Cambridge AS & A Level · Mathematics 9709 · Notes

# Sampling and estimation A2

Read online at **pastopic.com/cambridge/mathematics-9709/notes/sampling-and-estimation**

Last checked 29 September 2026

Read first: **Representation of data**, **The normal distribution**, **Linear combinations of random variables**

© 2026 Pastopic. Free for personal study. Please share the link, not the file.

## What you need to know

By the end of this topic you should be able to:

1. **Tell a sample from a population**, and explain why a sample has to be chosen at random.
2. **Say in simple words why a sampling method is not good enough**, and use random numbers (or dice, coins or a calculator) to choose a random sample. You don't need named methods such as quota or stratified sampling.
3. **Treat the sample mean $\bar{X}$ as a random variable** with $E(\bar{X}) = \mu$ and $\mathrm{Var}(\bar{X}) = \dfrac{\sigma^2}{n}$.
4. **Use the fact that $\bar{X}$ is exactly normal when $X$ is normal.**
5. **Use the Central Limit Theorem (CLT)** when $X$ is not known to be normal: for a large sample, $\bar{X}$ is approximately normal. An informal understanding is enough.
6. **Calculate unbiased estimates of the population mean and variance** from a sample, from the raw values or from totals such as $\Sigma x$ and $\Sigma x^2$, and know in simple terms what "unbiased" means.
7. **Find and interpret a confidence interval for a population mean**, when the population is normal with known variance, or when the sample is large.
8. **Find an approximate confidence interval for a population proportion** from a large sample.

Confidence intervals that use the $t$-distribution, and estimates from two samples, are Further Mathematics, not this course.

## Understand it

### Populations and samples

The **population** is everything you want to know about: every packet a factory fills, every student in a college. A **sample** is the part of it you actually measure. You take a sample because measuring everything would take too long or cost too much, or because measuring destroys the item (you can't test every fuse until it blows and still sell it).

A **random sample** is one where every member of the population has the same chance of being chosen, and every choice is independent of the others. Only then do the sample's results tell you about the whole population. A sample that leaves some members out, or makes some more likely than others, is **biased**: asking only people at a bus stop what they think of the bus service, or measuring something only in one season, gives answers that may not be typical of everyone.

## Choosing a random sample

Number the population $1$ to $N$ and use random numbers to pick the numbers.

- **Random digits.** If $N = 350$, read the digits in groups of three: $271, 804, 093, \ldots$ Keep $271$, ignore $804$ (there is no member $804$), keep $093$ (member $93$), and ignore any number already chosen or equal to $000$.
- **Scaling a calculator random number.** A calculator gives $u$ from $0.000$ to $0.999$. Multiply by $N$ and round up. With $N = 600$, $u = 0.358$ gives $214.8$, so member $215$. Member $k$ is chosen by every $u$ with $\frac{k-1}{N} < u \leqslant \frac{k}{N}$. There are $1000$ values of $u$ for $600$ members, so member $1$ is chosen only by $u = 0.001$, but member $2$ by $0.002$ and $0.003$. The members don't have equal chances, so this method is **not random**. (And $u = 0.000$ gives no member at all.)
- **Dice and coins.** A method is random only if every choice is **equally likely**. To pick one of three people with one dice: $1$ or $2$ gives the first, $3$ or $4$ the second, $5$ or $6$ the third. Counting heads on two coins is not fair: $P(0 \text{ heads}) = \frac{1}{4}$ but $P(1 \text{ head}) = \frac{1}{2}$.

## The sample mean is a random variable

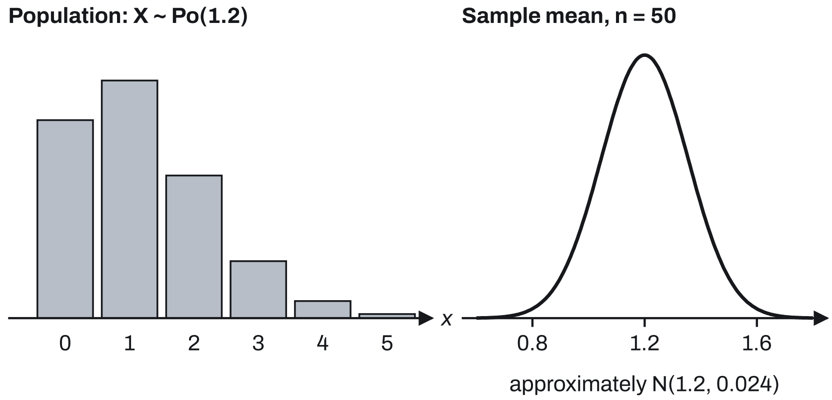
Take a sample of $n$ values from a population with mean $\mu$ and variance $\sigma^2$, and work out their mean $\bar{x}$. Take another sample and you get a different $\bar{x}$. So the sample mean $\bar{X}$ is a random variable with its own distribution. Using the rules for combining independent variables,

