STATISTICS AND PROBABILITY

CHAPTER 12

# Probability Distributions

12

SYLLABUS 4.7 4.8 4.9 4.12 4.14

If you measure the same kind of thing many times, you rarely get the same answer twice. Masses differ from packet to packet, and a repeated experiment gives a different count each time. A single number cannot describe a quantity that behaves this way. What describes it is the list of values the quantity can take, together with how likely each value is. That list is called a **distribution**, and this chapter covers the few that fit a great many situations.

A factory fills bags of rice, each meant to hold 500 grams, and weighs 5000 of them. No two masses are exactly equal. Group the readings, draw them as a graph, and a shape appears: most bags near 500 g, with fewer and fewer on either side, falling away evenly into one smooth hump. The shape is a bell.

The same curve appears where there is no rice at all. Astronomers found it in the *errors* of their telescope readings, and biologists in the heights of people. The reason is usually the same: a bag's mass is decided not by one thing but by many small effects, such as the size of the grains, a slight vibration in the machine and the exact moment the valve closes, and the end of the chapter shows why such a sum gives a bell.

A distribution turns uncertainty into something you can calculate with. From one you can predict an average result, measure how far individual results fall from that average, and put a number on the risk of an extreme value. That is why these curves are used in medicine, manufacturing and finance, and also why care is required: every answer depends on the distribution being the right description, and a wrong choice gives precise answers that are wrong.

By the end of this chapter you will be able to find the mean and variance of a discrete random variable, use the binomial distribution, find normal probabilities and work backwards with the inverse normal, standardise a value to compare across distributions, and work with a continuous random variable given by a density function.

# 12.1 Discrete Random Variables

A **random variable** is a number whose value depends on chance. The number of heads in three coin tosses, the score on a die, the number of defective items in a box: each is a rule that assigns a number to a random outcome. We write random variables with capital letters, $X$, and particular values with lower case, $x$.

A random variable is **discrete** when it can only take separate, countable values. To describe it fully, you list every value it can take alongside the probability of each. This list is its **probability distribution**.

Here are five discrete random variables, each with the values it can take.

- $X$ = the number of heads in three coin tosses. Values $0, 1, 2, 3$.
- $X$ = the score on one roll of a fair die. Values $1, 2, 3, 4, 5, 6$.
- $X$ = the number of defective items in a box of twenty. Values $0, 1, \ldots, 20$.
- $X$ = the number of students absent from a class of thirty. Values $0, 1, \ldots, 30$.
- $X$ = the number of tosses of a coin until the first head appears. Values $1, 2, 3, \ldots$, with no largest value.

The last one deserves attention. A discrete variable does not have to take a finite number of values. It only has to take separate values that can be listed in order, and that list may continue for ever. What rules a variable out is taking *every* value across a range, as a height or a waiting time does. Those are continuous variables, and §12.4 treats them.

DEF A **discrete random variable** $X$ takes countable values $x_1, x_2, \ldots$ Its **probability distribution** gives $P(X = x_i)$ for each value, subject to

$$P(X = x_i) \geq 0 \qquad \text{and} \qquad \sum_i P(X = x_i) = 1.$$

The second condition is the one the examples below use. Since some value must occur, all the probabilities add to exactly 1. This single fact lets you find a missing probability.

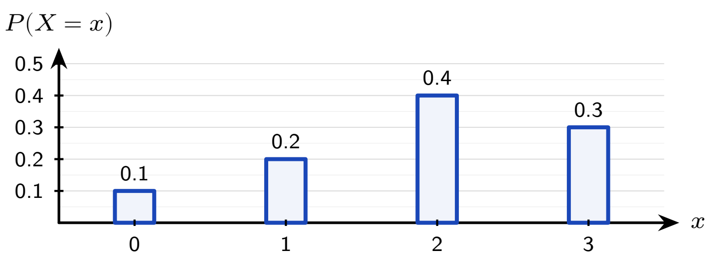

**Figure 12.1** The probability distribution of a discrete random variable. Each bar is the probability of one value, and the four bars must total exactly $1$.

**Worked through.** A discrete random variable $X$ has the distribution below.

| $x$ | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| $P(X = x)$ | 0.1 | 0.3 | $k$ | 0.2 |

To find $k$, use the fact that the probabilities must sum to 1:

$$0.1 + 0.3 + k + 0.2 = 1 \implies k = 0.4.$$

Then $P(X \geq 3) = P(X = 3) + P(X = 4) = 0.4 + 0.2 = 0.6.$

**EXAMPLE 12.1**

A random variable $X$ takes values $1, 2, 3, 4$ with $P(X = x) = kx$. Find $k$.
Sum the probabilities and set equal to 1:

$$k(1) + k(2) + k(3) + k(4) = 10k = 1 \implies k = \frac{1}{10}.$$

When the distribution is given by a formula, the total-probability condition still determines the constant.

**YOUR TURN 12.1** Discrete Random Variables

### Discrete or continuous

**1.** State whether each random variable is discrete or continuous.

(a) the number of emails received in a day

(b) the time a runner takes to finish a race

(c) the number of rolls of a die until the first six

(d) the mass of an apple

**Finding a missing probability**

**2.**

(a) For $x = 1, 2, 3$ the probabilities are $0.2$, $0.5$ and $k$. Find $k$.
(b) A variable $X$ has $P(X = x) = 0.15,\ 0.25,\ 0.35,\ k$ for $x = 0, 1, 2, 3$. Find $k$ and $P(X \le 1)$.
(c) A variable takes values $x = 0, 1, 2, 3$ with $P(X = x) = k(x + 1)$. Find $k$.
(d) A variable takes values $x = 1, 2, 3, 4$ with $P(X = x) = \dfrac{x^2}{30}$. Find $P(X = 3)$.
(e) For the distribution $P(X = x) = 0.1,\ 0.4,\ 0.3,\ 0.2$ at $x = 0, 1, 2, 3$, find $P(X > 1)$ and $P(1 \le X \le 2)$.

**3.**

(a) Show that $P(X = x) = \frac{1}{18}(4 + x)$ for $x = 1, 2, 3$ defines a probability distribution, and find $P(X \ge 2)$.
(b) A variable takes values $x = 1, 2, 3, 4$ with probabilities $0.1$, $a$, $0.3$, $b$. Given that $P(X \le 2) = 0.4$, find $a$ and $b$.
(c) Hence find $P(X > 2)$ for the distribution in part (b).
(d) A variable takes values $x = 1, 2, 3$ with $P(X = x) = k \cdot 2^x$. Find $k$.

**Checking that a distribution is valid**

**4.**

(a) Show that the values of $P(X = x) = \dfrac{2x - 1}{8}$ for $x = 0, 1, 2, 3$ add up to $1$, and explain why this is still not a probability distribution.
(b) A variable has $P(X = 1) = c$, $P(X = 2) = 0.5 - c$ and $P(X = 3) = 0.5$. Find the set of values of $c$ for which this is a probability distribution.

**Concept check**

**5.** A variable $X$ takes the values $1, 2, 3, 4$ with probabilities $0.1$, $0.3$, $0.4$ and $0.2$. A student writes: "$P(X < 3) = P(X \le 3) = 0.1 + 0.3 + 0.4 = 0.8$." Identify the error, and find the correct value of $P(X < 3)$.

# 12.2 Expected Value and Variance

If you played a game many times, what would you win on average per play? That long-run average is the **expected value** $E(X)$. It is not the most likely outcome, and it need not be a value that $X$ can take at all. It is the balance point of the distribution, found by weighting each value by its probability.

KEY · IB

$$E(X) = \mu = \sum_i x_i \, P(X = x_i)$$

This is exactly the mean from Ch. 10, with probabilities playing the role of relative frequencies. Over many trials, the relative frequency of each value approaches its probability, so the average approaches $E(X)$.

**Worked through.** Suppose a game costs $2 to play: you roll a fair die and win, in dollars, the number shown. To decide whether the game is worth playing, let $X$ be the amount won. Each face is equally likely:

$$E(X) = \tfrac{1}{6}(1+2+3+4+5+6) = \tfrac{21}{6} = 3.5.$$

You expect to win $3.50 per play but pay $2, so the expected net gain is $1.50. Over many games you gain money on average.

That last calculation names an idea the syllabus uses directly. A game is **fair** when the expected net gain is zero, so that neither side profits in the long run. Since the net gain is the winnings minus the stake, a fair stake is exactly the expected winnings.

KEY A game is **fair** when $E(\text{gain}) = 0$, that is, when the stake equals $E(\text{winnings})$.

In the worked-through game above, the stake of $2 is less than the expected winnings of $3.50, so the game is not fair: it favours the player.

The calculation so far ran forwards, from a complete distribution to its mean. Questions also run the other way. If a table hides two probabilities and the mean is given, you hold two facts about those unknowns: the probabilities total $1$, and the weighted sum of the values is the stated mean. Two facts, two unknowns, so write both as equations and solve them together.

