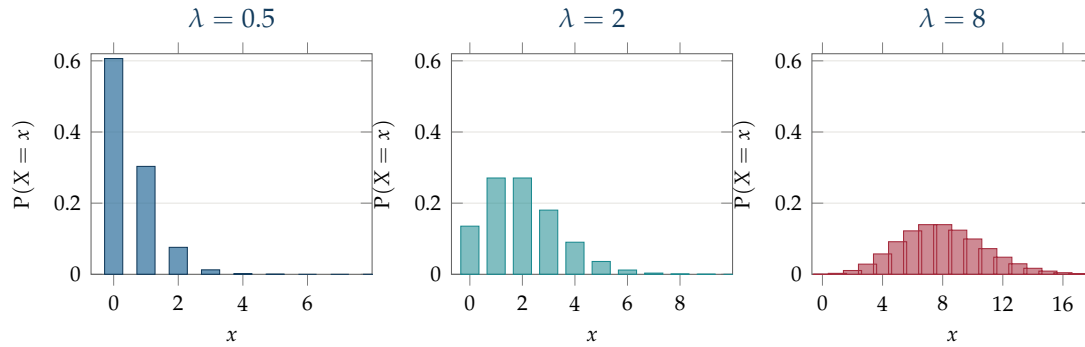


Figure 14.1: The Poisson distribution for three values of λ . For small λ it is strongly right-skewed with its peak at 0; as λ grows it becomes increasingly symmetric and bell-shaped — which is why the Normal approximation of Section 14.6 works for large λ .

14.4 Worked examples

Example 14.1 (Basic Poisson calculations)

If $X \sim \text{Poisson}(2)$, find (a) $P(X = 4)$; (b) $P(X \geq 3)$.

Solution. Here $\lambda = 2$ and $e^{-2} \approx 0.135335$.

(a) $P(X = 4)$.

$$P(X = 4) = \frac{e^{-2} 2^4}{4!} = \frac{e^{-2} \times 16}{24} = \frac{2e^{-2}}{3} \approx \frac{2 \times 0.135335}{3} \approx \boxed{0.0902}.$$

(b) $P(X \geq 3)$.

An infinite sum — so take the complement.

$$\begin{aligned} P(X \geq 3) &= 1 - P(X = 0) - P(X = 1) - P(X = 2) \\ &= 1 - e^{-2} \left[\frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!} \right] \\ &= 1 - e^{-2} [1 + 2 + 2] \\ &= 1 - 5e^{-2} \\ &\approx 1 - 5(0.135335) = 1 - 0.676675 \approx \boxed{0.3233}. \end{aligned}$$

Why the complement is not optional here.

$P(X \geq 3)$ is a sum of infinitely many terms. Only the complement gives a finite calculation. Always convert “at least” into $1 -$ “at most” for the Poisson.

Example 14.2 (Bacteria in a liquid)

The mean number of bacteria per millilitre of a liquid is known to be 4. Assuming that the number of bacteria follows a Poisson distribution, find the probability that, in 1 ml of liquid, there will be (a) no bacteria; (b) 4 bacteria; (c) less than 3 bacteria.

Solution. $X \sim \text{Poisson}(4)$, and $e^{-4} \approx 0.0183156$.

(a) No bacteria.

$$P(X = 0) = \frac{e^{-4}4^0}{0!} = e^{-4} \approx \boxed{0.0183}.$$

(b) Exactly 4 bacteria.

$$P(X = 4) = \frac{e^{-4}4^4}{4!} = \frac{256e^{-4}}{24} = \frac{32e^{-4}}{3} \approx 10.6667 \times 0.0183156 \approx \boxed{0.1954}.$$

Note this is the modal value, yet its probability is under 20% — with count data the mean is rarely a likely individual outcome.

(c) Less than 3 bacteria.

“Less than 3” means $X \in \{0, 1, 2\}$:

$$\begin{aligned} P(X < 3) &= e^{-4} \left[\frac{4^0}{0!} + \frac{4^1}{1!} + \frac{4^2}{2!} \right] \\ &= e^{-4} [1 + 4 + 8] = 13e^{-4} \\ &\approx 13 \times 0.0183156 \approx \boxed{0.2381}. \end{aligned}$$

Common pitfall: “Less than 3” excludes 3

For a discrete variable, strict and non-strict inequalities give genuinely different answers:

$$P(X < 3) = P(X \leq 2) \neq P(X \leq 3).$$

Here $P(X \leq 3) = P(X < 3) + P(X = 3) \approx 0.2381 + 0.1954 = 0.4335$ — almost double. Read the wording carefully. (In Chapter 15 you will see this distinction disappears for continuous variables.)

14.5 Scaling the interval

Property: Changing the interval length

If events occur at a rate of λ per unit interval, then the number of events in an interval of length t is $\text{Poisson}(\lambda t)$.

Intuition: Why λ scales linearly

Three millilitres of liquid contain three times as many bacteria on average as one millilitre. Since λ is the mean, it must triple. This is the additivity of the Poisson: the sum of independent $\text{Poisson}(\lambda_1)$ and $\text{Poisson}(\lambda_2)$ variables is $\text{Poisson}(\lambda_1 + \lambda_2)$.

Example 14.3 (Bacteria in 3 ml)

Using the data of the previous example, find the probability that in 3 ml of liquid there will be less than 2 bacteria.

Solution. Let Y be the number of bacteria in 3 ml. The mean scales:

$$\lambda_Y = 3 \times 4 = 12, \quad Y \sim \text{Poisson}(12).$$

$$\begin{aligned} P(Y < 2) &= P(Y = 0) + P(Y = 1) \\ &= e^{-12} \left[\frac{12^0}{0!} + \frac{12^1}{1!} \right] = 13e^{-12} \\ &\approx 13 \times 6.1442 \times 10^{-6} \approx \boxed{7.99 \times 10^{-5}}. \end{aligned}$$

Sense check.

Roughly one chance in 12,500. That is far smaller than the corresponding probability for 1 ml, which is reasonable: it would be surprising to find almost no bacteria in three times the volume when you expect twelve.

Common pitfall: Scale λ , never the probability

A frequent error is to compute the answer for 1 ml and then multiply the *probability* by 3. Probabilities do not scale like that — and the result could exceed 1. It is the **mean** λ that scales with the interval, and the probability must be recomputed from the new λ .

14.6 The Normal approximation to the Poisson

Theorem: Normal approximation

If $X \sim \text{Poisson}(\lambda)$ then $E(X) = \lambda$ and $\text{Var}(X) = \lambda$. Now for **large** λ ,

$$X \sim N(\lambda, \lambda) \quad \text{approximately.}$$

The usual rule of thumb is $\lambda \geq 10$, and preferably $\lambda \geq 20$. Because a discrete distribution is being replaced by a continuous one, a **continuity correction** is needed — see the worked example below and Section 16.5.1.

Example 14.4 (Radioactive disintegration)

A radioactive disintegration gives counts that follow a Poisson distribution with mean count per second of 25. Find the probability that in 1 second the count is between 23 and 27 inclusive, (a) using the Poisson distribution; (b) using the Normal approximation to the Poisson distribution.

Solution.

(a) Exact Poisson calculation.

$X \sim \text{Poisson}(25)$. We need

$$P(23 \leq X \leq 27) = \sum_{x=23}^{27} \frac{e^{-25} 25^x}{x!}.$$

Evaluating each term:

x	23	24	25	26	27	Sum
$P(X = x)$	0.0763	0.0795	0.0795	0.0765	0.0708	0.3826

$$P(23 \leq X \leq 27) \approx \boxed{0.3827}.$$

(Carrying full precision through the individual terms gives 0.38265; the table is rounded to four decimal places, so its column total differs in the final digit.)

(b) Normal approximation.

Since $\lambda = 25$ is large, take

$$X \approx N(\mu = 25, \sigma^2 = 25), \quad \sigma = 5.$$

Apply the continuity correction: the discrete set $\{23, 24, 25, 26, 27\}$ becomes the continuous interval $(22.5, 27.5)$.

$$P(22.5 < X < 27.5) = P\left(\frac{22.5 - 25}{5} < Z < \frac{27.5 - 25}{5}\right) = P(-0.5 < Z < 0.5).$$

From the standard normal table (Appendix B),

$$\Phi(0.5) = 0.6915, \quad \Phi(-0.5) = 1 - 0.6915 = 0.3085,$$

$$P(-0.5 < Z < 0.5) = 0.6915 - 0.3085 = \boxed{0.3830}.$$

Comparison.

Method	Answer	Error
Exact Poisson	0.3827	—
Normal, with continuity correction	0.3830	+0.0003
Normal, <i>without</i> correction	0.3108	−0.0719

The approximation with the correction is accurate to three decimal places. The version without the correction is wrong by more than 18% — which is exactly why the correction is not optional.

Binomial or Poisson? A decision rule

	Binomial	Poisson
Question form	“how many out of n ?”	“how many in an interval?”
Upper limit	n	none
Parameters	n and p	λ alone
Mean, variance	$np, np(1 - p)$	λ, λ
Typical wording	“in 20 trials”, “of 400 bolts”	“per hour”, “per ml”, “per week”

If the problem tells you both a number of trials *and* a probability, it is Binomial. If it gives you only an average rate, it is Poisson.

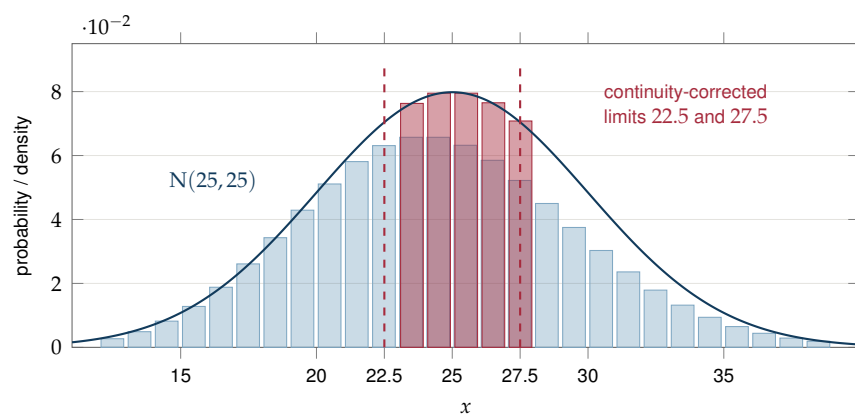


Figure 14.2: The $\text{Poisson}(25)$ bars with the $N(25, 25)$ density superimposed. Each bar has width 1, so its area equals its probability; the shaded bars run from 22.5 to 27.5, which is exactly the interval the Normal must be integrated over.

Continuous Random Variables

15.1 What is a continuous random variable?

Consider a random variable which can take any value in an interval. Such a random variable is *not* a discrete random variable, because the values in an interval cannot be placed in one-to-one correspondence with the counting numbers.

Definition: Continuous random variable

A random variable which can take any value in an interval is called a **continuous random variable**.

Common pitfall: For a continuous variable, $P(X = x) = 0$ for every x

This is the fact that makes continuous variables genuinely different, and it sounds absurd until you see why.

There are uncountably many points in an interval. If each individual point had probability $\varepsilon > 0$, then any finite collection of n points would carry probability $n\varepsilon$, which exceeds 1 once $n > 1/\varepsilon$ — contradicting Axiom 2. The only consistent assignment is $P(X = x) = 0$.

Two consequences follow.

1. Probability zero does **not** mean impossible. Some value does occur; each particular value simply had probability zero beforehand.
2. Strict and non-strict inequalities coincide:

$$P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b).$$

This is a genuine simplification — contrast the care needed with “less than 3” for the Poisson in Chapter 14.

15.2 The probability density function

Since individual values carry no probability, we describe a continuous variable by how densely probability is packed near each value.

Definition: Probability density function

The probability properties of a continuous random variable X are specified by its **probability density function** $f(x)$, which has the properties

1. $f(x) \geq 0$ for every $x \in \mathbb{R}$;
2. $\int_{-\infty}^{\infty} f(x) dx = 1$.

Definition: Probabilities as areas

The probability of the event $\{a < X \leq b\}$ is found from the density function f as

$$P(a < X \leq b) = \int_a^b f(x) dx.$$

Note that the density function of a continuous random variable is such that **probabilities are given by areas under its graph**.

The discrete/continuous dictionary

Every formula from Chapters 11–12 has a continuous counterpart obtained by replacing sums with integrals and probability functions with densities.

	Discrete	Continuous
Describes distribution	$P(x) = P(X = x)$	$f(x)$, a density
Total probability	$\sum_x P(x) = 1$	$\int_{-\infty}^{\infty} f(x) dx = 1$
Interval probability	$\sum_{a \leq x \leq b} P(x)$	$\int_a^b f(x) dx$
Expectation	$\sum_x xP(x)$	$\int x f(x) dx$
$E(X^2)$	$\sum_x x^2 P(x)$	$\int x^2 f(x) dx$
c.d.f.	$F(x) = \sum_{t \leq x} P(t)$	$F(x) = \int_{-\infty}^x f(t) dt$

Common pitfall: $f(x)$ is not a probability

A density can exceed 1. For example $f(x) = 4$ on $[0, 0.25]$ is a perfectly valid density (its area is $4 \times 0.25 = 1$), yet $f(0.1) = 4$. What must never exceed 1 is the *area*, not the height. The units of f are “probability per unit of x ”, which is why it is called a *density*.

15.3 Example: the exponential density**Example 15.1** (An exponential density and the area it represents)

Suppose the random variable T has density function given by

$$f_T(t) = 0 \quad \text{when } t < 0, \quad f_T(t) = \lambda e^{-\lambda t} \quad \text{when } t \geq 0.$$

Show the area corresponding to $P(1 < T \leq 2)$ on a graph.

Solution. The required probability is the area under the curve between $t = 1$ and $t = 2$:

$$P(1 < T \leq 2) = \int_1^2 \lambda e^{-\lambda t} dt = \left[-e^{-\lambda t} \right]_1^2 = e^{-\lambda} - e^{-2\lambda}.$$

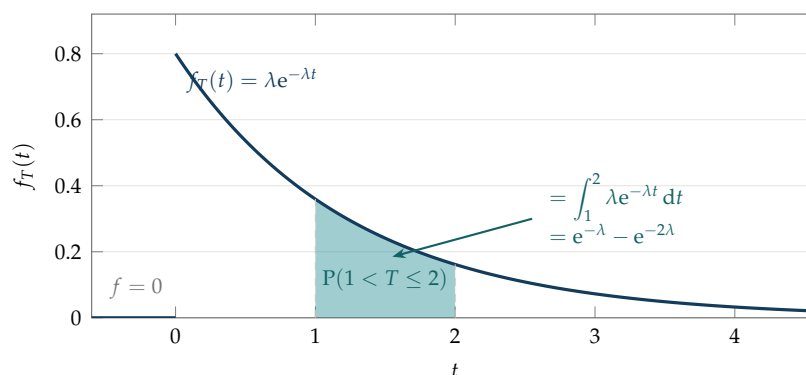


Figure 15.1: The exponential density with $\lambda = 0.8$. The shaded area is the probability. Note the density is 0 for $t < 0$, since the variable models a waiting time.

Recipe: Finding an unknown constant in a density

Almost every exam question of this type starts the same way: you are given $f(x) = k \cdot$ (something) on an interval and asked for k . The constant is always determined by condition (2):

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

Integrate over the interval where f is non-zero, set the result equal to 1, and solve for k . There is never more than one answer.

