Specialist Mathematics Units 3 & 4

Chapter 11 — Linear combinations of random variables and the sample mean

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.

Significant effort has been made to ensure this content is legal. Please send any areas of concern to admin@vwx.au with appropriate document/legal references so errors can be verified and changes be made.

Contents

Chapter 11 Linear combinations of random variables and the sample mean — 11A Linear functions of a random variable

Theory

A **random variable** X assigns a numerical value to each outcome of a random experiment. Two numbers summarise its behaviour: the **mean** (or expected value) $\mu = E(X)$, which locates the centre of the distribution, and the **variance** $\sigma^2 = Var(X)$, which measures 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 $p_i = P(X = x_i)$ and the second form for the variance is usually quicker to use.

This section studies what happens to the mean and variance when a random variable is transformed by a **linear function**

$$Y = aX + b,$$

for real constants a and b. Such a change of variable occurs whenever a measurement is rescaled or shifted: converting units, adding a fixed charge, or applying a per-unit rate.

Mean of a linear function.

Replacing each value x_i by $a x_i + b$ and using the distribution of X,

$$E(aX + b) = \sum_i (a x_i + b) p_i = a \sum_i x_i p_i + b \sum_i p_i.$$

The first sum is μ and the second is 1, so

$$E(aX + b) = a\mu + b.$$

The mean simply follows the transformation: scale it by a and shift it by b.

Variance and standard deviation of a linear function.

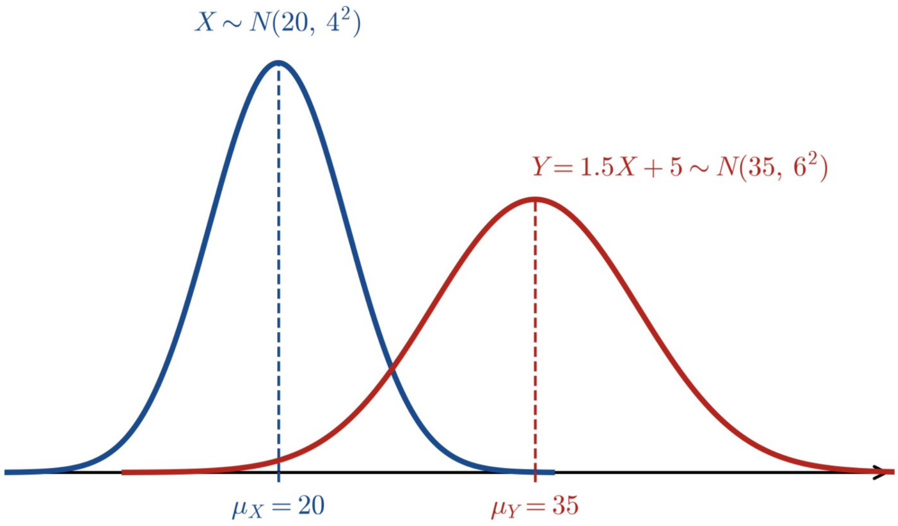
Using $Var(Y) = E\left[(Y - E(Y))^2\right]$ with $Y = aX + b$ and $E(Y) = a\mu + b$,

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

The additive constant b cancels, so

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

Two features are worth noting. A **shift** b moves every value by the same amount, so it changes the centre but leaves the spread untouched — the variance does not depend on b. A **scaling** a stretches all deviations from the mean by a factor a, so the variance is multiplied by a^2 and the standard deviation by $|a|$. The modulus appears because a standard deviation is never negative: multiplying by a negative a reflects the distribution but does not give it a negative spread. The diagram shows $X \sim N(20, 4^2)$ and its image $Y = 1.5X + 5$: the centre moves from 20 to 35, and the spread grows because $|a| = 1.5 > 1$.



Standardising a random variable.

A particularly useful linear function turns X into a variable with mean 0 and standard deviation 1. The **standardised variable** is

$$Z = \frac{X-\mu}{\sigma} = \frac{1}{\sigma}X - \frac{\mu}{\sigma},$$

which is linear with $a = \frac{1}{\sigma}$ and $b = -\frac{\mu}{\sigma}$. Applying the two rules,

$$E(Z) = \frac{1}{\sigma}\mu - \frac{\mu}{\sigma} = 0, Var(Z) = \left(\frac{1}{\sigma}\right)^2 \sigma^2 = 1.$$

So **every** random variable, whatever its mean and standard deviation, can be recentred and rescaled to one with mean 0 and variance 1. The value of Z is the **number of standard deviations** that X lies above (positive) or below (negative) its mean; it is often called the z-**score**.

Linear functions of a normal random variable.

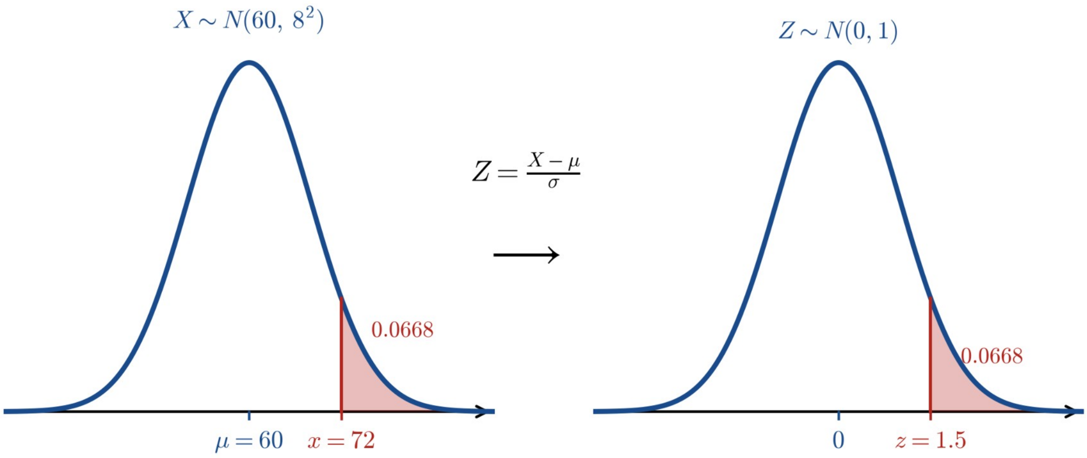
If X is **normally distributed**, $X \sim N(\mu, \sigma^2)$, then any linear function of X is also normal:

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

In particular the standardised variable is the **standard normal** variable, $Z \sim N(0,1)$. This is what makes normal probabilities easy to find: a probability about X is converted into one about Z by standardising the boundary,

$$P(X \le x) = P\left(Z \le \frac{x-\mu}{\sigma}\right),$$

and the required value of $P(Z \le z)$ is read from technology (or a standard normal table). The picture below standardises $X \sim N(60, 8^2)$ at $x = 72$: the boundary maps to $z = \frac{72-60}{8} = 1.5$, and the shaded upper-tail areas match, both equal to 0.0668.



Applications.

Linear functions arise whenever a random quantity is measured on a rescaled or shifted axis.

- **Unit conversion.** Temperature in degrees Fahrenheit is $F = \frac{9}{5}C + 32$, where C is the Celsius temperature. If C has mean μ and standard deviation σ, then $E(F) = \frac{9}{5}\mu + 32$ and $sd(F) = \frac{9}{5}\sigma$.
- **A fixed charge plus a per-unit rate.** A cost of the form $\text{cost} = b + aX$ (a fixed amount b plus a dollars per unit of X) has mean $a\mu + b$ and standard deviation $|a|\sigma$.
- **Rescaling a measurement.** Changing units (centimetres to metres, grams to kilograms) or applying a scale factor multiplies both the mean and the standard deviation by that factor.

In every case the mean transforms fully ($a\mu + b$) while the standard deviation responds only to the scaling ($|a|\sigma$), and — when the original variable is normal — the transformed variable is normal too.

Worked examples

Example 1 — scaffolded

A carnival spinner shows a score X with the probability distribution below. A prize of $Y = 5X + 2$ dollars is paid. Find $E(X)$ and $Var(X)$, then $E(Y)$, $Var(Y)$ and $sd(Y)$.

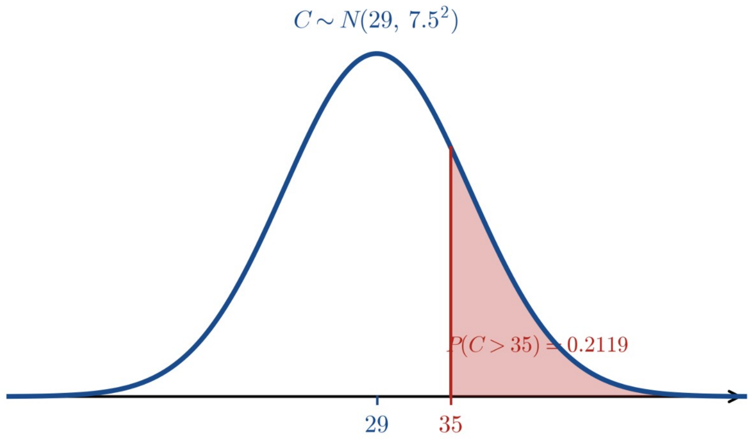
x	0	1	2	3
$P(X = x)$	0.4	0.3	0.2	0.1

Working	Comment
$E(X) = 0(0.4) + 1(0.3) + 2(0.2) + 3(0.1) = 1.$	Multiply each value by its probability and add.
$E(X^2) = 0^2(0.4) + 1^2(0.3) + 2^2(0.2) + 3^2(0.1) = 0 + 0.3$.	The variance also needs $E(X^2)$.
$Var(X) = E(X^2) - [E(X)]^2 = 2 - 1^2 = 1$, so $\sigma = 1$.	Use $Var(X) = E(X^2) - \mu^2$.
$E(Y) = E(5X + 2) = 5\mu + 2 = 5(1) + 2 = 7.$	Use $E(aX + b) = a\mu + b$ with $a = 5, b = 2$.
$Var(Y) = 5^2 Var(X) = 25(1) = 25.$	Use $Var(aX + b) = a^2\sigma^2$; the $+2$ has no effect.
$sd(Y) = \lvert 5 \rvert \sigma = 5(1) = 5.$	Use $sd(aX + b) = \lvert a \rvert \sigma$.

Example 2 — scaffolded

A courier's parcels have mass X kilograms, where $X \sim N(8, 2.5^2)$. The shipping cost is $C = 3X + 5$ dollars (a \$5 handling fee plus \$3 per kilogram). (a) Find the mean and standard deviation of C. (b) Find $P(C > 35)$.

Working	Comment
$E(C)=3E(X)+5=3(8)+5=29$.	Mean of $aX+b$ with $a=3, b=5$.
$sd(C)=\lvert 3\rvert\sigma=3(2.5)=7.5$, so $Var(C)=7.5^2=56.25$.	Only the scaling $a=3$ affects the spread.
X is normal, so $C=3X+5$ is normal: $C\sim N(29, 7.5^2)$.	A linear function of a normal variable is normal.
Standardise the boundary: $z=\dfrac{35-29}{7.5}=0.8$.	$z=\dfrac{c-\mu_C}{\sigma_C}$.
$P(C>35)=P(Z>0.8)=1-P(Z\le 0.8)=1-0.7881=$.	From technology, $P(Z\le 0.8)=0.7881$.



Example 3 — unscaffolded

Scores on an aptitude test are modelled by $X\sim N(60, 8^2)$. Write down the standardised variable Z, confirm that $E(Z)=0$ and $Var(Z)=1$, and find the probability that a randomly chosen score exceeds 72.

Standardising gives

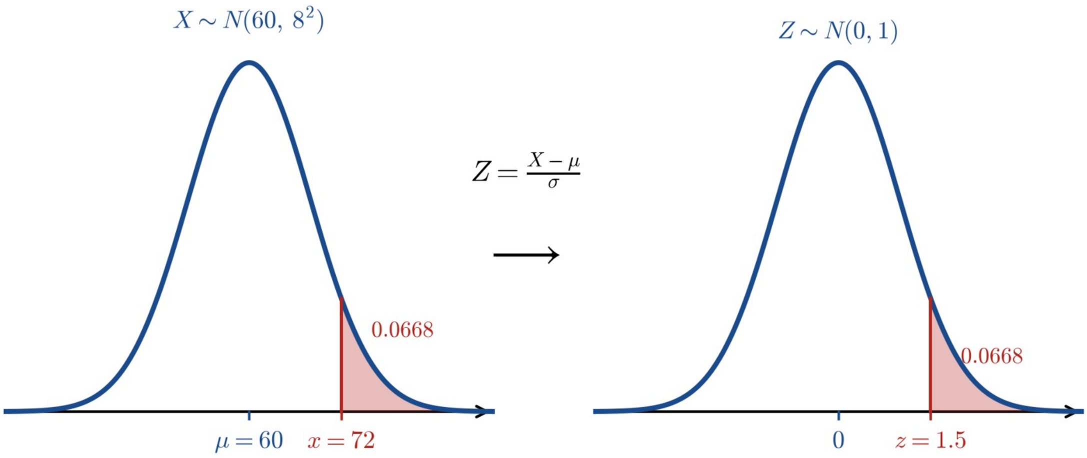
$$Z=\frac{X-60}{8}=\frac{1}{8}X-\frac{60}{8},$$

a linear function with $a=\frac{1}{8}$ and $b=-\frac{60}{8}=-7.5$. Applying the linear rules,

$$E(Z)=\frac{1}{8}(60)-7.5=7.5-7.5=0, Var(Z)=\left(\frac{1}{8}\right)^2(8^2)=\frac{64}{64}=1.$$

Because X is normal, $Z\sim N(0,1)$. The score 72 standardises to $z=\frac{72-60}{8}=1.5$, so

$$P(X>72)=P(Z>1.5)=1-P(Z\le 1.5)=1-0.9332=0.0668.$$



$\therefore Z = \dfrac{X-60}{8}$ has mean 0 and variance 1, and $P(X>72) \approx 0.0668$ (about 6.7 % of scores).

Example 4 — unscaffolded

On summer days the maximum temperature (in °C) at a town is modelled by $X \sim N(24, 4^2)$. Temperatures are reported in °F using $F = \frac{9}{5}X + 32$. Find the mean and standard deviation of F, and the probability that the reported maximum exceeds 82.4°F.

Here $a = \frac{9}{5} = 1.8$ and $b = 32$, so

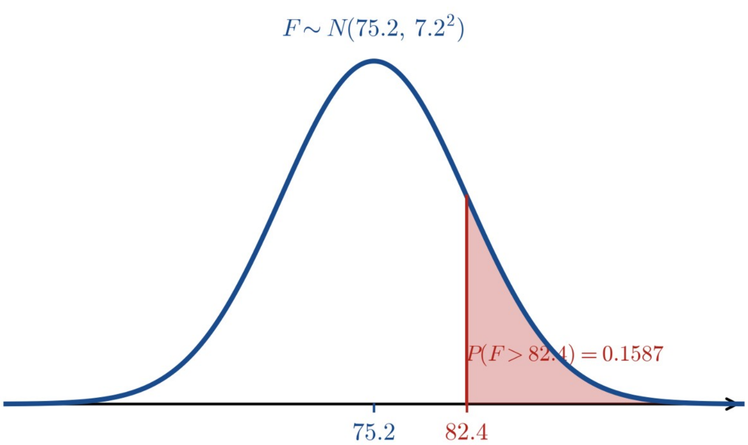
$$E(F) = 1.8(24) + 32 = 43.2 + 32 = 75.2,\ sd(F) = |1.8|(4) = 7.2.$$

Since X is normal, F is normal: $F \sim N(75.2, 7.2^2)$. Standardising the boundary,

$$z = \frac{82.4 - 75.2}{7.2} = \frac{7.2}{7.2} = 1,$$

so

$$P(F > 82.4) = P(Z > 1) = 1 - P(Z \le 1) = 1 - 0.8413 = 0.1587.$$



$\therefore$ the reported maximum has mean 75.2°F and standard deviation 7.2°F, and $P(F > 82.4) \approx 0.1587$.

Practice questions

Scaffolded

Q1. A discrete random variable X has the distribution below, and $Y = 10X - 5$.

x	1	2	3	4
$P(X=x)$	0.1	0.2	0.3	0.4

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

Q2. A machine cuts rods whose length X mm has mean 50 and standard deviation 6. Each rod is then coated, adding exactly 10 mm, so the finished length is $Y = X + 10$. (a) Find $E(Y)$. (b) Find $Var(Y)$. (c) Find $sd(Y)$. (d) Explain why the standard deviation of Y equals that of X.

Q3. A variable X has mean 80 and standard deviation 12, and is standardised using $Z = \frac{X-80}{12}$. (a) Write Z in the form $aX + b$, stating a and b. (b) Find $E(Z)$. (c) Find $Var(Z)$. (d) Find the z-score of $X = 98$.

Q4. Examination marks are modelled by $X \sim N(500, 50^2)$. (a) State μ and σ. (b) Show that $X < 560$ standardises to $Z < 1.2$. (c) Find $P(Z < 1.2)$ from technology, and hence $P(X < 560)$. (d) Find the probability that a mark exceeds 560.

Q5. Weekly electricity use X kWh is modelled by $X \sim N(30, 5^2)$. The bill is $B = 4X + 20$ dollars. (a) Find $E(B)$. (b) Find $sd(B)$. (c) Explain why B is normal, and state its distribution. (d) Find $P(B > 150)$.

Q6. The midday temperature (°C) is modelled by $X \sim N(20, 5^2)$; in °F it is $F = 1.8X + 32$. (a) Find $E(F)$. (b) Find $sd(F)$. (c) State the distribution of F. (d) Find $P(F > 86)$.

Unscaffolded

Q7. A discrete random variable X has the distribution below, and $Y = 3X - 1$. Find $E(Y)$, $Var(Y)$ and $sd(Y)$.

x	2	4	6
$P(X=x)$	0.25	0.5	0.25

Q8. A random variable X has mean 15 and variance 9. Find the mean, variance and standard deviation of $U = X + 7$ and of $V = X - 7$, and comment on the spread of each.

Q9. A random variable X has mean 10 and standard deviation 4. Find the mean, variance and standard deviation of $Y = -2X + 3$.

Q10. Adult heights (cm) are modelled by $X \sim N(170, 8^2)$. Find (a) $P(X > 182)$ and (b) $P(162 < X < 178)$.

Q11. Scores on a standardised test are modelled by $X \sim N(500, 100^2)$. Find the z-score of a score of 650 and of a score of 430, and state which score is more unusual.

Q12. A random variable X has mean 4 and standard deviation 2. A new variable $Y = aX + b$ (with $a > 0$) has mean 20 and standard deviation 6. Find a and b.

Q13. A car's fuel use X litres on a trip is modelled by $X \sim N(45, 5^2)$. The trip cost is $C = 1.8X + 15$ dollars. Find the mean and standard deviation of C, and $P(C > 114)$.

Q14. A random variable X has mean μ and standard deviation $\sigma > 0$. Using the rules for a linear function, show that $Z = \frac{X - \mu}{\sigma}$ has mean 0 and variance 1.

Q15. Daily temperatures reported in °F are modelled by $F \sim N(68, 9^2)$. The Celsius temperature is $C = \frac{F - 32}{1.8}$. Find the mean and standard deviation of C, and $P(C < 15)$.

Q16. Battery lifetimes (hours) are modelled by $X \sim N(40, 6^2)$. Find the lifetime that is exceeded by 90% of batteries.

Q17. A random variable $X \sim N(20, 4^2)$, and $Y = 50 - 2X$. Find the mean and standard deviation of Y, and $P(6 < Y < 18)$.

Q18. Raw test scores are modelled by $X \sim N(60, 5^2)$, and scaled scores are $Y = 1.5X + 10$. (a) Find the mean and standard deviation of Y. (b) Find $P(Y > 115)$. (c) Find the scaled score y that is exceeded by 95% of candidates.

Answers

Q1. $E(X)=3$, $Var(X)=1$; $E(Y)=25$, $Var(Y)=100$, $sd(Y)=10$. **Q2.** $E(Y)=60$, $Var(Y)=36$, $sd(Y)=6$; the shift moves the centre but not the spread. **Q3.** $a=\frac{1}{12}$, $b=-\frac{20}{3}$; $E(Z)=0$, $Var(Z)=1$; z-score of 98 is 1.5. **Q4.** $\mu=500$, $\sigma=50$; $z=1.2$; $P(X<560)=0.8849$; $P(X>560)=0.1151$. **Q5.** $E(B)=140$, $sd(B)=20$; $B\sim N(140, 20^2)$; $P(B>150)=0.3085$. **Q6.** $E(F)=68$, $sd(F)=9$; $F\sim N(68, 9^2)$; $P(F>86)=0.0228$. **Q7.** $E(Y)=11$, $Var(Y)=18$, $sd(Y)=3\sqrt{2}\approx 4.24$. **Q8.** U: mean 22, variance 9, sd 3. V: mean 8, variance 9, sd 3. Both have the same spread as X — a shift does not change the variance. **Q9.** $E(Y)=-17$, $Var(Y)=64$, $sd(Y)=8$. **Q10.** (a) 0.0668; (b) 0.6827. **Q11.** z-score of 650 is 1.5; of 430 is -0.7; 650 is more unusual (larger $|z|$). **Q12.** $a=3$, $b=8$. **Q13.** $E(C)=96$, $sd(C)=9$; $P(C>114)=0.0228$. **Q14.** $E(Z)=0$ and $Var(Z)=1$ (shown). **Q15.** $E(C)=20$, $sd(C)=5$; $P(C<15)=0.1587$. **Q16.** About 32.31 hours. **Q17.** $E(Y)=10$, $sd(Y)=8$; $P(6<Y<18)=0.5328$. **Q18.** (a) mean 100, sd 7.5; (b) 0.0228; (c) $y\approx 87.66$.

Fully-worked solutions

Q1. $E(X)=1(0.1)+2(0.2)+3(0.3)+4(0.4)=0.1+0.4+0.9+1.6=3$. $E(X^2)=1^2(0.1)+2^2(0.2)+3^2(0.3)+4^2(0.4)=0.1+0.8+2.7+6.4=10$, so $Var(X)=E(X^2)-[E(X)]^2=10-3^2=1$. With $Y=10X-5$ ($a=10, b=-5$), $E(Y)=10(3)-5=25$, $Var(Y)=10^2(1)=100$ and $sd(Y)=|10|\sqrt{1}=10$.