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}, \qquad E(\bar{X}) = \frac{n\mu}{n} = \mu, \qquad \operatorname{Var}(\bar{X}) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

The sample mean is centred on the true mean, and its spread gets smaller as $n$ grows: a bigger sample gives a more reliable mean. Its standard deviation is $\frac{\sigma}{\sqrt{n}}$, so to halve the spread you need four times the sample.

## Which distribution: normal, or the Central Limit Theorem

- **If $X$ is normal**, then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ **exactly**, for any $n$, even $n = 5$. No CLT is needed.
- **If $X$ is not normal, or its distribution is unknown**, the **Central Limit Theorem** says that for a large sample (say $n \geqslant 30$), $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ **approximately**, whatever the shape of the population.




So a Poisson or binomial variable, a variable with an unknown distribution, or a combination such as $2X + Y$ of non-normal variables, all need the CLT when you use a normal distribution for their sample mean. For $X \sim \mathrm{Po}(1.2)$ and $n = 50$: $\bar{X} \sim N(1.2,\ 0.024)$ approximately, since $\mathrm{Var}(X) = 1.2$ and $1.2 \div 50 = 0.024$.

## Unbiased estimates

Usually you don't know $\mu$ or $\sigma^2$, so you estimate them from the sample. An estimate is **unbiased** if the method gives the right answer **on average**: single samples come out too high or too low, but over many samples the errors cancel.

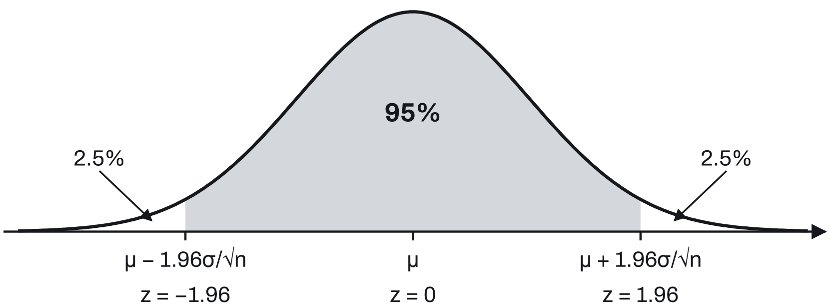
- The sample mean $\bar{x} = \dfrac{\Sigma x}{n}$ is an unbiased estimate of $\mu$.
- The ordinary variance of the sample, $\dfrac{\Sigma(x - \bar{x})^2}{n}$, is on average **too small**, because the values are spread around their own mean $\bar{x}$ rather than the true $\mu$. Dividing by $n - 1$ instead fixes this:

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} = \frac{1}{n - 1}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right).$$

A quick equivalent: $s^2 = \dfrac{n}{n - 1} \times \left(\dfrac{\Sigma x^2}{n} - \bar{x}^2\right)$, the sample variance multiplied by $\dfrac{n}{n - 1}$.

## Confidence intervals for a mean

A single estimate $\bar{x}$ is almost never exactly $\mu$. A **confidence interval** gives a range instead. Start from $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$: $95\%$ of sample means lie within $1.96$ standard deviations of $\mu$.



Turn that round: if $\bar{x}$ is within $1.96\frac{\sigma}{\sqrt{n}}$ of $\mu$, then $\mu$ is within $1.96\frac{\sigma}{\sqrt{n}}$ of $\bar{x}$. So the $95\%$ confidence interval is

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}.$$

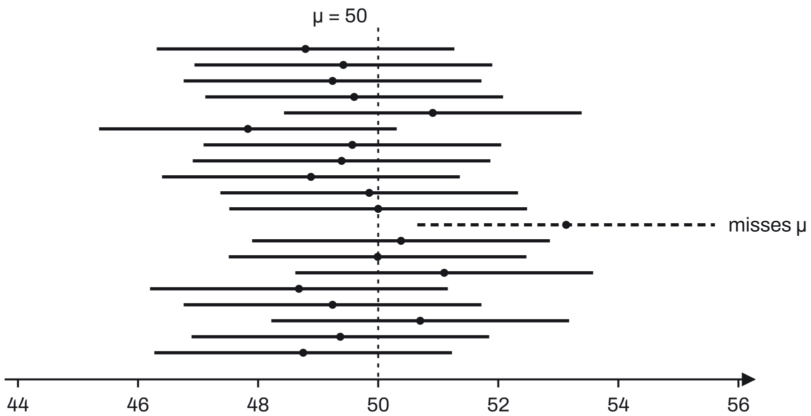
For other levels, change the $z$-value. For an $\alpha\%$ interval, $z$ cuts off $\frac{100-\alpha}{2}\%$ in each tail, so $\Phi(z) = \frac{1}{2}\left(1 + \frac{\alpha}{100}\right)$:

