Lecture 12 - **Permutation tests, the Jackknife and the Bootstrap**

BIOENG-210 Course Notes
Prof. Gioele La Manno

May 2025

Contents

1 Computer Age Statistical Inference

1.1 The Computational Revolution in Statistics

The field of statistics has undergone a profound transformation over the past few decades, driven largely by the exponential increase in computational power. This revolution has fundamentally changed how we approach statistical inference, moving beyond traditional analytical methods toward algorithm-based approaches that harness computational resources.

In the early days of statistics, methods were constrained by computational limitations. Statistical procedures needed to be mathematically tractable, often requiring simplifying assumptions about data distributions, independence, or homogeneity.

The construction of standard errors, confidence intervals, and hypothesis tests typically relied on analytical formulas that could be calculated efficiently with the tools at hand.

Today, we find ourselves in what Hastie, Tibshirani, and Friedman aptly called "the computer age of statistics" - an era where intensive computation allows us to tackle previously intractable problems through simulation and algorithmic approaches. Instead of deriving exact formulas for sampling distributions, we can simulate them directly.

This computational turn allows tacking situations where traditional methods struggle or fail:

- When standard parametric assumptions are violated (e.g., non-normality, heteroscedasticity)
- When deriving analytical solutions is not mathematically tractable (e.g., complex models, high-dimensional data)
- When the data structure is complex (e.g., measurements nested within subjects, within families, within populations)
- Problems requiring complex estimators without available analytical solutions (e.g., statistics derived from clustering algorithms)
- Estimation of complex parameters with intractable sampling distributions (e.g., entropy measures, performance scores)
- Small sample sizes where asymptotic approximations may be unreliable
- When dealing with censored, truncated, or otherwise irregularly sampled data

In such scenarios, computational approaches to inference provide robust alternatives to traditional methods. Rather than forcing our biological questions to fit within the constraints of analytical methods, we can develop computational procedures that directly address the specific challenges of our data.

In this lecture, we explore fundamental computational approaches to statistical inference that leverage the power of resampling: permutation tests, the jackknife, and the bootstrap (in both its non-parametric and parametric variants). These methods share a common philosophy - using computational algorithms to approximate sampling distributions and construct inference without relying on strong distributional assumptions or asymptotic approximations.

2 Permutation Tests

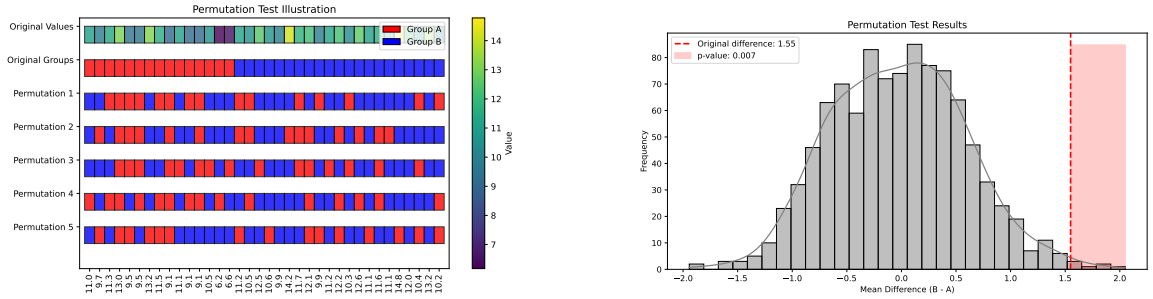
2.1 The Logic of Permutation Testing

Permutation tests offer a powerful, distribution-free approach to hypothesis testing that predates both the jackknife and bootstrap. The core insight, articulated by R.A. Fisher in the 1930s, is elegant in its simplicity: if a null hypothesis of no effect is true, then the observed group labels are

essentially arbitrary, and reassigning observations randomly should produce test statistics similar to what we observed.