15.4 Exercises from Handout 07, worked in full

Exercise 15.1 (A triangular density)

A continuous random variable has p.d.f. $f(x)$ where

$$f(x) = kx, \quad 0 \leq x \leq 4.$$

(a) Find the value of the constant k ; (b) sketch $y = f(x)$; (c) find $P(1 \leq X \leq 2.5)$.

Solution.

(a) Finding k .

The total area must be 1:

$$\int_0^4 kx dx = 1 \implies k \left[\frac{x^2}{2} \right]_0^4 = 1 \implies k \left(\frac{16}{2} - 0 \right) = 1 \implies 8k = 1,$$

$$\boxed{k = \frac{1}{8}} \quad \text{so} \quad f(x) = \frac{x}{8}, \quad 0 \leq x \leq 4.$$

Check with geometry: the graph is a triangle of base 4 and height $f(4) = 4/8 = 0.5$, so its area is $\frac{1}{2} \times 4 \times 0.5 = 1$. ✓

(b) The sketch.

See Figure 15.2. It is a straight line from $(0,0)$ to $(4,0.5)$, and zero elsewhere.

(c) $P(1 \leq X \leq 2.5)$.

$$\begin{aligned} P(1 \leq X \leq 2.5) &= \int_1^{2.5} \frac{x}{8} dx = \left[\frac{x^2}{16} \right]_1^{2.5} \\ &= \frac{(2.5)^2}{16} - \frac{1^2}{16} = \frac{6.25 - 1}{16} = \frac{5.25}{16} = \frac{21}{64} \approx \boxed{0.328}. \end{aligned}$$

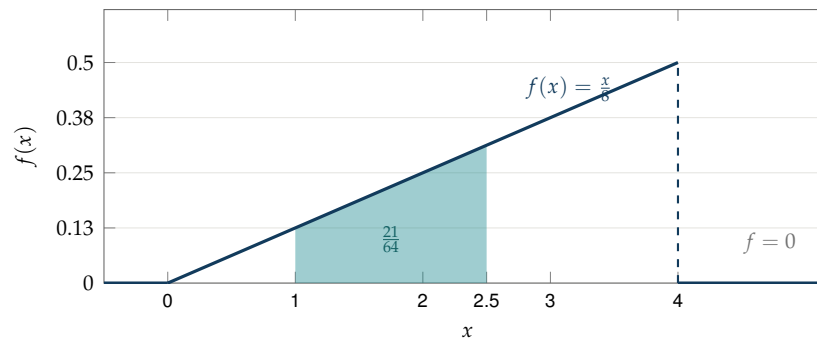


Figure 15.2: The density $f(x) = x/8$ on $[0, 4]$. The whole triangle has area 1; the shaded trapezium has area $21/64 \approx 0.328$.

Exercise 15.2 (A decreasing linear density)

The continuous random variable X has p.d.f. $f(x)$ where

$$f(x) = k(4 - x), \quad 1 \leq x \leq 3.$$

(a) Find the value of the constant k ; (b) sketch $y = f(x)$; (c) find $P(1.2 \leq X \leq 2.4)$.

Solution.

(a) Finding k .

$$\begin{aligned} \int_1^3 k(4 - x) dx &= 1 \\ k \left[4x - \frac{x^2}{2} \right]_1^3 &= 1 \\ k \left[\left(12 - \frac{9}{2} \right) - \left(4 - \frac{1}{2} \right) \right] &= 1 \\ k [7.5 - 3.5] &= 1 \\ 4k &= 1, \end{aligned}$$

$$\boxed{k = \frac{1}{4}} \quad \text{so} \quad f(x) = \frac{4 - x}{4}, \quad 1 \leq x \leq 3.$$

Geometric check: the region is a trapezium with parallel sides $f(1) = 3/4$ and $f(3) = 1/4$ and width 2, so its area is $\frac{1}{2}(0.75 + 0.25)(2) = 1$. ✓

(b) The sketch.

See Figure 15.3: a straight line falling from (1, 0.75) to (3, 0.25).

(c) $P(1.2 \leq X \leq 2.4)$.

$$\begin{aligned}
 P(1.2 \leq X \leq 2.4) &= \int_{1.2}^{2.4} \frac{4-x}{4} dx = \frac{1}{4} \left[4x - \frac{x^2}{2} \right]_{1.2}^{2.4} \\
 &= \frac{1}{4} \left[(9.6 - 2.88) - (4.8 - 0.72) \right] \\
 &= \frac{1}{4} [6.72 - 4.08] = \frac{2.64}{4} = \boxed{0.66}.
 \end{aligned}$$

Check by geometry.

The shaded region is a trapezium with parallel sides $f(1.2) = 2.8/4 = 0.7$ and $f(2.4) = 1.6/4 = 0.4$, and width $2.4 - 1.2 = 1.2$. Its area is $\frac{1}{2}(0.7 + 0.4)(1.2) = \frac{1}{2}(1.1)(1.2) = 0.66$. ✓

For linear densities the geometric check is faster than the integral, and worth doing every time.

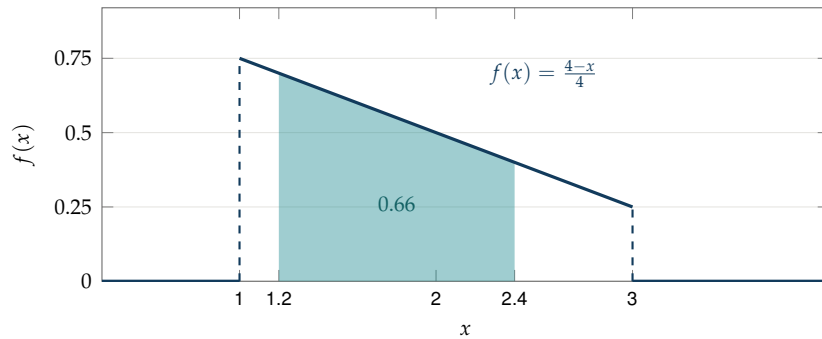


Figure 15.3: The density $f(x) = (4 - x)/4$ on $[1, 3]$. Two thirds of the probability lies in $[1.2, 2.4]$.

15.5 Expectation and variance

Definition: Expectation and variance of a continuous variable

If X is a continuous random variable with p.d.f. $f(x)$, then the expectation of X is

$$E(X) = \int x f(x) dx,$$

often denoted μ and referred to as the **mean** of X . The variance of X is

$$\text{Var}(X) = E[(X - \mu)^2] = E(X^2) - \mu^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2, \quad \text{where } \mu = E(X).$$

Every property proved in Chapter 12 — linearity, $\text{Var}(aX + b) = a^2 \text{Var}(X)$, and so on — carries over unchanged, because the proofs used only the *algebraic* structure of the sum, which the integral shares.

Example 15.2 (*Expectation and variance of a quadratic density*)

If X is a continuous random variable with p.d.f.

$$f(x) = \frac{3x^2}{64}, \quad 0 \leq x \leq 4,$$

find $E(X)$ and $\text{Var}(X)$.

Solution.

Preliminary check that f is a valid density.

$$\int_0^4 \frac{3x^2}{64} dx = \frac{3}{64} \left[\frac{x^3}{3} \right]_0^4 = \frac{3}{64} \cdot \frac{64}{3} = 1. \checkmark$$

The expectation.

$$\begin{aligned} E(X) &= \int_0^4 x \cdot \frac{3x^2}{64} dx = \frac{3}{64} \int_0^4 x^3 dx = \frac{3}{64} \left[\frac{x^4}{4} \right]_0^4 \\ &= \frac{3}{64} \cdot \frac{256}{4} = \frac{3}{64} \cdot 64 = \boxed{3}. \end{aligned}$$

The second moment.

$$\begin{aligned} E(X^2) &= \int_0^4 x^2 \cdot \frac{3x^2}{64} dx = \frac{3}{64} \int_0^4 x^4 dx = \frac{3}{64} \left[\frac{x^5}{5} \right]_0^4 \\ &= \frac{3}{64} \cdot \frac{1024}{5} = \frac{3072}{320} = 9.6. \end{aligned}$$

The variance.

$$\text{Var}(X) = E(X^2) - \mu^2 = 9.6 - 3^2 = 9.6 - 9 = \boxed{0.6},$$

$$\text{SD}(X) = \sqrt{0.6} \approx 0.775.$$

Interpretation.

The density $3x^2/64$ increases steeply, so probability is concentrated towards the right-hand end of $[0, 4]$. That is why the mean, 3, sits well above the midpoint 2 of the interval. The standard deviation of about 0.78 is small relative to the range of 4, confirming the mass is tightly bunched near the upper end.

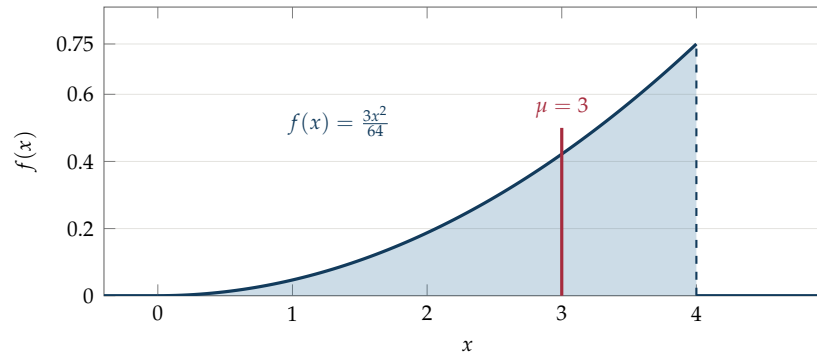


Figure 15.4: The density $3x^2/64$ on $[0, 4]$, with the mean $\mu = 3$ and the interval $\mu \pm \sigma$ marked. The rising density pulls the mean to the right of the interval's midpoint.

15.6 The cumulative distribution function

Definition: c.d.f. of a continuous variable

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt.$$

Conversely, wherever F is differentiable, $f(x) = F'(x)$: **the density is the derivative of the c.d.f.**

Property: Using the c.d.f.

$$P(a \leq X \leq b) = F(b) - F(a).$$

Example 15.3 (The c.d.f. of the triangular density)

For $f(x) = x/8$ on $[0, 4]$,

$$F(x) = \int_0^x \frac{t}{8} dt = \frac{x^2}{16}, \quad 0 \leq x \leq 4,$$

with $F(x) = 0$ for $x < 0$ and $F(x) = 1$ for $x > 4$. Check: $F(4) = 16/16 = 1$. ✓

Then

$$P(1 \leq X \leq 2.5) = F(2.5) - F(1) = \frac{6.25}{16} - \frac{1}{16} = \frac{5.25}{16} = \frac{21}{64},$$

agreeing with the direct integration above. ✓

Once you have F , every interval probability is a subtraction rather than a fresh integral — which is exactly why the Normal distribution is tabulated through its c.d.f. in Chapter 16.

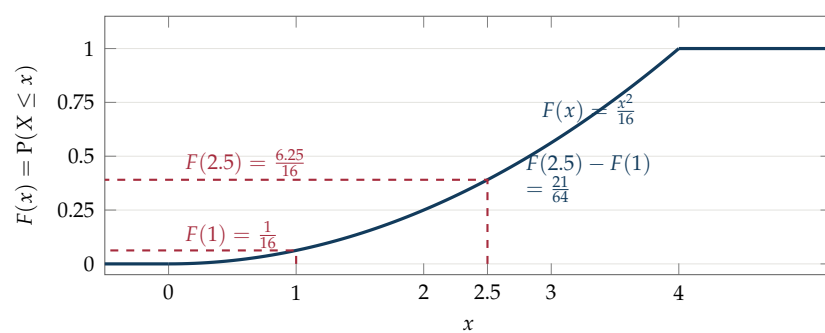


Figure 15.5: The c.d.f. corresponding to $f(x) = x/8$. Unlike the discrete case (Figure 11.4), this c.d.f. is *continuous* — no jumps, because no single value carries positive probability.

The Normal Distribution

16.1 Why it matters

The Normal distribution is the most important continuous distribution in statistics. Many measured quantities in the natural sciences follow a normal distribution — for example heights, masses, ages and examination results.

Definition: Probability density function of a normal variable

A continuous random variable X having p.d.f.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty,$$

is said to have a **normal distribution** with mean μ and variance σ^2 . The parameters of the distribution are μ and σ^2 , and we write $X \sim N(\mu, \sigma^2)$.

The distribution is **bell shaped** and **symmetrical** about $x = \mu$.

Property: Expectation and variance

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

$$E(X) = \mu \quad \text{and} \quad \text{Var}(X) = \sigma^2.$$

Intuition: Reading the formula

The formula looks forbidding but has only three moving parts.

- $(x - \mu)^2$ — the squared distance from the centre. Symmetric in $\pm$, which is why the curve is symmetric about μ .
- $\exp[- \cdot / 2\sigma^2]$ — the negative exponential makes the density fall away rapidly as you move from the centre, and σ^2 in the denominator controls *how* rapidly. Large σ , slow decay, wide curve.
- $1/(\sigma\sqrt{2\pi})$ — purely a normalising constant, chosen so the total area is exactly 1.

Two parameters, two jobs: μ slides the curve left and right, σ stretches or squeezes it.

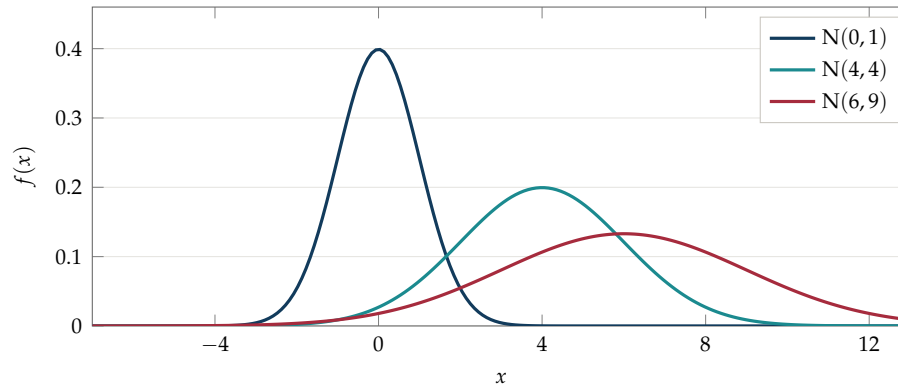


Figure 16.1: Three normal densities. μ determines where the peak sits; σ determines how wide and how tall it is. All three enclose area exactly 1, so a wider curve must be shorter.

16.2 The empirical rule

Property: The empirical rule

The area under the total curve is 1, and

- approximately **68%** of the distribution lies within $\pm 1\sigma$ of the mean;
- approximately **95%** of the distribution lies within $\pm 2\sigma$ of the mean;
- approximately **99.8%** of the distribution lies within $\pm 3\sigma$ of the mean.

Where these numbers come from

They are not new facts — they are readings from the standard normal table:

$$P(-1 < Z < 1) = 2\Phi(1) - 1 = 2(0.8413) - 1 = 0.6826,$$

$$P(-2 < Z < 2) = 2(0.9772) - 1 = 0.9544,$$

$$P(-3 < Z < 3) = 2(0.99865) - 1 = 0.9973.$$

The handout quotes the third as 99.8%; more precisely it is 99.7%. The rule is a memory aid for rough work; use the table when precision matters.

16.3 The standard normal distribution

There are infinitely many normal distributions, one for each (μ, σ^2) pair. Tabulating them all is impossible. The solution is to convert every normal variable to a single reference distribution.

