DegreeLive

Bachelor of Commerce (B2001) | Level 2 • Part B Major | Monash University

Statistical thinking

Course Code: ETC2420 • Module 2 Notes

Module 2: Computational Inference: Bootstrap, Permutation Tests & Resampling

1. Sampling Variability and the Limits of Analytical Formulas

Every sample statistic $\hat{\theta} = T(X_1, \ldots, X_n)$ computed from random data is itself a random variable. If an analyst repeated an Australian labour-force survey or sampled a fresh cohort of Melbourne retail stores under identical conditions, the numerical estimate $\hat{\theta}$ would differ from sample to sample. The probability distribution of $\hat{\theta}$ across all hypothetical independent samples of size n from the population distribution F is called the **sampling distribution** of $\hat{\theta}$.

1.1 Bias, Variance, Mean Squared Error, and Standard Error

To evaluate how accurately an estimator $\hat{\theta}$ targets an unknown population parameter θ, we decompose its expected squared error into systematic bias and sampling variance:

$$\text{Bias}(\hat{\theta}) = \text{E}[\hat{\theta}] - \theta, \qquad \text{MSE}(\hat{\theta}) = \text{E}[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2$$

Where: $\text{E}[\hat{\theta}]$ is the expectation of the estimator across repeated samples; $\text{Bias}(\hat{\theta})$ measures systematic over- or under-estimation; $\text{Var}(\hat{\theta})$ measures sample-to-sample dispersion around $\text{E}[\hat{\theta}]$; and $\text{MSE}(\hat{\theta})$ is the Mean Squared Error combining both sources of inaccuracy.

The **standard error** of an estimator, $\text{SE}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}$, quantifies the typical magnitude of sampling fluctuation. For the sample mean $\bar{X}$ of i.i.d. observations with finite population standard deviation σ, the exact analytical standard error is $\text{SE}(\bar{X}) = \sigma/\sqrt{n}$, consistently estimated by $s/\sqrt{n}$. Furthermore, by the Central Limit Theorem, $(\bar{X} - \mu)/(s/\sqrt{n})$ converges to $\text{N}(0, 1)$ as n grows large.

1.2 Why Analytical Asymptotics Break Down

Classical textbook inference relies heavily on closed-form variance formulas and large-sample Gaussian approximations. In applied data analysis, these analytical shortcuts frequently become intractable or misleading:

- **Non-linear or non-smooth estimators:** Statistics such as the sample median, interquartile range (IQR), 10% trimmed mean, Gini inequality index, Sharpe ratio, or ratio of two correlated totals $\hat{R} = \bar{Y}/\bar{X}$ either lack exact finite-sample standard error formulas or depend on unknown population density values. For example, the asymptotic variance of the sample median $\hat{m}$ is $\text{Var}(\hat{m}) \approx [4nf(m)^2]^{-1}$, which requires estimating the unknown density f at the true median m—a notoriously noisy task in small samples.
- **Ratio estimators and the Delta Method:** For a smooth non-linear function $g(\bar{X})$, a first-order Taylor expansion gives the Delta Method approximation $\text{Var}(g(\bar{X})) \approx [g'(\mu)]^2 \text{Var}(\bar{X})$. While useful asymptotically, this linearisation breaks down when sample size is modest or g exhibits strong curvature near μ.
- **Small samples with pronounced skewness:** Australian household wealth, personal income, and insurance claim severities exhibit extreme right skewness. When $n = 20$ or $n = 30$, the sampling distribution of $\bar{X}$ remains noticeably skewed, rendering symmetric t-intervals $\bar{x} \pm t_{n-1,\,0.975}(s/\sqrt{n})$ inaccurate in one or both tails.
- **Complex multi-stage estimation pipelines:** When an analyst imputes missing values, winsorises outliers, and fits a regularised model, analytical standard errors ignore uncertainty induced by earlier pipeline steps.

Computational resampling methods—principally the **bootstrap**, **jackknife**, and **permutation tests**—replace restrictive analytical derivations with direct computational simulation from the observed data.

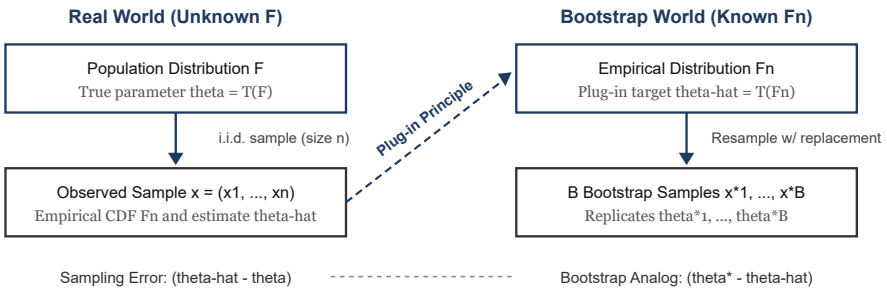
Degreelive

2. The Non-Parametric Bootstrap, Parametric Bootstrap, and Jackknife

Introduced by Bradley Efron in 1979, the **non-parametric bootstrap** operationalises a remarkably simple idea: since the true population distribution F is unknown, use the empirical distribution $\hat{F}_n$ of the observed sample as a stand-in for F. This substitution is known as the **plug-in principle**.

2.1 Efron's Plug-In Principle and Bootstrap Resampling

In the real world, an unknown population distribution F generates an i.i.d. sample $\mathbf{x} = (x_1, \dots, x_n)$, from which we compute $\hat{\theta} = s(\mathbf{x})$. In the **bootstrap world**, the empirical distribution $\hat{F}_n$ places equal probability mass $1/n$ on each observed data point x_i. Drawing a new i.i.d. sample of size n from $\hat{F}_n$ is algebraically identical to drawing n observations **uniformly at random with replacement** from the original sample $\{x_1, \dots, x_n\}$.



