Two samples and complex data structures, bootstrap-based inference

Outline

1. The two-sample problem
2. Bootstrapping regression models
3. Bootstrapping time series
4. Bootstrap confidence intervals
5. Permutation tests
6. Bootstrap hypothesis tests

The two-sample problem

Consider a statistical analysis scenario involving two distinct populations: a *Treatment* group and a *Control* group. Each group is characterized by a cumulative distribution function (cdf) that represents some relevant variable of interest.

Specifically, let F denote the cdf for the Treatment group, and G denote the cdf for the Control group. This setup allows to compare the effects of a treatment on the variable of interest between the two groups.

The observations from the *Treatment* group are denoted as $\mathbf{z} = (z_1, z_2, \ldots, z_m)$, where each z_i represents an individual observation and m is the total number of observations in the *Treatment* group. The empirical cumulative distribution function (ecdf) for these observations is represented by F_m.

Similarly, the *Control* group observations are denoted as $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, each y_i represents an individual observation in the *Control* group and n being the total number of observations. The ecdf for the *Control* group is denoted by G_n. The observations in the *Treatment* group are taken to be independent of those in the *Control* group.

The combined observed data from both groups are represented as $\mathbf{x} = (\mathbf{z}, \mathbf{y})$, comprising a total of $m + n$ observations.

Then, a bootstrap sample $\mathbf{x}^*$ is generated. This involves resampling with replacement m observations from the *Treatment* group, denoted as $\mathbf{z}^*$, and n independent observations from the *Control* group, also resampled with replacement, denoted as $\mathbf{y}^*$. Therefore, the bootstrap sample is represented as $\mathbf{x}^* = (\mathbf{z}^*, \mathbf{y}^*)$.

Example

Can a moth remember what it learned as a caterpillar? (Blackiston et al., 2008, PLoS ONE)

Fifth instar *Manduca sexta* caterpillars received an electrical shock associatively paired with a specific odor to create a conditioned odor aversion.

Air choice / Group	**Treatment**	**Control**
Clear air	32	25
Specific odor	9	21
Total	41	46

We consider the *odds ratio*, where the *odds* (of success) in a group is the fraction of the proportion of success divided by the proportion of failure.

- **Sucess**: choose clear air.
- **Treatment**:

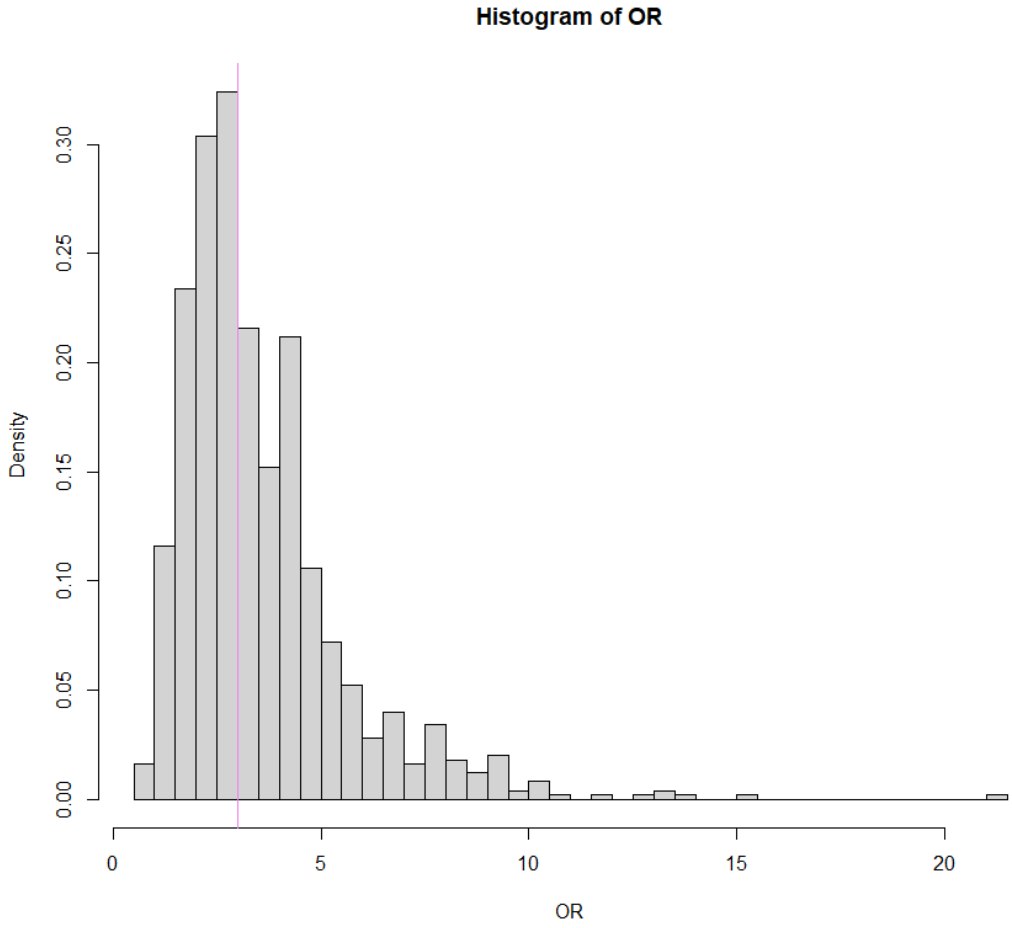
$$\widehat{p}_T = \frac{32}{41} \qquad \widehat{O}_T = \frac{\widehat{p}_T}{1 - \widehat{p}_T} = \tfrac{32}{9} = 3.56$$