Definition: Standardisation

Consider the random variable Z where

$$Z = \frac{X - \mu}{\sigma}.$$

If the random variable X has a normal distribution with mean μ and variance σ^2 , then the random variable Z has a **standard normal distribution** with mean 0 and variance 1. We write

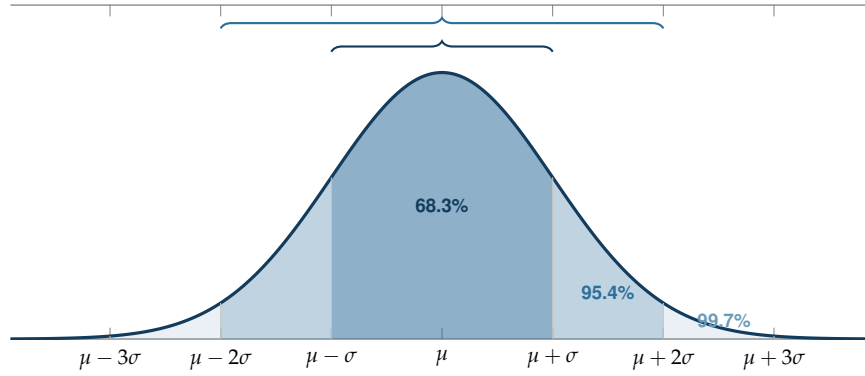


Figure 16.2: The empirical rule. Roughly two thirds of a normal population lies within one standard deviation of the mean, and virtually all of it within three.

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

Why Z has mean 0 and variance 1. By the linearity results of Chapter 12, with $a = 1/\sigma$ and $b = -\mu/\sigma$:

$$E(Z) = \frac{E(X) - \mu}{\sigma} = \frac{\mu - \mu}{\sigma} = 0,$$

$$\text{Var}(Z) = \frac{1}{\sigma^2} \text{Var}(X) = \frac{\sigma^2}{\sigma^2} = 1. \quad \square$$

Intuition: What a z-score means

z counts **how many standard deviations x lies above the mean**. A z -score of $+1.5$ means “one and a half standard deviations above average”, whatever the original units were.

This makes z -scores the natural way to compare across different scales. A student scoring $z = 1.2$ in Statistics and $z = 0.4$ in Physics performed relatively better in Statistics, regardless of the raw marks or the difficulty of the two papers.

Definition: The p.d.f. of Z

The p.d.f. of the standard normal variable is denoted $\phi(z)$, where

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

obtained by setting $\mu = 0$ and $\sigma = 1$ in the general formula.

Definition: The cumulative distribution function

The c.d.f. of Z is

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

This integral cannot be evaluated in closed form, so we refer to tables. See Appendix B.

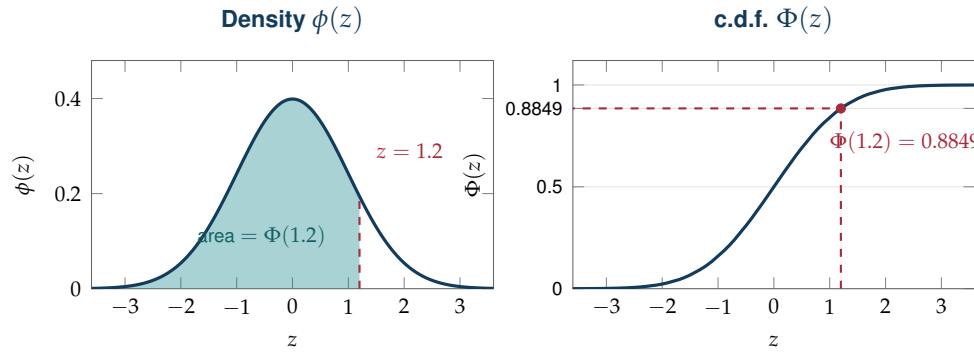


Figure 16.3: The standard normal density and its cumulative distribution function. $\Phi(z)$ is the *area* to the left of z under the density — the table in Appendix B lists these areas.

16.4 Using the table

Recipe: The standardisation procedure

To find $P(a < X < b)$ when $X \sim N(\mu, \sigma^2)$:

1. **Draw** the curve and shade the region you want. This step prevents more errors than any other.
2. **Standardise** both limits: $z_1 = \frac{a - \mu}{\sigma}$, $z_2 = \frac{b - \mu}{\sigma}$.
3. **Look up** $\Phi(z_1)$ and $\Phi(z_2)$.
4. **Subtract:** $P(a < X < b) = \Phi(z_2) - \Phi(z_1)$.

Property: The four table rules

Want	Use	Reason
$P(Z < z)$	$\Phi(z)$	direct from the table
$P(Z > z)$	$1 - \Phi(z)$	complement rule
$P(Z < -z)$	$1 - \Phi(z)$	symmetry about 0
$P(z_1 < Z < z_2)$	$\Phi(z_2) - \Phi(z_1)$	difference of areas

Intuition: Why $\Phi(-z) = 1 - \Phi(z)$

The curve is symmetric about 0, so the area to the left of $-z$ is a mirror image of the area to the right of $+z$. Since the areas left and right of $+z$ must sum to 1, the area right of z is $1 - \Phi(z)$ — and therefore so is the area left of $-z$. This is why tables only need to list positive z .

Example 16.1 (Standard normal lookups)

Given $Z \sim N(0, 1)$, find each of the following.

Required	Working	Answer
$P(Z < 1.25)$	$\Phi(1.25)$	0.8944
$P(Z > 1.25)$	$1 - 0.8944$	0.1056
$P(Z < -1.25)$	$1 - \Phi(1.25) = 1 - 0.8944$	0.1056
$P(-1.25 < Z < 1.25)$	$0.8944 - 0.1056$	0.7888
$P(0.5 < Z < 1.5)$	$\Phi(1.5) - \Phi(0.5) = 0.9332 - 0.6915$	0.2417

Example 16.2 (*Heights of students*)

The heights of a large group of students are normally distributed with mean $\mu = 168$ cm and standard deviation $\sigma = 8$ cm. Find the probability that a randomly chosen student is (a) shorter than 180 cm; (b) taller than 155 cm; (c) between 160 and 176 cm.

Solution.

(a) Shorter than 180 cm.

$$z = \frac{180 - 168}{8} = \frac{12}{8} = 1.5,$$

$$P(X < 180) = \Phi(1.5) = 0.9332.$$

(b) Taller than 155 cm.

$$z = \frac{155 - 168}{8} = \frac{-13}{8} = -1.625 \approx -1.63,$$

$$P(X > 155) = P(Z > -1.63) = \Phi(1.63) = 0.9484.$$

(c) Between 160 and 176 cm.

$$z_1 = \frac{160 - 168}{8} = -1, \quad z_2 = \frac{176 - 168}{8} = +1,$$

$$P(160 < X < 176) = \Phi(1) - \Phi(-1) = 0.8413 - 0.1587 = 0.6826.$$

This is the empirical rule's 68%, arrived at by calculation — the interval 160 to 176 is exactly $\mu \pm \sigma$. ✓

Example 16.3 (*Working backwards: finding x from a probability*)

Using the same height distribution, the tallest 10% of students are to be selected for a basketball trial. What is the minimum height required?

Solution. We need x such that $P(X > x) = 0.10$, equivalently $P(X < x) = 0.90$.

Find the z -value.

Search the *body* of the table for 0.90. The closest entry is $\Phi(1.28) = 0.8997$, so $z \approx 1.28$.

Convert back to x .

Rearranging $z = (x - \mu)/\sigma$ gives $x = \mu + z\sigma$:

$$x = 168 + 1.28 \times 8 = 168 + 10.24 = 178.2 \text{ cm.}$$

Students of 178.2 cm or taller qualify.

Note on direction.

Forward problems (given x , find a probability) go $x \rightarrow z \rightarrow \Phi$. Inverse problems (given a probability, find x) go $\Phi \rightarrow z \rightarrow x$, and you search the *inside* of the table rather than its margins.

16.5 The Normal approximation to the Binomial

Theorem: Normal approximation to the Binomial

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

$$X \sim N(np, np(1-p)).$$

The usual condition is $np > 5$ and $n(1-p) > 5$.

Intuition: Why it works

A Binomial variable is a sum of n independent trials. The Central Limit Theorem says that sums of many independent contributions tend towards a normal shape regardless of what the individual contributions look like. Figure 13.1 already showed $\text{Bin}(10, 0.5)$ looking distinctly bell-like; by $n = 100$ the resemblance is very close.

The condition $np > 5$ and $n(1-p) > 5$ ensures the distribution is not pressed up against either boundary 0 or n , where the Binomial is skewed and the symmetric normal fits badly.

16.5.1 The continuity correction

Always apply the continuity correction

A Binomial or Poisson variable takes only whole-number values; a normal variable takes any real value. Each integer k is therefore represented by the interval $(k - 0.5, k + 0.5)$.

Discrete event	Continuous interval
$X = k$	$k - 0.5 < X < k + 0.5$
$X \leq k$	$X < k + 0.5$
$X < k$	$X < k - 0.5$
$X \geq k$	$X > k - 0.5$
$X > k$	$X > k + 0.5$
$a \leq X \leq b$	$a - 0.5 < X < b + 0.5$

The rule of thumb: **expand the interval by 0.5 at each end that is included; shrink it by 0.5 at each end that is excluded.**

Example 16.4 (Approximating a Binomial probability)

A fair coin is tossed 100 times. Find the probability of obtaining between 45 and 55 heads inclusive.

Solution. $X \sim \text{Bin}(100, 0.5)$. Since $np = 50 > 5$ and $n(1 - p) = 50 > 5$, use

$$X \approx N(\mu = 50, \sigma^2 = 25), \quad \sigma = 5.$$

Continuity correction.

$45 \leq X \leq 55$ becomes $44.5 < X < 55.5$.

Standardise.

$$z_1 = \frac{44.5 - 50}{5} = -1.1, \quad z_2 = \frac{55.5 - 50}{5} = +1.1.$$

Look up and subtract.

$$P(44.5 < X < 55.5) = \Phi(1.1) - \Phi(-1.1) = 0.8643 - 0.1357 = 0.7286.$$

Accuracy.

The exact Binomial value is 0.7287 — the approximation is correct to four decimal places. Without the continuity correction we would have used $z = \pm 1.0$ and obtained 0.6826, an error of 6%.

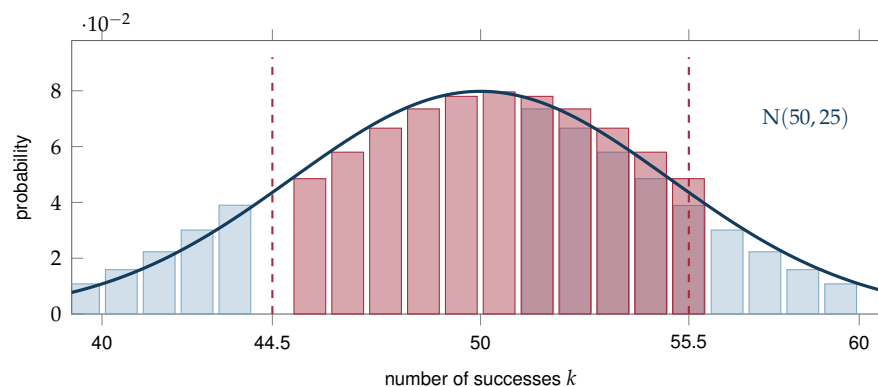


Figure 16.4: $\text{Bin}(100, 0.5)$ with the $N(50, 25)$ density overlaid. Because each bar has width 1, the shaded bars for $45 \leq k \leq 55$ occupy exactly 44.5 to 55.5 — which is what the continuity correction encodes.

16.6 Choosing the right distribution

Distribution	Type	Parameters	Mean	Variance
$\text{Bin}(n, p)$	discrete	n, p	np	$np(1 - p)$
$\text{Poisson}(\lambda)$	discrete	λ	λ	λ
$N(\mu, \sigma^2)$	continuous	μ, σ^2	μ	σ^2

Common pitfall: Three errors that cost marks every year

1. **Forgetting the continuity correction** when approximating a discrete distribution by the Normal. This is the most frequently lost mark in the whole course.

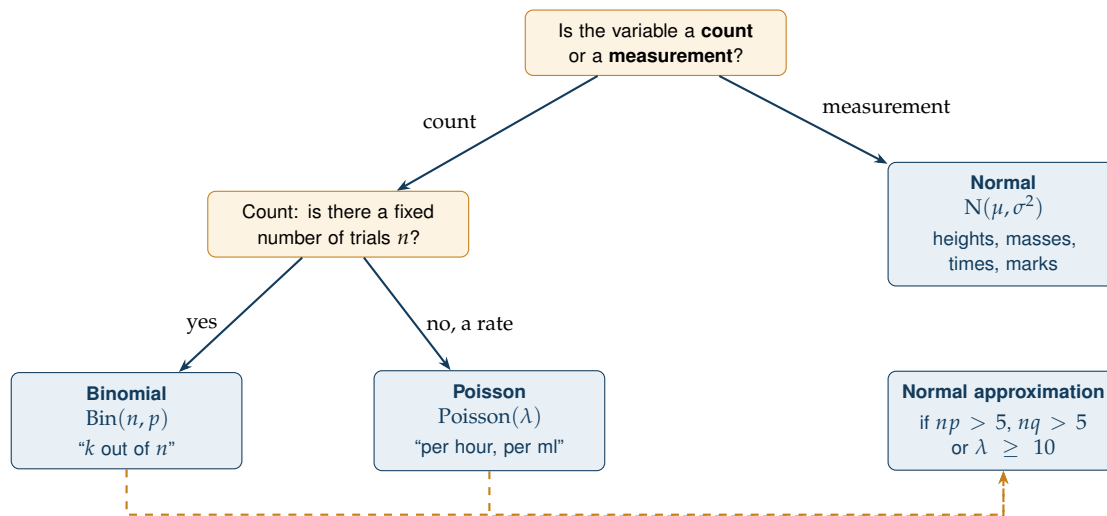


Figure 16.5: A decision tree for the three distributions in the syllabus. The dashed arrows are the large-parameter approximations of Sections 14.6 and 16.5.

2. **Confusing σ with σ^2 .** The notation $N(\mu, \sigma^2)$ names the *variance*, but the standardisation formula divides by the *standard deviation*. $N(50, 25)$ has $\sigma = 5$, not 25.
3. **Not drawing the curve.** A quick sketch with the region shaded makes it immediately obvious whether you want $\Phi(z)$ or $1 - \Phi(z)$. Skipping it is a false economy.

PART IV

Sampling and Further Distributions

Linear Combinations of Normal Variables

17.1 Why this chapter exists

Chapter 16 told you everything about *one* normal variable. Real problems rarely involve one. A journey has an outward leg and a return leg; a total cost is a sum of component costs; a profit is a difference between revenue and expenditure. This chapter answers the question: **if X and Y are normal, what is the distribution of $X + Y$, of $X - Y$, of $3X$?**

The answer is remarkably clean, and it rests on a property that is special to the normal distribution.

The closure property

The normal family is **closed** under linear combinations: any sum, difference or multiple of independent normal variables is *itself normal*. Nothing else in this course behaves so well. Add two Binomials with different p and you do not get a Binomial; add two uniforms and you get a triangle. Add two normals and you get a normal.

This is one of the main reasons the normal distribution dominates statistical practice.

17.2 Sum and difference of two independent normal variables

Theorem: Sum and difference

If X and Y are two **independent** normal variables such that

$$X \sim N(\mu_1, \sigma_1^2) \quad \text{and} \quad Y \sim N(\mu_2, \sigma_2^2),$$

then

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

$$X - Y \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

Where the parameters come from. The *normality* of the result is a genuine theorem requiring calculus beyond this course. But the two parameters follow immediately from Chapter 12.

