DegreeLive

Bachelor of Commerce (B2001) | Level 1 • Part A Core | Monash University

Business and economic statistics

Course Code: ETC1000 • Module 3 Notes

Module 3: Statistical Inference: Point Estimation, Confidence Intervals and Hypothesis Testing

3.1 Principles of Point Estimation

In Module 2, population parameters such as the true mean μ or true default rate p were treated as known constants in order to deduce the sampling behaviour of sample statistics. In real commercial and economic research, however, population parameters are unknown and unobserved. **Statistical inference** reverses the direction of reasoning: analysts use a single observed random sample to estimate unknown population parameters and quantify the precision and reliability of those estimates.

Estimators versus Estimates and Finite-Sample Properties

An **estimator** ($\hat{\theta}$) is the general mathematical rule or formula—viewed as a random variable prior to data collection—applied to sample data $X_1, \dots, X_n$ to estimate a target population parameter θ. Once sample numbers are collected and plugged into the formula, the resulting specific numerical value is called a **point estimate**. For example, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the estimator of μ, whereas $\bar{x} = \text{AUD } 640.00$ computed from a Melbourne rental survey is the point estimate.

Because many competing formulas could be proposed to estimate the same parameter (for instance, sample mean $\bar{X}$ versus sample median versus midrange), mathematical statisticians evaluate estimators against three rigorous quality criteria:

- **Unbiasedness:** An estimator $\hat{\theta}$ is *unbiased* if the mean of its sampling distribution equals the true parameter across all possible parameter values: $E(\hat{\theta}) = \theta$. The quantity $\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$ measures systematic over- or under-estimation. As noted in Module 1, $\bar{X}$ is unbiased for μ, $\hat{p} = X/n$ is unbiased for p, and $s^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ is unbiased for σ^2, whereas dividing the sum of squares by n produces a downward-biased estimator of σ^2.
- **Efficiency (Minimum Variance):** Among all unbiased estimators of θ, the estimator with the smallest sampling variance $\text{Var}(\hat{\theta})$ is termed *efficient* (or Minimum Variance Unbiased). For a normal population, both the sample mean and sample median are unbiased estimators of μ, yet $\text{Var}(\bar{X}) = \sigma^2/n$ is strictly smaller than $\text{Var}(\text{Median}) \approx \frac{\pi\sigma^2}{2n} \approx 1.57\frac{\sigma^2}{n}$, making $\bar{X}$ more efficient.
- **Consistency:** An estimator $\hat{\theta}_n$ is *consistent* if it converges in probability to the true parameter θ as the sample size grows arbitrarily large ($\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1$ for every $\varepsilon > 0$). A sufficient condition for consistency of an unbiased estimator is that $\lim_{n\to\infty} \text{Var}(\hat{\theta}_n) = 0$. Because $\text{Var}(\bar{X}) = \sigma^2/n \to 0$ and $\text{Var}(\hat{p}) = p(1-p)/n \to 0$, both $\bar{X}$ and $\hat{p}$ are consistent estimators.

Constructing Estimators: Method of Moments and Maximum Likelihood Intuition

In econometrics and business statistics, estimators are derived systematically rather than guessed. The **Method of Moments** equates sample moments (such as $\bar{X}$) to population moments (μ) and solves for the parameter. **Maximum Likelihood Estimation (MLE)** selects the parameter value $\hat{\theta}_{\text{MLE}}$ that maximises the joint probability (or likelihood function $L(\theta \mid x_1, \dots, x_n)$) of having observed the actual sample data. For a normal distribution, MLE reproduces $\hat{\mu}_{\text{MLE}} = \bar{X}$ and yields $\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum(X_i - \bar{X})^2$, which is slightly downward-biased in finite samples but consistent and asymptotically efficient as $n \to \infty$.

Target Parameter	Standard Estimator	Expected Value	Unbiased?	Consistent?
Population Mean (μ)	$\bar{X} = \frac{1}{n}\sum X_i$	μ	Yes	Yes
Population Proportion (p)	$\hat{p} = X/n$	p	Yes	Yes
Population Variance (σ^2)	$s^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$	σ^2	Yes	Yes
Population Variance (σ^2)	$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum(X_i - \bar{X})^2$	$\frac{n-1}{n}\sigma^2$	No (downward-biased)	Yes

3.2 Interval Estimation: Confidence Intervals for the Population Mean

A point estimate provides no indication of how close it is likely to lie to the unknown parameter; due to sampling variability, a continuous point estimate such as $\bar{x}$ has probability zero of equalling μ exactly. A **confidence interval (CI)** resolves this limitation by pairing the point estimate with a **margin of error (E)**, constructing a plausible range $[L, U]$ at a chosen

degreelive

confidence level $(1-\alpha)100\%$, where $\alpha \in (0,1)$ is the tail error probability (commonly $\alpha = 0.10, 0.05, 0.01$ for 90%, 95%, and 99% confidence).

