Estimators

Statistics turns the probability story around: instead of deducing data from a known population, we want to infer the population from data.

Definition.
- A **population parameter** is a fixed (but typically unknown) number describing the population, e.g. the mean μ, the variance σ^2, a proportion p.
- A **statistic** is any quantity computed from the sample alone — it must not involve unknown population parameters. As a random sample varies, a statistic is a *random variable* with its own distribution.
- An **estimator** of a parameter θ is a statistic used to estimate it, written with hat notation: $\hat{\theta}$.

Definition. An estimator $\hat{\theta}$ is **unbiased** if

$$\mathbb{E}[\hat{\theta}] = \theta$$

i.e. on average, over many repeated samples, it hits the true value — it has no systematic tendency to over- or under-estimate.

> **Remark.** The idea is perfectly accessible and worth understanding properly: it pays off immediately in the $n-1$ mystery below.

Theorem ($\overline{X}$ is unbiased for μ)

For a random sample $X_1, \ldots, X_n$ from any population with mean μ,

$$\mathbb{E}[\overline{X}] = \mu$$

i.e. the sample mean is an unbiased estimator of the population mean.

The proof is a one-line application of the linearity of expectation — write it out yourself.

$$\mathbb{E}[\overline{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n}(n\mu) = \mu$$

> **Remark.** Unbiasedness is not automatic. The sample *median*, for instance, is in general **not** an unbiased estimator of the population median (consider a skewed population: the sample median of small samples is systematically dragged about by the skew). Each proposed estimator must earn the property.

> **Remark** (Comparing estimators)**.** For a symmetric population both the sample mean and sample median are unbiased estimators of the centre — so which is better? The natural tiebreaker is *variance*: an estimator that is unbiased *and* tightly concentrated is more useful than an unbiased but wildly scattered one. For normal populations $\mathrm{Var}[\overline{X}] = \sigma^2/n$ beats the sample median's variance ($\approx \pi\sigma^2/2n$), which is why we use the mean. This idea — efficiency of estimators — is the start of a beautiful university topic.

Example (OCR S4, June 2011)

The continuous random variable U has unknown mean μ and known variance σ^2. In order to estimate μ, two random samples, one of 4 observations of U and the other of 6 observations of U, are taken. The sample means are denoted by $\overline{U}_4$ and $\overline{U}_6$ respectively. One estimator S, given by $S = \frac{1}{2}\left(\overline{U}_4 + \overline{U}_6\right)$, is proposed.

(a) Show that S is unbiased and find $\mathrm{Var}[S]$ in terms of σ^2.

(b) A second estimator T of the form $a\overline{U}_4 + b\overline{U}_6$ is proposed, where a and b are chosen such that T is an unbiased estimator for μ with the smallest possible variance. Find the values of a and b and the corresponding variance of T.

(c) State, giving a reason, which of S and T is the better estimator.

(d) Compare the efficiencies of this preferred estimator and the mean of all 10 observations.

Recall $\mathbb{E}[\overline{U}_n] = \mu$ and $Var[\overline{U}_n] = \sigma^2/n$, and that the two sample means are independent.

(a) $\mathbb{E}[S] = \frac{1}{2}\left(\mathbb{E}[\overline{U}_4] + \mathbb{E}[\overline{U}_6]\right) = \frac{1}{2}(\mu + \mu) = \mu$, *so S is unbiased. By independence,*

$$Var[S] = \frac{1}{4}\left(\frac{\sigma^2}{4} + \frac{\sigma^2}{6}\right) = \frac{5\sigma^2}{48}$$

(b) $\mathbb{E}[T] = (a+b)\mu$, *and unbiasedness for* every μ *forces $a + b = 1$. Then*

$$Var[T] = \frac{a^2\sigma^2}{4} + \frac{(1-a)^2\sigma^2}{6}$$

Minimise $f(a) = \frac{a^2}{4} + \frac{(1-a)^2}{6}$: $f'(a) = \frac{a}{2} - \frac{1-a}{3} = 0$ gives $3a = 2(1-a)$, so $a = \frac{2}{5}$, $b = \frac{3}{5}$ (and $f'' = \frac{1}{2} + \frac{1}{3} > 0$, a minimum). Then

$$Var[T] = \frac{(2/5)^2\sigma^2}{4} + \frac{(3/5)^2\sigma^2}{6} = \frac{\sigma^2}{25} + \frac{3\sigma^2}{50} = \frac{\sigma^2}{10}$$

(c) Both are unbiased, and $Var[T] = \frac{\sigma^2}{10} < \frac{5\sigma^2}{48} = Var[S]$, so T is the better estimator.