Notice the subtle but vital role of the bootstrap replicates: the bootstrap does *not* generate new information beyond the sample, so the centre of the bootstrap replicates $\bar{\theta}^*$ cannot improve upon $\hat{\theta}$ as an estimate of location unless bias-adjusted. What the bootstrap accurately mimics—provided $\hat{F}_n$ is a good approximation to F—is the **shape and spread** of the sampling fluctuations $(\hat{\theta} - \theta)$ via the observable fluctuations $(\theta^* - \hat{\theta})$. Notice also that any individual bootstrap sample contains roughly $1 - (1 - 1/n)^n \approx 1 - e^{-1} \approx 63.2\%$ of the distinct original observations (some repeated multiple times), leaving about 36.8% out-of-bag.

2.2 Bootstrap Algorithm, Standard Error, and Bias Estimation

Given an original i.i.d. sample $\mathbf{x} = (x_1, \dots, x_n)$ and estimator $\hat{\theta} = s(\mathbf{x})$, the non-parametric bootstrap executes the following steps:

1. For $b = 1, 2, \dots, B$: draw a bootstrap sample $\mathbf{x}^{*(b)} = (x_1^{*(b)}, \dots, x_n^{*(b)})$ of size n with replacement from $\{x_1, \dots, x_n\}$.
2. Compute the b-th bootstrap replicate statistic $\hat{\theta}^{*(b)} = s(\mathbf{x}^{*(b)})$.
3. Approximate the sampling standard error of $\hat{\theta}$ using the sample standard deviation of the B bootstrap replicates:

$$\mathrm{SE}_{\mathrm{boot}}(\hat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^{*(b)} - \bar{\theta}^*\right)^2}, \qquad \bar{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*(b)}$$

Where: B is the number of bootstrap resamples; $\hat{\theta}^{*(b)}$ is the statistic evaluated on the b-th resample; $\bar{\theta}^*$ is the average of the B bootstrap replicates; and $\mathrm{SE}_{\mathrm{boot}}(\hat{\theta})$ is the non-parametric bootstrap standard error.

The bootstrap also estimates finite-sample bias by comparing the bootstrap expectation $\bar{\theta}^*$ against the plug-in target $\hat{\theta}$:

$$\widehat{\mathrm{Bias}}_{\mathrm{boot}}(\hat{\theta}) = \bar{\theta}^* - \hat{\theta}, \qquad \hat{\theta}_{\mathrm{BC}} = \hat{\theta} - \widehat{\mathrm{Bias}}_{\mathrm{boot}}(\hat{\theta}) = 2\hat{\theta} - \bar{\theta}^*$$

Where: $\widehat{\mathrm{Bias}}_{\mathrm{boot}}(\hat{\theta})$ estimates the finite-sample bias $\mathrm{E}[\hat{\theta}] - \theta$; and $\hat{\theta}_{\mathrm{BC}}$ is the bias-corrected point estimate. Note that bias correction subtracts an estimated noise term and can increase variance slightly; analysts typically apply bias correction only when the estimated bias exceeds roughly $0.25\,\mathrm{SE}_{\mathrm{boot}}$.

Because the bootstrap uses B finite Monte Carlo draws from $\hat{F}_n$ rather than enumerating all n^n possible resamples, $\mathrm{SE}_{\mathrm{boot}}$ exhibits simulation variability proportional to $1/\sqrt{B}$. For standard error estimation, $B = 1{,}000$ to $2{,}000$ is usually sufficient; for tail quantiles in 95% or 99% confidence intervals, $B = 5{,}000$ to $10{,}000$ is recommended to keep Monte Carlo jitter negligible.

Worked Example 2.1: Bootstrap Standard Error for Median Daily Cafe Revenue

Suppose a Melbourne suburban cafe records daily takings (in AUD) across $n = 5$ randomly selected weekdays: $\mathbf{x} = (1{,}420,\ 1{,}580,\ 1{,}650,\ 1{,}810,\ 2{,}940)$. Notice the outlier day of $\$2{,}940$, which makes the median a more representative measure of typical daily revenue than the mean. The ordered original sample gives the sample median $\hat{\theta} = 1{,}650$.

To illustrate the exact mechanics, suppose we draw $B = 6$ bootstrap resamples of size $n = 5$ with replacement (in practice, software uses $B \geq 2{,}000$):

Resample b	Sorted Bootstrap Sample $\mathbf{x}^{*(b)}$	Bootstrap Median $\hat{\theta}^{*(b)}$	Deviation $(\hat{\theta}^{*(b)} - \bar{\theta}^{*})$
1	(1,420, 1,580, 1,580, 1,650, 2,940)	1,580	−138.33
2	(1,580, 1,650, 1,810, 1,810, 2,940)	1,810	+91.67
3	(1,420, 1,420, 1,650, 1,810, 1,810)	1,650	−68.33
4	(1,650, 1,650, 1,810, 2,940, 2,940)	1,810	+91.67
5	(1,420, 1,580, 1,650, 1,650, 1,810)	1,650	−68.33
6	(1,580, 1,810, 1,810, 1,810, 2,940)	1,810	+91.67

Step 1: Compute the bootstrap mean $\bar{\theta}^{*} = (1{,}580 + 1{,}810 + 1{,}650 + 1{,}810 + 1{,}650 + 1{,}810)/6 = 10{,}310/6 = 1{,}718.33$.

