

Sampling Distributions

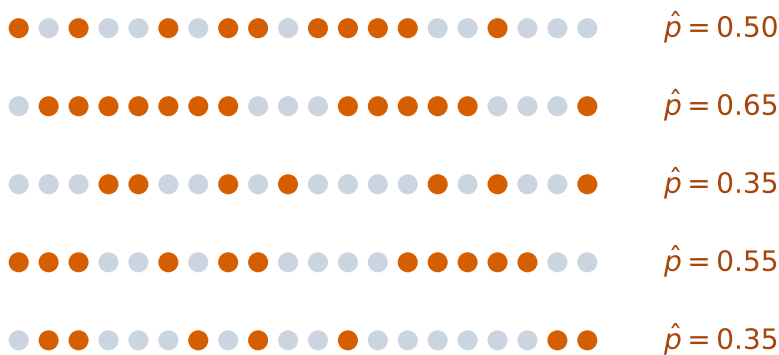
AP Statistics

Why Two Samples Never Match: Sampling Variability

A **statistic** (like a sample mean $\bar{x}$ or sample proportion $\hat{p}$) is computed from a sample and **varies** from sample to sample – this is **sampling variability**.

A **parameter** (μ or p) is the fixed truth about the population. The **sampling distribution** is the distribution of a statistic over all possible samples of a given size – it is the bridge from one sample to inference.

population: 40% say yes



same population,
same method,
different answers

each row: a random sample of 20; orange = yes

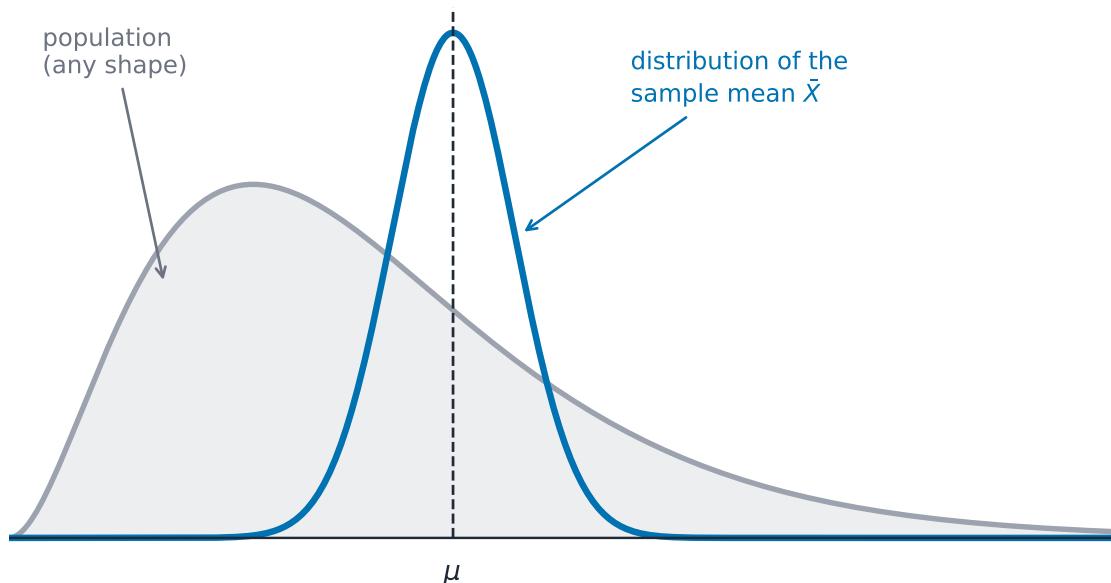
Five random samples from one population give five different sample proportions

The Normal Curve as a Model for a Statistic

For large enough samples, many sampling distributions are approximately **normal**. That lets us describe a statistic by a **center** (its mean), a **spread** (its **standard error**), and a normal shape – and then compute how likely a given sample result is.