**EXAMPLE 12.2**

A discrete random variable $X$ has the distribution below, and $E(X) = 1.8$. Find $a$ and $b$.

| $x$ | **0** | **1** | **2** | **3** |
|---|---|---|---|---|
| $P(X = x)$ | 0.15 | $a$ | $b$ | 0.25 |

The total-probability condition gives one equation:

$$0.15 + a + b + 0.25 = 1 \implies a + b = 0.6.$$

The stated mean gives a second, by weighting each value by its probability:

$$0(0.15) + 1(a) + 2(b) + 3(0.25) = 1.8 \implies a + 2b = 1.05.$$

Subtracting the first equation from the second removes $a$ and leaves $b = 0.45$, so $a = 0.6 - 0.45 = 0.15$.
Check both conditions on the finished table: $0.15 + 0.15 + 0.45 + 0.25 = 1$, and $0(0.15) + 1(0.15) + 2(0.45) + 3(0.25) = 1.8$, so the pair satisfies the two facts it was built from.

### 12.2.1 Variance and standard deviation HL

Expected value locates the centre, and **variance** measures the spread, exactly as in descriptive statistics. The variance is the expected squared distance from the mean. A short calculation turns that definition into a form that is easier to use.

KEY · IB HL

$$\text{Var}(X) = E((X-\mu)^2) = E(X^2) - \mu^2, \qquad \text{where } E(X^2) = \sum_i x_i^2\, P(X = x_i)$$

The second form, $E(X^2) - \mu^2$, is usually the quicker one. Find the mean of the squares, then subtract the square of the mean. The standard deviation is $\sigma = \sqrt{\text{Var}(X)}$.

**Worked through.** To find the variance and standard deviation of the variable $X$ below, start with the mean.

| $x$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $P(X = x)$ | 0.1 | 0.2 | 0.4 | 0.3 |

The mean is

$$\mu = 0(0.1) + 1(0.2) + 2(0.4) + 3(0.3) = 1.9.$$

Then $E(X^2)$:

$$E(X^2) = 0^2(0.1) + 1^2(0.2) + 2^2(0.4) + 3^2(0.3) = 0.2 + 1.6 + 2.7 = 4.5.$$

So $\text{Var}(X) = 4.5 - 1.9^2 = 4.5 - 3.61 = 0.89$, and $\sigma = \sqrt{0.89} \approx 0.943$.

### 12.2.2 Linear transformations HL

If a random variable is scaled and shifted, its mean and variance change in a predictable way, exactly as those of a data set do. For constants $a$ and $b$:

KEY · IB HL

$$E(aX+b) = aE(X)+b, \qquad \text{Var}(aX+b) = a^2\,\text{Var}(X)$$

The shift $b$ moves the mean but leaves the spread unchanged. The scale factor $a$ stretches the mean by $a$ and the variance by $a^2$, because variance is built from squared distances.

**EXAMPLE 12.3**

For the variable above, $E(X) = 1.9$ and $\text{Var}(X) = 0.89$. Let $Y = 2X+3$. Find $E(Y)$ and $\text{Var}(Y)$.

$$E(Y) = 2(1.9)+3 = 6.8, \qquad \text{Var}(Y) = 2^2(0.89) = 3.56.$$

The $+3$ leaves the variance unchanged. Only the factor of $2$ affects it, and it does so squared.

## YOUR TURN 12.2 Expected Value and Variance

### Expected value and fair games

**6.**

(a) Find $E(X)$ for $x = 1, 2, 3$ with $P(X = x) = 0.2,\ 0.3,\ 0.5$.

(b) Find $E(X)$ for $x = 0, 1, 2, 3$ with $P(X = x) = 0.1,\ 0.4,\ 0.3,\ 0.2$.

(c) A fair coin is tossed. The player wins \$3 for heads and loses \$2 for tails. Find the player's expected gain per toss.

(d) In a game you gain \$5 with probability $\frac{1}{8}$, and otherwise lose \$1. Find the expected gain, and state whether the game is fair.

**7.** A stall at a fair runs a prize wheel. The wheel pays \$8 with probability $0.1$, \$2 with probability $0.3$, and nothing otherwise.

(a) Find the expected amount won per spin, and the stake that would make the game fair.

(b) The stall charges \$2 to spin. Find the expected gain per spin for a player, and the expected loss over 50 spins.

**8.** A fair die is rolled. You win \$10 for a six, \$4 for a four or a five, and nothing otherwise. Find the stake that makes the game fair, and state whether a stake of \$4 favours the player or the organiser.

**Unknown probabilities from the mean**

**9.**

(a) A variable takes the values $1, 2, 3, 4$ with probabilities $0.2$, $a$, $b$, $0.1$, and $E(X) = 2.3$. Find $a$ and $b$.

(b) A game pays \$10 with probability $p$, \$2 with probability $q$, and nothing with probability $0.5$. The fair stake is \$3. Find $p$ and $q$.

**Variance and standard deviation**

**10.** Find $E(X^2)$, Var$(X)$ and the standard deviation for each distribution.

(a) $x = 0, 1, 2, 3$ with $P(X = x) = 0.1,\ 0.4,\ 0.3,\ 0.2$

(b) $x = -1, 0, 2$ with $P(X = x) = 0.3,\ 0.2,\ 0.5$

**Linear transformations**

**11.**

(a) Given $E(X) = 4$ and Var$(X) = 2$, find $E(3X - 1)$ and Var$(3X - 1)$.

(b) A variable has $E(X) = 3$ and Var$(X) = 5$. Find $E(X^2)$, Var$(5 - 2X)$ and the standard deviation of $5 - 2X$.

**Concept check**

**12.** Given that Var$(X) = 4$, a student writes: "*Var*$(2X + 3) = 2 \times 4 + 3 = 11$." Identify the two errors, and find the correct value.

**13.** A fair die is rolled and $X$ is the score, so $E(X) = 3.5$. A student says: "*this must be wrong, because the die can never show* $3.5$." Explain why $E(X) = 3.5$ is correct, and state what it means.

## 12.3 The Binomial Distribution

Many situations consist of the same experiment repeated: ten coin flips, twenty items checked for defects, twelve multiple-choice questions answered by guessing. When each repetition is independent and has only two possible outcomes, the number of "successes" follows one particular distribution, the **binomial**.

Four conditions define it. Check each one before reaching for the distribution.

- **Fixed number of trials.** The value of $n$ is decided in advance, not by the results.
- **Two outcomes only.** Every trial ends as a success or a failure, with nothing in between.
- **Constant probability.** Every trial has the same probability of success $p$.
- **Independence.** The result of one trial does not change the probabilities for any other.

When all four hold, write $X \sim B(n,p)$.

> DEF If $X$ counts the successes in $n$ independent trials each with success probability $p$, then $X \sim B(n,p)$, and
>
> $$P(X=r) = {}^{n}C_r p^r (1-p)^{n-r}, \qquad r = 0,1,2,\dots,n.$$

Every factor in that formula has a meaning. The term $p^r$ accounts for the $r$ successes, $(1-p)^{n-r}$ for the remaining trials failing, and ${}^{n}C_r$ for the number of different orders in which those successes can occur. In practice you evaluate the whole expression on a GDC: use the binomial pdf for $P(X=r)$ and the binomial cdf for $P(X \le r)$.

**Worked through.** A student guesses on a 10-question multiple-choice test, each question having 4 options. The number correct is $X \sim B(10, 0.25)$, since each guess is correct with probability $\frac{1}{4}$. The binomial pdf gives the probability of exactly 3 correct:

$$P(X=3) = {}^{10}C_3 (0.25)^3 (0.75)^7 \approx 0.250.$$

The binomial cdf gives the probability of at most 2 correct:

$$P(X \le 2) = P(0) + P(1) + P(2) \approx 0.526.$$

Even with 10 questions, scoring 2 or fewer is more likely than not. Guessing rarely produces a good score.

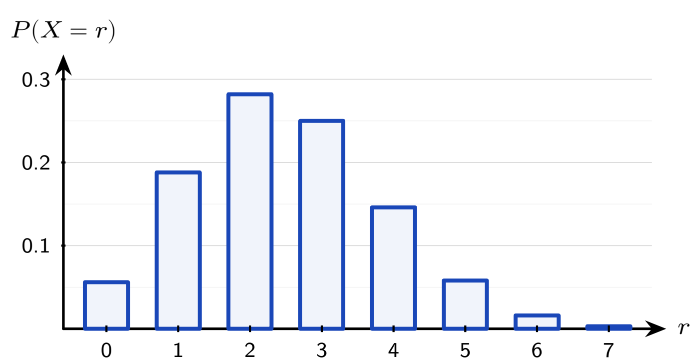


**Figure 12.2** The distribution $B(10,\ 0.25)$, the guessing student above. The tallest bar is at $r=2$ $(0.282)$, with $r=3$ close behind $(0.250)$, and the long tail to the right shows why a high score by guessing is so unlikely.