| **Level** | $90\%$ | $92\%$ | $94\%$ | $95\%$ | $96\%$ | $98\%$ | $99\%$ |
|---|---|---|---|---|---|---|---|
| $\Phi(z)$ | $0.95$ | $0.96$ | $0.97$ | $0.975$ | $0.98$ | $0.99$ | $0.995$ |
| $z$ | $1.645$ | $1.751$ | $1.881$ | $1.960$ | $2.054$ | $2.326$ | $2.576$ |

The values for $90\%$, $95\%$, $98\%$ and $99\%$ are in the critical values table; find the others by reading the normal table backwards.

If $\sigma$ is unknown and the sample is large, use the unbiased estimate $s$ in its place, and the CLT covers a population that is not normal.

**What the interval means.** $\mu$ is a fixed number; it is the interval that changes from sample to sample. A $95\%$ interval is made by a method that catches $\mu$ in $95\%$ of samples. It is **not** a range containing $95\%$ of the population's values.




Because each interval catches $\mu$ with probability $0.95$, independently, the number of intervals out of $r$ that contain $\mu$ is $B(r, 0.95)$.

**Width.** The width is $2z\frac{\sigma}{\sqrt{n}}$. A higher confidence level (bigger $z$) gives a wider interval; a bigger sample gives a narrower one.

## Confidence intervals for a proportion

If $x$ of a large sample of $n$ have some property, estimate the population proportion $p$ by $\hat{p} = \frac{x}{n}$. The sample proportion is approximately $N\left(p, \frac{p(1-p)}{n}\right)$, so the approximate interval is

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

It is **approximate** twice over: the normal distribution only approximates the distribution of $\hat{p}$, and $\hat{p}$ stands in for the unknown $p$ in the variance.

## Key facts and formulas

"Given" means it is in the MF19 list of formulae and tables in the exam. "Learn it" means you must remember it.

| Formula | Status |
|---|---|
| Unbiased estimate of $\mu$: $\bar{x} = \dfrac{\Sigma x}{n}$ | Given |
| Unbiased estimate of $\sigma^2$: $s^2 = \dfrac{\Sigma(x-\bar{x})^2}{n-1} = \dfrac{1}{n-1}\left(\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}\right)$ | Given |
| Central Limit Theorem: $\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$ approximately for large $n$ | Given |
| Sample proportion: approximately $N\left(p, \dfrac{p(1-p)}{n}\right)$ | Given |
| $E(\bar{X}) = \mu$, $\mathrm{Var}(\bar{X}) = \dfrac{\sigma^2}{n}$; exactly normal if $X$ is normal | Learn it |
| Coded data: $\bar{x} = a + \dfrac{\Sigma(x-a)}{n}$, and $s^2 = \dfrac{1}{n-1}\left(\Sigma(x-a)^2 - \dfrac{(\Sigma(x-a))^2}{n}\right)$ | Learn it |
| Confidence interval for $\mu$: $\bar{x} \pm z\dfrac{\sigma}{\sqrt{n}}$ (use $s$ for $\sigma$ if unknown and $n$ large) | Learn it |
| Confidence interval for $p$: $\hat{p} \pm z\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$, with $\hat{p} = \dfrac{x}{n}$ | Learn it |
| $z$ for an $\alpha\%$ interval: $\Phi(z) = \frac{1}{2}\left(1 + \frac{\alpha}{100}\right)$; level from $z$: $\alpha\% = 2\Phi(z) - 1$ | Learn it |
| Width $= 2 \times z \times$ standard error; centre $=$ the sample mean (or $\hat{p}$) | Learn it |
| Critical values: $P(Z \leqslant z) = p$, e.g. $p = 0.975 \Rightarrow z = 1.960$ | Given |
| Binomial $B(n, p)$: $\mu = np$, $\sigma^2 = np(1-p)$; Poisson $\mathrm{Po}(\lambda)$: $\mu = \sigma^2 = \lambda$ | Given |