For the mean, linearity of expectation (which needs no independence) gives

$$E(X \pm Y) = E(X) \pm E(Y) = \mu_1 \pm \mu_2.$$

For the variance, independence lets variances add, and $\text{Var}(-Y) = (-1)^2 \text{Var}(Y) = \text{Var}(Y)$, so

$$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y) = \sigma_1^2 + \sigma_2^2$$

in *both* cases. □

Common pitfall: The variances add even for a difference

Look at the two boxed lines again. The means subtract, but the variances **add** in both. Writing $\sigma_1^2 - \sigma_2^2$ for the difference is the most common error in this chapter, and it can produce a negative “variance” — an immediate signal that something has gone wrong.

The reason was given in Chapter 12: variance measures uncertainty, and combining two uncertain quantities can only produce something *more* uncertain, never less. Subtracting does not cancel randomness; it compounds it.

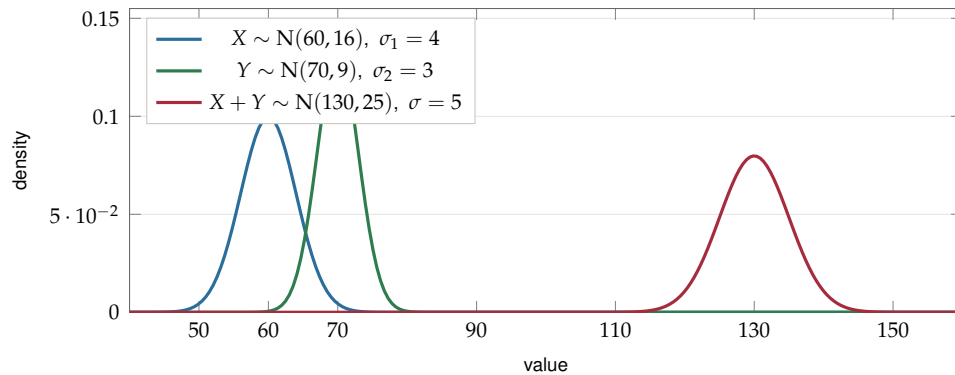


Figure 17.1: Adding two independent normals. The means add ($60 + 70 = 130$) and the *variances* add ($16 + 9 = 25$), so the new standard deviation is $\sqrt{25} = 5$ — *not* $4 + 3 = 7$. Standard deviations never add; variances do.

Pythagoras for standard deviations

Since $\sigma^2 = \sigma_1^2 + \sigma_2^2$, we get

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}.$$

This is exactly the hypotenuse of a right triangle with legs σ_1 and σ_2 . In the figure above, $\sigma_1 = 4$ and $\sigma_2 = 3$ give $\sigma = 5$ — the 3–4–5 triangle. Combining independent uncertainties is a *geometric* operation, not an arithmetic one.

17.3 Worked examples**Example 17.1** (*Sums and differences of two normals*)

If $X \sim N(60, 16)$ and $Y \sim N(70, 9)$, find (a) $P(X + Y < 140)$; (b) $P(120 < X + Y < 135)$; (c) $P(Y - X > 7)$.

Solution.

(a) $P(X + Y < 140)$.

First identify the distribution of the sum:

$$(X + Y) \sim N(60 + 70, 16 + 9), \quad \text{i.e.} \quad R = X + Y \sim N(130, 25), \quad \sigma = 5.$$

We require

$$P(R < 140) = P\left(Z < \frac{140 - 130}{5}\right) = P(Z < 2) = \boxed{0.9772}.$$

(b) $P(120 < X + Y < 135)$.

Using the same $R \sim N(130, 25)$:

$$z_1 = \frac{120 - 130}{5} = -2, \quad z_2 = \frac{135 - 130}{5} = +1,$$

$$P(120 < R < 135) = P(-2 < Z < 1) = \Phi(1) - \Phi(-2) = 0.8413 - 0.0228 = \boxed{0.8185}.$$

(c) $P(Y - X > 7)$.

Now form the *difference* — note the order: $Y - X$, so the means subtract in that order, but the variances still add.

$$(Y - X) \sim N(70 - 60, 9 + 16), \quad \text{i.e.} \quad T = Y - X \sim N(10, 25), \quad \sigma = 5.$$

$$P(Y - X > 7) = P(T > 7) = P\left(Z > \frac{7 - 10}{5}\right) = P(Z > -0.6).$$

By symmetry, $P(Z > -0.6) = \Phi(0.6) = \boxed{0.7257}$.

Sense check on (c).

The mean of $Y - X$ is 10, and we are asking for the probability that it exceeds 7 — a value *below* the mean. The answer must therefore exceed 0.5, and 0.7257 does. ✓

17.4 Multiples of normal variables

Theorem: Multiples

If X is a normal variable such that $X \sim N(\mu, \sigma^2)$, then for any constant a ,

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

This is Chapter 12's scaling rules — $E(aX) = a E(X)$ and $\text{Var}(aX) = a^2 \text{Var}(X)$ — combined with the fact that scaling a normal keeps it normal.

Common pitfall: $3X$ is not the same as $X_1 + X_2 + X_3$

This distinction is worth real care.

Quantity	Meaning	Variance
$3X$	one observation, tripled	$9\sigma^2$
$X_1 + X_2 + X_3$	three independent observations, added	$3\sigma^2$

Both have mean 3μ , but tripling a single measurement triples its error too, whereas three separate measurements partially cancel one another. For $\sigma = 5$: $\text{SD}(3X) = 15$ but $\text{SD}(X_1 + X_2 + X_3) = 5\sqrt{3} \approx 8.66$.

Read the problem carefully. “Three times the weight of one box” is $3X$; “the total weight of three boxes” is $X_1 + X_2 + X_3$.

Example 17.2 (A multiple of a normal variable)

If $X \sim N(50, 25)$, find $P(3X > 160)$.

Solution. Scaling by $a = 3$:

$$3X \sim N(3 \times 50, 3^2 \times 25) = N(150, 225), \quad \sigma = 15.$$

$$P(3X > 160) = P\left(Z > \frac{160 - 150}{15}\right) = P(Z > 0.667) = 1 - \Phi(0.67) = 1 - 0.7486 = \boxed{0.2523}.$$

An alternative route.

You could instead divide through by 3 first:

$$P(3X > 160) = P\left(X > \frac{160}{3}\right) = P(X > 53.33) = P\left(Z > \frac{53.33 - 50}{5}\right) = P(Z > 0.667). \checkmark$$

Both give the same z , as they must — the z -score is unchanged by rescaling, which is precisely what makes standardisation work.

17.5 The general linear combination

Putting the two theorems together gives the result you should actually memorise.

Theorem: General linear combination

If $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are independent, and a, b, c are constants, then

$$aX + bY + c \sim N(a\mu_1 + b\mu_2 + c, a^2\sigma_1^2 + b^2\sigma_2^2).$$

Everything in one line

$$\text{mean: } a\mu_1 + b\mu_2 + c \qquad \text{variance: } a^2\sigma_1^2 + b^2\sigma_2^2$$

Note the asymmetry that causes all the trouble: the constant c shifts the mean but *vanishes* from the variance, and the coefficients enter the mean linearly but the variance *squared* — so a negative b subtracts from the mean while still adding to the variance.

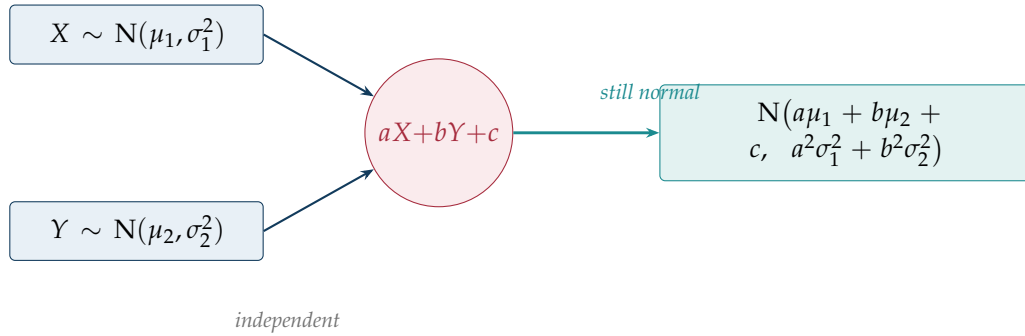


Figure 17.2: The closure property in schematic form. Any linear combination of independent normals is normal, with parameters given by the two rules above.

17.6 An extended exercise

Exercise 17.1 (*Mr Jones at the library*)

Each weekday Mr Jones walks to the local library to read the newspapers. The time he takes to walk to and from the library is a normal variable with mean 15 minutes and standard deviation 2 minutes. The time he spends in the library is a normal variable with mean 25 minutes and standard deviation $\sqrt{12}$ minutes. Find the probability that, on a particular day,

- (i) Mr Jones is away from the house for more than 45 minutes;
- (ii) Mr Jones spends more time travelling than in the library.

Solution.

Set up the variables.

Let X be the time he takes to walk to and from the library, and Y the time he spends in the library. Converting the standard deviations to variances:

$$\begin{aligned} X &\sim N(15, 4) & (\sigma_X = 2 \Rightarrow \sigma_X^2 = 4), \\ Y &\sim N(25, 12) & (\sigma_Y = \sqrt{12} \Rightarrow \sigma_Y^2 = 12). \end{aligned}$$

(i) Away from the house for more than 45 minutes.

“Away from the house” is the *total* time, travelling plus reading:

$$X + Y \sim N(15 + 25, 4 + 12) = N(40, 16), \quad \sigma = 4.$$

$$P(X + Y > 45) = P\left(Z > \frac{45 - 40}{4}\right) = P(Z > 1.25) = 1 - \Phi(1.25) = 1 - 0.8944 = \boxed{0.1056}.$$

So on about 11% of weekdays he is out for more than three quarters of an hour.

(ii) More time travelling than in the library.

The event “ $X > Y$ ” cannot be handled directly, but it can be rewritten as a statement about a *single* variable:

$$P(X > Y) = P(X - Y > 0).$$

Now form the difference — means subtract, variances add:

$$X - Y \sim N(15 - 25, 4 + 12) = N(-10, 16), \quad \sigma = 4.$$

$$P(X - Y > 0) = P\left(Z > \frac{0 - (-10)}{4}\right) = P(Z > 2.5) = 1 - \Phi(2.5) = 1 - 0.9938 = \boxed{0.0062}.$$

Interpretation.

Only about 0.6% of days — roughly one weekday in 160, or about three times a year. That is intuitively right: he averages 25 minutes reading against 15 minutes walking, so travelling would have to be unusually long *and* reading unusually short on the same day.

The technique worth extracting.

Comparisons between two random variables are **always** converted into a statement about their difference:

$$\begin{aligned} P(X > Y) &= P(X - Y > 0), \\ P(X < Y) &= P(X - Y < 0), \\ P(|X - Y| < d) &= P(-d < X - Y < d). \end{aligned}$$

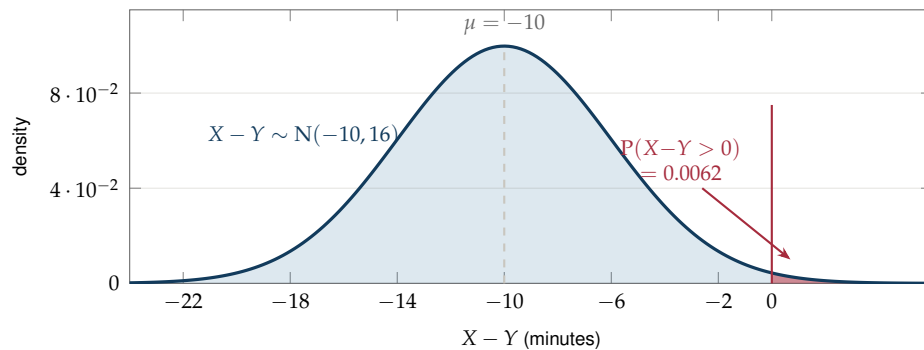


Figure 17.3: Mr Jones, part (ii). The difference $X - Y$ is centred on -10 minutes, so the event $X > Y$ sits 2.5 standard deviations into the right tail.

Recipe: Handling any linear-combination problem

1. **Name** each underlying variable and write down its distribution. Convert any standard deviation you are given into a *variance*.
2. **Translate** the question into a single new variable — a sum, a difference, or a multiple.
3. **Combine** the parameters: means combine linearly (with signs); variances combine with *squared* coefficients and *always add*.
4. **Standardise** the new variable and use the table exactly as in Chapter 16.

Step 1 is where most marks are lost: the problem usually states a standard deviation, while every combination rule is written in terms of variances.

Sampling Distributions and the Central Limit Theorem

18.1 Closing the loop

Chapter 1 posed the central problem of statistics: you measure a sample and want to say something about a population you did not measure. Everything since has been building the machinery. This chapter delivers the answer.

The key realisation is this. A statistic such as $\bar{X}$ is computed from randomly chosen data, so **it is itself a random variable**. Draw a different sample and you get a different $\bar{x}$. That variability has a distribution — called a *sampling distribution* — and once we know it, we can quantify exactly how far a sample mean is likely to stray from the population mean.

Definition: Sampling distribution

The **sampling distribution** of a statistic is the probability distribution of that statistic over all possible samples of a given size n drawn from the population.

Intuition: Two different distributions, easily confused

Keep these apart at all costs.

	Population distribution	Sampling distribution of $\bar{X}$
Describes	individual observations X	sample means $\bar{X}$
Mean	μ	μ (the same)
Variance	σ^2	σ^2/n (smaller)
Shape	anything at all	normal, for large n

The two share a centre but not a spread, and for large n the second is normal regardless of what the first looks like.

18.2 The distribution of the sample mean

Theorem: Sample mean from a normal population

If $X_1, X_2, \dots, X_n$ is a random sample of size n taken from a normal distribution with mean μ and variance σ^2 , such that $X \sim N(\mu, \sigma^2)$, then the distribution of $\bar{X}$ is **also normal**, with mean μ and variance σ^2/n :

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

Proof. Write $\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$. Each X_i is normal, and by the closure property of Chapter 17 the sum of independent normals is normal, as is any multiple of one. So $\bar{X}$ is normal. Its parameters follow from Chapter 12:

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n}(n\mu) = \mu,$$

and, using independence so that variances add, together with $\text{Var}(aX) = a^2 \text{Var}(X)$ with $a = 1/n$,

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}. \quad \square$$

Definition: Standard error

The standard deviation of the sampling distribution of $\bar{X}$ is called the **standard error of the mean**:

$$\text{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

The $\sqrt{n}$ law

The standard error falls as $\sigma/\sqrt{n}$, *not* as σ/n . This has a blunt practical consequence:

Sample size n	25	100	400	1600
Standard error (for $\sigma = 10$)	2.00	1.00	0.50	0.25

To halve your uncertainty you must quadruple your sample. Precision is expensive, and it gets more expensive the more of it you already have. This single fact governs the economics of survey design, clinical trials and quality control.

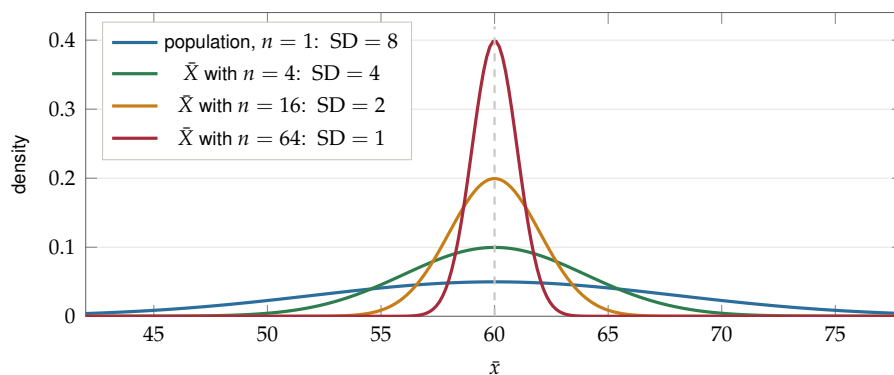


Figure 18.1: Sampling distributions of $\bar{X}$ for a $N(60, 64)$ population. All are centred on $\mu = 60$; each quadrupling of n halves the spread. Larger samples give means that cluster ever more tightly around the truth.