Frequentist Interpretation of Confidence Intervals

Properly interpreting a 95% confidence interval is essential in executive reporting. Before the sample is drawn, the random interval $[\bar{X} - E,\ \bar{X} + E]$ has probability 0.95 of bracketing the fixed unknown parameter μ. Once a specific numerical sample yields an observed interval—say, $[\text{AUD } 620.19,\ \text{AUD } 659.81]$—the parameter μ is either inside that fixed interval or outside it. Thus, the correct frequentist interpretation is: *if the identical sampling and interval-construction procedure were repeated independently across many samples of size n, approximately 95% of the resulting intervals would contain the true population mean* μ. Analysts must avoid the common misconception that "there is a 95% probability that μ lies inside this specific fixed interval", because μ is a fixed constant, not a random variable.

Case 1: Population Standard Deviation (σ) Known (z-Interval)

When σ is known from historical census benchmarks or calibrated manufacturing tolerances, and either the population is normal or $n \geq 30$ (by the Central Limit Theorem), the $(1-\alpha)100\%$ confidence interval for μ is:

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Where: $\bar{x}$ is the sample mean, $z_{\alpha/2}$ is the upper $(\alpha/2)$ critical value from the Standard Normal distribution (1.6449 for 90%, 1.9600 for 95%, and 2.5758 for 99% confidence), σ is the known population standard deviation, n is the sample size, and $E = z_{\alpha/2}(\sigma/\sqrt{n})$ is the margin of error. Increasing confidence $(1-\alpha)$ widens the interval (larger $z_{\alpha/2}$), whereas increasing n narrows the interval at rate $1/\sqrt{n}$.

Illustrative z-Interval & Sample-Size Planning Example: An automated parcel-sorting depot in Sydney records a long-run process standard deviation of $\sigma = 1.80$ kg for regional freight consignments. A sample of $n = 36$ consignments yields $\bar{x} = 14.20$ kg. For a 95% CI ($z_{0.025} = 1.96$), $\text{SE} = 1.80/\sqrt{36} = 0.30$ kg, $E = 1.96(0.30) = 0.588$ kg, giving $[13.61,\ 14.79]$ kg. If management requires halving the margin of error to $E = 0.25$ kg at 95% confidence, the required sample size is $n = \lceil (1.96 \times 1.80/0.25)^2 \rceil = \lceil (14.112)^2 \rceil = \lceil 199.15 \rceil = 200$ consignments.

Case 2: Population Standard Deviation (σ) Unknown (t-Interval) and Student's t-Distribution

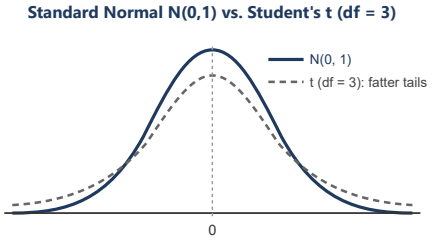
In nearly all empirical business studies, σ is unknown alongside μ and must be estimated by the sample standard deviation s. Substituting s for σ introduces an additional source of sampling variability, causing the standardised statistic $T = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ to follow **Student's t-distribution** with $\nu = n - 1$ **degrees of freedom** (discovered by William Sealy Gosset in 1908) whenever the underlying population is approximately normal (or n is large).

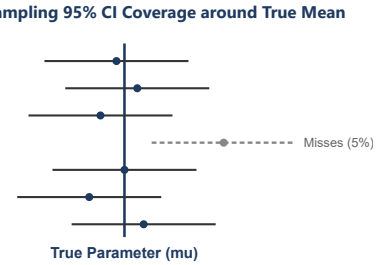
The Student's t-distribution shares the symmetry and zero mean of $N(0,1)$, but exhibits **heavier (fatter) tails** reflecting the added uncertainty of estimating σ with s. As the degrees of freedom $\nu = n - 1$ increase, s becomes a more precise estimate of σ, and the t_ν distribution converges smoothly to $N(0,1)$. With σ unknown, the $(1-\alpha)100\%$ confidence interval for μ is:

$$\bar{x} \pm t_{\alpha/2,\,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

Where: $\bar{x}$ is the sample mean, $t_{\alpha/2,\,n-1}$ is the upper $(\alpha/2)$ quantile of the t-distribution with $n-1$ degrees of freedom, s is the sample standard deviation, n is the sample size, and $s/\sqrt{n}$ is the estimated standard error of the mean.

DegreeLive





Worked Numerical Example 1: Inner-Melbourne Weekly Apartment Rents

A Victorian housing policy analyst samples $n = 25$ recently leased one-bedroom apartments in inner Melbourne and records a sample mean weekly rent of $\bar{x} = \text{AUD } 640.00$ with sample standard deviation $s = \text{AUD } 48.00$. Historical data indicate weekly rents in this segment are approximately symmetric without extreme skewness. Construct and interpret a 95% confidence interval for the true population mean weekly rent μ.

Step 1: Identify the appropriate distribution and critical value. Because σ is unknown and s is estimated from $n = 25$ observations, we use the t-distribution with $\nu = n - 1 = 24$ degrees of freedom. For 95% confidence, $\alpha = 0.05$ and $\alpha/2 = 0.025$. From the t-table, $t_{0.025,\,24} = 2.0639$.