- **Control**:

$$\widehat{p}_C = \frac{25}{46} \qquad \widehat{O}_C = \frac{\widehat{p}_C}{1 - \widehat{p}_C} = \tfrac{25}{21} = 1.19$$

- **Odds ratio**:

$$\widehat{OR} = \frac{\widehat{O}_T}{\widehat{O}_C} = \frac{3.56}{1.19} = 2.99$$



Bootstrapping regression models

This section introduces the concept of bootstrapping in the context of regression models.

Consider a data set with n observations, each consisting of $k+1$ tuples, denoted $(\mathbf{x}_i^t, y_i) = (x_{i1}, \ldots, x_{ik}, y_i)$. These observations represent the independent variables (or covariates) $\mathbf{x}_i^t$ and the dependent variable y_i.

The dependent variable, y, is the response or outcome that the regression model attempts to predict, based on the values of the independent variables.

The independent variables, $x_1, \ldots, x_k$, are the k covariates that the model uses to make predictions about y.

The dependent variable, $E[y_i|\mathbf{x}_i]$, as a linear function of the independent variables, given by

$$E[y_i|\mathbf{x}_i] = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the regression coefficients to be estimated.

ε_i represents the random error term associated with the i-th observation, capturing the deviation of the observed values from those predicted by the linear model.

The **simple linear regression** model is a special case with only one independent variable:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

The **multiple linear regression** model generalizes to multiple independent variables:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

Linear regression model assumptions and matrix model

For linear regression models to provide valid inferences, certain assumptions must be satisfied:

- **Homogeneity**: The expected value of the error term is zero, $E[\varepsilon_i] = 0$.
- **Independence**: The error terms are independent of each other, ε_i is independent of ε_j for all $i \neq j$.
- **Homoscedasticity**: The variance of the error terms is constant across observations, $Var[\varepsilon_i] = \sigma^2$, not dependent on i.
- **Normality**: The error terms are normally distributed.

Matrix model To facilitate estimation and hypothesis testing in multiple linear regression, the model can be expressed in matrix form:

Design matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

Regression coefficients vector $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

Response vector $\mathbf{y}$:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Error vector $\boldsymbol{\varepsilon}$:

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Using matrix notation, the regression model can be compactly written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The ordinary least squares (OLS) estimator for the regression coefficients is given by $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$.

Bootstrapping pairs

In the context of regression analysis, bootstrapping can be particularly useful for estimating the distribution of the regression coefficients, especially when the theoretical distribution is complex or unknown.

When we perform a bootstrapping in regression, we often focus on resampling pairs of observations. Each observation consists of a vector of predictor variables $\mathbf{x}_i$ and a corresponding response variable y_i. By resampling pairs $(\mathbf{x}_i^t, y_i)$, we maintain the relationship between the predictor variables and the response variable, which is crucial to accurately estimate the regression model parameters during the bootstrap process.