(d) The mean of all 10 *observations is unbiased with variance $\frac{\sigma^2}{10}$ — the same efficiency as T. In fact they are the* same statistic*: weighting each sample mean by its sample size, $\frac{2}{5}\overline{U}_4 + \frac{3}{5}\overline{U}_6 = \frac{4\overline{U}_4 + 6\overline{U}_6}{10}$, which is the mean of all* 10 *observations. (Moral: pooling the raw data is optimal, and the naive S over-weights the smaller sample.)*

Estimating the Variance: the Mystery of $n-1$

The obvious estimator of σ^2 is the variance of the sample, computed "the usual way":

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Surprisingly, this is *biased*: it systematically underestimates σ^2.

Theorem

For a random sample $X_1, \ldots, X_n$ from a population with mean μ and variance σ^2,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = \frac{n-1}{n}\sigma^2$$

Intuition first: the deviations are measured from $\overline{X}$, not from the true mean μ. But $\overline{X}$ is computed from this very sample, so it sits *closer to the data than μ does* — indeed $\overline{X}$ is exactly the value minimising $\sum(x_i - a)^2$. Measuring spread about the sample's own centre therefore comes out too small, by precisely the factor $\frac{n-1}{n}$.

The proof is two lines of expectation algebra: expand $\sum(X_i - \overline{X})^2$, then take expectations using $\mathbb{E}[Y^2] = \mathrm{Var}[Y] + (\mathbb{E}[Y])^2$.

First expand:

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum X_i^2 - n\overline{X}^2$$

Now take expectations, using $\mathbb{E}[X_i^2] = Var[X_i] + (\mathbb{E}[X_i])^2 = \sigma^2 + \mu^2$ and likewise $\mathbb{E}[\overline{X}^2] = \frac{\sigma^2}{n} + \mu^2$:

$$\mathbb{E}\Big[\sum(X_i - \overline{X})^2\Big] = n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) = (n-1)\sigma^2$$

Dividing by n gives the theorem. Notice the bias enters exactly through $Var[\overline{X}] = \sigma^2/n$ — the wobble of the sample mean itself.

The fix is now obvious: divide by $n-1$ instead of n.

Definition. The **unbiased estimate of the population variance** from a sample is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)$$

Equivalently: compute the variance "the usual way" and then multiply by $\frac{n}{n-1}$.

> **Remark** (OCR terminology warning)**.** For OCR, the phrase **sample variance** means s^2, the *unbiased estimate of the population variance* — it is **not** the same as the variance of the sample computed with divisor n! Read questions with this in mind, and when in doubt in FM work, use s^2.

Tip (Calculator)

Your calculator's statistics mode gives both: σ_x uses divisor n (variance of the data in hand), while s_x uses divisor $n-1$ (the unbiased estimate). For estimation, hypothesis tests and confidence intervals you want s_x.

Example

A random sample of five measurements (in mm) of a machined part is

$$4.8, \quad 5.2, \quad 5.5, \quad 4.9, \quad 5.6$$

Find unbiased estimates of the population mean and variance.

$\sum x = 26.0$ and $\sum x^2 = 135.70$, with $n = 5$.

$$\bar{x} = \frac{26.0}{5} = 5.2$$

$$s^2 = \frac{1}{4}\left(135.70 - \frac{26.0^2}{5}\right) = \frac{1}{4}(135.70 - 135.20) = 0.125$$

So the unbiased estimates are $\hat{\mu} = 5.2$ mm and $\hat{\sigma}^2 = 0.125$ mm^2. (Check: the divisor-n variance is 0.1, and $\frac{n}{n-1} \times 0.1 = \frac{5}{4} \times 0.1 = 0.125$.)

Example (In class)

A random sample of 100 observations from a population gives $\sum x = 512$ and $\sum x^2 = 2843$. Find unbiased estimates of the population mean and variance.

