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.

> **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 $\text{Var}[S]$ 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.

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 $\mathbb{E}[Y^2] = \mathrm{Var}[Y] + (\mathbb{E}[Y])^2$.

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.

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).

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.

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.

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.

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).

Example (Width and sample size)

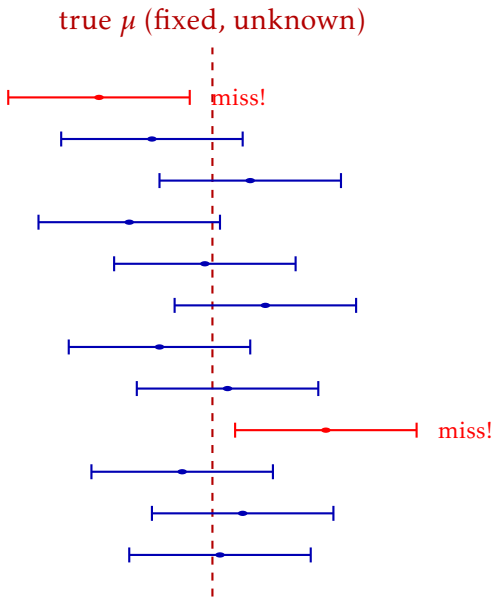
Measurements of a quantity have standard deviation $\sigma = 12$. How large a sample is needed so that a 90% confidence interval for the mean has width less than 4?

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 $\mu + 1.6$.

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