## 12.3.1 Mean and variance of the binomial

The mean could be found by summing $r\,P(X=r)$, but a formula gives it directly. For $n$ trials each succeeding with probability $p$, the results are clean.

KEY · IB For $X \sim B(n,p)$:

$$E(X) = np, \qquad \text{Var}(X) = np(1-p)$$

The mean is the result you would expect: $10$ trials at probability $0.25$ give $2.5$ successes on average. The variance is largest when $p = 0.5$, which is the value at which a single trial is least predictable.

**EXAMPLE 12.4**

A factory's items are defective with probability 0.05, independently. A sample of 20 is taken. Find the expected number of defectives and the probability that at least one is defective.
Let $X \sim B(20, 0.05)$. The expected number is

$$E(X) = 20 \times 0.05 = 1.$$

For "at least one," use the complement:

$$P(X \geq 1) = 1 - P(X = 0) = 1 - (0.95)^{20} \approx 0.642.$$

Even with a low defect rate, a sample of 20 is more likely than not to contain a defective item. The probability of "at least one" rises quickly as the sample size increases.

CAUTION Check independence and a fixed $n$ before using the binomial. Drawing items *without replacement* breaks independence, since each draw changes the next probability, so that count is not binomial.

## 12.3.2 Finding an unknown $n$ or $p$

Every binomial calculation so far began with $n$ and $p$ and produced a probability. A question can supply the probability instead and ask for one of the parameters. Both reversals start in the same place, at

$$P(X = 0) = {}^nC_0\, p^0 (1-p)^n = (1-p)^n,$$

a term with no combinatorial factor, since ${}^nC_0 = 1$. It is a single power, so it can be solved by logarithms when $n$ is the unknown and by a root when $p$ is.

**Worked through.** Each ticket in a charity draw wins a prize with probability $0.1$, independently of every other ticket. To find the least number of tickets that must be bought for the probability of winning at least one prize to exceed $0.95$, let $n$ be the number of tickets, so the number of prizes is $X \sim B(n, 0.1)$ and the complement gives

$$P(X \geq 1) = 1 - P(X = 0) = 1 - (0.9)^n.$$

The requirement $1-(0.9)^n > 0.95$ rearranges to $(0.9)^n < 0.05$. Take logarithms of both sides:

$$n \ln 0.9 < \ln 0.05 \implies n > \frac{\ln 0.05}{\ln 0.9} \approx 28.4,$$

where the inequality reverses because $\ln 0.9$ is negative. Tickets are counted in whole numbers, so $n = 29$.

Confirm at both ends: $(0.9)^{29} \approx 0.047$ and $(0.9)^{28} \approx 0.052$, so 29 tickets bring the failure probability below $0.05$ and 28 do not.

**EXAMPLE 12.5**

The number of faulty components in a batch of six is $X \sim B(6, p)$. Given that $P(X = 0) = 0.262$, find $p$.

With $r = 0$ the binomial formula reduces to a single power:

$$(1-p)^6 = 0.262.$$

Take the sixth root of both sides:

$$1 - p = 0.262^{1/6} \approx 0.800 \implies p \approx 0.200.$$

The root returns $1-p$ rather than $p$, so do not stop at $0.800$; substituting back gives $(0.8)^6 \approx 0.262$, which confirms both the value and the final subtraction.

**YOUR TURN 12.3** The Binomial Distribution

### Binomial probabilities

**14.** GDC

(a) $X \sim B(8, 0.5)$. Find $P(X = 4)$.
(b) $X \sim B(10, 0.3)$. Find $P(X = 2)$.
(c) $X \sim B(6, 0.4)$. Find $P(X \leq 2)$.
(d) $X \sim B(12, 0.25)$. Find $P(X \geq 1)$.

**15.** GDC In a factory, $8\%$ of items are defective, independently of one another. A sample of 25 items is taken.

(a) Find the probability that exactly 2 items are defective.
(b) Find the probability that more than 3 items are defective.
(c) Find the expected number of defective items in the sample.

**16.** GDC A basketball player scores each free throw with probability $0.7$, independently of the others. She takes 15 free throws.

(a) Find the probability that she scores at least 12.
(b) Find the probability that she scores between 8 and 11 inclusive.

### Mean and variance

**17.**

(a) $X \sim B(20, 0.1)$. Find the mean and the variance.
(b) $X \sim B(n, p)$ with $E(X) = 6$ and $\text{Var}(X) = 4.2$. Find $n$ and $p$.

### When the binomial applies

**18.**

(a) A box holds 10 items, 3 of them defective, and 4 items are drawn without replacement. Explain why the number of defective items drawn is not binomial.
(b) A student counts how many of the 30 days in June are rainy. Give one reason why $B(30, p)$ may not be a good model for this count.

**Finding an unknown $n$ or $p$**

**19.**

(a) Each item made by a machine is faulty with probability $0.03$, independently. Find the least number of items that must be checked for the probability of finding at least one faulty item to exceed $0.9$.

(b) $X \sim B(5, p)$ and $P(X = 0) = 0.07776$. Find $p$.

**20.**

(a) $X \sim B(n, 0.2)$. Find the least value of $n$ for which $P(X = 0) < 0.01$.

(b) $X \sim B(4, p)$ and $P(X = 4) = 0.1296$. Find $p$.

**Concept check**

**21.** GDC $X \sim B(10, 0.4)$. A student writes:
"$P(X \geq 3) = 1 - P(X \leq 3) \approx 1 - 0.382 = 0.618$." Identify the error, and find the correct value of $P(X \geq 3)$.

## 12.4 Continuous Random Variables HL

A discrete variable jumps between separate values. A continuous one, such as a height, a mass or a waiting time, can take any value in a range. You cannot list a probability for each value, because there are infinitely many of them. Something has to replace the list.

The bags of rice show what replaces it. Group the 5000 masses into bars one gram wide. The probability of a bag falling in a given group is then the *area* of that group's bar. Now halve the bar width. There are twice as many bars, each half as wide and about as tall as before, and the total area is unchanged at 1. Halve it again, and again. The tops of the bars approach a smooth curve, and the area under that curve is still 1.

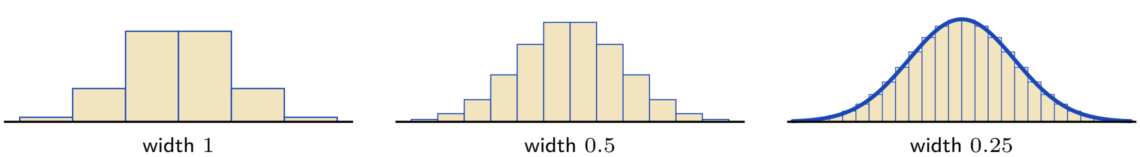


**Figure 12.3** The same 5000 masses grouped into bars of decreasing width. Every version encloses total area $1$; as the bars narrow, their tops approach a smooth curve. That curve is the density.

So probability for a continuous variable is **area**, and the curve whose area gives it is called a **probability density function**, written $f(x)$. Two conditions make $f$ a valid density: it is never negative, and the total area under it is 1.

DEF HL A **probability density function** $f(x)$ satisfies

$$f(x) \geq 0 \quad \text{for all } x, \qquad \int_{-\infty}^{\infty} f(x)\,dx = 1,$$

and probabilities are areas: $P(a \leq X \leq b) = \displaystyle\int_a^b f(x)\,dx.$

The symbol $\int$ is new. It is the notation for **the area under a curve**, and $\int_a^b f(x)\,dx$ means the area under $y = f(x)$ between $x = a$ and $x = b$. Ch. 15 defines it and shows how to evaluate one, and the examples below use those methods; if you meet this section before Ch. 15, read each integral as an area and return to the calculations afterwards.

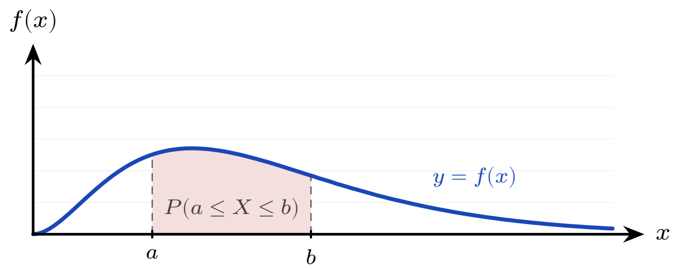


**Figure 12.4** For a continuous variable, a probability is an area. The shaded region between $a$ and $b$ is $P(a \leq X \leq b)$, and the whole area under the curve is $1$.

One consequence of this looks strange at first.