Definition 2.1 (Permutation Test)**.** A permutation test involves:

1. Calculate the test statistic T for the original data
2. Generate M permutations of the data consistent with the null hypothesis (typically by shuffling group labels)
3. Calculate the test statistic T_m for each permutation
4. The p-value is the proportion of permutations where T_m is as or more extreme than T



(a) Permutation test illustration

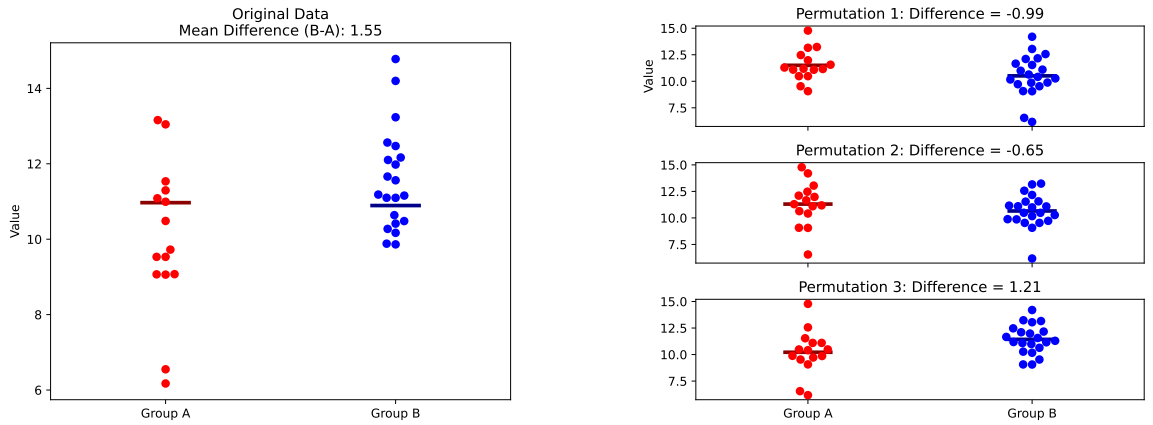
(b) Permutation test histogram

Figure 1: **Permutation test illustration**

This approach provides exact p-values (for discrete test statistics) or arbitrarily precise approximations (as M increases) without distributional assumptions.

2.2 Permutation Testing for Simple Comparisons

Consider the fundamental biological question of whether two groups differ. Under traditional approaches, we might use a t-test, which assumes normally distributed data with equal variances. With permutation testing, we can relax these assumptions:



(a) Data, two gropus mean comparison

(b) Different labels permutations

Figure 2: **Permutation test using beeswarm plots**

1. Calculate the observed difference in means (or another statistic) between groups

2. Randomly reassign observations to groups many times, maintaining the original group sizes
3. For each permutation, recalculate the test statistic
4. Determine how extreme the observed statistic is relative to this permutation distribution

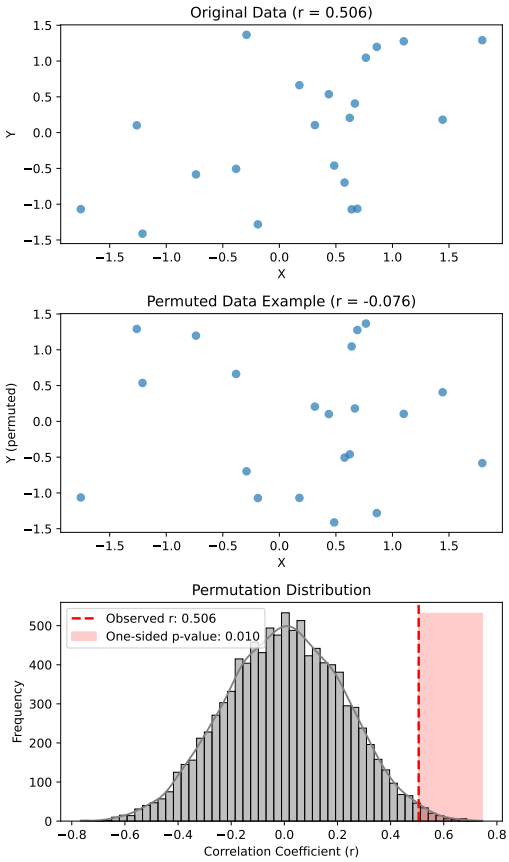
This approach is particularly valuable when:

- Sample sizes are small (where normality assumptions are most critical)
- Data exhibit skewness or outliers that violate normality assumptions
- Variances differ substantially between groups