Q2. With $Y=X+10$ we have $a=1$ and $b=10$. Then $E(Y)=1(50)+10=60$, $Var(Y)=1^2 Var(X)=6^2=36$ and $sd(Y)=|1|(6)=6$. The standard deviation is unchanged because adding a constant shifts every length by the same 10 mm; the distances of values from the mean — and hence the spread — are not affected.

Q3. Writing $Z=\frac{X-80}{12}=\frac{1}{12}X-\frac{80}{12}$, so $a=\frac{1}{12}$ and $b=-\frac{80}{12}=-\frac{20}{3}$. Then $E(Z)=\frac{1}{12}(80)-\frac{20}{3}=\frac{80}{12}-\frac{80}{12}=0$ and $Var(Z)=\left(\frac{1}{12}\right)^2(12^2)=1$. The z-score of $X=98$ is $z=\frac{98-80}{12}=\frac{18}{12}=1.5$.

Q4. (a) $\mu=500$ and $\sigma=50$. (b) Standardising the boundary, $z=\frac{560-500}{50}=\frac{60}{50}=1.2$, so $X<560$ is equivalent to $Z<1.2$. (c) From technology $P(Z<1.2)=0.8849$, hence $P(X<560)=0.8849$. (d) The complement is $P(X>560)=1-0.8849=0.1151$.

Q5. With $B=4X+20$ ($a=4, b=20$), $E(B)=4(30)+20=140$ and $sd(B)=|4|(5)=20$. Because X is normal and B is a linear function of X, B is normal: $B\sim N(140, 20^2)$. Standardising, $z=\frac{150-140}{20}=0.5$, so $P(B>150)=P(Z>0.5)=1-0.6915=0.3085$.

Q6. With $F=1.8X+32$, $E(F)=1.8(20)+32=36+32=68$ and $sd(F)=|1.8|(5)=9$. As a linear function of the normal variable X, $F\sim N(68, 9^2)$. Standardising, $z=\frac{86-68}{9}=2$, so $P(F>86)=P(Z>2)=1-0.9772=0.0228$.

Q7. $E(X)=2(0.25)+4(0.5)+6(0.25)=0.5+2+1.5=4$. $E(X^2)=2^2(0.25)+4^2(0.5)+6^2(0.25)=1+8+9=18$, so $Var(X)=18-4^2=2$. With $Y=3X-1$ ($a=3, b=-1$), $E(Y)=3(4)-1=11$, $Var(Y)=3^2(2)=18$ and $sd(Y)=|3|\sqrt{2}=3\sqrt{2}\approx 4.24$.

Q8. For $U=X+7$: $E(U)=15+7=22$, $Var(U)=1^2(9)=9$, $sd(U)=3$. For $V=X-7$: $E(V)=15-7=8$, $Var(V)=1^2(9)=9$, $sd(V)=3$. Each is only a shift ($a=1$), so both keep the variance 9 and standard deviation 3 of X: the centre moves but the spread is identical for X, U and V.

Q9. With $Y=-2X+3$ ($a=-2, b=3$), $E(Y)=-2(10)+3=-17$, $Var(Y)=(-2)^2(4^2)=4(16)=64$ and $sd(Y)=|-2|(4)=8$. The negative scaling reflects the distribution but, through $|a|$ and a^2, keeps the spread positive.

Q10. Here $\mu=170$, $\sigma=8$. (a) $z=\frac{182-170}{8}=1.5$, so $P(X>182)=P(Z>1.5)=1-0.9332=0.0668$. (b) The endpoints standardise to $z=\frac{162-170}{8}=-1$ and $z=\frac{178-170}{8}=1$, so $P(162<X<178)=P(-1<Z<1)=0.8413-0.1587=0.6827$.

Q11. The z-score of 650 is $\frac{650-500}{100}=1.5$; the z-score of 430 is $\frac{430-500}{100}=-0.7$. Since $|1.5|>|-0.7|$, the score 650 lies further from the mean in standard-deviation units and is therefore the more unusual result.

Q12. The standard deviation depends only on the scaling: $sd(Y)=|a|\sigma$ gives $|a|(2)=6$, so $|a|=3$, and as $a>0$, $a=3$. The mean gives $E(Y)=a(4)+b=20$, so $3(4)+b=20$, hence $b=20-12=8$. Thus $Y=3X+8$.

Q13. With $C=1.8X+15$, $E(C)=1.8(45)+15=81+15=96$ and $sd(C)=|1.8|(5)=9$; as a linear function of the normal X, $C\sim N(96,9^2)$. Standardising, $z=\frac{114-96}{9}=2$, so $P(C>114)=P(Z>2)=1-0.9772=0.0228$.

Q14. Write $Z=\frac{X-\mu}{\sigma}=\frac{1}{\sigma}X-\frac{\mu}{\sigma}$, a linear function with $a=\frac{1}{\sigma}$ and $b=-\frac{\mu}{\sigma}$. Then $E(Z)=a\mu+b=\frac{1}{\sigma}\mu-\frac{\mu}{\sigma}=0$ and $Var(Z)=a^2\sigma^2=\left(\frac{1}{\sigma}\right)^2\sigma^2=1$. Hence any random variable standardises to one with mean 0 and variance 1.

Q15. Write $C=\frac{F-32}{1.8}=\frac{1}{1.8}F-\frac{32}{1.8}$, so $a=\frac{1}{1.8}=\frac{5}{9}$ and $b=-\frac{32}{1.8}$. Then $E(C)=\frac{68-32}{1.8}=\frac{36}{1.8}=20$ and $sd(C)=\frac{1}{1.8}(9)=5$; as a linear function of the normal F, $C\sim N(20,5^2)$. Standardising, $z=\frac{15-20}{5}=-1$, so $P(C<15)=P(Z<-1)=0.1587$.

Q16. "Exceeded by 90% of batteries" means $P(X>x)=0.90$, so $P(X<x)=0.10$ and x is the 10th percentile. The standard normal value with $P(Z<z)=0.10$ is $z=-1.2816$ (from technology). Then

$$x=\mu+z\sigma=40+(-1.2816)(6)=40-7.689=32.31.$$

So about 32.31 hours are exceeded by 90% of batteries.

Q17. With $Y=50-2X$ ($a=-2, b=50$), $E(Y)=-2(20)+50=10$ and $sd(Y)=|-2|(4)=8$; as a linear function of the normal X, $Y\sim N(10,8^2)$. The endpoints standardise to $z=\frac{6-10}{8}=-0.5$ and $z=\frac{18-10}{8}=1$, so

$$P(6<Y<18)=P(-0.5<Z<1)=0.8413-0.3085=0.5328.$$

Q18. (a) With $Y=1.5X+10$, $E(Y)=1.5(60)+10=100$ and $sd(Y)=|1.5|(5)=7.5$, so $Y\sim N(100,7.5^2)$. (b) Standardising, $z=\frac{115-100}{7.5}=2$, so $P(Y>115)=P(Z>2)=1-0.9772=0.0228$. (c) "Exceeded by 95% of candidates" means $P(Y>y)=0.95$, so $P(Y<y)=0.05$; the standard normal value with $P(Z<z)=0.05$ is $z=-1.6449$. Then

$$y=100+(-1.6449)(7.5)=100-12.336=87.66.$$

So a scaled score of about 87.66 is exceeded by 95% of candidates.

Theory

An earlier section studied a **linear function** $Y = aX + b$ of a *single* random variable X, and found $E(Y) = a\mu + b$ and $Var(Y) = a^2\sigma^2$. This section combines **two or more** random variables — for example a total mass built from several independent parts, or the difference between two independent measurements. Throughout, a random variable X has **mean** $\mu = E(X)$ and **variance** $\sigma^2 = Var(X)$, with **standard deviation** $\sigma = sd(X) = \sqrt{Var(X)}$.

Linearity of expectation.

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

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

In particular $E(X+Y) = E(X) + E(Y)$ and $E(X-Y) = E(X) - E(Y)$. A defining feature of the mean is that this rule holds **whether or not X and Y are independent** — expected values always add. Adding a constant c simply shifts the mean: $E(aX + bY + c) = aE(X) + bE(Y) + c$.

Variance of a linear combination of independent variables.

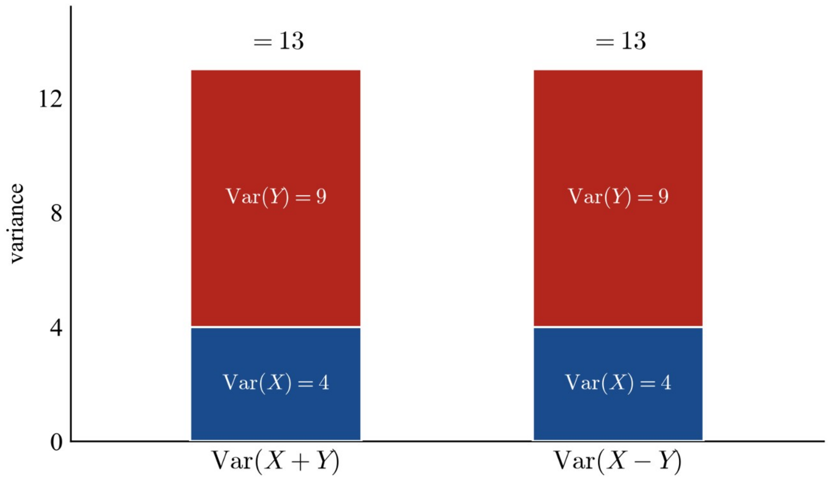
Variances behave differently. When X and Y are **independent**,

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y).$$

The coefficients enter **squared**, so each term is non-negative. Taking $a = b = 1$ and then $a = 1, b = -1$ gives the two most important special 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**, not subtracted. Uncertainty *accumulates* — combining two independent variables can only make the spread larger, whether you add them or subtract them, and $Var(X-Y) = Var(X+Y)$. Forgetting this and writing $Var(X) - Var(Y)$ for a difference is a common and costly error. The diagram shows the two variances stacking to the same total for both $X+Y$ and $X-Y$.



Because a constant has no spread, an added constant leaves the variance unchanged: $Var(aX + bY + c) = a^2 Var(X) + b^2 Var(Y)$. The standard deviation is the square root,

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

so **standard deviations do not add** — it is the variances that combine.

General linear combination.

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

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

which holds with no restriction. If in addition the X_i are **independent**, the variance is

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

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

Sums of identically distributed copies.

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

$$E\left(\sum_{i=1}^{n} X_i\right) = n\mu, Var\left(\sum_{i=1}^{n} X_i\right) = n\sigma^2,$$

so the sum $X_1 + X_2 + \ldots + X_n$ has standard deviation $\sqrt{n}\sigma$.

A key distinction: nX versus $X_1 + X_2 + \ldots + X_n$.

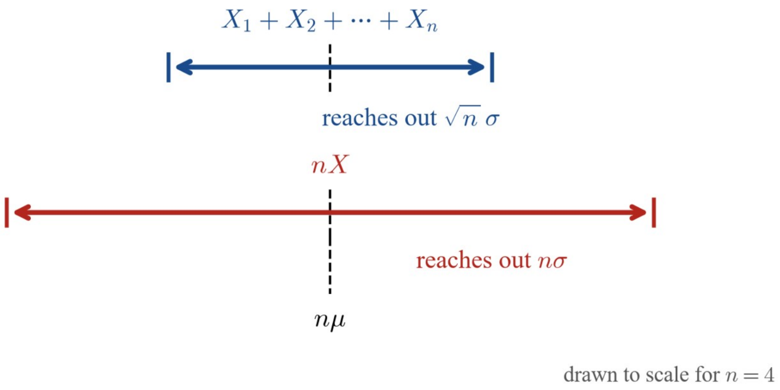
These two expressions are easy to confuse but describe genuinely different quantities.

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

Both have the same mean $n\mu$, but their spreads differ by a factor of n in the variance.

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

Intuitively, 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. Mistaking one for the other is a favourite examination trap. The diagram compares the two spreads about their common mean $n\mu$: the sum reaches out only $\sqrt{n}\sigma$, while nX reaches out $n\sigma$.



What these rules settle, and what they do not.

The rules above fix the **centre** and the **spread** of a linear combination, using nothing but the means and variances of the parts. They settle nothing about the **shape** of its distribution, so on their own they cannot deliver a probability such as $P(X+Y>30)$: two combinations can share a mean and a variance and still spread their probability quite differently. No question in this section therefore asks for a probability: every numerical answer is a mean, a variance or a standard deviation, or something read off from them — an unknown coefficient, a sample size, a ratio of spreads.

There is one important family for which the shape *is* settled — independent **normal** variables — and the probability calculations that unlocks are the subject of the next section.

Applications.

- **Totals of independent parts.** The combined mass, length or time built from several independent components has mean equal to the sum of the means, and variance equal to the sum of the variances.
- **Differences and comparisons.** To compare two independent processes — by how much one typically exceeds the other, and how variable that gap is — 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 or portfolio $a_1X_1+a_2X_2+\ldots$ 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	10	4
Y	6	9

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

Working	Comment
$E(X+Y)=E(X)+E(Y)=10+6=16$.	Means always add, independent or not.
$Var(X+Y)=Var(X)+Var(Y)=4+9=13$.	For independent variables the variances add.
$E(X-Y)=E(X)-E(Y)=10-6=4$.	The mean of a difference subtracts.
$Var(X-Y)=Var(X)+Var(Y)=4+9=13$.	Still a sum: the coefficient -1 squares to $+1$.

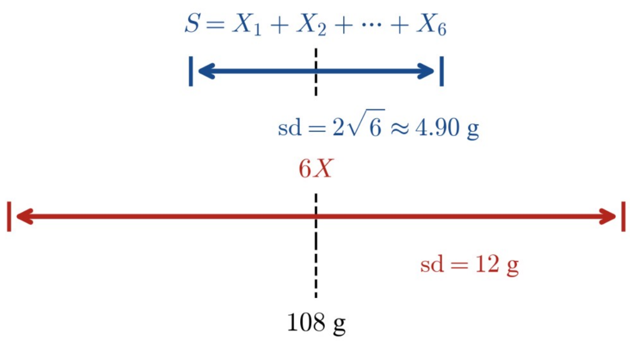
Working	Comment
$E(2X+3Y)=2E(X)+3E(Y)=2(10)+3(6)=38$.	Use $E(aX+bY)=aE(X)+bE(Y)$.
$Var(2X+3Y)=2^2 Var(X)+3^2 Var(Y)=4(4)+9($.	Square each coefficient before multiplying by the variance.

Notice that $X+Y$ and $X-Y$ have the **same** variance 13; only the means differ.

Example 2 — scaffolded

A café fills paper cups with coffee; the mass in one independently filled cup is X grams, with mean $\mu=18$ and standard deviation $\sigma=2$ (so $\sigma^2=4$). A tray holds $n=6$ independently filled cups. (a) Find the mean and standard deviation of the total mass $S=X_1+X_2+\ldots+X_6$. (b) Find the mean and standard deviation of $6X$ (six times the mass of one cup), and explain why the answers differ from part (a).

Working	Comment
$E(S)=6\mu=6(18)=108$ g.	Sum of n iid copies: mean $n\mu$.
$Var(S)=6\sigma^2=6(4)=24$, so $sd(S)=\sqrt{24}=2\sqrt{6}\approx 4.90$ g.	Variance of the sum is $n\sigma^2$; sd is $\sqrt{n}\sigma$.
$E(6X)=6\mu=6(18)=108$ g.	Scaling one cup by 6: mean 6μ, same as $E(S)$.
$Var(6X)=6^2\sigma^2=36(4)=144$, so $sd(6X)=6\sigma=12$ g.	Variance of nX is $n^2\sigma^2$; sd is $n\sigma$.
The means agree (108 g) but the spreads do not: $sd(S)\approx 4.90$ against $sd(6X)=12$.	Six *separate* cups partly average out; six times *one* cup magnifies its single deviation.



same mean, spreads in the ratio $\sqrt{6}:1$

Example 3 — unscaffolded

A café's daily sales are X coffees, with mean 240 and standard deviation 6, and Y cakes, with mean 60 and standard deviation 5; the two counts are independent. Each coffee returns 5 dollars and each cake 8 dollars, and the fixed daily cost is 600 dollars, so the daily profit in dollars is $P=5X+8Y-600$. Find the mean, variance and standard deviation of P.

The profit is a linear combination of two variables plus a constant, and the constant simply shifts the mean:

$$E(P)=5E(X)+8E(Y)-600=5(240)+8(60)-600=1200+480-600=1080.$$

The constant contributes no variability at all, and — because the counts are independent — each coefficient enters squared:

$$Var(P)=5^2 Var(X)+8^2 Var(Y)=25(6^2)+64(5^2)=25(36)+64(25)=900+1600=2500,$$

so $sd(P)=\sqrt{2500}=50$ dollars.

Notice that cakes contribute the larger share of the variance (1600 against 900) even though far fewer are sold and their sales are the less variable of the two: it is the coefficient 8, squared, that magnifies their day-to-day fluctuation.

$\therefore E(P)=1080$ dollars, $Var(P)=2500$ and $sd(P)=50$ dollars.

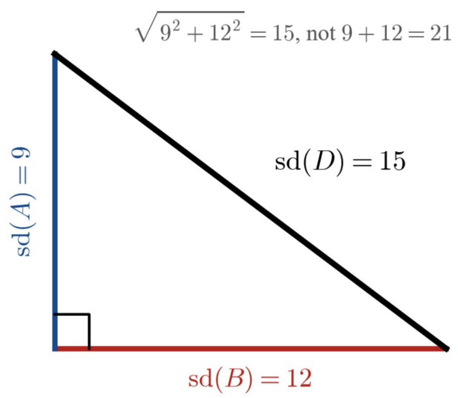
Example 4 — unscaffolded

Two independent couriers deliver a parcel. Courier A's time has mean 40 minutes and standard deviation 9 minutes; courier B's time has mean 58 minutes and standard deviation 12 minutes. Let $D=B-A$ be the amount by which courier B is slower. Find the mean, variance and standard deviation of D, and explain why the probability that courier B arrives first cannot be found from this information alone.

The mean of the difference subtracts, but — because A and B are independent — the variances **add**:

$$E(D)=58-40=18, Var(D)=12^2+9^2=144+81=225,$$

so $sd(D)=\sqrt{225}=15$ minutes. On average courier B takes 18 minutes longer, and that gap varies with standard deviation 15 minutes — a spread wider than either courier's own (12 and 9 minutes), because both sources of variability feed into it. The standard deviations combine like the perpendicular sides of a right-angled triangle, $\sqrt{9^2+12^2}=15$; they certainly do not add to 21.



Courier B arrives first exactly when $D<0$, a boundary that sits 1.2 standard deviations below the mean of D. But a mean and a standard deviation fix only the centre and the spread of D, not its shape, and different shapes place different amounts of probability below that boundary. Without a distributional model for A and B the probability cannot be evaluated.

$\therefore E(D)=18$ minutes, $Var(D)=225$ and $sd(D)=15$ minutes; the probability that B arrives first needs a distribution, not merely a mean and a variance.

Practice questions

Scaffolded

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

Variable	Mean	Variance
X	8	5
Y	3	2

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

(d) Find $sd(3X+2Y)$.

Q2. A pallet is loaded with $n=20$ independent bags of rice; one bag has mass X with mean 5 kg and standard deviation 0.2 kg (so $\sigma^2=0.04$). Let $S=X_1+\ldots+X_{20}$ be the total mass. (a) Find $E(S)$. (b) Find $Var(S)$. (c) Find $sd(S)$. (d) Find $Var(20X)$ and explain why it differs from $Var(S)$.

Q3. A batch of fruit punch is mixed from 2 litres of juice A and 3 litres of juice B, after which 50 grams of sugar is stirred in. Juice A carries X grams of sugar per litre, with mean 30 and standard deviation 4; juice B carries Y grams per litre, with mean 12 and standard deviation 5; and X and Y are independent. The total sugar in one batch is $S=2X+3Y+50$ grams. (a) Find $E(S)$. (b) Find $Var(S)$. (c) Find $sd(S)$. (d) Explain why the 50 grams of added sugar changes $E(S)$ but not $Var(S)$.

Q4. Two independent filling machines deliver masses X grams and Y grams. Machine X has mean 250 and standard deviation 4; machine Y has mean 244 and standard deviation 3. Let $D=X-Y$. (a) Find $E(D)$. (b) Find $Var(D)$ and $sd(D)$. (c) Explain why $Var(D)$ is the sum, not the difference, of the two variances. (d) Machine Y is recalibrated so that its mean becomes 250 grams, its standard deviation unchanged. State the new $E(D)$ and $sd(D)$.

Q5. Independent variables X_1, X_2, X_3 have the means and variances below, and $W=2X_1-X_2+3X_3+5$.

Variable	Mean	Variance
X_1	4	1
X_2	6	4
X_3	10	2

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

Q6. A random variable X has mean $\mu=15$ and variance $\sigma^2=9$. Consider $n=4$ independent copies X_1, X_2, X_3, X_4. (a) Find $E(X_1+X_2+X_3+X_4)$ and $Var(X_1+X_2+X_3+X_4)$. (b) Find $E(4X)$ and $Var(4X)$. (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 $E(X)=12$, $sd(X)=3$, $E(Y)=7$ and $sd(Y)=4$. Find the mean and standard deviation of $X+Y$ and of $X-Y$, and comment on their spreads.

Q8. Independent variables X_1, X_2, X_3 are identically distributed, each with mean 20 and variance 5. Find the mean, variance and standard deviation of the sum $S=X_1+X_2+X_3$ and of $3X$, and contrast them.

Q9. Independent random variables X and Y have $E(X)=6$, $Var(X)=9$, $E(Y)=10$ and $Var(Y)=4$. Find the positive value of a for which $Var(aX+Y)=40$, and hence find $E(aX+Y)$ for that value of a.

Q10. The mass of a filled crate is the mass of the empty crate plus the mass of its contents, and those two masses are independent. An empty crate has mean mass 4 kg and standard deviation 1.2 kg, while a filled crate has mean mass 30 kg and standard deviation 2 kg. Find the mean and standard deviation of the mass of the contents.

Q11. Independent variables X and Y have mean 5, variance 2 and mean 8, variance 3 respectively. A revenue is $R=4X+7Y$. Find $E(R)$, $Var(R)$ and $sd(R)$.