- A bar of zero width has zero area, so $P(X = c) = 0$ for every single value $c$.
- This does not mean the value is impossible. It means that no exact value carries a share of the total. A person's height is never exactly $165$ cm to unlimited precision, only somewhere in a range near it.
- The endpoints therefore contribute nothing, so $P(a \leq X \leq b)$ and $P(a < X < b)$ are the same number. Here you may be careless about $<$ and $\leq$ in a way you cannot be with a discrete variable.

The expected value and variance carry over from the discrete case, with integrals replacing sums.

KEY · IB HL

$$E(X) = \int_{-\infty}^{\infty} x\,f(x)\,dx, \qquad \text{Var}(X) = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2$$

Compare these with the discrete versions in §12.2. The structure is identical, and only the machinery has changed: a sum over a list of values has become an area over a range of them. As before, the standard deviation is $\sigma = \sqrt{\text{Var}(X)}$, and it is the number to quote when you want a spread in the same units as $X$.

Three numbers describe where a density sits, and each is read off the curve in its own way.

- The **mode** is the value at which $f$ is highest, that is, the peak of the curve.
- The **median** is the value that cuts the area in half, with $0.5$ on each side.
- The **mean** is the balance point, $\int x f(x)\,dx$.

For a symmetric curve with a single peak all three sit together. For a lopsided one they do not, and the density worked through below shows all three.

**Worked through.** Take the density $f(x) = kx$ for $0 \le x \le 2$, and $f(x) = 0$ elsewhere. To find $k$, use the fact that the total area must be 1:

$$\int_0^2 kx\,dx = k\left[\tfrac{x^2}{2}\right]_0^2 = 2k = 1 \implies k = \tfrac{1}{2}.$$

Then

$$P(X < 1) = \int_0^1 \tfrac{x}{2}\,dx = \left[\tfrac{x^2}{4}\right]_0^1 = \tfrac{1}{4}.$$

The mean:

$$E(X) = \int_0^2 x \cdot \tfrac{x}{2}\,dx = \int_0^2 \tfrac{x^2}{2}\,dx = \left[\tfrac{x^3}{6}\right]_0^2 = \tfrac{8}{6} = \tfrac{4}{3}.$$

For the variance, first find $E(X^2)$, then subtract $\mu^2$:

$$E(X^2) = \int_0^2 x^2 \cdot \tfrac{x}{2}\,dx = \int_0^2 \tfrac{x^3}{2}\,dx = \left[\tfrac{x^4}{8}\right]_0^2 = 2,$$

$$\text{Var}(X) = 2 - \left(\tfrac{4}{3}\right)^2 = 2 - \tfrac{16}{9} = \tfrac{2}{9}, \qquad \sigma = \sqrt{\tfrac{2}{9}} = \tfrac{\sqrt{2}}{3} \approx 0.471.$$

The **median** $m$ splits the area in half:

$$\int_0^m \tfrac{x}{2}\,dx = \tfrac{1}{2} \implies \tfrac{m^2}{4} = \tfrac{1}{2} \implies m = \sqrt{2}.$$

The **mode** is where $f$ is greatest. Since $f(x) = \frac{x}{2}$ increases across $[0, 2]$, the mode is at the right end, $x = 2$.

Collect the three centres: mode $2$, median $\sqrt{2} \approx 1.414$, mean $\frac{4}{3} \approx 1.333$. Mean $<$ median $<$ mode: the density rises towards $x = 2$, so its long tail lies to the *left*. That is negative skew, in the language of Ch. 10.

CAUTION The mean, median, and mode of a continuous variable usually differ. The mode maximises $f(x)$; the median halves the area; the mean is the balance point $\int x f(x)\,dx$. For a symmetric density with a single peak all three coincide; for a skewed density they generally differ.

### 12.4.1 Piecewise densities

A density need not be a single formula. Any function that is non-negative and encloses total area 1 qualifies, including one built from pieces. The only new requirement is care with the arithmetic: integrals must be split at the joins, and the median may lie in either piece.

**Worked through.** A continuous random variable has density

$$f(x) = \begin{cases} \dfrac{2x}{3} & 0 \le x \le 1, \\ \dfrac{3-x}{3} & 1 < x \le 3, \\ 0 & \text{otherwise.} \end{cases}$$

To verify that $f$ is a valid density, find $P(X > 2)$ and find the median, start with the two conditions for a density. Both pieces are non-negative, and the total area splits at $x = 1$:

$$\int_0^1 \tfrac{2x}{3}\,dx + \int_1^3 \tfrac{3-x}{3}\,dx = \Big[\tfrac{x^2}{3}\Big]_0^1 + \tfrac{1}{3}\Big[3x - \tfrac{x^2}{2}\Big]_1^3 = \tfrac{1}{3} + \tfrac{2}{3} = 1.\checkmark$$

For $P(X > 2)$ only the second piece matters:

$$P(X > 2) = \int_2^3 \tfrac{3-x}{3}\,dx = \tfrac{1}{3}\Big[3x - \tfrac{x^2}{2}\Big]_2^3 = \tfrac{1}{6}.$$

For the median, first locate it: the area up to $x = 1$ is $\frac{1}{3} < \frac{1}{2}$, so the median lies in the second piece. Collect the remaining area there:

$$\tfrac{1}{3} + \int_1^m \tfrac{3-x}{3}\,dx = \tfrac{1}{2} \implies 3m - \tfrac{m^2}{2} - \tfrac{5}{2} = \tfrac{1}{2} \implies m^2 - 6m + 6 = 0,$$

giving $m = 3 - \sqrt{3} \approx 1.27$ (taking the root inside the interval). The mode is $x = 1$, where the two pieces meet at the peak of the triangle.

### 12.4.2 Modes by calculus

When a density rises and then falls, its mode lies inside the interval, and you find it by the standard method: differentiate and solve $f'(x) = 0$. A mode can also sit at an endpoint or at a corner, as in the two densities above, so always confirm where the maximum actually is before answering.

**Worked through.** A continuous random variable has density $f(x) = kx^2(3 - x)$ for $0 \le x \le 3$, and zero elsewhere. To find $k$, the mode and $E(X)$, start from the fact that the total area is 1:

$$\int_0^3 k(3x^2 - x^3)\,dx = k\Big[x^3 - \tfrac{x^4}{4}\Big]_0^3 = k\left(27 - \tfrac{81}{4}\right) = \tfrac{27k}{4} = 1 \implies k = \tfrac{4}{27}.$$

For the mode, maximise $f$:

$$f'(x) = k(6x - 3x^2) = 3kx(2 - x) = 0 \implies x = 0 \text{ or } x = 2.$$

Since $f$ rises on $(0,2)$ and falls on $(2,3)$, the mode is $x = 2$. Finally,

$$E(X) = \int_0^3 k(3x^3 - x^4)\,dx = \tfrac{4}{27}\left[\tfrac{3x^4}{4} - \tfrac{x^5}{5}\right]_0^3 = \tfrac{4}{27}\cdot\tfrac{243}{20} = \tfrac{9}{5}.$$

Mean $1.8$, mode $2$: the longer tail of the density lies to the left of the peak, so the balance point sits below it.

**YOUR TURN 12.4** Continuous Random Variables HL

**Finding $k$ and probabilities**

**22.**

(a) A density is $f(x) = kx^2$ for $0 \le x \le 3$, and zero elsewhere. Find $k$.
(b) A density is $f(x) = k$ for $0 \le x \le 4$, and zero elsewhere. Find $k$ and $P(X < 1)$.
(c) A density is $f(x) = \dfrac{x}{8}$ for $0 \le x \le 4$, and zero elsewhere. Find $P(X > 2)$.
(d) For the density in part (c), write down $P(X = 2)$, and explain why $P(X \le 2) = P(X < 2)$.

**Mean, variance, median and mode**

**23.**

(a) For the density $f(x) = \dfrac{x}{8}$, $0 \le x \le 4$, find $E(X)$.
(b) For the same density, find Var$(X)$.
(c) A density is $f(x) = 3x^2$ for $0 \le x \le 1$, and zero elsewhere. Find the median.

**24.** A continuous random variable has density $f(x) = kx(4 - x)$ for $0 \le x \le 4$, and zero elsewhere.

(a) Find $k$.
(b) Write down the mode of this density, and justify your answer.
(c) Find $E(X)$.
(d) Find Var$(X)$.

**25.** A continuous random variable has density $f(x) = k(4x - x^3)$ for $0 \le x \le 2$, and zero elsewhere.

(a) Find $k$.
(b) Use calculus to find the mode, giving your answer in exact form.
(c) Find $E(X)$.

**Piecewise densities**

**26.** A variable has density $f(x) = x$ for $0 \le x \le 1$, $f(x) = \frac{1}{2}$ for $1 < x \le 2$, and zero elsewhere.

(a) Verify that the total area is 1, and find $P(X < 1.5)$.

(b) Find the median.

