Specialist Mathematics Units 1 & 2

Chapter 7 — Simulation, sampling and sampling distributions

Aboriginal and Torres Strait Islander content has been excluded as accuracy cannot be guaranteed, and it is better to be respectful than cover the entire curriculum inaccurately.

Contents

Chapter 7 Simulation, sampling and sampling distributions — 7A Mean, variance and standard deviation of a discrete random variable

Theory

In Mathematical Methods the language of probability is used to attach a number $P(A)$ to an *event*. Here we take a further step: to each outcome of a random experiment we attach a **numerical value**, and then study how those values are distributed. The object that carries the value is a random variable, and this section builds the algebra of two numbers that summarise it — its **mean**, which says where the values are centred, and its **variance** (with the associated **standard deviation**), which says how widely they are spread. Everything that follows assumes the basic probability of Methods; what is new is the systematic treatment of *expectation*.

Discrete random variables and their probability distribution.

A **random variable** X assigns a real number to every outcome of a random experiment. It is **discrete** when its possible values can be listed separately, $x_1, x_2, \ldots, x_n$ (for us, a finite list). The behaviour of X is captured by its **probability distribution** (or **probability function**), which records the probability of each value,

$$p(x) = P(X = x).$$

Because these are probabilities of the mutually exclusive outcomes $X = x_1$, $X = x_2, \ldots$, they obey the two conditions carried over from Methods: every probability lies between 0 and 1, and the probabilities of all the values add to 1,

$$\sum p(x) = 1,$$

the sum running over every value in the list. A distribution is usually presented as a table. For instance, if X is the score on a weighted four-face die,

x	1	2	3	4
$p(x)$	0.3	0.4	0.2	0.1

then the four probabilities sum to 1, as they must. The condition $\sum p(x) = 1$ is often used in reverse: if one probability in a table is unknown, it is found by subtracting the others from 1.

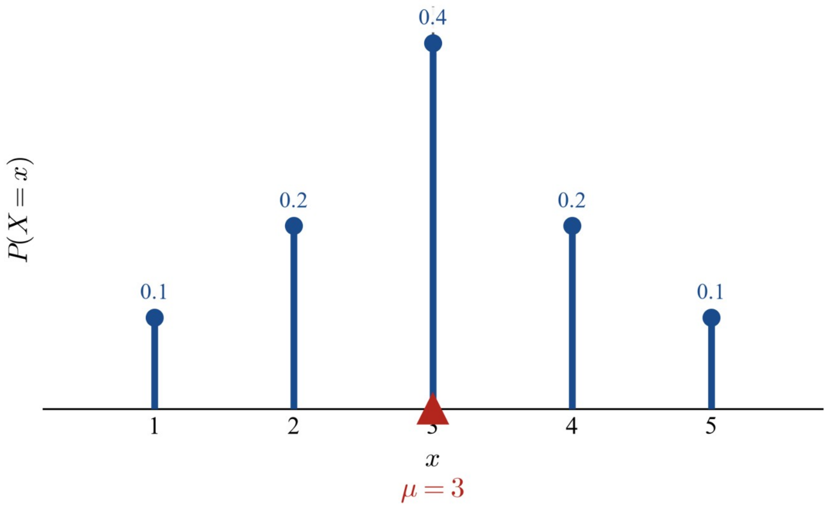
The mean, or expected value.

The **mean** (also called the **expected value**) of a discrete random variable is

$$E(X) = \mu = \sum x\, p(x).$$

It is a **weighted average** of the possible values, each value weighted by how likely it is. If the experiment were repeated a great many times, the average of the observed values would settle down to μ; this is the sense in which μ is "expected". Note that μ need not itself be an attainable value of X.

A useful mental picture is a **balance point**. Imagine the probabilities as masses placed at the values x along a ruler; the mean μ is the point at which the ruler balances — the centre of mass of the distribution.



For the distribution shown, which is symmetric about 3, the balance point is $\mu=3$; the fulcrum sits exactly under the centre.

The expected value of a function of X.

The averaging idea extends to any quantity built from X. If g is a function, then $g(X)$ is itself a random variable, and its mean is obtained by weighting each value $g(x)$ by the probability $p(x)$ of the x that produced it:

$$E(g(X))=\sum g(x)\,p(x).$$

The single most important case takes $g(x)=x^2$, giving

$$E(X^2)=\sum x^2\,p(x).$$

This is needed for the variance below. It is essential to keep $E(X^2)$ and $(E(X))^2$ apart: in general they are **not** equal, and their difference is exactly what measures spread.

The variance and standard deviation.

The **variance** measures how far the values of X typically fall from the mean. It is defined as the expected value of the *squared deviation* $(X-\mu)^2$:

$$Var(X)=\sigma^2=E\left((X-\mu)^2\right)=\sum(x-\mu)^2\,p(x).$$

Squaring keeps every contribution non-negative and removes the sign of the deviation, so $Var(X)\geq 0$ always. Computing it straight from the definition is laborious, so we simplify. Expanding the square and splitting the sum,

$$Var(X)=\sum\left(x^2-2\mu x+\mu^2\right)p(x)=\sum x^2\,p(x)-2\mu\sum x\,p(x)+\mu^2\sum p(x).$$

The three sums are $E(X^2)$, then μ, then 1, so

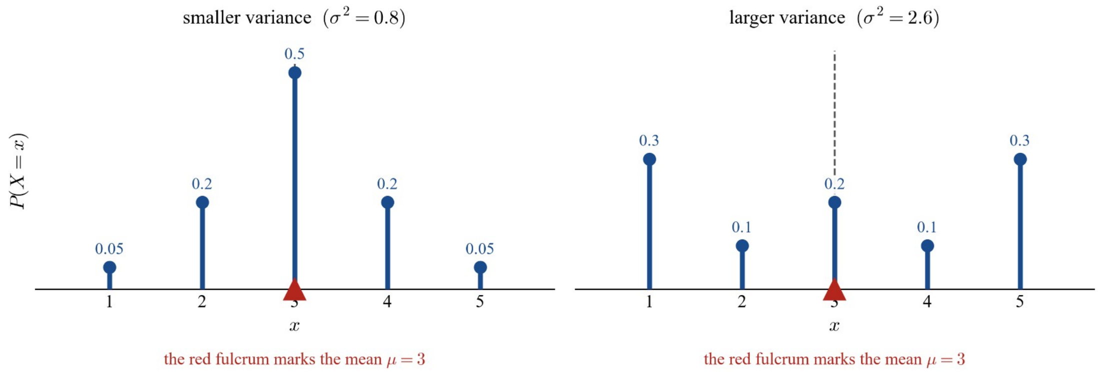
$$Var(X)=E(X^2)-2\mu\cdot\mu+\mu^2\cdot 1=E(X^2)-\mu^2.$$

This **working formula**, $Var(X)=E(X^2)-\mu^2$, is almost always the quicker route: find μ and $E(X^2)$, then subtract.

Because the variance is built from *squared* deviations, its units are the square of the units of X. To return to the original scale we take the positive square root, the **standard deviation**,

$$\sigma=sd(X)=\sqrt{Var(X)}.$$

The standard deviation is the headline measure of spread: a small σ means the probability is concentrated near μ, a large σ means it is dispersed. The two distributions below share the same mean $\mu=3$ but have different variances, $\sigma^2=0.8$ on the left and $\sigma^2=2.6$ on the right.



Linear change of variable: scaling and shifting.

A random variable is frequently replaced by a **linear function** of itself,

$$Y=aX+b,$$

for constants a and b. This happens whenever a measurement is rescaled or shifted — converting units, applying a per-item rate on top of a fixed charge, or recentring a score. We need the mean and variance of Y in terms of those of X.

For the mean, weight each transformed value $ax+b$ by $p(x)$ and split the sum:

$$E(aX+b)=\sum(ax+b)p(x)=a\sum x\,p(x)+b\sum p(x)=a\mu+b,$$

using $\sum x\,p(x)=\mu$ and $\sum p(x)=1$. Hence

$$E(aX+b)=aE(X)+b.$$

The mean simply follows the transformation.

For the variance, write $Y=aX+b$ with $E(Y)=a\mu+b$, and use the definition:

$$Var(aX+b)=E\left((Y-E(Y))^2\right)=E\left((aX+b-a\mu-b)^2\right)=E\left(a^2(X-\mu)^2\right)=a^2E\left((X-\mu)^2\right).$$

The constant b has cancelled, and the factor a^2 comes out of the expectation, so

$$Var(aX+b)=a^2Var(X)\text{ and }sd(aX+b)=|a|\sigma.$$

The modulus appears because a standard deviation is never negative: a negative scale factor reflects the distribution but cannot give it a negative spread.

These two rules have a clear meaning. A **shift** by b moves every value — and so the centre — by the same amount, but leaves the distances between values untouched; the spread, and therefore the variance, does not depend on b at all. A **scaling** by a stretches every deviation from the mean by the factor a, so the standard deviation is multiplied by $|a|$ and the variance by a^2. Two special cases are worth recording: multiplying gives $E(aX)=aE(X)$ and $Var(aX)=a^2Var(X)$, while merely adding a constant gives $E(X+b)=E(X)+b$ and $Var(X+b)=Var(X)$.

Worked examples

Example 1 — scaffolded

A community garden plot is checked each morning and X is the number of ripe tomatoes picked, with the probability distribution below. Find $E(X)$, $E(X^2)$, $Var(X)$ and the standard deviation σ.

x	1	2	3	4
$p(x)$	1/8	3/8	3/8	1/8

Working	Comment
$1/8+3/8+3/8+1/8=1.$	Check the probabilities form a valid distribution.
$E(X)=1\cdot 1/8+2\cdot 3/8+3\cdot 3/8+4\cdot 1/8=1+6+9+$.	The mean is the weighted average $\sum x\,p(x)$.
$E(X^2)=1^2\cdot 1/8+2^2\cdot 3/8+3^2\cdot 3/8+4^2\cdot 1/8=1+12$.	Weight each *squared* value by $p(x)$.
$Var(X)=E(X^2)-\mu^2=7-(5/2)^2=7-25/4=3/4.$	Use $Var(X)=E(X^2)-\mu^2$.
$\sigma=\sqrt{3/4}=\frac{\sqrt{3}}{2}\approx 0.87.$	The standard deviation is the positive square root.

Example 2 — scaffolded

A citizen-science app records X, the number of native bees visiting a grevillea flower in one minute. One probability is missing from the table below. The app awards $P=4X+3$ points for the minute. Find the missing probability, then $E(X)$ and $Var(X)$, and hence $E(P)$, $Var(P)$ and $sd(P)$.

x	0	1	2	3
$p(x)$	0.2	?	0.3	0.1

Working	Comment
$P(X=1)=1-(0.2+0.3+0.1)=0.4.$	The probabilities must sum to 1.
$E(X)=0(0.2)+1(0.4)+2(0.3)+3(0.1)=0.4+0.6+($.	Weighted average of the values.
$E(X^2)=0^2(0.2)+1^2(0.4)+2^2(0.3)+3^2(0.1)=0.4+1.$.	Needed for the variance.
$Var(X)=2.5-1.3^2=2.5-1.69=0.81$, so $\sigma=0.9$.	$E(X^2)-\mu^2$, then take the root.
$E(P)=E(4X+3)=4\mu+3=4(1.3)+3=8.2.$	Apply $E(aX+b)=a\mu+b$ with $a=4, b=3$.
$Var(P)=4^2 Var(X)=16(0.81)=12.96;$ $sd(P)=\lvert 4\rvert\sigma=4(0.9)=3.6.$	The $+3$ does not affect the spread.

Example 3 — unscaffolded

A short warm-up quiz has three multiple-choice questions. Let X be the number a student answers incorrectly, with the distribution below. The mark awarded is $S=20-4X$. Find the mean and standard deviation of X, and of S.

x	0	1	2	3
$p(x)$	0.2	0.2	0.3	0.3

The mean and second moment are

$$E(X)=0(0.2)+1(0.2)+2(0.3)+3(0.3)=0.2+0.6+0.9=1.7\,,$$

$$E(X^2)=0(0.2)+1(0.2)+4(0.3)+9(0.3)=0.2+1.2+2.7=4.1\,,$$

so $Var(X)=4.1-1.7^2=4.1-2.89=1.21$ and $\sigma=\sqrt{1.21}=1.1$. Now $S=20-4X$ is the linear function with $a=-4$ and $b=20$, so

$$E(S) = -4(1.7) + 20 = 13.2, Var(S) = (-4)^2(1.21) = 19.36, sd(S) = |-4|(1.1) = 4.4.$$

The negative scale factor reflects the scores (a higher X gives a lower S) but, through $|a|$, keeps the spread positive.

$\therefore$ X has mean 1.7 and standard deviation 1.1, while the mark S has mean 13.2 and standard deviation 4.4.

Example 4 — unscaffolded

On an online marketplace a seller's items receive a whole-number rating $X \in \{1, 2, 3, 4, 5\}$. Two probabilities are unknown, and the mean rating is known to be 3.

x	1	2	3	4	5
$p(x)$	0.1	p	0.3	q	0.1

The platform pays the seller $R = 8X - 5$ dollars per rating. Find p and q, the variance of X, and the standard deviation of R.

The probabilities sum to 1, giving $0.1 + p + 0.3 + q + 0.1 = 1$, so

$$p + q = 0.5.$$

The mean is 3, giving $1(0.1) + 2p + 3(0.3) + 4q + 5(0.1) = 3$, that is $0.1 + 2p + 0.9 + 4q + 0.5 = 3$, so

$$2p + 4q = 1.5 \Rightarrow p + 2q = 0.75.$$

Subtracting the first equation from this gives $q = 0.25$, and then $p = 0.25$. With the full distribution,

$$E(X^2) = 1(0.1) + 4(0.25) + 9(0.3) + 16(0.25) + 25(0.1) = 0.1 + 1 + 2.7 + 4 + 2.5 = 10.3,$$

so $Var(X) = 10.3 - 3^2 = 1.3$. For $R = 8X - 5$ (here $a = 8$),

$$Var(R) = 8^2(1.3) = 83.2, sd(R) = |8|\sqrt{1.3} = 8\sqrt{1.3} \approx 9.12.$$

$\therefore$ $p = q = 0.25$, $Var(X) = 1.3$, and the payment R has standard deviation $8\sqrt{1.3} \approx \$9.12$.

Practice questions

Scaffolded

Q1. The number of dogs X arriving at an off-leash park gate in a one-minute window has the distribution below.

x	0	1	2	3
$p(x)$	0.5	0.2	0.2	0.1

(a) Find $E(X)$.
(b) Find $E(X^2)$.
(c) Find $Var(X)$.
(d) Find the standard deviation σ, correct to two decimal places.

Q2. A scratch-card gives a voucher worth X dollars, where $X \in \{2, 4, 6, 8\}$. One probability is missing.

x	2	4	6	8
$p(x)$	0.1	?	0.3	0.2

(a) Find the missing probability $P(X = 4)$.
(b) Find $E(X)$.
(c) Find $E(X^2)$ and hence $Var(X)$.

(d) Find the standard deviation σ, correct to two decimal places.

Q3. A market stall serves X customers in a five-minute period, with the distribution below. Its takings for the period are $T=6X+2$ dollars.

x	2	3	4	5
$p(x)$	0.4	0.3	0.2	0.1

(a) Find $E(X)$.
(b) Find $E(X^2)$ and hence $Var(X)$.
(c) Find $E(T)$.
(d) Find $Var(T)$ and $sd(T)$.

Q4. In a phone game a player begins with 50 points and loses 10 points for each of X penalties, so the points remaining are $W=50-10X$. The distribution of X is below.

x	0	1	2	3
$p(x)$	0.2	0.3	0.3	0.2

(a) Find $E(X)$ and $Var(X)$.
(b) Find $E(W)$.
(c) Find $Var(W)$ and $sd(W)$.
(d) Explain why the "$+50$" in W has no effect on $sd(W)$, whereas the "-10" does.

Q5. The mass X kilograms of recycling a household puts out each week has mean 12 and standard deviation 4. For each variable below, state its mean, variance and standard deviation. (a) $A=X+9$. (b) $B=3X$. (c) $C=3X+9$. (d) $D=-2X+5$.

Q6. A raw test score X has mean 5 and standard deviation 2. It is rescaled to $Y=aX+b$, with $a>0$, so that Y has mean 23 and standard deviation 8. (a) Use the standard deviation to find a. (b) Use the mean to find b. (c) Write down the rule for Y.