Step 2: Compute the sum of squared deviations and bootstrap SE:

$$\sum_{b=1}^{6} \left(\hat{\theta}^{*(b)} - \bar{\theta}^{*}\right)^2 = (-138.33)^2 + 3(91.67)^2 + 2(-68.33)^2 = 19{,}136.11 + 25{,}208.33 + 9{,}338.89 = 53{,}683.33$$

$$\mathrm{SE}_{\mathrm{boot}} = \sqrt{\frac{53{,}683.33}{6-1}} = \sqrt{10{,}736.67} \approx 103.62$$

Where: $\mathrm{SE}_{\mathrm{boot}} \approx \$103.62$ is the bootstrap standard error of the sample median across these 6 resamples. Observe that the outlier $\$2{,}940$ barely perturbs the bootstrap medians—most resamples yield $\$1{,}580$, $\$1{,}650$, or $\$1{,}810$—demonstrating why the median has much smaller sampling variability than the mean under heavy outliers.

2.3 Parametric Bootstrap and Jackknife Resampling

When strong domain knowledge supports a specific parametric family $F(x;\theta)$—for instance, fitting an Exponential or Gamma model to call-centre waiting times—the **parametric bootstrap** replaces the empirical distribution $\hat{F}_n$ with the fitted parametric distribution $F(x;\hat{\theta}_{\mathrm{MLE}})$. Instead of resampling rows from $\mathbf{x}$, the algorithm simulates B synthetic datasets $\mathbf{x}^{*(b)} \overset{\text{i.i.d.}}{\sim} F(\cdot;\hat{\theta}_{\mathrm{MLE}})$, refits the MLE $\hat{\theta}^{*(b)}$ on each synthetic dataset, and evaluates the spread of $\hat{\theta}^{*(b)}$ or any derived quantity such as the 99-th percentile tail risk. The parametric bootstrap produces smoother quantile estimates in small samples and overcomes the discrete support limitations of $\hat{F}_n$, provided the parametric family itself is not misspecified.

Predating the bootstrap, Quenouille and Tukey's **jackknife** is a deterministic leave-one-out resampling method. Let $\hat{\theta}_{(-i)} = s(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ denote the statistic computed with the i-th observation deleted, and let $\bar{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{(-i)}$. The jackknife bias and standard error estimators are:

DegreeLive

$$\widehat{\text{Bias}}_{\text{jack}} = (n-1)\,(\bar{\theta}_{(\cdot)} - \hat{\theta}), \qquad \widehat{\text{SE}}_{\text{jack}}(\hat{\theta}) = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(-i)} - \bar{\theta}_{(\cdot)})^2}$$

Where: $\hat{\theta}_{(-i)}$ is the i-th leave-one-out replicate; $\bar{\theta}_{(\cdot)}$ is the average of all n leave-one-out replicates; and the inflation factor $(n-1)$ reflects the strong overlap between samples differing in only one observation. Note that the jackknife works well for smooth statistics (means, variances, regression coefficients) and supplies the acceleration parameter a in BC_a intervals, but fails on non-smooth statistics like the sample median where deleting a single point produces erratic jumps.

Worked Example 2.2: Jackknife Bias and Standard Error for Sample Variance Ratio

Suppose $n = 4$ observations of weekly percentage return on an ASX index are $\mathbf{x} = (1.0,\ 2.0,\ 3.0,\ 6.0)$, with sample mean $\bar{x} = 3.0$ and biased MLE variance $\hat{\sigma}^2_{\text{MLE}} = \frac{1}{4}[(-2)^2 + (-1)^2 + 0^2 + 3^2] = \frac{14}{4} = 3.50$. Evaluate the jackknife bias and standard error of $\hat{\theta} = \hat{\sigma}^2_{\text{MLE}}$.

- Delete $x_1 = 1.0$: subset $(2, 3, 6)$, mean $11/3$, $\hat{\theta}_{(-1)} = \frac{1}{3}[(2-11/3)^2 + (3-11/3)^2 + (6-11/3)^2] = \frac{26}{9} \approx 2.8889$.
- Delete $x_2 = 2.0$: subset $(1, 3, 6)$, mean $10/3$, $\hat{\theta}_{(-2)} = \frac{1}{3}[(-7/3)^2 + (-1/3)^2 + (8/3)^2] = \frac{114}{27} \approx 4.2222$.
- Delete $x_3 = 3.0$: subset $(1, 2, 6)$, mean 3.0, $\hat{\theta}_{(-3)} = \frac{1}{3}[(-2)^2 + (-1)^2 + 3^2] = \frac{14}{3} \approx 4.6667$.
- Delete $x_4 = 6.0$: subset $(1, 2, 3)$, mean 2.0, $\hat{\theta}_{(-4)} = \frac{1}{3}[(-1)^2 + 0^2 + 1^2] = \frac{2}{3} \approx 0.6667$.

The average is $\bar{\theta}_{(\cdot)} = (2.8889 + 4.2222 + 4.6667 + 0.6667)/4 = 3.1111$. Thus:

$$\widehat{\text{Bias}}_{\text{jack}} = (4-1)(3.1111 - 3.5000) = 3(-0.3889) = -1.1667$$

$$\hat{\theta}_{\text{jack, BC}} = 3.5000 - (-1.1667) = 4.6667 = \frac{14}{3} = s^2$$

Remarkably, the jackknife bias correction automatically converts the biased MLE variance $\frac{1}{4}\sum(x_i - \bar{x})^2 = 3.50$ into the exact unbiased sample variance $s^2 = \frac{1}{3}\sum(x_i - \bar{x})^2 = 4.6667$! Meanwhile, $\widehat{\text{SE}}_{\text{jack}} = \sqrt{\frac{3}{4}[(-0.2222)^2 + (1.1111)^2 + (1.5556)^2 + (-2.4444)^2]} = \sqrt{\frac{3}{4}(9.6790)} \approx 2.694$.

3. Bootstrap Confidence Intervals and Structural Resampling

Converting a bootstrap distribution $\{\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(B)}\}$ into a $(1-\alpha)$ confidence interval can be performed in several ways. Choosing carefully among these methods matters whenever the sampling distribution of $\hat{\theta}$ is skewed or biased.