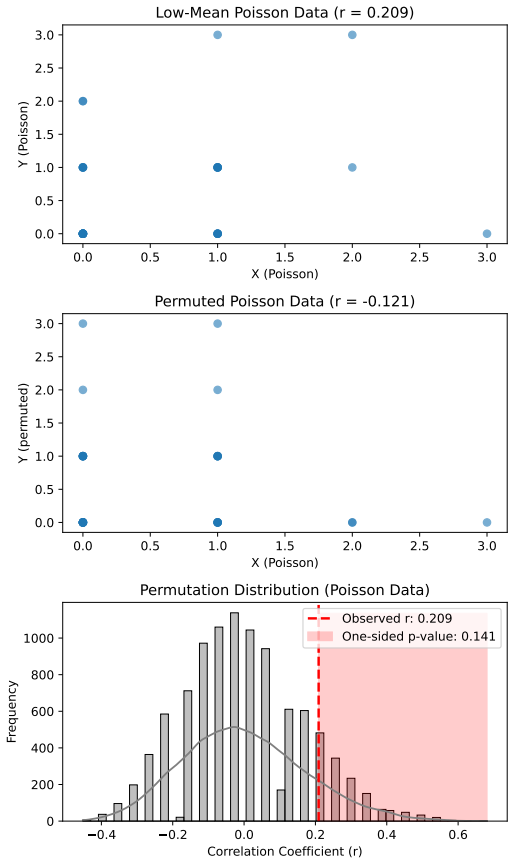
For the simplest cases, permutation tests often give results similar to traditional tests when their assumptions are met. The power of permutation testing emerges most clearly when we move beyond these simple scenarios.

2.3 Permutation Testing for correlation

Permutation tests can also be applied to assess the significance of correlation coefficients. For example, consider a dataset with two continuous variables, X and Y, where we want to test the null hypothesis that there is no correlation between them. The steps are as follows:



(a) Permutation test for correlation



(b) Permutation test for correlation

Figure 3: **Permutation test for correlation**

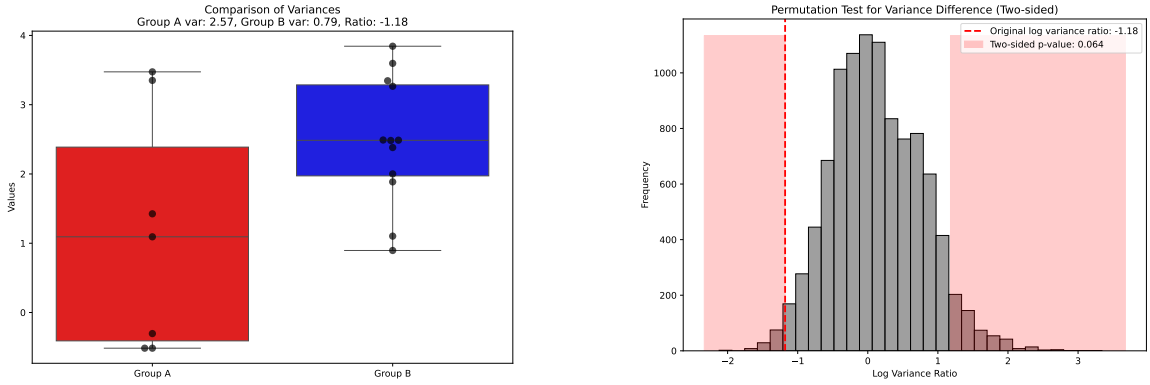
1. Calculate the observed correlation coefficient r_{XY} between X and Y.
2. Randomly permute the values of one variable (e.g., Y) while keeping the other variable (X) fixed.
3. Calculate the correlation coefficient for each permutation, resulting in a distribution of correlation coefficients under the null hypothesis.
4. Determine how extreme the observed correlation coefficient is relative to this permutation distribution.

The example of the figure above shows two datasets/examples. In the first one we have N=15 points with a correlation of 0.5, which turns out to be statistically significant. The second one is a dataset whose both marginal distributions are Poisson with low mean, note how the null distribution is not normal and not centered at zero.

2.4 Complex Null Hypotheses

The true value of permutation testing lies in its flexibility to handle scenarios where traditional approaches struggle. Consider a complex multi-group experiment with different sample sizes, heterogeneous variances, and non-normal distributions. Constructing a valid parametric test would require numerous assumptions and approximations, but the permutation framework remains straightforward.

The permutation approach extends naturally to custom test statistics designed for specific biological questions. For instance, we might be interested in whether the variance (rather than the mean) differs between groups. (this requires centering before permutation) By defining an appropriate test statistic and permuting the data according to the null hypothesis, we can construct valid tests for these questions without deriving analytical sampling distributions.