Unscaffolded

Q7. A raffle prize X (in dollars) has the distribution below. Find $E(X)$, $Var(X)$ and σ (to two decimal places).

x	5	10	15	20
$p(x)$	0.5	0.3	0.1	0.1

Q8. In a card game the net gain X dollars per round has the distribution below. Find $E(X)$, $Var(X)$ and σ (to two decimal places).

x	-2	-1	0	1	2
$p(x)$	0.1	0.2	0.4	0.2	0.1

Q9. The number of daily sign-ups X to an app has the distribution below, and a promotion pays a reward $Y=5X+8$ dollars. Find $E(Y)$, $Var(Y)$ and $sd(Y)$ (to two decimal places).

x	0	1	2	3	4
$p(x)$	0.3	0.25	0.2	0.15	0.1

Q10. The number of text messages X a commuter sends on a train trip has the distribution below, with one probability missing.

x	0	1	2	3	4
$p(x)$	0.15	0.25	?	0.2	0.1

Find the missing probability, then $E(X)$ and $Var(X)$.

Q11. The number of cars X carried on a small ferry run has the distribution below, with two unknown probabilities, and the mean is 2.1.

x	0	1	2	3	4
$p(x)$	0.15	p	0.25	q	0.1

Find p and q. Each run charges a toll $T=10X+5$ dollars; find $E(T)$ and $sd(T)$ (to two decimal places).

Q12. A random variable X has mean 8 and variance 5. For each variable below, find its mean, variance and standard deviation (leave surds exact). (a) $2X+1$. (b) $5-3X$.

Q13. A bakery's number of unsold pies X at closing has the distribution below. The day's profit is $P=40-2X$ dollars. Find $E(P)$ and $sd(P)$ (to two decimal places).

x	0	1	2	3	4
$p(x)$	0.4	0.3	0.15	0.1	0.05

Q14. A stall game pays a prize X dollars with the distribution below, and charges $\$c$ to play, so a player's profit is $X-c$.

x	0	2	5	10
$p(x)$	0.5	0.3	0.15	0.05

(a) Find $E(X)$.
(b) Find the value of c that makes the game **fair** (expected profit 0).
(c) Find the variance of a player's profit, explaining why it equals $Var(X)$.

Q15. A sensor reading X has mean 10 and standard deviation 6. Find constants $a>0$ and b so that $W=aX+b$ has mean 0 and standard deviation 18, and write down the rule for W.

Q16. A five-sector spinner is labelled $1,2,3,4,5$, and the probability of a sector is proportional to its label: $p(x)=cx$ for $x=1,2,3,4,5$. (a) Find the constant c. (b) Find $E(X)$. (c) Find $Var(X)$.

Q17. A random variable X has mean 4 and variance 9. A linear function $Y=aX+b$ with $a<0$ has mean 30 and variance 36. Find a and b.

Q18. A gelato cart records X, the number of scoops per order, with the distribution below (one probability missing). Each order is priced $C=4X+2$ dollars.

x	1	2	3	4
$p(x)$	0.1	0.4	?	0.1

(a) Find the missing probability.
(b) Find $E(X)$, $Var(X)$ and $sd(X)$ (to two decimal places).
(c) Find $E(C)$ and $sd(C)$.
(d) A rival cart prices orders $D=2X+50$ dollars. Which of C and D has the greater spread of prices? Justify your answer.

Answers

Q1. (a) 0.9; (b) 1.9; (c) 1.09; (d) $\sigma=\sqrt{1.09}\approx1.04$. **Q2.** (a) 0.4; (b) 5.2; (c) $E(X^2)=30.4$, $Var(X)=3.36$; (d) $\sigma\approx1.83$. **Q3.** (a) 3; (b) $E(X^2)=10$, $Var(X)=1$; (c) $E(T)=20$; (d) $Var(T)=36$, $sd(T)=6$. **Q4.** (a) $E(X)=1.5$, $Var(X)=1.05$; (b) $E(W)=35$; (c) $Var(W)=105$, $sd(W)=\sqrt{105}\approx10.25$; (d) see solution. **Q5.** (a) mean 21, variance 16, sd 4; (b) mean 36, variance 144, sd 12; (c) mean 45, variance 144, sd 12; (d) mean -19, variance 64, sd 8. **Q6.** (a) $a=4$; (b) $b=3$; (c) $Y=4X+3$. **Q7.** $E(X)=9$, $Var(X)=24$, $\sigma=2\sqrt{6}\approx4.90$. **Q8.** $E(X)=0$, $Var(X)=1.2$, $\sigma=\sqrt{1.2}\approx1.10$. **Q9.** $E(Y)=15.5$, $Var(Y)=43.75$, $sd(Y)=5\sqrt{7}/2\approx6.61$. **Q10.** Missing probability 0.3; $E(X)=1.85$, $Var(X)=1.4275$. **Q11.** $p=0.15$, $q=0.35$; $E(T)=26$, $sd(T)=\sqrt{149}\approx12.21$. **Q12.** (a) mean 17, variance 20, sd $2\sqrt{5}\approx4.47$; (b) mean -19, variance 45, sd $3\sqrt{5}\approx6.71$. **Q13.** $E(P)=37.8$, $sd(P)=\sqrt{139}/5\approx2.36$. **Q14.** (a) $E(X)=1.85$; (b) $c=1.85$; (c) $Var(\text{profit})=Var(X)=6.5275$. **Q15.** $a=3$, $b=-30$; $W=3X-30$. **Q16.** (a) $c=1/15$; (b) $E(X)=11/3\approx3.67$; (c) $Var(X)=14/9\approx1.56$. **Q17.** $a=-2$, $b=38$. **Q18.** (a) 0.4; (b) $E(X)=2.5$, $Var(X)=0.65$, $sd(X)\approx0.81$; (c) $E(C)=12$, $sd(C)=2\sqrt{65}/5\approx3.22$; (d) C has the greater spread.

Fully-worked solutions

Q1. The probabilities sum to 1, so the table is valid. (a) $E(X)=0(0.5)+1(0.2)+2(0.2)+3(0.1)=0.2+0.4+0.3=0.9$. (b) $E(X^2)=0(0.5)+1(0.2)+4(0.2)+9(0.1)=0.2+0.8+0.9=1.9$. (c) $Var(X)=E(X^2)-\mu^2=1.9-0.9^2=1.9-0.81=1.09$. (d) $\sigma=\sqrt{1.09}\approx1.04$.

Q2. (a) $P(X=4)=1-(0.1+0.3+0.2)=0.4$. (b) $E(X)=2(0.1)+4(0.4)+6(0.3)+8(0.2)=0.2+1.6+1.8+1.6=5.2$. (c) $E(X^2)=4(0.1)+16(0.4)+36(0.3)+64(0.2)=0.4+6.4+10.8+12.8=30.4$, so $Var(X)=30.4-5.2^2=30.4-27.04=3.36$. (d) $\sigma=\sqrt{3.36}=2\sqrt{21}/5\approx1.83$.

Q3. (a) $E(X)=2(0.4)+3(0.3)+4(0.2)+5(0.1)=0.8+0.9+0.8+0.5=3$. (b) $E(X^2)=4(0.4)+9(0.3)+16(0.2)+25(0.1)=1.6+2.7+3.2+2.5=10$, so $Var(X)=10-3^2=1$. (c) With $T=6X+2$ ($a=6, b=2$), $E(T)=6(3)+2=20$. (d) $Var(T)=6^2 Var(X)=36(1)=36$ and $sd(T)=|6|\sqrt{1}=6$; the constant $+2$ does not affect the spread.

Q4. (a) $E(X)=0(0.2)+1(0.3)+2(0.3)+3(0.2)=0.3+0.6+0.6=1.5$; $E(X^2)=0(0.2)+1(0.3)+4(0.3)+9(0.2)=0.3+1.2+1.8=3.3$, so $Var(X)=3.3-1.5^2=3.3-2.25=1.05$. (b) With $W=50-10X$ ($a=-10, b=50$), $E(W)=-10(1.5)+50=35$. (c) $Var(W)=(-10)^2(1.05)=105$ and $sd(W)=|-10|\sqrt{1.05}=\sqrt{105}\approx10.25$. (d) Adding 50 shifts every possible score by the same amount, so the distances of the scores from their mean — and hence the spread — are unchanged; the variance does not depend on b. The factor -10 multiplies every deviation by 10 (its size), so the standard deviation is multiplied by $|-10|=10$.

Q5. In each case use $E(aX+b)=a\mu+b$, $Var(aX+b)=a^2\sigma^2$ and $sd=|a|\sigma$, with $\mu=12$, $\sigma=4$, $\sigma^2=16$. (a) $A=X+9$: mean $12+9=21$, variance $1^2(16)=16$, sd 4. (b) $B=3X$: mean $3(12)=36$, variance $3^2(16)=144$, sd $3(4)=12$. (c) $C=3X+9$: mean $3(12)+9=45$, variance 144, sd 12 (the $+9$ changes only the mean). (d) $D=-2X+5$: mean $-2(12)+5=-19$, variance $(-2)^2(16)=64$, sd $|-2|(4)=8$.

Q6. (a) The standard deviation depends only on the scaling: $sd(Y)=|a|\sigma$ gives $|a|(2)=8$, so $|a|=4$, and as $a>0$, $a=4$. (b) The mean gives $a(5)+b=23$, so $4(5)+b=23$, hence $b=23-20=3$. (c) $Y=4X+3$.

Q7. $E(X)=5(0.5)+10(0.3)+15(0.1)+20(0.1)=2.5+3+1.5+2=9$. $E(X^2)=25(0.5)+100(0.3)+225(0.1)+400(0.1)=12.5+30+22.5+40=105$, so $Var(X)=105-9^2=24$ and $\sigma=\sqrt{24}=2\sqrt{6}\approx4.90$.

Q8. By the symmetry of the table, $E(X)=-2(0.1)-1(0.2)+0+1(0.2)+2(0.1)=0$.
$E(X^2)=4(0.1)+1(0.2)+0+1(0.2)+4(0.1)=0.4+0.2+0.2+0.4=1.2$, so $Var(X)=1.2-0^2=1.2$ and $\sigma=\sqrt{1.2}\approx1.10$.

Q9. First $E(X)=0(0.3)+1(0.25)+2(0.2)+3(0.15)+4(0.1)=0.25+0.4+0.45+0.4=1.5$;
$E(X^2)=0+1(0.25)+4(0.2)+9(0.15)+16(0.1)=0.25+0.8+1.35+1.6=4$, so $Var(X)=4-1.5^2=1.75$. With $Y=5X+8$ ($a=5,b=8$), $E(Y)=5(1.5)+8=15.5$, $Var(Y)=5^2(1.75)=43.75$ and $sd(Y)=|5|\sqrt{1.75}=5\sqrt{7}/2\approx6.61$.

Q10. The missing probability is $1-(0.15+0.25+0.2+0.1)=0.3$. Then
$E(X)=0(0.15)+1(0.25)+2(0.3)+3(0.2)+4(0.1)=0.25+0.6+0.6+0.4=1.85$;
$E(X^2)=0+1(0.25)+4(0.3)+9(0.2)+16(0.1)=0.25+1.2+1.8+1.6=4.85$, so
$Var(X)=4.85-1.85^2=4.85-3.4225=1.4275$.

Q11. Summing to 1: $0.15+p+0.25+q+0.1=1$, so $p+q=0.5$. The mean is 2.1:
$0+p+2(0.25)+3q+4(0.1)=2.1$, that is $p+3q+0.9=2.1$, so $p+3q=1.2$. Subtracting, $2q=0.7$, giving $q=0.35$ and $p=0.15$. Then $E(X^2)=0+1(0.15)+4(0.25)+9(0.35)+16(0.1)=0.15+1+3.15+1.6=5.9$, so $Var(X)=5.9-2.1^2=5.9-4.41=1.49$. With $T=10X+5$, $E(T)=10(2.1)+5=26$ and $sd(T)=|10|\sqrt{1.49}=\sqrt{149}\approx12.21$.

Q12. Use $\mu=8$, $Var(X)=5$, $\sigma=\sqrt{5}$. (a) $2X+1$: mean $2(8)+1=17$, variance $2^2(5)=20$, sd $|2|\sqrt{5}=2\sqrt{5}\approx4.47$. (b) $5-3X$ (that is $-3X+5$): mean $-3(8)+5=-19$, variance $(-3)^2(5)=45$, sd $|-3|\sqrt{5}=3\sqrt{5}\approx6.71$.

Q13. $E(X)=0(0.4)+1(0.3)+2(0.15)+3(0.1)+4(0.05)=0.3+0.3+0.3+0.2=1.1$;
$E(X^2)=0+1(0.3)+4(0.15)+9(0.1)+16(0.05)=0.3+0.6+0.9+0.8=2.6$, so $Var(X)=2.6-1.1^2=1.39$.
With $P=40-2X$ ($a=-2,b=40$), $E(P)=-2(1.1)+40=37.8$ and $sd(P)=|-2|\sqrt{1.39}=\sqrt{139}/5\approx2.36$.

Q14. (a) $E(X)=0(0.5)+2(0.3)+5(0.15)+10(0.05)=0.6+0.75+0.5=1.85$. (b) The profit is $X-c$, with expected value $E(X-c)=E(X)-c=1.85-c$. The game is fair when this is 0, so $c=1.85$; a fair charge equals the expected prize. (c) The profit $X-c$ is a shift of X (scale $a=1$, shift $b=-c$), so $Var(X-c)=1^2Var(X)=Var(X)$.
Now $E(X^2)=0+4(0.3)+25(0.15)+100(0.05)=1.2+3.75+5=9.95$, so
$Var(X)=9.95-1.85^2=9.95-3.4225=6.5275$. Subtracting the fixed charge shifts every profit equally and so leaves the spread unchanged.

Q15. The standard deviation fixes the scale: $sd(W)=|a|(6)=18$, so $|a|=3$, and as $a>0$, $a=3$. The mean fixes the shift: $E(W)=a(10)+b=0$, so $3(10)+b=0$, giving $b=-30$. Hence $W=3X-30$.

Q16. (a) The probabilities must sum to 1: $c(1+2+3+4+5)=15c=1$, so $c=1/15$. (b) Weighting each value by $p(x)=x/15$, $E(X)=1/15(1(1)+2(2)+3(3)+4(4)+5(5))=1+4+9+16+25/15=55/15=11/3$. (c)
$E(X^2)=1/15(1^2(1)+2^2(2)+3^2(3)+4^2(4)+5^2(5))=1+8+27+64+125/15=225/15=15$, so
$Var(X)=15-(11/3)^2=15-121/9=135-121/9=14/9\approx1.56$.

Q17. The variance fixes the size of the scale: $Var(Y)=a^2(9)=36$, so $a^2=4$ and $a=\pm2$; since $a<0$, $a=-2$. The mean fixes the shift: $E(Y)=a(4)+b=30$, so $-2(4)+b=30$, giving $b=30+8=38$. Thus $Y=-2X+38$.

Q18. (a) $P(X=3)=1-(0.1+0.4+0.1)=0.4$. (b)
$E(X)=1(0.1)+2(0.4)+3(0.4)+4(0.1)=0.1+0.8+1.2+0.4=2.5$;
$E(X^2)=1(0.1)+4(0.4)+9(0.4)+16(0.1)=0.1+1.6+3.6+1.6=6.9$, so
$Var(X)=6.9-2.5^2=6.9-6.25=0.65$ and $sd(X)=\sqrt{0.65}\approx0.81$. (c) With $C=4X+2$, $E(C)=4(2.5)+2=12$ and $sd(C)=|4|\sqrt{0.65}=4\sqrt{0.65}=2\sqrt{65}/5\approx3.22$. (d) The spread of a linear function depends only on the size of the

scale factor. For $C=4X+2$ it is $sd(C)=4\sigma$, while for $D=2X+50$ it is $sd(D)=2\sigma$. Since $4>2$, C has the greater spread of prices (here $sd(C)\approx 3.22$ against $sd(D)\approx 1.61$), even though D has the larger mean.

Theory

A **random variable** X attaches a number to each outcome of a random experiment. Two numbers summarise how it behaves: the **mean** (or **expected value**) $\mu = E(X)$, which locates the centre of the distribution, and the **variance** $\sigma^2 = Var(X)$, which measures its spread. The **standard deviation** is $\sigma = sd(X) = \sqrt{Var(X)}$. For a discrete X these are computed from its probability distribution by

$$E(X) = \sum_i x_i p_i, Var(X) = \sum_i (x_i - \mu)^2 p_i = E(X^2) - \mu^2,$$

where $p_i = P(X = x_i)$; the second form for the variance is usually the quicker to apply. The previous section built random samples by repeating an experiment; this section asks what happens to the mean and variance when random variables are **scaled, shifted, added and subtracted** — the arithmetic that governs sample totals, differences of two measurements, and every weighted blend of independent quantities.