Q12. A single item has mass with mean 250 g and standard deviation 4 g, and items are packed independently. (a) Find the mean and standard deviation of the total mass of 16 items. (b) How many independent items give the total a standard deviation of 20 g?

Q13. A blended fuel has energy content $B = 0.5X + 0.3Y + 0.2Z$ megajoules per litre, where the energy contents X, Y and Z (in MJ/L) of the three independent components have means 34, 28 and 22 and standard deviations 2, 3 and 5 respectively. Find $E(B)$, $Var(B)$ and $sd(B)$.

Q14. The service times of two baristas are A, with mean 180 seconds and standard deviation 20 seconds, and B, with mean 150 seconds and standard deviation 15 seconds. (a) Assuming A and B are independent, find the mean and standard deviation of the total $A + B$. (b) Suppose instead that A and B are **not** independent. Which of the two answers in part (a) can still be relied on, and why?

Q15. Independent variables X, Y, Z have means 10, 6, 2 and variances 4, 9, 1 respectively. For $W = X - 2Y + 3Z$, find $E(W)$, $Var(W)$ and $sd(W)$.

Q16. Each morning a depot receives 6 crates, each of mean mass 20 kg with standard deviation 4 kg, and dispatches 4 pallets, each of mean mass 25 kg with standard deviation 5 kg. All ten masses are independent. Let D be the total mass received minus the total mass dispatched. Find $E(D)$, $Var(D)$ and $sd(D)$.

Q17. A drinks can is filled to a mean of 330 mL with standard deviation 5 mL, cans being filled independently. (a) For a pack of 6 cans the total is $S = X_1 + \ldots + X_6$. Find $E(S)$ and $sd(S)$, giving the standard deviation in exact form. (b) A student instead (wrongly) models the total as $6X$, six times a single can. Find $E(6X)$ and $sd(6X)$, and give the exact ratio $sd(6X) : sd(S)$. (c) For a pack of n cans, state the ratio $Var(nX) : Var(S)$.

Q18. A car's outbound trip time X has mean 50 minutes and standard deviation 8 minutes, and its return trip time Y has mean 46 minutes and standard deviation 6 minutes, independently. (a) For the round trip $T = X + Y$, find $E(T)$ and $sd(T)$. (b) For the difference $D = X - Y$, find $E(D)$ and $sd(D)$, and say what the sign of $E(D)$ tells you. (c) Explain why $E(T)$ and $sd(T)$ are not by themselves enough to find $P(T > 108)$, and state what further assumption about X and Y would make that probability calculable.

Answers

Q1. (a) $E=11$, $Var=7$; (b) $E=5$, $Var=7$; (c) $E=30$, $Var=53$; (d) $sd=\sqrt{53}\approx7.28$. **Q2.** $E(S)=100$ kg, $Var(S)=0.8$, $sd(S)=\sqrt{0.8}\approx0.894$ kg; $Var(20X)=16$ (one bag scaled by 20, not 20 separate bags). **Q3.** $E(S)=146$ g; $Var(S)=289$; $sd(S)=17$ g; a constant shifts the mean but has no spread of its own. **Q4.** $E(D)=6$ g; $Var(D)=25$, $sd(D)=5$ g; (d) $E(D)=0$ g with $sd(D)=5$ g unchanged. **Q5.** $E(W)=37$, $Var(W)=26$, $sd(W)=\sqrt{26}\approx5.10$. **Q6.** Sum: $E=60$, $Var=36$, $sd=6$; $4X$: $E=60$, $Var=144$, $sd=12$. **Q7.** $X+Y$: mean 19, $sd=5$; $X-Y$: mean 5, $sd=5$; equal spread, since the variances $9+16=25$ add either way. **Q8.** S: mean 60, variance 15, $sd=\sqrt{15}\approx3.87$; $3X$: mean 60, variance 45, $sd=3\sqrt{5}\approx6.71$. **Q9.** $a=2$; $E(2X+Y)=22$. **Q10.** Mean 26 kg; variance 2.56, so $sd=1.6$ kg. **Q11.** $E(R)=76$, $Var(R)=179$, $sd(R)=\sqrt{179}\approx13.38$. **Q12.** (a) mean 4000 g, $sd=16$ g; (b) $n=25$. **Q13.** $E(B)=29.8$ MJ/L; $Var(B)=2.81$; $sd(B)=\sqrt{2.81}\approx1.68$ MJ/L. **Q14.** (a) mean 330 s, $Var=625$, $sd=25$ s; (b) only the mean — expectations add regardless of independence, but the variance rule requires it. **Q15.** $E(W)=4$, $Var(W)=49$, $sd(W)=7$. **Q16.** $E(D)=20$ kg; $Var(D)=196$; $sd(D)=14$ kg. **Q17.** (a) $E(S)=1980$ mL, $sd(S)=5\sqrt{6}\approx12.25$ mL; (b) $E(6X)=1980$ mL, $sd(6X)=30$ mL, ratio $\sqrt{6}:1\approx2.449:1$; (c) $n:1$. **Q18.** (a) $E(T)=96$ min, $sd(T)=10$ min; (b) $E(D)=4$ min, $sd(D)=10$ min — the outbound leg is on average 4 minutes the slower; (c) the mean and standard deviation do not fix the shape of the distribution of T; if X and Y were independent *normal* variables, T would be normal and the probability could be found.

Fully-worked solutions

Q1. With X, Y independent: $E(X)=8$, $Var(X)=5$, $E(Y)=3$, $Var(Y)=2$. (a) $E(X+Y)=8+3=11$ and $Var(X+Y)=5+2=7$. (b) $E(X-Y)=8-3=5$ and $Var(X-Y)=5+2=7$ — the variances add here too, because the coefficient -1 squares to $+1$. (c) $E(3X+2Y)=3(8)+2(3)=24+6=30$ and $Var(3X+2Y)=3^2(5)+2^2(2)=45+8=53$. (d) $sd(3X+2Y)=\sqrt{53}\approx7.28$.

Q2. One bag: $\mu=5$, $\sigma^2=0.04$, $n=20$. (a) $E(S)=20\mu=20(5)=100$ kg. (b) $Var(S)=20\sigma^2=20(0.04)=0.8$. (c) $sd(S)=\sqrt{0.8}\approx0.894$ kg. (d) $Var(20X)=20^2\sigma^2=400(0.04)=16$. This is the variance of 20 times a **single** bag, whose one deviation is magnified by 20 (variance $\times 20^2$). The total S instead sums 20 **separate** independent bags, whose deviations partly cancel, so its variance grows only by a factor of 20. Hence $Var(20X)=16$ is far larger than $Var(S)=0.8$.

Q3. $S=2X+3Y+50$ with X, Y independent, $Var(X)=4^2=16$ and $Var(Y)=5^2=25$. (a) $E(S)=2(30)+3(12)+50=60+36+50=146$ g. (b) $Var(S)=2^2(16)+3^2(25)=4(16)+9(25)=64+225=289$. (c) $sd(S)=\sqrt{289}=17$ g. (d) The added sugar is a fixed 50 g, identical in every batch. It shifts every value of $2X+3Y$ up by 50, so the mean rises by 50; but a constant has no variability of its own and shifting a distribution does not stretch it, so the spread — and hence the variance — is unchanged.

Q4. X: mean 250, $Var(X)=4^2=16$; Y: mean 244, $Var(Y)=3^2=9$; independent, $D=X-Y$. (a) $E(D)=250-244=6$ g. (b) $Var(D)=16+9=25$, so $sd(D)=\sqrt{25}=5$ g. (c) In $D=X-Y$ the coefficient of Y is -1, and the variance rule squares every coefficient: $(-1)^2=1$. The two independent variabilities therefore accumulate rather than cancel, giving $Var(X-Y)=Var(X)+Var(Y)$. (d) Changing a mean shifts a distribution without changing its spread, so $E(D)=250-250=0$ g while $Var(D)=16+9=25$ and $sd(D)=5$ g are unchanged.

Q5. Independent X_1, X_2, X_3; $W=2X_1-X_2+3X_3+5$. (a) $E(W)=2(4)-1(6)+3(10)+5=8-6+30+5=37$. (b) The added constant 5 contributes no variance, and each coefficient is squared: $Var(W)=2^2(1)+(-1)^2(4)+3^2(2)=4+4+18=26$. (c) $sd(W)=\sqrt{26}\approx5.10$.

Q6. $\mu=15$, $\sigma^2=9$, $n=4$. (a) $E(X_1+X_2+X_3+X_4)=4\mu=60$ and $Var(X_1+X_2+X_3+X_4)=4\sigma^2=4(9)=36$. (b) $E(4X)=4\mu=60$ and $Var(4X)=4^2\sigma^2=16(9)=144$. (c) $sd(X_1+\ldots+X_4)=\sqrt{36}=6$ and $sd(4X)=\sqrt{144}=12$. (d) Both means are $4\mu=60$, but the sum of four independent copies has variance $4\sigma^2$ (deviations partly offset), while $4X$ scales one observation's deviation by 4, giving variance $4^2\sigma^2=16\sigma^2$ — four times as large.

Q7. Independent, so $Var(X)=3^2=9$ and $Var(Y)=4^2=16$. $X+Y$: mean $12+7=19$, variance $9+16=25$, so $sd=5$. $X-Y$: mean $12-7=5$, variance $9+16=25$, so $sd=5$. The two spreads are identical: whether the variables are added or subtracted, the independent variances add to 25, giving standard deviation 5 in both cases.

Q8. iid, each mean 20, variance 5, with $n=3$. $S=X_1+X_2+X_3$: mean $3(20)=60$, variance $3(5)=15$, so $sd=\sqrt{15}\approx3.87$. $3X$: mean $3(20)=60$, variance $3^2(5)=45$, so $sd=\sqrt{45}=3\sqrt{5}\approx6.71$. Same mean 60; but $3X$ has three times the variance of S (45 against 15), because it magnifies a single observation rather than summing three independent ones.

Q9. With X, Y independent, the coefficient of X is squared:

$$Var(aX+Y)=a^2Var(X)+1^2Var(Y)=9a^2+4.$$

Setting $9a^2+4=40$ gives $9a^2=36$, so $a^2=4$ and the positive solution is $a=2$. Then, by linearity of expectation, $E(2X+Y)=2(6)+10=12+10=22$.

Q10. Write $F=K+C$, where K is the empty-crate mass and C the contents' mass, with K and C independent. Means always add, so

$$E(C)=E(F)-E(K)=30-4=26\,\text{kg}.$$

Because K and C are independent the variances add, $Var(F)=Var(K)+Var(C)$, so

$$Var(C)=2^2-1.2^2=4-1.44=2.56,$$

giving $sd(C)=\sqrt{2.56}=1.6$ kg. (Note that it is the *variances* that subtract here, never the standard deviations: $2-1.2=0.8\neq1.6$.)

Q11. $R=4X+7Y$ with X, Y independent. $E(R)=4(5)+7(8)=20+56=76$. $Var(R)=4^2(2)+7^2(3)=16(2)+49(3)=32+147=179$. $sd(R)=\sqrt{179}\approx13.38$.

Q12. One item: $\mu=250$, $\sigma=4$ (so $\sigma^2=16$), items independent. (a) The total of 16 items has mean $16(250)=4000$ g and variance $16(16)=256$, so $sd=\sqrt{256}=16$ g. (b) The total of n items has standard deviation $\sqrt{n}\sigma=4\sqrt{n}$. Setting $4\sqrt{n}=20$ gives $\sqrt{n}=5$, so $n=25$ items.

Q13. $B=0.5X+0.3Y+0.2Z$ with X, Y, Z independent and variances $2^2=4$, $3^2=9$, $5^2=25$.

$$E(B)=0.5(34)+0.3(28)+0.2(22)=17+8.4+4.4=29.8\,\text{MJ/L}.$$

Each coefficient is squared, and squaring a fraction makes it smaller:

$$Var(B)=0.5^2(4)+0.3^2(9)+0.2^2(25)=0.25(4)+0.09(9)+0.04(25)=1+0.81+1=2.81,$$

so $sd(B)=\sqrt{2.81}\approx1.68$ MJ/L. The blend is much less variable than its most erratic component (sd 5), because the fractional weights shrink every contribution.

Q14. $Var(A)=20^2=400$ and $Var(B)=15^2=225$. (a) $E(A+B)=180+150=330$ s. Assuming independence, $Var(A+B)=400+225=625$, so $sd(A+B)=\sqrt{625}=25$ s. (b) Only the mean survives. Linearity of expectation, $E(A+B)=E(A)+E(B)$, holds for **any** two random variables, so $E(A+B)=330$ s regardless. The rule $Var(A+B)=Var(A)+Var(B)$, by contrast, is exactly the rule that needs independence: if the two baristas' times tend to rise and fall together (a busy period slows both) the total would vary more than 25 s, and if one tends to be quick when the other is slow it would vary less. Without independence the variance cannot be found from the two individual variances alone.

Q15. $W = X - 2Y + 3Z$ with X, Y, Z independent. $E(W) = 10 - 2(6) + 3(2) = 10 - 12 + 6 = 4$. $Var(W) = 1^2(4) + (-2)^2(9) + 3^2(1) = 4 + 36 + 9 = 49$, so $sd(W) = \sqrt{49} = 7$. (Each coefficient is squared, so the -2 contributes $+36$.)

Q16. Let $R = Y_1 + \ldots + Y_6$ be the total received (each Y_i with mean 20 kg and variance $4^2 = 16$) and $P = Z_1 + \ldots + Z_4$ the total dispatched (each Z_j with mean 25 kg and variance $5^2 = 25$). All ten masses are independent, so

$$E(R) = 6(20) = 120, Var(R) = 6(16) = 96,$$

$$E(P) = 4(25) = 100, Var(P) = 4(25) = 100.$$

For $D = R - P$ the means subtract but the variances add:

$$E(D) = 120 - 100 = 20\,\text{kg}, Var(D) = 96 + 100 = 196,$$

so $sd(D) = \sqrt{196} = 14$ kg. (Equivalently, D is a single linear combination of the ten independent masses with coefficients $+1$ six times and -1 four times, and every coefficient squares to 1.)

Q17. One can: $\mu = 330$, $\sigma = 5$ (so $\sigma^2 = 25$), cans independent. (a) $S = X_1 + \ldots + X_6$ sums 6 independent cans, so $E(S) = 6(330) = 1980$ mL and $Var(S) = 6(25) = 150$, giving $sd(S) = \sqrt{150} = 5\sqrt{6} \approx 12.25$ mL. (b) $6X$ scales a **single** can, so $E(6X) = 6(330) = 1980$ mL — the same mean — but $Var(6X) = 6^2(25) = 900$ and $sd(6X) = 30$ mL. The ratio is

$$\frac{sd(6X)}{sd(S)} = \frac{30}{5\sqrt{6}} = \frac{6}{\sqrt{6}} = \sqrt{6} \approx 2.449,$$

so $sd(6X) : sd(S) = \sqrt{6} : 1$. The wrong model overstates the spread by a factor of about 2.4. (c) In general $Var(nX) = n^2\sigma^2$ and $Var(S) = n\sigma^2$, so

$$Var(nX) : Var(S) = n^2\sigma^2 : n\sigma^2 = n : 1.$$

With $n = 6$ this gives $900 : 150 = 6 : 1$, matching part (b), whose standard deviations are in the ratio $\sqrt{6} : 1$.

Q18. X: mean 50, $Var(X) = 8^2 = 64$; Y: mean 46, $Var(Y) = 6^2 = 36$; independent. (a) $E(T) = 50 + 46 = 96$ min and $Var(T) = 64 + 36 = 100$, so $sd(T) = \sqrt{100} = 10$ min. (b) $E(D) = 50 - 46 = 4$ min and $Var(D) = 64 + 36 = 100$ as well (the variances add for a difference too), so $sd(D) = 10$ min. The mean is positive, so on average the outbound leg is the slower of the two, by 4 minutes. (c) A mean and a standard deviation fix only where the distribution of T is centred and how far it typically spreads; they say nothing about its **shape**. Many different distributions share the mean 96 and standard deviation 10, and they assign different amounts of probability beyond 108, so $P(T > 108)$ is not determined. If it were also given that X and Y are **normally** distributed, then T — a sum of independent normal variables — would itself be normal, with mean 96 and standard deviation 10, and the probability could be found by standardising. That is the step taken in the next section.

Theory

The previous section established two rules for combining random variables. For any random variables $X_1, X_2, \ldots, X_n$ with means $\mu_i = E(X_i)$ and constants $a_1, a_2, \ldots, a_n$, the mean of a linear combination is

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

and, **provided the variables are independent**, the variance is

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

where $\sigma_i^2 = Var(X_i)$. These formulas give the centre and spread of a combination whatever the shape of the distributions. This section adds the decisive extra fact for the **normal** case: the combination is not merely *some* distribution with that mean and variance — it is itself a **normal** distribution. That is what lets us calculate probabilities.

A linear combination of independent normals is normal.

If $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are **independent** normal random variables and a, b are constants, then

$$aX + bY \sim N\left(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2\right).$$

More generally, for independent normals $X_1, \ldots, X_n$ with $X_i \sim N(\mu_i, \sigma_i^2)$ and constants $a_1, \ldots, a_n$, the linear combination $a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$ is normally distributed, with

$$\text{mean} = \sum_{i=1}^{n} a_i \mu_i, \text{variance} = \sum_{i=1}^{n} a_i^2 \sigma_i^2.$$

This preservation of normality is special to the normal family — it does **not** hold for most distributions. Independence is what allows the variances to add with no cross terms. (A single normal shifted and scaled, $aX + b$, is the case of one variable, and is normal too, as seen earlier.)

Computing probabilities for a combination.

Once the combination W is known to be normal with mean μ_W and standard deviation σ_W, every probability is found by **standardising** the boundary and reading the standard normal $Z \sim N(0, 1)$ from technology:

$$P(W \leq w) = P\left(Z \leq \frac{w - \mu_W}{\sigma_W}\right).$$

The routine is always the same:

1. identify the coefficients a_i;
2. compute the mean $\mu_W = \sum a_i \mu_i$;
3. compute the variance $\sigma_W^2 = \sum a_i^2 \sigma_i^2$, then σ_W;
4. state $W \sim N(\mu_W, \sigma_W^2)$;
5. standardise the boundary and evaluate the required area.

The difference of two independent normals.

A case that arises constantly is the **difference** $X-Y$ — the combination with $a=1$ and $b=-1$. Its mean is $\mu_X-\mu_Y$, but its variance is

$$Var(X-Y)=(1)^2\sigma_X^2+(-1)^2\sigma_Y^2=\sigma_X^2+\sigma_Y^2,$$

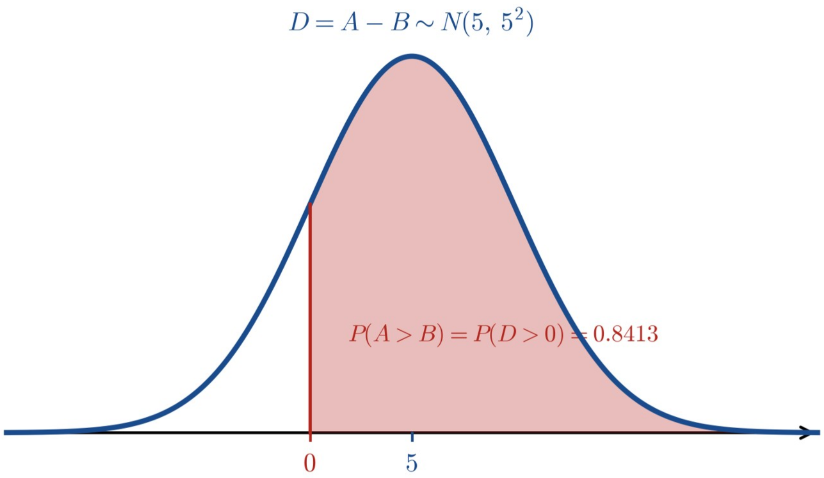
so the variances **add**, exactly as they do for a sum. Subtracting does not subtract the spread; uncertainty always accumulates. Hence for independent normals,

$$X-Y\sim N\left(\mu_X-\mu_Y,\sigma_X^2+\sigma_Y^2\right).$$

This is the tool for **comparing** two normal quantities. To find the chance that X exceeds Y, rewrite the event using the single variable $D=X-Y$:

$$P(X>Y)=P(X-Y>0)=P(D>0).$$

The picture shows $A\sim N(30,4^2)$ and $B\sim N(25,3^2)$: their difference is $D=A-B\sim N(5,5^2)$, and the shaded area $P(D>0)=0.8413$ is the probability that A exceeds B.



The sum of n identical independent normals.

If $X_1,X_2,\ldots,X_n$ are independent and **identically distributed** (iid), each $X_i\sim N(\mu,\sigma^2)$, then taking every $a_i=1$,

$$X_1+X_2+\ldots+X_n\sim N\left(n\mu,n\sigma^2\right),$$

with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$. The total of n measurements is normal; its mean grows in proportion to n, but its standard deviation grows only like $\sqrt{n}$.

Distinguishing X_1+X_2 from $2X$.

These expressions look alike but describe different random variables. X_1+X_2 is the sum of **two separate independent draws**; $2X$ is **one draw, doubled**. Both have mean 2μ, but their variances differ:

$$Var(X_1+X_2)=\sigma^2+\sigma^2=2\sigma^2, Var(2X)=2^2\sigma^2=4\sigma^2.$$

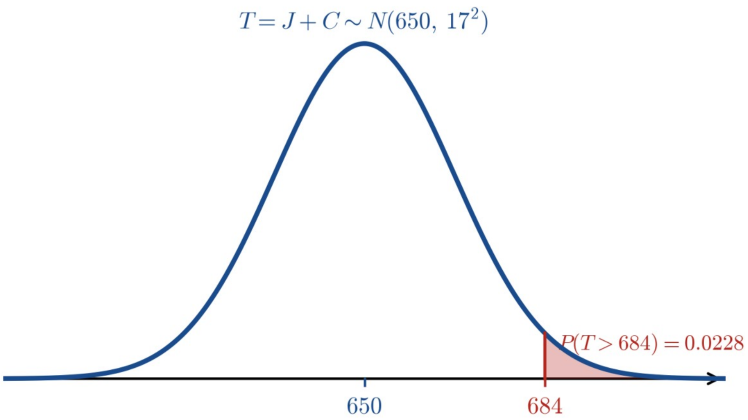
So $2X$ has **twice** the variance — standard deviation 2σ against $\sigma\sqrt{2}$. Doubling a single value amplifies its one random deviation, whereas adding two independent values lets their fluctuations partly cancel. Whenever a total is written down, ask whether it is a sum of independent draws or a multiple of a single draw: the means agree, but the spreads — and hence the probabilities — do not.