Example (OCR S2, June 2009)

The continuous random variable R has the distribution $N(\mu, \sigma^2)$. The results of 100 observations of R are summarised by

$$\sum r = 3360.0, \qquad \sum r^2 = 115782.84$$

(a) Calculate an unbiased estimate of μ and an unbiased estimate of σ^2.

(b) The mean of 9 observations of R is denoted by $\overline{R}$. Calculate an estimate of $\mathbb{P}(\overline{R} > 32.0)$.

(c) Explain whether you need to use the Central Limit Theorem in your answer to part (b).

(a) $\hat{\mu} = \bar{r} = \frac{3360.0}{100} = 33.6$ and

$$s^2 = \frac{1}{99}\left(115782.84 - \frac{3360.0^2}{100}\right) = \frac{2886.84}{99} = 29.16$$

(b) Using the estimates in place of μ and σ^2:

$$\overline{R} \sim N\left(33.6, \frac{29.16}{9}\right) = N(33.6, 1.8^2)$$

$$\mathbb{P}(\overline{R} > 32.0) = \mathbb{P}\left(Z > \frac{32.0 - 33.6}{1.8}\right) = \mathbb{P}(Z > -0.889) = 0.813$$

(c) No. R itself is normally distributed, so $\overline{R}$ is exactly *normal for any sample size — the Central Limit Theorem (which concerns non-normal parents) is not needed, and $n = 9$ being small doesn't matter.*

Textbook Exercises: [CUP.S] Ch 8 §3, 5; [S2] Ch 4 §4.7

Confidence Intervals

A point estimate like $\bar{x} = 5.2$ carries no sense of its own precision. Better to report an *interval* of plausible values for μ, together with how confident we are in the procedure.

Definition. A $C\%$ **confidence interval** for the population mean is

$$\left(\bar{x} - z\frac{\sigma}{\sqrt{n}},\ \bar{x} + z\frac{\sigma}{\sqrt{n}}\right) \qquad \text{i.e.} \qquad \bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

where z is the value such that $\mathbb{P}(-z < Z < z) = C\%$ for $Z \sim N(0,1)$, found by inverse normal.

Fact (Common critical values) —

Confidence level	90%	95%	99%
z	1.645	1.960	2.576

Each z is the inverse normal of $\frac{1+C}{2}$, e.g. $\Phi^{-1}(0.975) = 1.960$. Don't memorise blindly — know how to find z for, say, 98%.

Fact (The three cases — mirroring the hypothesis tests) — The formula $\bar{x} \pm z\sigma/\sqrt{n}$ may be used when:

1. the sample is drawn from a **normal population with known** (given or assumed) **variance** — exact, any n;
2. the sample is **large**, from **any population with known variance** — approximate, by the Central Limit Theorem;
3. the sample is **large**, from any population with **unknown variance** — replace σ by s, the square root of the unbiased estimate s^2.

Example (Case 1)

The mass of a machine component is normally distributed with standard deviation 4 g. A random sample of 25 components has mean mass 72.4 g. Find a 95% confidence interval for the population mean mass.

Population normal, $\sigma = 4$ known: case 1.

$$72.4 \pm 1.96 \times \frac{4}{\sqrt{25}} = 72.4 \pm 1.568$$

giving the interval $(70.8,\ 74.0)$ *g (3 s.f.).*

Example (Case 3, and commenting on a claim)

A supplier claims that the mean nicotine content of a batch of cigarettes is 22 mg. A laboratory measures a random sample of 80 cigarettes, obtaining

$$\sum x = 1672 \qquad \sum x^2 = 35\,632$$

(a) Find a 95% confidence interval for the population mean nicotine content.

(b) Comment on the supplier's claim.

(a) $\bar{x} = \frac{1672}{80} = 20.9$ *and*

$$s^2 = \frac{1}{79}\left(35\,632 - \frac{1672^2}{80}\right) = \frac{687.2}{79} = 8.699 \text{ (4 s.f.)}, \quad s = 2.949$$

The variance is unknown but $n = 80$ *is large, so (CLT, case 3):*

$$20.9 \pm 1.96 \times \frac{2.949}{\sqrt{80}} = 20.9 \pm 0.646$$

giving the interval $(20.3,\ 21.5)$ *mg (3 s.f.).*

(b) 22 lies outside the 95% confidence interval, so there is evidence to suggest that the population mean nicotine content differs from 22 mg — the data casts doubt on the supplier's claim. (Note the cautious phrasing: the claim is not "disproved".)