The Central Limit Theorem

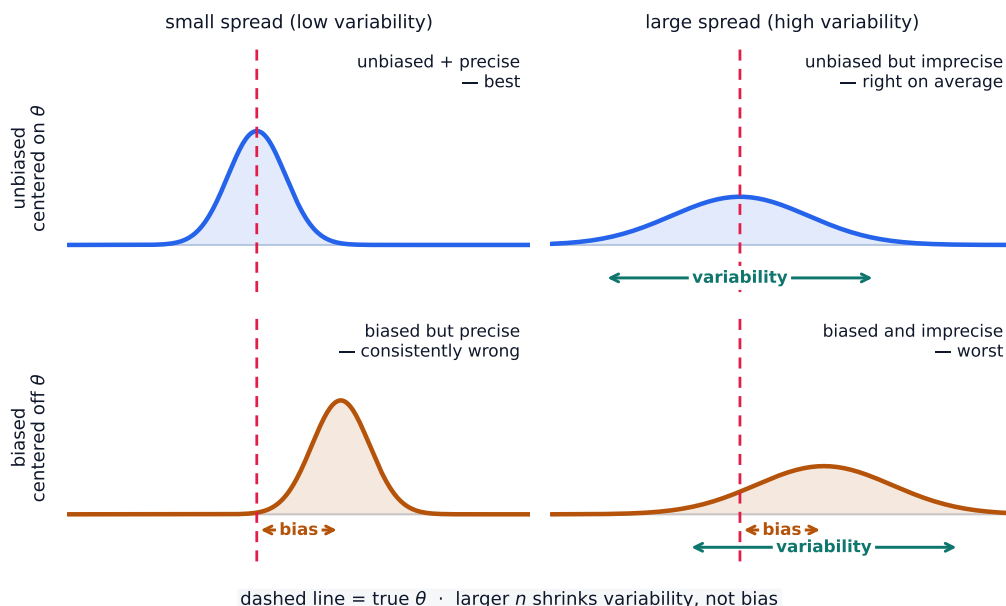
The **Central Limit Theorem** (CLT): for a **sample mean**, if the sample size n is large enough (a common rule is $n \geq 30$), the sampling distribution of $\bar{x}$ is approximately **normal**, *regardless* of the population's shape. The larger n , the more normal and the tighter the distribution.



The sample mean is nearly normal whatever the shape of the population

Good Guesses and Bad Guesses: Bias

A statistic is **unbiased** if the mean of its sampling distribution equals the parameter – it is correct **on average**. Bias is about the center being off; **variability** is about the spread. A good estimator is both unbiased (centered right) and low-variability (precise); larger samples reduce variability but do not fix bias from bad sampling.



*Bias and variability are separate faults. Only the top-left estimator is both centered on θ and tight; the bottom-left one is precise but **consistently wrong**, which no amount of extra data will fix.*

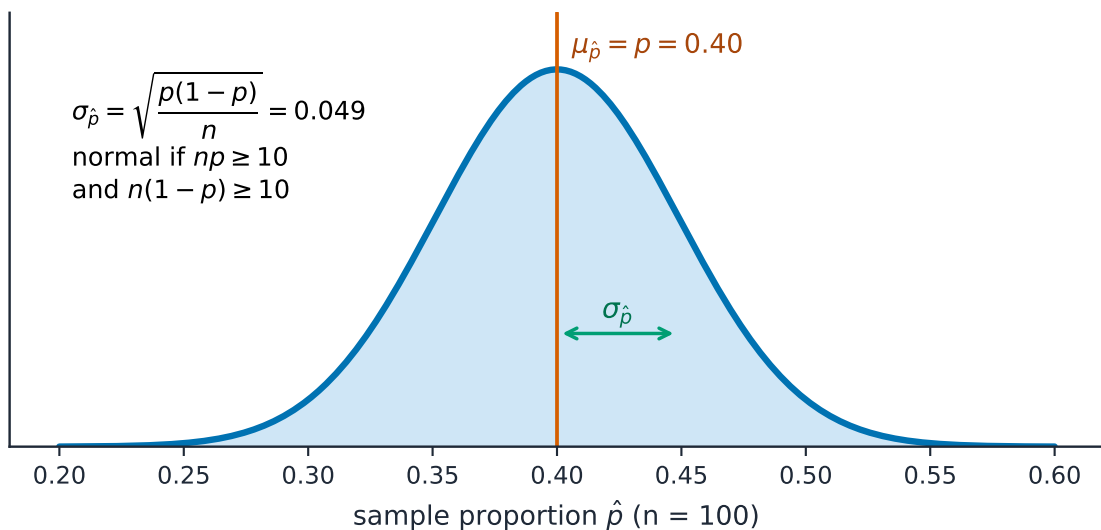
The Sampling Distribution of a Sample Proportion