## How to do it

In the Paper 6 questions we have tagged for this topic (76 questions, 2021–2025; 65 different, because some papers share questions), the types came up this often. 31 of the 65 mix two or more types, so the counts add up to more than 65.

| Question type | Questions |
|---|---|
| Unbiased estimates of the mean and variance | 26 |
| Confidence interval for a population mean | 19 |
| Interpreting confidence intervals | 14 |

| Question type | Questions |
|---|---|
| Explaining the Central Limit Theorem | 13 |
| Confidence interval for a proportion | 12 |
| The distribution of the sample mean | 12 |
| Sampling methods and randomness | 11 |

## 1. Unbiased estimates of the mean and variance (26 questions)

1. **Mean:** $\bar{x} = \dfrac{\Sigma x}{n}$. Keep at least 4 significant figures for the next step.
2. **Variance:** $s^2 = \dfrac{1}{n-1}\left(\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}\right)$. Write it with the numbers in, so the method mark is clear.
3. Give both to 3 significant figures (or exactly, as a fraction).

For example, $n = 40$, $\Sigma x = 1060$, $\Sigma x^2 = 28\,450$: $\bar{x} = 26.5$ and $s^2 = \frac{1}{39}\left(28\,450 - \frac{1060^2}{40}\right) = \frac{360}{39} = 9.23$.

**Variations you will meet:**

- **Raw data** (a list of values): find $\Sigma x$ and $\Sigma x^2$ with your calculator, then use the same formulas.
- **A frequency table:** $\Sigma x \to \Sigma fx$, $\Sigma x^2 \to \Sigma fx^2$, $n \to \Sigma f$.
- **Coded data**, given $\Sigma(x-a)$ and $\Sigma(x-a)^2$: find the mean and variance of $x - a$ with the usual formulas, then add $a$ to the mean only. The variance is unchanged by subtracting $a$. For example, $n = 40$, $\Sigma(x - 20) = 36$, $\Sigma(x-20)^2 = 250$: $\bar{x} = 20 + \frac{36}{40} = 20.9$ and $s^2 = \frac{1}{39}\left(250 - \frac{36^2}{40}\right) = 5.58$.
- **Working backwards to $\Sigma x^2$.** Given $n$, $\bar{x}$ and $s^2$: $\Sigma x^2 = (n-1)s^2 + n\bar{x}^2$. With $n = 50$, $\bar{x} = 12$, $s^2 = 4$: $\Sigma x^2 = 49 \times 4 + 50 \times 144 = 7396$.
- **An unknown value $a$ in the data.** Write $\Sigma x$ and $\Sigma x^2$ in terms of $a$, put them into the $s^2$ formula, set it equal to the given value and solve the quadratic. Keep the root that fits the condition given (for example $a > 0$).
- **"Show that the estimate is ..."**: write the full formula with the numbers in, then the value, to more figures than the answer given.
- **Comment on the estimates.** A variance that is far too small (or negative) for the context suggests a mistake in $\Sigma x$ or $\Sigma x^2$. Estimates from a small sample, or a sample that isn't random, may not be reliable.
- **The next part of the question says "assume these are the true values"**: then use $\bar{x}$ as $\mu$ and $s^2$ as $\sigma^2$ in a probability or a hypothesis test.

## 2. Confidence interval for a population mean (19 questions)

1. **Find** $\bar{x}$, and $\sigma$ (given) or $s$ (from the data, for a large sample).
2. **Find** $z$ for the level: $\Phi(z) = \frac{1}{2}\left(1 + \frac{\alpha}{100}\right)$.
3. **Interval:** $\bar{x} \pm z\dfrac{\sigma}{\sqrt{n}}$. Don't forget the $\sqrt{n}$.
4. **Write it as an interval** to 3 significant figures, such as "$44.5$ to $46.5$".

**Variations you will meet (working backwards):**

- **From the interval to $\bar{x}$ and $\sigma$.** The centre is $\bar{x} = \frac{1}{2}(\text{lower} + \text{upper})$. Half the width is $z\frac{\sigma}{\sqrt{n}}$; solve that for $\sigma$ (the population standard deviation, or its estimate).
- **Find $n$ from the width:** $\sqrt{n} = \dfrac{2z\sigma}{\text{width}}$. $n$ must be a whole number, so round to the nearest integer.
- **Find the level $\alpha$ from the width or one end:** find $z$ from the half-width, then $\alpha\% = 2\Phi(z) - 1$. For example, if the upper end is $\bar{x} + 0.9$ and $\frac{s}{\sqrt{n}} = 0.5$, then $z = 1.8$, $\Phi(1.8) = 0.9641$, and $2(0.9641) - 1 = 0.928$, a $92.8\%$ interval. Don't stop at $\Phi(z)$.
- **Wider or narrower?** Without calculating, compare $n$ and the level: a smaller sample, or a higher level, gives a wider interval.