Worked examples

Example 1 — scaffolded

A filled jar's mass is the sum of the empty jar's mass $J \sim N(150, 8^2)$ g and the contents' mass $C \sim N(500, 15^2)$ g, which are independent. Let $T = J + C$. (a) State the distribution of T. (b) Find $P(T > 684)$.

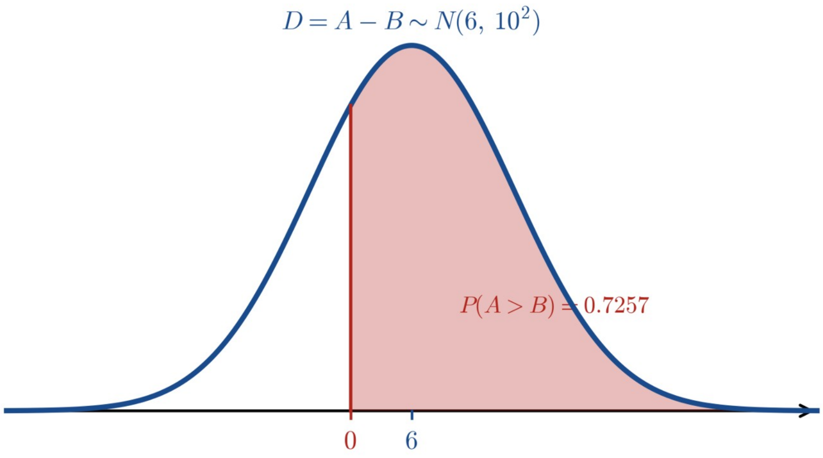
Working	Comment
$E(T) = E(J) + E(C) = 150 + 500 = 650$.	Means always add: $E(J+C) = \mu_J + \mu_C$.
$Var(T) = \sigma_J^2 + \sigma_C^2 = 8^2 + 15^2 = 64 + 225 = 289$.	Independent, so the variances add with no cross term.
$sd(T) = \sqrt{289} = 17$.	The standard deviation is the square root of the variance.
J and C are independent normals, so $T = J + C \sim N(650, 17^2)$.	A linear combination of independent normals is normal.
Standardise the boundary: $z = \dfrac{684 - 650}{17} = \dfrac{34}{17} = 2$.	$z = \dfrac{w - \mu_W}{\sigma_W}$.
$P(T > 684) = P(Z > 2) = 1 - P(Z \le 2) = 1 - 0.9772 = 0.0$.	From technology $P(Z \le 2) = 0.9772$.



Example 2 — scaffolded

Two brands of battery have lifetimes (hours) $A \sim N(40, 6^2)$ and $B \sim N(34, 8^2)$, independent. One battery of each brand is chosen at random. Find the probability that the brand-A battery lasts longer than the brand-B battery, $P(A > B)$.

Working	Comment
Let $D = A - B$; the event $A > B$ is the event $D > 0$.	Compare two normals through their difference.
$E(D) = \mu_A - \mu_B = 40 - 34 = 6$.	Means subtract for a difference.
$Var(D) = \sigma_A^2 + \sigma_B^2 = 6^2 + 8^2 = 36 + 64 = 100$, so $sd(D) = 10$.	For a difference the variances still **add**.
A, B are independent normals, so $D = A - B \sim N(6, 10^2)$.	Difference of independent normals is normal.
Standardise the boundary $D = 0$: $z = \dfrac{0 - 6}{10} = -0.6$.	$z = \dfrac{0 - \mu_D}{\sigma_D}$.
$P(A > B) = P(D > 0) = P(Z > -0.6) = 1 - P(Z \le -0.6$.	From technology $P(Z \le -0.6) = 0.2743$.



Example 3 — unscaffolded

A lift is rated to carry a total mass of up to 720 kg. Nine adults, whose masses (kg) are independent and each modelled by $X \sim N(70, 15^2)$, step in. Let $S = X_1 + X_2 + \ldots + X_9$ be their total mass. Find the distribution of S and the probability that the total mass exceeds the rating.

Because the nine masses are iid normal, the total is normal, with

$$E(S) = 9\mu = 9(70) = 630, Var(S) = 9\sigma^2 = 9(15^2) = 9(225) = 2025,$$

so $sd(S) = \sqrt{2025} = 45$ and $S \sim N(630, 45^2)$. Standardising the boundary 720,

$$z = \frac{720 - 630}{45} = \frac{90}{45} = 2,$$

so

$$P(S > 720) = P(Z > 2) = 1 - 0.9772 = 0.0228.$$

$\therefore$ the total mass is $S \sim N(630, 45^2)$, and $P(S > 720) \approx 0.0228$ — only about a 2.3% chance that the rated mass is exceeded.

Example 4 — unscaffolded

A machine fills bags of flour whose mass (g) is $X \sim N(500, 20^2)$. Compare two quantities: the combined mass $X_1 + X_2$ of two independently filled bags, and twice the mass $2X$ of a single bag. Find each distribution and the probability that each exceeds 1040 g.

Both quantities have mean $2(500) = 1000$ g, but their spreads differ. For two independent bags,

$$Var(X_1 + X_2) = \sigma^2 + \sigma^2 = 2(20^2) = 800, sd = \sqrt{800} = 20\sqrt{2} \approx 28.28,$$

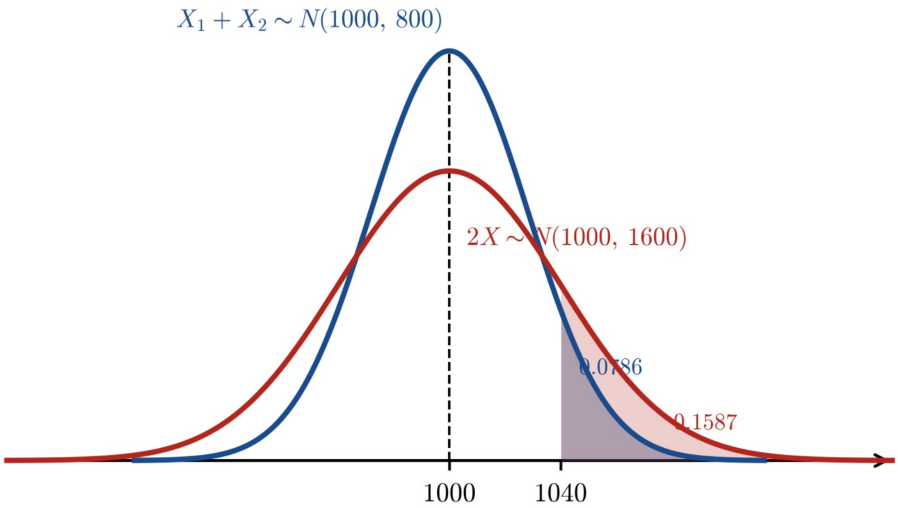
so $X_1 + X_2 \sim N(1000, 800)$. For one bag doubled,

$$Var(2X) = 2^2\sigma^2 = 4(20^2) = 1600, sd = 40,$$

so $2X \sim N(1000, 1600)$. Doubling a single bag gives the larger standard deviation. Standardising 1040 in each,

$$P(X_1 + X_2 > 1040) = P\left(Z > \frac{1040 - 1000}{20\sqrt{2}}\right) = P(Z > 1.4142) = 1 - 0.9214 = 0.0786,$$

$$P(2X > 1040) = P\left(Z > \frac{1040 - 1000}{40}\right) = P(Z > 1) = 1 - 0.8413 = 0.1587.$$



$\therefore X_1+X_2 \sim N(1000,800)$ and $2X \sim N(1000,1600)$ share the mean 1000 but not the spread, so $P(2X>1040)=0.1587$ is more than double $P(X_1+X_2>1040)=0.0786$: a multiple of one draw is more variable than a sum of independent draws.

Practice questions

Scaffolded

Q1. The heights of two stacked components are $A \sim N(40,3^2)$ cm and $B \sim N(25,4^2)$ cm, independent. Let $T=A+B$. (a) Find $E(T)$. (b) Find $Var(T)$ and $sd(T)$. (c) State the distribution of T. (d) Find $P(T<70)$.

Q2. $X \sim N(10,4^2)$ and $Y \sim N(8,1.5^2)$ are independent normal random variables, and $W=X+2Y$. (a) Find $E(W)$. (b) Find $Var(W)$ and $sd(W)$. (c) State the distribution of W. (d) Find $P(W>36)$.

Q3. Two independent normal random variables are $X \sim N(50,6^2)$ and $Y \sim N(42,8^2)$. Let $D=X-Y$. (a) Find $E(D)$. (b) Find $Var(D)$ and $sd(D)$. (c) State the distribution of D. (d) Hence find $P(X>Y)$.

Q4. A vending machine dispenses cans whose masses (g) are independent, each $X \sim N(375,5^2)$. A pack contains 4 cans; let $S=X_1+X_2+X_3+X_4$. (a) Find $E(S)$. (b) Find $Var(S)$ and $sd(S)$. (c) State the distribution of S. (d) Find $P(S<1490)$.

Q5. A machine fills bags of sugar whose mass (g) is $X \sim N(1000,15^2)$. (a) State the distribution of X_1+X_2, the total mass of two independently filled bags. (b) State the distribution of $2X$, twice the mass of a single bag. (c) Which of X_1+X_2 and $2X$ has the greater standard deviation, and why? (d) Find $P(2X>2060)$.

Q6. $X \sim N(120,1.5^2)$ and $Y \sim N(80,4^2)$ are independent, and $W=2X+Y$. (a) Find $E(W)$. (b) Find $Var(W)$ and $sd(W)$. (c) State the distribution of W. (d) Find $P(W<312)$.

Unscaffolded

Q7. Independent normal random variables are $X \sim N(20,5^2)$ and $Y \sim N(30,12^2)$. Find the distribution of $X+Y$ and hence $P(X+Y>63)$.

Q8. The daily takings (in hundreds of dollars) at two independent stalls are $X \sim N(100,6^2)$ and $Y \sim N(90,8^2)$. Find the probability that the first stall takes more than the second, $P(X>Y)$.

Q9. Two machines fill bottles; the volumes delivered are independent, $X \sim N(505,4^2)$ mL and $Y \sim N(500,3^2)$ mL. Let $D=X-Y$. Find $P(-5<X-Y<15)$.

Q10. A baker's tray holds 16 muffins whose masses (g) are independent, each $X \sim N(45, 5^2)$. Let S be the total mass of the tray. Find the distribution of S and $P(S > 760)$.

Q11. A component's mass (g) is $X \sim N(500, 10^2)$. (a) Explain the difference between $X_1 + X_2$ (two independent components) and $2X$ (one component, doubled), and state the standard deviation of each. (b) Find $P(X_1 + X_2 < 980)$ and $P(2X < 980)$, and comment.

Q12. An investor's return (per cent) is $R = 0.6X + 0.4Y$, where $X \sim N(20, 5^2)$ and $Y \sim N(10, 10^2)$ are the independent returns of two funds. Find $E(R)$, $sd(R)$ and $P(R > 21)$.

Q13. In a race the times (s) of two runners are independent, $X \sim N(50, 3^2)$ and $Y \sim N(54, 4^2)$. The first runner wins if $X < Y$. Find the probability that the first runner wins.

Q14. A pallet holds 25 bags of cement whose masses (kg) are independent, each $X \sim N(25, 0.6^2)$. Let S be the total mass. Find the total mass that is exceeded by 90% of pallets.

Q15. A shop's daily revenue and cost (dollars) are independent, $R \sim N(1000, 40^2)$ and $C \sim N(600, 30^2)$. The profit is $P = R - C$. Find the distribution of P, then $P(P > 450)$ and $P(P < 300)$.

Q16. A box holds 4 chocolate bars whose masses (g) are independent, each $X \sim N(50, 4^2)$. Let S be the total mass of the box. (a) Find $P(S > 216)$. (b) Find the total mass that is exceeded by 99% of boxes.

Q17. A three-course meal has course energies (in units of 100 kJ) that are independent normal random variables $X \sim N(5, 1^2)$, $Y \sim N(8, 2^2)$ and $Z \sim N(3, 2^2)$. Let $T = X + Y + Z$ be the total energy. Find (a) $P(T > 22)$ and (b) $P(10 < T < 19)$.

Q18. Two machines each fill jars to a mean of 250 g, with independent masses $A \sim N(250, 8^2)$ and $B \sim N(250, 6^2)$. Let $D = A - B$. (a) State the distribution of D. (b) Find $P(|A - B| > 10)$. (c) Find the value d such that $P(|A - B| > d) = 0.05$.

Answers

Q1. $E(T)=65$; $Var(T)=25$, $sd(T)=5$; $T\sim N(65,5^2)$; $P(T<70)=0.8413$. **Q2.** $E(W)=26$; $Var(W)=25$, $sd(W)=5$; $W\sim N(26,5^2)$; $P(W>36)=0.0228$. **Q3.** $E(D)=8$; $Var(D)=100$, $sd(D)=10$; $D\sim N(8,10^2)$; $P(X>Y)=0.7881$. **Q4.** $E(S)=1500$; $Var(S)=100$, $sd(S)=10$; $S\sim N(1500,10^2)$; $P(S<1490)=0.1587$. **Q5.** (a) $X_1+X_2\sim N(2000,450)$; (b) $2X\sim N(2000,900)$; (c) $2X$ (sd $30>15\sqrt{2}\approx 21.21$), because doubling one draw amplifies its single deviation; (d) $P(2X>2060)=0.0228$. **Q6.** $E(W)=320$; $Var(W)=25$, $sd(W)=5$; $W\sim N(320,5^2)$; $P(W<312)=0.0548$. **Q7.** $X+Y\sim N(50,13^2)$; $P(X+Y>63)=0.1587$. **Q8.** $P(X>Y)=0.8413$. **Q9.** $P(-5<X-Y<15)=0.9545$. **Q10.** $S\sim N(720,20^2)$; $P(S>760)=0.0228$. **Q11.** (a) X_1+X_2: sd $10\sqrt{2}\approx 14.14$; $2X$: sd 20. (b) $P(X_1+X_2<980)=0.0786$, $P(2X<980)=0.1587$; the doubled single draw is more variable, so more probability sits in its tail. **Q12.** $E(R)=16$, $sd(R)=5$; $P(R>21)=0.1587$. **Q13.** $P(X<Y)=0.7881$. **Q14.** $S\sim N(625,3^2)$; about 621.16 kg. **Q15.** $P\sim N(400,50^2)$; $P(P>450)=0.1587$; $P(P<300)=0.0228$. **Q16.** (a) $P(S>216)=0.0228$; (b) about 181.39 g. **Q17.** (a) $P(T>22)=0.0228$; (b) $P(10<T<19)=0.8186$. **Q18.** (a) $D\sim N(0,10^2)$; (b) $P(|A-B|>10)=0.3173$; (c) $d=19.60$ g.

Fully-worked solutions

Q1. Means add and, by independence, variances add: $E(T)=40+25=65$ and $Var(T)=3^2+4^2=9+16=25$, so $sd(T)=5$. As a sum of independent normals, $T=A+B\sim N(65,5^2)$. Standardising, $z=\dfrac{70-65}{5}=1$, so $P(T<70)=P(Z<1)=0.8413$.

Q2. With $W=X+2Y$ ($a=1$ on X, $a=2$ on Y), $E(W)=1(10)+2(8)=10+16=26$ and $Var(W)=1^2(4^2)+2^2(1.5^2)=16+4(2.25)=16+9=25$, so $sd(W)=5$. As a linear combination of independent normals, $W\sim N(26,5^2)$. Standardising, $z=\dfrac{36-26}{5}=2$, so $P(W>36)=P(Z>2)=1-0.9772=0.0228$.

Q3. For the difference $D=X-Y$: $E(D)=50-42=8$, and the variances add, $Var(D)=6^2+8^2=36+64=100$, so $sd(D)=10$ and $D\sim N(8,10^2)$. Then $P(X>Y)=P(D>0)$; standardising $D=0$, $z=\dfrac{0-8}{10}=-0.8$, so $P(X>Y)=P(Z>-0.8)=1-P(Z\le -0.8)=1-0.2119=0.7881$.

Q4. The four masses are iid normal, so S is normal with $E(S)=4(375)=1500$ and $Var(S)=4(5^2)=100$, giving $sd(S)=10$ and $S\sim N(1500,10^2)$. Standardising, $z=\dfrac{1490-1500}{10}=-1$, so $P(S<1490)=P(Z<-1)=0.1587$.

Q5. (a) Two independent bags give $X_1+X_2\sim N(2000,15^2+15^2)=N(2000,450)$, with $sd=15\sqrt{2}\approx 21.21$. (b) One bag doubled gives $2X\sim N(2000,2^2\cdot 15^2)=N(2000,900)$, with $sd=30$. (c) $2X$ has the greater standard deviation ($30>21.21$): doubling one draw multiplies its single deviation by 2, whereas a sum of two independent draws lets the deviations partly cancel. (d) Standardising in $2X\sim N(2000,900)$, $z=\dfrac{2060-2000}{30}=2$, so $P(2X>2060)=P(Z>2)=1-0.9772=0.0228$.

Q6. With $W=2X+Y$, $E(W)=2(120)+80=320$ and $Var(W)=2^2(1.5^2)+1^2(4^2)=4(2.25)+16=9+16=25$, so $sd(W)=5$ and $W\sim N(320,5^2)$. Standardising, $z=\dfrac{312-320}{5}=-1.6$, so $P(W<312)=P(Z<-1.6)=0.0548$.

Q7. $E(X+Y)=20+30=50$ and $Var(X+Y)=5^2+12^2=25+144=169$, so $sd=13$ and $X+Y\sim N(50,13^2)$. Standardising, $z=\dfrac{63-50}{13}=1$, so $P(X+Y>63)=P(Z>1)=1-0.8413=0.1587$.

Q8. Let $D = X - Y$: $E(D) = 100 - 90 = 10$ and $Var(D) = 6^2 + 8^2 = 100$, so $sd(D) = 10$ and $D \sim N(10, 10^2)$. Then $P(X > Y) = P(D > 0)$; standardising, $z = \frac{0 - 10}{10} = -1$, so $P(X > Y) = P(Z > -1) = 1 - 0.1587 = 0.8413$.

Q9. Let $D = X - Y$: $E(D) = 505 - 500 = 5$ and $Var(D) = 4^2 + 3^2 = 16 + 9 = 25$, so $sd(D) = 5$ and $D \sim N(5, 5^2)$. The endpoints standardise to $z = \frac{-5 - 5}{5} = -2$ and $z = \frac{15 - 5}{5} = 2$, so

$$P(-5 < X - Y < 15) = P(-2 < Z < 2) = 0.97725 - 0.02275 = 0.9545.$$

Q10. The 16 masses are iid normal, so the total is normal with $E(S) = 16(45) = 720$ and $Var(S) = 16(5^2) = 400$, giving $sd(S) = 20$ and $S \sim N(720, 20^2)$. Standardising, $z = \frac{760 - 720}{20} = 2$, so $P(S > 760) = P(Z > 2) = 1 - 0.9772 = 0.0228$.

Q11. (a) $X_1 + X_2$ adds **two separate independent** components, so its variance is $10^2 + 10^2 = 200$ and $sd = 10\sqrt{2} \approx 14.14$; $2X$ **doubles one** component, so its variance is $2^2(10^2) = 400$ and $sd = 20$. Both have mean 1000. (b) In $X_1 + X_2 \sim N(1000, 200)$, $z = \frac{980 - 1000}{10\sqrt{2}} = -1.4142$, so $P(X_1 + X_2 < 980) = P(Z < -1.4142) = 1 - 0.9214 = 0.0786$. In $2X \sim N(1000, 400)$, $z = \frac{980 - 1000}{20} = -1$, so $P(2X < 980) = P(Z < -1) = 0.1587$. The doubled single draw is more variable, so more probability lies beyond 980 .

Q12. With $R = 0.6X + 0.4Y$, $E(R) = 0.6(20) + 0.4(10) = 12 + 4 = 16$ and $Var(R) = 0.6^2(5^2) + 0.4^2(10^2) = 0.36(25) + 0.16(100) = 9 + 16 = 25$, so $sd(R) = 5$ and $R \sim N(16, 5^2)$. Standardising, $z = \frac{21 - 16}{5} = 1$, so $P(R > 21) = P(Z > 1) = 1 - 0.8413 = 0.1587$.

Q13. Let $D = X - Y$: $E(D) = 50 - 54 = -4$ and $Var(D) = 3^2 + 4^2 = 9 + 16 = 25$, so $sd(D) = 5$ and $D \sim N(-4, 5^2)$. The first runner wins if $X < Y$, i.e. $D < 0$; standardising, $z = \frac{0 - (-4)}{5} = 0.8$, so $P(X < Y) = P(D < 0) = P(Z < 0.8) = 0.7881$.

Q14. The 25 masses are iid normal, so S is normal with $E(S) = 25(25) = 625$ and $Var(S) = 25(0.6^2) = 25(0.36) = 9$, giving $sd(S) = 3$ and $S \sim N(625, 3^2)$. "Exceeded by 90% of pallets" means $P(S > s) = 0.90$, so $P(S < s) = 0.10$ and s is the 10th percentile. The standard normal value with $P(Z < z) = 0.10$ is $z = -1.2816$, so

$$s = 625 + (-1.2816)(3) = 625 - 3.845 = 621.16.$$

So about 621.16 kg is exceeded by 90% of pallets.

Q15. For the profit $P = R - C$: $E(P) = 1000 - 600 = 400$ and, as a difference of independent normals, $Var(P) = 40^2 + 30^2 = 1600 + 900 = 2500$, so $sd(P) = 50$ and $P \sim N(400, 50^2)$. Then $z = \frac{450 - 400}{50} = 1$ gives $P(P > 450) = P(Z > 1) = 1 - 0.8413 = 0.1587$, and $z = \frac{300 - 400}{50} = -2$ gives $P(P < 300) = P(Z < -2) = 0.0228$.