Step 2: Compute the standard error, margin of error, and interval endpoints.

$$\text{SE}(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{48.00}{\sqrt{25}} = \frac{48.00}{5} = \text{AUD } 9.60$$

$$E = t_{0.025,\,24} \times \text{SE}(\bar{x}) = 2.0639 \times 9.60 = \text{AUD } 19.81$$

$$95\%\ \text{CI} = 640.00 \pm 19.81 = [\text{AUD } 620.19,\ \text{AUD } 659.81]$$

Where: $\text{SE}(\bar{x}) = 9.60$ is the estimated standard error of the sample mean, $t_{0.025,24} = 2.0639$ is the two-sided 95% t-critical value, and $E = 19.81$ is the half-width margin of error. The analyst reports with 95% confidence that the true mean weekly rent for inner-Melbourne one-bedroom apartments lies between A$620.19 and A$659.81.

Reference Critical Values of Student's t-Distribution and Precision Trade-Offs

The table below contrasts two-sided $t_{\alpha/2,\,\nu}$ critical values across degrees of freedom ν, demonstrating how t-multipliers converge toward Standard Normal $z_{\alpha/2}$ values as sample size grows:

Degrees of Freedom ($\nu = n - 1$)	90% CI ($t_{0.05,\nu}$)	95% CI ($t_{0.025,\nu}$)	99% CI ($t_{0.005,\nu}$)
5	2.0150	2.5706	4.0321
10	1.8125	2.2281	3.1693
15	1.7531	2.1315	2.9467
20	1.7247	2.0860	2.8453
24	1.7109	2.0639	2.7969
30	1.6973	2.0423	2.7500
60	1.6706	2.0003	2.6603
∞ (Standard Normal z)	1.6449	1.9600	2.5758

DegreeLive

Holding sample data fixed ($\bar{x} = 640$, $s = 48$, $n = 25$), widening confidence from 95% ($t = 2.0639$) to 99% ($t = 2.7969$) expands the margin of error from A$19.81 to A$26.85 (interval $[\text{AUD } 613.15,\ \text{AUD } 666.85]$), illustrating the inescapable trade-off between higher certainty and lower precision for a given sample size.

3.3 Confidence Intervals for Population Proportions and Sample Size Determination

Market researchers, auditors, and public-policy analysts frequently estimate population proportions (p), such as the share of Australian businesses experiencing cyber incidents, brand awareness percentages, or invoice error rates.

Large-Sample Confidence Interval for Population Proportion (p)

By the Central Limit Theorem (Section 2.7), when the sample contains at least 10 observed successes and 10 observed failures ($n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$), the sampling distribution of $\hat{p}$ is approximately normal. Replacing the unknown p in the standard error with its sample estimate $\hat{p}$ gives the large-sample Wald $(1-\alpha)100\%$ confidence interval for p:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Where: $\hat{p} = x/n$ is the sample proportion of successes, $z_{\alpha/2}$ is the Standard Normal critical value, n is the sample size, and $\sqrt{\hat{p}(1-\hat{p})/n}$ is the estimated standard error of $\hat{p}$. When p is very close to 0 or 1 (such as rare compliance breaches in banking audits) or n is modest, the Wald interval suffers poor coverage and can even overshoot $[0, 1]$; in those regimes, software computes the **Wilson score interval**, which centres and shrinks the interval naturally within $(0, 1)$.

Sample Size Planning Before Data Collection

Before commissioning a costly commercial survey or operational audit, managers specify a desired confidence level $(1-\alpha)$ and a maximum tolerable margin of error E, then solve backward for the minimum required sample size n (always rounding *up* to the next integer so the margin-of-error guarantee is met):

- **For estimating a population mean (μ):** Requires a planning estimate of σ (from a pilot study or historical data):

$$n = \left\lceil \left(\frac{z_{\alpha/2}\,\sigma}{E} \right)^2 \right\rceil$$

  Where: $z_{\alpha/2}$ is the normal critical value, σ is the planning standard deviation, E is the maximum allowable dollar or unit margin of error, and $\lceil \cdot \rceil$ denotes rounding up to the nearest integer.

- **For estimating a population proportion (p):** Uses a prior planning estimate p^* if available; if no reliable prior estimate exists, substituting $p^* = 0.50$ maximises the product $p^*(1-p^*) = 0.25$ and yields a conservative upper bound guaranteeing E for any true p:

$$n = \left\lceil \frac{z_{\alpha/2}^2\, p^*(1-p^*)}{E^2} \right\rceil$$

  Where: p^* is the planning proportion (set to 0.50 if unknown) and E is the target decimal margin of error (e.g., 0.025 for ± 2.5 percentage points).

Worked Numerical Example 2: PayID Adoption Interval and Survey Sizing

Part A (99% CI for Proportion): A payments industry survey of $n = 600$ randomly sampled Australian small businesses finds that $x = 438$ regularly accept PayID real-time bank transfers. Construct a 99% confidence interval for the national proportion p of small businesses using PayID.

Here $\hat{p} = 438/600 = 0.7300$. Normality holds comfortably since $n\hat{p} = 438 \geq 10$ and $n(1-\hat{p}) = 162 \geq 10$. For 99% confidence, $\alpha = 0.01$ and $z_{0.005} = 2.5758$:

DegreeLive

$$\text{SE}(\hat{p}) = \sqrt{\frac{0.73(0.27)}{600}} = \sqrt{0.0003285} = 0.01812$$

$$E = 2.5758 \times 0.01812 = 0.0467 \implies 99\%\ \text{CI} = 0.7300 \pm 0.0467 = [0.6833,\ 0.7767]$$

Where: $\hat{p} = 0.7300$ is the sample proportion and $E = 0.0467$ is the margin of error. With 99% confidence, between 68.33% and 77.67% of Australian small businesses accept PayID.