**27.** A variable has density

$$f(x) = \begin{cases} \dfrac{x}{3} & 0 \le x \le 2, \\ \dfrac{2(3-x)}{3} & 2 < x \le 3, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Verify that $f$ is a valid density.

(b) Find the median, giving your answer in exact form.

(c) Find $P(X > 2.5)$.

**Concept check**

**28.** A density is $f(x) = 2x$ for $0 \le x \le 1$, and zero elsewhere. A student says: "*$f(0.8) = 1.6$, which is greater than 1, so this cannot be a probability density.*" Explain why the student is wrong.

## 12.5 The Normal Distribution

If the histogram of bag masses is refined far enough, the tops of the bars approach a smooth curve, which is close to the **normal distribution**. Everything from the last section still applies to it: probability is area, and the total area is $1$.

What earns this one curve a section of its own is how often it occurs. Heights, masses, measurement errors and examination marks often follow it approximately. The reason was given at the start of the chapter: when a quantity is the sum of many small independent effects, none of which dominates, its distribution is close to normal. That situation is common, so the curve is used widely in statistics.

The name can mislead. *Normal* here does not mean usual, correct or healthy. It is a nineteenth-century label that survived, and a great many real quantities are not normal at all. Incomes, city populations and waiting times are all strongly skewed. Before using the curve, check that the data actually resemble it.

Every normal curve, whatever its mean and spread, has the same four properties.

- It is symmetric about the mean, so the mean, median and mode are equal and all sit at the peak.
- It has a single maximum, at the centre, and falls away on both sides.
- The tails approach the horizontal axis but never meet it. Every interval of values, however far from the mean, therefore has a positive probability, however small.
- The total area beneath it is $1$, as it is for any density.

Two numbers fix a normal distribution completely, and each does a different job.

- The **mean** $\mu$ locates the centre. Changing $\mu$ slides the whole curve left or right without altering its shape at all.
- The **standard deviation** $\sigma$ sets the width. Increasing $\sigma$ stretches the curve sideways, and because the area underneath must stay equal to 1, stretching it sideways also flattens it. A large $\sigma$ therefore gives a wide, low bell and a small $\sigma$ a narrow, tall one.

We write $X \sim N(\mu, \sigma^2)$. Note that the second entry is the *variance*, not the standard deviation, so take a square root before using $\sigma$.

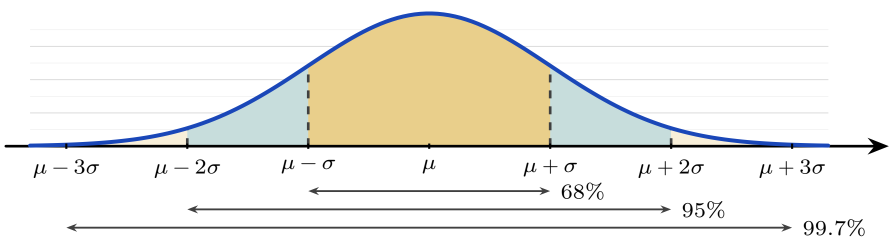


**Figure 12.5** The normal curve, symmetric about $\mu$. The three shaded bands reach one, two and three standard deviations either side of the mean, and hold about $68\%$, $95\%$ and $99.7\%$ of the values. Almost nothing lies beyond three standard deviations.

The shape of the normal curve gives a quick approximation, the **68–95–99.7 rule**. It lets you judge, from $\mu$ and $\sigma$ alone, where most of the data lie, without a calculator.

KEY **The 68–95–99.7 rule.** For $X \sim N(\mu, \sigma^2)$, approximately

$$68\% \text{ of values lie in } \mu \pm \sigma, \qquad 95\% \text{ in } \mu \pm 2\sigma, \qquad 99.7\% \text{ in } \mu \pm 3\sigma.$$

Read in reverse, the same rule measures how surprising a value is. Falling outside two standard deviations happens to about 5% of values, roughly one in twenty. Falling outside three happens to about $0.3\%$, roughly one in 370. That is why a reading three standard deviations from the mean is often treated, in quality control for example, as a signal that something has changed, rather than as ordinary variation.

The curve has an equation, and although you will not have to use it, seeing it explains where the shape comes from.

KEY **The normal density** (not assessed):

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

Read the exponent first, because it carries the whole shape. The quantity $(x-\mu)^2$ is a parabola: it is zero at $x=\mu$ and grows as $x$ moves away from the mean, and it depends only on the *distance* from $\mu$, not on the direction. The minus sign turns that growth into decay, since $e^{-u}$ is largest when $u=0$ and shrinks towards zero as $u$ grows. So the curve is tallest at the mean, falls away at the same rate on both sides, and never quite reaches the axis. The $2\sigma^2$ underneath sets the speed of the fall: a large $\sigma$ makes the exponent grow slowly, giving a wide, low bell. The factor $\dfrac{1}{\sigma\sqrt{2\pi}}$ in front changes no part of the shape; it is the number that makes the total area exactly $1$.

This formula is not in the formula booklet, and you are not expected to use it. Normal probabilities in this course are found with a GDC, or from symmetry. It is here so that the bell is a consequence of something rather than a shape you are asked to accept.

**Worked through.** The heights of adult women are modelled by $X \sim N(165, 6^2)$ (in cm). Both $P(X<172)$ and $P(158<X<172)$ are areas under the normal curve, read from a GDC with $\mu=165$, $\sigma=6$: $P(X<172)\approx 0.878$ and $P(158<X<172)\approx 0.757$. The second answer is larger than the 68% of the rule of thumb, because 158 and 172 are about $1.17$ standard deviations either side of the mean, a little more than one.

Questions ask for three shapes of area, and a calculator handles all three from the same two entries. An area to the left needs a lower limit far below the mean, an area to the right needs an upper limit far above it, and an area between two values needs both limits as given.

**EXAMPLE 12.6**

The masses of eggs from a farm are modelled by $X \sim N(58, 4^2)$, in grams. Find (a) $P(X>65)$, (b) $P(54<X<62)$.

Enter $\mu=58$ and $\sigma=4$ on a GDC.

(a) A right-hand tail has no natural upper limit, so supply one far above the mean, such as $10^{99}$, and take the area from $65$ upwards:

$$P(X>65)\approx 0.0401.$$

(b) Here both limits are given:

$$P(54<X<62)\approx 0.683.$$

The limits in (b) sit exactly one standard deviation either side of $58$, so the second answer is the $68\%$ of the rule of thumb, computed rather than estimated.

A probability also answers a question about a whole batch. If each of $N$ items follows the same normal distribution independently, the count that falls in some range is binomial with that probability, and its expected value is $N$ times the probability. So the expected number in a batch is a normal calculation followed by one multiplication.

**EXAMPLE 12.7**

The lifetime of a light bulb is modelled by $X \sim N(1200, 100^2)$, in hours. A hospital installs 400 of these bulbs. Find (a) the probability that a bulb fails before 1000 hours, (b) the expected number of the 400 that fail before 1000 hours.

(a) With $\mu = 1200$ and $\sigma = 100$, a lower limit far below the mean gives

$$P(X < 1000) \approx 0.0228.$$

(b) Multiply the batch size by that probability, carried at more digits than the rounded $0.0228$ of part (a):

$$400 \times 0.022750 = 9.100,$$

so about $9$ bulbs are expected to fail before 1000 hours.

Round only at the end, and read the answer as a long-run average: a particular batch of 400 may contain any number of early failures.

## YOUR TURN 12.5 The Normal Distribution

### Shape and the 68–95–99.7 rule

**29.** The curves of $A \sim N(40, 3^2)$, $B \sim N(40, 6^2)$ and $C \sim N(55, 3^2)$ are drawn on the same axes.

(a) State which two curves have the same shape, and describe how they are related.
(b) State which curve has the lowest peak, and explain why.

**30.** Heights are modelled by $X \sim N(170, 8^2)$, in cm. Without a GDC, use the 68–95–99.7 rule to estimate the percentage of heights

(a) between $162$ cm and $178$ cm;
(b) between $146$ cm and $194$ cm;
(c) above $186$ cm.

**Normal probabilities**

**31.** GDC

(a) $X \sim N(50, 10^2)$. Find $P(X < 60)$.
(b) $X \sim N(100, 15^2)$. Find $P(X > 120)$.
(c) $X \sim N(20, 4^2)$. Find $P(18 < X < 26)$.
(d) $Z \sim N(0, 1)$. Find $P(Z < 1.5)$.
(e) $X \sim N(500, 50^2)$. Find $P(X < 400)$.
(f) $X \sim N(30, 16)$. Find $P(X > 34)$.

**32.** GDC The masses of packets of rice are modelled by $X \sim N(500, 25^2)$, in grams.

(a) Find the probability that a packet weighs more than 540 g.
(b) Find $P(450 < X < 575)$, and compare your answer with the value the 68–95–99.7 rule gives.
(c) In a batch of 800 packets, find the expected number weighing more than 540 g.
(d) Write down the range of masses within which about $95\%$ of packets lie.