Q16. The four masses are iid normal, so $S \sim N(200, 64)$ with $E(S) = 4(50) = 200$, $Var(S) = 4(4^2) = 64$ and $sd(S) = 8$. (a) $z = \frac{216 - 200}{8} = 2$, so $P(S > 216) = P(Z > 2) = 1 - 0.9772 = 0.0228$. (b) "Exceeded by 99% of boxes" means $P(S > s) = 0.99$, so $P(S < s) = 0.01$; the standard normal value with $P(Z < z) = 0.01$ is $z = -2.3263$, so

$$s = 200 + (-2.3263)(8) = 200 - 18.610 = 181.39.$$

So about 181.39 g is exceeded by 99 % of boxes.

Q17. The total $T = X + Y + Z$ is a sum of independent normals, so it is normal with $E(T) = 5 + 8 + 3 = 16$ and $Var(T) = 1^2 + 2^2 + 2^2 = 1 + 4 + 4 = 9$, giving $sd(T) = 3$ and $T \sim N(16, 3^2)$. (a) $z = \frac{22-16}{3} = 2$, so $P(T > 22) = P(Z > 2) = 1 - 0.9772 = 0.0228$. (b) The endpoints standardise to $z = \frac{10-16}{3} = -2$ and $z = \frac{19-16}{3} = 1$, so

$$P(10 < T < 19) = P(-2 < Z < 1) = 0.84134 - 0.02275 = 0.8186.$$

Q18. (a) For $D = A - B$: $E(D) = 250 - 250 = 0$, and the variances add, $Var(D) = 8^2 + 6^2 = 64 + 36 = 100$, so $sd(D) = 10$ and $D \sim N(0, 10^2)$. (b) By symmetry about 0,

$$P(|A - B| > 10) = 2P(D > 10) = 2P(Z > 1) = 2(1 - 0.84134) = 0.3173.$$

(c) $P(|A - B| > d) = 0.05$ means $P(D > d) = 0.025$ by symmetry, so d standardises to the value with $P(Z < z) = 0.975$, namely $z = 1.9600$. Then

$$d = 0 + 1.9600(10) = 19.60.$$

So $|A - B|$ exceeds about 19.60 g only 5 % of the time.

Theory

A **random sample** of size n from a population is a set of random variables $X_1, X_2, \ldots, X_n$ — one for each observation — that are **independent** and each have the **same distribution** as the population. If the population has mean μ and standard deviation σ (so variance σ^2), then each X_i satisfies

$$E(X_i) = \mu, Var(X_i) = \sigma^2, i = 1, 2, \ldots, n.$$

Before the sample is collected the individual values are uncertain, so each X_i is a random variable; only after measuring do they take particular numerical values.

The quantity of interest is usually not a single observation but their **average**. The **sample mean**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is itself a random variable, because it is built from the random variables $X_1, \ldots, X_n$. Different samples give different values of $\bar{X}$, so $\bar{X}$ has its own distribution — the **sampling distribution of the mean** — with its own mean and standard deviation. This section finds them, using the rules for a linear combination of independent random variables.

Mean of the sample mean.

Since $\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ is a linear combination of the X_i, the mean follows from the rule $E(a_1 X_1 + \cdots + a_n X_n) = a_1 E(X_1) + \cdots + a_n E(X_n)$ with every coefficient $a_i = \frac{1}{n}$:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}(n\mu) = \mu.$$

So $E(\bar{X}) = \mu$: the sample mean is centred on the population mean, whatever the sample size. We say $\bar{X}$ is an **unbiased** estimator of μ.

Variance and standard deviation of the sample mean.

Because the X_i are **independent**, the variance of their sum is the sum of the variances — $Var(a_1 X_1 + \cdots + a_n X_n) = a_1^2 Var(X_1) + \cdots + a_n^2 Var(X_n)$ — and each coefficient $\frac{1}{n}$ contributes a factor $\frac{1}{n^2}$:

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}.$$

Taking the square root,

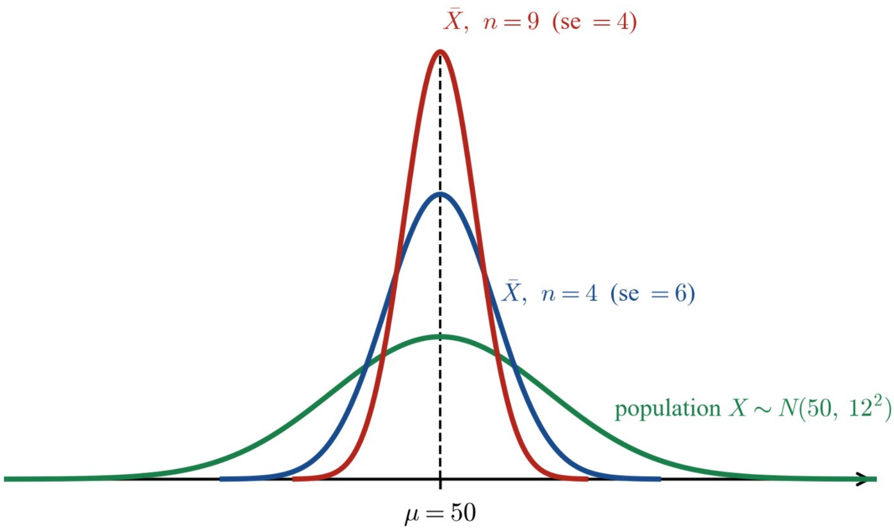
$$Var(\bar{X}) = \frac{\sigma^2}{n}, sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

The standard deviation of the sample mean, $\frac{\sigma}{\sqrt{n}}$, is called the **standard error** of the mean.

Interpretation: the sample mean is less variable.

The two results say that averaging leaves the centre unchanged but **shrinks the spread**. A single observation has standard deviation σ; the mean of n observations has standard deviation $\frac{\sigma}{\sqrt{n}}$, smaller by a factor $\frac{1}{\sqrt{n}}$. High and low values in a sample tend to cancel when averaged, so $\bar{X}$ clusters more tightly around μ than any individual X_i does.

Because the factor is $\frac{1}{\sqrt{n}}$ and not $\frac{1}{n}$, the gain slows as n grows: to **halve** the standard error the sample size must be **quadrupled**. For instance $n=4$ halves it ($\sqrt{4}=2$), $n=9$ divides it by 3, and $n=100$ divides it by 10. The diagram shows a population $X \sim N(50, 12^2)$ together with the sampling distribution of $\bar{X}$ for $n=4$ (standard error 6) and $n=9$ (standard error 4): each curve keeps the same centre $\mu=50$ but is taller and narrower as n increases.



Sampling from a normal population.

If the population itself is **normal**, $X \sim N(\mu, \sigma^2)$, then $\bar{X}$ is a linear combination of independent normal random variables, and such a combination is again normal. Together with the mean and variance found above,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{exactly, for every } n.$$

No approximation is needed: for a normal population the sampling distribution of the mean is **exactly** normal. (For a non-normal population $\bar{X}$ is only *approximately* normal, and only for large n — the central limit theorem, treated separately.)

Probabilities about $\bar{X}$ are then found exactly as for any normal variable, by standardising with the standard error in place of σ:

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1), P(\bar{X} \le \bar{x}) = P\left(Z \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right).$$

The only change from a single-observation calculation is that σ is replaced by the standard error $\frac{\sigma}{\sqrt{n}}$ in the denominator.

Finding the sample size.

Because the standard error decreases with n, a sample can be made large enough to achieve a required precision. Setting $\frac{\sigma}{\sqrt{n}}$ at or below a target value and solving gives the smallest suitable n; since n must be a whole number, the result is **rounded up**. Similarly, requiring $\bar{X}$ to lie within a stated distance of μ with a given probability leads, after standardising, to an inequality for n.

Worked examples

Example 1 — scaffolded

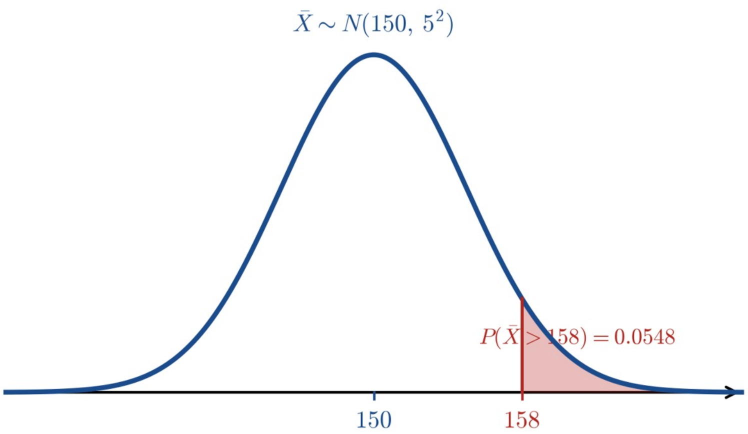
A population has mean $\mu=50$ and standard deviation $\sigma=12$. A random sample of size $n=9$ is taken and its mean $\bar{X}$ recorded. Find $E(\bar{X})$, $Var(\bar{X})$ and the standard error $sd(\bar{X})$, and compare the spread of $\bar{X}$ with that of a single observation.

Working	Comment
$E(\bar{X})=\mu=50$.	The sample mean is centred on the population mean.
$Var(\bar{X})=\frac{\sigma^2}{n}=\frac{12^2}{9}=\frac{144}{9}=16$.	Variance of the mean is σ^2/n.
$sd(\bar{X})=\frac{\sigma}{\sqrt{n}}=\frac{12}{\sqrt{9}}=\frac{12}{3}=4$.	Standard error is $\sigma/\sqrt{n}$; equivalently $\sqrt{16}=4$.
A single observation has $sd=12$; the mean has $sd=4=\frac{12}{3}$.	Smaller by the factor $\frac{1}{\sqrt{9}}=\frac{1}{3}$.
So $\bar{X}$ is one-third as spread out as a single value, though centred at the same place.	Averaging reduces variability without shifting the centre.

Example 2 — scaffolded

The masses of apples from an orchard are normally distributed with mean $\mu=150$ g and standard deviation $\sigma=20$ g. A random sample of $n=16$ apples is weighed. Find the probability that the sample mean mass exceeds 158 g.

Working	Comment
$sd(\bar{X})=\frac{\sigma}{\sqrt{n}}=\frac{20}{\sqrt{16}}=\frac{20}{4}=5$.	Standard error replaces σ.
The population is normal, so $\bar{X}\sim N(150,5^2)$ exactly.	Mean of a normal population is exactly normal.
Standardise: $z=\frac{158-150}{5}=\frac{8}{5}=1.6$.	$z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$.
$P(\bar{X}>158)=P(Z>1.6)=1-P(Z\le 1.6)$.	Upper tail.
$=1-0.9452=0.0548$.	From technology, $P(Z\le 1.6)=0.9452$.



Example 3 — unscaffolded

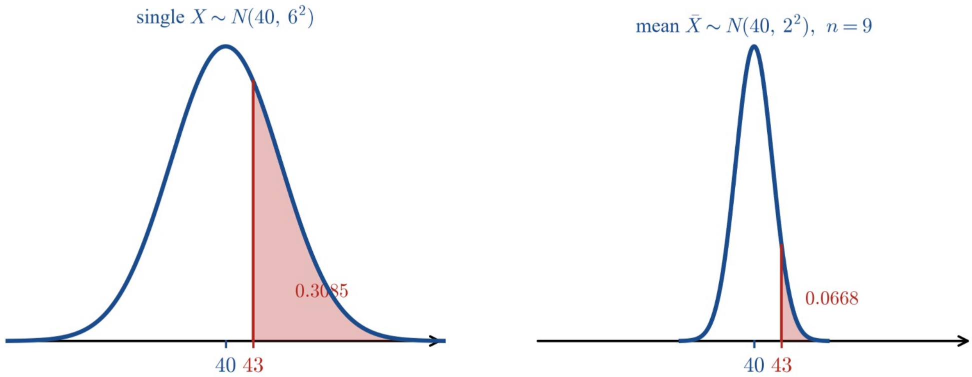
The time to complete a task is normally distributed with mean 40 minutes and standard deviation 6 minutes, $X\sim N(40,6^2)$. (a) Find the probability that one randomly chosen worker takes more than 43 minutes. (b) A random sample of 9 workers is taken; find the probability that their mean time exceeds 43 minutes. Compare the two results.

For a **single** observation the spread is $\sigma = 6$, so standardising gives $z = \dfrac{43-40}{6} = 0.5$ and

$$P(X > 43) = P(Z > 0.5) = 1 - 0.6915 = 0.3085.$$

For the **sample mean** of $n = 9$, the standard error is $\dfrac{\sigma}{\sqrt{n}} = \dfrac{6}{3} = 2$, so $\bar{X} \sim N(40, 2^2)$ exactly (the population is normal). Standardising, $z = \dfrac{43-40}{2} = 1.5$ and

$$P(\bar{X} > 43) = P(Z > 1.5) = 1 - 0.9332 = 0.0668.$$



$\therefore$ a single worker exceeds 43 minutes with probability 0.3085, but the mean of nine workers exceeds 43 minutes with probability only 0.0668. The sample mean is far less likely to stray from 40, because its standard error 2 is one-third of the population's 6.

Example 4 — unscaffolded

A quantity is normally distributed in a population with standard deviation $\sigma = 10$. How large must a random sample be so that the sample mean lies within 2 units of the population mean with probability at least 0.95; that is, $P(|\bar{X} - \mu| < 2) \geq 0.95$?

Standardising the event $|\bar{X} - \mu| < 2$ with the standard error $\dfrac{\sigma}{\sqrt{n}} = \dfrac{10}{\sqrt{n}}$,

$$|\bar{X} - \mu| < 2 \Leftrightarrow |Z| < \frac{2}{10/\sqrt{n}} = \frac{2\sqrt{n}}{10} = \frac{\sqrt{n}}{5}.$$

The central probability $P(|Z| < z^*) = 0.95$ requires $z^* = 1.96$ (from technology, $P(Z \leq 1.96) = 0.975$). Hence

$$\frac{\sqrt{n}}{5} \geq 1.96 \quad \sqrt{n} \geq 9.8 \quad n \geq 96.04.$$

Because n must be a whole number, round up.

$\therefore$ a sample of at least $n = 97$ observations makes the sample mean lie within 2 of μ with probability at least 0.95.

Practice questions

Scaffolded

Q1. A population has mean $\mu = 60$ and standard deviation $\sigma = 20$. A random sample of size $n = 25$ is taken. (a) State $E(\bar{X})$. (b) Find $Var(\bar{X})$. (c) Find the standard error $sd(\bar{X})$. (d) By what factor is the standard deviation of $\bar{X}$ smaller than that of a single observation?

Q2. A normal population has $X \sim N(100, 15^2)$, and a random sample of size $n=9$ is taken. (a) State the distribution of $\bar{X}$. (b) Find $sd(\bar{X})$. (c) Standardise $\bar{X}=105$. (d) Find $P(\bar{X}>105)$.

Q3. A normal population has $X \sim N(25, 8^2)$, and a random sample of size $n=16$ is taken. (a) Find $sd(\bar{X})$. (b) Standardise the endpoints 22 and 28. (c) Find $P(22<\bar{X}<28)$.

Q4. A population has standard deviation $\sigma=24$. A random sample of size n is taken, and the standard error of the mean must be at most 4. (a) Write the standard error in terms of n. (b) Set up the inequality $\frac{24}{\sqrt{n}} \le 4$. (c) Solve for $\sqrt{n}$, then for n. (d) State the smallest suitable sample size.

Q5. Test scores are modelled by $X \sim N(500, 80^2)$, and a random sample of size $n=64$ is taken. (a) State the distribution of $\bar{X}$. (b) Find $sd(\bar{X})$. (c) Find $P(\bar{X}<490)$. (d) Find $P(\bar{X}>515)$.

Q6. A normal population has $X \sim N(20, 5^2)$, and a random sample of size $n=25$ is taken. (a) Find $P(X>22)$ for a single observation. (b) Find $sd(\bar{X})$. (c) Find $P(\bar{X}>22)$. (d) Explain why the two probabilities differ.

Unscaffolded

Q7. A population has mean 40 and standard deviation 9. For a random sample of size $n=81$, find $E(\bar{X})$, $Var(\bar{X})$ and $sd(\bar{X})$.

Q8. Heights are modelled by $X \sim N(70, 12^2)$. A random sample of $n=36$ is taken. Find $P(\bar{X}<66)$.

Q9. The volume of drink in a bottle is $X \sim N(1.5, 0.4^2)$ litres. A random sample of $n=25$ bottles is taken. Find $P(\bar{X}>1.6)$.

Q10. A population has standard deviation $\sigma=30$. Find the smallest sample size n for which the standard error of the mean is at most 4.

Q11. A normal population has $X \sim N(250, 40^2)$. A random sample of $n=100$ is taken. Find $P(246<\bar{X}<258)$.

Q12. A normal population has $X \sim N(60, 10^2)$. (a) Find $P(X>63)$ for a single observation. (b) For a random sample of $n=16$, find $P(\bar{X}>63)$.

Q13. A population has standard deviation $\sigma=18$. Find the smallest sample size n for which the standard error of the mean is less than 2.5.

Q14. A random sample $X_1, X_2, \ldots, X_n$ is taken from a population with mean μ and variance σ^2. Using $\bar{X}=\frac{1}{n}(X_1+X_2+\cdots+X_n)$ and the rules for linear combinations of independent random variables, show that $E(\bar{X})=\mu$ and $Var(\bar{X})=\frac{\sigma^2}{n}$.

Q15. A quantity is normally distributed with standard deviation $\sigma=6$. Find the smallest sample size n so that $P(|\bar{X}-\mu|<1.5) \ge 0.99$.

Q16. A machine fills bags whose net mass is $X \sim N(505, 8^2)$ grams. A quality inspector weighs a random sample of $n=4$ bags. Find the probability that the sample mean mass is less than 500 g.

Q17. The mass of a newborn lamb is $X \sim N(3.2, 0.5^2)$ kg. A random sample of $n=25$ lambs is taken. (a) Find $P(\bar{X}>3.35)$. (b) Find the value c such that $P(\bar{X}>c)=0.90$.

Q18. A population is normal with mean $\mu = 80$ and standard deviation $\sigma = 20$. (a) For a random sample of size $n = 16$, find $P(\acute{X} < 75)$. (b) Find the smallest sample size n so that $sd(\acute{X}) \leq 2$. (c) Using that sample size, find $P(78 < \acute{X} < 82)$.

Answers

Q1. $E(\bar{X})=60$; $Var(\bar{X})=16$; $sd(\bar{X})=4$; smaller by a factor $\frac{1}{5}$. **Q2.** $\bar{X}\sim N(100,5^2)$; $sd(\bar{X})=5$; $z=1$; $P(\bar{X}>105)=0.1587$. **Q3.** $sd(\bar{X})=2$; $z=-1.5$ and $z=1.5$; $P(22<\bar{X}<28)=0.8664$. **Q4.** $\frac{24}{\sqrt{n}}$; $\sqrt{n}\geq 6$, so $n\geq 36$; smallest $n=36$. **Q5.** $\bar{X}\sim N(500,10^2)$; $sd(\bar{X})=10$; $P(\bar{X}<490)=0.1587$; $P(\bar{X}>515)=0.0668$. **Q6.** (a) 0.3446; (b) $sd(\bar{X})=1$; (c) 0.0228; (d) $\bar{X}$ is less variable (standard error 1 vs 5), so it is far less likely to exceed 22. **Q7.** $E(\bar{X})=40$, $Var(\bar{X})=1$, $sd(\bar{X})=1$. **Q8.** 0.0228. **Q9.** 0.1056. **Q10.** $n=57$. **Q11.** 0.8186. **Q12.** (a) 0.3821; (b) 0.1151. **Q13.** $n=52$. **Q14.** $E(\bar{X})=\mu$ and $Var(\bar{X})=\frac{\sigma^2}{n}$ (shown). **Q15.** $n=107$. **Q16.** 0.1056. **Q17.** (a) 0.0668; (b) $c\approx 3.07$ kg. **Q18.** (a) 0.1587; (b) $n=100$; (c) 0.6827.

Fully-worked solutions

Q1. With $\mu=60$, $\sigma=20$, $n=25$: $E(\bar{X})=\mu=60$; $Var(\bar{X})=\frac{\sigma^2}{n}=\frac{20^2}{25}=\frac{400}{25}=16$; and $sd(\bar{X})=\frac{\sigma}{\sqrt{n}}=\frac{20}{\sqrt{25}}=\frac{20}{5}=4$. The standard deviation of $\bar{X}$ is 4, against 20 for a single observation, so it is smaller by the factor $\frac{1}{\sqrt{25}}=\frac{1}{5}$.

Q2. The population is normal, so $\bar{X}$ is exactly normal with $sd(\bar{X})=\frac{15}{\sqrt{9}}=\frac{15}{3}=5$, giving $\bar{X}\sim N(100,5^2)$. Standardising, $z=\frac{105-100}{5}=1$, so $P(\bar{X}>105)=P(Z>1)=1-0.8413=0.1587$.

Q3. The standard error is $sd(\bar{X})=\frac{8}{\sqrt{16}}=\frac{8}{4}=2$, so $\bar{X}\sim N(25,2^2)$. The endpoints standardise to $z=\frac{22-25}{2}=-1.5$ and $z=\frac{28-25}{2}=1.5$, hence

$$P(22<\bar{X}<28)=P(-1.5<Z<1.5)=0.9332-0.0668=0.8664.$$

Q4. The standard error is $\frac{\sigma}{\sqrt{n}}=\frac{24}{\sqrt{n}}$. Requiring it to be at most 4 gives $\frac{24}{\sqrt{n}}\leq 4$, so $\sqrt{n}\geq\frac{24}{4}=6$, hence $n\geq 36$. The smallest whole-number sample size is $n=36$ (and then the standard error is exactly $\frac{24}{6}=4$).

Q5. The population is normal, so $\bar{X}$ is exactly normal with $sd(\bar{X})=\frac{80}{\sqrt{64}}=\frac{80}{8}=10$, giving $\bar{X}\sim N(500,10^2)$. Then $z=\frac{490-500}{10}=-1$, so $P(\bar{X}<490)=P(Z<-1)=0.1587$; and $z=\frac{515-500}{10}=1.5$, so $P(\bar{X}>515)=P(Z>1.5)=1-0.9332=0.0668$.