(a) Permutation test for variance

(b) Permutation test for variance histogram

Figure 4: **Permutation test for variance**

3 The Jackknife Method

3.1 Standard errors, bias and confidence intervals "in general"?

In the past lectures on inference, we got used to the idea that there exist a formula for computing the standard error of a statistic. But let's rethink of the situation where we could derive such formulas, we started from the standard error of the mean:

$$\hat{se}(\bar{X}) = \frac{s}{\sqrt{n}} \quad \text{where } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{1}$$

If you recall, this formula was justified by the Central Limit Theorem. We also saw the standard error on the regression coefficients that was given by the formula:

$$\hat{se}(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} \quad \text{where } \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{2}$$

This had already a more complicated derivation, but it was still possible to derive it under the assumption of normality of the errors.

So how do I estimate the standard error of a statistic that is not a mean or a regression coefficient? This could be a score, a measure of distance or any other statistic that is not a simple linear function of the data. In other words, we want to estimate

$$\hat{se}(t(\mathbf{X})) \quad \text{where } t(\mathbf{X}) \text{ is a statistic of interest} \tag{3}$$

How can this be done with few assumption for a general (and maybe complicated) statistic? For example one that require a complex algorithm to compute? Similar question is relevant for the estimation of the bias of a statistic or the confidence intervals.

A classical statistics did not go super far with this, the typical approach would be to derive the asymptotic distribution of the statistic and then use an analytical method called the "delta method".

3.1.1 The delta method

The delta method is a powerful technique for deriving the standard error of a general statistic. It's based on Taylor series expansions and provides a way to approximate the variance (and thus standard error) of a statistic $t(\mathbf{X})$ when a direct formula isn't available.

When we have a statistic $t(\mathbf{X})$ that can be expressed as a differentiable function of another statistic for which we know the standard error, the delta method allows us to derive its standard error.

For example, if our statistic $t(\mathbf{X})$ can be written as $g(\hat{\theta})$, where $\hat{\theta}$ is a statistic with known standard error $se(\hat{\theta})$, we can approximate the standard error of $t(\mathbf{X})$ using a first-order Taylor series:

$$g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta} - \theta) \tag{4}$$

Taking the variance of both sides:

$$\text{Var}(g(\hat{\theta})) \approx \text{Var}(g(\theta) + \nabla g(\theta)^T(\hat{\theta} - \theta)) \tag{5}$$

Since $g(\theta)$ is a constant, this simplifies to:

$$\text{Var}(g(\hat{\theta})) \approx \text{Var}(\nabla g(\theta)^T(\hat{\theta} - \theta)) = \nabla g(\theta)^T \cdot \text{Var}(\hat{\theta}) \cdot \nabla g(\theta) \tag{6}$$

In practice, we substitute our estimate $\hat{\theta}$ for the unknown θ:

$$\text{Var}(g(\hat{\theta})) \approx \nabla g(\hat{\theta})^T \cdot \text{Var}(\hat{\theta}) \cdot \nabla g(\hat{\theta}) = \nabla g(\hat{\theta})^T \cdot \Sigma \cdot \nabla g(\hat{\theta}) \tag{7}$$

where Σ represents the covariance matrix of $\hat{\theta}$.

Therefore, the standard error becomes:

$$\hat{se}(g(\hat{\theta})) \approx \sqrt{\nabla g(\hat{\theta})^T \cdot \Sigma \cdot \nabla g(\hat{\theta})} \tag{8}$$

in the univariate case it simplifies to

$$\hat{se}(g(\hat{\theta})) \approx |g'(\hat{\theta})| \cdot \hat{se}(\hat{\theta}) \tag{9}$$

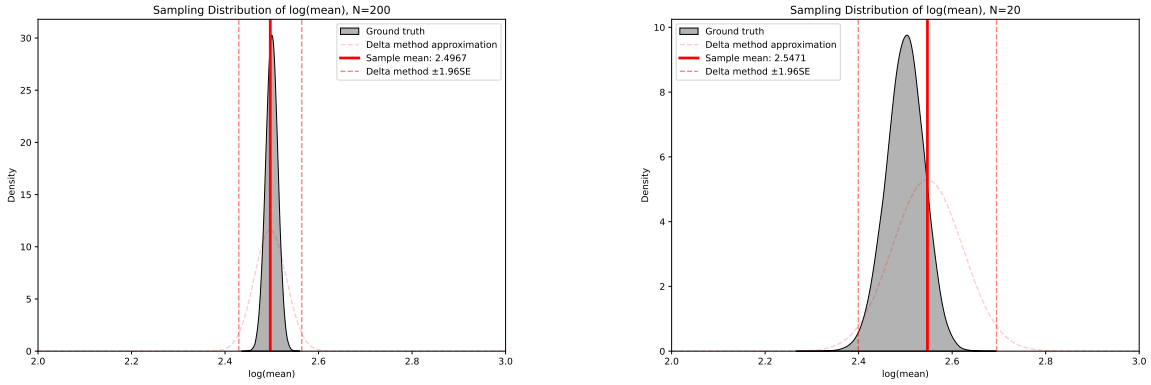
Let's see a typical example of this method applied to the logarithm of the sample mean:

$$t(\mathbf{x}) = \log(\bar{x}) \quad \text{where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{10}$$

The delta method gives us:

$$\hat{se}(\log(\bar{x})) \approx \frac{1}{\bar{x}} \cdot \hat{se}(\bar{x}) = \frac{s}{\bar{x}\sqrt{n}} \tag{11}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is the sample variance.



(a) Sampling distribution $\log(\bar{x})$ (N=200)

(b) Sampling distribution $\log(\bar{x})$ (N=20)

Figure 5: **Delta method example**

This method can be extended to more complex statistics, but it requires that the statistic be a smooth, differentiable function. For statistics that involve non-smooth operations (like max or median) or for complex algorithmic statistics, the delta method may not be applicable.

Before computational methods became widespread, statisticians were limited to statistics where the standard error could be derived analytically using approaches like the delta method. This mathematical complexity often restricted the types of statistics that could be practically used in research.

3.2 Origins of the Jackknife and Basic Concept

The jackknife, introduced by Maurice Quenouille in 1949 and further developed by John Tukey in 1958, represents an early example of computational statistics. Its name evokes the versatile pocket tool - a simple yet effective instrument that can solve many problems.

The core idea is elegantly straightforward: to understand the stability of a statistic, systematically leave out each observation one at a time and recalculate the statistic. This creates a collection of estimates that reveals how sensitive the result is to individual data points.

For example, when applying the jackknife to a simple statistic like the sample mean, we examine how each data point's removal affects the average. With more complex statistics, like a correlation coefficient between gene expression profiles, the jackknife helps assess whether the correlation is driven by a small subset of influential observations or is robust across the dataset.

Definition 3.1 (Jackknife Estimate)**.** Given a dataset $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ and a statistic $\hat{\theta} = t(\mathbf{X})$, the jackknife estimates are:

$$\hat{\theta}_{(i)} = t(\mathbf{X}_{(i)})$$

where $\mathbf{X}_{(i)}$ represents the dataset with the i-th observation removed.

Algorithm 1 Jackknife Method

Input: Data $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, Statistic function $t(\cdot)$
Output: Jackknife estimates, Standard error, Bias-corrected estimate
Compute full-sample estimate $\hat{\theta} = t(\mathbf{X})$
for $i = 1$ to n **do**
    Create leave-one-out sample $\mathbf{X}_{(i)} = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$
    Compute jackknife replicate $\hat{\theta}_{(i)} = t(\mathbf{X}_{(i)})$
end for
Compute average of jackknife estimates: $\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{(i)}$
Compute jackknife standard error: $\hat{se}_{jack} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2}$
Compute jackknife bias estimate: $\hat{bias}_{jack} = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta})$
Compute bias-corrected estimate: $\hat{\theta}_{corrected} = \hat{\theta} - \hat{Bias}_{jack}$

The jackknife approach offers several key advantages:

- It can be applied in an automatic way to any statistic, all its needed is an algorithm to compute the statistic
- It requires minimal assumptions about the underlying distribution
- It provides a systematic way to assess estimate stability
- It can be applied to complex statistics where analytical formulas for standard errors are unavailable

3.3 Estimating Standard Errors

The primary application of the jackknife is estimating the standard error of a statistic without making distributional assumptions.

The jackknife estimate of standard error is:

$$\hat{se}_{jack}(\hat{\theta}) = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2}$$

where $\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{(i)}$ is the average of the jackknife estimates.

This formula has an intuitive interpretation: it measures how much the statistic fluctuates when individual observations are removed, scaled to reflect the appropriate sampling variation.

If you are wondering where fudge factor $\frac{n-1}{n}$, it is selected so that $\hat{se}_{jack}$ is exactly equal to the analytical formula for the standard error of the mean when doing the appropriate substitutions.