Scaling and shifting one variable.

For real constants a and b, the transformed variable $aX + b$ has

$$E(aX + b) = a\mu + b, Var(aX + b) = a^2\sigma^2, sd(aX + b) = |a|\sigma.$$

The mean follows the transformation completely, but the variance responds only to the *scaling*: replacing each value x_i by $a x_i + b$ multiplies every deviation from the mean by a, so the squared deviations — and hence the variance — are multiplied by a^2, while the shift b moves every value equally and leaves the spread untouched. The modulus appears in the standard deviation because a spread is never negative: a negative a reflects the distribution but does not give it a negative width. In particular, doubling a single observation gives

$$Var(2X) = 2^2\sigma^2 = 4\sigma^2,$$

a fact to be contrasted sharply below with adding two separate observations.

Linearity of expectation.

For **any** two random variables X and Y and any real constants a and b,

$$E(aX + bY) = aE(X) + bE(Y).$$

This follows directly from the definition of expectation applied to the joint distribution: summing $ax + by$ against the probabilities $P(X = x, Y = y)$ splits into $a\sum x P(X = x)$ and $b\sum y P(Y = y)$, which are $aE(X)$ and $bE(Y)$. The decisive feature is that this identity holds **whether or not X and Y are independent** — expected values always combine linearly. In particular

$$E(X + Y) = E(X) + E(Y), E(X - Y) = E(X) - E(Y),$$

and an added constant c merely shifts the mean, $E(aX + bY + c) = aE(X) + bE(Y) + c$.

Variance of a linear combination of independent variables.

Variances behave quite differently, and here **independence** is essential. Write $\mu_X = E(X)$ and $\mu_Y = E(Y)$. Expanding the definition of variance,

$$Var(aX + bY) = E\left[(aX + bY - a\mu_X - b\mu_Y)^2\right] = E\left[(a(X - \mu_X) + b(Y - \mu_Y))^2\right].$$

Multiplying out the square gives three terms,

$$Var(aX+bY)=a^2E\left[(X-\mu_X)^2\right]+2abE\left[(X-\mu_X)(Y-\mu_Y)\right]+b^2E\left[(Y-\mu_Y)^2\right].$$

The first and third expectations are $Var(X)$ and $Var(Y)$. The middle **cross term** carries the factor $E\left[(X-\mu_X)(Y-\mu_Y)\right]$. When X and Y are **independent**, $P(X=x,Y=y)=P(X=x)P(Y=y)$, so $E(XY)=E(X)E(Y)$ and therefore

$$E\left[(X-\mu_X)(Y-\mu_Y)\right]=E(XY)-\mu_X\mu_Y=\mu_X\mu_Y-\mu_X\mu_Y=0.$$

The cross term vanishes, leaving

$$Var(aX+bY)=a^2Var(X)+b^2Var(Y)(X,Y\text{ independent}).$$

Each coefficient enters **squared**, so every term is non-negative. Taking $a=b=1$ and then $a=1,b=-1$ gives the two most important cases:

$$Var(X+Y)=Var(X)+Var(Y),Var(X-Y)=Var(X)+Var(Y).$$

Read the second identity carefully. For a **difference** the coefficient of Y is -1, and $(-1)^2=1$, so the two variances are still **added**. Independent uncertainty *accumulates*: combining two independent variables can only widen the spread, whether they are added or subtracted, and $Var(X-Y)=Var(X+Y)$. Writing $Var(X)-Var(Y)$ for a difference is a common and costly error. Because a constant has no spread, an added constant leaves the variance unchanged, $Var(aX+bY+c)=a^2Var(X)+b^2Var(Y)$, and the standard deviation is the square root,

$$sd(aX+bY)=\sqrt{a^2Var(X)+b^2Var(Y)},$$

so it is the **variances**, never the standard deviations, that combine.

The general linear combination.

The rules extend to any number of variables. Let $X_1,X_2,\ldots,X_n$ have means μ_i and variances σ_i^2, and let $a_1,a_2,\ldots,a_n$ be constants. The expected value of the linear combination is

$$E\left(\sum_{i=1}^{n}a_iX_i\right)=\sum_{i=1}^{n}a_i\mu_i,$$

which holds with no restriction, while if the X_i are **independent** the variance is

$$Var\left(\sum_{i=1}^{n}a_iX_i\right)=\sum_{i=1}^{n}a_i^2\sigma_i^2.$$

Again each coefficient is squared, so a negative coefficient contributes exactly as its positive counterpart would.

Sums of independent identically distributed copies.

A very common special case is a **sample total**: n **independent and identically distributed** (iid) copies $X_1,X_2,\ldots,X_n$, each with the same mean μ and variance σ^2. Setting every $a_i=1$, $\mu_i=\mu$ and $\sigma_i^2=\sigma^2$ in the general rules gives

$$E\left(X_1+X_2+\ldots+X_n\right)=n\mu,Var\left(X_1+X_2+\ldots+X_n\right)=n\sigma^2,$$

so the total has standard deviation $\sqrt{n}\sigma$. The mean grows in direct proportion to n, but the standard deviation grows only like $\sqrt{n}$ — a fact at the heart of the sampling distributions studied next.

The crucial distinction: nX versus $X_1+X_2+\ldots+X_n$.

These two expressions are easily confused yet describe genuinely different quantities.

- $n\,X$ is n times a **single** observation. As a scaling of one variable, $Var(n\,X) = n^2\sigma^2$ and $sd(n\,X) = n\,\sigma$.
- $X_1 + X_2 + \ldots + X_n$ is a sum of n **separate, independent** observations, with $Var = n\,\sigma^2$ and $sd = \sqrt{n}\,\sigma$.

Quantity	Mean	Variance	Standard deviation
$n\,X$	$n\,\mu$	$n^2\sigma^2$	$n\,\sigma$
$X_1 + X_2 + \ldots + X_n$	$n\,\mu$	$n\,\sigma^2$	$\sqrt{n}\,\sigma$

Both have the same mean $n\,\mu$, but their variances differ by a factor of n. Scaling one observation by n magnifies its single deviation from the mean by n, so the variance grows by n^2; adding n independent observations lets high and low values partly offset one another, so the variance grows only by n. Consider a spinner scoring X points with values 3, 5, 7 and probabilities $1/4$, $1/2$, $1/4$; a short calculation gives $\mu = 5$ and $\sigma^2 = 2$. Then two independent spins have $Var(X_1 + X_2) = 2\,\sigma^2 = 4$, whereas doubling one spin has $Var(2\,X) = 4\,\sigma^2 = 8$ — twice as large, even though both average 10. Both quantities *can* land on 10: the sum when a low spin meets a high one or when both show 5, and $2\,X$ whenever $X = 5$. The difference lies in how they miss. Because $2\,X$ doubles a single face it has no value between 6 and 10, nor between 10 and 14, so a miss is always by the full 4 points; the sum, by contrast, is off-centre by only 2 points whenever exactly one spin shows 5, and reaches ± 4 only when both spins land on the same extreme. Mistaking one expression for the other is a favourite examination trap.

Building the distribution of a sum.

Means and variances are only summaries: for independent discrete variables the whole **probability distribution** of a sum can be written down as well. Independence makes each *ordered* pair of values carry the product of its two probabilities, so

$$P(X_1 + X_2 = s) = \sum_{x_1 + x_2 = s} P(X_1 = x_1)\,P(X_2 = x_2)$$

that is, list every ordered pair (x_1, x_2), multiply the two probabilities, and add the products of all pairs sharing the same total s. If X has k possible values there are k^2 ordered pairs to sort, and a total reachable in several ways collects several products; the same method extends to three or more independent variables, over ordered triples and beyond.

Scaling needs no such work. The variable $2\,X$ takes the value $2\,x$ exactly when X takes the value x, so

$$P(2\,X = 2\,x) = P(X = x)$$

and its distribution is simply the distribution of X with every **value** doubled and every **probability** unchanged. That is the sharpest way to see the contrast above: $2\,X$ has no more possible values than X, merely stretched away from the centre, whereas $X_1 + X_2$ has *more* values — and the new ones sit in the gaps between them, where the two spins have partly cancelled. Example 2 builds both distributions in full for the spinner just described.

Applications.

- **Sample totals.** The total mass, length or score built from n independent, identically distributed items has mean $n\,\mu$ and variance $n\,\sigma^2$.
- **Differences and comparisons.** To compare two independent processes — which is larger, or by how much one exceeds the other — work with $X - Y$: its mean is $\mu_X - \mu_Y$, but its variance is $\sigma_X^2 + \sigma_Y^2$.
- **Weighted blends.** A recipe, portfolio or score $a_1 X_1 + a_2 X_2 + \ldots$ mixing independent ingredients has mean $a_1\mu_1 + a_2\mu_2 + \ldots$ and variance $a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ldots$.

In every case the means combine by the plain linear rule, while — for independent variables — the *variances* combine with squared coefficients, and it is the variance, never the standard deviation, that adds.

Worked examples

Example 1 — scaffolded

Two independent random variables X and Y have the means and variances shown.

Variable	Mean	Variance
X	11	6
Y	7	3

Find the mean and variance of (a) $X+Y$, (b) $X-Y$, and (c) $3X-2Y$, and state $sd(3X-2Y)$.

Working	Comment
$E(X+Y)=E(X)+E(Y)=11+7=18$.	Means always add, independent or not.
$Var(X+Y)=Var(X)+Var(Y)=6+3=9$.	For independent variables the variances add.
$E(X-Y)=E(X)-E(Y)=11-7=4$.	The mean of a difference subtracts.
$Var(X-Y)=Var(X)+Var(Y)=6+3=9$.	Still a sum: the coefficient -1 squares to $+1$.
$E(3X-2Y)=3E(X)-2E(Y)=3(11)-2(7)=33-$.	Use $E(aX+bY)=aE(X)+bE(Y)$.
$Var(3X-2Y)=3^2Var(X)+(-2)^2Var(Y)=9(6)+4$.	Square each coefficient before multiplying by the variance.
$sd(3X-2Y)=\sqrt{66}\approx 8.12$.	Take the square root of the variance.

Notice that $X+Y$ and $X-Y$ share the **same** variance 9; only their means differ.

Example 2 — scaffolded

An arcade token, when spun, shows a score X points with the distribution below.

x	3	5	7
$P(X=x)$	0.25	0.5	0.25

(a) Find $E(X)$ and $Var(X)$.

(b) Two tokens are spun independently; find the mean and standard deviation of the total score $S=X_1+X_2$.

(c) Find the mean and standard deviation of $2X$ (twice the score of a single token), and explain why the spread differs from part (b).

(d) Write down the probability distribution of S and the probability distribution of $2X$, and compare them.

Working	Comment
$E(X)=3(0.25)+5(0.5)+7(0.25)=0.75+2.5+1.75=$.	Multiply each value by its probability and add.
$E(X^2)=9(0.25)+25(0.5)+49(0.25)=2.25+12.5+1$.	The variance also needs $E(X^2)$.
$Var(X)=E(X^2)-\mu^2=27-5^2=2$, so $\sigma=\sqrt{2}$.	Use $Var(X)=E(X^2)-\mu^2$.
$E(S)=2\mu=2(5)=10$.	Sum of $n=2$ iid copies: mean $n\mu$.
$Var(S)=2\sigma^2=2(2)=4$, so $sd(S)=\sqrt{4}=2$.	Variance of the sum is $n\sigma^2$; sd is $\sqrt{n}\sigma$.
$E(2X)=2\mu=10$, the same mean as S.	Scaling one token by 2: mean 2μ.
$Var(2X)=2^2\sigma^2=4(2)=8$, so $sd(2X)=\sqrt{8}=2\sqrt{2}\approx 2.83$.	Variance of nX is $n^2\sigma^2$; sd is $n\sigma$.
The $3\times 3=9$ ordered pairs give $P(S=6)=1/4\cdot 1/4=1/16$ and	Only $3+3$ makes 6; 8 comes from $3+5$ or $5+3$.

Working	Comment
$P(S=8)=2(1/4\cdot 1/2)=4/16$.	
$P(S=10)=2(1/4\cdot 1/4)+1/2\cdot 1/2=2/16+4/16=6/$.	Three ordered pairs total 10: $3+7$, $7+3$ and $5+5$.
By symmetry $P(S=12)=4/16$ and $P(S=14)=1/16$.	The five probabilities add to $16/16=1$.
$2X$ takes 6, 10, 14 with probabilities $1/4$, $1/2$, $1/4$.	Doubling changes the values, not the probabilities.

The means agree (10 points), but $sd(S)=2$ is smaller than $sd(2X)=2\sqrt{2}$: two *separate* spins partly average out, while doubling *one* spin magnifies its single deviation. Writing both distributions in sixteenths makes the comparison direct:

s	6	8	10	12	14
$P(S=s)$	$1/16$	$4/16$	$6/16$	$4/16$	$1/16$

t	6	10	14
$P(2X=t)$	$4/16$	$8/16$	$4/16$

Both are symmetric about 10, and $2X$ is in fact *on* the centre rather more often than S is ($8/16$ against $6/16$). What separates them is the size of a miss. The sum has the intermediate values 8 and 12, which it takes with probability $4/16+4/16=8/16$, and it reaches the extremes 6 and 14 only when both spins agree, with probability just $2/16$. Since $2X$ has no intermediate values at all, every miss costs it the full 4 points, and it misses with probability $8/16$. Averaging the squared distances from 10 gives $Var(S)=4$ against $Var(2X)=8$.

Example 3 — unscaffolded

In a game, a red spinner scores X points and a blue spinner scores Y points, the two spins being independent. Their distributions are

x	2	6	10
$P(X=x)$	0.25	0.5	0.25

y	1	3	5
$P(Y=y)$	0.25	0.5	0.25

The blue score counts double, so the game total is $W=X+2Y$. Find $E(W)$, $Var(W)$ and $sd(W)$.

For the red spinner, $E(X)=2(0.25)+6(0.5)+10(0.25)=6$ and $E(X^2)=4(0.25)+36(0.5)+100(0.25)=44$, so $Var(X)=44-6^2=8$. For the blue spinner, $E(Y)=1(0.25)+3(0.5)+5(0.25)=3$ and $E(Y^2)=1(0.25)+9(0.5)+25(0.25)=11$, so $Var(Y)=11-3^2=2$. Because the two spins are independent, the mean combines linearly and the variance combines with squared coefficients:

$$E(W)=E(X)+2E(Y)=6+2(3)=12,$$

$$Var(W)=Var(X)+2^2 Var(Y)=8+4(2)=16,$$

so $sd(W)=\sqrt{16}=4$.

$\therefore W=X+2Y$ has mean 12, variance 16 and standard deviation 4.

Example 4 — unscaffolded

In a darts-style game each throw scores X points, where $X=4$ (an outer hit) with probability 0.6 and $X=9$ (an inner hit) with probability 0.4. A round consists of 6 **independent** throws, giving a total $T=X_1+X_2+\ldots+X_6$. Find the mean, variance and standard deviation of T. A player claims the round is "just $6X$" and so has standard deviation 6σ; explain why this is wrong.

One throw has $E(X)=4(0.6)+9(0.4)=2.4+3.6=6$ and $E(X^2)=16(0.6)+81(0.4)=9.6+32.4=42$, so $Var(X)=42-6^2=6$ and $\sigma=\sqrt{6}$. As a sum of 6 iid throws,

$$E(T)=6\mu=6(6)=36,\, Var(T)=6\sigma^2=6(6)=36,$$

so $sd(T)=\sqrt{36}=6$. The player's model $6X$ instead scales a **single** throw: it has the same mean $6\mu=36$, but $Var(6X)=6^2\sigma^2=36(6)=216$, giving $sd(6X)=6\sqrt{6}\approx 14.70$ — more than twice as large. The six throws are separate independent observations whose deviations partly cancel, so the true total is far less variable than six copies of one throw.

$\therefore T$ has mean 36, variance 36 and standard deviation 6, not the variance 216 that the incorrect model $6X$ would give.

Practice questions

Scaffolded

Q1. Independent random variables X and Y have the means and variances below.