3.1 Five Bootstrap Confidence Interval Constructions

1. Normal-Approximation (Standard) Bootstrap Interval

Uses SE_{boot} inside the classical symmetric Gaussian interval:

$$\text{CI}_{\text{norm}} = [\hat{\theta} - z_{1-\alpha/2}\,\text{SE}_{\text{boot}},\ \ \hat{\theta} + z_{1-\alpha/2}\,\text{SE}_{\text{boot}}]$$

Where: $\hat{\theta}$ is the original sample estimate; $z_{1-\alpha/2}$ is the standard Normal quantile (e.g., 1.96 for 95%); and SE_{boot} is the bootstrap standard error. This interval relies on $\hat{\theta}$ being unbiased and approximately Normal, so it sacrifices the bootstrap's ability to capture asymmetry.

DegreeLive

2. Percentile Bootstrap Interval

Reads the $\alpha/2$ and $1-\alpha/2$ empirical quantiles directly from the sorted bootstrap replicates $\hat{\theta}^*_{(1)} \leq \cdots \leq \hat{\theta}^*_{(B)}$:

$$\text{CI}_{\text{perc}} = [\hat{\theta}^*_{(\alpha/2)},\ \hat{\theta}^*_{(1-\alpha/2)}]$$

Where: $\hat{\theta}^*_{(q)}$ denotes the q-th empirical quantile of the B bootstrap replicates. The percentile interval is transformation-respecting (taking logs or exponentiating the endpoints matches bootstrapping on the transformed scale) and naturally accommodates skewness when $\hat{\theta}$ is nearly unbiased.

3. Basic (Pivotal / Reverse-Percentile) Bootstrap Interval

Uses the bootstrap distribution of the root $\delta^* = \hat{\theta}^* - \hat{\theta}$ to approximate the true error $\delta = \hat{\theta} - \theta$. Requiring $P\left(\delta^*_{\alpha/2} \leq \hat{\theta} - \theta \leq \delta^*_{1-\alpha/2}\right) \approx 1-\alpha$ and solving for θ reverses the order of the quantiles:

$$\text{CI}_{\text{basic}} = [2\hat{\theta} - \hat{\theta}^*_{(1-\alpha/2)},\ 2\hat{\theta} - \hat{\theta}^*_{(\alpha/2)}]$$

Where: $\hat{\theta}$ is the original estimate; and $\hat{\theta}^*_{(1-\alpha/2)}$ and $\hat{\theta}^*_{(\alpha/2)}$ are the upper and lower quantiles of the bootstrap replicates. Notice the reflection around $\hat{\theta}$: if the bootstrap distribution has a long right tail ($\hat{\theta}^*_{(1-\alpha/2)} - \hat{\theta}$ is large), the basic interval extends further to the *left* of $\hat{\theta}$ because a right-skewed sampling distribution implies that large upward errors $\hat{\theta} > \theta$ are more plausible, meaning θ may lie substantially below an observed $\hat{\theta}$.

4. Studentised (Bootstrap-t) Interval

Instead of pivoting $\hat{\theta} - \theta$, the bootstrap-t pivots the approximately scale-free Studentised statistic $t^*_b = (\hat{\theta}^{*(b)} - \hat{\theta})/\text{SE}^{*(b)}$, where $\text{SE}^{*(b)}$ is the standard error estimated inside resample b (either analytically or via an inner nested bootstrap). Using empirical quantiles $t^*_{(\alpha/2)}$ and $t^*_{(1-\alpha/2)}$ yields:

$$\text{CI}_{\text{stud}} = [\hat{\theta} - t^*_{(1-\alpha/2)}\,\text{SE}_{\text{boot}},\ \hat{\theta} - t^*_{(\alpha/2)}\,\text{SE}_{\text{boot}}]$$

Where: $t^*_{(q)}$ is the q-th quantile of the simulated bootstrap-t pivots. Like the classical t-interval, studentisation automatically widens intervals in small samples to account for variance estimation uncertainty while preserving skewness, achieving second-order accuracy $O(n^{-1})$.

5. Bias-Corrected and Accelerated (BC_a) Interval

Developed by Efron (1987), the BC_a interval adjusts the percentile endpoints α_1, α_2 using two parameters: a **bias-correction factor** z_0 measuring median bias, and an **acceleration constant** a computed from jackknife replicates $\hat{\theta}_{(-i)}$, which captures how the standard error of $\hat{\theta}$ varies with θ:

$$z_0 = \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} \text{I}(\hat{\theta}^{*(b)} < \hat{\theta})\right), \qquad a = \frac{\sum_{i=1}^{n}\left(\bar{\theta}_{(\cdot)} - \hat{\theta}_{(-i)}\right)^3}{6\left[\sum_{i=1}^{n}\left(\bar{\theta}_{(\cdot)} - \hat{\theta}_{(-i)}\right)^2\right]^{3/2}}$$

$$\alpha_1 = \Phi\left(z_0 + \frac{z_0 + z_{\alpha/2}}{1 - a(z_0 + z_{\alpha/2})}\right), \qquad \alpha_2 = \Phi\left(z_0 + \frac{z_0 + z_{1-\alpha/2}}{1 - a(z_0 + z_{1-\alpha/2})}\right)$$

DegreeLive

Where: Φ and Φ^{-1} are the standard Normal CDF and quantile function; z_0 adjusts for median bias (when $z_0 = a = 0$, $\alpha_1 = \alpha/2$ and $\alpha_2 = 1 - \alpha/2$, recovering the ordinary percentile CI); and $\text{CI}_{\text{BCa}} = [\hat{\theta}^*_{(\alpha_1)},\ \hat{\theta}^*_{(\alpha_2)}]$. BC_a intervals achieve second-order accuracy without needing nested bootstraps and serve as the standard choice in modern statistical software.