3.4 The Jackknife for Bias Estimation

Beyond standard errors, the jackknife can also estimate the bias of a statistic. The bias estimate is:

$$\hat{bias}_{jack}(\hat{\theta}) = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta})$$

This allows us to construct bias-corrected estimates:

$$\hat{\theta}_{corrected} = \hat{\theta} - \hat{bias}_{jack}(\hat{\theta}) = n\hat{\theta} - (n-1)\hat{\theta}_{(\cdot)}$$

Bias correction is particularly valuable for ratio estimators, variance components, and other statistics that exhibit systematic bias in finite samples.

While the jackknife provides valuable insights, it has limitations. It can perform poorly for non-smooth statistics (like medians) and provides limited information about the shape of the sampling distribution beyond the standard error. These limitations motivated the development of more comprehensive resampling approaches.

3.5 Theoretical Foundations

Why does the jackknife method work?

The theoretical justification comes from examining how the jackknife approximates the influence function of a statistic. The intuition is that when we remove a single observation, we are measuring that observation's contribution to the overall estimate.

The jackknife is that it can be seen just as an automatic numerical way to compute the Taylor aproximation as in the delta method. To see this, we can rewrite the jackknife standard error as:

$$\begin{aligned}
\hat{se}_{jack}(\hat{\theta}) &= \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2} \\
&= \sqrt{\frac{n-1}{n^2}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2 \cdot n} \\
&= \sqrt{\frac{1}{n^2}\sum_{i=1}^{n} n(n-1)(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2}
\end{aligned}$$

Now, defining $D_i = \frac{\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}}{1/\sqrt{n(n-1)}}$ as the scaled difference:

$$\begin{aligned}
\hat{se}_{jack}(\hat{\theta}) &= \sqrt{\frac{1}{n^2}\sum_{i=1}^{n} D_i^2} \\
&= \frac{1}{n}\sqrt{\sum_{i=1}^{n} D_i^2}
\end{aligned}$$

D_i are approximate *directional derivatives*, measures of how fast the statistic $t(\mathbf{X})$ is changing as we decrease the weight on data point X_i. So $\text{se}^2_{\text{jack}}$ is proportional to the sum of squared derivatives of $t(\mathbf{X})$ in the n component directions.

This directly connects to the delta method formula we saw earlier:

$$\hat{se}(g(\hat{\theta})) \approx \sqrt{\nabla g(\hat{\theta})^T \cdot \Sigma \cdot \nabla g(\hat{\theta})} \tag{12}$$

The jackknife effectively estimates this same quantity, but does so numerically through resampling rather than analytically.

3.6 Limitations of the Jackknife

The jackknife has some important limitations. In particular, it tends to underestimate variance for non-smooth statistics and can break down completely for certain estimators:

1. **Non-smooth statistics:** For statistics like the median or other quantiles, removing a single observation might not change the estimate at all if it's not near the quantile of interest. This leads to artificially small variance estimates.
2. **Insufficient exploration:** The jackknife creates only n resamples, all with exactly $n-1$ observations. This limited exploration of possible samples might not adequately capture the full sampling distribution.
3. **Edge effects:** For statistics that are particularly sensitive to extreme values, the jackknife can either overestimate variance (if removing an extreme value drastically changes the estimate) or underestimate it (if the statistic is robust to such removal).

This limitation motivated the development of more comprehensive resampling approaches which explore a wider range of possible resamples.

4 The Bootstrap Method

4.1 Fundamental Concept: Resampling with Replacement

While the jackknife and permutation tests represented early advances in computational statistics, the bootstrap, introduced by Bradley Efron in 1979, marked a true paradigm shift. Its core insight is that we can simulate the sampling process by resampling with replacement from the original data, effectively treating our observed sample as a stand-in for the unknown population.

Definition 4.1 (Bootstrap Sample)**.** Given a dataset $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, a bootstrap sample $\mathbf{X}^* = \{X_1^*, X_2^*, \ldots, X_n^*\}$ is created by sampling n observations with replacement from $\mathbf{X}$.