Variable	Mean	Variance
X	9	8
Y	4	5

(a) Find $E(X+Y)$ and $Var(X+Y)$.
(b) Find $E(X-Y)$ and $Var(X-Y)$.
(c) Find $E(2X+3Y)$ and $Var(2X+3Y)$.
(d) Find $sd(2X+3Y)$.

Q2. A carnival game pays out X tickets each play, with mean 12 and standard deviation 5 (so $\sigma^2=25$); plays are independent. A child plays $n=9$ times, for a total of $S=X_1+X_2+\ldots+X_9$ tickets. (a) Find $E(S)$. (b) Find $Var(S)$. (c) Find $sd(S)$. (d) Find $Var(9X)$ and explain why it differs from $Var(S)$.

Q3. Independent random variables X and Y have the distributions below.

x	4	6	8
$P(X=x)$	0.25	0.5	0.25

y	8	12
$P(Y=y)$	0.5	0.5

(a) Find $E(X)$ and $Var(X)$.
(b) Find $E(Y)$ and $Var(Y)$.
(c) Find $E(X+Y)$ and $Var(X+Y)$.
(d) Find $E(2X-Y)$, $Var(2X-Y)$ and $sd(2X-Y)$.

Q4. Two independent machines produce outputs X and Y, where X has mean 500 and variance 36, and Y has mean 480 and variance 64. Let $D = X - Y$. (a) Find $E(D)$. (b) Find $Var(D)$ and $sd(D)$. (c) Explain why $Var(D)$ is a **sum**, not a difference, of the two variances. (d) State $Var(X+Y)$ and compare it with $Var(X-Y)$.

Q5. Independent variables X_1, X_2, X_3 have the means and variances below, and $W = 3X_1 - 2X_2 + X_3 + 4$.

Variable	Mean	Variance
X_1	5	2
X_2	8	3
X_3	2	1

(a) Find $E(W)$.
(b) Find $Var(W)$.
(c) Find $sd(W)$.

Q6. A random variable X has mean $\mu = 20$ and variance $\sigma^2 = 7$. Consider $n = 3$ independent copies X_1, X_2, X_3. (a) Find $E(X_1 + X_2 + X_3)$ and $Var(X_1 + X_2 + X_3)$. (b) Find $E(3X)$ and $Var(3X)$. (c) Find the standard deviation of each. (d) Explain why the two variances differ although the means are equal.

Unscaffolded

Q7. Independent variables X and Y have $E(X) = 30$, $sd(X) = 5$, $E(Y) = 18$ and $sd(Y) = 12$. Find the mean and standard deviation of $X + Y$ and of $X - Y$, and comment on their spreads.

Q8. A random variable X has mean 15 and variance 10, and $X_1, X_2, \ldots, X_5$ are 5 independent copies of it. Find the mean, variance and standard deviation of the sum $S = X_1 + X_2 + \ldots + X_5$ and of $5X$, and contrast them.

Q9. A random variable X has the distribution below.

x	5	10	15
$P(X = x)$	0.3	0.4	0.3

(a) Find $E(X)$ and $Var(X)$, and hence the mean and standard deviation of the total $S = X_1 + X_2 + X_3 + X_4$ of 4 independent copies of X.
(b) For two independent copies, write down the probability distribution of $X_1 + X_2$ and the probability distribution of $2X$, and use them to compare the two spreads.

Q10. Two independent delivery routes take times X minutes on route P and Y minutes on route Q, where X has mean 34 and standard deviation 7, and Y has mean 28 and standard deviation 4. For $D = X - Y$, find $E(D)$ and $sd(D)$, and comment on whether route P is reliably the slower one.

Q11. Independent variables X and Y have mean 6, variance 4 and mean 9, variance 5 respectively. A revenue is $R = 5X + 2Y$. Find $E(R)$, $Var(R)$ and $sd(R)$.

Q12. A single component has mass with mean 50 g and standard deviation 3 g, and components are packed independently. (a) Find the mean and standard deviation of the total mass of 16 components. (b) How many independent components give the total a standard deviation of 15 g?

Q13. A courier's charge is $C = 8 + 3X + 2Y$ points, where X has mean 4 and variance 2, and Y has mean 5 and variance 3, the two being independent. Find $E(C)$, $Var(C)$ and $sd(C)$.

Q14. Let X and Y be independent random variables with variances $Var(X)$ and $Var(Y)$. Using the definition of variance and the independence of X and Y, show that $Var(X - Y) = Var(X) + Var(Y)$.

Q15. Independent variables X, Y, Z have means 7, 4, 10 and variances 3, 5, 2 respectively. For $W = 2X - Y + 4Z$, find $E(W)$, $Var(W)$ and $sd(W)$.

Q16. Each round of a game awards X points, where $X=2$ or $X=10$ each with probability 0.5. Find $E(X)$ and $Var(X)$, and hence the mean, variance and standard deviation of the total $S=X_1+X_2+X_3+X_4$ over 4 independent rounds.

Q17. A random variable X takes the value 1 or 7, each with probability 0.5. (a) Find $E(X)$ and $Var(X)$. (b) For two independent copies, find the mean and standard deviation of X_1+X_2. (c) Find the mean and standard deviation of $2X$. (d) Write down the probability distribution of X_1+X_2 and the probability distribution of $2X$, and use them to explain why the two have different spreads even though their means are equal.

Q18. A workshop performs two independent tasks, taking X minutes and Y minutes, where X has mean 25 and standard deviation 4, and Y has mean 15 and standard deviation 3. (a) For the total time $T=X+Y$, find $E(T)$ and $sd(T)$. (b) For the difference $D=X-Y$, find $E(D)$ and $sd(D)$. (c) The first task is repeated on n independent items, giving a total time $M=X_1+X_2+\ldots+X_n$. Find n so that $sd(M)=24$ minutes.

Answers

Q1. (a) $E=13$, $Var=13$; (b) $E=5$, $Var=13$; (c) $E=30$, $Var=77$; (d) $sd=\sqrt{77}\approx 8.77$. **Q2.** (a) 108; (b) 225; (c) 15; (d) $Var(9X)=2025$ — nine times *one* play, not nine separate plays. **Q3.** (a) $E(X)=6$, $Var(X)=2$; (b) $E(Y)=10$, $Var(Y)=4$; (c) $E=16$, $Var=6$; (d) $E=2$, $Var=12$, $sd=2\sqrt{3}\approx 3.46$. **Q4.** (a) 20; (b) $Var(D)=100$, $sd(D)=10$; (c) the coefficient -1 squares to $+1$, so independent variances accumulate; (d) $Var(X+Y)=100$, identical to $Var(X-Y)$. **Q5.** (a) 5; (b) 31; (c) $\sqrt{31}\approx 5.57$. **Q6.** (a) $E=60$, $Var=21$; (b) $E=60$, $Var=63$; (c) $\sqrt{21}\approx 4.58$ and $3\sqrt{7}\approx 7.94$; (d) $3X$ scales one observation (variance $9\sigma^2$), the sum adds three independent ones (variance $3\sigma^2$). **Q7.** $X+Y$: mean 48, $sd=13$; $X-Y$: mean 12, $sd=13$; equal spread, since the variances $25+144=169$ add either way. **Q8.** S: mean 75, variance 50, $sd=5\sqrt{2}\approx 7.07$; $5X$: mean 75, variance 250, $sd=5\sqrt{10}\approx 15.81$. **Q9.** (a) $E(X)=10$, $Var(X)=15$; S: mean 40, variance 60, $sd=2\sqrt{15}\approx 7.75$; (b) X_1+X_2 takes $10, 15, 20, 25, 30$ with probabilities $0.09, 0.24, 0.34, 0.24, 0.09$, while $2X$ takes $10, 20, 30$ with probabilities $0.3, 0.4, 0.3$ — both have mean 20, but $Var(X_1+X_2)=30$ against $Var(2X)=60$, so the sum is the narrower. **Q10.** $E(D)=6$ min, $sd(D)=\sqrt{65}\approx 8.06$ min; since the spread (8.06) exceeds the mean gap (6), route P is not reliably slower — the two times overlap considerably. **Q11.** $E(R)=48$, $Var(R)=120$, $sd(R)=2\sqrt{30}\approx 10.95$. **Q12.** (a) mean 800 g, $sd=12$ g; (b) $n=25$. **Q13.** $E(C)=30$, $Var(C)=30$, $sd(C)=\sqrt{30}\approx 5.48$. **Q14.** See solution. **Q15.** $E(W)=50$, $Var(W)=49$, $sd(W)=7$. **Q16.** $E(X)=6$, $Var(X)=16$; S: mean 24, variance 64, $sd=8$. **Q17.** (a) $E(X)=4$, $Var(X)=9$; (b) mean 8, $sd=3\sqrt{2}\approx 4.24$; (c) mean 8, $sd=6$; (d) X_1+X_2 takes $2, 8, 14$ with probabilities $0.25, 0.5, 0.25$, while $2X$ takes $2, 14$ with probabilities $0.5, 0.5$ — the sum puts half its probability on the centre 8, a value $2X$ can never take, so the sum is less spread out. **Q18.** (a) $E(T)=40$, $sd(T)=5$; (b) $E(D)=10$, $sd(D)=5$; (c) $n=36$.

Fully-worked solutions

Q1. With X, Y independent: $E(X)=9$, $Var(X)=8$, $E(Y)=4$, $Var(Y)=5$. (a) $E(X+Y)=9+4=13$ and $Var(X+Y)=8+5=13$. (b) $E(X-Y)=9-4=5$ and $Var(X-Y)=8+5=13$ — the variances add here too, since the coefficient -1 squares to $+1$. (c) $E(2X+3Y)=2(9)+3(4)=18+12=30$ and $Var(2X+3Y)=2^2(8)+3^2(5)=32+45=77$. (d) $sd(2X+3Y)=\sqrt{77}\approx 8.77$.

Q2. One play: $\mu=12$, $\sigma^2=25$, and $n=9$ independent plays. (a) $E(S)=9\mu=9(12)=108$. (b) $Var(S)=9\sigma^2=9(25)=225$. (c) $sd(S)=\sqrt{225}=15$. (d) $Var(9X)=9^2\sigma^2=81(25)=2025$. This is the variance of 9 times a **single** play, whose one deviation is magnified by 9 (variance $\times 9^2$). The total S instead sums 9 **separate** independent plays, whose deviations partly cancel, so its variance grows only by a factor of 9. Hence $Var(9X)=2025$ is far larger than $Var(S)=225$.

Q3. (a) $E(X)=4(0.25)+6(0.5)+8(0.25)=6$ and $E(X^2)=16(0.25)+36(0.5)+64(0.25)=38$, so $Var(X)=38-6^2=2$. (b) $E(Y)=8(0.5)+12(0.5)=10$ and $E(Y^2)=64(0.5)+144(0.5)=104$, so $Var(Y)=104-10^2=4$. (c) $E(X+Y)=6+10=16$ and $Var(X+Y)=2+4=6$. (d) $E(2X-Y)=2(6)-10=2$ and $Var(2X-Y)=2^2(2)+(-1)^2(4)=8+4=12$, so $sd(2X-Y)=\sqrt{12}=2\sqrt{3}\approx 3.46$.

Q4. X mean 500, variance 36; Y mean 480, variance 64; independent, $D=X-Y$. (a) $E(D)=500-480=20$. (b) $Var(D)=36+64=100$, so $sd(D)=\sqrt{100}=10$. (c) Subtracting Y carries the coefficient -1, and $(-1)^2=1$; because the variables are independent their variabilities accumulate, so $Var(X-Y)=Var(X)+Var(Y)$ — a sum, not a difference. (d) $Var(X+Y)=36+64=100$, identical to $Var(X-Y)$: a sum and a difference of the same two independent variables have the same spread.

Q5. Independent X_1, X_2, X_3; $W=3X_1-2X_2+X_3+4$. (a) $E(W)=3(5)-2(8)+1(2)+4=15-16+2+4=5$. (b) The added constant 4 contributes no variance, and each coefficient is squared: $Var(W)=3^2(2)+(-2)^2(3)+1^2(1)=18+12+1=31$. (c) $sd(W)=\sqrt{31}\approx 5.57$.

Q6. $\mu=20$, $\sigma^2=7$, $n=3$. (a) $E(X_1+X_2+X_3)=3\mu=60$ and $Var(X_1+X_2+X_3)=3\sigma^2=3(7)=21$. (b) $E(3X)=3\mu=60$ and $Var(3X)=3^2\sigma^2=9(7)=63$. (c) $sd(X_1+X_2+X_3)=\sqrt{21}\approx 4.58$ and $sd(3X)=\sqrt{63}=3\sqrt{7}\approx 7.94$. (d) Both means are $3\mu=60$, but the sum of three independent copies has variance $3\sigma^2$ (deviations partly offset), while $3X$ scales one observation's deviation by 3, giving variance $3^2\sigma^2=9\sigma^2$ — three times as large.

Q7. Independent, so $Var(X)=5^2=25$ and $Var(Y)=12^2=144$. $X+Y$: mean $30+18=48$, variance $25+144=169$, so $sd=13$. $X-Y$: mean $30-18=12$, variance $25+144=169$, so $sd=13$. The two spreads are identical: whether the variables are added or subtracted, the independent variances add to 169, giving standard deviation 13 in both cases.

Q8. iid, each mean 15, variance 10, with $n=5$. $S=X_1+\ldots+X_5$: mean $5(15)=75$, variance $5(10)=50$, so $sd=\sqrt{50}=5\sqrt{2}\approx 7.07$. $5X$: mean $5(15)=75$, variance $5^2(10)=250$, so $sd=\sqrt{250}=5\sqrt{10}\approx 15.81$. Same mean 75, but $5X$ has five times the variance of S (250 against 50), because it magnifies a single observation rather than summing five independent ones.

Q9. (a) $E(X)=5(0.3)+10(0.4)+15(0.3)=1.5+4+4.5=10$ and $E(X^2)=25(0.3)+100(0.4)+225(0.3)=7.5+40+67.5=115$, so $Var(X)=115-10^2=15$. The total of 4 independent copies has

$$E(S)=4(10)=40, Var(S)=4(15)=60,$$

so $sd(S)=\sqrt{60}=2\sqrt{15}\approx 7.75$. (b) Each ordered pair (x_1, x_2) has probability $P(X=x_1)P(X=x_2)$, and pairs with the same total are collected: $P(X_1+X_2=10)=0.3^2=0.09$; $P(X_1+X_2=15)=2(0.3)(0.4)=0.24$; $P(X_1+X_2=20)=2(0.3)(0.3)+0.4^2=0.18+0.16=0.34$; $P(X_1+X_2=25)=2(0.4)(0.3)=0.24$; $P(X_1+X_2=30)=0.3^2=0.09$.

s	10	15	20	25	30
$P(X_1+X_2=s)$	0.09	0.24	0.34	0.24	0.09

Doubling instead keeps the probabilities and doubles the values:

t	10	20	30
$P(2X=t)$	0.3	0.4	0.3

Both have mean 20, but $Var(X_1+X_2)=2(15)=30$ while $Var(2X)=2^2(15)=60$. The distributions show why: $2X$ takes only the three values 10, 20, 30, so a miss of the centre always costs 10, and it misses with probability 0.6. The sum has the extra values 15 and 25, taken with probability $0.24+0.24=0.48$, which are only 5 from the centre, and it reaches the extremes 10 and 30 with probability just $0.09+0.09=0.18$.

Q10. $Var(X)=7^2=49$ and $Var(Y)=4^2=16$. For $D=X-Y$, $E(D)=34-28=6$ minutes and $Var(D)=49+16=65$, so $sd(D)=\sqrt{65}\approx 8.06$ minutes. Although route P is on average 6 minutes slower, the standard deviation of the difference, about 8.06 minutes, is larger than that gap, so D is often negative: route P is **not** reliably the slower one, and the two times overlap considerably.

Q11. $R=5X+2Y$ with X, Y independent. $E(R)=5(6)+2(9)=30+18=48$. $Var(R)=5^2(4)+2^2(5)=100+20=120$. $sd(R)=\sqrt{120}=2\sqrt{30}\approx 10.95$.

Q12. One component: $\mu=50$, $\sigma=3$ (so $\sigma^2=9$), components independent. (a) The total of 16 components has mean $16(50)=800$ g and variance $16(9)=144$, so $sd=\sqrt{144}=12$ g. (b) The total of n components has standard deviation $\sqrt{n}\sigma=3\sqrt{n}$. Setting $3\sqrt{n}=15$ gives $\sqrt{n}=5$, so $n=25$ components.