Method	Formula	Captures Skewness	Accuracy Order	Key Caveat
Normal Approx.	$\hat{\theta} \pm z_{1-\alpha/2}\text{SE}_{\text{boot}}$	No (symmetric)	$O(n^{-1/2})$	Poor coverage when sampling distribution is skewed
Percentile	$[\hat{\theta}^*_{(\alpha/2)},\ \hat{\theta}^*_{(1-\alpha/2)}]$	Yes	$O(n^{-1/2})$	Undercovers if $\hat{\theta}$ has finite-sample bias
Basic (Pivotal)	$[2\hat{\theta} - \hat{\theta}^*_{(1-\alpha/2)},\ 2\hat{\theta} - \hat{\theta}^*_{(\alpha/2)}]$	Yes (reflected)	$O(n^{-1/2})$	Can occasionally extend slightly outside bounded support
Studentised (t)	$[\hat{\theta} - t^*_{(1-\alpha/2)}\text{SE},\ \hat{\theta} - t^*_{(\alpha/2)}\text{SE}]$	Yes	$O(n^{-1})$	Requires inner SE calculation; unstable under heavy outliers
BC_a	$[\hat{\theta}^*_{(\alpha_1)},\ \hat{\theta}^*_{(\alpha_2)}]$ via (z_0, a)	Yes	$O(n^{-1})$	Requires $B \geq 2{,}000$ plus n jackknife evaluations

Worked Example 3.1: Contrasting Percentile and Basic Bootstrap 95% CIs

Suppose an analyst estimates the 90-th percentile of Australian small-business invoice settlement times (in days) from $n = 60$ invoices, obtaining $\hat{\theta} = 48.0$ days. Running $B = 5{,}000$ bootstrap resamples yields a right-skewed replicate distribution whose 2.5% and 97.5% quantiles are $\hat{\theta}^*_{(0.025)} = 41.5$ days and $\hat{\theta}^*_{(0.975)} = 61.0$ days.

Step 1: Compute the 95% Percentile Bootstrap CI directly from the replicate quantiles:

$$\text{CI}_{\text{perc}} = [41.5,\ 61.0] \text{ days}$$

Where: the lower margin is $48.0 - 41.5 = 6.5$ days below $\hat{\theta}$, and the upper margin is $61.0 - 48.0 = 13.0$ days above $\hat{\theta}$.

Step 2: Compute the 95% Basic (Pivotal) Bootstrap CI by reflecting around $2\hat{\theta} = 96.0$:

$$\text{CI}_{\text{basic}} = [2(48.0) - 61.0,\ \ 2(48.0) - 41.5] = [35.0,\ 54.5] \text{ days}$$

Where: the lower margin extends 13.0 days below $\hat{\theta}$ and the upper margin extends 6.5 days above $\hat{\theta}$. Notice how the two methods disagree on direction of asymmetry; when $\hat{\theta}$ is a location/scale quantile with skewness-dependent variance, BC_a reconciles this discrepancy automatically.

3.2 Regression, Dependent Data, and Bootstrap Failure Modes

When data possess structure beyond a single i.i.d. scalar sample, the resampling unit must preserve the dependence structure of the DGP:

Data / Model Structure	Resampling Strategy	Key Validity Condition
Two independent groups	Resample separately within Group A (n_A) and Group B (n_B)	Independence between and within groups
Regression: Cases (Paired)	Resample full rows $(y_i, \mathbf{x}_i)$ with replacement	Valid under heteroskedasticity and random $\mathbf{x}_i$
Regression: Residual	Fix $\mathbf{x}_i$, resample centred residuals e_i^* to form $y_i^* = \mathbf{x}_i'\hat{\beta} + e_i^*$	Requires homoskedastic, i.i.d. error terms

DegreeLive

Data / Model Structure	Resampling Strategy	Key Validity Condition
Time series (autocorrelated)	Moving or stationary **Block Bootstrap** (blocks of length ℓ)	Stationarity; block length ℓ balances dependence vs. variance
Clustered / Panel data	Cluster bootstrap (resample entire clusters/firms/schools)	Independence across clusters; preserves within-cluster correlation

Finally, the naive non-parametric bootstrap **fails** in several well-known regimes: (1) **extreme order statistics** such as $\hat{\theta} = \max(X_1, \ldots, X_n)$ for a $\text{Uniform}(0, \theta)$ distribution, because bootstrap resamples can never exceed the observed maximum and put probability $1 - (1 - 1/n)^n \approx 63.2\%$ on $\hat{\theta}^* = \hat{\theta}$; (2) **heavy-tailed distributions with infinite variance** (e.g., Cauchy or Pareto with tail index ≤ 2), where standard resamples fail to converge unless subsampling ($m \ll n$ without replacement) is used; and (3) **very small samples** ($n < 10$) where $\hat{F}_n$ is too coarse to represent F.

4. Permutation and Randomisation Hypothesis Tests

While the bootstrap targets estimation uncertainty (standard errors and confidence intervals) without assuming a null hypothesis, **permutation tests** (originally conceived by R. A. Fisher in the 1930s) target hypothesis testing directly by constructing the exact null reference distribution through re-labelling.

4.1 The Sharp Null Hypothesis and Exchangeability

Consider comparing two groups—for example, a randomised trial comparing Treatment ($D_i = 1$, n_1 units) against Control ($D_i = 0$, n_0 units). Under the **sharp null hypothesis** H_0, treatment has literally zero effect on every single individual unit: each unit's observed outcome y_i would have been identical regardless of which group it was assigned to. Equivalently, in an observational setting comparing two populations, $H_0 : F_1 = F_0$ implies that the combined $N = n_1 + n_0$ observations are **exchangeable** under H_0: any partition of the pooled outcomes into n_1 Treatment labels and n_0 Control labels is equally likely.