Example 18.1 (Sample mean from a normal population)

A random sample of size 15 is taken from a normal distribution with mean 60 and standard deviation 4. Find the probability that the mean of the sample is less than 58.

Solution.

The population.

$$X \sim N(60, 16) \quad (\sigma = 4 \Rightarrow \sigma^2 = 16).$$

The sampling distribution.

For a sample of size $n = 15$,

$$\bar{X} \sim N\left(60, \frac{16}{15}\right), \quad \text{SD}(\bar{X}) = \frac{4}{\sqrt{15}} \approx 1.0328.$$

Standardise and read the table.

$$P(\bar{X} < 58) = P\left(Z < \frac{58 - 60}{4/\sqrt{15}}\right) = P\left(Z < \frac{-2}{1.0328}\right) = P(Z < -1.936).$$

By symmetry, $P(Z < -1.936) = 1 - \Phi(1.94) = 1 - 0.9738 = \boxed{0.0264}.$

Contrast with a single observation.

For one individual, $P(X < 58) = P(Z < -0.5) = 0.3085$ — nearly one chance in three. For the mean of fifteen, it is one chance in thirty-eight. Averaging has made an outcome that is unremarkable for an individual quite unusual for a group. This is the whole reason samples are informative.

18.3 The Central Limit Theorem

The previous theorem required the population to be normal. The Central Limit Theorem removes that requirement, and it is the most important result in the subject.

Theorem: Central Limit Theorem

If $X_1, X_2, \dots, X_n$ is a random sample of size n taken from **any** distribution with mean μ and variance σ^2 , then for large n the distribution of the sample mean is **approximately normal** with mean μ and variance σ^2/n :

$$\bar{X} \overset{\text{approx}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right).$$

Note: the approximation gets better as n gets larger.

Why this is extraordinary

Read the statement again and notice what it does *not* say. It does not say the population must be normal, or symmetric, or continuous, or bell-shaped. It says **any** distribution with a finite mean and variance — skewed, discrete, bimodal, whatever — produces approximately normal sample means.

This is why the normal distribution appears everywhere in nature. Any quantity that is effectively an average or a sum of many small independent contributions — human height, measurement error, examination totals — will be approximately normal, no matter what the individual contributions look like.

It is also why the normal table in Appendix B is enough to handle problems about populations whose distribution you do not know at all.

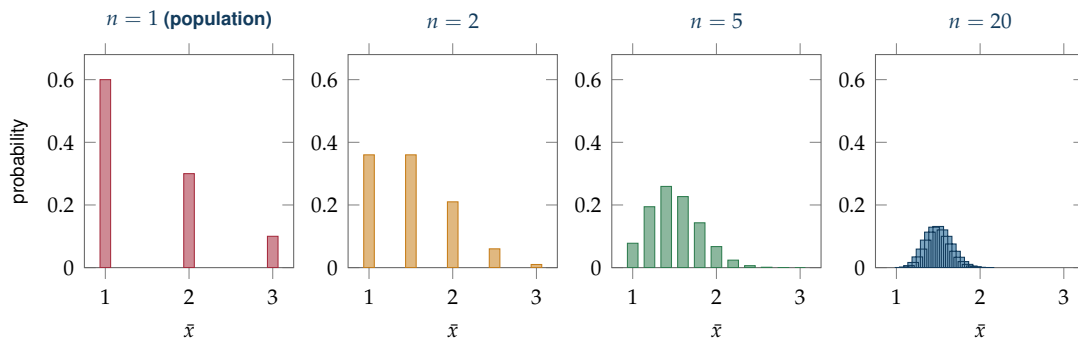


Figure 18.2: The Central Limit Theorem in action. The population (left) is strongly right-skewed and takes only three values: $P(1) = 0.6$, $P(2) = 0.3$, $P(3) = 0.1$, with $\mu = 1.5$ and $\sigma = 0.671$. As n grows the distribution of $\bar{X}$ becomes symmetric and bell-shaped, and its spread shrinks as $\sigma/\sqrt{n}$: from 0.671 to 0.474 to 0.300 to 0.150.

Common pitfall: How large is “large”?

The conventional threshold is $n \geq 30$, but it is a rule of thumb, not a theorem.

- If the population is already *normal*, the result is exact for *every* n , even $n = 1$. No approximation is involved.
- If the population is roughly symmetric, $n \approx 10$ is usually plenty — look at the $n = 5$ panel above.
- If the population is severely skewed or has heavy tails, $n = 30$ may not be enough; $n = 50$ or more may be required.

Note also what the theorem does *not* claim: the *individual* observations do not become normal, however large n is. Only the *average* does.

Example 18.2 (CLT applied to a Binomial population)

If a random sample of size 30 is taken from the $\text{Bin}(9, 0.5)$ distribution, find the probability that the sample mean exceeds 5.

Solution.

Population parameters.

Here $X \sim \text{Bin}(9, 0.5)$, a *discrete* population. From Chapter 13,

$$\mu = np = 9 \times 0.5 = 4.5, \quad \sigma^2 = npq = 9 \times 0.5 \times 0.5 = 2.25.$$

Apply the CLT.

The population is Binomial, not normal — so the exact theorem does not apply, but with $n = 30$ the Central Limit Theorem does. Hence

$$\bar{X} \stackrel{\text{approx}}{\sim} N\left(4.5, \frac{2.25}{30}\right) = N(4.5, 0.075), \quad \text{SD}(\bar{X}) = \sqrt{0.075} \approx 0.2739.$$

Standardise.

$$P(\bar{X} > 5) = P\left(Z > \frac{5 - 4.5}{0.2739}\right) = P(Z > 1.826) = 1 - \Phi(1.83) = 1 - 0.9664 = \boxed{0.0340}.$$

Two things to notice.

First, **no continuity correction is used here**. The correction of Chapter 16 applies when a normal replaces a discrete distribution for the *count itself*. Here we are modelling a *mean of thirty counts*, which can take many closely spaced values $(0, 1/30, 2/30, \dots)$, so the discreteness is negligible. Second, the population $\text{Bin}(9, 0.5)$ is symmetric, which is why $n = 30$ gives an excellent approximation.

18.4 Two theorems, one picture

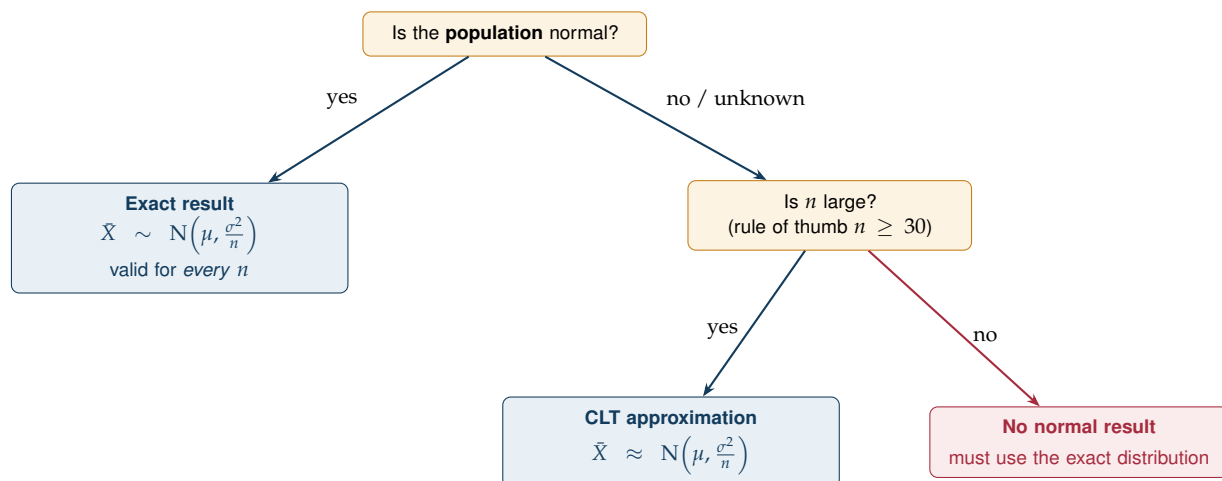


Figure 18.3: Choosing between the exact sampling-distribution result and the Central Limit Theorem. Both give the same formula — the difference is whether it is exact or approximate.

Recipe: Any sample-mean problem in four steps

1. Write down the **population** mean μ and variance σ^2 . If the population is Binomial or Poisson, get these from the formulas of Chapters 13–14 rather than being given them directly.
2. Form the sampling distribution: $\bar{X} \sim N(\mu, \sigma^2/n)$, exactly if the population is normal, approximately by the CLT otherwise.
3. Compute the standard error $\sigma/\sqrt{n}$ — *this*, not σ , is the denominator of the z-score.
4. Standardise: $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$, and read the table.

Common pitfall: Dividing by σ instead of $\sigma / \sqrt{n}$

This is the defining error of the chapter. If a question asks about a *sample mean* and your z-score uses σ alone, you have answered a question about a single individual instead.

A quick diagnostic: a z-score for a sample mean should generally be *larger* in magnitude than the corresponding one for an individual, because $\sigma / \sqrt{n} < \sigma$. In the size-15 example above, an individual gave $z = -0.5$ while the mean gave $z = -1.94$. If your sample-mean z comes out smaller than the individual one, check the denominator.

Further Discrete Distributions

19.1 Four more models

Chapter 13 handled the question “how many successes in n fixed trials?” and Chapter 14 “how many events in an interval?”. This chapter completes the discrete toolkit with four more distributions, each answering a different question about the same underlying idea of repeated trials.

Distribution	The question it answers	Section
Discrete Uniform	“which of k equally likely outcomes?”	19.2
Geometric	“how many trials until the <i>first</i> success?”	19.3
Negative Binomial	“how many trials until the <i>kth</i> success?”	19.4
Hypergeometric	“how many successes when sampling <i>without replacement</i> ?”	19.5

Intuition: The organising idea

Binomial, Geometric and Negative Binomial all describe the *same* sequence of independent trials — they differ only in what you decide to **fix** and what you let vary.

	Fixed in advance	Random
Binomial	the number of trials n	the number of successes
Geometric	one success	the number of trials
Negative Binomial	k successes	the number of trials

The Hypergeometric is the odd one out: it is the Binomial with the independence assumption removed.

19.2 The Discrete Uniform distribution

Definition: Discrete Uniform distribution

Take the sample space $S = \{1, 2, 3, \dots, k\}$ with a probability measure that assigns **equal** probability to all outcomes — that is, all outcomes are equally likely. Define the random variable X by $X(i) = i$ for $i = 1, 2, \dots, k$. Then

$$P(X = x) = \frac{1}{k}, \quad x = 1, 2, \dots, k.$$

This is the formal version of the symmetry principle from Chapter 6. A fair die is the discrete uniform distribution with $k = 6$.

Property: Moments of the Uniform distribution

$$\text{Mean } \mu = \frac{1 + 2 + \cdots + k}{k} = \frac{\frac{1}{2}k(k+1)}{k} = \frac{k+1}{2}, \quad \text{Variance } \sigma^2 = \frac{k^2 - 1}{12}.$$

Proof. The mean uses the standard sum $1 + 2 + \cdots + k = \frac{1}{2}k(k+1)$, as displayed. For the variance, use $1^2 + 2^2 + \cdots + k^2 = \frac{1}{6}k(k+1)(2k+1)$:

$$\begin{aligned} E(X^2) &= \frac{1}{k} \cdot \frac{k(k+1)(2k+1)}{6} = \frac{(k+1)(2k+1)}{6}, \\ \text{Var}(X) &= \frac{(k+1)(2k+1)}{6} - \left(\frac{k+1}{2}\right)^2 = \frac{(k+1)[2(2k+1) - 3(k+1)]}{12} \\ &= \frac{(k+1)(4k+2-3k-3)}{12} = \frac{(k+1)(k-1)}{12} = \frac{k^2 - 1}{12}. \end{aligned}$$

□

Example 19.1 (A fair die)

For $k = 6$:

$$\mu = \frac{6+1}{2} = 3.5, \quad \sigma^2 = \frac{36-1}{12} = \frac{35}{12} \approx 2.917, \quad \sigma \approx 1.708.$$

The mean of 3.5 is the familiar illustration that an expected value need not be an attainable value.

19.3 The Geometric distribution**Definition: Geometric distribution**

Consider a sequence of independent and identically distributed (iid) Bernoulli trials with $P(\{S\}) = p$. The variable of interest is the **number of trials that has to be performed until the first success**.

Random variable: X is the number of trials on which the first success occurs.

Distribution: for $X = x$ there must be a run of $(x-1)$ failures followed by a success, so

$$P(X = x) = p(1-p)^{x-1}, \quad x = 1, 2, 3, \dots; \quad 0 < p < 1.$$

Intuition: Reading the formula off the sequence

The event $X = x$ corresponds to exactly one sequence:

$$\underbrace{F F \cdots F}_{x-1 \text{ failures}} S.$$

By independence its probability is $(1-p)^{x-1} \times p$. Unlike the Binomial there is **no combinatorial coefficient**, because only one arrangement is possible — the success must come last, by definition.

Property: Moments of the Geometric distribution

$$\text{Mean } \mu = \frac{1}{p}, \quad \text{Variance } \sigma^2 = \frac{1-p}{p^2}.$$

The mean is obvious once stated

If an event happens one time in five ($p = 0.2$), you expect to wait $1/0.2 = 5$ trials for it. If a coin is fair, you expect $1/0.5 = 2$ tosses to see the first head. The mean of a geometric is simply the reciprocal of the success probability — worth remembering because it needs no calculation. Note also that the probabilities *always* decrease: $P(X = 1) = p$ is the largest value, whatever p is. The single most likely waiting time is always one trial.

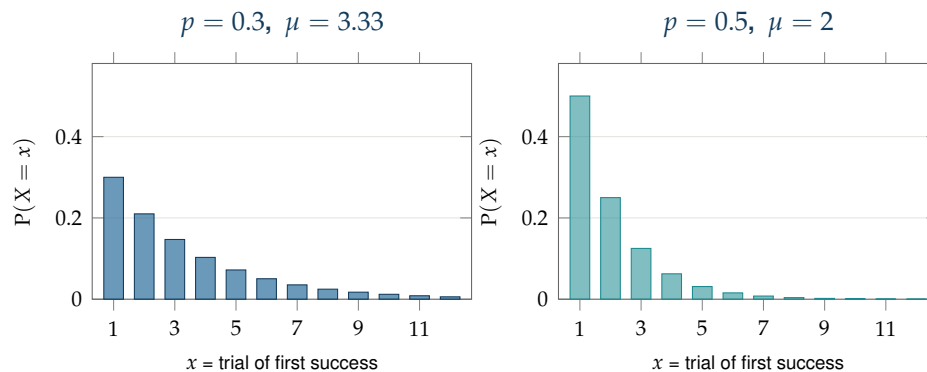


Figure 19.1: The Geometric distribution always decays: each bar is a factor $(1 - p)$ smaller than the one before it. A larger p means a faster decay and a shorter expected wait.