Q6. (a) For a single observation, $z=\frac{22-20}{5}=0.4$, so $P(X>22)=P(Z>0.4)=1-0.6554=0.3446$. (b) The standard error is $sd(\bar{X})=\frac{5}{\sqrt{25}}=\frac{5}{5}=1$. (c) Now $z=\frac{22-20}{1}=2$, so $P(\bar{X}>22)=P(Z>2)=1-0.9772=0.0228$. (d) The two differ because $\bar{X}$ has standard error 1 while a single value has standard deviation 5; the sample mean is much more tightly concentrated about 20, so it exceeds 22 far less often.

Q7. With $\mu=40$, $\sigma=9$, $n=81$: $E(\bar{X})=40$; $Var(\bar{X})=\frac{9^2}{81}=\frac{81}{81}=1$; and $sd(\bar{X})=\frac{9}{\sqrt{81}}=\frac{9}{9}=1$.

Q8. The standard error is $\frac{12}{\sqrt{36}}=\frac{12}{6}=2$, so $\bar{X}\sim N(70,2^2)$. Standardising, $z=\frac{66-70}{2}=-2$, hence $P(\bar{X}<66)=P(Z<-2)=0.0228$.

Q9. The standard error is $\frac{0.4}{\sqrt{25}}=\frac{0.4}{5}=0.08$, so $\bar{X}\sim N(1.5,0.08^2)$. Standardising, $z=\frac{1.6-1.5}{0.08}=1.25$, hence $P(\bar{X}>1.6)=P(Z>1.25)=1-0.8944=0.1056$.

Q10. Requiring the standard error at most 4: $\frac{30}{\sqrt{n}}\leq 4$, so $\sqrt{n}\geq\frac{30}{4}=7.5$, hence $n\geq 7.5^2=56.25$. Because n must be a whole number, round up: the smallest sample size is $n=57$ (then $\frac{30}{\sqrt{57}}\approx 3.97\leq 4$, while $n=56$ gives $\frac{30}{\sqrt{56}}\approx 4.01>4$).

Q11. The standard error is $\frac{40}{\sqrt{100}}=\frac{40}{10}=4$, so $\bar{X}\sim N(250,4^2)$. The endpoints standardise to $z=\frac{246-250}{4}=-1$ and $z=\frac{258-250}{4}=2$, hence

$$P(246<\bar{X}<258)=P(-1<Z<2)=0.97725-0.15866=0.81859\approx 0.8186.$$

Keep the two areas to five decimal places until the end: rounding them to four decimal places first gives $0.9772-0.1587=0.8185$, which is wrong in the fourth decimal place.

Q12. (a) For a single observation, $z=\frac{63-60}{10}=0.3$, so $P(X>63)=P(Z>0.3)=1-0.6179=0.3821$. (b) For $n=16$ the standard error is $\frac{10}{\sqrt{16}}=\frac{10}{4}=2.5$, so $\bar{X}\sim N(60,2.5^2)$. Then $z=\frac{63-60}{2.5}=1.2$, so $P(\bar{X}>63)=P(Z>1.2)=1-0.8849=0.1151$.

Q13. Requiring the standard error below 2.5: $\frac{18}{\sqrt{n}}<2.5$, so $\sqrt{n}>\frac{18}{2.5}=7.2$, hence $n>51.84$. The smallest whole-number sample size is $n=52$ (then $\frac{18}{\sqrt{52}}\approx 2.496<2.5$, while $n=51$ gives $\frac{18}{\sqrt{51}}\approx 2.521>2.5$).

Q14. Writing $\bar{X}=\frac{1}{n}(X_1+X_2+\cdots+X_n)$, each coefficient is $a_i=\frac{1}{n}$. By the linear-combination rule for the mean, and since each $E(X_i)=\mu$ (there are n equal terms),

$$E(\bar{X})=\frac{1}{n}E(X_1)+\cdots+\frac{1}{n}E(X_n)=\frac{1}{n}(\mu+\mu+\cdots+\mu)=\frac{1}{n}(n\mu)=\mu.$$

By the rule for the variance of a combination of **independent** variables (squares of the coefficients),

$$Var(\bar{X})=\frac{1}{n^2}Var(X_1)+\cdots+\frac{1}{n^2}Var(X_n)=\frac{1}{n^2}(n\sigma^2)=\frac{\sigma^2}{n}.$$

Hence $E(\bar{X})=\mu$ and $Var(\bar{X})=\frac{\sigma^2}{n}$.

Q15. Standardising $|\bar{X}-\mu|<1.5$ with standard error $\frac{6}{\sqrt{n}}$,

$$|Z|<\frac{1.5}{6/\sqrt{n}}=\frac{1.5\sqrt{n}}{6}=\frac{\sqrt{n}}{4}.$$

The requirement $P(|Z|<z^*)=0.99$ needs $z^*=2.5758$ (from technology, $P(Z\le 2.5758)=0.995$). So $\frac{\sqrt{n}}{4}\ge 2.5758$, giving $\sqrt{n}\ge 4(2.5758)=10.3032$ and $n\ge 10.3032^2=106.16$. Rounding up, the smallest sample size is $n=107$.

Q16. The standard error is $\frac{8}{\sqrt{4}}=\frac{8}{2}=4$, so $\acute{X}\sim N(505,4^2)$. Standardising, $z=\frac{500-505}{4}=-1.25$, hence $P(\acute{X}<500)=P(Z<-1.25)=0.1056$.

Q17. The standard error is $\frac{0.5}{\sqrt{25}}=\frac{0.5}{5}=0.1$, so $\acute{X}\sim N(3.2,0.1^2)$. (a) $z=\frac{3.35-3.2}{0.1}=1.5$, so $P(\acute{X}>3.35)=P(Z>1.5)=1-0.9332=0.0668$. (b) $P(\acute{X}>c)=0.90$ means $P(\acute{X}<c)=0.10$, so c has z-score $z=-1.2816$ (since $P(Z<-1.2816)=0.10$). Then

$$c=\mu+z\,sd(\acute{X})=3.2+(-1.2816)(0.1)=3.2-0.12816=3.07184\approx 3.07\text{ kg}.$$

Q18. (a) For $n=16$ the standard error is $\frac{20}{\sqrt{16}}=\frac{20}{4}=5$, so $\acute{X}\sim N(80,5^2)$; then $z=\frac{75-80}{5}=-1$, so $P(\acute{X}<75)=P(Z<-1)=0.1587$. (b) Requiring $\frac{20}{\sqrt{n}}\le 2$ gives $\sqrt{n}\ge 10$, so $n\ge 100$; the smallest sample size is $n=100$. (c) With $n=100$ the standard error is $\frac{20}{\sqrt{100}}=2$, so $\acute{X}\sim N(80,2^2)$. The endpoints standardise to $z=\frac{78-80}{2}=-1$ and $z=\frac{82-80}{2}=1$, hence

$$P(78<\acute{X}<82)=P(-1<Z<1)=0.84134-0.15866=0.68268\approx 0.6827,$$

again carrying five decimal places in the two areas rather than rounding them first.

Chapter 11 Linear combinations of random variables and the sample mean — 11E Investigating the distribution of the sample mean using simulation

Theory

For a random sample $X_1, X_2, \ldots, X_n$ drawn from a population with mean μ and standard deviation σ, the **sample mean**

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

is itself a random variable: a different sample gives a different value $\bar{x}$. The previous section established its exact behaviour — that

$$E(\bar{X}) = \mu, sd(\bar{X}) = \frac{\sigma}{\sqrt{n}},$$

and that $\bar{X}$ is **exactly normal** when the population is normal. This section rebuilds that same picture **experimentally**, by **simulation**, and then uses simulation to see what happens when the population is *not* normal.

Building the distribution of $\bar{X}$ by simulation.

A single sample produces just one number, $\bar{x}$, which tells us little about the random variable $\bar{X}$. To see the whole **distribution** of $\bar{X}$ we repeat the sampling many times:

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 — say 1000 or more **repetitions** — keeping n fixed.
3. Collect all the recorded $\bar{x}$ values and display them in a **histogram**.

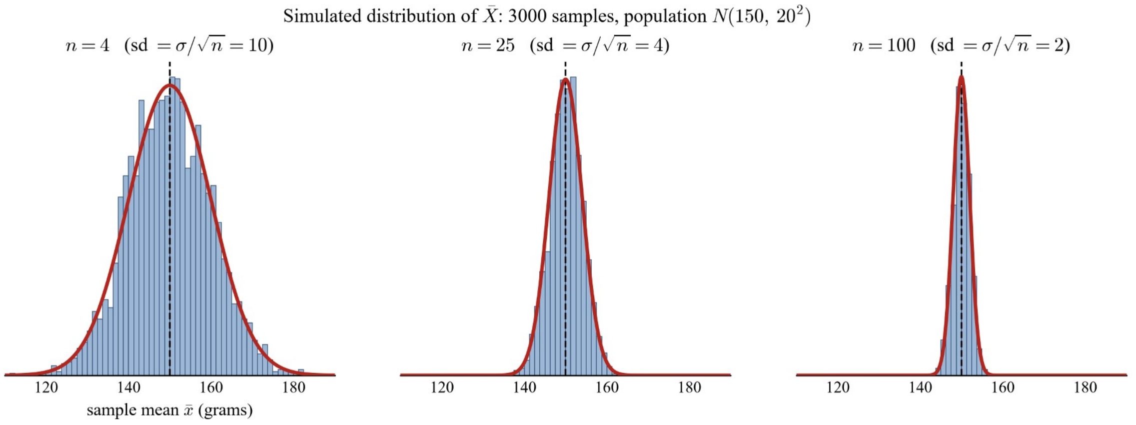
The resulting histogram is a **simulated sampling distribution of the sample mean**. Because it is built from real (simulated) data it is only an *approximation* to the true distribution of $\bar{X}$, but with enough repetitions it becomes a very accurate one.

What a large simulation reveals.

Three features stand out, and each matches the theory exactly.

- **Centre.** The histogram is centred on the population mean μ: the average of all the simulated $\bar{x}$ values is close to μ, because $E(\bar{X}) = \mu$.
- **Spread.** The histogram has standard deviation close to $\frac{\sigma}{\sqrt{n}}$, which is **smaller** than the population spread σ by a factor $\frac{1}{\sqrt{n}}$.
- **Shape.** When the population is normal the histogram is bell-shaped — a normal curve.

The clearest way to see the effect of the sample size is to run the simulation for several values of n from the **same** population. The diagram below shows 3000 simulated sample means for $n=4$, $n=25$ and $n=100$, drawn from a population of apple masses modelled by $X \sim N(150, 20^2)$ grams. As n increases the histogram stays centred on 150 but becomes **narrower**, and the theoretical normal curve $N(\mu, \sigma^2/n)$ (in red) sits neatly over each one.



The narrowing is governed entirely by $\frac{\sigma}{\sqrt{n}}$: multiplying n by 25 (from 4 to 100) divides the spread by $\sqrt{25}=5$ (from 10 to 2).

sample size n	1	4	25	100
theoretical $sd(\bar{X})=\frac{\sigma}{\sqrt{n}}$	20	10	4	2

The table below compares the theory with the three simulations shown above: in each case the simulated centre is close to $\mu=150$ and the simulated spread is close to $\frac{\sigma}{\sqrt{n}}$, as expected.

n	simulated centre	simulated sd	theory: μ	theory: $\sigma/\sqrt{n}$
4	149.9	9.98	150	10
25	150.2	3.95	150	4
100	150.0	2.01	150	2

Reading a simulated histogram.

A simulated histogram of sample means can be read the other way round to *estimate* the population figures:

- its **centre** estimates the population mean μ;
- its **standard deviation** estimates $\frac{\sigma}{\sqrt{n}}$, so multiplying that spread by $\sqrt{n}$ estimates the population standard deviation σ;
- its **shape** indicates whether a normal model for $\bar{X}$ is reasonable.

Sample size n versus number of repetitions.

Two different numbers control a simulation, and it is essential to keep them apart.

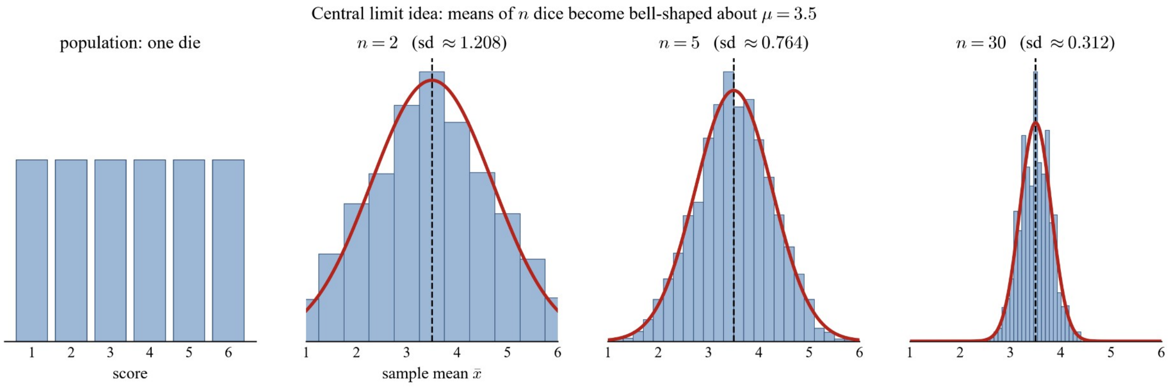
- The **sample size** n is how many observations go into each $\bar{x}$. It controls the *actual distribution* of $\bar{X}$: increasing n shrinks the spread to $\frac{\sigma}{\sqrt{n}}$ and (for a non-normal population) makes the shape more bell-like.
- The **number of repetitions** is how many sample means are collected. It controls only how *smoothly and reliably* the histogram traces that distribution. Using more repetitions makes the histogram less jagged and brings it closer to the true model, but it does **not** move the centre or change the spread — those are fixed by μ, σ and n.

Non-normal populations: the central limit idea.

Simulation is most illuminating when the population is **not** normal, because then the exact "normal in, normal out" rule no longer applies. Remarkably, the sample mean still becomes approximately normal as n grows. The diagram below starts from a fair die, whose score is **uniform** on $\{1,2,3,4,5,6\}$ — as far from bell-shaped as possible — with

$$\mu=3.5,\sigma=\sqrt{35/12}\approx 1.708.$$

Histograms of 3000 sample means are shown for $n=2$, $n=5$ and $n=30$.



The mean of two dice already has a triangular shape; by $n=5$ it is clearly mound-shaped; and by $n=30$ it is very close to the normal curve $N\left(3.5,\sigma^2/30\right)$. Throughout, the centre stays at $\mu=3.5$ and the spread follows $\frac{\sigma}{\sqrt{n}}$ (about 1.208, then 0.764, then 0.312). This is the **central limit theorem** idea, stated informally:

> For a random sample of size n from *any* population with mean μ and standard deviation σ, the sample mean is **approximately normal** for large n,
>
> $$\acute{X}\approx N\left(\mu,\frac{\sigma^2}{n}\right),$$
>
> and the approximation improves as n increases. A common rule of thumb is that $n\geq 30$ is usually large enough.

If the population is already normal the result is exact for every n; if it is not, the normal model for $\acute{X}$ is an approximation that is good once n is large enough.

Using the simulated distribution to estimate probabilities.

Because the simulated distribution of $\acute{X}$ is so well described by $N\left(\mu,\sigma^2/n\right)$, probabilities about a sample mean are found exactly as for any normal variable — by standardising with the sample-mean standard deviation $\frac{\sigma}{\sqrt{n}}$:

$$P\left(\acute{X}>a\right)=P\left(Z>\frac{a-\mu}{\sigma/\sqrt{n}}\right),Z\sim N(0,1).$$

For a normal population this is exact; for a non-normal population it is the central-limit approximation, and is reliable when n is large.

Worked examples

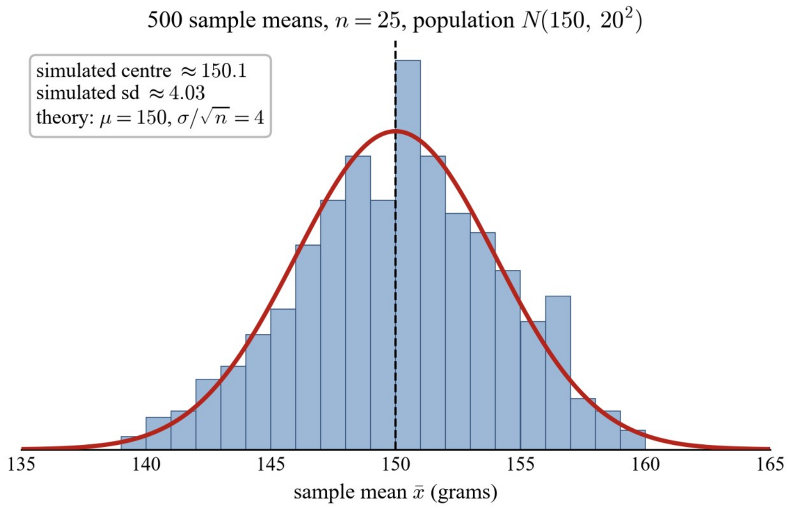
Example 1 — scaffolded

A machine fills sugar bags whose mass X grams is modelled by $X\sim N\left(500,8^2\right)$. A simulation draws 1000 samples of size $n=16$ and records the mean $\acute{x}$ of each. Predict the centre, spread and shape of the histogram of these sample means, then compare with the simulation, which gave an average of the sample means of 500.2 and a standard deviation of 1.98.

Working	Comment
Centre: $E(\acute{X})=\mu=500$.	The histogram is centred on the population mean.
Spread: $sd(\acute{X})=\frac{\sigma}{\sqrt{n}}=\frac{8}{\sqrt{16}}=\frac{8}{4}=2$.	Divide the population sd by $\sqrt{n}$.
Shape: the population is normal, so $\acute{X}\sim N(500,2^2)$ exactly — the histogram is bell-shaped.	A sample mean of a normal population is exactly normal.
Compare: simulated centre $500.2\approx 500$ and simulated sd $1.98\approx 2$.	The simulation agrees with the theory to within the expected variation.

Example 2 — scaffolded

Apple masses are modelled by $X\sim N(150,20^2)$ grams. The histogram below shows the means of 500 samples of size $n=25$. (a) State the centre and standard deviation the histogram should have. (b) The simulation gave an average of the sample means of 150.1 and a standard deviation of 4.03. Are these consistent with the theory? (c) If instead samples of size $n=100$ were used, by what factor would the spread of the histogram change?



Working	Comment
(a) Centre $=\mu=150$; spread $=\frac{\sigma}{\sqrt{n}}=\frac{20}{\sqrt{25}}=\frac{20}{5}=4$.	Read the centre from μ and the spread from $\sigma/\sqrt{n}$.
(b) Simulated centre $150.1\approx 150$ and simulated sd $4.03\approx 4$, so the results are consistent.	Both simulated values are close to the theoretical ones.
(c) New spread $=\frac{20}{\sqrt{100}}=\frac{20}{10}=2$.	Recompute $\sigma/\sqrt{n}$ with $n=100$.
The spread changes from 4 to 2, a factor of $\frac{1}{2}$.	Since $\sqrt{100/25}=2$, the spread halves.

Example 3 — unscaffolded

A fair six-sided die is rolled repeatedly; each score X is **uniform** on $\{1,2,3,4,5,6\}$ with mean $\mu=3.5$ and standard deviation $\sigma=\sqrt{35/12}\approx 1.708$. A game uses the mean $\acute{X}$ of $n=30$ rolls. Explain why $\acute{X}$ is approximately normal even though a single roll is not, state the distribution of $\acute{X}$, and estimate $P(\acute{X}>4)$.

A single roll is uniform, not normal, so no exact normal result applies to one roll. However, $n=30$ is large, so by the central limit theorem the sample mean is approximately normal. Its centre and spread are

$$E(\acute{X})=\mu=3.5,\ sd(\acute{X})=\frac{\sigma}{\sqrt{n}}=\frac{1.708}{\sqrt{30}}\approx 0.3118,$$

so $\acute{X}\approx N(3.5,0.3118^2)$. Standardising the boundary,

$$z = \frac{4 - 3.5}{0.3118} \approx 1.604,$$

so

$$P(\bar{X} > 4) = P(Z > 1.604) = 1 - P(Z \le 1.604) \approx 1 - 0.9456 = 0.0544.$$

$\therefore \bar{X} \approx N(3.5, 0.3118^2)$ and $P(\bar{X} > 4) \approx 0.0544$ — about 5.4% of samples of 30 rolls have a mean above 4.

Example 4 — unscaffolded

A quantity X in a population has mean 20 and standard deviation 6 (its distribution is right-skewed, not normal). Three simulations are run: **A** uses 200 samples of size $n = 9$; **B** uses 5000 samples of size $n = 9$; **C** uses 200 samples of size $n = 36$. Describe how the histogram of sample means compares between A and B, and between A and C, giving the theoretical spread in each case.

Simulations A and B use the **same** sample size $n = 9$, so their sample means have the same theoretical distribution: both are centred at $\mu = 20$ with spread

$$\frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{9}} = \frac{6}{3} = 2.$$

The only difference is the number of repetitions: B collects 5000 means rather than 200, so B's histogram is **smoother** and traces the true distribution more closely, but its centre and spread are the same as A's.

Simulation C keeps 200 repetitions but increases the sample size to $n = 36$, which changes the actual distribution of $\bar{X}$. Its spread is

$$\frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{36}} = \frac{6}{6} = 1,$$

half that of A, so C's histogram is **narrower** — though still centred at 20 and, with the larger n, closer to a normal shape.

$\therefore$ more **repetitions** (A $\rightarrow$ B) only smooth the histogram, leaving the centre 20 and spread 2 unchanged, while a larger **sample size** (A $\rightarrow$ C) narrows the spread from 2 to 1.

Practice questions

Scaffolded

Q1. Test scores are modelled by $X \sim N(70, 12^2)$. A simulation draws 800 samples of size $n = 9$ and records each mean. (a) State the centre the histogram of sample means should have. (b) Find its standard deviation $\frac{\sigma}{\sqrt{n}}$. (c) State its expected shape. (d) The simulation gave an average of the sample means of 69.8 and a standard deviation of 4.05. Are these consistent with the theory?