Because group labels carry no signal under H_0, there are $\binom{N}{n_1}$ equally probable assignments. Evaluating the test statistic T (such as difference in means $\bar{y}_1 - \bar{y}_0$, difference in medians, or a trimmed difference) across all partitions yields the exact finite-sample null distribution of T—without assuming Normality, equal variances, or large n.

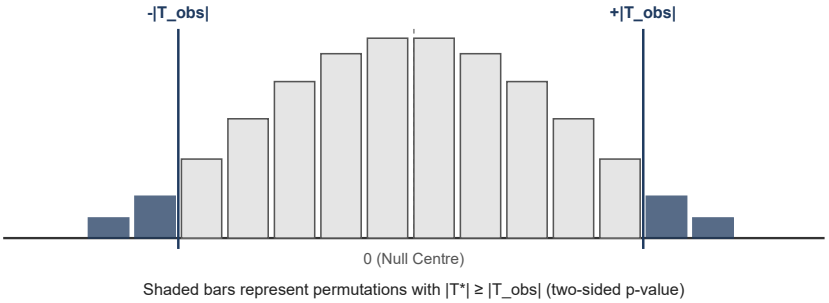
4.2 Exact and Monte Carlo Permutation p-Values

When $\binom{N}{n_1}$ is small (e.g., $\binom{10}{5} = 252$), every permutation can be enumerated exhaustively. When N is larger (e.g., $\binom{40}{20} \approx 1.38 \times 10^{11}$), we approximate the permutation distribution by randomly shuffling the group labels B times (typically $B = 10{,}000$) and computing $T^{*(b)}$ for each shuffle. Including the observed data partition itself as one valid arrangement ensures exact Type I error control via the $+1$ rule:

$$p_{\text{perm}} = \frac{1 + \sum_{b=1}^{B} \mathrm{I}\left(\left|T^{*(b)}\right| \geq \left|T_{\text{obs}}\right|\right)}{B + 1}$$

Where: T_{obs} is the test statistic computed on the actual observed assignment; $T^{*(b)}$ is the test statistic under the b-th random permutation of labels; $\mathrm{I}(\cdot)$ is the indicator function equal to 1 if true and 0 otherwise; and B is the number of random shuffles. Adding 1 to both numerator and denominator guarantees that p_{perm} can never equal zero and preserves α-level validity even for moderate B.

Degreelive



Worked Example 4.1: Exact Two-Group Permutation Test for Checkout Redesign

An Australian e-commerce retailer conducts a pilot randomisation of $N = 6$ regional shoppers into two checkout interfaces: Control A ($n_A = 3$) and Redesigned B ($n_B = 3$), measuring checkout completion time (in seconds). Observed times are:

- **Group A (Control):** $44,\ 51,\ 49$ (mean $\bar{y}_A = 48.0$ s).
- **Group B (Redesign):** $36,\ 40,\ 38$ (mean $\bar{y}_B = 38.0$ s).

The observed difference in means is $T_{\text{obs}} = \bar{y}_A - \bar{y}_B = 48.0 - 38.0 = +10.0$ seconds. Under H_0 (interface has no effect on any shopper), the pooled set of 6 times is $\{36, 38, 40, 44, 49, 51\}$, whose grand sum is 258.

Step 1: Count total combinations: $\binom{6}{3} = \frac{6\times5\times4}{3\times2\times1} = 20$ equally likely allocations of 3 items to Group A. Because $\bar{y}_A - \bar{y}_B = \frac{S_A}{3} - \frac{258-S_A}{3} = \frac{2S_A-258}{3}$, where S_A is the sum of the 3 observations assigned to Group A, $|T^*| \geq 10.0$ occurs whenever $S_A \geq 144$ or $S_A \leq 114$.

Step 2: Enumerate the extreme subsets of size 3 from $\{36, 38, 40, 44, 49, 51\}$: only $\{44, 49, 51\}$ achieves $S_A = 144$ ($T^* = +10.0$), and only $\{36, 38, 40\}$ achieves $S_A = 114$ ($T^* = -10.0$). The next closest subsets $\{40, 49, 51\}$ ($S_A = 140, T^* = +7.33$) and $\{36, 38, 44\}$ ($S_A = 118, T^* = -7.33$) fall strictly inside $(-10, +10)$.

Step 3: Compute the exact two-sided permutation p-value:

$$p_{\text{exact}} = \frac{2}{20} = 0.10$$

Where: 2 is the number of allocations out of $\binom{6}{3} = 20$ producing $|T^*| \geq 10.0$. Even though every observation in B is strictly faster than every observation in A, $n_A = n_B = 3$ can never yield a two-sided exact p-value below $2/20 = 0.10$. This reveals a critical design insight: small sample sizes impose a mathematical floor $2/\binom{N}{n_1}$ on achievable permutation significance regardless of effect size.

4.3 Paired Sign-Flipping, Correlation Permutations, and Unequal Variances

The permutation logic generalises seamlessly beyond two independent groups:

- **Testing association (Pearson or Spearman correlation):** Given n pairs (x_i, y_i), under H_0 of independence between X and Y, holding x_i fixed and randomly permuting the y_i values breaks any true pairing while keeping both marginal distributions intact. Recomputing $r^{*(b)}$ across B shuffles yields an exact non-parametric test of zero association.
- **Paired designs (One-sample randomisation / sign-flipping test):** When each subject receives both treatments (producing paired differences $d_i = y_{i,1} - y_{i,0}$), under the sharp null of zero treatment effect each difference d_i is symmetric around zero, meaning its sign ± 1 is equally likely across 2^n possible sign assignments.
- **Studentised permutation test under heteroskedasticity:** Standard permutation tests of $\bar{y}_1 - \bar{y}_0$ require exchangeability under H_0 ($F_1 = F_0$, including equal variances). If variances differ sharply between two groups even when means are equal, permuting raw labels can inflate Type I error. Using the Welch studentised statistic $T_{\text{Welch}} = (\bar{y}_1 - \bar{y}_0) / \sqrt{s_1^2/n_1 + s_0^2/n_0}$ inside the permutation loop restores asymptotic Type I error validity even when variances are unequal.