Part B (Required Sample Size): Next quarter, the researcher wants to estimate adoption in regional Queensland with 95% confidence ($z_{0.025} = 1.96$) and a margin of error no wider than $E = \pm 0.025$ (± 2.5 percentage points), assuming no reliable prior regional rate ($p^* = 0.50$):

$$n = \left\lceil \frac{(1.96)^2 \times 0.50 \times 0.50}{(0.025)^2} \right\rceil = \left\lceil \frac{3.8416 \times 0.25}{0.000625} \right\rceil = \lceil 1,536.64 \rceil = 1,537 \text{ businesses}$$

Where: 1,537 is the minimum sample size required to guarantee a $\pm 2.5\%$ margin of error at 95% confidence regardless of the true regional adoption rate.

3.4 Foundations of Hypothesis Testing

While confidence intervals estimate a parameter's plausible range, managers frequently need to test a specific quantitative claim, benchmark, or policy target—for example, whether an operational redesign reduced delivery defects, or whether average audit processing time exceeds a service-level guarantee. **Hypothesis testing** formalises this decision problem using proof-by-contradiction logic analogous to a criminal trial.

Formulating Null (H_0) and Alternative (H_1) Hypotheses

Every hypothesis test pairs two mutually exclusive, exhaustive statements about a population parameter θ (such as μ or p)—never about sample statistics:

- **Null Hypothesis (H_0):** Represents the status quo, baseline benchmark, or "no effect / no difference" claim. It always contains an equality relation ($=$, $\leq$, or $\geq$), written canonically with the equality boundary $\theta = \theta_0$. Under the testing framework, H_0 is maintained as provisionally true unless sample evidence overwhelmingly contradicts it.
- **Alternative (Research) Hypothesis (H_1 or H_a):** Represents the claim the researcher seeks evidence to support, containing a strict inequality ($\neq$, $>$, or $<$). The direction of H_1 determines the test tail structure:
  - **Two-tailed test:** $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ (detects departure in either direction).
  - **Right-tailed (upper-tailed) test:** $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ (detects an increase above θ_0).
  - **Left-tailed (lower-tailed) test:** $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$ (detects a decrease below θ_0).

Type I Error, Type II Error, and Statistical Power

Because decisions rest on random samples rather than complete censuses, any decision rule can err in two distinct ways, summarised in the decision matrix below:

Statistical Decision	H_0 is Actually True	H_0 is Actually False (H_1 True)
Fail to Reject H_0	Correct Decision (Probability $1 - \alpha$)	**Type II Error** (False Negative, Probability β)
Reject H_0	**Type I Error** (False Positive, Probability α)	Correct Decision: **Statistical Power** ($1 - \beta$)

The **significance level (α)** caps the maximum tolerable probability of a Type I error (wrongly rejecting a true H_0), conventionally set at $\alpha = 0.05$ (5%), $\alpha = 0.01$ (1%), or $\alpha = 0.10$ (10%). The probability of failing to detect a genuinely false H_0 is β, and $1 - \beta$ is the **statistical power** of the test. For a fixed sample size n, lowering α (say from 0.05 to 0.01) makes rejecting

degreelive

H_0 harder and therefore *increases* β (reduces power). The only way to reduce both α and β simultaneously is to increase the sample size n (or improve measurement precision to lower σ).

In Australian commerce and regulation, selecting α balances asymmetric error costs. When the Therapeutic Goods Administration (TGA) evaluates whether a new pharmaceutical outperforms placebo (H_0: no efficacy), a Type I error approves an ineffective or unsafe drug, justifying a stringent $\alpha = 0.01$ or 0.025 one-sided threshold. Conversely, in an aviation safety fatigue screen (H_0: component meets stress specification), a Type II error risks missing a defective turbine blade, prompting engineers to demand high power ($1 - \beta \geq 0.90$) and a more conservative rejection threshold (e.g., $\alpha = 0.10$).

Worked Calculation of Type II Error (β) and Statistical Power ($1 - \beta$)

Suppose an Australian beverage bottler specifies a target mean fill volume of $\mu_0 = 500$ mL with known process standard deviation $\sigma = 10$ mL. Quality inspectors test $H_0 : \mu \geq 500$ versus $H_1 : \mu < 500$ at $\alpha = 0.05$ using a random sample of $n = 25$ bottles (SE $= 10/\sqrt{25} = 2.0$ mL). Compute β and statistical power if the true mean fill actually drops to $\mu_1 = 496$ mL.