For a sample proportion $\hat{p}$ from an SRS: the mean is p (unbiased), and the standard deviation is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

This spread has two names: it is the **standard deviation** of the sampling distribution, and it is called the **standard error** once you must estimate it from the sample (replacing p by $\hat{p}$) – which is exactly what the later inference units do. It is approximately **normal** when $np \geq 10$ and $n(1-p) \geq 10$ (the **Large Counts** condition), and the 10% condition ($n \leq 0.10N$) keeps the observations near-independent.

Worked example. Suppose 40% of voters favor a measure ($p = 0.4$) and you sample $n = 100$. The standard error is $\sigma_{\hat{p}} = \sqrt{\frac{0.4(0.6)}{100}} = 0.049$. The chance a sample gives $\hat{p} > 0.5$ is $z = \frac{0.5 - 0.4}{0.049} = 2.04$, so $P(\hat{p} > 0.5) \approx 0.02$ – a majority in the sample would be surprising.



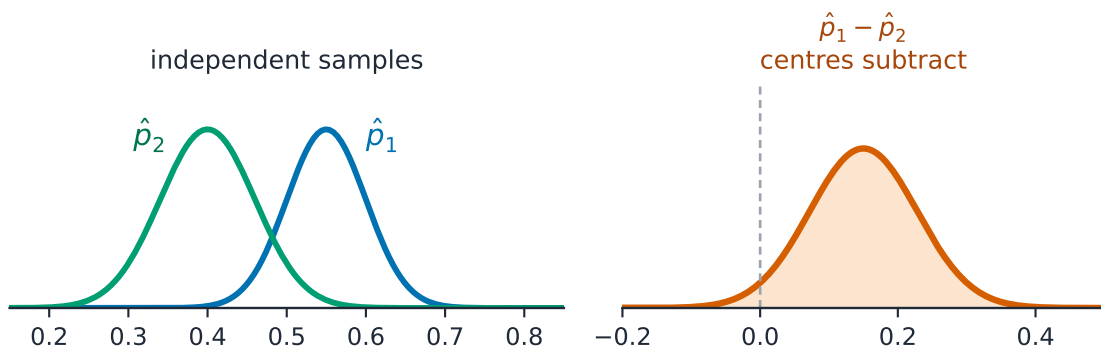
The sampling distribution of $\hat{p}$ is centred at p with standard deviation $\sqrt{p(1-p)/n}$

Comparing Two Groups: Difference of Sample Proportions

For $\hat{p}_1 - \hat{p}_2$ from two independent samples: the mean is $p_1 - p_2$, and because the samples are independent the **variances add**:

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

It is approximately normal when the Large Counts condition holds in **both** samples.



$$\sigma = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

The distribution of $\hat{p}_1 - \hat{p}_2$: centres subtract and variances add

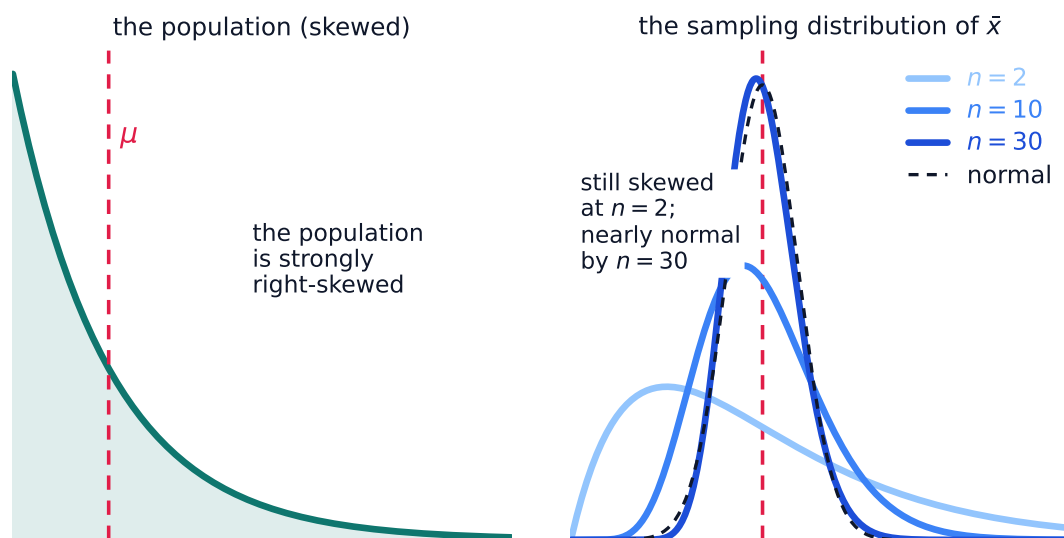
The Sampling Distribution of a Sample Mean