## 3. Interpreting confidence intervals (14 questions)

Each $c$ confidence interval (for example $c = 0.95$) contains the true mean with probability $c$, and intervals from independent samples are independent.

- **All $r$ intervals contain $\mu$:** $c^r$. **None of them:** $(1-c)^r$.
- **Exactly $k$ of $r$:** binomial, $\binom{r}{k}c^k(1-c)^{r-k}$. With intervals at **different** levels, list the cases: for a $90\%$ and a $95\%$ interval, $P(\text{exactly one contains } \mu) = 0.9 \times 0.05 + 0.1 \times 0.95$.
- **Expected number** of $r$ intervals that contain $\mu$: $rc$.
- **The largest $r$ with $P(\text{all contain}) > q$:** solve $c^r > q$ with logarithms, $r < \dfrac{\ln q}{\ln c}$ (the inequality turns round because $\ln c < 0$), then take the whole number below.
- **Wholly above (or below) $\mu$:** only one tail, so $\frac{1}{2}(1-c)$; for $98\%$ that is $0.01$.
- **Conditional probabilities.** "Given at least one of two intervals contains $\mu$, find $P(\text{both})$": $\dfrac{c^2}{1-(1-c)^2}$. Two intervals from the **same** sample, one at a higher level, are nested: the narrower one is inside the wider one. So $P(\text{narrow contains} \mid \text{wide contains}) = \dfrac{c_{\text{narrow}}}{c_{\text{wide}}}$.
- **Widths at two levels from the same sample.** Width is proportional to $z$, so the ratio of widths is the ratio of the $z$-values. If one interval is $k$ times as wide as another, divide its $z$ by $k$, then use $2\Phi(z) - 1$.
- **What it means.** "96% of the population lie in this interval" is wrong: the interval is about the **mean**, not individual values.
- **Comment on a claim.** If the claimed value is outside the interval, the claim is not supported; if it is inside, the data don't contradict it.

## 4. Explaining the Central Limit Theorem (13 questions)

Decide which of two cases you are in, and name the **population** distribution each time:

- **The CLT was needed** because the population distribution is not known (or is not normal, such as binomial or Poisson), and the sample is large, so the sample mean is approximately normal.
- **The CLT was not needed** because the population itself is normally distributed, so the sample mean is exactly normal.

**Variations you will meet:**

- **"Which part of your working uses the CLT?"**: using a $z$-value from the normal table for the sample mean.
- **"Under what circumstances would the CLT not be needed?"**: if the population were known to be normal.
- **"Why is it valid to use the CLT here?"**: because the sample is large (give the $n$).
- **"Would it make a difference if the population were normal?"**: for a large sample, no; the sample mean was already approximately normal.
- **"Why would the method fail for a small sample?"**: the CLT would not apply, so the sample mean need not be normal.

## 5. Confidence interval for a proportion (12 questions)

1. $\hat{p} = \dfrac{x}{n}$.
2. $z$ for the level, as before.
3. Interval: $\hat{p} \pm z\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$, given as an interval of **proportions** (not numbers of people).

**Variations you will meet:**

- **Find the level from one end.** From the upper end $U$, work out $\hat{p}$ and $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, solve $\hat{p} + z \times \text{that} = U$ for $z$, then $2\Phi(z) - 1$. From both ends, $\hat{p}$ is the midpoint.
- **Find $\hat{p}$ (or $x$) from the width.** $2z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = w$ gives $\hat{p}(1-\hat{p}) = n\left(\frac{w}{2z}\right)^2$, a quadratic with **two** roots that add up to $1$ (if $0.3$ works, so does $0.7$). Multiply by $n$ if the question asks for the number $x$.
- **Find $n$ and $\hat{p}$ from both ends**: $\hat{p}$ is the midpoint, then solve the half-width equation for $n$.
- **Narrower interval from the same data?** Use a lower confidence level.

## 6. The distribution of the sample mean (12 questions)

1. **Find $\mu$ and $\sigma^2$ for one value of $X$.** From $N(\mu, \sigma^2)$ directly; from $B(n, p)$, $np$ and $np(1-p)$; from $\mathrm{Po}(\lambda)$, both are $\lambda$; from a combination such as $2X + Y$, use $E$ and $\mathrm{Var}$ rules first.
2. **State the distribution:** $\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$, exactly if $X$ is normal, approximately (CLT) otherwise.
3. **Standardise** with the standard deviation $\dfrac{\sigma}{\sqrt{n}}$: $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$, and use the table.