**Concept check**

**33.** $X \sim N(50, 10^2)$. A student says: "*about $68\%$ of values lie within one standard deviation of the mean, so $P(X < 60) \approx 0.68$.*" Identify the error, and estimate $P(X < 60)$ correctly without a GDC.

## 12.6 Standardisation and the Inverse Normal

Every normal distribution is the same bell, just shifted and scaled. **Standardising** expresses any value as the number of standard deviations it sits from its mean. That number is the **$z$-value**, and it converts any normal variable to the **standard normal** $Z \sim N(0, 1)$.

KEY · IB

$$z = \frac{x - \mu}{\sigma}$$

Standardising does two things at once. Subtracting $\mu$ moves the centre of the distribution to $0$. Dividing by $\sigma$ then rescales the spread, so that one unit of $z$ means one standard deviation. The units cancel in that division, and that is what makes comparison possible: $z$ is a pure number with no units attached, so a mass in grams and a time in seconds become directly comparable once standardised.

Read a $z$-value as follows.

- $z > 0$ means the value lies above the mean, and $z < 0$ means it lies below.

- $z = 0$ means the value is exactly the mean.
- The size of $z$ is the distance from the mean, measured in standard deviations. A $z$-value of $1.5$ means "one and a half standard deviations above the mean", whatever the original units were.

The 68–95–99.7 rule of §12.5 now reads very simply. About $68\%$ of values have $|z| < 1$, about $95\%$ have $|z| < 2$, and about $99.7\%$ have $|z| < 3$. Standardising therefore lets you compare values drawn from different normal distributions, and it is also how you work backwards from a probability to a value.

**Worked through.** Exam scores are $N(60, 12^2)$. To see how many standard deviations above the mean a score of 78 lies, standardise it:

$$z = \frac{78 - 60}{12} = \frac{18}{12} = 1.5.$$

The score sits $1.5$ standard deviations above the mean. A student scoring 78 on a different exam with $N(70, 4^2)$ would have $z = 2$, and so be higher *relative to their cohort*, despite the equal raw mark.

### 12.6.1 The inverse normal

Often the question runs in the other direction. Instead of being given a value and asked for a probability, you are given a probability and asked for the value that produces it. "What score is the 90th percentile?" is a question of this kind. The **inverse normal** on a GDC takes a probability and returns the corresponding value $x$.

One detail needs care. Unless you choose another tail setting, a calculator's inverse normal uses the area to the *left* of the unknown value, but examination questions are often phrased in terms of the area to the right. Converting between the two is a single subtraction, and forgetting it produces an answer on the wrong side of the mean.

KEY **Using the inverse normal.**

1. Write down the probability you are given, and decide whether it is the area to the left or to the right of the unknown value.
2. If it is a right-hand area, subtract it from $1$. The calculator needs the left-hand area.
3. Enter that left-hand area together with $\mu$ and $\sigma$.
4. Read off $x$, then check that it falls on the expected side of the mean.

Step 4 is worth the few seconds it takes. An area greater than $0.5$ must return a value above the mean, and an area less than $0.5$ must return one below it. This single check catches the subtraction error before it costs marks.

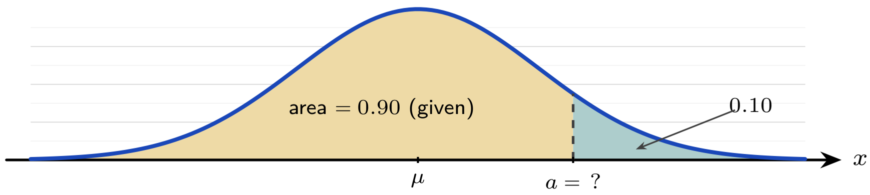


**Figure 12.6** The inverse normal reverses the usual question. Instead of being given $a$ and finding the area, you are given the area and must find $a$. A calculator expects the area to the *left* of $a$ (gold), so a question phrased as a right-hand tail (teal), here $0.10$, must be turned into its left-hand partner $0.90$ first.

**Worked through.** To find the height $a$ below which 90% of women fall, for heights $X \sim N(165, 6^2)$, you need $a$ with $P(X < a) = 0.90$. This is already a left-hand area, so the inverse normal gives

$$a \approx 172.7 \text{ cm}.$$

Equivalently, the 90th percentile sits at $z \approx 1.282$, so $a = 165 + 1.282(6) \approx 172.7$, matching.

## 12.6.2 Finding an unknown $\mu$ or $\sigma$

The inverse normal also lets you recover a missing parameter. Suppose you know one probability and the value attached to it, but either $\mu$ or $\sigma$ is missing. The method has three steps.

1. Use the inverse normal on the *standard* curve, $N(0, 1)$, to turn the probability into a $z$-value.
2. Substitute that $z$, the known value $x$, and whichever parameter you have into $z = \dfrac{x - \mu}{\sigma}$.
3. Solve the resulting equation for the one unknown that remains.

Multiplied through by $\sigma$, the equation becomes $x = \mu + z\sigma$, which is linear in $\mu$ and $\sigma$, so this last step is ordinary algebra.

**Worked through.** A machine fills bottles so that volumes are normal with $\sigma = 5$ mL. To find the mean $\mu$ given that 80% of bottles contain less than 20 mL, apply the three steps. Here $P(X < 20) = 0.80$. The inverse normal for the standard curve gives the matching $z$:

$$z = 0.8416, \qquad \text{so} \qquad 0.8416 = \frac{20 - \mu}{5}.$$

Solving, $20 - \mu = 4.21$, hence $\mu \approx 15.8$ mL.

## 12.6.3 Both parameters unknown

When neither $\mu$ nor $\sigma$ is known, a single probability statement is not enough, because one equation cannot determine two unknowns. Two statements are required.

Rearranging $z = \dfrac{x - \mu}{\sigma}$ into $x = \mu + z\sigma$ shows why this works. Each statement produces one equation of that form, linear in $\mu$ and $\sigma$, and a pair of linear equations in two unknowns can be solved simultaneously in the usual way.

**Worked through.** The masses of a species of fish are normally distributed. It is known that 5% of the fish have mass below 50 g and 10% have mass above 80 g. To find $\mu$ and $\sigma$, translate each statement into a $z$-value with the inverse normal:

$$P(X < 50) = 0.05 \implies z_1 = -1.645, \qquad P(X < 80) = 0.90 \implies z_2 = 1.282.$$

Each gives an equation $x = \mu + z\sigma$:

$$50 = \mu - 1.645\sigma, \qquad 80 = \mu + 1.282\sigma.$$

Subtracting, $30 = 2.926\sigma$, so $\sigma \approx 10.3$ g. Substituting back, $\mu = 50 + 1.645(10.25) \approx 66.9$ g.

Note the signs carefully: a value *below* the mean has a negative $z$. One wrong sign here changes both answers, so check each $z$ against the picture before solving.

**YOUR TURN 12.6** Standardisation and the Inverse Normal

**Standardising and the inverse normal**

**34.** GDC

(a) $X \sim N(60, 12^2)$. Find the $z$-value of $x = 84$.
(b) $X \sim N(70, 5^2)$. Find the value $x$ for which $P(X < x) = 0.95$.
(c) $X \sim N(80, 10^2)$. Find the value $x$ for which $P(X > x) = 0.15$.
(d) For $X \sim N(50, 8^2)$, find the lower quartile, the upper quartile, and the interquartile range.

**Finding $\mu$ or $\sigma$**

**35.** GDC

(a) $X \sim N(\mu, 8^2)$ with $P(X < 50) = 0.30$. Find $\mu$.
(b) $X \sim N(\mu, 6^2)$ with $P(X > 40) = 0.20$. Find $\mu$.
(c) $X \sim N(100, \sigma^2)$ with $P(X < 90) = 0.25$. Find $\sigma$.

**36.** GDC Scores are normally distributed with mean 50. Given that $10\%$ of scores exceed 60, find the standard deviation $\sigma$.

**Comparing with $z$-values**

**37.** GDC Two examinations have marks $A \sim N(65, 10^2)$ and $B \sim N(72, 6^2)$. A student scores 75 in $A$ and 80 in $B$. Find both $z$-values, and state which is the better performance relative to the cohort.

**Both parameters unknown**

**38.** GDC

(a) $X \sim N(\mu, \sigma^2)$ with $P(X < 20) = 0.10$ and $P(X < 35) = 0.90$. Find $\mu$ and $\sigma$.
(b) $X \sim N(\mu, \sigma^2)$ with $P(X < 30) = 0.20$ and $P(X > 52) = 0.05$. Find $\mu$ and $\sigma$.