Q13. $C=8+3X+2Y$ with X, Y independent. $E(C)=8+3(4)+2(5)=8+12+10=30$. The constant 8 adds no variance, and each coefficient is squared: $Var(C)=3^2(2)+2^2(3)=18+12=30$, so $sd(C)=\sqrt{30}\approx 5.48$.

Q14. Write $\mu_X=E(X)$ and $\mu_Y=E(Y)$; then $E(X-Y)=\mu_X-\mu_Y$. By the definition of variance,

$$Var(X-Y)=E\left[\left((X-Y)-(\mu_X-\mu_Y)\right)^2\right]=E\left[\left((X-\mu_X)-(Y-\mu_Y)\right)^2\right].$$

Expanding the square,

$$Var(X-Y)=E\left[(X-\mu_X)^2\right]-2E\left[(X-\mu_X)(Y-\mu_Y)\right]+E\left[(Y-\mu_Y)^2\right].$$

The first and last terms are $Var(X)$ and $Var(Y)$. Because X and Y are independent, $E(XY)=E(X)E(Y)=\mu_X\mu_Y$, and so

$$E\left[(X-\mu_X)(Y-\mu_Y)\right]=E(XY)-\mu_X\mu_Y=\mu_X\mu_Y-\mu_X\mu_Y=0.$$

The middle term therefore vanishes, leaving $Var(X-Y)=Var(X)+Var(Y)$, as required.

Q15. $W=2X-Y+4Z$ with X, Y, Z independent. $E(W)=2(7)-1(4)+4(10)=14-4+40=50$. $Var(W)=2^2(3)+(-1)^2(5)+4^2(2)=12+5+32=49$, so $sd(W)=\sqrt{49}=7$. (Each coefficient is squared, so the -1 contributes $+5$.)

Q16. $E(X)=2(0.5)+10(0.5)=6$ and $E(X^2)=4(0.5)+100(0.5)=52$, so $Var(X)=52-6^2=16$. The total of 4 independent rounds has

$$E(S)=4(6)=24,\ Var(S)=4(16)=64,$$

so $sd(S)=\sqrt{64}=8$.

Q17. (a) $E(X)=1(0.5)+7(0.5)=4$ and $E(X^2)=1(0.5)+49(0.5)=25$, so $Var(X)=25-4^2=9$. (b) For two independent copies, $E(X_1+X_2)=2(4)=8$ and $Var(X_1+X_2)=2(9)=18$, so $sd(X_1+X_2)=\sqrt{18}=3\sqrt{2}\approx 4.24$. (c) $E(2X)=2(4)=8$ and $Var(2X)=2^2(9)=36$, so $sd(2X)=\sqrt{36}=6$. (d) Each of the four ordered pairs (x_1,x_2) has probability $0.5\times 0.5=0.25$, so collecting equal totals gives $P(X_1+X_2=2)=0.25$ (from $1+1$), $P(X_1+X_2=8)=0.25+0.25=0.5$ (from $1+7$ and $7+1$) and $P(X_1+X_2=14)=0.25$ (from $7+7$). Doubling keeps the probabilities and doubles the values, so $2X$ takes 2 and 14, each with probability 0.5.

s	2	8	14
$P(X_1+X_2=s)$	0.25	0.5	0.25

t	2	14
$P(2X=t)$	0.5	0.5

Half of the sum’s probability sits on the central value 8, which $2X$ can never take: a mixed pair of spins cancels, but a doubled value cannot. Forced always to an extreme, $2X$ is more spread out (variance 36) than the sum (variance 18), even though both average 8.

Q18. $Var(X)=4^2=16$ and $Var(Y)=3^2=9$. (a) $T=X+Y$: $E(T)=25+15=40$ and $Var(T)=16+9=25$, so $sd(T)=\sqrt{25}=5$ minutes. (b) $D=X-Y$: $E(D)=25-15=10$ and $Var(D)=16+9=25$ (the variances add for a difference), so $sd(D)=5$ minutes. (c) The total of n independent copies of X has standard deviation $\sqrt{n}(4)=4\sqrt{n}$. Setting $4\sqrt{n}=24$ gives $\sqrt{n}=6$, so $n=36$.

Theory

Much of statistics is about learning what is true of a whole **population** from a much smaller **sample** that we can actually observe. Before any real data are collected, a great deal can be understood by *modelling* the chance mechanism on a computer or with a simple physical device and watching how the numbers we compute behave. This section sets up the language of **random experiments**, shows how a **simulation** imitates one, and draws the central distinction of inferential statistics — between a fixed **population parameter** and a varying **sample statistic**.

Random experiments, outcomes and events.

A **random experiment** is a process that can be repeated under the same conditions, whose individual result cannot be predicted in advance, but for which the complete list of possible results *is* known beforehand. Rolling a die, tossing three coins, or drawing a card are all random experiments. The set of all possible results is the **sample space**, also called the **event space**, written S; a single element of S is an **outcome**; and an **event** is any subset of S — a collection of outcomes we choose to single out.

For example, rolling a four-sided die with faces $1, 2, 3, 4$ has sample space (event space) $S = \{1, 2, 3, 4\}$, the outcome "a 3" is the element 3, and the event "an even number" is the subset $\{2, 4\}$. Outcomes are **equally likely** when each has the same chance; they need not be. It is the known structure of the sample space that makes a random experiment possible to *model*.

Simulation.

A **simulation** is a model of a random experiment that reproduces its probabilities using a simpler, more convenient mechanism — a coin, a die, a **spinner**, a table of **random digits**, or **technology** (the random-number generator of a calculator or spreadsheet). A well-built simulation is run many times, each run being one **trial**, and the results are recorded as if the real experiment had been performed.

To construct a simulation:

- identify the outcomes of the experiment and their probabilities;
- choose a random device and **assign** to each outcome a share of the device in proportion to its probability;
- run the trials, reading the device, and record the outcome of each.

With **random digits** $0, 1, \ldots, 9$ (each equally likely), an outcome of probability $0.3 = 3/10$ is modelled by assigning it three of the ten digits, say $0, 1, 2$; the remaining digits 3–9 model the complementary outcome. Probabilities needing finer resolution use **two-digit** numbers 00–99: a probability of 0.12 is modelled by the twelve numbers 00–11. A **spinner** achieves the same effect geometrically — its sectors have central angles proportional to the probabilities, so an outcome of probability 0.25 is given a sector of $0.25 \times 360 = 90$, that is, a 90° sector. Technology instead returns a decimal spread evenly over $[0, 1)$, and an outcome of probability p is recorded whenever the returned value is less than p.

Populations, parameters and sample statistics.

A **population** is the entire collection of individuals or values under study; a **sample** is the subset of the population actually observed. A sample is a **random sample** when it is selected so that chance alone decides which members are included, with no systematic bias; a simulation produces a random sample drawn from the *model* of the population.

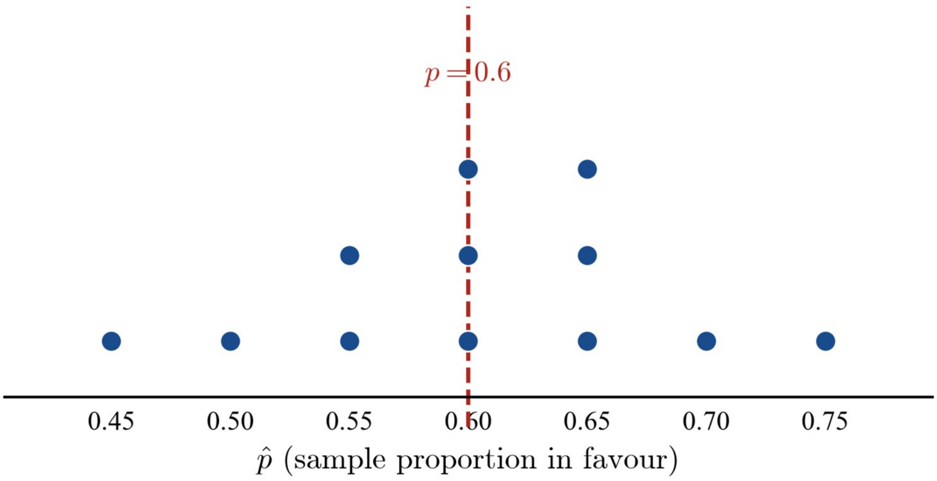
A numerical summary of the *whole population* is a **population parameter**. It is a single fixed number — generally unknown — such as the population mean μ or the population proportion p. A numerical summary computed from a *sample* is a **sample statistic**, such as the **sample mean** $\acute{x}$ or the **sample proportion** $\hat{p}$. The crucial difference is this:

Quantity	Symbol	Computed from	Value
Population mean	μ	the whole population	fixed (a parameter)

Quantity	Symbol	Computed from	Value
Sample mean	$\bar{x}$	one sample	varies sample to sample (a statistic)
Population proportion	p	the whole population	fixed (a parameter)
Sample proportion	$\hat{p}$	one sample	varies sample to sample (a statistic)

Because a different sample generally gives a different value, a sample statistic is itself a random quantity: the spread of its values across repeated samples is called **sampling variability**. A parameter, by contrast, does not change — it is a property of the population, not of any sample.

The dot plot below illustrates this. Twelve random samples of 20 voters were simulated from a population in which a fixed proportion $p=0.6$ favour a proposal; each sample gives a sample proportion $\hat{p}=\frac{\text{number in favour}}{20}$.



The twelve statistics scatter from 0.45 to 0.75 about the fixed parameter $p=0.6$ (the dashed line). No single $\hat{p}$ need equal p, yet the values cluster around it: their average is $\frac{7.25}{12}\approx 0.604$, close to p.

Computing a sample mean and a sample proportion.

For a sample of n numerical values $x_1, x_2, \ldots, x_n$, the **sample mean** is their average,

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

When the data are grouped in a frequency table, with each value x occurring with frequency f, this is computed as

$$\bar{x}=\frac{\sum f x}{\sum f}.$$

For a categorical sample in which each member either does or does not have a feature of interest (a “success”), the **sample proportion** is

$$\hat{p}=\frac{\text{number of successes}}{\text{sample size}}=\frac{x}{n}.$$

If each member is coded 1 for a success and 0 otherwise, then the number of successes is the sum of the codes and the sample proportion is exactly the sample mean of the $0/1$ data: $\hat{p}$ is a sample mean in disguise.

Representing simulation results.

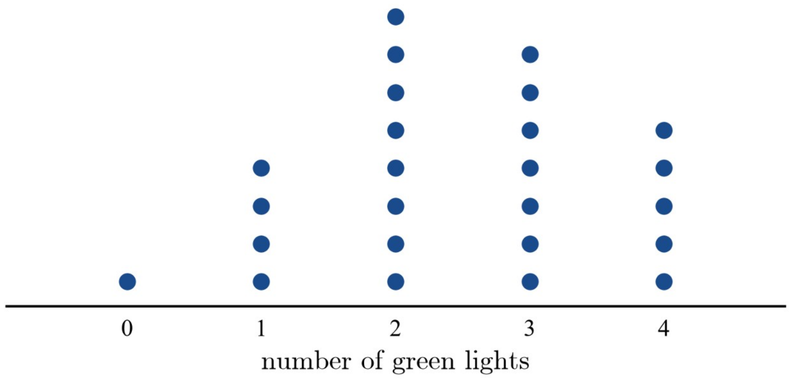
The results of many trials are summarised by counting. The **frequency** of an outcome is the number of trials giving it; its **relative frequency** is

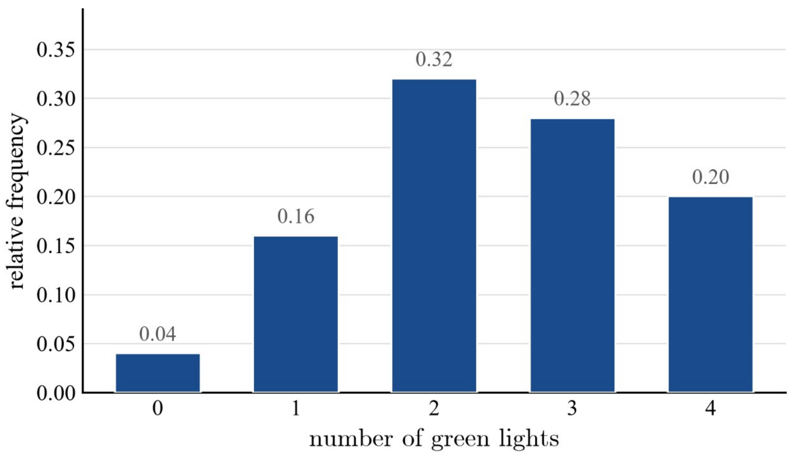
$$\text{relative frequency} = \frac{\text{frequency}}{\text{total number of trials}}.$$

Relative frequencies lie between 0 and 1 and sum to 1 over all outcomes, and they serve as an *estimate* of the underlying probability: as the number of trials grows, the relative frequency of an outcome tends to settle towards its true probability.

A **dot plot** displays a small or moderate dataset by stacking one dot for each observation above its value on a number line. It shows the shape, centre and spread of the data at a glance. The two diagrams below summarise a simulation of 25 commuter trips, each recording how many of four traffic signals were green:

Number green	0	1	2	3	4
Frequency	1	4	8	7	5





The dot plot (left-hand values first) shows the stacked counts; the relative-frequency plot rescales them to $0.04, 0.16, 0.32, 0.28, 0.20$, which sum to 1. The sample mean number of green signals is

$$\acute{x} = \frac{0(1)+1(4)+2(8)+3(7)+4(5)}{25} = \frac{61}{25} = 2.44.$$

This single simulated dataset gives one value of the statistic $\acute{x}$; a fresh set of 25 trips would generally give another.

Worked examples

Example 1 — scaffolded

A native wildflower seed germinates with probability 0.3; this population proportion $p=0.3$ is a fixed number. Model the germination of 20 seeds with single random digits, using the digits

$$7\,0\,4\,2\,9\,1\,5\,3\,8\,0\,6\,2\,1\,9\,4\,7\,0\,3\,5\,2$$

(read left to right, one per seed). Find the sample proportion that germinate and compare it with p.

Working	Comment
$p=0.3=3/10$, so let the digits $0,1,2$ mean "germinates" and 3–9 mean "does not".	Three of the ten equally-likely digits are assigned to germination, matching 0.3.
Each of the 20 digits is one seed; tally those in $\{0,1,2\}$.	The remaining digits 3–9 are the seeds that do not germinate.
The digits in $\{0,1,2\}$ occur 8 times among the 20.	Counting across the list gives 8 germinations.
Sample proportion $\hat{p}=\dfrac{8}{20}=0.4$.	Sample proportion = successes ÷ sample size.
Here $\hat{p}=0.4$ while $p=0.3$.	The statistic $\hat{p}$ differs from the parameter p — sampling variability.

Example 2 — scaffolded

A quality inspector opens 20 boxes of memory sticks and records the number of faulty sticks per box, summarised below.

Faults	0	1	2	3	4
Frequency	6	7	4	2	1

Find (a) the sample mean number of faults per box, (b) the sample proportion of boxes with at least one faulty stick, and (c) the relative-frequency distribution.

Working	Comment
Sample size $n=6+7+4+2+1=20$.	The sample size is the sum of the frequencies.
$\acute{x}=\dfrac{0(6)+1(7)+2(4)+3(2)+4(1)}{20}=\dfrac{25}{20}=1.25$.	Sample mean from a frequency table, $\acute{x}=\dfrac{\sum f\,x}{\sum f}$.
Boxes with at least one fault $=20-6=14$, so $\hat{p}=\dfrac{14}{20}=0.7$.	"At least one" is the complement of "zero faults".
Relative frequencies: $6/20,7/20,4/20,2/20,1/20=0.30,0.35,0.20,0.10$.	Each is frequency $\div\, n$; these sum to 1.

Example 3 — unscaffolded

A board game uses a spinner with three outcomes: **advance** (probability 0.5), **hold** (probability 0.2) and **retreat** (probability 0.3). Describe how both a spinner and a two-digit random-number rule could model one spin, and — given that a simulation of 50 spins produced advance 27, hold 8 and retreat 15 — find the relative frequency of each outcome and compare it with the given probabilities.

A spinner is divided into three sectors with central angles proportional to the probabilities: advance $0.5\times360=180$, hold $0.2\times360=72$ and retreat $0.3\times360=108$ — that is, sectors of 180°, 72° and 108°, which sum to 360°. Equivalently, using two-digit random numbers 00–99, assign 00–49 (fifty numbers) to advance, 50–69 (twenty numbers) to hold and 70–99 (thirty numbers) to retreat, matching 0.5, 0.2 and 0.3.