Example 19.2 (The fourth child is the first son)

Suppose that the probability of having a male or a female child is equal. Find the probability that a family's fourth child is their first son.

Solution. Let a “success” be the birth of a son, so $p = \frac{1}{2}$, and let X be the number of the child on which the first son occurs. Then X is geometric, and we require $P(X = 4)$: three daughters followed by a son.

$$P(X = 4) = p(1 - p)^{4-1} = \left(\frac{1}{2}\right) \left(\frac{1}{2}\right)^3 = \left(\frac{1}{2}\right)^4 = \frac{1}{16} = \boxed{0.0625}.$$

Common pitfall: The memoryless property

The Geometric distribution has a striking feature:

$$P(X > m + n \mid X > m) = P(X > n).$$

Having already waited m trials without a success tells you *nothing* about how much longer you must wait. The trials are independent, so the process has no memory of past failures. This is why the “gambler’s fallacy” — believing that a run of failures makes a success “due” — is wrong. A coin that has landed tails ten times running is still exactly as likely to land heads next toss as it ever was.

19.4 The Negative Binomial distribution

Definition: Negative Binomial distribution

This is a generalisation of the geometric distribution.

Random variable: X is the number of trials on which the k th success occurs, where k is a positive integer.

Distribution:

$$P(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \quad x = k, k+1, \dots; \quad 0 < p < 1.$$

We say that X has the negative binomial (k, p) distribution.

The **Geometric distribution** is the **special case** of this random variable when $k = 1$.

Where the formula comes from. For the k th success to fall on trial x , two things must happen simultaneously:

1. Trial x itself is a success — probability p .
2. Among the first $x - 1$ trials there are exactly $k - 1$ successes and $(x - 1) - (k - 1) = x - k$ failures. The number of arrangements is $\binom{x-1}{k-1}$, and each has probability $p^{k-1}(1-p)^{x-k}$.

Multiplying, and combining the p^{k-1} with the final p :

$$P(X = x) = \binom{x-1}{k-1} p^{k-1} (1-p)^{x-k} \times p = \binom{x-1}{k-1} p^k (1-p)^{x-k}. \quad \square$$

Setting $k = 1$ gives $\binom{x-1}{0} p (1-p)^{x-1} = p(1-p)^{x-1}$, the geometric formula. $\checkmark$

Common pitfall: The binomial coefficient is $\binom{x-1}{k-1}$, not $\binom{x}{k}$

The last trial is *not* free to move — it must be a success, or the count would have reached k earlier. So only the first $x - 1$ trials are arranged, and only $k - 1$ successes are placed among them. Writing $\binom{x}{k}$ is the standard error here, and it over-counts by including sequences in which the k th success arrives too early.

Property: Moments of the Negative Binomial distribution

$$\text{Mean } \mu = \frac{k}{p}, \quad \text{Variance } \sigma^2 = \frac{k(1-p)}{p^2}.$$

Both are exactly k times the geometric values, which makes sense: waiting for the k th success is waiting for the first success, k times over, and those waits are independent.

Example 19.3 (Striking oil)

An oil company conducts a geological study that indicates that an exploratory oil well should have a 20% chance of striking oil. What is the probability that the third strike comes on the seventh well drilled?

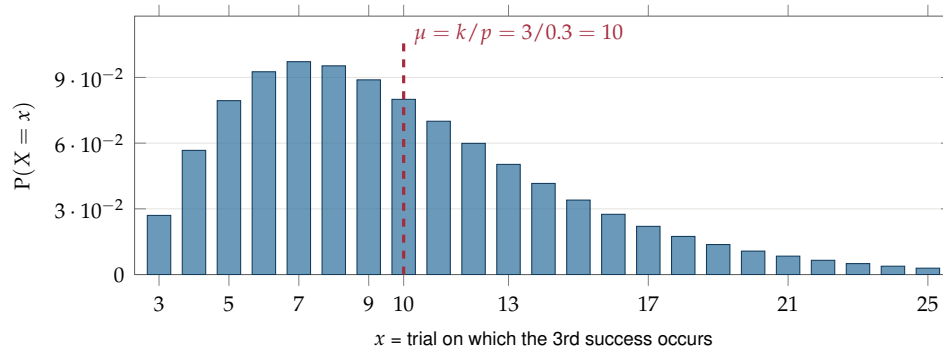


Figure 19.2: Negative Binomial with $k = 3$, $p = 0.3$. Unlike the Geometric it has an interior peak (at $x = 7$ here) because the first $k - 1$ successes must arrive before the last one. The distribution remains right-skewed, so the mean 10 lies well above the mode.

Solution. Here $x = 7$, $k = 3$, $p = 0.2$.

$$P(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k},$$

$$P(X = 7) = \binom{7-1}{3-1} (0.2)^3 (1-0.2)^{7-3} = \binom{6}{2} (0.2)^3 (0.8)^4.$$

With $\binom{6}{2} = 15$, $(0.2)^3 = 0.008$ and $(0.8)^4 = 0.4096$:

$$P(X = 7) = 15 \times 0.008 \times 0.4096 = \boxed{0.049}.$$

Sense check.

The expected number of wells needed for three strikes is $k/p = 3/0.2 = 15$, so getting there by the seventh well is comfortably ahead of schedule — a probability around 5% is plausible for a specific early value.

19.5 The Hypergeometric distribution

Definition: Hypergeometric distribution

This is the **finite population equivalent of the binomial distribution**, in the following sense. Suppose objects are selected at random, one after another, **without replacement**, from a finite population consisting of k successes and $N - k$ failures. **The trials are not independent.**

Random variable: X is the number of successes in a sample of size n from a population of size N that has k successes and $(N - k)$ failures.

Distribution:

$$P(X = x) = \frac{\binom{k}{x} \binom{N-k}{n-x}}{\binom{N}{n}}, \quad x = 0, 1, 2, \dots, \min(n, k).$$

Intuition: Reading the formula as a counting argument

This is nothing but the classical formula $P = n(A)/n(\Omega)$ from Chapter 6, with the counting done by Chapter 7's combinations:

$$P(X = x) = \frac{\overbrace{\binom{k}{x}}^{\text{choose } x \text{ of the } k \text{ successes}} \times \overbrace{\binom{N-k}{n-x}}^{\text{choose } n-x \text{ of the } N-k \text{ failures}}}{\underbrace{\binom{N}{n}}_{\text{all possible samples of size } n}}.$$

It is exactly the reasoning used for the marbles in Chapter 7 — the Hypergeometric is simply that argument written in general.

Property: Moments of the Hypergeometric distribution

$$\text{Mean } \mu = \frac{nk}{N}, \quad \text{Variance } \sigma^2 = n \cdot \frac{k}{N} \cdot \left(1 - \frac{k}{N}\right) \cdot \frac{N-n}{N-1} = \frac{nk(N-k)(N-n)}{N^2(N-1)}.$$

Common pitfall: A correction to Handout 12

The handout gives the variance as

$$\sigma^2 = \frac{nk(N-k)(N-n)}{p^2},$$

which cannot be right: the Hypergeometric has no parameter p at all, and the expression is not even dimensionally consistent. The correct denominator is $N^2(N-1)$.

Verify with the deck example below ($N = 20$, $k = 6$, $n = 5$):

$$\sigma^2 = \frac{5 \times 6 \times 14 \times 15}{20^2 \times 19} = \frac{6300}{7600} = 0.8289,$$

which agrees with a direct computation from the probability table.

The finite population correction

Write $p = k/N$ for the proportion of successes in the population. Then

$$\underbrace{\sigma^2 = np(1-p)}_{\text{Binomial}} \quad \text{versus} \quad \underbrace{\sigma^2 = np(1-p) \cdot \frac{N-n}{N-1}}_{\text{Hypergeometric}}.$$

The extra factor $\frac{N-n}{N-1}$ is called the **finite population correction**. It is always less than or equal to 1, so sampling without replacement always gives *less* variability than sampling with replacement.

Two limiting cases make it intuitive:

- If $n = N$ you sample the whole population, the factor is 0, and the variance is 0 — you know the answer exactly, with no randomness left.
- If N is enormous relative to n , the factor is close to 1 and the Hypergeometric is well approximated by $\text{Bin}(n, k/N)$. Removing a few items from a huge population barely changes its composition. The usual rule of thumb is $n < 0.05N$.

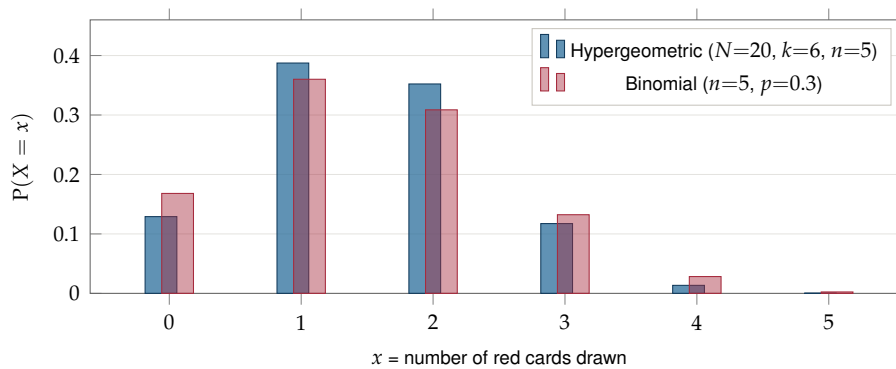


Figure 19.3: Sampling without replacement (blue) against sampling with replacement (red), for the same population proportion $p = 6/20 = 0.3$. Both have mean 1.5, but the Hypergeometric is more concentrated: variance 0.829 against 1.050. Drawing without replacement makes extreme results less likely.

Example 19.4 (Red cards from a small deck)

A deck of cards contains 20 cards: 6 red cards and 14 black cards. 5 cards are drawn randomly without replacement. What is the probability that exactly 4 red cards are drawn?

Solution. The probability of choosing exactly 4 red cards is

$$P(X = 4) = \frac{\text{number of samples with 4 red cards and 1 black card}}{\text{number of possible 5-card samples}}.$$

With $N = 20, k = 6, n = 5, x = 4$:

$$P(X = 4) = \frac{\binom{6}{4} \binom{20-6}{5-4}}{\binom{20}{5}} = \frac{\binom{6}{4} \binom{14}{1}}{\binom{20}{5}} = \frac{15 \times 14}{15\,504} = \frac{210}{15\,504} = \boxed{0.0135}.$$

A note on the handout's wording.

Handout 12 writes the denominator as “# of possible 4 card samples”. It should read *5-card* samples — the sample size is $n = 5$, and the arithmetic $\binom{20}{5} = 15\,504$ confirms this is what was intended.

Sense check.

The mean is $\mu = nk/N = 5 \times 6/20 = 1.5$ red cards, so drawing 4 is well above average and should be rare — about one time in 74. ✓

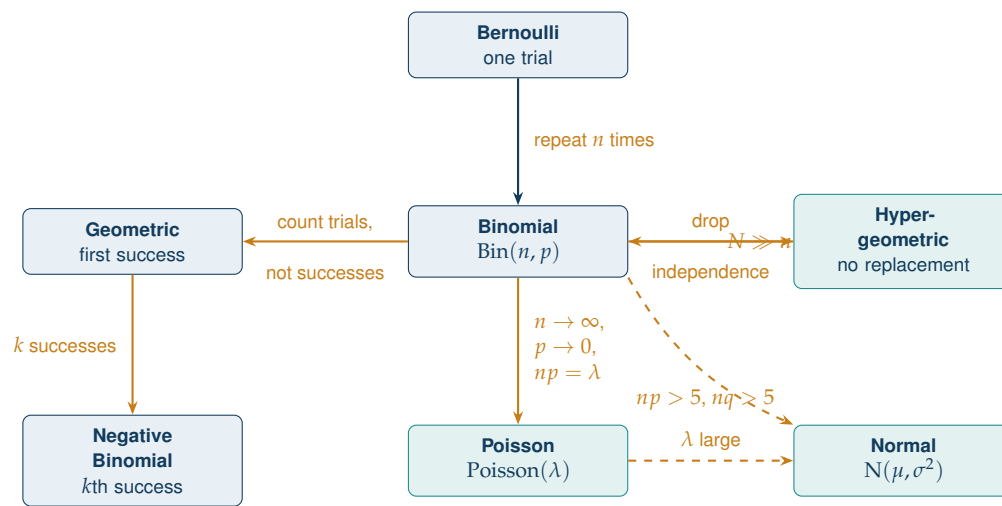
19.6 How the discrete distributions relate

Figure 19.4: The family tree of the discrete distributions in this course. Solid arrows are exact relationships; dashed arrows are large-parameter approximations. Every model here descends from a single Bernoulli trial.

19.7 Summary table

Distribution	Probability function	Mean	Variance
Uniform on $\{1, \dots, k\}$	$P(X = x) = \frac{1}{k},$ $x = 1, \dots, k$	$\frac{k+1}{2}$	$\frac{k^2 - 1}{12}$
Geometric (p)	$P(X = x) = p(1-p)^{x-1},$ $x = 1, 2, \dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negative Binomial (k, p)	$P(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k},$ $x = k, k+1, \dots$	$\frac{k}{p}$	$\frac{k(1-p)}{p^2}$
Hypergeometric (N, k, n)	$P(X = x) = \frac{\binom{k}{x} \binom{N-k}{n-x}}{\binom{N}{n}}$	$\frac{nk}{N}$	$\frac{nk(N-k)(N-n)}{N^2(N-1)}$

A pattern in the means

Every mean in this chapter has an interpretation you can state in words, which is the fastest way to remember them.

Geometric $1/p$	"one success takes $1/p$ trials"
Negative Binomial k/p	" k successes take k times as long"
Hypergeometric nk/N	"sample size $\times$ population proportion"
Binomial np	"number of trials $\times$ chance per trial"

Note that the Hypergeometric and Binomial means are identical when $p = k/N$ — only their variances differ.

Revision Exercises

20.1 Choosing the right distribution

By far the most common reason for losing marks at this stage is not computational error but **misidentification**: applying a Binomial formula to a Geometric problem, or a Binomial to one that needs a Hypergeometric. Before attempting the exercises, work through the diagnostic below.

Recipe: Four questions that identify any distribution in this course

1. **Am I counting successes, or counting trials?**
Successes $\rightarrow$ Binomial or Hypergeometric. Trials $\rightarrow$ Geometric or Negative Binomial.
2. **Is the number of trials fixed in advance?**
Yes $\rightarrow$ Binomial. No, I stop at a success $\rightarrow$ Geometric / Negative Binomial.
3. **Is there replacement (or an effectively infinite population)?**
Yes $\rightarrow$ Binomial. No, and the population is small $\rightarrow$ Hypergeometric.
4. **Am I given a rate per unit time or space instead of a number of trials?**
Yes $\rightarrow$ Poisson.