1. **Find the critical sample-mean boundary under H_0:** For a left-tailed test at $\alpha = 0.05$, $z_{0.05} = -1.6449$, so H_0 is rejected whenever $\bar{X} < \mu_0 - 1.6449(\text{SE}) = 500 - 1.6449(2.0) = 496.71$ mL, and H_0 is *not* rejected whenever $\bar{X} \geq 496.71$ mL.
2. **Evaluate β under the true alternative $\mu_1 = 496$ mL:** A Type II error occurs when we fail to reject H_0 despite $\mu_1 = 496$ holding:

$$z_\beta = \frac{496.71 - 496.00}{2.0} = +0.3550 \implies \beta = P(\bar{X} \geq 496.71 \mid \mu = 496) = 1 - \Phi(0.3550) = 0.3613 \ (36.13\%)$$

Where: $\beta = 0.3613$ is the probability of missing the 4 mL under-fill defect. Consequently, the statistical power of the $n = 25$ inspection plan to detect a 4 mL shortfall is $1 - \beta = 1 - 0.3613 = 0.6387$ (63.87%).

3.5 The p-Value Approach, Critical Value Approach, and CI Duality

A test statistic standardises the discrepancy between the observed sample estimate and the hypothesised null value θ_0 in standard-error units: Test Statistic $= (\text{Estimate} - \theta_0)/\text{SE}$. Two equivalent decision methods convert that test statistic into a formal conclusion.

Critical Value versus p-Value Decision Rules

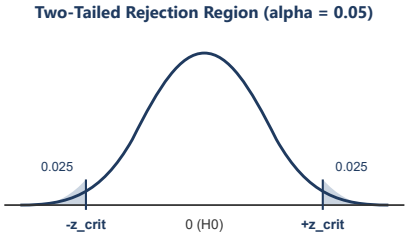
In the **critical value approach**, the analyst partitions the sampling distribution under H_0 into a non-rejection region and a **rejection region** of total area α (split as $\alpha/2$ in each tail for two-tailed tests, or placed entirely in the upper/lower tail for one-tailed tests). If the observed test statistic falls inside the rejection region, H_0 is rejected.

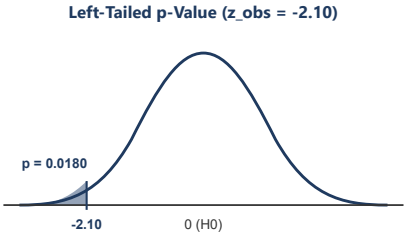
In the modern ***p*-value approach**, the analyst computes the exact probability, *assuming H_0 is true*, of observing a test statistic at least as extreme (in the direction supporting H_1) as the value actually observed in the sample:

- Right-tailed ($H_1 : \theta > \theta_0$): $p\text{-value} = P(T \geq t_{\text{obs}} \mid H_0)$.
- Left-tailed ($H_1 : \theta < \theta_0$): $p\text{-value} = P(T \leq t_{\text{obs}} \mid H_0)$.
- Two-tailed ($H_1 : \theta \neq \theta_0$): $p\text{-value} = 2 \times P(T \geq |t_{\text{obs}}| \mid H_0)$.

Universal Decision Rule: Reject H_0 at significance level α if and only if $p\text{-value} \leq \alpha$; otherwise, fail to reject H_0. Analysts must avoid misinterpreting the p-value as "the probability that H_0 is true". Rather, it is a conditional probability $P(\text{Data as extreme or more extreme} \mid H_0 \text{ is true})$ measuring evidence incompatibility against H_0.

degreelive





Statistical Significance versus Practical Significance and CI Duality

A subtle but crucial duality connects two-sided hypothesis tests and confidence intervals: a two-tailed test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ rejects H_0 at significance level α if and only if θ_0 falls *outside* the corresponding $(1-\alpha)100\%$ confidence interval. Furthermore, analysts must distinguish **statistical significance** (whether an observed difference is unlikely due to pure sampling noise) from **practical (economic) significance** (whether the effect magnitude matters financially). With huge digital transaction datasets ($n = 500,000$), an average basket lift of A\$0.12 can yield $p < 0.001$ (statistically significant) while failing to cover the engineering cost of deploying a website change (economically trivial).

To prevent misreporting in corporate analytics and economic briefs, the table below contrasts frequentist misconceptions against correct statistical interpretations:

Statement	Common Misconception (Incorrect)	Rigorous Frequentist Interpretation (Correct)
p-value $= 0.018$	"There is a 1.8% probability that H_0 is true."	If H_0 were true, a sample discrepancy at least this extreme would occur in 1.8% of repeated samples.
Fail to reject H_0 ($p = 0.14$)	"We have proven that H_0 is true (zero effect)."	Sample data lack sufficient statistical evidence against H_0 at level α; a small real effect may remain undetected (Type II risk).
95% CI = $[620, 660]$	"95% of apartments rent between A\$620 and A\$660."	The interval estimates the *population mean* μ, not individual observations; 95% of intervals built by this method cover μ.

3.6 One-Sample Hypothesis Tests for Means and Proportions

Applying hypothesis testing systematically follows a five-step discipline: (1) state H_0 and H_1 clearly in terms of population parameters; (2) specify α and select the test statistic after verifying distributional assumptions; (3) determine the critical value or p-value rule; (4) compute the test statistic from sample data; and (5) state both the statistical decision and its plain-language business conclusion.

One-Sample Tests for Mean (μ) and Proportion (p)

- **One-sample z-test for μ (σ known):**

$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Where: $\bar{x}$ is the sample mean, μ_0 is the hypothesised null mean, σ is the known population standard deviation, and n is the sample size.