Example (In class: case 2, and finding z)

A large sample of 40 observations is taken from a population which is *not* normally distributed, but whose standard deviation is known to be 2.5. The sample mean is 15.2. Find a 98% confidence interval for the population mean, justifying the use of the normal distribution.

Deriving the interval

Where does $\bar{x} \pm z\sigma/\sqrt{n}$ come from? It is the statement $\mathbb{P}(-z < Z < z) = C\%$ about the standardised sample mean, with the inequality unwrapped to put μ in the middle. Run the algebra yourself before revealing it.

Suppose $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, *so that* $Z = \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. *For 95%, the inverse normal gives* $\mathbb{P}(-1.96 < Z < 1.96) = 0.95$*:*

$$\mathbb{P}\left(-1.96 < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

Rearranging the inequality for μ*:*

$$\mathbb{P}\left(\overline{X} - 1.96\,\frac{\sigma}{\sqrt{n}} \;<\; \mu \;<\; \overline{X} + 1.96\,\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

The confidence interval is this random *interval* $\overline{X} \pm 1.96\,\sigma/\sqrt{n}$, *evaluated at the sample in hand.*

Remark (Confidence intervals and two-tail tests)**.** A value μ_0 lies outside a 95% confidence interval exactly when a two-tail hypothesis test of $H_0 : \mu = \mu_0$ at the 5% level would reject H_0. The interval is precisely the set of null hypotheses the data would *not* reject — which is why "comment on the claim" questions can be answered straight from the interval.

Example (OCR Further Stats, November 2021)

A random sample of 160 observations of a random variable X is selected. The sample can be summarised as follows.

$$n = 160, \qquad \sum x = 2688, \qquad \sum x^2 = 48\,398$$

(a) Calculate unbiased estimates of $\mathbb{E}[X]$ and $\mathrm{Var}[X]$.

(b) Find a 99% confidence interval for $\mathbb{E}[X]$, giving the end-points of the interval correct to 4 significant figures.

(c) Explain whether it was necessary to use the Central Limit Theorem in answering part (a); part (b).

(a) $\hat{\mu} = \bar{x} = \frac{2688}{160} = 16.8$ *and*

$$s^2 = \frac{1}{159}\left(48\,398 - \frac{2688^2}{160}\right) = \frac{160}{159} \times 20.2475 = 20.37 \text{ (4 s.f.)}$$

(b) Unknown variance, $n = 160$ *large: case 3, with* $z = \Phi^{-1}(0.995) = 2.576$.

$$16.8 \pm 2.576 \times \sqrt{\frac{20.37}{160}} = 16.8 \pm 0.9192$$

giving the interval $(15.88,\ 17.72)$ *(4 s.f.).*

(c) Not in part (a): unbiasedness of $\bar{x}$ *and* s^2 *holds for* any *distribution — no normality is involved. Yes in part (b):* X *is not stated to be normal, so the (approximate) normality of* $\overline{X}$*, on which the interval rests, comes from the Central Limit Theorem, justified because* $n = 160$ *is large.*

Example (Width and sample size)

Measurements of a quantity have standard deviation $\sigma = 12$. How large a sample is needed so that a 90% confidence interval for the mean has width less than 4?

The width of the interval is $2 \times 1.645 \times \dfrac{12}{\sqrt{n}}$. *We need*

$$\frac{2 \times 1.645 \times 12}{\sqrt{n}} < 4 \quad \Longleftrightarrow \quad \sqrt{n} > \frac{2 \times 1.645 \times 12}{4} = 9.87 \quad \Longleftrightarrow \quad n > 97.4$$