For the 50 recorded spins the relative frequencies are

$$\frac{27}{50}=0.54,\frac{8}{50}=0.16,\frac{15}{50}=0.30.$$

These sample statistics are close to the fixed probabilities 0.5, 0.2 and 0.3: the first two sit a little away from them, while the relative frequency of retreat happens to land exactly on 0.3. Increasing the number of spins would tend to reduce the remaining differences.

$\therefore$ the simulated relative frequencies 0.54, 0.16, 0.30 estimate the probabilities 0.5, 0.2, 0.3.

Example 4 — unscaffolded

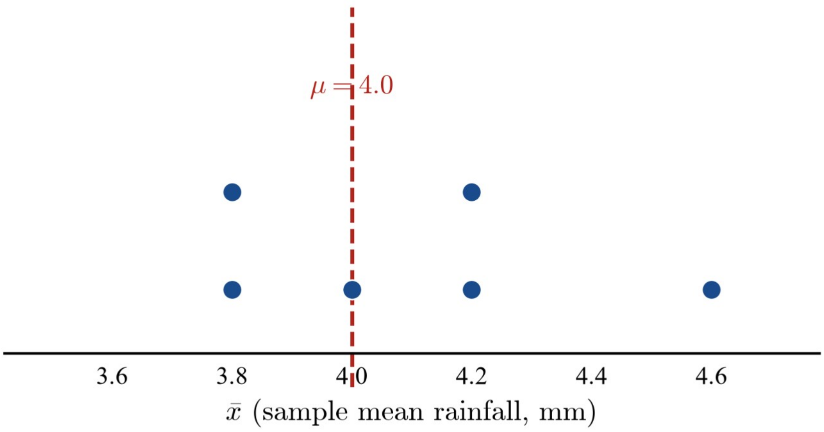
The daily rainfall (in mm) at a weather station has population mean μ. A modeller draws six random samples of five days each and computes the six sample means, in mm:

$$3.8,4.2,3.8,4.6,4.0,4.2.$$

The station's long-run mean is in fact $\mu=4.0$ mm. Represent the six sample means on a dot plot and describe how the sample statistic behaves relative to the parameter.

The mean of the six sample means is

$$\frac{3.8+4.2+3.8+4.6+4.0+4.2}{6}=\frac{24.6}{6}=4.1\,\text{mm}.$$



Each sample mean $\acute{x}$ is a sample statistic: it changes from sample to sample, here ranging from 3.8 to 4.6 mm, and is scattered about the fixed parameter $\mu=4.0$ mm (the dashed line). No single $\acute{x}$ **need** equal μ — here the fifth sample mean, 4.0 mm, happens to, while the other five do not — yet the values cluster around it, and their own average 4.1 mm is close to μ.

$\therefore$ the six sample means vary about the fixed population mean $\mu=4.0$ mm (here from 3.8 to 4.6 mm, averaging 4.1 mm).

Practice questions

Scaffolded

Q1. A fair four-sided die with faces $1,2,3,4$ is rolled and a coin is tossed. (a) List the sample space. (b) State the number of outcomes and whether they are equally likely. (c) Write the event $E=$ "an even number together with a head" as a set of outcomes. (d) State the number of outcomes in E.

Q2. A basketballer makes a free throw with probability 0.6. A single-digit simulation models 15 shots. (a) Give a random-digit rule for one shot. (b) Using the digits 381609427518603, carry out the simulation. (c) State the number of shots made. (d) Find the sample proportion $\hat{p}$ made and compare it with 0.6.

Q3. The number of rainy days in each of 12 randomly chosen weeks was

$$2,4,1,3,0,5,2,3,4,1,2,3.$$

(a) Find the sample mean number of rainy days per week. (b) Find the sample proportion of weeks with at least 3 rainy days. (c) Find the relative frequency of weeks with exactly 3 rainy days.

Q4. A simulation of 40 trials of "the number of heads in 3 tosses of a coin" gave the frequencies below.

Heads	0	1	2	3
Frequency	6	14	13	7

(a) Construct the relative-frequency distribution.
(b) Verify that the relative frequencies sum to 1.
(c) Find the sample mean number of heads.

Q5. A weather model has probabilities sunny 0.55, cloudy 0.30 and rainy 0.15. (a) Find the central angle of each sector of a spinner representing the model. (b) Give a two-digit random-number rule for the model. (c) A simulation of 60 days gave sunny 31, cloudy 20, rainy 9. Find the relative frequency of each, to 3 decimal places where needed.

Q6. In a large batch of lollies a fixed proportion $p=0.25$ are red. Eight random samples of 20 lollies each are taken; the numbers of red lollies are

$$5,6,4,7,5,3,6,4.$$

(a) Find the eight sample proportions $\hat{p}$. (b) State which quantity is the population parameter and which are the sample statistics. (c) Find the mean of the eight sample proportions. (d) Determine whether any sample proportion equals p.

Unscaffolded

Q7. Two discs are drawn, one after the other without replacement, from a bag of discs numbered 1 to 5. Taking order into account, how many outcomes are in the sample space, and how many outcomes make up the event "the two numbers add to 6"?

Q8. A tram is late with probability 0.2. Using the rule "digits 0 and 1 mean late", the random digits

$$4170382159063172804169503$$

simulate 25 trams. Find the number of late trams and the sample proportion late.

Q9. The numbers of eggs laid in a week by 15 hens were

$$5,6,7,6,5,4,6,7,5,6,5,7,6,4,5.$$

Find the sample mean and the sample proportion of hens laying at least 6 eggs.

Q10. A simulation of 50 trials of "the number of sixes in 5 rolls of a fair die" gave the frequencies below.

Sixes	0	1	2	3	4	5
Frequency	20	18	8	3	1	0

Find (a) the relative-frequency distribution, (b) the sample mean number of sixes, and (c) the relative frequency of "at least one six".

Q11. A random experiment has four outcomes with probabilities 0.40, 0.25, 0.20 and 0.15. (a) Find the central angles of the four sectors of a spinner modelling it. (b) Give a two-digit random-number rule for the four outcomes.

Q12. A spinner is designed so that the probability of "win" is 0.35. In a simulation of 80 spins, "win" occurred 24 times. Find the relative frequency of winning and comment on how it compares with the design probability 0.35.

Q13. In a sample of 30 manufactured parts, each part is coded 1 if it passes inspection and 0 if it fails; 21 parts pass. Show that the sample proportion passing equals the sample mean of the $0/1$ data, and state its value.

Q14. Components are defective with probability 0.12. Using two-digit random numbers with the rule "00–11 means defective", the numbers

$$47\,03\,88\,81\,52\,09\,76\,34\,60\,11\,92\,07\,41\,23\,80\,55\,70\,68\,39\,64\,71\,86\,12\,50\,29$$

simulate 25 components. Find the number defective and the sample proportion defective.

Q15. In a town the mean age of residents is $\mu = 38.0$ years. A random sample of six residents has ages $41, 29, 52, 35, 44, 31$ years. Find the sample mean age and state whether it equals μ.

Q16. Two fair six-sided dice are rolled and the event of interest is "the sum is at least 9". In a simulation of 40 rolls this event occurred 11 times. (a) Find the relative-frequency estimate of the probability. (b) Find the exact theoretical probability and compare.

Q17. A call centre answers a call within 20 seconds with probability 0.7. Using the rule "digits 0–6 mean answered in time", the digits

$$5\,2\,8\,1\,6\,9\,0\,3\,7\,4\,2\,6\,8\,5\,1\,9\,3\,0\,6\,7\,4\,2\,8\,5\,1\,6\,9\,0\,3\,7$$

simulate 30 calls. Find the sample proportion answered in time and the relative frequency of calls not answered in time.

Q18. In a seed lot a fixed proportion $p = 0.8$ of seeds are viable. Ten random samples of 25 seeds each are simulated; the numbers of viable seeds are

$$21, 19, 20, 22, 18, 23, 20, 21, 19, 20.$$

(a) Find the ten sample proportions. (b) Find the overall proportion viable across all 250 seeds and the mean of the ten sample proportions. (c) State the range of the sample proportions and explain what this illustrates about a statistic and a parameter.

Answers

Q1. (a) $\{(1,\mathrm{H}),(1,\mathrm{T}),(2,\mathrm{H}),(2,\mathrm{T}),(3,\mathrm{H}),(3,\mathrm{T}),(4,\mathrm{H}),(4,\mathrm{T})\}$; (b) 8, equally likely; (c) $E=\{(2,\mathrm{H}),(4,\mathrm{H})\}$; (d) 2. **Q2.** (a) digits 0–5 made, 6–9 missed; (c) 9 made; (d) $\hat{p}=9/15=0.6$ (equal to 0.6 here). **Q3.** (a) 2.5; (b) 0.5; (c) 0.25. **Q4.** (a) $0.15, 0.35, 0.325, 0.175$; (b) sum $=1$; (c) 1.525. **Q5.** (a) 198°, 108°, 54°; (b) 00–54 sunny, 55–84 cloudy, 85–99 rainy; (c) $0.517, 0.333, 0.150$. **Q6.** (a) $0.25, 0.30, 0.20, 0.35, 0.25, 0.15, 0.30, 0.20$; (b) $p=0.25$ is the parameter, the eight $\hat{p}$ are statistics; (c) 0.25; (d) yes — the first and fifth (both 0.25). **Q7.** 20 outcomes; 4 in the event. **Q8.** 8 late; $\hat{p}=8/25=0.32$. **Q9.** $\acute{x}=5.6$; $\hat{p}=8/15\approx 0.53$. **Q10.** (a) $0.40, 0.36, 0.16, 0.06, 0.02, 0.00$; (b) 0.94; (c) 0.6. **Q11.** (a) 144°, 90°, 72°, 54°; (b) 00–39, 40–64, 65–84, 85–99. **Q12.** 0.30; below the design probability 0.35 (sampling variability). **Q13.** $\hat{p}=21/30=0.7$, equal to the mean of the $0/1$ data. **Q14.** 4 defective; $\hat{p}=4/25=0.16$. **Q15.** $\acute{x}=116/3\approx 38.7$ years; not equal to $\mu=38.0$. **Q16.** (a) $11/40=0.275$; (b) $10/36=5/18\approx 0.278$; close. **Q17.** $\hat{p}=21/30=0.7$; relative frequency not in time $=0.3$. **Q18.** (a) $0.84, 0.76, 0.80, 0.88, 0.72, 0.92, 0.80, 0.84, 0.76, 0.80$; (b) overall $203/250=0.812$, mean of $\hat{p}=0.812$; (c) range 0.72 to 0.92 — the statistic $\hat{p}$ varies while $p=0.8$ is fixed.

Fully-worked solutions

Q1. Each outcome is a (die, coin) pair, so the sample space is

$$S=\{(1,\mathrm{H}),(1,\mathrm{T}),(2,\mathrm{H}),(2,\mathrm{T}),(3,\mathrm{H}),(3,\mathrm{T}),(4,\mathrm{H}),(4,\mathrm{T})\},$$

which has $4\times 2=8$ outcomes, all equally likely (fair die, fair coin). The event "an even number together with a head" needs an even die-face and a head, giving $E=\{(2,\mathrm{H}),(4,\mathrm{H})\}$, which has 2 outcomes.

Q2. (a) Since $0.6=6/10$, let the digits $0,1,2,3,4,5$ mean "made" and $6,7,8,9$ mean "missed". (b)–(c) Among the 15 digits 381609427518603, those in $\{0,\dots,5\}$ number 9 (the others, $8,6,9,7,8,6$, are the six misses). (d) The sample proportion made is $\hat{p}=9/15=0.6$, which happens to equal the probability 0.6 exactly for this sample.

Q3. (a) The total is $2+4+1+3+0+5+2+3+4+1+2+3=30$ over $n=12$ weeks, so $\acute{x}=30/12=2.5$. (b) Weeks with at least 3 rainy days are those valued $4,3,5,3,4,3$ — reading across, there are 6 of them, so $\hat{p}=6/12=0.5$. (c) The value 3 occurs 3 times, a relative frequency of $3/12=0.25$.

Q4. (a) Dividing each frequency by 40: $6/40=0.15$, $14/40=0.35$, $13/40=0.325$, $7/40=0.175$. (b) $0.15+0.35+0.325+0.175=1$. (c) The sample mean number of heads is

$$\acute{x}=\frac{0(6)+1(14)+2(13)+3(7)}{40}=\frac{61}{40}=1.525.$$

Q5. (a) Each central angle is the probability times 360: sunny $0.55\times 360=198$, cloudy $0.30\times 360=108$, rainy $0.15\times 360=54$ — that is, 198°, 108° and 54° (summing to 360°). (b) With numbers 00–99, assign 00–54 (55 numbers) to sunny, 55–84 (30 numbers) to cloudy and 85–99 (15 numbers) to rainy. (c) The relative frequencies are $31/60\approx 0.517$, $20/60\approx 0.333$ and $9/60=0.150$.

Q6. (a) Dividing each red count by 20:

$$5/20, 6/20, 4/20, 7/20, 5/20, 3/20, 6/20, 4/20=0.25, 0.30, 0.20, 0.35, 0.25, 0.15, 0.30, 0.20.$$

(b) The fixed batch proportion $p=0.25$ is the population parameter; the eight $\hat{p}$ values are sample statistics. (c) Their mean is $0.25+0.30+0.20+0.35+0.25+0.15+0.30+0.20/8=2.00/8=0.25$ (equivalently, $40/160=0.25$ from all 160 lollies). (d) Yes — the first and fifth sample proportions are both 0.25, equal to p.

Q7. Drawing two of the five numbered discs in order, without replacement, gives $5\times 4=20$ ordered outcomes. The pairs summing to 6 are $(1,5),(5,1),(2,4),(4,2)$ — the pair $(3,3)$ is impossible without replacement — so the event has 4 outcomes.

Q8. With "digits 0 and 1 mean late", count the digits in $\{0,1\}$ among the 25: these occur 8 times. Hence the number late is 8 and the sample proportion late is $\hat{p}=8/25=0.32$ (larger than the probability 0.2, illustrating sampling variability).

Q9. The total number of eggs is $5+6+7+6+5+4+6+7+5+6+5+7+6+4+5=84$ over $n=15$ hens, so $\acute{x}=84/15=5.6$. Hens laying at least 6 eggs are those valued 6 or 7; there are 8 of them, giving $\hat{p}=8/15\approx 0.53$.

Q10. (a) Dividing each frequency by 50: $0.40, 0.36, 0.16, 0.06, 0.02, 0.00$. (b) The sample mean is

$$\acute{x}=\frac{0(20)+1(18)+2(8)+3(3)+4(1)+5(0)}{50}=\frac{47}{50}=0.94\,.$$

(c) "At least one six" has frequency $18+8+3+1=30$, a relative frequency of $30/50=0.6$.

Q11. (a) Multiplying each probability by 360: $0.40\times 360=144$, $0.25\times 360=90$, $0.20\times 360=72$, $0.15\times 360=54$ — sectors of 144°, 90°, 72° and 54° (summing to 360°). (b) With numbers 00–99 assign 00–39 (40 numbers), 40–64 (25 numbers), 65–84 (20 numbers) and 85–99 (15 numbers) to the four outcomes respectively.

Q12. The relative frequency of winning is $24/80=0.30$. This sample statistic is a little below the design probability 0.35; such a gap is ordinary sampling variability, and running more spins would tend to bring the relative frequency closer to 0.35.

Q13. With 21 ones and 9 zeros in the 30 codes, the sum of the codes is 21, so the sample mean is $\acute{x}=21/30$. The sample proportion passing is also $\hat{p}=21/30$; the two are identical because the number of successes equals the sum of the 0 / 1 codes. Their common value is $21/30=0.7$.

Q14. With "00–11 means defective", find the two-digit numbers that are at most 11: these are $03, 09, 11, 07$ — four of them. Hence the number defective is 4 and the sample proportion defective is $\hat{p}=4/25=0.16$ (above the probability 0.12; sampling variability).

Q15. The sample total is $41+29+52+35+44+31=232$ over $n=6$ residents, so $\acute{x}=232/6=116/3\approx 38.7$ years. This sample statistic does not equal the fixed population parameter $\mu=38.0$ years — a sample mean need not match the population mean.