- **One-sample t-test for μ (σ unknown):**

$$t_{\text{stat}} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \sim t_{n-1} \text{ under } H_0$$

Where: s is the sample standard deviation and $n - 1$ is the degrees of freedom, assuming normality (or large n via the CLT).

- **One-sample z-test for proportion (p):** Crucially, because the test statistic is evaluated *under the assumption that* $H_0 : p = p_0$ *is true*, the standard error uses the null proportion p_0 rather than $\hat{p}$ (provided $np_0 \geq 10$ and $n(1-p_0) \geq 10$):

$$z_{\text{stat}} = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$

Where: $\hat{p} = x/n$ is the sample proportion, p_0 is the hypothesised null proportion, and n is the sample size.

Worked Numerical Example 3: Superannuation Administrative Fees (t-Test)

A consumer advocacy group investigates whether the average annual fixed administration fee charged across Australian balanced MySuper accounts exceeds the advertised industry benchmark of A$450 per year. A random sample of $n = 16$ funds yields $\bar{x} = \text{AUD } 474.00$ and $s = \text{AUD } 42.00$. Test at $\alpha = 0.05$ assuming fee levels are approximately normally distributed.

Step 1 (Hypotheses): $H_0 : \mu \leq 450$ versus $H_1 : \mu > 450$ (right-tailed test).

Step 2 (Test Statistic & Critical Value): Because σ is unknown and $n = 16$, use a t-test with $\nu = 16 - 1 = 15$ degrees of freedom. For a right-tailed test at $\alpha = 0.05$, the critical value is $t_{0.05,\,15} = +1.7531$ (reject H_0 if $t_{\text{stat}} \geq 1.7531$).

Step 3 (Calculation):

$$\text{SE}(\bar{x}) = \frac{42.00}{\sqrt{16}} = \frac{42.00}{4} = \text{AUD } 10.50 \qquad \Longrightarrow \qquad t_{\text{stat}} = \frac{474.00 - 450.00}{10.50} = \frac{24.00}{10.50} = +2.2857$$

Where: $\text{SE}(\bar{x}) = 10.50$ is the estimated standard error and $t_{\text{stat}} = +2.2857$ is the observed t-statistic. From t-tables with $df = 15$, $t_{0.025,15} = 2.1315$ and $t_{0.01,15} = 2.6025$, so $0.01 < p\text{-value} < 0.025$ (exact p-value $= 0.0186$).

Step 4 (Decision & Conclusion): Since $t_{\text{stat}} = 2.2857 > 1.7531$ (equivalently $p = 0.0186 \leq 0.05$), we reject H_0. There is statistically significant evidence at the 5% level that average annual administration fees exceed A$450.

Worked Numerical Example 4: Online Cart Abandonment (z-Test for Proportion)

Historically, a Melbourne online fashion retailer experienced a 65% ($p_0 = 0.65$) checkout cart-abandonment rate. After introducing one-click Apple Pay and PayID checkout, an analyst tracks $n = 400$ independent checkout sessions and observes 240 abandoned carts ($\hat{p} = 240/400 = 0.60$). Test at $\alpha = 0.05$ whether abandonment has fallen significantly.

Step 1 (Hypotheses & Conditions): $H_0 : p \geq 0.65$ versus $H_1 : p < 0.65$ (left-tailed test). Normality holds since $np_0 = 400(0.65) = 260 \geq 10$ and $n(1-p_0) = 140 \geq 10$.

Step 2 (Calculation): Compute the null standard error and z-statistic:

$$\text{SE}_0 = \sqrt{\frac{0.65(0.35)}{400}} = \sqrt{0.00056875} = 0.02385 \qquad \Longrightarrow \qquad z_{\text{stat}} = \frac{0.60 - 0.65}{0.02385} = -2.0966$$

Where: $\text{SE}_0 = 0.02385$ is the standard error evaluated under $p_0 = 0.65$ and $z_{\text{stat}} = -2.0966$ is the test statistic. The left-tailed p-value is $\Phi(-2.0966) = 0.0180$.

Step 3 (Conclusion): Since $p\text{-value} = 0.0180 \leq 0.05$ (and $z_{\text{stat}} = -2.0966 < -z_{0.05} = -1.6449$), reject H_0. The checkout redesign reduced cart abandonment below the historical 65% benchmark.

Assumption Diagnostics and Robustness of t-Procedures

Before relying on any small-sample t-test or t-confidence interval ($n < 30$), analysts inspect histograms, boxplots, or normal Q-Q plots of the sample data (or paired differences). While the t-procedure is remarkably **robust to mild non-normality** whenever distributions are roughly symmetric, severe skewness or extreme outliers in small samples invalidate normal-theory p-values by

DegreeLive

distorting both $\bar{x}$ and s. When strong skewness persists for small n, analysts either apply a monotonic transformation (such as analysing $\ln(\text{income})$ or $\ln(\text{house price})$) or employ distribution-free **non-parametric tests** such as the Wilcoxon signed-rank test (for one-sample or paired comparisons) or the Mann–Whitney U test (for two independent samples).