For a sample mean $\bar{x}$ from an SRS: the mean is μ (unbiased), and the standard deviation is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Its shape is normal if the population is normal, or approximately normal for large n by the CLT. Note the spread shrinks like $\sqrt{n}$ – quadrupling the sample halves the standard error.

Worked example. A population has $\mu = 70$ and $\sigma = 12$. For samples of $n = 36$, the sampling distribution of $\bar{x}$ is centered at 70 with standard error $\frac{12}{\sqrt{36}} = 2$. The chance a sample mean exceeds 73 is $z = \frac{73 - 70}{2} = 1.5$, so $P(\bar{x} > 73) \approx 0.067$.



CLT: for large n , $\bar{x} \approx \text{Normal}(\mu, \sigma/\sqrt{n})$ even when the population is not normal

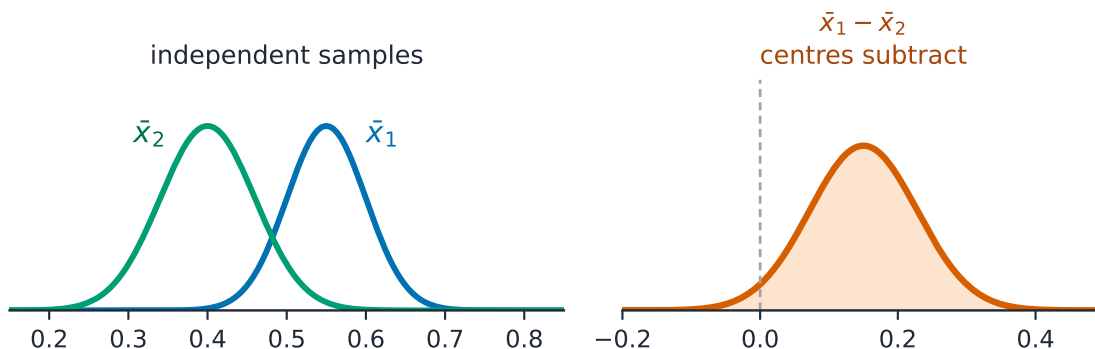
The population on the left is strongly skewed, yet every sampling distribution of $\bar{x}$ is centered at μ . A larger n shrinks the standard error $\sigma/\sqrt{n}$, so the curve gets taller and narrower – and it also straightens: still clearly skewed at $n=2$, almost exactly normal (dashed) by $n=30$.

Comparing Two Groups: Difference of Sample Means

For $\bar{x}_1 - \bar{x}_2$ from two independent samples: the mean is $\mu_1 - \mu_2$, and (independent, so variances add)

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

This is the foundation for two-sample inference in the next units.



$$\sigma = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The distribution of $\bar{x}_1 - \bar{x}_2$: centres subtract and variances add

Exam tips

- A **sampling distribution** is the distribution of a statistic over many samples, **centered on the true parameter**.
- The **Central Limit Theorem**: for a large enough sample the sample mean is approximately normal, even if the population is not.
- Larger samples give **less** variability (a smaller standard error).
- Check the conditions (random, independent/10%, large enough) before using a normal model.
- Keep straight what varies – the statistic – versus the fixed parameter.