For example, $X \sim N(50, 36)$ and $n = 16$: $\bar{X} \sim N(50, 2.25)$, so $P(\bar{X} > 52) = 1 - \Phi\left(\frac{2}{1.5}\right) = 1 - \Phi(1.333) = 0.0912$.

**Variations you will meet:**

- **Working backwards:** given $P(\bar{X} < a)$, find $z$ from the table, then $a = \mu + z\frac{\sigma}{\sqrt{n}}$. For a probability below $0.5$, $z$ is negative.

- **An unknown $\sigma$ from a $z$-value:** set $\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ equal to the given $z$ and solve.
- **Poisson or binomial populations:** no continuity correction is expected for a sample mean.
- **Between two values:** $\Phi(z_2) - \Phi(z_1)$, as for any normal variable.

### 7. Sampling methods and randomness (11 questions)

- **"Why is this method unsatisfactory?"**: say who is left out or over-represented, in context: people chosen from one place may share views that others don't; values from one part of the year may not be typical of the whole year; the first few results in a row are not a random sample.
- **Choose with random digits:** read the digits in groups of the right length, skip numbers too big, $000$, and repeats. List the ones you keep.
- **Scaling random numbers:** member $k$ comes from $\frac{k-1}{N} < u \leqslant \frac{k}{N}$; list the $u$ values in that range. Unequal numbers of $u$ values means unequal chances: not random.
- **Re-using digits** (starting again from the second digit, say) repeats numbers already chosen, so it isn't random.
- **Dice or coins:** show the probabilities are equal (random) or not (not random), with numbers. To choose two people from three, choose which one to **leave out**: $1$–$2$, $3$–$4$, $5$–$6$.
- **"Why use a sample?"**: testing destroys the item, or the population is too big to measure all of it.
- **"What condition makes the estimates reliable?"**: a random sample (and, for accuracy, a large one).

## Worked examples

The bold codes show where marks are earned, as in Cambridge mark schemes. The number is how many marks.

- **M1** method mark: for a correct method, even if the numbers slip.
- **A1** accuracy mark: for a correct answer; it needs the M mark before it.
- **B1** mark for a correct result on its own, with no method needed.

### Example 1

The times, $t$ minutes, that a random sample of $120$ customers waited on a helpline are summarised by

$$n = 120, \qquad \Sigma t = 5460, \qquad \Sigma t^2 = 252\,300.$$

(a) Calculate unbiased estimates of the population mean and variance of the waiting times. **[3]**

(b) Calculate a $95\%$ confidence interval for the population mean waiting time. **[3]**

(c) Explain why the Central Limit Theorem was needed in part (b). **[1]**

**(a)** $\bar{t} = \dfrac{5460}{120} = 45.5$. **B1**

$$s^2 = \frac{1}{119}\left(252\,300 - \frac{5460^2}{120}\right) = \frac{252\,300 - 248\,430}{119} = \frac{3870}{119} = 32.5\ (3\text{ s.f.})$$

**M1** for the correct formula with $n - 1 = 119$, **A1**.

**(b)** $z = 1.96$. **B1**

$$45.5 \pm 1.96\sqrt{\frac{32.521}{120}} = 45.5 \pm 1.020.$$

**M1** for the form $\bar{t} \pm z\sqrt{\frac{s^2}{n}}$. The interval is $44.5$ to $46.5$ minutes (3 s.f.). **A1**

**(c)** The population distribution of waiting times is not known, so the CLT is needed to say that the sample mean is approximately normally distributed (the sample is large). **B1**

## Example 2

In a random sample of $250$ households in a town, $90$ have a pet.

(a) Calculate an approximate $94\%$ confidence interval for the proportion of households in the town that have a pet. **[3]**

(b) In a second random sample of $300$ households in a different town, $x$ households have a garden. An approximate $90\%$ confidence interval for the proportion with a garden has width $0.08225$. Find the two possible values of $x$. **[4]**

**(a)** $\hat{p} = \dfrac{90}{250} = 0.36$. $\Phi(z) = 0.97$, so $z = 1.881$. **B1**

$$0.36 \pm 1.881\sqrt{\frac{0.36 \times 0.64}{250}} = 0.36 \pm 0.0571.$$

**M1**. The interval is $0.303$ to $0.417$ (3 s.f.). **A1**

**(b)** With $\hat{p} = \dfrac{x}{300}$ and $z = 1.645$ (**B1**):

$$2 \times 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{300}} = 0.08225 \ \Rightarrow\ \sqrt{\frac{\hat{p}(1-\hat{p})}{300}} = 0.025 \ \Rightarrow\ \hat{p}(1-\hat{p}) = 0.1875.$$