3.7 Two-Sample Inference: Independent Samples and Paired Comparisons

Business experiments and comparative economic evaluations routinely compare two populations rather than one fixed benchmark—for instance, comparing customer spend across two cities, evaluating productivity before and after software deployment, or contrasting conversion rates in an A/B trial. Choosing the correct two-sample procedure hinges on whether the two samples are **independent** or **paired (dependent)**.

Independent Samples: Welch's Unpooled versus Pooled Two-Sample t-Test

When two random samples of sizes n_1 and n_2 are drawn independently from two distinct populations with means μ_1, μ_2 and unknown variances σ_1^2, σ_2^2, modern statistical practice defaults to **Welch's (unpooled) two-sample t-test** because it does not require the restrictive assumption that $\sigma_1^2 = \sigma_2^2$:

$$t_{\text{Welch}} = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where: $\bar{x}_1, \bar{x}_2$ are the two sample means, D_0 is the hypothesised difference $\mu_1 - \mu_2$ (typically 0), s_1^2, s_2^2 are the sample variances, and n_1, n_2 are the sample sizes. Welch's test uses the Satterthwaite approximate degrees of freedom:

$$\nu \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Where: ν is the effective degrees of freedom (typically rounded down to an integer if using printed tables, or left exact in software; bounded between $\min(n_1 - 1, n_2 - 1)$ and $n_1 + n_2 - 2$). Only when strong evidence supports equal population variances ($\sigma_1^2 = \sigma_2^2$) should analysts employ the **pooled-variance t-test**, which replaces s_1^2 and s_2^2 with the weighted pooled variance $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ and exact $df = n_1 + n_2 - 2$.

Paired-Samples t-Test for Dependent Observations

When observations in two groups are naturally linked—such as measuring the same 10 bank branches before and after staff training, or matching twin retail stores across control and promotional weeks—samples are **dependent (paired)**. Treating paired data as independent wastes statistical power because unit-to-unit baseline heterogeneity inflates s_1^2 and s_2^2. Instead, compute the within-pair difference $d_i = x_{1i} - x_{2i}$ for each of the n pairs and run a standard one-sample t-test on the differences:

$$t_{\text{paired}} = \frac{\bar{d} - \mu_{d,0}}{s_d / \sqrt{n}} \sim t_{n-1} \text{ under } H_0 \qquad \text{and} \qquad \text{CI for } \mu_d: \ \bar{d} \pm t_{\alpha/2,\, n-1} \left(\frac{s_d}{\sqrt{n}}\right)$$

Where: $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$ is the mean of the n paired differences, s_d is the sample standard deviation of the differences, $\mu_{d,0}$ is the hypothesised mean difference (usually 0), and n is the number of matched pairs.

Two-Sample z-Test for Difference Between Proportions ($p_1 - p_2$)

To test $H_0 : p_1 - p_2 = 0$ across two large independent binomial samples, combine successes under H_0 into the **pooled sample proportion** $\bar{p} = (x_1 + x_2)/(n_1 + n_2)$:

DegreeLive

$$z_{\text{stat}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where: $\hat{p}_1 = x_1/n_1$ and $\hat{p}_2 = x_2/n_2$ are the two sample proportions and $\bar{p}$ is the pooled proportion under $H_0 : p_1 = p_2$. (By contrast, constructing a confidence interval for $p_1 - p_2$ does not assume $p_1 = p_2$ and therefore uses the unpooled standard error $\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$.)

Worked Numerical Example 5: Independent Store Comparison and Paired Branch Productivity

Part A (Welch's Independent Two-Sample *t*-Test): A retail group compares mean daily sales (A$ thousands) between its Sydney flagship ($n_1 = 18$ trading days, $\bar{x}_1 = 24.6$, $s_1 = 3.2$) and Brisbane flagship ($n_2 = 15$ trading days, $\bar{x}_2 = 21.8$, $s_2 = 2.8$). Test $H_0 : \mu_1 - \mu_2 = 0$ versus $H_1 : \mu_1 - \mu_2 \neq 0$ at $\alpha = 0.05$.

$$\text{SE}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{3.2^2}{18} + \frac{2.8^2}{15}} = \sqrt{0.5689 + 0.5227} = \sqrt{1.0916} = 1.0448$$

$$t_{\text{Welch}} = \frac{24.6 - 21.8}{1.0448} = \frac{2.8}{1.0448} = +2.6800$$

Evaluating Satterthwaite's degrees of freedom yields $\nu \approx 30.91$ (rounded down conservatively to 30). At $\alpha = 0.05$ (two-tailed), $t_{0.025,\,30} = 2.0423$. Since $|2.6800| > 2.0423$ ($p \approx 0.0117 \leq 0.05$), reject H_0: mean daily revenue differs significantly between Sydney and Brisbane.