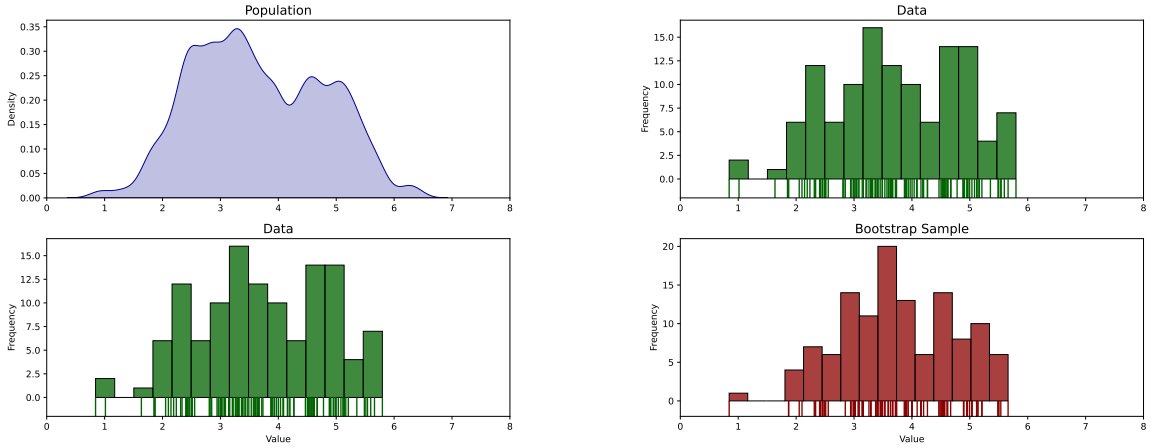
By resampling with replacement, some original observations appear multiple times in a bootstrap sample while others are omitted. This mimics the natural variation we would see if we could draw multiple samples from the original population.

4.2 The Bootstrap Algorithm

The general bootstrap procedure for estimating the sampling distribution of a statistic $\hat{\theta} = t(\mathbf{X})$ follows these steps:

1. Generate B bootstrap samples $\mathbf{X}^{*1}, \mathbf{X}^{*2}, \ldots, \mathbf{X}^{*B}$ by sampling n observations with replacement from the original dataset $\mathbf{X}$.
2. Calculate the statistic of interest for each bootstrap sample: $\hat{\theta}^{*b} = t(\mathbf{X}^{*b})$ for $b = 1, 2, \ldots, B$.
3. Use the empirical distribution of $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}$ to approximate the sampling distribution of $\hat{\theta}$.

The resulting bootstrap distribution provides a computational approximation of the sampling distribution without requiring analytical derivations or distributional assumptions. From this distribution, we can estimate:



(a) Bootstrap sampling illustration

(b) Bootstrap sampling illustration

Figure 6: **Bootstrap sampling illustration**

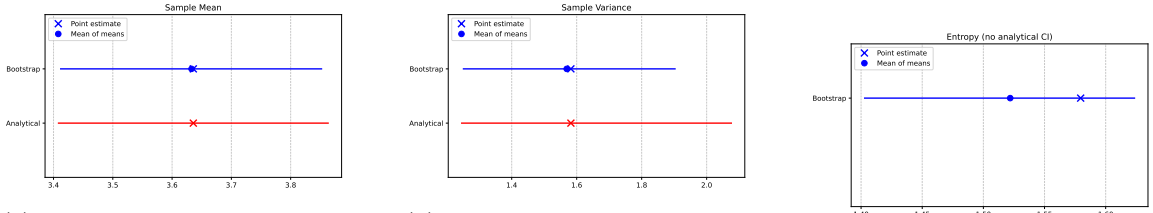
- Standard errors: $\hat{SE}_{boot}(\hat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{*b} - \bar{\theta}^*)^2}$, where $\bar{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b}$
- Confidence intervals: Using percentiles of the bootstrap distribution
- Bias: $\hat{Bias}_{boot}(\hat{\theta}) = \bar{\theta}^* - \hat{\theta}$

4.3 Validation: Bootstrap for Simple Statistics

For simple statistics like means and variances, the bootstrap often produces results very similar to traditional methods when their assumptions are met. This concordance is reassuring and serves as a validation that the bootstrap approach works correctly.

4.4 Bootstrap Confidence Intervals

Beyond standard error estimation, the bootstrap excels at constructing confidence intervals, especially when the sampling distribution is asymmetric or non-normal.



(a) Bootstrap vs. analytical: CI sample mean

(b) Bootstrap vs. analytical: CI sample variance

(c) Bootstrap CI Entropy

Figure 7: **Bootstrap estimates compared to analytical results**

These comparisons show that for simple cases where analytical solutions exist, the bootstrap provides equivalent information. The real power of the bootstrap emerges when we move beyond these simple cases to scenarios where analytical approaches become difficult or impossible.