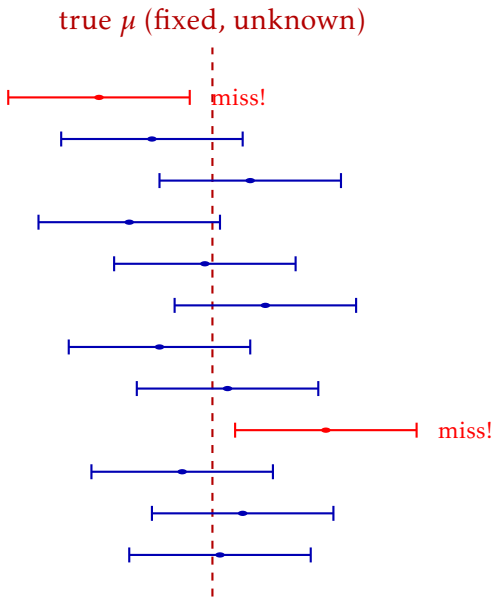
Since n must be a whole number, the smallest sufficient sample size is $n = 98$. *(Notice the diminishing returns: width shrinks like* $1/\sqrt{n}$*, so halving the width costs four times the data.)*

Textbook Exercises: [CUP.S] Ch 9 §2; [S3&4] S3 Ch 3

Interpreting a Confidence Interval

Remark (The crucial misconception)**.** A 95% confidence interval does **not** mean "there is a 95% chance that μ lies in this interval". The population mean μ is a *fixed number*, not a random one: a particular interval such as $(70.8, 74.0)$ either contains μ or it doesn't — there is no probability left. What *is* random is the interval: different samples give different intervals.

Fact (Correct interpretation) — The 95% describes the *procedure*, not any single interval: under repeated sampling, 95% of the confidence intervals constructed this way would capture the true mean μ. Our confidence is in the method that generated the interval — like trusting a fisherman who nets the fish 95% of the time, while knowing nothing about today's particular cast.



Twelve samples, twelve 95% *confidence intervals (dot = sample mean). The intervals move; μ does not. In the long run* 95% *of intervals capture μ — but any one interval either has, or hasn't.*

Exercise. A student writes: "My 95% confidence interval is $(20.3, 21.5)$, so $\mathbb{P}(20.3 < \mu < 21.5) = 0.95$." Explain precisely what is wrong, and rewrite the sentence correctly.

The randomness of the interval is exactly what the final part of this past-paper question exploits.

Example (OCR Further Stats, June 2023)

A club secretary collects data about the time, T minutes, needed to process the details of a new member. The mean of T is denoted by μ and the variance of T is denoted by σ^2. The results of a random sample of 40 observations of T are summarised as follows.

$$n = 40, \qquad \sum t = 396.0, \qquad \sum t^2 = 4271.40$$

(a) Determine a 99% confidence interval for μ.

(b) The secretary discovers that over a long period the value of σ^2 is in fact 10.0. The secretary collects an independent random sample of 50 observations of T and constructs a new 99% confidence interval for μ based on this sample of size 50, but using $\sigma^2 = 10.0$. Find the probability that this new confidence interval contains the value $\mu + 1.6$.

(a) $\bar{t} = \frac{396.0}{40} = 9.9$ *and*

$$s^2 = \frac{1}{39}\left(4271.40 - \frac{396.0^2}{40}\right) = \frac{40}{39} \times 8.775 = 9.0$$

Unknown variance, $n = 40$ *large (CLT, case 3),* $z = 2.576$*:*

$$9.9 \pm 2.576 \times \sqrt{\frac{9.0}{40}} = 9.9 \pm 1.222$$

giving the interval $(8.68,\ 11.12)$ *minutes (3 s.f.).*

(b) Treat the new interval as random*: it is* $\overline{T} \pm 2.576\sqrt{10.0/50} = \overline{T} \pm 1.152$*, where* $\overline{T} \sim N(\mu, 10.0/50)$*, i.e. standard deviation* $\sqrt{0.2} = 0.4472$*. The interval contains the fixed number* $\mu + 1.6$ *exactly when*

$$\overline{T} - 1.152 < \mu + 1.6 < \overline{T} + 1.152 \quad \Longleftrightarrow \quad 0.448 < \overline{T} - \mu < 2.752$$

Since $\overline{T} - \mu \sim N(0, 0.2)$*,*

$$\mathbb{P}\left(\frac{0.448}{0.4472} < Z < \frac{2.752}{0.4472}\right) = \mathbb{P}(1.002 < Z < 6.154) = 1.0000 - 0.8418 = 0.158 \text{ (3 s.f.)}$$

(The question only makes sense because the interval *wobbles while* $\mu + 1.6$ *stays fixed — the heart of the correct interpretation above.)*

Textbook Exercises: [CUP.S] Ch 9 §2; [S3&4] S3 Ch 3