**M1** for the width equation. So $\hat{p}^2 - \hat{p} + 0.1875 = 0$ (**M1** for a quadratic), giving $\hat{p} = 0.25$ or $0.75$, and $x = 75$ or $225$. **A1**

## Example 3

The number of emails, $X$, that a worker receives in a ten-minute period has the distribution $\mathrm{Po}(3.6)$. A random sample of $80$ ten-minute periods is taken.

(a) Find the probability that the sample mean is greater than $3.8$. **[4]**

(b) Find the value of $a$ such that $P(\bar{X} < a) = 0.9$. **[2]**

(c) State why the Central Limit Theorem was needed in part (a). **[1]**

**(a)** $E(X) = \mathrm{Var}(X) = 3.6$, so $\bar{X} \sim N\left(3.6,\ \frac{3.6}{80}\right) = N(3.6,\ 0.045)$ approximately. **B1** for the mean, **B1** for the variance.

$$z = \frac{3.8 - 3.6}{\sqrt{0.045}} = \frac{0.2}{0.21213} = 0.943.$$

**M1** for standardising with $\sqrt{0.045}$. $P(\bar{X} > 3.8) = 1 - \Phi(0.943) = 0.173$ (3 s.f.). **A1**

**(b)** $\Phi(z) = 0.9$ gives $z = 1.282$. $a = 3.6 + 1.282 \times 0.21213 = 3.87$ (3 s.f.). **M1 A1**

**(c)** The population ($X$) has a Poisson distribution, not a normal one. **B1**

### Example 4

The lengths of bolts made by a machine have standard deviation $4.2$ mm. A random sample of $n$ bolts gives a $98\%$ confidence interval for the mean length of $26.08$ mm to $28.52$ mm, correct to 4 significant figures.

(a) Find the sample mean and the value of $n$. **[4]**

(b) Five more random samples are taken, and a $98\%$ confidence interval is found from each. Find the probability that at least four of these intervals contain the population mean. **[2]**

(c) Find the probability that the interval in the question lies wholly below the population mean. **[1]**

**(a)** $\bar{x} = \frac{1}{2}(26.08 + 28.52) = 27.3$ mm. **B1**

Half the width is $\frac{1}{2}(28.52 - 26.08) = 1.22$, and $z = 2.326$ (**B1**):

$$2.326 \times \frac{4.2}{\sqrt{n}} = 1.22 \ \Rightarrow\ \sqrt{n} = \frac{2.326 \times 4.2}{1.22} = 8.008 \ \Rightarrow\ n = 64.1.$$

**M1**. $n$ is a whole number, so $n = 64$. **A1**

**(b)** The number containing the mean is $B(5, 0.98)$. **M1**

$$0.98^5 + 5 \times 0.98^4 \times 0.02 = 0.9039 + 0.0922 = 0.996\ (3\text{ s.f.})$$

**A1**

**(c)** The $2\%$ of intervals that miss are split evenly between the two sides, so $\frac{1}{2} \times 0.02 = 0.01$. **B1**

## Common mistakes

- **Using the biased variance.** Dividing by $n$ instead of $n-1$ loses the method mark, and later marks too. Examiners see this every year, and also a mix of the two formulas (for example $\frac{1}{n-1}\Sigma x^2 - \bar{x}^2$). Learn one correct form and write it out.

- **Rounding the mean too early.** Using $61.4$ instead of $61.43$ as the centre of an interval can change the answer's third figure. Keep $4$ or more significant figures until the end.
- **Forgetting** $\sqrt{n}$, or using $\sigma^2$ where $\sigma$ belongs. The interval uses $\frac{\sigma}{\sqrt{n}}$, and standardising a sample mean does too.
- **The wrong $z$-value.** For $98\%$, $z = 2.326$ (cut off $1\%$ each side), not $2.054$. Check with $\Phi(z) = \frac{1}{2}\left(1 + \frac{\alpha}{100}\right)$.
- **Not giving an interval.** "$\bar{x} \pm 1.02$" or a single number loses the last mark; write both ends.
- **Stopping at $\Phi(z)$ when finding a level.** $\Phi(z) = 0.9641$ does not mean a $96\%$ interval: the level is $2\Phi(z) - 1$.
- **Vague CLT answers.** "The distribution is unknown" or "it is not normal" is not enough: say the **population** distribution. And don't say the CLT is needed "because the sample is large" when the question asks why it is needed; the reason is the unknown population.
- **Saying the interval contains $96\%$ of the values.** It is about the mean, not individuals.
- **Multiplying instead of powering.** The chance that three $95\%$ intervals all contain $\mu$ is $0.95^3$, not $0.95 \times 3$.
- **Using the whole tail for "wholly above".** For a $98\%$ interval it is $0.01$, not $0.02$.
- **Random numbers:** keeping a repeat or a number too big, or saying a method is "biased" without saying why (which numbers are more likely, or which were already used).