4.5 Beyond the Basics: Bootstrap for Complex Scenarios

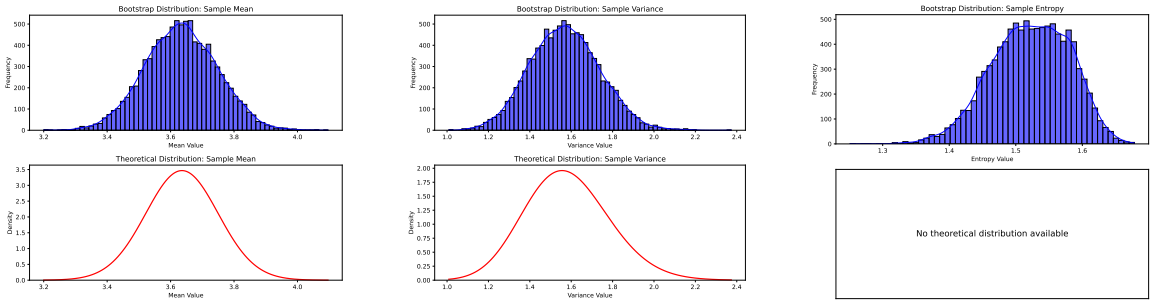
Unlike the jackknife, which provides primarily standard error estimates, the bootstrap gives us **the full empirical distribution of the statistic.**

This enables inference beyond standard errors, including:

- Assessing skewness, kurtosis, and other distributional properties
- Constructing confidence intervals that account for asymmetry
- Testing hypotheses about complex parameters
- Quantifying uncertainty for algorithmic estimators

Let's have a look at some examples.

This versatility makes the bootstrap particularly valuable for complex biological analyses where traditional methods are inadequate.



(a) Bootstrap vs. theoretical (b) Bootstrap vs. theoretical (c) Bootstrap CI Entropy

Figure 8: **Bootstrap estimates compared to theoretical results**

4.5.1 Percentile Method

The simplest approach uses percentiles of the bootstrap distribution directly:

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

where $\hat{\theta}^*_{\alpha/2}$ and $\hat{\theta}^*_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ percentiles of the bootstrap distribution.

This method is intuitive and preserves range restrictions (e.g., correlations between -1 and 1), but can have poor coverage properties when the estimator is biased or its distribution is skewed.

4.6 Parametric Bootstrap: When Sample Size is Small

One key limitation of the standard (nonparametric) bootstrap is its reliance on the sample being representative of the population. When sample sizes are small, resampling from the observed data may not adequately capture population variability. In such cases, we can leverage parametric assumptions through the parametric bootstrap.

Definition 4.2 (Parametric Bootstrap)**.** The parametric bootstrap involves:

1. Fit a parametric model to the observed data and estimate its parameters $\hat{\theta}$
2. Generate synthetic datasets by sampling from the fitted model with parameters $\hat{\theta}$
3. Calculate the statistic of interest for each synthetic dataset
4. Use the distribution of these statistics to approximate the sampling distribution

When the parametric assumptions are reasonable, this approach can provide more reliable inference than the nonparametric bootstrap for small samples. Instead of resampling from a small set of observed values, we sample from a continuous distribution that reflects our understanding of the data-generating process. Parametric families act as regularizers, smoothing out the raw data and de-emphasizing outliers.

4.6.1 Example: Small Sample Gene Expression Analysis

Consider analyzing differential expression with only 3 replicates per condition:

1. Fit a negative binomial model to the observed counts, estimating mean $\hat{\mu}_i$ and dispersion $\hat{\phi}_i$
2. Generate B synthetic datasets by sampling from Negative Binomial$(\hat{\mu}_i, \hat{\phi}_i)$
3. Calculate differential expression statistics for each synthetic dataset
4. Use the resulting distribution for confidence intervals or hypothesis tests

The parametric bootstrap is particularly valuable when:

- Sample sizes are too small for reliable nonparametric resampling
- We have strong theoretical reasons to believe a specific distributional form
- The parameters of interest are part of a well-defined statistical model
- We want to investigate model uncertainty rather than just sampling uncertainty

However, this approach depends critically on the correctness of the assumed model. If the model is misspecified, inference based on the parametric bootstrap may be misleading. When possible, comparing results from parametric and nonparametric approaches can provide insight into model adequacy.