Q2. A population has standard deviation $\sigma = 15$. Consider the histogram of sample means for various sample sizes. (a) Find its standard deviation for $n = 9$. (b) Find it for $n = 25$. (c) Find it for $n = 225$. (d) By what factor does the spread change as n increases from 9 to 225?

Q3. A population has mean 40 and standard deviation 10. (a) Find the standard deviation of the sample-mean histogram for $n = 4$. (b) Find it for $n = 100$. (c) A student keeps $n = 100$ but increases the number of samples from 500 to 5000. Does the centre or the spread of the histogram change? (d) What does increasing the number of repetitions improve?

Q4. A spinner's score has mean 5 and standard deviation 2, and its distribution is **not** normal. A simulation draws 400 samples of size $n=50$. (a) State the centre of the histogram of sample means. (b) Find its standard deviation $\frac{\sigma}{\sqrt{n}}$ (to 3 decimal places). (c) State its expected shape, and justify it. (d) Using the normal model for $\bar{X}$, estimate $P(\bar{X}<4.5)$.

Q5. A population is modelled by $X\sim N(500,20^2)$. Samples of size $n=16$ are taken. (a) State $E(\bar{X})$. (b) Find $sd(\bar{X})$. (c) State the distribution of $\bar{X}$. (d) Find $P(\bar{X}>508)$.

Q6. A population has standard deviation $\sigma=24$. Two simulated histograms of sample means are produced: one for $n=16$ and one for $n=64$. (a) Find the standard deviation of the histogram for $n=16$. (b) Find it for $n=64$. (c) Which histogram is narrower? (d) Both histograms are centred on the same value. Explain why.

Unscaffolded

Q7. A population is modelled by $X\sim N(250,30^2)$. A simulation of 1000 samples of size $n=36$ gave an average of the sample means of 250.4 and a standard deviation of 4.90. State the centre, spread and shape the histogram should have, and say whether the simulation is consistent with the theory.

Q8. A population has standard deviation 45. Find the sample size n for which the histogram of sample means has standard deviation equal to 5.

Q9. Waiting times at a clinic are right-skewed (not normal) with mean 8 minutes and standard deviation 5 minutes. For samples of size $n=40$, explain why the sample mean is approximately normal, state the distribution of $\bar{X}$, and estimate $P(\bar{X}<7)$.

Q10. Two simulations of the **same** population use the **same** sample size but 100 and $10\,000$ repetitions. Describe what differs between the two resulting histograms and what stays the same.

Q11. A population is modelled by $X\sim N(1000,50^2)$. For samples of size $n=25$, find $P(990<\bar{X}<1015)$.

Q12. A fair die is rolled $n=100$ times; the score is uniform with mean 3.5 and standard deviation $\sigma=\sqrt{35/12}\approx 1.708$. Find the standard deviation of $\bar{X}$ and estimate $P(\bar{X}>3.6)$.

Q13. A simulation of 2000 samples of size $n=25$ from an unknown population gave an average of the sample means of 61.9 and a standard deviation of the sample means of 2.02. Use these to estimate the population mean and the population standard deviation.

Q14. A population has standard deviation 18. Find the smallest sample size n for which the histogram of sample means has standard deviation at most 3.

Q15. A population is modelled by $X\sim N(210,42^2)$. Student A simulates 500 samples of size $n=4$; Student B simulates 500 samples of size $n=36$. Find the standard deviation of each student's histogram, state which is narrower and by what factor, and find $P(\bar{X}>220)$ for each student's sample mean.

Q16. A population that is **not** normal has mean 12 and standard deviation 4. For samples of size $n=64$, estimate $P(\bar{X}>13)$, and state the result you are relying on.

Q17. A simulated sampling distribution of $\bar{X}$ for samples of size $n=49$ has standard deviation 1.4. Find the population standard deviation, and predict the standard deviation of the histogram if the sample size is increased to $n=196$.

Q18. The number of parcels a depot receives per day has mean 20 and standard deviation 9, and is right-skewed (not normal). Let $\bar{X}$ be the mean daily number over $n=36$ days. (a) Explain why $\bar{X}$ is approximately normal. (b) State the distribution of $\bar{X}$. (c) Find $P(\bar{X}>22)$. (d) Find $P(19<\bar{X}<21)$.

Answers

Q1. (a) 70; (b) 4; (c) approximately normal (bell-shaped); (d) yes — $69.8 \approx 70$ and $4.05 \approx 4$. **Q2.** (a) 5; (b) 3; (c) 1; (d) a factor of $\frac{1}{5}$. **Q3.** (a) 5; (b) 1; (c) no; (d) it makes the histogram smoother and closer to the true model. **Q4.** (a) 5; (b) 0.283; (c) approximately normal, by the central limit theorem ($n=50$ is large); (d) 0.0385. **Q5.** (a) 500; (b) 5; (c) $\bar{X} \sim N(500, 5^2)$; (d) 0.0548. **Q6.** (a) 6; (b) 3; (c) the $n=64$ histogram; (d) $E(\bar{X})=\mu$ for every n, so the centre does not depend on n. **Q7.** Centre 250, sd 5, shape normal; consistent, since $250.4 \approx 250$ and $4.90 \approx 5$. **Q8.** $n=81$. **Q9.** By the central limit theorem $\bar{X} \approx N(8, 0.7906^2)$; $P(\bar{X}<7) \approx 0.1030$. **Q10.** More repetitions make the histogram smoother and closer to the true distribution; the centre and spread are unchanged. **Q11.** 0.7745. **Q12.** $sd(\bar{X}) \approx 0.1708$; $P(\bar{X}>3.6) \approx 0.2791$. **Q13.** $\mu \approx 61.9$; $\sigma \approx 2.02 \times \sqrt{25} = 10.1$. **Q14.** $n=36$. **Q15.** A: 21; B: 7; B is narrower, by a factor of $\frac{1}{3}$; $P(\bar{X}>220) \approx 0.3170$ for A and ≈ 0.0766 for B. **Q16.** 0.0228, relying on the central limit theorem. **Q17.** $\sigma = 1.4 \times 7 = 9.8$; for $n=196$ the spread is 0.7 (halved). **Q18.** (a) $n=36 \geq 30$, so the central limit theorem applies; (b) $\bar{X} \approx N(20, 1.5^2)$; (c) 0.0912; (d) 0.4950.

Fully-worked solutions

Q1. The histogram of sample means is centred on the population mean, so (a) its centre is $\mu=70$. (b) Its standard deviation is $\frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{9}} = \frac{12}{3} = 4$. (c) Because the population is normal, $\bar{X}$ is exactly normal, so the histogram is bell-shaped. (d) The simulated centre 69.8 is close to 70 and the simulated standard deviation 4.05 is close to 4, so the simulation is consistent with the theory.

Q2. The spread of the sample-mean histogram is $\frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{n}}$. (a) For $n=9$: $\frac{15}{3}=5$. (b) For $n=25$: $\frac{15}{5}=3$. (c) For $n=225$: $\frac{15}{15}=1$. (d) The spread falls from 5 to 1, a factor of $\frac{1}{5}$; this matches $\frac{1}{\sqrt{225/9}} = \frac{1}{\sqrt{25}} = \frac{1}{5}$.

Q3. The spread is $\frac{10}{\sqrt{n}}$. (a) For $n=4$: $\frac{10}{2}=5$. (b) For $n=100$: $\frac{10}{10}=1$. (c) Increasing the number of repetitions does **not** change the centre or the spread: those are fixed by μ, σ and n, all unchanged here. (d) More repetitions make the histogram smoother and bring it closer to the true distribution of $\bar{X}$.

Q4. (a) The centre is $\mu=5$. (b) The standard deviation is $\frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{50}} \approx 0.283$. (c) Although the population is not normal, $n=50$ is large, so by the central limit theorem the histogram is approximately normal. (d) Using $\bar{X} \approx N(5, 0.283^2)$, standardise: $z = \frac{4.5-5}{0.283} \approx -1.768$, so $P(\bar{X}<4.5) = P(Z<-1.768) \approx 0.0385$.

Q5. (a) $E(\bar{X}) = \mu = 500$. (b) $sd(\bar{X}) = \frac{20}{\sqrt{16}} = \frac{20}{4} = 5$. (c) As the mean of a normal population, $\bar{X} \sim N(500, 5^2)$ exactly. (d) Standardising, $z = \frac{508-500}{5} = \frac{8}{5} = 1.6$, so $P(\bar{X}>508) = P(Z>1.6) = 1-0.9452 = 0.0548$.

Q6. The spread is $\frac{24}{\sqrt{n}}$. (a) For $n=16$: $\frac{24}{4}=6$. (b) For $n=64$: $\frac{24}{8}=3$. (c) The $n=64$ histogram is narrower, since a larger sample size gives a smaller spread. (d) The centre of the histogram is $E(\bar{X})=\mu$ for every sample size, so both histograms are centred on the same population mean regardless of n.

Q7. The histogram should be centred on $\mu = 250$, with standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{30}{\sqrt{36}} = \frac{30}{6} = 5$, and — because the population is normal — a normal (bell) shape. The simulated centre 250.4 is close to 250 and the simulated standard deviation 4.90 is close to 5, so the simulation is consistent with the theory.

Q8. Require $\frac{\sigma}{\sqrt{n}} = \frac{45}{\sqrt{n}} = 5$. Then $\sqrt{n} = \frac{45}{5} = 9$, so $n = 9^2 = 81$.

Q9. The population is not normal, but $n = 40$ is large, so by the central limit theorem the sample mean is approximately normal. Its centre is $\mu = 8$ and its standard deviation is $\frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{40}} \approx 0.7906$, so $\acute{X} \approx N(8, 0.7906^2)$. Standardising, $z = \frac{7-8}{0.7906} \approx -1.265$, so $P(\acute{X} < 7) = P(Z < -1.265) \approx 0.1030$.

Q10. Because the population and the sample size are the same, the true distribution of $\acute{X}$ is identical in both simulations, so both histograms have the **same** centre and spread. The only difference is the number of repetitions: the $10\,000$-repetition histogram is smoother and traces the true distribution of $\acute{X}$ more closely, while the 100-repetition histogram is more jagged. Increasing the repetitions does not shift the centre or change the spread.

Q11. Here $\acute{X} \sim N(1000, \sigma_{\acute{X}}^2)$ with $\sigma_{\acute{X}} = \frac{50}{\sqrt{25}} = \frac{50}{5} = 10$. The endpoints standardise to $z = \frac{990-1000}{10} = -1$ and $z = \frac{1015-1000}{10} = 1.5$, so

$$P(990 < \acute{X} < 1015) = P(-1 < Z < 1.5) = 0.9332 - 0.1587 = 0.7745.$$

Q12. The standard deviation of the sample mean is $\frac{\sigma}{\sqrt{n}} = \frac{1.708}{\sqrt{100}} = \frac{1.708}{10} \approx 0.1708$. As $n = 100$ is large, $\acute{X} \approx N(3.5, 0.1708^2)$. Standardising, $z = \frac{3.6-3.5}{0.1708} \approx 0.586$, so $P(\acute{X} > 3.6) = P(Z > 0.586) = 1 - 0.7209 = 0.2791$.

Q13. The centre of the histogram estimates the population mean, so $\mu \approx 61.9$. The spread of the histogram estimates $\frac{\sigma}{\sqrt{n}}$, so $\frac{\sigma}{\sqrt{25}} \approx 2.02$ gives $\sigma \approx 2.02 \times \sqrt{25} = 2.02 \times 5 = 10.1$.

Q14. Require $\frac{\sigma}{\sqrt{n}} = \frac{18}{\sqrt{n}} \le 3$. Then $\sqrt{n} \ge \frac{18}{3} = 6$, so $n \ge 36$. The smallest such sample size is $n = 36$ (which gives a spread of exactly 3).

Q15. Each histogram has standard deviation $\frac{42}{\sqrt{n}}$. For Student A ($n = 4$): $\frac{42}{2} = 21$. For Student B ($n = 36$): $\frac{42}{6} = 7$. Student B's histogram is narrower; the spread falls from 21 to 7, a factor of $\frac{1}{3}$ (matching $\frac{1}{\sqrt{36/4}} = \frac{1}{3}$). The population is normal, so $\acute{X}$ is exactly normal for each student. For A, $z = \frac{220-210}{21} \approx 0.476$, so $P(\acute{X} > 220) = P(Z > 0.476) = 1 - 0.6830 = 0.3170$; for B, $z = \frac{220-210}{7} \approx 1.429$, so $P(\acute{X} > 220) = P(Z > 1.429) = 1 - 0.9234 = 0.0766$. The narrower histogram places far less of its area above 220.

Q16. The population is not normal, but $n = 64$ is large, so by the central limit theorem $\acute{X} \approx N(12, \sigma_{\acute{X}}^2)$ with $\sigma_{\acute{X}} = \frac{4}{\sqrt{64}} = \frac{4}{8} = 0.5$. Standardising, $z = \frac{13-12}{0.5} = 2$, so $P(\acute{X} > 13) = P(Z > 2) = 1 - 0.9772 = 0.0228$.

Q17. The histogram's standard deviation is $\frac{\sigma}{\sqrt{n}}$, so $\frac{\sigma}{\sqrt{49}}=1.4$ gives $\sigma=1.4\times 7=9.8$. For $n=196$, $\frac{\sigma}{\sqrt{196}}=\frac{9.8}{14}=0.7$. Quadrupling n (from 49 to 196) halves the spread, from 1.4 to 0.7.

Q18. (a) The population is not normal, but $n=36\geq 30$ is large, so by the central limit theorem $\acute{X}$ is approximately normal. (b) Its centre is $\mu=20$ and its standard deviation is $\frac{\sigma}{\sqrt{n}}=\frac{9}{\sqrt{36}}=\frac{9}{6}=1.5$, so $\acute{X}\approx N(20,1.5^2)$. (c) Standardising, $z=\frac{22-20}{1.5}\approx 1.333$, so $P(\acute{X}>22)=P(Z>1.333)=1-0.9088=0.0912$. (d) The endpoints standardise to $z=\frac{19-20}{1.5}\approx -0.667$ and $z=\frac{21-20}{1.5}\approx 0.667$, so

$$P(19<\acute{X}<21)=P(-0.667<Z<0.667)=0.7475-0.2525=0.4950.$$

Chapter 11 Linear combinations of random variables and the sample mean — 11F The distribution of the sample mean

Theory

Suppose $X_1, X_2, \dots, X_n$ is a **random sample** of size n drawn from a population with mean μ and standard deviation σ: the X_i are **independent** and each has the same distribution as the population variable X. The **sample mean**

$$\acute{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is itself a random variable — a different sample gives a different value of $\acute{X}$ — and its probability distribution is called the **sampling distribution of the sample mean**. Earlier sections established, from the rules for linear combinations, the two summary numbers of this distribution:

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

So the sample mean is centred on the **same** value μ as the population, but it is **less spread out**: its standard deviation $\sigma / \sqrt{n}$ (sometimes called the **standard error** of the mean) shrinks as the sample size grows. This section identifies the full distribution of $\acute{X}$ and uses it to compute probabilities.

The distribution of $\acute{X}$ for a normal population.

If the population itself is **normal**, $X \sim N(\mu, \sigma^2)$, then $\acute{X}$ is a linear combination of independent normal variables, and a linear combination of independent normal variables is normal. Hence the sample mean is **exactly normal**:

$$X \sim N(\mu, \sigma^2) \Rightarrow \acute{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

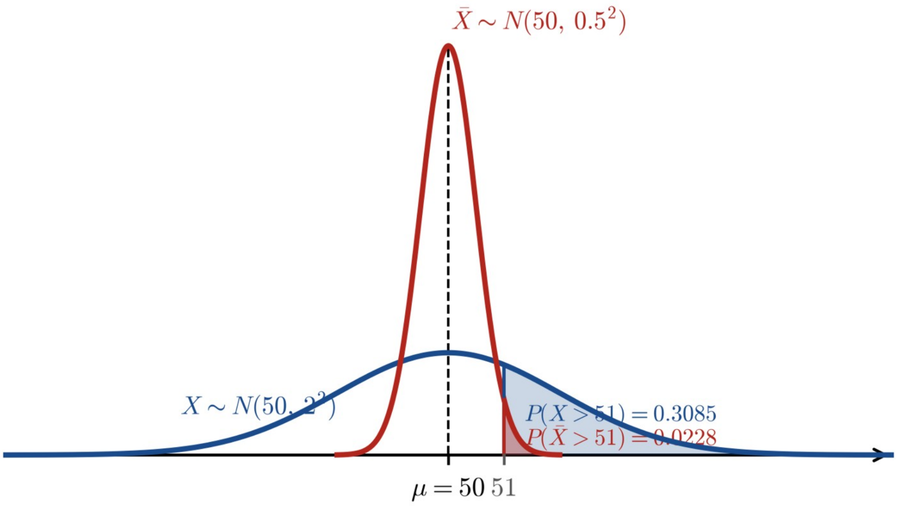
This holds for **every** sample size n, however small. The curve of $\acute{X}$ has the same centre μ as the population but is taller and narrower, by the factor $1 / \sqrt{n}$.

Large samples — the central limit theorem.

What if the population is **not** normal — or its shape is unknown? Remarkably, the sample mean is still **approximately normal**, provided the sample is large enough. This is the **central limit theorem** (CLT):

> For a random sample of size n from **any** population with mean μ and standard deviation σ, the sample mean is approximately normal for large n,
>
> $$\acute{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right),$$
>
> and the approximation improves as n increases.

A common working rule is that $n \geq 30$ is “large enough” for the approximation to be good; the more skewed the population, the larger the n required. (We state the theorem and apply it; no proof is expected at this level.) The two cases combine into a single practical rule: **treat $\acute{X}$ as normal** with mean μ and standard deviation $\sigma / \sqrt{n}$ — exactly when the population is normal, and approximately when the sample is large. The diagram below shows a population variable $X \sim N(50, 2^2)$ together with the sample mean $\acute{X} \sim N(50, 0.5^2)$ for a sample of size $n = 16$: same centre, but the sample mean is four times narrower because $\sigma / \sqrt{n} = 2 / \sqrt{16} = 0.5$.



Computing probabilities about $\bar{X}$.

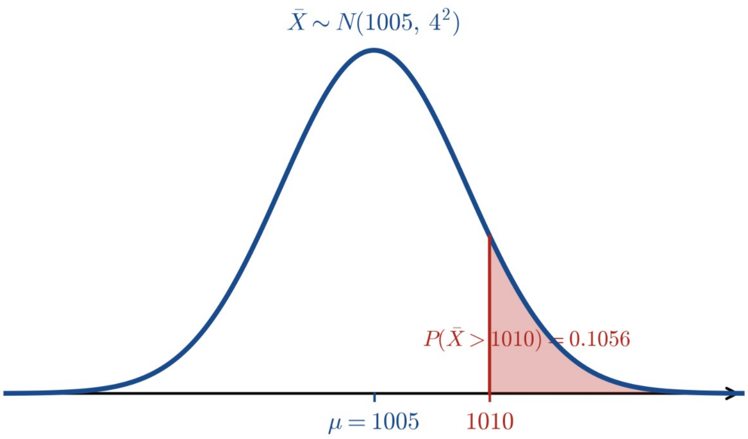
Because $\bar{X}$ is (exactly or approximately) normal, probabilities about it are found by **standardising**, exactly as for any normal variable — but with the standard deviation $\sigma/\sqrt{n}$ in place of σ:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

A boundary value c for $\bar{X}$ maps to the z-score $z = \dfrac{c-\mu}{\sigma/\sqrt{n}}$, and then

$$P(\bar{X} > c) = P\left(Z > \frac{c-\mu}{\sigma/\sqrt{n}}\right),\ P(a < \bar{X} < b) = P\left(\frac{a-\mu}{\sigma/\sqrt{n}} < Z < \frac{b-\mu}{\sigma/\sqrt{n}}\right).$$

The required standard-normal areas are read from technology. The picture shows the sampling distribution $\bar{X} \sim N(1005, 4^2)$ of Example 1, with the upper-tail probability $P(\bar{X} > 1010) = 0.1056$ shaded.



The sample size needed for a given accuracy.

A frequent design question is: *how large must the sample be so that $\bar{X}$ is within a given distance d of the population mean μ, with a stated probability p?* We require

$$P(|\bar{X} - \mu| < d) = p.$$

Standardising, $|\bar{X} - \mu| < d$ is the same as $|Z| < \dfrac{d}{\sigma/\sqrt{n}} = \dfrac{d\sqrt{n}}{\sigma}$. If z^* is the positive value with $P(-z^* < Z < z^*) = p$, that is

$$z^* = \Phi^{-1}\left(\frac{1+p}{2}\right),$$

then we need $\dfrac{d\sqrt{n}}{\sigma} = z^*$, so

$$n = \left(\frac{z^* \sigma}{d}\right)^2.$$

Because n must be a whole number and a larger sample can only sharpen the estimate, we always **round up** to the next integer. (Halving the required distance d multiplies the necessary sample size by four, since n depends on $1/d^2$.)

A single observation versus the sample mean.

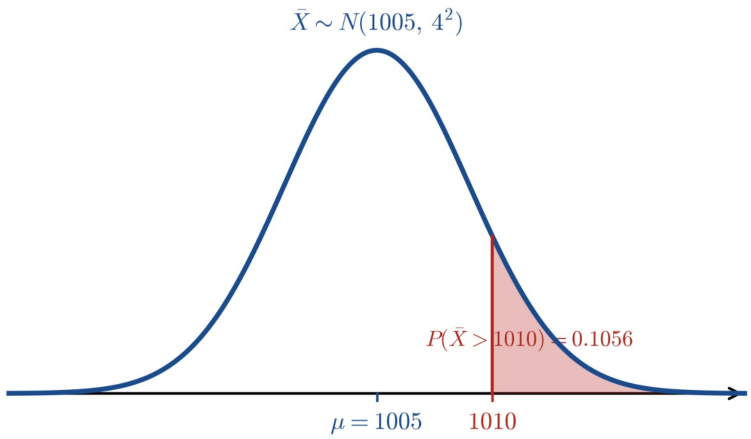
It is worth contrasting a probability about the **sample mean** $\bar{X}$ with the same statement about a **single observation** X. Both are centred on μ, but X has standard deviation σ while $\bar{X}$ has the much smaller $\sigma/\sqrt{n}$. So a fixed departure from μ is far **less** likely for the sample mean than for one observation: averaging cancels the extremes. For instance, with $X \sim N(50, 2^2)$ a single value exceeds 51 with probability $P(X > 51) = 0.3085$, whereas the mean of $n = 16$ values exceeds 51 with probability only $P(\bar{X} > 51) = 0.0228$. Always check whether a question asks about **one** measurement (use σ) or about the **mean of** n measurements (use $\sigma/\sqrt{n}$).