## How to get the marks

- **Estimates:** the mean usually earns a mark on its own; the variance mark needs the correct unbiased formula with the numbers in. Give the variance, not the standard deviation, unless asked.
- **Intervals:** there is usually one mark for the correct $z$ (seen), one for the form $\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$ with any $z$, and one for the final interval to 3 significant figures, written as an interval.
- **Show your $z$.** Write "$z = 2.054$" so the mark can be given even if the arithmetic slips.
- **"Show that"** means the answer is given: write the formula with numbers and a value with more figures than the one printed.
- **Explanations** earn one mark each and must be specific and in context: name the population, the part of the method, or the group left out. "It is biased" alone rarely scores.
- **"Without calculation"** means a reason in words, such as "smaller $n$, so wider".
- **"Correct to the nearest integer"** (a level or an $n$): round only at the end, and give a whole number.
- **Probabilities of intervals:** write the expression, such as $0.95^r > 0.5$ or $4 \times 0.94^3 \times 0.06$, before the answer.
- **Accuracy:** 3 significant figures unless told otherwise, from working kept to at least $4$.

## Quick check

1. Find unbiased estimates of the population mean and variance from this random sample: $12.4$, $11.8$, $13.1$, $12.9$, $12.2$, $11.6$.

2. $X \sim N(50, 36)$. Find the probability that the mean of a random sample of $9$ values is less than $48$. Was the Central Limit Theorem needed?
3. Members of a club are numbered $1$ to $420$. A sample is chosen by reading these random digits in groups of three: $3\,8\,2\,9\,1\,7\,0\,4\,5\,4\,1\,5\,6\,3\,3\,0\,8\,2$. Write down the numbers of the first four members chosen.
4. A population has standard deviation $3$. From a random sample of $50$ values, a confidence interval for the population mean has width $1.5$. Find the confidence level, correct to 1 decimal place.
5. An approximate $95\%$ confidence interval for a proportion is $0.312$ to $0.408$. Find the sample proportion and the sample size.

## Answers

1. $\Sigma x = 74$, $\Sigma x^2 = 914.42$, $n = 6$. $\bar{x} = \frac{74}{6} = 12.3$ (3 s.f.). $s^2 = \frac{1}{5}\left(914.42 - \frac{74^2}{6}\right) = \frac{1}{5}(1.7533) = 0.351$ (3 s.f.).
2. $\bar{X} \sim N\left(50, \frac{36}{9}\right) = N(50, 4)$. $P(\bar{X} < 48) = \Phi\left(\frac{48-50}{2}\right) = \Phi(-1) = 1 - 0.8413 = 0.159$. No: $X$ is normal, so $\bar{X}$ is exactly normal.
3. The groups are $382$, $917$, $045$, $415$, $633$, $082$. Ignore $917$ and $633$ (bigger than $420$). The members are $382$, $45$, $415$ and $82$.
4. Half the width is $0.75 = z \times \frac{3}{\sqrt{50}}$, so $z = \frac{0.75\sqrt{50}}{3} = 1.768$. $\Phi(1.768) = 0.9615$, and $2(0.9615) - 1 = 0.923$: a $92.3\%$ interval.
5. $\hat{p} = \frac{1}{2}(0.312 + 0.408) = 0.36$. Half the width is $0.048 = 1.96\sqrt{\frac{0.36 \times 0.64}{n}}$, so $\frac{0.2304}{n} = \left(\frac{0.048}{1.96}\right)^2$ and $n = 384.16$, so $n = 384$.

## Past questions to try

- **9709/61/M/J/23 Q4**: $n$ from an interval's width, an unsuitable sample, and several intervals.
- **9709/62/M/J/24 Q2**: random digits, then estimates from totals.
- **9709/63/O/N/21 Q3**: the confidence level from the width of a proportion interval.
- **9709/63/M/J/21 Q4**: estimates, an interval, the CLT and an interval wholly above the mean.