Degree Live

Worked Example 4.2: Exact Paired Sign-Flipping Test on Store Productivity

Five Melbourne retail stores trial an automated stock-shelving tool for one month, recording paired weekly labour hours saved minus baseline: $\mathbf{d} = (+4.0,\ +6.0,\ -1.0,\ +5.0,\ +6.0)$. Here $\sum d_i = 20.0$ and observed mean difference $\bar{d}_{\text{obs}} = 20.0/5 = +4.0$ hours/week.

Step 1: Count total sign assignments: there are $2^5 = 32$ equally likely vectors of signs $(s_1, \dots, s_5) \in \{-1, +1\}^5$ under H_0. Note that $\bar{d}^* \geq 4.0$ requires signed sum $S^* \geq 20.0$, and $\bar{d}^* \leq -4.0$ requires $S^* \leq -20.0$.

Step 2: Identify extreme sign assignments:

- All positive $(+4, +6, +1, +5, +6)$: sum $= +22$ ($|\bar{d}^*| = 4.4 \geq 4.0$).
- Observed signs $(+4, +6, -1, +5, +6)$: sum $= +20$ ($|\bar{d}^*| = 4.0 \geq 4.0$).
- Flipping any other single element reduces the sum below 20 (e.g., flipping -4 drops sum from 22 to 14).
- By symmetry, the two mirrored negative assignments $(-4, -6, -1, -5, -6)$ and $(-4, -6, +1, -5, -6)$ yield sums -22 and -20.

Step 3: Compute the exact two-sided paired randomisation p-value: $p_{\text{exact}} = \frac{4}{32} = 0.125$.

5. Monte Carlo Simulation and Computational Power Studies

Whereas the bootstrap resamples from an empirical sample $\hat{F}_n$, **Monte Carlo simulation** generates synthetic datasets directly from a fully specified probabilistic model F_θ using pseudo-random numbers. Monte Carlo methods enable analysts to compute intractable probabilities, evaluate the finite-sample performance of estimators, and design adequately powered studies.

5.1 Inverse Transform Sampling and Acceptance-Rejection

Modern statistical environments generate high-quality pseudo-random draws $U \sim \text{Uniform}(0, 1)$. The **Probability Integral Transform** guarantees that if X has continuous strictly increasing CDF F, then $U = F(X) \sim \text{Uniform}(0, 1)$. Inverting this relationship gives **inverse transform sampling**: to simulate an observation from target CDF F, draw $u \sim \text{Uniform}(0, 1)$ and set:

$$x = F^{-1}(u)$$

Where: $u \in (0, 1)$ is a standard uniform draw; and F^{-1} is the quantile function (inverse CDF) of the target distribution.

Worked Example 5.1: Simulating Exponential Insurance Claim Inter-Arrivals

Suppose fire-damage claims arrive at an Australian insurer at rate $\lambda = 0.25$ claims per day ($\text{Exp}(0.25)$, mean 4 days). Because $F(x) = 1 - e^{-\lambda x}$, solving $u = 1 - e^{-\lambda x}$ yields $F^{-1}(u) = -\frac{1}{\lambda}\ln(1 - u)$. Given four pseudo-random uniform draws $u_1 = 0.18,\ u_2 = 0.72,\ u_3 = 0.55,\ u_4 = 0.91$:

- $x_1 = -4\ln(0.82) \approx 0.794$ days; $\quad x_2 = -4\ln(0.28) \approx 5.092$ days;
- $x_3 = -4\ln(0.45) \approx 3.194$ days; $\quad x_4 = -4\ln(0.09) \approx 9.632$ days.

When F^{-1} lacks a closed form, **acceptance-rejection sampling** draws a candidate y from an easy proposal density $g(y)$ satisfying $f(y) \leq Mg(y)$ for a constant $M \geq 1$, draws $u \sim \text{Uniform}(0, 1)$, and accepts $x = y$ if $u \leq \frac{f(y)}{Mg(y)}$ (otherwise rejecting and repeating). On average, one draw is accepted every M proposals.

5.2 Monte Carlo Integration and Variance Reduction

Suppose we wish to evaluate a complex expectation or probability $p = P(g(\mathbf{X}) \in A) = \text{E}[\text{I}(g(\mathbf{X}) \in A)]$ where analytical integration is impossible (for example, the probability that a multi-asset ASX portfolio experiences a drawdown exceeding 15% over 20 trading days under correlated heavy-tailed shocks). Generating M independent synthetic trajectories $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(M)}$ yields the Monte Carlo estimator and its simulation standard error:

DegreeLive

$$\hat{p}_M = \frac{1}{M}\sum_{m=1}^{M} \mathrm{I}(g(\mathbf{X}^{(m)}) \in A), \qquad \mathrm{SE}_{\mathrm{MC}}(\hat{p}_M) = \sqrt{\frac{\hat{p}_M(1-\hat{p}_M)}{M}}$$

Where: M is the number of independent Monte Carlo replications; $\hat{p}_M$ is the simulated event proportion; and $\mathrm{SE}_{\mathrm{MC}}$ measures the precision of the Monte Carlo approximation itself. Note that $\mathrm{SE}_{\mathrm{MC}} \leq 0.5/\sqrt{M}$: achieving a margin of error below 0.005 requires $M \geq 10{,}000$ runs, whereas achieving 0.001 requires $M = 250{,}000$ runs.

To improve precision without multiplying M, analysts employ **variance-reduction techniques**: *antithetic variates* pair each uniform draw U with $1-U$ to induce negative correlation between paired replicates, shrinking the variance of their average; and *control variates* subtract a correlated zero-mean random variable whose exact expectation is known analytically.