Q16. (a) The relative-frequency estimate is $11/40=0.275$. (b) For two dice there are 36 equally-likely outcomes; those with sum at least 9 are the 10 outcomes giving sums $9, 10, 11, 12$, so the theoretical probability is $10/36=5/18\approx 0.278$. The simulation estimate 0.275 is close to the theoretical 0.278.

Q17. With "digits 0–6 mean answered in time", count the digits in $\{0,\ldots,6\}$ among the 30: there are 21 (the remaining 9 digits are 7, 8 or 9). Hence the sample proportion answered in time is $\hat{p}=21/30=0.7$, and the relative frequency of calls not answered in time is $9/30=0.3$.

Q18. (a) Dividing each viable count by 25:

$$21/25, 19/25, 20/25, 22/25, 18/25, 23/25, 20/25, 21/25, 19/25, 20/25=0.84, 0.76, 0.80, 0.88, 0.72, 0.92, 0.80, 0.8$$

(b) The total viable is $21+19+20+22+18+23+20+21+19+20=203$ out of 250 seeds, an overall proportion of $203/250=0.812$; the mean of the ten sample proportions is the same, $203/250=0.812$. (c) The sample proportions range from 0.72 to 0.92: the statistic $\hat{p}$ varies from sample to sample, whereas the population parameter $p=0.8$ is a single fixed value that the statistics scatter around.

Chapter 7 Simulation, sampling and sampling distributions — 7D The distribution of sample means

Theory

A quantity measured across a whole **population** — the mean height of every Year 11 student in a state, the mean mass of every apricot in an orchard — is usually impossible to find directly: the population is too large to measure in full. Instead we take a **sample**, measure that, and use the sample to *estimate* the population quantity. This section studies how the mean of a sample behaves: how close it tends to sit to the true population mean, and how that behaviour improves as the sample grows.

Parameters and statistics.

A number that describes the whole population is a **population parameter**; it is a fixed (if unknown) constant. For a population of N values the two parameters we need are the **population mean** μ and the **population standard deviation** σ,

$$\mu=\frac{x_1+x_2+\cdots+x_N}{N},\sigma=\sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_N-\mu)^2}{N}}.$$

The standard deviation here divides by N (not $N-1$): it is the spread of the *whole population*. A number computed from a **sample** is a **sample statistic**; the one we study is the **sample mean**. For a sample $x_1, x_2, \ldots, x_n$ of size n it is

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}.$$

The essential distinction is that a parameter is **fixed** while a statistic **varies** from sample to sample: draw a new sample and $\bar{x}$ generally changes. Because $\bar{x}$ is our best single guess at the unknown μ, we call it an **estimate** of μ.

The sample mean as a random variable.

Before a sample is taken, each observation is a random variable, so the sample mean is itself a random variable, written $\bar{X}$:

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

Here $X_1, X_2, \ldots, X_n$ is a **random sample** — the draws are **independent** and each has the same distribution as the population. Throughout this section samples are taken **with replacement** (each item is returned before the next is drawn), which keeps the draws independent. A particular sample produces a particular value $\bar{x}$; the probability distribution of all the values $\bar{X}$ can take is the **sampling distribution of the sample mean**.

Building the sampling distribution by simulation.

One sample gives just one number and tells us little about $\bar{X}$. To see the whole distribution we repeat the sampling many times. This is naturally done with technology (a statistics package or a short program), following the same three steps by hand or by machine:

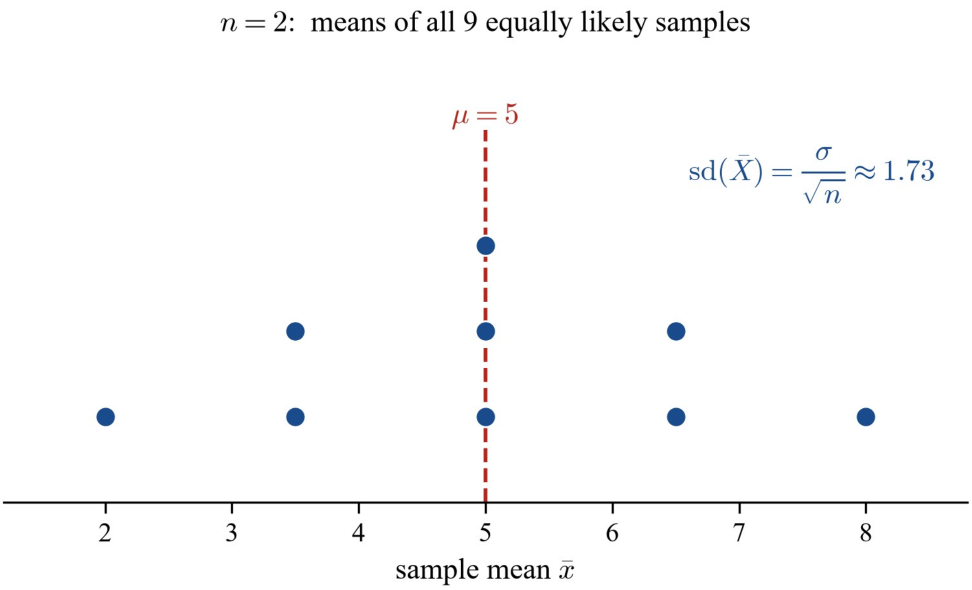
1. Draw a random sample of size n from the population and record its mean $\bar{x}$.
2. Repeat step 1 a large number of times, keeping the sample size n **fixed**.
3. Display all the recorded values $\bar{x}$ in a **dot plot**.

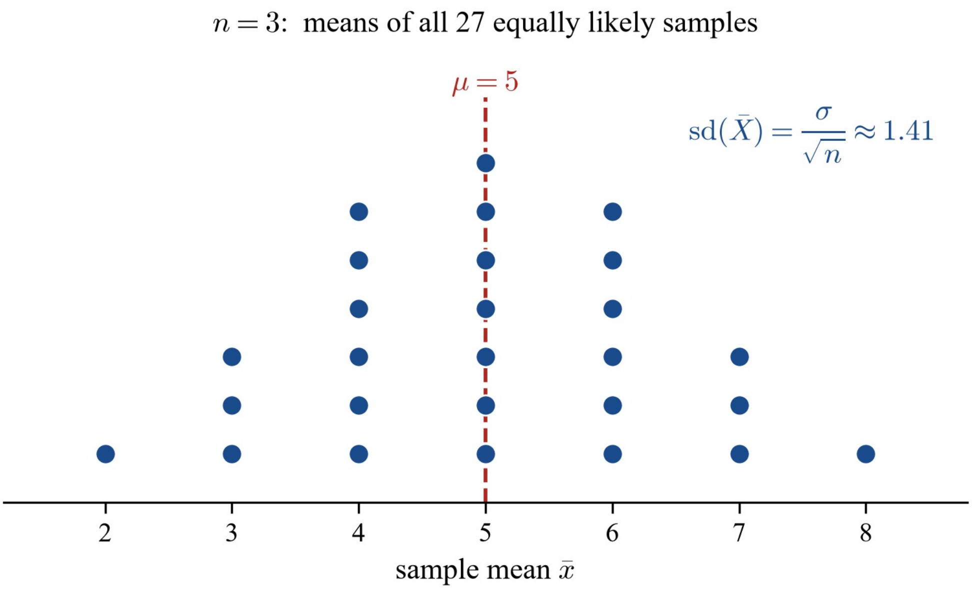
The resulting dot plot is a **simulated sampling distribution of** $\bar{X}$. The more repetitions we use, the more closely it traces the true distribution; taking *every* possible sample reproduces the sampling distribution **exactly**. The diagrams below are built that exact way, from a small population it is possible to enumerate completely.

A worked population. Consider three equally likely cards showing 2, 5 and 8, drawn with replacement. Its parameters are

$$\mu=\frac{2+5+8}{3}=5,\sigma=\sqrt{\frac{(2-5)^2+(5-5)^2+(8-5)^2}{3}}=\sqrt{\frac{18}{3}}=\sqrt{6}\approx 2.449.$$

For a sample of size $n=2$ there are $3^2=9$ equally likely ordered samples; their means take the values $2,3.5,5,6.5,8$ with frequencies $1,2,3,2,1$. For $n=3$ there are $3^3=27$ samples, with means $2,3,4,5,6,7,8$ and frequencies $1,3,6,7,6,3,1$. Each dot below is one sample's mean.





Two features stand out, and they are the heart of the topic.

- **Centre.** Both dot plots are centred exactly on the population mean $\mu=5$. The average of all the sample means is 5 whatever the sample size.
- **Spread.** The $n=3$ distribution is **more tightly clustered** about the centre than the $n=2$ distribution. Although both still range from 2 to 8, only about one-third of the $n=2$ means (3 of 9) lie within one unit of 5, against about 70 % of the $n=3$ means (19 of 27). The larger sample also gives a more **bell-shaped** picture.

The mean and standard deviation of $\acute{X}$.

These observations are exact, not accidental. For a random sample of size n from a population with mean μ and standard deviation σ,

$$E(\acute{X})=\mu, Var(\acute{X})=\frac{\sigma^2}{n}, SD(\acute{X})=\frac{\sigma}{\sqrt{n}}.$$

The first result says the sample mean is centred on the very quantity it estimates: on average it is neither too high nor too low. The second and third say the sample mean is **less spread out** than the population, by the factor $\frac{1}{\sqrt{n}}$; the quantity $\frac{\sigma}{\sqrt{n}}$ is sometimes called the **standard error** of the mean. The worked population confirms both: enumerating the $n=2$ samples gives $E(\acute{X})=5=\mu$ and $SD(\acute{X})=\sqrt{3}=\frac{\sqrt{6}}{\sqrt{2}}=\frac{\sigma}{\sqrt{2}}$, exactly as the formula predicts. (At this level we establish the results empirically, by simulation and by small enumerations like this one, rather than by formal proof.)

The effect of the sample size.

Because $SD(\acute{X})=\frac{\sigma}{\sqrt{n}}$, increasing the sample size shrinks the spread of the sampling distribution while leaving its centre at μ. For the worked population ($\sigma=\sqrt{6}$):

n	1	2	3	4	6	9
$SD(\acute{X})=\frac{\sigma}{\sqrt{n}}$	$\sqrt{6}\approx 2.449$	$\sqrt{3}\approx 1.732$	$\sqrt{2}\approx 1.414$	$\frac{\sqrt{6}}{2}\approx 1.225$	1	$\frac{\sqrt{6}}{3}\approx 0.816$

The spread falls steadily: quadrupling n halves it (because $\sqrt{4}=2$), so larger samples give sample means that are packed more and more closely around the true mean. This is why a bigger sample gives a more reliable estimate of μ.

Reading a sampling-distribution dot plot.

The relationship can be read the other way to *estimate* population figures from a simulation. In a dot plot of sample means,

- the **centre** estimates the population mean μ;
- the **spread** equals $\frac{\sigma}{\sqrt{n}}$, so multiplying it by $\sqrt{n}$ estimates the population standard deviation σ;
- the **shape** becomes more bell-like as n grows.

Throughout, keep the two roles apart: the **sample size** n fixes the true spread $\sigma/\sqrt{n}$ of the sampling distribution, while the **number of repetitions** only controls how smoothly a simulation traces that distribution — more repetitions make a neater dot plot but do not move its centre or change its spread.

Worked examples

Example 1 — scaffolded

A game uses three equally likely cards showing 2, 5 and 8 points. A card is drawn, its value recorded, then returned, so repeated draws are independent. (a) Find the population mean μ and standard deviation σ. (b) A sample of four draws gives $8, 2, 8, 5$; find the sample mean $\acute{x}$. (c) State which of μ, σ and $\acute{x}$ are population parameters and which is a sample statistic, and say what $\acute{x}$ estimates.

Working	Comment
$\mu=\frac{2+5+8}{3}=\frac{15}{3}=5.$	The population mean averages all three values.
$\sigma=\sqrt{\frac{(2-5)^2+(5-5)^2+(8-5)^2}{3}}=\sqrt{\frac{9+0+9}{3}}=\sqrt{6}\approx 2.4$	Population standard deviation: divide by $N=3$.

.

Working	Comment
$\bar{x}=\frac{8+2+8+5}{4}=\frac{23}{4}=5.75$.	The mean of this one sample of size $n=4$.
$\mu=5$ and $\sigma=\sqrt{6}$ are population **parameters** (fixed); $\bar{x}=5.75$ is a sample **statistic** (varies with the sample).	$\bar{x}=5.75$ is an estimate of $\mu=5$.

Example 2 — scaffolded

For the same three cards, a sample of size $n=2$ is drawn with replacement and its mean $\bar{X}$ is recorded. (a) List all equally likely samples and their means. (b) Tabulate the sampling distribution of $\bar{X}$. (c) Find $E(\bar{X})$ and $SD(\bar{X})$. (d) Confirm that $E(\bar{X})=\mu$ and $SD(\bar{X})=\frac{\sigma}{\sqrt{2}}$.

Working	Comment
The 9 ordered samples give means 2; $3.5, 3.5$; $5, 5, 5$; $6.5, 6.5$; 8.	From $(2,2)$; $(2,5),(5,2)$; $(2,8),(5,5),(8,2)$; $(5,8),(8,5)$; $(8,8)$.
Distribution: $\bar{x}=2, 3.5, 5, 6.5, 8$ with $P=\frac{1}{9},\frac{2}{9},\frac{3}{9},\frac{2}{9},\frac{1}{9}$.	Group equal means; the nine probabilities sum to 1.
$E(\bar{X})=\frac{2(1)+3.5(2)+5(3)+6.5(2)+8(1)}{9}=\frac{45}{9}=5$.	The mean of the sampling distribution.
$E(\bar{X}^2)=\frac{2^2(1)+3.5^2(2)+5^2(3)+6.5^2(2)+8^2(1)}{9}=\frac{25}{9}$.	Needed for the variance.
$Var(\bar{X})=28-5^2=3$, so $SD(\bar{X})=\sqrt{3}\approx 1.732$.	Using $Var(\bar{X})=E(\bar{X}^2)-[E(\bar{X})]^2$.
$E(\bar{X})=5=\mu$ and $\frac{\sigma}{\sqrt{2}}=\frac{\sqrt{6}}{\sqrt{2}}=\sqrt{3}=SD(\bar{X})$.	Both results agree with the formulae.

Example 3 — unscaffolded

A large squad of swimmers has lap times, in seconds, with mean $\mu=84$ and standard deviation $\sigma=21$. Random samples of lap times are taken and the sample mean $\bar{X}$ recorded. Describe how the dot plot of sample means for samples of size $n=4$ compares with the one for $n=36$, giving the centre and spread in each case.

Both dot plots are centred on the population mean, because $E(\bar{X})=\mu=84$ for every sample size. Their spreads are

$$SD(\bar{X})=\frac{\sigma}{\sqrt{n}}=\frac{21}{\sqrt{4}}=\frac{21}{2}=10.5\,(n=4),\ \frac{21}{\sqrt{36}}=\frac{21}{6}=3.5\,(n=36).$$

So increasing the sample size from 4 to 36 (a factor of 9) leaves the centre at 84 but divides the spread by $\sqrt{9}=3$, from 10.5 down to 3.5. The $n=36$ dot plot is therefore much narrower and more sharply bell-shaped, clustering the sample means tightly around 84.

$\therefore$ both dot plots are centred at 84; the spread is 10.5 for $n=4$ and 3.5 for $n=36$, so the larger sample gives a narrower, more concentrated distribution.

Example 4 — unscaffolded

A population has mean 90 and standard deviation $\sigma=26$. (a) Find $SD(\bar{X})$ for samples of size $n=4$. (b) Find the sample size needed for the sampling distribution of $\bar{X}$ to have standard deviation 2.

For part (a), the spread of the sampling distribution is

$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{26}{\sqrt{4}} = \frac{26}{2} = 13.$$

For part (b), set the spread equal to 2 and solve for n:

$$\frac{26}{\sqrt{n}} = 2 \Rightarrow \sqrt{n} = \frac{26}{2} = 13 \Rightarrow n = 13^2 = 169.$$

$\therefore SD(\bar{X}) = 13$ when $n = 4$, and a sample of size $n = 169$ is needed to reduce it to 2.

Practice questions

Scaffolded

Q1. A box contains four tokens numbered 2, 4, 6 and 8; one is drawn, recorded and returned. (a) Find the population mean μ. (b) Find the population standard deviation σ. (c) A sample of size 3 gives 4, 8, 4; find $\bar{x}$. (d) State which of μ, σ, $\bar{x}$ are parameters and which is a statistic.