Part B (Paired-Samples *t*-Test): An Australian bank measures daily mortgage applications processed across $n = 10$ regional branches before and after deploying an automated document-verification platform. The paired differences ($d_i = \text{After}_i - \text{Before}_i$) are: $+4, +6, +2, +5, +7, +3, +5, +1, +6, +4$. Test whether mean processing volume increased ($H_0 : \mu_d \leq 0$ versus $H_1 : \mu_d > 0$) at $\alpha = 0.01$.

$$\bar{d} = \frac{43}{10} = 4.30, \qquad s_d = \sqrt{\frac{\sum_{i=1}^{10}(d_i - 4.30)^2}{9}} = 1.8886, \qquad \text{SE}(\bar{d}) = \frac{1.8886}{\sqrt{10}} = 0.5972$$

$$t_{\text{paired}} = \frac{4.30 - 0}{0.5972} = +7.2001$$

Where: $\bar{d} = 4.30$ is the average daily gain per branch and $s_d = 1.8886$ is the standard deviation of within-branch differences. With $df = n - 1 = 9$ and $\alpha = 0.01$ (right-tailed), the critical value is $t_{0.01,\,9} = 2.8214$. Since $7.2001 > 2.8214$ ($p < 0.0001$), reject H_0. Moreover, a 95% confidence interval for the mean productivity lift is $4.30 \pm 2.2622(0.5972) = [2.95,\ 5.65]$ applications per day.

Part C (Two-Sample Proportion *z*-Test & CI): An Australian e-commerce platform runs a randomised A/B trial comparing checkout conversion under its existing layout ($n_1 = 500$, $x_1 = 110$ conversions, $\hat{p}_1 = 0.220$) against a streamlined Express Checkout layout ($n_2 = 500$, $x_2 = 140$ conversions, $\hat{p}_2 = 0.280$). Test $H_0 : p_1 - p_2 = 0$ versus $H_1 : p_1 - p_2 \neq 0$ at $\alpha = 0.05$ and construct a 95% CI for $p_1 - p_2$.

DegreeLive

$$\bar{p} = \frac{110 + 140}{500 + 500} = 0.250, \qquad \text{SE}_{\text{pooled}} = \sqrt{0.25(0.75)\left(\frac{1}{500} + \frac{1}{500}\right)} = 0.02739$$

$$z_{\text{stat}} = \frac{0.220 - 0.280}{0.02739} = -2.1909 \implies p\text{-value} = 2 \times \Phi(-2.1909) = 0.0285 \leq 0.05$$

Because $p = 0.0285 \leq 0.05$, we reject H_0: conversion rates differ significantly. Using the unpooled standard error $\sqrt{0.22(0.78)/500 + 0.28(0.72)/500} = 0.02732$, the 95% CI for $p_1 - p_2$ is $(0.220 - 0.280) \pm 1.96(0.02732) = [-0.1135,\ -0.0065]$, indicating Express Checkout lifts conversion by between 0.65 and 11.35 percentage points.

Part D (Pooled-Variance Independent *t*-Test): Suppose a logistics operator compares order-picking times (minutes) between its Melbourne ($n_1 = 12$, $\bar{x}_1 = 18.4$, $s_1 = 2.1$) and Sydney ($n_2 = 10$, $\bar{x}_2 = 20.2$, $s_2 = 2.4$) distribution centres, where long-term engineering audits support equal process variances ($\sigma_1^2 = \sigma_2^2$). Testing $H_0 : \mu_1 - \mu_2 = 0$ versus $H_1 : \mu_1 - \mu_2 \neq 0$ at $\alpha = 0.05$ ($df = 12 + 10 - 2 = 20$, $t_{0.025,\,20} = 2.0860$):

$$s_p^2 = \frac{(12-1)(2.1)^2 + (10-1)(2.4)^2}{12 + 10 - 2} = \frac{48.51 + 51.84}{20} = 5.0175 \implies s_p = \sqrt{5.0175} = 2.2400$$

$$t_{\text{pooled}} = \frac{18.4 - 20.2}{2.2400\sqrt{\frac{1}{12} + \frac{1}{10}}} = \frac{-1.80}{0.9591} = -1.8768$$

Where: $s_p^2 = 5.0175$ is the pooled sample variance and 0.9591 is the pooled standard error. Since $|-1.8768| = 1.8768 < 2.0860$ ($p \approx 0.0752 > 0.05$), we fail to reject H_0 at the 5% level: there is insufficient statistical evidence of a difference in mean picking time between the Melbourne and Sydney centres.

Summary Reference Matrix of Hypothesis Testing Procedures

Inference Setting	Parameter	Test Statistic	Null Distribution & Conditions
One-sample mean (σ known)	μ	$(\bar{x} - \mu_0)/(\sigma/\sqrt{n})$	$N(0,1)$; normal pop. or $n \geq 30$
One-sample mean (σ unknown)	μ	$(\bar{x} - \mu_0)/(s/\sqrt{n})$	t_{n-1}; normal pop. or large n
One-sample proportion	p	$(\hat{p} - p_0)/\sqrt{p_0(1-p_0)/n}$	$N(0,1)$; $np_0 \geq 10, n(1-p_0) \geq 10$
Two independent means (unequal var.)	$\mu_1 - \mu_2$	$(\bar{x}_1 - \bar{x}_2)/\sqrt{s_1^2/n_1 + s_2^2/n_2}$	t_ν (Welch-Satterthwaite df)
Paired / matched-pair means	μ_d	$\bar{d}/(s_d/\sqrt{n})$	t_{n-1}; paired differences normal or large n
Two independent proportions	$p_1 - p_2$	$(\hat{p}_1 - \hat{p}_2)/\sqrt{\bar{p}(1-\bar{p})(1/n_1 + 1/n_2)}$	$N(0,1)$; expected cell counts ≥ 10

Degreelive