**Concept check**

**39.** GDC $X \sim N(120, 15^2)$. To find $x$ with $P(X > x) = 0.2$, a student enters an area of $0.2$ into the inverse normal and obtains $x \approx 107.4$. Identify the error, find the correct value of $x$, and explain how a quick check would have caught the mistake.

Chapter 12 · Probability Distributions

# Reflections

TOK The same situation can often be described by more than one model: a binomial count can be approximated by a normal curve (Challenge Question 28), and a real data set is never exactly either.

What criteria should decide between two models that both fit? Simplicity, accuracy, and agreement with existing theory can point different ways, and the choice is made by people.

IM In 1846 the Belgian astronomer Adolphe Quetelet found this curve in the chest measurements of 5738 Scottish soldiers, and took it as support for an idea he had introduced in 1835, *l'homme moyen*, the "average man." He treated this statistical fiction as an ideal that real people deviate from, although no actual person has all the average values at once. The normal curve is a fact about variation; the "average man" was an interpretation added to it. When we summarise a population by its mean, are we describing a real thing, or creating one?

RW The frequent appearance of the bell curve has an explanation: the central limit theorem. It says that the sum of many independent random effects, none of them dominant, tends toward a normal distribution, whatever the shape of the individual effects. This is why measurement errors, heights, and the means of large samples are often approximately normal. De Moivre found the curve as a limit of the binomial in 1733, and Laplace gave a first version of the theorem in 1810; general proofs came from Lyapunov (1901) and Lindeberg (1922).

EXT The binomial becoming normal is a theorem, not a loose comparison. Plot $B(n, 0.5)$ for increasing $n$ and the discrete bars take the shape of the continuous bell. Galton built a machine, the quincunx, to show it physically: balls dropped through rows of pegs, each bounce an independent left-or-right trial, pile up at the bottom in a binomial distribution whose shape is close to normal. The binomial and normal distributions of this chapter are connected: as $n$ increases, the first approaches the second.

# End-of-Chapter Problems
# Chapter 12: Probability Distributions

**1.** A discrete random variable $X$ has the distribution below.

| $x$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $P(X = x)$ | 0.2 | $k$ | 0.35 | 0.15 |

(a) Find $k$. [2]
(b) Find $E(X)$. [2]
(c) Find Var$(X)$. [3]

**2.** A charity raffle sells 500 tickets at $2 each. There is one prize of $300 and five prizes of $50.

(a) Find the expected winnings (before cost) for a single ticket. [3]
(b) Hence find the expected gain or loss per ticket, and comment. [2]

**3.** The random variable $X$ has $E(X) = 5$ and Var$(X) = 4$.

(a) Find $E(X^2)$. [1]
(b) The variable $Y = aX + b$, where $a > 0$, has $E(Y) = 0$ and Var$(Y) = 1$. Find $a$ and $b$. [3]

**4.** GDC In a production line, 4% of items are faulty, independently. Items are packed in boxes of 40. Let $X$ be the number of faulty items in a box.

(a) State the distribution of $X$. [1]
(b) Find $P(2 \leq X \leq 4)$. [2]
(c) Find the probability that $X$ is greater than its mean. [2]
(d) Write down the most likely number of faulty items in a box. [1]

**5.** GDC A fair six-sided die is rolled 30 times. Let $X$ be the number of sixes.

(a) State the distribution of $X$ and find its mean. [2]
(b) Find $P(X = 5)$. [2]
(c) Find $P(X \geq 7)$. [2]

**6.** A continuous random variable has density $f(x) = k(x + 1)$ for $0 \leq x \leq 2$, and zero elsewhere.

(a) Find $k$. [2]
(b) Find $P(X < 1)$. [2]

(c) Find $E(X)$. [3]

(d) Find the median and the mode. [3]

**7.** GDC The heights of a species of plant are $X \sim N(165, 9^2)$ in cm.

(a) Find $P(X < 150)$. [2]

(b) Find $P(150 < X < 180)$. [2]

(c) Find the proportion of plants taller than 175 cm. [2]

**8.** GDC Test scores are normally distributed with mean 500 and standard deviation 80.

(a) Find the score exceeded by 30% of candidates. [2]

(b) The middle 90% of scores lie between $a$ and $b$, symmetric about the mean. Find $a$ and $b$. [3]

**9.** The random variable $X$ is normally distributed with $P(X < 20) = 0.3$ and $P(X > 40) = 0.3$.

(a) Write down the mean of $X$. [1]

(b) Find $P(20 < X < 40)$. [1]

(c) Find $P(X < 40 \mid X > 20)$. [3]

**10.** GDC Reaction times are $X \sim N(250, 40^2)$ in milliseconds. A reaction is "slow" if it exceeds 320 ms.

(a) Find the probability that a single reaction is slow. [2]

(b) Thirty reactions are recorded independently. Find the expected number of slow reactions. [2]

(c) Find the probability that at least one of the thirty is slow. [3]

**11.** A continuous random variable has density $f(x) = \frac{1}{8}(4 - x)$ for $0 \le x \le 4$, and zero elsewhere.

(a) Verify that $f$ is a valid probability density function. [3]

(b) Find $P(X < 2)$. [2]

(c) Find the mode. [1]

**12.** A continuous random variable $X$ has probability density function $f(x) = ax + b$ for $0 \le x \le 2$, and $f(x) = 0$ elsewhere. It is known that $E(X) = \frac{11}{9}$.

(a) Find the values of $a$ and $b$. [5]

(b) Find $P(X > 1)$. [2]

**13.** A discrete random variable $X$ takes the values $1, 2, 3, 4$ with $P(X = 1) = 0.2$, $P(X = 2) = a$, $P(X = 3) = b$, and $P(X = 4) = 0.1$. It is known that $E(X) = 2.4$.

(a) Find the values of $a$ and $b$. [4]

(b) Find $P(X > 2 \mid X \geq 2)$. [3]

**14.** A stall game costs $\$c$ to play. A bag holds 3 red and 7 blue counters, and the player draws two counters without replacement. The player wins \$10 if both counters are red, \$2 if exactly one is red, and nothing otherwise.

(a) Find the expected winnings from a single play. [2]

(b) Find the value of $c$ for which the game is fair. [1]

(c) The stall sets $c = 2$. Find the player's expected loss over 100 plays. [2]

**15.** GDC A gardener plants 12 seeds. Each seed germinates with probability 0.85, independently of the others. Let $X$ be the number that germinate.

(a) State two conditions that make a binomial model appropriate here. [2]

(b) Find $P(X = 10)$. [2]

(c) Find $P(X \geq 10)$. [2]

(d) Find the mean and standard deviation of $X$. [2]

**16.** GDC A darts player hits the bullseye with probability 0.3 on each throw, independently.

(a) In 8 throws, find the probability of exactly 3 bullseyes. [2]

(b) Find the least number of throws needed for the probability of at least one bullseye to exceed 0.98. [3]

**17.** GDC The random variable $X \sim B(10, p)$ satisfies $P(X = 0) = 0.0282$.

(a) Find the value of $p$. [3]

(b) Find $P(X = 3)$. [2]

(c) Write down $E(X)$ and Var$(X)$. [2]

**18.** GDC The lifetime of a battery is $X \sim N(150, 18^2)$, in hours.

(a) Find the probability that a battery lasts less than 125 hours. [2]

(b) The manufacturer wants no more than 2% of batteries to fail before a guaranteed time of $t$ hours. Find the largest possible value of $t$. [2]

(c) Given that a battery has lasted 160 hours, find the probability that it lasts more than 180 hours. [2]

**19.** GDC The masses of apples are $X \sim N(160, 20^2)$, in grams. The heaviest 8% are sold as premium, and the lightest 5% are rejected.

(a) Find the minimum mass of a premium apple. [2]
(b) Find the maximum mass of a rejected apple. [2]
(c) Write down the proportion of apples that are neither premium nor rejected. [1]

**20.** GDC Journey times are normally distributed with mean 25 minutes and standard deviation $\sigma$. It is known that $P(X < 30) = 0.85$.

(a) Find $\sigma$. [3]
(b) Hence write down $P(X < 20)$, justifying your answer. [2]

**21.** GDC The masses of loaves of bread are normally distributed with unknown mean $\mu$ and standard deviation $\sigma$. It is known that 2% of loaves have mass below 300 g and 5% have mass above 340 g.