Q2. For a population of three equally likely values 1, 4 and 7 (drawn with replacement), a sample of size $n = 2$ is taken. (a) List all equally likely samples of size 2 and their means. (b) Tabulate the sampling distribution of $\bar{X}$. (c) Find $E(\bar{X})$ and $SD(\bar{X})$. (d) Verify that $SD(\bar{X}) = \frac{\sigma}{\sqrt{2}}$, given $\sigma = \sqrt{6}$.

Q3. A population has standard deviation $\sigma = 28$. For the sampling distribution of $\bar{X}$, find $SD(\bar{X})$ when (a) $n = 4$; (b) $n = 49$; (c) $n = 196$; (d) By what factor does the spread change as n increases from 4 to 196?

Q4. Two dot plots of sample means come from the **same** population (standard deviation $\sigma = 30$): one uses $n = 4$, the other uses $n = 25$. (a) Find the spread $SD(\bar{X})$ of each. (b) How many times as wide as the $n = 25$ plot is the $n = 4$ plot? (c) Explain why changing n leaves the centre of the plot where it is. (d) Which dot plot is more bell-shaped?

Q5. A simulation records the means of eight samples from a population:

$$4.6, 5.1, 4.8, 5.3, 4.7, 5.0, 5.2, 4.9.$$

(a) Find the mean of these eight sample means. (b) Use it to estimate the population mean μ. (c) Explain why this is only an estimate. (d) State what happens to the spread of the sample means if the sample size is increased.

Q6. A population has mean 50 and standard deviation $\sigma = 33$. (a) Find $SD(\bar{X})$ for $n = 9$. (b) Find $SD(\bar{X})$ for $n = 36$. (c) Describe how the dot plot of sample means changes as n goes from 9 to 36. (d) Does the centre of the dot plot depend on n? Explain.

Unscaffolded

Q7. A spinner has two equally likely outcomes, 5 and 11, and is spun three times (with replacement). Find the sampling distribution of $\bar{X}$, then find $E(\bar{X})$ and $SD(\bar{X})$, and confirm $SD(\bar{X}) = \frac{\sigma}{\sqrt{3}}$ (here $\sigma = 3$).

Q8. A population has standard deviation $\sigma = 18$. Find $SD(\bar{X})$ for $n = 9$ and for $n = 81$, and state the factor by which the spread changes.

Q9. A population has standard deviation $\sigma = 32$. Find the sample size n for which the sampling distribution of $\bar{X}$ has standard deviation 4.

Q10. A population has standard deviation $\sigma = 30$. Find the sample size n for which $SD(\bar{X}) = 2.5$.

Q11. A population has standard deviation $\sigma = 8$. Dot plots of sample means are made for $n = 5$ and $n = 20$. (a) Find $SD(\bar{X})$ for each (to 3 decimal places). (b) By what factor does the spread change? (c) Which dot plot is more tightly clustered about the population mean?

Q12. A set of five cards is numbered $1, 2, 3, 4, 5$; a card is drawn, recorded and returned. (a) Find μ and σ. (b) A sample of size 4 gives $2, 5, 3, 2$; find $\acute{x}$. (c) State what $\acute{x}$ estimates, and whether it is a parameter or a statistic.

Q13. For the five cards of Q12, a sample of size $n=2$ is taken with replacement. (a) Give the sampling distribution of $\acute{X}$ as a table of values and probabilities. (b) Find $E(\acute{X})$ and $SD(\acute{X})$. (c) Confirm $SD(\acute{X})=\frac{\sigma}{\sqrt{2}}$, given $\sigma=\sqrt{2}$.

Q14. Explain, in terms of μ and $\frac{\sigma}{\sqrt{n}}$, what happens to the centre and the spread of the sampling distribution of $\acute{X}$ as the sample size n increases. Illustrate with a population of standard deviation $\sigma=35$ by comparing $n=1$ with $n=25$.

Q15. A population has standard deviation $\sigma=63$. Find the smallest sample size n for which the sampling distribution of $\acute{X}$ has standard deviation at most 9.

Q16. A dot plot of the means of many samples of size $n=16$ is centred at 4.8 and has spread 0.75. (a) Estimate the population mean μ. (b) Using $SD(\acute{X})=\frac{\sigma}{\sqrt{n}}$, estimate the population standard deviation σ. (c) Predict the spread of the dot plot if the sample size is increased to $n=64$.

Q17. A population has standard deviation $\sigma=48$. Copy and complete the table of $SD(\acute{X})=\frac{\sigma}{\sqrt{n}}$, then describe the trend.

n	1	4	9	16	36
$SD(\acute{X})$					

Q18. Return to the three cards showing 2, 5 and 8, drawn with replacement ($\mu=5$, $\sigma=\sqrt{6}$). (a) State μ and σ. (b) Find $SD(\acute{X})$ for $n=2$ and for $n=8$. (c) By what factor does the spread change from $n=2$ to $n=8$? (d) Describe how the dot plot of sample means changes from $n=2$ to $n=8$.

Answers

Q1. (a) $\mu=5$; (b) $\sigma=\sqrt{5}\approx 2.236$; (c) $\bar{x}=\frac{16}{3}\approx 5.33$; (d) μ,σ are parameters, $\bar{x}$ is a statistic. **Q2.** (b) $\bar{x}=1,2.5,4,5.5,7$ with $P=1/9,2/9,3/9,2/9,1/9$; (c) $E(\bar{X})=4$, $SD(\bar{X})=\sqrt{3}\approx 1.732$; (d) $\sqrt{6}/\sqrt{2}=\sqrt{3}$. **Q3.** (a) 14; (b) 4; (c) 2; (d) a factor of $1/7$. **Q4.** (a) 15 and 6; (b) 2.5 times as wide; (c) the centre is $E(\bar{X})=\mu$, which does not involve n; (d) the $n=25$ plot. **Q5.** (a) 4.95; (b) $\mu\approx 4.95$; (c) a different set of samples gives different means; (d) it decreases (as $\sigma/\sqrt{n}$). **Q6.** (a) 11; (b) 5.5; (c) same centre (50), spread halves, more bell-shaped; (d) no — the centre is μ for every n. **Q7.** $\bar{x}=5,7,9,11$ with $P=1/8,3/8,3/8,1/8$; $E(\bar{X})=8$, $SD(\bar{X})=\sqrt{3}\approx 1.732=3/\sqrt{3}$. **Q8.** 6 and 2; a factor of $1/3$. **Q9.** $n=64$. **Q10.** $n=144$. **Q11.** (a) 3.578 and 1.789; (b) a factor of $1/2$; (c) the $n=20$ plot. **Q12.** (a) $\mu=3$, $\sigma=\sqrt{2}\approx 1.414$; (b) $\bar{x}=3$; (c) a statistic; it estimates μ. **Q13.** (a) $\bar{x}=1,1.5,2,2.5,3,3.5,4,4.5,5$ with frequencies $1,2,3,4,5,4,3,2,1$ (out of 25); (b) $E(\bar{X})=3$, $SD(\bar{X})=1$; (c) $\sqrt{2}/\sqrt{2}=1$. **Q14.** Centre stays at μ; spread decreases as $\sigma/\sqrt{n}$. For $\sigma=35$: $n=1$ gives 35, $n=25$ gives 7. **Q15.** $n=49$. **Q16.** (a) $\mu\approx 4.8$; (b) $\sigma\approx 3$; (c) 0.375. **Q17.** $48,24,16,12,8$; the spread decreases as n increases. **Q18.** (a) $\mu=5$, $\sigma=\sqrt{6}$; (b) $\sqrt{3}\approx 1.732$ and $\frac{\sqrt{6}}{\sqrt{8}}\approx 0.866$; (c) a factor of $1/2$; (d) same centre (5), spread halved, more bell-shaped.

Fully-worked solutions

Q1. (a) $\mu=\frac{2+4+6+8}{4}=\frac{20}{4}=5$. (b) The squared deviations are $(2-5)^2=9$, $(4-5)^2=1$, $(6-5)^2=1$, $(8-5)^2=9$, so $\sigma=\sqrt{\frac{9+1+1+9}{4}}=\sqrt{\frac{20}{4}}=\sqrt{5}\approx 2.236$. (c) $\bar{x}=\frac{4+8+4}{3}=\frac{16}{3}\approx 5.33$. (d) μ and σ describe the whole population, so they are parameters; $\bar{x}$ is computed from a sample, so it is a statistic, and it estimates μ.

Q2. (a) The 9 ordered samples give means 1; $2.5,2.5$; $4,4,4$; $5.5,5.5$; 7, from $(1,1)$; $(1,4),(4,1)$; $(1,7),(4,4),(7,1)$; $(4,7),(7,4)$; $(7,7)$. (b) Collecting equal means, $\bar{x}=1,2.5,4,5.5,7$ with probabilities $1/9,2/9,3/9,2/9,1/9$. (c) $E(\bar{X})=\frac{1(1)+2.5(2)+4(3)+5.5(2)+7(1)}{9}=\frac{36}{9}=4$, and $E(\bar{X}^2)=\frac{1+12.5+48+60.5+49}{9}=\frac{171}{9}=19$, so $Var(\bar{X})=19-16=3$ and $SD(\bar{X})=\sqrt{3}\approx 1.732$. (d) $\frac{\sigma}{\sqrt{2}}=\frac{\sqrt{6}}{\sqrt{2}}=\sqrt{3}$, which matches.

Q3. The spread is $\frac{\sigma}{\sqrt{n}}=\frac{28}{\sqrt{n}}$. (a) $\frac{28}{2}=14$. (b) $\frac{28}{7}=4$. (c) $\frac{28}{14}=2$. (d) The spread falls from 14 to 2, a factor of $1/7$ — matching $\frac{1}{\sqrt{196/4}}=\frac{1}{7}$.

Q4. (a) $SD(\bar{X})=\frac{30}{\sqrt{4}}=15$ for $n=4$ and $\frac{30}{\sqrt{25}}=6$ for $n=25$. (b) The $n=4$ plot is $\frac{15}{6}=2.5$ times as wide as the $n=25$ plot. (c) The centre is $E(\bar{X})=\mu$, an expression that does not involve n, so altering the sample size cannot move it. (d) The larger sample, $n=25$, gives the more bell-shaped dot plot.

Q5. (a) The mean is $\frac{4.6+5.1+4.8+5.3+4.7+5.0+5.2+4.9}{8}=\frac{39.6}{8}=4.95$. (b) Because the centre of a sampling distribution estimates μ, we estimate $\mu\approx 4.95$. (c) It is only an estimate because the sample means vary: a different set of samples would give a slightly different average. (d) A larger sample size reduces the spread of the sample means, since $SD(\bar{X})=\frac{\sigma}{\sqrt{n}}$ decreases as n grows.

Q6. (a) $SD(\acute{X})=\frac{33}{\sqrt{9}}=\frac{33}{3}=11$. (b) $\frac{33}{\sqrt{36}}=\frac{33}{6}=5.5$. (c) The centre stays at the population mean 50, while the spread halves from 11 to 5.5, so the dot plot becomes narrower and more bell-shaped. (d) No: the centre is $E(\acute{X})=\mu$ for every sample size, so it does not depend on n.

Q7. With outcomes 5 and 11, $\mu=8$ and $\sigma=\sqrt{\frac{9+9}{2}}=3$. Over $2^3=8$ equally likely samples of size 3, the number of 11 s is $0,1,2,3$ with frequencies $1,3,3,1$, giving means $5,7,9,11$. So $\acute{x}=5,7,9,11$ with $P=1/8,3/8,3/8,1/8$. Then $E(\acute{X})=\frac{5+21+27+11}{8}=\frac{64}{8}=8$ and $E(\acute{X}^2)=\frac{25+147+243+121}{8}=\frac{536}{8}=67$, so $Var(\acute{X})=67-64=3$ and $SD(\acute{X})=\sqrt{3}\approx1.732$. This equals $\frac{\sigma}{\sqrt{3}}=\frac{3}{\sqrt{3}}=\sqrt{3}$.

Q8. $SD(\acute{X})=\frac{18}{\sqrt{9}}=6$ for $n=9$ and $\frac{18}{\sqrt{81}}=2$ for $n=81$. The spread changes by a factor of $1/3$ (matching $\frac{1}{\sqrt{81/9}}=\frac{1}{3}$).

Q9. Require $\frac{32}{\sqrt{n}}=4$, so $\sqrt{n}=\frac{32}{4}=8$ and $n=64$.

Q10. Require $\frac{30}{\sqrt{n}}=2.5$, so $\sqrt{n}=\frac{30}{2.5}=12$ and $n=12^2=144$.

Q11. (a) $SD(\acute{X})=\frac{8}{\sqrt{5}}\approx3.578$ for $n=5$ and $\frac{8}{\sqrt{20}}\approx1.789$ for $n=20$. (b) The spread changes by the factor $\frac{1}{\sqrt{20/5}}=\frac{1}{\sqrt{4}}=\frac{1}{2}$. (c) The $n=20$ dot plot is more tightly clustered about the mean, since its spread is smaller.

Q12. (a) $\mu=\frac{1+2+3+4+5}{5}=3$, and with squared deviations $4,1,0,1,4$, $\sigma=\sqrt{\frac{10}{5}}=\sqrt{2}\approx1.414$. (b) $\acute{x}=\frac{2+5+3+2}{4}=\frac{12}{4}=3$. (c) $\acute{x}$ is a sample statistic; it estimates the population mean $\mu=3$.

Q13. (a) For $n=2$ there are $5^2=25$ equally likely samples; the sample mean equals $\frac{\text{sum}}{2}$, where the sum runs from 2 to 10. The sums occur with frequencies $1,2,3,4,5,4,3,2,1$, so

$$\acute{x}=1,1.5,2,2.5,3,3.5,4,4.5,5, P=1/25,2/25,3/25,4/25,5/25,4/25,3/25,2/25,1/25.$$

(b) By symmetry $E(\acute{X})=3$, and $Var(\acute{X})=\frac{\sigma^2}{n}=\frac{2}{2}=1$, so $SD(\acute{X})=1$. (c) $\frac{\sigma}{\sqrt{2}}=\frac{\sqrt{2}}{\sqrt{2}}=1$, which matches.

Q14. The centre of the sampling distribution is $E(\acute{X})=\mu$, which does **not** depend on n, so it stays fixed as n increases. The spread is $SD(\acute{X})=\frac{\sigma}{\sqrt{n}}$, which **decreases** as n grows, because averaging more observations cancels the extremes. For $\sigma=35$: at $n=1$ the spread is $\frac{35}{\sqrt{1}}=35$ (the population's own spread), while at $n=25$ it is $\frac{35}{\sqrt{25}}=7$ — five times smaller, with the same centre.

Q15. Require $\frac{63}{\sqrt{n}}\le9$, so $\sqrt{n}\ge\frac{63}{9}=7$ and $n\ge49$. The smallest such sample size is $n=49$ (giving spread exactly 9).

Q16. (a) The centre of a sampling distribution estimates the population mean, so $\mu \approx 4.8$. (b) The spread equals $\frac{\sigma}{\sqrt{n}}$, so $\frac{\sigma}{\sqrt{16}} = 0.75$ gives $\sigma \approx 0.75 \times 4 = 3$. (c) For $n = 64$, the spread is $\frac{\sigma}{\sqrt{64}} = \frac{3}{8} = 0.375$.

Q17. With $SD(\acute{X}) = \frac{48}{\sqrt{n}}$: for $n = 1, 4, 9, 16, 36$ the values are $48, 24, 16, 12, 8$ respectively. The spread decreases as n increases (each time n is multiplied by a square, the spread is divided by that square root), while the centre stays at μ.

n	1	4	9	16	36
$SD(\acute{X})$	48	24	16	12	8

Q18. (a) $\mu = 5$ and $\sigma = \sqrt{6} \approx 2.449$. (b) $SD(\acute{X}) = \frac{\sqrt{6}}{\sqrt{2}} = \sqrt{3} \approx 1.732$ for $n = 2$, and $\frac{\sqrt{6}}{\sqrt{8}} = \sqrt{\frac{6}{8}} = \sqrt{0.75} \approx 0.866$ for $n = 8$. (c) The spread changes by the factor $\frac{1}{\sqrt{8/2}} = \frac{1}{\sqrt{4}} = \frac{1}{2}$. (d) The dot plot keeps the same centre 5 but its spread halves, so it becomes narrower and more bell-shaped, clustering the sample means more tightly about 5.