Worked examples

Example 1 — scaffolded

A machine fills bags of sugar so that the mass X grams of a bag is normally distributed, $X \sim N(1005, 12^2)$. A quality inspector weighs a random sample of $n = 9$ bags and records the sample mean $\bar{X}$. (a) State the distribution of $\bar{X}$. (b) Find $P(\bar{X} > 1010)$.

Working	Comment
$E(\bar{X}) = \mu = 1005$; $sd(\bar{X}) = \dfrac{\sigma}{\sqrt{n}} = \dfrac{12}{\sqrt{9}} = \dfrac{12}{3} = 4$.	Sample-mean mean and standard deviation.
The population is normal, so $\bar{X}$ is **exactly** normal: $\bar{X} \sim N(1005, 4^2)$.	For a normal population $\bar{X} \sim N(\mu, \sigma^2/n)$ for any n.
Standardise the boundary: $z = \dfrac{1010 - 1005}{4} = \dfrac{5}{4} = 1.25$.	$z = \dfrac{c - \mu}{\sigma/\sqrt{n}}$.
$P(\bar{X} > 1010) = P(Z > 1.25) = 1 - P(Z \le 1.25) = 1 - 0.89$.	From technology, $P(Z \le 1.25) = 0.8944$.



Example 2 — scaffolded

At a busy café the time X minutes a customer waits to be served has mean $\mu = 5$ and standard deviation $\sigma = 2.4$; the distribution is right-skewed, not normal. A random sample of $n = 36$ customers is taken and the sample mean waiting time $\acute{X}$ is recorded. (a) Explain why $\acute{X}$ is approximately normal, and state its distribution. (b) Find $P(4.5 < \acute{X} < 5.5)$.

Working	Comment
$n = 36 \geq 30$ is large, so by the central limit theorem $\acute{X}$ is approximately normal.	The CLT applies even though the population is not normal.
$E(\acute{X}) = 5$; $sd(\acute{X}) = \frac{2.4}{\sqrt{36}} = \frac{2.4}{6} = 0.4$, so $\acute{X} \approx N(5, 0.4^2)$.	Mean μ, standard error $\sigma / \sqrt{n}$.
Standardise both endpoints: $z_1 = \frac{4.5 - 5}{0.4} = -1.25$, $z_2 = \frac{5.5 - 5}{0.4} = 1.25$.	Symmetric interval about the mean.
$P(4.5 < \acute{X} < 5.5) = P(-1.25 < Z < 1.25) = 0.89435 - 0.1$.	$P(Z \leq 1.25) = 0.89435$, $P(Z \leq -1.25) = 0.10565$; subtract first, round last.

Example 3 — unscaffolded

Bottles filled by a production line have volume with standard deviation $\sigma = 6$ mL. A technician wants the mean volume $\acute{X}$ of a random sample to be within 2 mL of the true mean μ with probability 0.95. Find the smallest sample size n needed.

The requirement is $P(|\acute{X} - \mu| < 2) = 0.95$. Standardising, this is $P(|Z| < z^*) = 0.95$, so $z^* = \Phi^{-1}\left(\frac{1 + 0.95}{2}\right) = \Phi^{-1}(0.975) = 1.959964$. Setting the standardised distance equal to z^*,

$$\frac{d\sqrt{n}}{\sigma} = z^* \Rightarrow n = \left(\frac{z^* \sigma}{d}\right)^2 = \left(\frac{1.959964 \times 6}{2}\right)^2 = (5.87989)^2 = 34.573.$$

Since n must be a whole number and the probability requirement is a *minimum*, we round **up**.

$\therefore$ the smallest sample size is $n = 35$ bottles.

Example 4 — unscaffolded

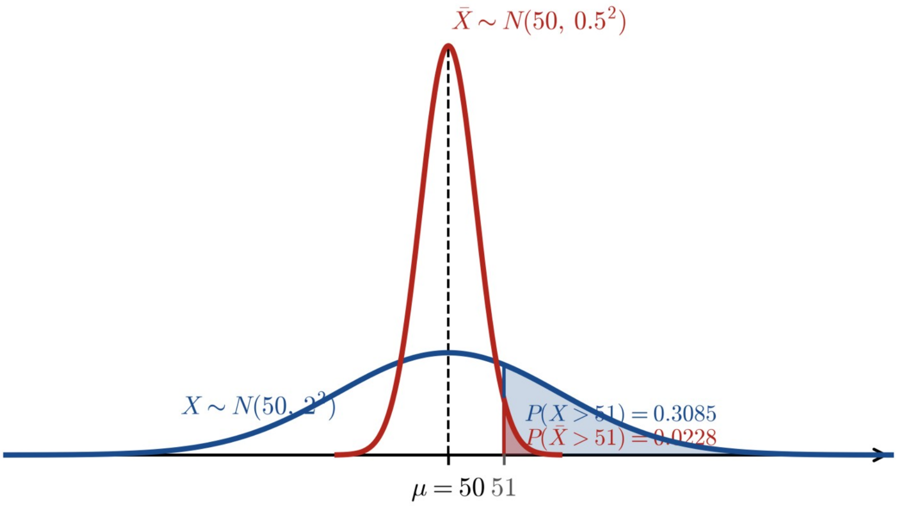
The lengths of manufactured components are modelled by $X \sim N(50, 2^2)$ mm. (a) Find the probability that a single component is longer than 51 mm. (b) Find the probability that the mean length of a random sample of $n = 16$ components exceeds 51 mm, and compare it with part (a).

For a **single** component the standard deviation is $\sigma = 2$, so $z = \frac{51 - 50}{2} = 0.5$ and

$$P(X > 51) = P(Z > 0.5) = 1 - 0.6915 = 0.3085.$$

For the **sample mean** of $n = 16$ the standard deviation is $\frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{16}} = 0.5$, so $\acute{X} \sim N(50, 0.5^2)$ and $z = \frac{51 - 50}{0.5} = 2$, giving

$$P(\acute{X} > 51) = P(Z > 2) = 1 - 0.9772 = 0.0228.$$



$\therefore P(X > 51) = 0.3085$ but $P(\bar{X} > 51) = 0.0228$: the sample mean is far less likely to lie 1 mm above μ, because its standard deviation is four times smaller than a single component's.

Practice questions

Scaffolded

Q1. Battery lifetimes are modelled by $X \sim N(40, 6^2)$ hours. A random sample of $n = 9$ batteries is tested and the sample mean lifetime $\bar{X}$ is found. (a) Find $E(\bar{X})$ and $sd(\bar{X})$. (b) State the distribution of $\bar{X}$. (c) Standardise the value $\bar{X} = 43$. (d) Find $P(\bar{X} > 43)$.

Q2. The mass of cereal in a box is modelled by $X \sim N(500, 8^2)$ grams. A random sample of $n = 16$ boxes is weighed. (a) Find $sd(\bar{X})$. (b) State the distribution of $\bar{X}$. (c) Standardise the endpoints 496 and 504. (d) Find $P(496 < \bar{X} < 504)$.

Q3. The number of emails a worker receives per day has mean $\mu = 20$ and standard deviation $\sigma = 9$; the distribution is not normal. A random sample of $n = 81$ days is taken. (a) Explain why $\bar{X}$ is approximately normal. (b) State the mean and standard deviation of $\bar{X}$. (c) Standardise the value $\bar{X} = 22$. (d) Find $P(\bar{X} > 22)$.

Q4. A population has standard deviation $\sigma = 10$. A researcher wants the sample mean $\bar{X}$ to be within 3 of the population mean μ with probability 0.95. (a) Write the requirement as a probability statement about $|\bar{X} - \mu|$. (b) Find z^* such that $P(-z^* < Z < z^*) = 0.95$. (c) Solve $\dfrac{z^* \sigma}{\sqrt{n}} = 3$ for n. (d) State the smallest sample size.

Q5. The mass of an apple is modelled by $X \sim N(150, 20^2)$ grams. (a) Find the probability that a single apple has mass more than 160 g. (b) For a random sample of $n = 25$ apples, find $sd(\bar{X})$. (c) Find $P(\bar{X} > 160)$. (d) Comment on the comparison between (a) and (c).

Q6. A coffee machine dispenses volumes modelled by $X \sim N(250, 5^2)$ mL. A random sample of $n = 4$ cups is measured. (a) Find $sd(\bar{X})$. (b) State the distribution of $\bar{X}$. (c) Standardise the value $\bar{X} = 247$. (d) Find $P(\bar{X} < 247)$.

Unscaffolded

Q7. Bags of flour have mass $X \sim N(1000, 15^2)$ grams. For a random sample of $n = 9$ bags, find $P(\bar{X} > 1007)$.

Q8. The duration X of a phone call at a call centre has mean $\mu = 3.5$ minutes and standard deviation $\sigma = 1.2$ minutes, with a skewed distribution. For a random sample of $n = 36$ calls, find $P(3.3 < \bar{X} < 3.7)$.

Q9. A filling process has standard deviation $\sigma = 8$ mL. Find the smallest sample size n so that the sample mean is within 2 mL of the population mean with probability 0.90.

Q10. Examination marks are modelled by $X \sim N(72, 10^2)$. (a) Find the probability that a single student scores below 66. (b) Find the probability that the mean mark of a class of $n = 16$ students is below 66, and compare it with part (a).

Q11. Drink volumes are modelled by $X \sim N(330, 4^2)$ mL. For a random sample of $n = 16$ cans, find $P(329 < \bar{X} < 332)$.

Q12. A machine part has a dimension with mean $\mu = 15$ and standard deviation $\sigma = 4$; the distribution is unknown. For a random sample of $n = 64$ parts, find $P(\bar{X} > 15.75)$.

Q13. A measuring instrument has standard deviation $\sigma = 5$. Find the smallest sample size n so that the sample mean is within 1 of the true mean with probability 0.99.

Q14. Resistor values are modelled by $X \sim N(100, 5^2)$ ohms. (a) Find the probability that a single resistor exceeds 102 ohms. (b) For a random sample of $n = 16$ resistors, find $P(\bar{X} > 102)$, and compare it with (a).

Q15. Bags of rice have mass $X \sim N(500, 10^2)$ grams. For a random sample of $n = 25$ bags, find the value c such that $P(\bar{X} > c) = 0.10$.

Q16. Components have weight with standard deviation $\sigma = 2$ g. A quality manager wants the mean weight of a random sample to be within 0.5 g of the true mean weight with probability 0.95. Find the smallest sample size required.

Q17. Steel rods have length $X \sim N(20, 1.5^2)$ cm. (a) For a random sample of $n = 9$ rods, find $P(19.5 < \bar{X} < 20.5)$. (b) For a random sample of $n = 36$ rods, find $P(19.5 < \bar{X} < 20.5)$, and comment on the effect of the larger sample.

Q18. The fill of a sugar bag is modelled by $X \sim N(1000, 8^2)$ grams. (a) Find the probability that a single bag contains less than 990 g. (b) For a carton of $n = 16$ bags, find the probability that the mean fill is less than 996 g. (c) Find the smallest number of bags n so that the sample mean is within 3 g of 1000 g with probability 0.95.

Answers

Q1. (a) $E(\acute{X})=40$, $sd(\acute{X})=2$; (b) $\acute{X}\sim N(40,2^2)$; (c) $z=1.5$; (d) $P(\acute{X}>43)=0.0668$. **Q2.** (a) $sd(\acute{X})=2$; (b) $\acute{X}\sim N(500,2^2)$; (c) $z=-2$ and $z=2$; (d) $P(496<\acute{X}<504)=0.9545$. **Q3.** (a) $n=81$ is large, so the central limit theorem applies; (b) mean 20, sd 1; (c) $z=2$; (d) $P(\acute{X}>22)=0.0228$. **Q4.** (a) $P(|\acute{X}-\mu|<3)=0.95$; (b) $z^*=1.959964$; (c) $n=42.68$; (d) $n=43$. **Q5.** (a) $P(X>160)=0.3085$; (b) $sd(\acute{X})=4$; (c) $P(\acute{X}>160)=0.0062$; (d) the sample mean is much less likely to exceed 160 (smaller spread). **Q6.** (a) $sd(\acute{X})=2.5$; (b) $\acute{X}\sim N(250,2.5^2)$; (c) $z=-1.2$; (d) $P(\acute{X}<247)=0.1151$. **Q7.** $P(\acute{X}>1007)=0.0808$. **Q8.** $P(3.3<\acute{X}<3.7)=0.6827$. **Q9.** $n=44$. **Q10.** (a) $P(X<66)=0.2743$; (b) $P(\acute{X}<66)=0.0082$ — much smaller for the sample mean. **Q11.** $P(329<\acute{X}<332)=0.8186$. **Q12.** $P(\acute{X}>15.75)=0.0668$. **Q13.** $n=166$. **Q14.** (a) $P(X>102)=0.3446$; (b) $P(\acute{X}>102)=0.0548$ — much smaller for the sample mean. **Q15.** $c=502.56$ g. **Q16.** $n=62$. **Q17.** (a) $P(19.5<\acute{X}<20.5)=0.6827$; (b) 0.9545 — the larger sample concentrates $\acute{X}$ nearer μ. **Q18.** (a) $P(X<990)=0.1056$; (b) $P(\acute{X}<996)=0.0228$; (c) $n=28$.

Fully-worked solutions

Q1. The population is normal, so $\acute{X}$ is exactly normal. (a) $E(\acute{X})=\mu=40$ and $sd(\acute{X})=\frac{6}{\sqrt{9}}=\frac{6}{3}=2$. (b) $\acute{X}\sim N(40,2^2)$. (c) $z=\frac{43-40}{2}=\frac{3}{2}=1.5$. (d) $P(\acute{X}>43)=P(Z>1.5)=1-0.9332=0.0668$.

Q2. (a) $sd(\acute{X})=\frac{8}{\sqrt{16}}=\frac{8}{4}=2$. (b) The population is normal, so $\acute{X}\sim N(500,2^2)$. (c) $z_1=\frac{496-500}{2}=-2$ and $z_2=\frac{504-500}{2}=2$. (d) $P(496<\acute{X}<504)=P(-2<Z<2)=0.97725-0.02275=0.9545$.

Q3. (a) The population is not normal, but $n=81$ is large, so by the central limit theorem $\acute{X}$ is approximately normal. (b) $E(\acute{X})=20$ and $sd(\acute{X})=\frac{9}{\sqrt{81}}=\frac{9}{9}=1$, so $\acute{X}\approx N(20,1^2)$. (c) $z=\frac{22-20}{1}=2$. (d) $P(\acute{X}>22)=P(Z>2)=1-0.9772=0.0228$.

Q4. (a) The requirement is $P(|\acute{X}-\mu|<3)=0.95$. (b) The symmetric interval containing central probability 0.95 has $z^*=\Phi^{-1}(0.975)=1.959964$. (c) Setting $\frac{z^*\sigma}{\sqrt{n}}=3$ gives $\sqrt{n}=\frac{z^*\sigma}{3}=\frac{1.959964\times 10}{3}=6.5332$, so $n=6.5332^2=42.68$. (d) Rounding up, the smallest sample size is $n=43$.

Q5. (a) For a single apple, $z=\frac{160-150}{20}=0.5$, so $P(X>160)=P(Z>0.5)=1-0.6915=0.3085$. (b) $sd(\acute{X})=\frac{20}{\sqrt{25}}=\frac{20}{5}=4$. (c) $z=\frac{160-150}{4}=2.5$, so $P(\acute{X}>160)=P(Z>2.5)=1-0.9938=0.0062$. (d) The sample mean is far less likely to exceed 160 g than a single apple, because its standard deviation (4) is much smaller than the population's (20): averaging cancels the extremes.

Q6. (a) $sd(\acute{X})=\frac{5}{\sqrt{4}}=\frac{5}{2}=2.5$. (b) The population is normal, so $\acute{X}\sim N(250,2.5^2)$ exactly (valid even though $n=4$ is small). (c) $z=\frac{247-250}{2.5}=-1.2$. (d) $P(\acute{X}<247)=P(Z<-1.2)=0.1151$.

Q7. $sd(\acute{X})=\frac{15}{\sqrt{9}}=\frac{15}{3}=5$, and since the population is normal, $\acute{X}\sim N(1000,5^2)$. Standardising, $z=\frac{1007-1000}{5}=1.4$, so $P(\acute{X}>1007)=P(Z>1.4)=1-0.9192=0.0808$.

Q8. The distribution is skewed, but $n=36\geq 30$ is large, so by the central limit theorem $\acute{X}$ is approximately normal with mean 3.5 and $sd(\acute{X})=\frac{1.2}{\sqrt{36}}=\frac{1.2}{6}=0.2$. The endpoints standardise to $z_1=\frac{3.3-3.5}{0.2}=-1$ and $z_2=\frac{3.7-3.5}{0.2}=1$, so $P(3.3<\acute{X}<3.7)=P(-1<Z<1)=0.84134-0.15866=0.6827$.

Q9. For "within 2 mL with probability 0.90", $z^*=\Phi^{-1}\left(\frac{1+0.90}{2}\right)=\Phi^{-1}(0.95)=1.644854$. Then

$$n=\left(\frac{z^*\sigma}{d}\right)^2=\left(\frac{1.644854\times 8}{2}\right)^2=(6.5794)^2=43.29.$$

Rounding up, $n=44$.

Q10. (a) For a single student, $z=\frac{66-72}{10}=-0.6$, so $P(X<66)=P(Z<-0.6)=0.2743$. (b) For the class mean of $n=16$, $sd(\acute{X})=\frac{10}{\sqrt{16}}=2.5$, so $z=\frac{66-72}{2.5}=-2.4$ and $P(\acute{X}<66)=P(Z<-2.4)=0.0082$. The class mean falling below 66 is much less likely than a single student doing so, because the sample mean has one-quarter of the population's spread.

Q11. $sd(\acute{X})=\frac{4}{\sqrt{16}}=1$, and the population is normal, so $\acute{X}\sim N(330,1^2)$. The endpoints standardise to $z_1=\frac{329-330}{1}=-1$ and $z_2=\frac{332-330}{1}=2$, so $P(329<\acute{X}<332)=P(-1<Z<2)=0.97725-0.15866=0.8186$.

Q12. The distribution is unknown, but $n=64$ is large, so by the central limit theorem $\acute{X}$ is approximately normal with mean 15 and $sd(\acute{X})=\frac{4}{\sqrt{64}}=\frac{4}{8}=0.5$. Standardising, $z=\frac{15.75-15}{0.5}=1.5$, so $P(\acute{X}>15.75)=P(Z>1.5)=1-0.9332=0.0668$.

Q13. For "within 1 with probability 0.99", $z^*=\Phi^{-1}\left(\frac{1+0.99}{2}\right)=\Phi^{-1}(0.995)=2.575829$. Then

$$n=\left(\frac{z^*\sigma}{d}\right)^2=\left(\frac{2.575829\times 5}{1}\right)^2=(12.8791)^2=165.87.$$

Rounding up, $n=166$.

Q14. (a) For a single resistor, $z=\frac{102-100}{5}=0.4$, so $P(X>102)=P(Z>0.4)=1-0.6554=0.3446$. (b) For the sample mean of $n=16$, $sd(\acute{X})=\frac{5}{\sqrt{16}}=1.25$, so $z=\frac{102-100}{1.25}=1.6$ and $P(\acute{X}>102)=P(Z>1.6)=1-0.9452=0.0548$. The sample mean is far less likely to exceed 102 ohms than a single resistor, because its standard deviation is four times smaller.

Q15. The population is normal, so $\acute{X}\sim N(500,2^2)$ with $sd(\acute{X})=\frac{10}{\sqrt{25}}=2$. We need $P(\acute{X}>c)=0.10$, i.e. $P(\acute{X}\leq c)=0.90$, so c is the 90th percentile. The standard-normal value with $P(Z\leq z)=0.90$ is $z=\Phi^{-1}(0.90)=1.281552$. Then

$$c=\mu+z\,sd(\bar{X})=500+1.281552\times 2=500+2.563=502.56\text{ g}.$$

Q16. For "within 0.5 g with probability 0.95", $z^*=\Phi^{-1}(0.975)=1.959964$. Then

$$n=\left(\frac{z^*\sigma}{d}\right)^2=\left(\frac{1.959964\times 2}{0.5}\right)^2=(7.8399)^2=61.46.$$

Rounding up, $n=62$.

Q17. The population is normal, so $\bar{X}$ is exactly normal in each case. (a) For $n=9$, $sd(\bar{X})=\frac{1.5}{\sqrt{9}}=\frac{1.5}{3}=0.5$; the endpoints standardise to $z=\frac{19.5-20}{0.5}=-1$ and $z=\frac{20.5-20}{0.5}=1$, so $P(19.5<\bar{X}<20.5)=P(-1<Z<1)=0.6827$. (b) For $n=36$, $sd(\bar{X})=\frac{1.5}{\sqrt{36}}=\frac{1.5}{6}=0.25$; the endpoints standardise to $z=\frac{19.5-20}{0.25}=-2$ and $z=\frac{20.5-20}{0.25}=2$, so $P(19.5<\bar{X}<20.5)=P(-2<Z<2)=0.9545$. The larger sample makes $\bar{X}$ more tightly concentrated about μ, so the mean is more likely to fall within 0.5 cm of 20.

Q18. (a) For a single bag, $z=\frac{990-1000}{8}=-1.25$, so $P(X<990)=P(Z<-1.25)=0.1056$. (b) For the carton mean of $n=16$, $sd(\bar{X})=\frac{8}{\sqrt{16}}=2$, so $z=\frac{996-1000}{2}=-2$ and $P(\bar{X}<996)=P(Z<-2)=0.0228$. (c) For "within 3 g with probability 0.95", $z^*=\Phi^{-1}(0.975)=1.959964$, so

$$n=\left(\frac{z^*\sigma}{d}\right)^2=\left(\frac{1.959964\times 8}{3}\right)^2=(5.2266)^2=27.32,$$

which rounds up to $n=28$ bags.