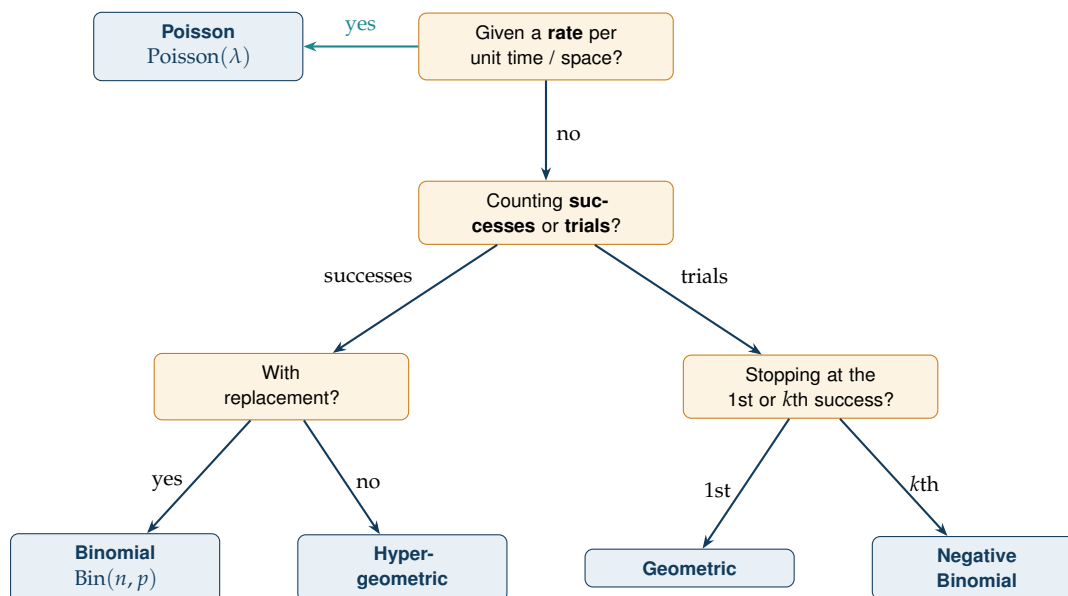


Figure 20.1: A decision tree covering every discrete distribution in the syllabus. Follow it from the top for each exercise below before writing anything down.

Wording that gives the answer away

Phrase in the question	Distribution
"a batch of ten", "in n trials", "out of 400"	Binomial
"per minute", "per ml", "mean rate of"	Poisson
"until the first", "the first ... occurs on"	Geometric
"his last ... at the x th", "the k th ... comes on"	Negative Binomial
"without replacement", small stated population size	Hypergeometric
"heights", "times", "masses" — a measurement	Normal
"the mean of a sample of size n "	Normal via Ch. 18

20.2 The exercises

Exercise 20.1 (Defective items in a batch)

Suppose that 1% of the items made by a certain machine are defective. To keep a check on the quality of the output, a batch of ten items is inspected occasionally. What is the probability that the next ten items inspected include *more than one* defective?

Solution.

Identify the distribution.

A fixed number of trials ($n = 10$), each item defective or not, with constant probability $p = 0.01$. This is **Binomial**. Let X be the number of defectives in a sample of 10, so

$$X \sim \text{Bin}(10, 0.01).$$

Translate "more than one".

"More than one" means $X \geq 2$, an infinite-looking sum of cases. Take the complement:

$$P(X > 1) = 1 - P(X = 0) - P(X = 1).$$

Compute the two terms.

$$P(X = 0) = \binom{10}{0} (0.01)^0 (0.99)^{10} = (0.99)^{10} = 0.904382,$$

$$P(X = 1) = \binom{10}{1} (0.01)^1 (0.99)^9 = 10 \times 0.01 \times 0.913517 = 0.091352.$$

Combine.

$$P(X > 1) = 1 - 0.904382 - 0.091352 = \boxed{0.00427}.$$

Interpretation.

About one inspection in 234 raises the alarm even though the machine is performing to specification. That is the *false alarm rate* of this quality control scheme, and it is exactly the kind of number an engineer needs before deciding what to do when two defectives turn up.

Exercise 20.2 (*Selling candy bars*)

Pat is required to sell candy bars to raise money for the 6th grade field trip. There is a 40% chance of him selling a candy bar at each house. He has to sell 5 candy bars in all. What is the probability he sells his *last* candy bar at the 11th house?

Solution.

Identify the distribution.

We need the trial on which the **5th success** occurs — the number of trials is random, not fixed. This is **Negative Binomial**, with

$$x = 11, \quad k = 5, \quad p = 0.4.$$

Apply the formula.

$$P(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k},$$

so the probability that he sells his last candy bar at the 11th house is

$$P(X = 11) = \binom{10}{4} (0.4)^5 (1-0.4)^6 = \binom{10}{4} (0.4)^5 (0.6)^6.$$

Evaluate.

With $\binom{10}{4} = 210$, $(0.4)^5 = 0.01024$ and $(0.6)^6 = 0.046656$:

$$P(X = 11) = 210 \times 0.01024 \times 0.046656 = \boxed{0.1003}.$$

Sense check.

The expected number of houses is $k/p = 5/0.4 = 12.5$, so 11 houses is slightly better than average — and a probability of about 10% for one specific value near the centre of the distribution is entirely reasonable.

Why not Binomial?

A tempting wrong answer is $\binom{11}{5} (0.4)^5 (0.6)^6 = 0.2207$. That computes the probability of 5 sales in 11 houses in any order, which includes sequences where the 5th sale happened at house 7 and houses 8–11 were all failures. The question specifies that the *last* sale is at house 11, which pins the final trial down and reduces the coefficient from $\binom{11}{5}$ to $\binom{10}{4}$.

Exercise 20.3 (*The first defective item*)

Products produced by a machine have a 3% defective rate. What is the probability that the first defective occurs in the fifth item inspected?

Solution.

Identify the distribution.

We want the trial of the **first** success (here a “success” is finding a defective). This is **Geometric** with $p = 0.03$. Let X be the number of trials until the first success.

Apply the formula.

$$P(X = 5) = p(1 - p)^{5-1} = 0.03(1 - 0.03)^4 = 0.03(0.97)^4.$$

With $(0.97)^4 = 0.885293$:

$$P(X = 5) = 0.03 \times 0.885293 = \boxed{0.0266}.$$

Sense check.

The expected wait is $1/p = 1/0.03 \approx 33$ items, so the first defective appearing as early as the fifth item is somewhat unlikely — a probability of about 2.7% fits. Note also that $P(X = 5)$ must be smaller than $P(X = 1) = 0.03$, since the geometric probabilities always decrease. ✓

Common pitfall: A correction to Handout 12

The handout gives the answer to the exercise above as 0.072. That value corresponds to a defective rate of $p = 0.3$, not the stated 3%:

$$0.3(0.7)^4 = 0.3 \times 0.2401 = 0.0720, \quad \text{whereas} \quad 0.03(0.97)^4 = 0.0266.$$

A decimal point appears to have been lost in the arithmetic. With $p = 0.03$ the correct answer is **0.0266**.

The sense check above catches this immediately: with a 3% defect rate you expect to wait about 33 items, so a 7% chance of the first defective landing on item 5 specifically would be far too high.

Exercise 20.4 (Three sixes in seven throws)

An ordinary die is thrown seven times. Find the probability of obtaining exactly three sixes.

Solution.**Identify the distribution.**

A fixed number of trials ($n = 7$), each a six or not, with constant probability $p = \frac{1}{6}$, and the throws are independent. This is **Binomial**. Let X be the number of sixes obtained:

$$X \sim \text{Bin}(7, \frac{1}{6}).$$

Apply the formula.

$$P(X = 3) = \binom{7}{3} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^4 = 35 \times \frac{1}{216} \times \frac{625}{1296}.$$

Evaluate.

$$P(X = 3) = \frac{35 \times 625}{216 \times 1296} = \frac{21\,875}{279\,936} = \boxed{0.07814}.$$

Sense check.

The expected number of sixes is $np = 7/6 \approx 1.17$, so three sixes is about twice the average — a probability under 10% is what one would expect.

Exercise 20.5 (*Calls at a call centre*)

A call centre receives calls at a mean rate of 3 per minute. Find the probability that during a randomly selected 3 minutes there will be no calls.

Solution.

Identify the distribution.

We are given a **rate** per unit time, with events randomly scattered in time. This is **Poisson**.

Scale the interval.

The rate is 3 calls per minute, and the interval of interest is 3 minutes. As in Section 14.5, it is the *mean* that scales with the interval:

$$\lambda = 3 \times 3 = 9.$$

Let X be the number of calls in 3 minutes, so $X \sim \text{Poisson}(9)$.

Apply the formula.

$$P(X = 0) = \frac{e^{-9} 9^0}{0!} = e^{-9} = \boxed{0.000123}.$$

Interpretation.

Roughly one chance in 8100 — a three-minute silence at this call centre would be a genuinely notable event. Note how sharply the answer depends on getting the scaling right: had you used $\lambda = 3$ instead of 9, you would have obtained $e^{-3} = 0.0498$, too large by a factor of 400.

20.3 A consolidated worked comparison

The three exercises above involving repeated trials at a fixed probability look superficially alike. Setting them side by side shows precisely what distinguishes them.

	Exercise 20.1 defectives in ten	Exercise 20.3 first defective	Exercise 20.2 last candy bar
What is fixed?	10 items inspected	one success sought	5 successes sought
What is random?	number of defectives	number of items	number of houses
Distribution	$\text{Bin}(10, 0.01)$	Geometric (0.03)	Neg. Bin. (5, 0.4)
Coefficient	$\binom{10}{x}$	none	$\binom{x-1}{4}$
Mean	$np = 0.1$	$1/p \approx 33.3$	$k/p = 12.5$
Answer	0.00427	0.0266	0.1003

The single most useful diagnostic

Ask: **does the question tell me how many trials there will be?**

If yes, the number of trials is fixed and the successes are random — so the answer is Binomial (or Hypergeometric, if there is no replacement). If no, the question is telling you how many *successes* to wait for, so the trials are random — and the answer is Geometric or Negative Binomial.

“A batch of ten items” fixes the trials. “Until the first defective” does not.

20.4 Further practice

The following additional problems use the same models. Answers are given immediately; work them before reading.

Exercise 20.6 (Mixed practice)

- A committee of 4 is chosen at random from 7 women and 5 men. What is the probability that exactly 3 are women?
- A basketball player has a free-throw success rate of 70%. What is the probability that her second miss occurs on her sixth attempt?
- A book of 200 pages contains 40 typographical errors, scattered at random. What is the probability that a given page contains no errors?
- The masses of bags of rice are normally distributed with mean 5.02 kg and standard deviation 0.04 kg. What is the probability that the mean mass of a random sample of 16 bags is less than 5.00 kg?

Solution.

(a) Hypergeometric.

Sampling without replacement from a small finite population of $N = 12$ with $k = 7$ women, $n = 4$:

$$P(X = 3) = \frac{\binom{7}{3}\binom{5}{1}}{\binom{12}{4}} = \frac{35 \times 5}{495} = \frac{175}{495} = \frac{35}{99} \approx \boxed{0.354}.$$

(b) Negative Binomial.

Here a “success” is a *miss*, so $p = 0.30$, and we want the 2nd success on trial 6: $k = 2$, $x = 6$.

$$P(X = 6) = \binom{5}{1} (0.3)^2 (0.7)^4 = 5 \times 0.09 \times 0.2401 \approx \boxed{0.108}.$$

Note the trap: the stated rate (70%) is for the outcome we are *not* counting. Always define “success” to match the event being counted.

(c) Poisson.

Errors are randomly scattered over pages, at a mean rate of $40/200 = 0.2$ per page, so $X \sim \text{Poisson}(0.2)$:

$$P(X = 0) = e^{-0.2} \approx \boxed{0.819}.$$

(d) Sampling distribution of the mean.

The population is normal, so Chapter 18’s exact result applies:

$$\bar{X} \sim N\left(5.02, \frac{0.04^2}{16}\right), \quad \text{SD}(\bar{X}) = \frac{0.04}{\sqrt{16}} = 0.01.$$

$$P(\bar{X} < 5.00) = P\left(Z < \frac{5.00 - 5.02}{0.01}\right) = P(Z < -2) = 1 - 0.9772 = \boxed{0.0228}.$$

Compare with a single bag: $P(X < 5.00) = P(Z < -0.5) = 0.3085$. The mean of sixteen bags is far more tightly controlled than any individual bag — the practical reason quality control samples rather than testing one unit.

PART V

Appendices

Master Formula Sheet

Part I — Descriptive statistics

Quantity	Formula / rule
Arithmetic mean	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$
Weighted mean	$\bar{X}_w = \frac{\sum w_i x_i}{\sum w_i}$
Median, n odd	the $\left(\frac{n+1}{2}\right)^{\text{th}}$ ordered value
Median, n even	mean of the $\left(\frac{n}{2}\right)^{\text{th}}$ and $\left(\frac{n}{2} + 1\right)^{\text{th}}$ ordered values
Quartiles	$Q_k = \frac{k(n+1)}{4}$ th value; $Q_1 = P_{25}$, $Q_2 = P_{50} = M$, $Q_3 = P_{75}$
Percentile P_k	compute $nk/100$; if an integer a , average the a th and $(a+1)$ th values; if a fraction, take the value in position $\lceil nk/100 \rceil$
Range	$R = \max - \min$
Interquartile range	$IQR = Q_3 - Q_1$
Sample variance	$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$
Standard deviation	$s = \sqrt{s^2}$
Linear transformation	$y = kx + c \Rightarrow \bar{y} = k\bar{x} + c, \quad s_y = k s_x$
Pearson's skewness	$\frac{\text{mean} - \text{mode}}{s} \approx \frac{3(\text{mean} - \text{median})}{s}$

Part II — Probability

Result	Statement
Axiom 1	$P(A) \geq 0$
Axiom 2	$P(\Omega) = 1$
Axiom 3	$P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$
Complement	$P(A') = 1 - P(A)$
Addition rule	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Classical probability	$P(A) = \frac{n(A)}{n(\Omega)}$ (equally likely outcomes only)
Relative frequency	$P(A) \approx f_r(A)/r$ for large r
De Morgan	$(A \cup B)' = A' \cap B'$; $(A \cap B)' = A' \cup B'$
Permutations	${}^n P_r = \frac{n!}{(n-r)!}$
Combinations	$\binom{n}{r} = \frac{n!}{r!(n-r)!}$
Conditional probability	$P(A B) = \frac{P(A \cap B)}{P(B)}$, $P(B) > 0$
Multiplication rule	$P(A \cap B) = P(A B) P(B) = P(B A) P(A)$
Total probability	$P(A) = \sum_i P(A B_i) P(B_i)$, $\{B_i\}$ a partition
Bayes' theorem	$P(A_i B) = \frac{P(B A_i) P(A_i)}{\sum_j P(B A_j) P(A_j)}$
Independence	$P(A \cap B) = P(A) P(B)$, equivalently $P(A B) = P(A)$
Mutual independence	all pairwise products <i>and</i> $P(A \cap B \cap C) = P(A) P(B) P(C)$
"At least one"	$P(\geq 1 \text{ success}) = 1 - (1 - p)^n$ (independent trials)

Part III — Random variables

Quantity	Discrete	Continuous
Normalisation	$\sum_x P(x) = 1$	$\int_{-\infty}^{\infty} f(x) dx = 1$
Interval probability	$\sum_{a \leq x \leq b} P(x)$	$\int_a^b f(x) dx$
c.d.f.	$F(x) = \sum_{t \leq x} P(t)$	$F(x) = \int_{-\infty}^x f(t) dt$, $f = F'$
Expectation	$E(X) = \sum_x x P(x)$	$E(X) = \int x f(x) dx$
Second moment	$E(X^2) = \sum_x x^2 P(x)$	$E(X^2) = \int x^2 f(x) dx$
Variance	$\text{Var}(X) = E[(X - \mu)^2] = E(X^2) - \mu^2$	

Rule	Statement
Linearity of E	$E(aX + b) = a E(X) + b$
Additivity of E	$E(X + Y) = E(X) + E(Y)$ (always, no independence needed)
Scaling of Var	$\text{Var}(aX + b) = a^2 \text{Var}(X)$
Additivity of Var	$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$ (independent X, Y only)