(a) Write down two equations for $\mu$ and $\sigma$ using standardised values. [3]
(b) Find $\mu$ and $\sigma$. [3]

**22.** A continuous random variable has density $f(x) = kx(2-x)^2$ for $0 \leq x \leq 2$, and zero elsewhere.

(a) Show that $k = \frac{3}{4}$. [2]
(b) Use calculus to find the mode of $X$. [2]
(c) Find $E(X)$. [2]
(d) Find Var$(X)$. [3]

**23.** GDC Scores on an examination are $X \sim N(55, 10^2)$. The pass mark is 45.

(a) Find the probability that a randomly chosen candidate passes. [2]
(b) Six candidates are chosen at random. Find the probability that exactly five of them pass. [3]
(c) Find the probability that at least five of them pass. [2]

**24.** A continuous random variable has density

$$f(x) = \begin{cases} k & 0 \leq x \leq 1, \\ k(2-x) & 1 < x \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Show that $k = \frac{2}{3}$. [3]
(b) Sketch the graph of $f$. [2]
(c) Find $P(X > 1.5)$. [2]
(d) Find the median of $X$. [2]
(e) Find $E(X)$. [3]

**25.** A discrete random variable $X$ has probability distribution $P(X = x) = k(5 - x)$ for $x = 1, 2, 3, 4$.

(a) Show that $k = \frac{1}{10}$. [2]
(b) Find $E(X)$. [2]
(c) Find $E(X^2)$, and hence show that $\text{Var}(X) = 1$. [3]
(d) In a game, a player pays a stake of $\$c$, a value of $X$ is generated, and the player receives $\$X^2$. Find the value of $c$ for which the game is fair. [2]
(e) The game is played five times, independently. Find the probability that $X = 1$ in exactly two of the five games. [3]
(f) Given that $X = 1$ in at least one of the five games, find the probability that $X = 1$ in exactly two of them. [3]

**26.** GDC In 1846 a Belgian astronomer published the chest measurements, in inches, of 5738 Scottish militiamen, and used them as evidence that human measurements follow a normal distribution. His table is reproduced below.

| Chest (in) | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
|---|---|---|---|---|---|---|---|---|
| Frequency | 3 | 18 | 81 | 185 | 420 | 749 | 1073 | 1079 |
| Chest (in) | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 |
| Frequency | 934 | 658 | 370 | 92 | 50 | 21 | 4 | 1 |

(a) Show that the mean chest measurement is $39.8$ inches, to 3 significant figures. [3]
(b) The standard deviation of the measurements is $2.05$ inches. Assuming $X \sim N(39.8,\, 2.05^2)$, find $P(X < 38.5)$. [2]
(c) The measurements were recorded to the nearest inch, so a soldier recorded as 38 inches or less had an actual chest measurement below $38.5$ inches. Find the proportion of the 5738 soldiers recorded as 38 inches or less, and compare it with your answer to part (b). [3]
(d) Find the proportion of soldiers whose measurement lies within one standard deviation of the mean, and compare it with the value predicted by the normal model. [3]
(e) Comment on how well the normal distribution describes this real data set. [2]

## Challenge Questions

**27.** GDC **Being ahead after $n$ games.** At a fair, a stall offers this game. You pay \$2, roll two fair dice, and are paid \$20 if the total is 12, \$5 if the total is 10 or 11, and nothing otherwise. Let $G$ be your net gain from one game, the payment received minus the \$2 stake. The games are independent, and after $n$ games you are *ahead* if your total net gain is greater than $0$.

(a) Write down the probability distribution of $G$, and show that $E(G) = -\frac{3}{4}$ and $\text{Var}(G) = \frac{625}{48}$. [4]

(b) Find the probability that you are ahead after $1$ game, and the probability that you are ahead after $5$ games. [5]

(c) The table gives the probability of being ahead after $n$ games for larger $n$, calculated exactly by computer.

| $n$ | $10$ | $50$ | $200$ |
|---|---|---|---|
| $P(\text{ahead})$ | $0.198$ | $0.0711$ | $0.00342$ |

Using these values and part (b), describe how the probability of being ahead changes as $n$ increases, and make a conjecture about its behaviour for large $n$. [2]

(d) Let $T_n$ be your total net gain after $n$ games. You may use the facts that $E(T_n) = nE(G)$ and that the standard deviation of $T_n$ is $\sigma\sqrt{n}$, where $\sigma^2 = \text{Var}(G)$. Assuming that $T_n$ is approximately normally distributed, show that

$$P(\text{ahead after } n \text{ games}) \approx P\left(Z > \frac{0.75\sqrt{n}}{\sigma}\right),$$

where $Z \sim N(0, 1)$. Evaluate this for $n = 10$, $50$ and $200$, and compare with the table. [5]

(e) The standard deviation of $T_n$ increases with $n$. Explain why the probability of being ahead nevertheless decreases towards $0$. [2]

(f) Using the normal approximation, find the least number of games after which the stall is at least $99\%$ sure to have made a profit. [3]

(g) Exact calculation shows that the stall needs about $156$ games, not the number found in part (f), and at $n = 200$ the normal approximation gives about half the exact probability in the table. Suggest a reason for these differences. Then investigate how changing the payments, while keeping $E(G) = -\frac{3}{4}$, changes how often players finish ahead, and discuss which design a stall owner might prefer. [3]

**28.** GDC **From the binomial to the bell.** The Reflections claim that a binomial distribution comes to look normal as $n$ grows. This question tests that claim, and finds the adjustment that makes it work.

Throughout, let $X \sim B(50, 0.5)$ and let $Y \sim N(\mu, \sigma^2)$ have the same mean and variance as $X$.

(a) Write down $E(X)$ and $\text{Var}(X)$, and hence state $\mu$ and $\sigma$ for $Y$. [2]

(b) Find $P(20 \leq X \leq 30)$ using the binomial. [2]

(c) Find $P(20 < Y < 30)$ using the normal. Comment on how well it matches part (b). [3]

(d) $X$ takes whole-number values, while $Y$ takes every value in between. The bar of the histogram for $X = 20$ covers the interval from $19.5$ to $20.5$. Using this idea, find $P(19.5 < Y < 30.5)$ and compare it with part (b). [4]

(e) Repeat parts (b), (c) and (d) for $X \sim B(10, 0.5)$ and the range $4 \leq X \leq 6$. [4]

(f) State which of the two normal calculations should be used, and describe how the quality of the approximation changes as $n$ increases. [3]

(g) Many textbooks state that the normal approximation may be used when $np > 5$ and $n(1-p) > 5$. Test this rule: for several binomial distributions with $np$ close to 5, such as $B(50, 0.1)$ and $B(10, 0.5)$, compare an exact binomial probability with its adjusted normal approximation. Discuss whether $np$ alone decides how good the approximation is, and suggest how the rule might be improved. [3]

**29. The most likely value.** Let $X \sim B(n,p)$ with $0 < p < 1$. This question finds the mode of $X$, the value it takes most often.

(a) Show that for $r = 1, 2, \dots, n$,

$$\frac{P(X=r)}{P(X=r-1)} = \frac{(n-r+1)\,p}{r\,(1-p)}.$$

[3]

(b) Hence show that $P(X = r) \geq P(X = r-1)$ if and only if $r \leq (n+1)p$. [3]

(c) Deduce that the mode of $X$ is $\lfloor (n+1)p \rfloor$ when $(n+1)p$ is not an integer, and use this to find the most likely number of successes when $X \sim B(20, 0.3)$. [3]

(d) Explain what happens when $(n+1)p$ is an integer, and illustrate your answer with $X \sim B(9, 0.5)$. [3]

(e) Show that the mode found in part (c) differs from the mean $np$ by less than 1. Then investigate, for several values of $n$ and $p$, how far the median of $X$ can be from $np$, and state a conjecture. [3]

**30.** GDC **Mode, median and mean of one family.** For a constant $a > 0$, a continuous random variable $X$ has probability density function

$$f(x) = \begin{cases} kx^a(1-x), & 0 \leq x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

(a) Show that $k = (a+1)(a+2)$. [2]

(b) Show that the mode of $X$ is $\dfrac{a}{a+1}$ and that $E(X) = \dfrac{a+1}{a+3}$. [4]

(c) Show that, for $0 \leq x \leq 1$, $P(X \leq x) = (a+2)x^{a+1} - (a+1)x^{a+2}$. For $a = 2$ and for $a = \frac{1}{2}$, find the median of $X$ and write the mode, median and mean in increasing order. Make a conjecture about the order when $a > 1$ and when $0 < a < 1$. [5]

(d) Prove that the mean is less than the mode if and only if $a > 1$, and state what happens when $a = 1$. [3]

(e) Show that $P\left(X \leq \dfrac{a}{a+1}\right) = 2\left(\dfrac{a}{a+1}\right)^{a+1}$. You may assume that $g(t) = \left(1 - \frac{1}{t}\right)^t$ is increasing for $t > 1$, and that $g(t) \to e^{-1}$ as $t \to \infty$. Deduce that the median is less than the mode if and only if $a > 1$. Find the limit of $P\left(X \leq \frac{a}{a+1}\right)$ as $a \to \infty$, and interpret it. [5]

(f) A proof that the mean is less than the median when $a > 1$ is harder. Check this numerically for several values of $a$. Then investigate either the wider family

$f(x) = kx^a(1-x)^b$, or how well the rule of thumb mode $-$ mean $\approx 3($median $-$ mean$)$ describes this family. [3]