5.3 Designing a Monte Carlo Study for Coverage and Power

Statistical researchers use Monte Carlo simulation as a laboratory to test whether inference procedures behave as advertised when textbook assumptions are violated. A disciplined simulation study specifies true parameters θ_0 and sample sizes n, simulates M independent datasets using a fixed deterministic seed, applies candidate estimators and CIs to each dataset, and aggregates empirical bias, RMSE, empirical coverage probability $\frac{1}{M}\sum \mathrm{I}(L^{(m)} \leq \theta_0 \leq U^{(m)})$, and rejection rates (Type I error under H_0; power under H_1).

Sample Size n (Data ~ Exp(1), $\mu_0 = 1$)	**Classical 95% t-CI Coverage**	**Percentile Bootstrap 95% Coverage**	**BC$_a$ Bootstrap 95% Coverage**
$n = 15$ (strongly skewed $\bar{X}$)	$\approx 90.2\%$ (undercovers)	$\approx 91.4\%$	$\approx 93.6\%$
$n = 30$ (moderately skewed $\bar{X}$)	$\approx 92.6\%$	$\approx 93.2\%$	$\approx 94.4\%$
$n = 100$ (CLT regime)	$\approx 94.5\%$	$\approx 94.6\%$	$\approx 94.9\%$

As summarised in this benchmark comparison for skewed Exponential data, classical t-intervals undercover substantially at $n = 15$ because one-sided tail skewness breaks Gaussian symmetry, whereas second-order BC$_a$ intervals adapt their left and right tail bounds to restore coverage rapidly toward the nominal 95% target.

Worked Example 2.3: Parametric Bootstrap for Exponential Tail Risk

Recall the Melbourne call-centre waiting times from Module 1 ($n = 10$, sample mean $\bar{x} = 1.90$ minutes, $\hat{\lambda}_{\mathrm{MLE}} = 1/1.90 = 0.5263$ calls/min). Suppose operations management needs a 95% confidence interval not for the mean waiting time, but for the 90-th percentile waiting time $\theta = Q_{0.90} = -\frac{1}{\lambda}\ln(1-0.90) = \frac{\ln(10)}{\lambda} \approx 4.375$ minutes—a tail SLA target.

Step 1: Specify the parametric bootstrap DGP as $\mathrm{Exp}(\hat{\lambda} = 0.5263)$.

Step 2: For $b = 1, \ldots, B$ (e.g., $B = 5{,}000$), simulate $n = 10$ synthetic observations $x_1^{*(b)}, \ldots, x_{10}^{*(b)} \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(0.5263)$ using inverse transform draws $-1.90\ln(1-u_i)$, compute the replicate sample mean $\bar{x}^{*(b)}$ and MLE $\hat{\lambda}^{*(b)} = 1/\bar{x}^{*(b)}$, and record the parametric quantile replicate $\hat{\theta}^{*(b)} = \ln(10)\,\bar{x}^{*(b)}$.

Step 3: Notice that because $\bar{X}^* \sim \mathrm{Gamma}(10,\ 10\hat{\lambda})$ analytically here, we can benchmark the simulation directly: $\mathrm{E}[\hat{\theta}^*] = 4.375$ minutes and $\mathrm{SE}(\hat{\theta}^*) = 4.375/\sqrt{10} = 1.383$ minutes, with exact 2.5% and 97.5% quantiles of the $\mathrm{Gamma}(10, 10)$ pivot yielding a skewed 95% parametric bootstrap CI of $[2.10,\ 7.47]$ minutes—appropriately asymmetric and much wider on the upper tail than a naive Normal interval $4.375 \pm 1.96(1.383) = [1.66,\ 7.09]$.

5.4 Computational Power Analysis Under Non-Normal Distributions

Analytical sample-size formulas assume Normal outcomes with known variance. When outcomes are skewed, zero-inflated, or heavy-tailed—such as daily customer spend on an Australian e-commerce site where 60% spend $\$0$ and spending buyers follow a Log-Normal distribution—analysts determine statistical power via **Monte Carlo power simulation**:

1. **Formulate realistic baseline and alternative DGPs:** Simulate Control observations $Y_{0,i}$ from the empirical or fitted skewed distribution (e.g., zero-inflated Log-Normal) and Treatment observations $Y_{1,i}$ incorporating the hypothesised

DegreeLive

uplift δ.

2. **Run M simulated trials at candidate sample size n:** On each synthetic trial $m = 1, \dots, M$ (say $M = 2{,}000$), apply the exact planned hypothesis test (e.g., Welch t-test, Mann-Whitney test, or permutation test) and record $p^{(m)}$.
3. **Estimate power and iterate n:** Compute $\widehat{\text{Power}}(n) = \frac{1}{M}\sum_{m=1}^{M} \mathrm{I}(p^{(m)} \le \alpha)$, with Monte Carlo SE $\sqrt{\widehat{\text{Power}}(1-\widehat{\text{Power}})/M}$. Increase n until $\widehat{\text{Power}}(n) \ge 0.80$.

DGP Scenario ($n_1 = n_0 = 40$, $\alpha = 0.05$)	Welch t-Test Power	Permutation (Mean) Power	Wilcoxon Rank-Sum Power
Symmetric Normal ($\delta = 0.5\sigma$)	59.6%	59.4%	57.1% (slight loss)
Heavy-tailed Student-t_3 (outliers)	41.2%	42.0%	61.8% (much higher)
Right-skewed Log-Normal (multiplicative lift)	46.5%	47.3%	68.4% (much higher)

This Monte Carlo comparison demonstrates why simulation-based power analysis is indispensable: under symmetric Normal data, parametric and permutation tests are nearly indistinguishable, whereas under heavy-tailed or Log-Normal commercial data, rank-based or log-transformed permutation tests achieve far higher power while strictly controlling Type I error.

DegreeLive