The distributions

Distribution	Probability function	Mean	Variance
Bin(n, p)	$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$ $k = 0, 1, \dots, n$	np	$np(1-p)$
Poisson(λ)	$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!},$ $x = 0, 1, 2, \dots$	λ	λ
Uniform on $\{1, \dots, k\}$	$P(X = x) = \frac{1}{k},$ $x = 1, \dots, k$	$\frac{k+1}{2}$	$\frac{k^2-1}{12}$
Geometric (p)	$P(X = x) = p(1-p)^{x-1},$ $x = 1, 2, \dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negative Binomial (k, p)	$P(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k},$ $x = k, k+1, \dots$	$\frac{k}{p}$	$\frac{k(1-p)}{p^2}$
Hypergeometric (N, k, n)	$P(X = x) = \frac{\binom{k}{x} \binom{N-k}{n-x}}{\binom{N}{n}}$	$\frac{nk}{N}$	$\frac{nk(N-k)(N-n)}{N^2(N-1)}$
$N(\mu, \sigma^2)$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)},$ $-\infty < x < \infty$	μ	σ^2

Combinations of normal variables and sampling

Result	Statement
Sum of independent normals	$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
Difference of independent normals	$X - Y \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$ (variances still add)
Multiple of a normal	$aX \sim N(a\mu, a^2\sigma^2)$
General linear combination	$aX + bY + c \sim N(a\mu_1 + b\mu_2 + c, a^2\sigma_1^2 + b^2\sigma_2^2)$
Comparing two variables	$P(X > Y) = P(X - Y > 0)$
Combining standard deviations	$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$ (Pythagoras)
Sample mean, normal population	$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ exactly, for every n
Central Limit Theorem	$\bar{X} \overset{\text{approx}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$ for any population, n large ($n \geq 30$)
Standard error of the mean	$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
z-score for a sample mean	$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ (not σ alone)

Normal distribution procedures

Task	Method
Standardise	$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$
Un-standardise	$x = \mu + z\sigma$
$P(X < a)$	$\Phi\left(\frac{a - \mu}{\sigma}\right)$
$P(X > a)$	$1 - \Phi\left(\frac{a - \mu}{\sigma}\right)$
$P(a < X < b)$	$\Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$
Negative z	$\Phi(-z) = 1 - \Phi(z)$
Empirical rule	68.3% within $\mu \pm \sigma$; 95.4% within $\mu \pm 2\sigma$; 99.7% within $\mu \pm 3\sigma$
Normal approx. to Binomial	$\text{Bin}(n, p) \approx N(np, np(1-p))$ if $np > 5$ and $n(1-p) > 5$
Normal approx. to Poisson	$\text{Poisson}(\lambda) \approx N(\lambda, \lambda)$ if $\lambda \geq 10$
Continuity correction	$X = k \rightarrow (k-0.5, k+0.5)$; $X \leq k \rightarrow X < k + 0.5$; $X \geq k \rightarrow X > k - 0.5$

Useful constants

e^{-1}	e^{-2}	e^{-3}	e^{-4}	e^{-5}	e^{-12}
0.36788	0.13534	0.049787	0.018316	0.0067379	6.144×10^{-6}

n	0	1	2	3	4	5	6	7
$n!$	1	1	2	6	24	120	720	5040

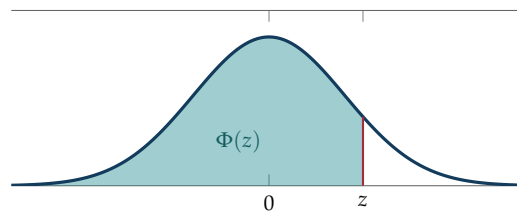
Pascal's triangle (rows $n = 0$ to 6), giving $\binom{n}{r}$:

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1

Standard Normal Distribution Table

How to read this table

The table gives $\Phi(z) = P(Z \leq z)$, the area under the standard normal curve to the **left** of z , for $z \geq 0$.



Recipe: Locating a value

To find $\Phi(1.63)$: go down the first column to the row **1.6**, then across to the column headed **0.03**. The entry is 0.9484.

For other cases use the four rules:

$$P(Z > z) = 1 - \Phi(z), \quad \Phi(-z) = 1 - \Phi(z), \quad P(z_1 < Z < z_2) = \Phi(z_2) - \Phi(z_1).$$

For an **inverse** problem (given a probability, find z), search the *body* of the table for the probability and read off the row and column labels.

Table of $\Phi(z) = P(Z \leq z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Critical values

Frequently needed percentage points, for inverse problems.

$\Phi(z)$	0.90	0.95	0.975	0.99	0.995	0.999	0.9995
z	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	3.2905
Central area	50%	68.27%	80%	90%	95%	99%	99.73%
$\pm z$	0.6745	1.0000	1.2816	1.6449	1.9600	2.5758	3.0000

The two you should memorise

$z = 1.96$ for a central 95% and $z = 2.58$ for a central 99%. These two numbers appear in every confidence interval you will ever construct.

Answers, Notation and Errata

Answer summary

Every exercise in these notes is solved in full where it appears. This table collects the final answers for quick checking.

Ref.	Problem	Answer
<i>Chapter 3 — Measures of location</i>		
Ex. 3.1	Quartiles of 50 exam marks	$Q_1 = 67, P_{58} = 77.5, Q_3 = 86, M = 75.5$
<i>Chapter 4 — Measures of dispersion</i>		
Ex. 4.2	Dispersion of the 50 marks	$R = 59, \text{IQR} = 19, \bar{x} = 74.92, s^2 = 201.50, s = 14.20$
<i>Chapter 6 — Axioms of probability</i>		
Ex. 6.1	Coin probability table	0, 0.4, 0.6, 1
Ex. 6.2	Die probability table	0, 0.3, 0.2, 0.5, 0.5, 0.7, 0.8, 1
Ex. 6.3	Two marbles from 3R, 2B	(a) {RR, RB, BB} (b) 3/5 (c) 7/10
Ex. 6.4	Drawing a seven	(a) 1/13 (b) 12/13
Ex. 6.5	Coin and die	(a) 1/2 (b) 1/3 (c) 1/6 (d) 2/3
Ex. 6.6	Survey of 1500 homes	(a) 2/15 (b) 14/15
Ex. 6.7	One card, three questions	(1) 1/4 (2) 10/13 (3) 3/13
<i>Chapter 8 — Conditional probability</i>		
Ex. 8.1	Three coin tosses	(a) 1/2 (b) 1/2 (c) 3/8 (d) 3/4
<i>Chapter 9 — Total probability and Bayes</i>		
Ex. 9.1	Ms. Ferrari's car	$P(S) = 0.78$
<i>Chapter 10 — Independence</i>		
Ex. 10.1	H_1, H_2, H_3 and A	(a) mutually independent (b) pairwise but <i>not</i> mutually independent
<i>Chapter 11 — Random variables</i>		
Ex. 11.1	Toss until two agree	$\Omega_X = \{0, 1, 2, \dots\}, \Omega_Y = \{2, 3, 4, \dots\}$
<i>Chapter 12 — Expectation and variance</i>		
Ex. 12.1	$X \in \{-2, \dots, 2\}$	(a) $E(X) = 0$ (b) $E(X^2) = 2.6$
Ex. 12.2	The $\pm k$ gamble	$E(X) = 0, \text{Var}(X) = k^2, \sigma_X = k $
Ex. 12.3	Die gamble	$E(X) = \$5$ ($\sigma_X \approx \$21.41$)
<i>Chapter 13 — Binomial</i>		
Ex. 13.1	Fair coin, 3 tosses	(a) 1/8 (b) 3/8 (c) 7/8 (d) 1/2
Ex. 13.2	Fair die, 5 tosses	(a) $625/3888 \approx 0.161$ (b) $3125/3888 \approx 0.804$ (c) $763/3888 \approx 0.196$
Ex. 13.3	Family of 4 children	(a) 15/16 (b) 7/8
Ex. 13.4	400 bolts, $p = 0.1$	(a) mean = 40 (b) SD = 6
Ex. 13.5	Mean 5.76, SD 1.92	$n = 16, p = 0.36, P(X = 6) \approx 0.201$
<i>Chapter 15 — Continuous random variables</i>		
Ex. 15.1	$f(x) = kx$ on $[0, 4]$	(a) $k = 1/8$ (c) $21/64 \approx 0.328$
Ex. 15.2	$f(x) = k(4 - x)$ on $[1, 3]$	(a) $k = 1/4$ (c) 0.66
<i>Chapter 17 — Linear combinations of normals</i>		
Ex. 17.1	Mr Jones at the library	(i) 0.1056 (ii) 0.0062
<i>Chapter 20 — Revision exercises</i>		
Ex. 20.1	Defectives in a batch of ten, $\text{Bin}(10, 0.01)$	$P(X > 1) = 0.00427$
Ex. 20.2	Last candy bar at the 11th house	0.1003
Ex. 20.3	First defective on the 5th item	0.0266 (handout: 0.072 — see errata)
Ex. 20.4	First defective on the 5th item	0.0266 (handout: 0.072 — see errata)

Worked *examples* (as opposed to exercises) with numerical answers:

Example	Answer
Counters without replacement (Ch. 8)	(a) $7/10$ (b) $7/30$ (c) $7/15$
Bag of trick coins (Ch. 9)	$P(H) = 5/12$
Laboratory test (Ch. 9)	$P(A B) \approx 0.165$
Three plants (Ch. 9)	(a) $P(D) = 0.027$ (b) $5/9 \approx 0.556$
Rectangle machine (Ch. 11)	$P_L: 0.5, 0.5; P_W: 0.32, 0.28, 0.40; P_X:$ $0.12, 0.33, 0.40, 0.15$
Perimeter moments (Ch. 12)	$E(X) = 9.16, \text{Var}(X) = 3.1344$
Poisson(2) (Ch. 14)	(a) 0.0902 (b) 0.3233
Bacteria, 1 ml (Ch. 14)	(a) 0.0183 (b) 0.1954 (c) 0.2381
Bacteria, 3 ml (Ch. 14)	7.99×10^{-5}
Radioactive counts (Ch. 14)	(a) 0.3827 exact (b) 0.3830 by Normal approximation
$f(x) = 3x^2/64$ (Ch. 15)	$E(X) = 3, \text{Var}(X) = 0.6$
Student heights (Ch. 16)	(a) 0.9332 (b) 0.9484 (c) 0.6826; top 10% at 178.2 cm
Coin tossed 100 times (Ch. 16)	$P(45 \leq X \leq 55) \approx 0.7286$
Sums of normals (Ch. 17)	(a) 0.9772 (b) 0.8185 (c) 0.7257
Multiple $3X$ (Ch. 17)	$P(3X > 160) = 0.2523$
Sample of 15 (Ch. 18)	$P(\bar{X} < 58) = 0.0264$
CLT from Bin(9, 0.5) (Ch. 18)	$P(\bar{X} > 5) = 0.0340$
Fourth child, first son (Ch. 19)	$1/16 = 0.0625$
Third oil strike on well 7 (Ch. 19)	0.049
Four red cards of five (Ch. 19)	$210/15504 \approx 0.0135$

Notation index

Symbol	Meaning
Ω, S	outcome space (sample space) of an experiment
ω	a single outcome, an element of Ω
$\mathcal{G}$	the domain of the probability function: a set of events
$\emptyset$	the empty set; the impossible event
A'	the complement of A (also written A^c or A')
$A \cup B$	union: A or B or both
$A \cap B$	intersection: both A and B
$n(A)$	the number of outcomes in A
$P(A)$	probability of the event A
$P(A B)$	conditional probability of A given B
X, Y, Z	random variables (capital letters)
x	a particular value of X (lower case)
Ω_X	the sample space of the random variable X
$P_X(x), P(x)$	probability function, $P(X = x)$
$f(x)$	probability density function
$F(x)$	cumulative distribution function, $P(X \leq x)$
$E(X), \mu$	expectation (mean) of X
$\text{Var}(X), \sigma^2$	variance of X
$\sigma, \text{SD}(X)$	standard deviation of X
$\bar{x}, s^2, s$	sample mean, variance and standard deviation
$\binom{n}{r}, {}^nC_r$	number of combinations of r from n
nP_r	number of permutations of r from n
$\text{Bin}(n, p)$	Binomial distribution
$\text{Poisson}(\lambda)$	Poisson distribution
$N(\mu, \sigma^2)$	Normal distribution with mean μ , variance σ^2
Z	the standard normal variable, $N(0, 1)$
$\phi(z), \Phi(z)$	standard normal density and c.d.f.
$\perp$	independence
$\bar{X}$	the sample mean, regarded as a random variable
$\sigma / \sqrt{n}$	the standard error of the mean
N, k, n	Hypergeometric: population size, successes in the population, sample size
iid	independent and identically distributed

Errata in the source handouts

Four corrections were noted while preparing these notes. All are flagged in the main text at the point where they arise.

Source	As printed	Correction
Handout 02, Exercise 3(b)	"Answer = 5/10" for the probability of drawing 1 red and 1 blue marble	Should be $6/10 = 3/5$. The handout's own enumerated extended sample space lists six mixed pairs. With 5/10 the three colour outcomes sum to 9/10, violating Axiom 2. See Section 6.5.
Handout 02, "Exhaustive events"	" $P(A \cup B) = S$ and $P(A \cup B) = 1$ "	A probability is a number, not a set. The intended statement is $A \cup B = \Omega$, which implies $P(A \cup B) = 1$. See Section 5.2.
Handout 12, Hypergeometric moments	$\sigma^2 = \frac{nk(N-k)(N-n)}{p^2}$	The Hypergeometric has no parameter p , and the expression is dimensionally inconsistent. The denominator should be $N^2(N-1)$: $\sigma^2 = \frac{nk(N-k)(N-n)}{N^2(N-1)}$. See Section 19.5.
Handout 12, Revision Exercise 3	" $P(X = 5) = 0.03(1 - 0.03)^4 = 0.072$ "	The stated expression is right but the arithmetic is not: $0.03(0.97)^4 = 0.0266$. The printed value 0.072 corresponds to $p = 0.3$, so a decimal point appears to have been lost. See Chapter 20.

Four further points where these notes are more precise than the slides, rather than correcting an outright error:

- Handout 08 states that approximately 99.8% of a normal distribution lies within $\pm 3\sigma$. The exact figure is 99.73%, conventionally quoted as 99.7%.
- Handout 05 writes $Z = XY$ without specifying the operation. It is read here as the *product* XY , which gives $\Omega_Z = \{0, 2\}$ for three coin tosses.
- Handout 12's Hypergeometric example writes the denominator as "# of possible 4 card samples". Since $n = 5$ and the arithmetic uses $\binom{20}{5} = 15\,504$, this should read *5-card* samples.
- Handout 10 is byte-identical to Handout 09. The Poisson material it contains is therefore presented once, in Chapter 14.

Source handouts

Handout	Topic	Chapters here
01	Introduction, data types, summary statistics	1–4
02	Experiments, events, axioms of probability	5–7
03	Conditional probability, total probability, Bayes	8–9
04	Independence	10
05	Discrete random variables, expectation, variance	11–12
06	Binomial distribution	13
07	Continuous random variables	15
08	Normal distribution	16
09	Poisson distribution, Normal approximation	14, 16
10	(byte-identical duplicate of Handout 09)	—
11	Combinations of normal variables, distribution of the sample mean, Central Limit Theorem	17–18
12	Uniform, Geometric, Negative Binomial and Hypergeometric distributions; revision exercises	19–20

Handouts 01–03, 11 and 12 by Dr Dilshani Tissera; Handouts 04–09 by Dr Dilanthi Gamage Dona, adapted and updated from the original lecture materials. Department of Statistics, University of Colombo.