In this case, bootstrapping over the rows of the dataset implies that each sampled pair, or $k+1$-tuple, is treated as an independent observation. This approach is particularly relevant in the simple linear regression model, where the goal is to understand the linear relationship between a single predictor variable and the response variable. However, this method can also be extended to multiple regression models involving more than one predictor variable.

Algorithm

1. **Sample with replacement:** Randomly select n pairs from the original dataset $\{(\mathbf{x}_1^t, y_1), \ldots, (\mathbf{x}_n^t, y_n)\}$ to form a bootstrap sample. This sample is denoted as $(\mathbf{x}^t, y)^*$, where each element is a pair of predictor variables and the response variable.

2. **Estimate the model parameters:** For each bootstrap sample $(\mathbf{x}^t, y)^*$, fit a regression model to estimate its parameters. This involves finding the best-fitting line (in the case of simple linear regression) or hyperplane (in the case of multiple regression) that describes the relationship between $\mathbf{x}^t$ and y within the bootstrap sample.

3. **Repeat the process:** Steps 1 and 2 are repeated a large number of times (e.g., 1000 or more iterations) to generate a distribution of the estimated model parameters. This distribution can then be used to assess the variability of the parameter estimates and to construct confidence intervals.

Bootstrapping residuals (semiparametric bootstrap)

Bootstrapping residuals, also known as the semiparametric bootstrap, is a resampling technique used primarily in the context of regression models. Unlike traditional bootstrapping methods that resample observations directly, this approach focuses on resampling the residuals (the differences between observed values and those predicted by the model). This method assumes that the model structure is correct, but does not assume a specific distribution for the residuals.

The process involves several key steps, starting with the estimation of regression coefficients and concluding with the generation of new bootstrap samples for further analysis.

Algorithm

1. **Estimate the model parameters:** Begin by fitting the regression model to the original dataset to estimate the coefficients, denoted as $\widehat{\beta}$. This step involves finding the best-fitting line (or hyperplane in multiple dimensions) that minimizes the difference between observed and predicted values.

2. **Compute residuals:** Calculate residuals, which are the differences between observed values (y_i) and the values predicted by the model. The formula for calculating each residual $\widehat{\varepsilon}_i$ is given by:

$$\widehat{\varepsilon}_i = y_i - \left(\widehat{\beta}_0 + \sum_{j=1}^{k} \widehat{\beta}_j x_{ij} \right)$$

   where y_i is the observed value, $\widehat{\beta}_0$ is the intercept, $\widehat{\beta}_j$ are the estimated coefficients, and x_{ij} are the predictor variables.

3. **Generate bootstrap residuals:** Sample with replacement n residuals from the set of computed residuals $\{\widehat{\varepsilon}_1, \widehat{\varepsilon}_2, \ldots, \widehat{\varepsilon}_n\}$ to create a new set of residuals $\{\widehat{\varepsilon}_1^*, \widehat{\varepsilon}_2^*, \ldots, \widehat{\varepsilon}_n^*\}$. These sampled residuals will be used to generate new bootstrap samples.

4. **Create bootstrap samples:** For each of the n observations in the original dataset, add the bootstrap residuals to the predicted values based on the estimated coefficients. This results in a new dataset:

$$\left\{ \left(\mathbf{x}_1^t, \widehat{\beta}_0 + \sum_{j=1}^{k} \widehat{\beta}_j x_{1j} + \widehat{\varepsilon}_1^* \right), \ldots, \left(\mathbf{x}_n^t, \widehat{\beta}_0 + \sum_{j=1}^{k} \widehat{\beta}_j x_{nj} + \widehat{\varepsilon}_n^* \right) \right\}$$

where $\mathbf{x}_i^t$ represents the predictor variables for the i-th observation, and $\widehat{\varepsilon}_i^*$ are the bootstrap residuals.

5. **Estimate parameters from bootstrap samples:** Finally, use each of the generated bootstrap samples to re-estimate the parameters of the regression model. This step is repeated multiple times (each time with a new set of bootstrap residuals) to obtain a distribution of the estimator which can be used to assess its variability and construct confidence intervals.

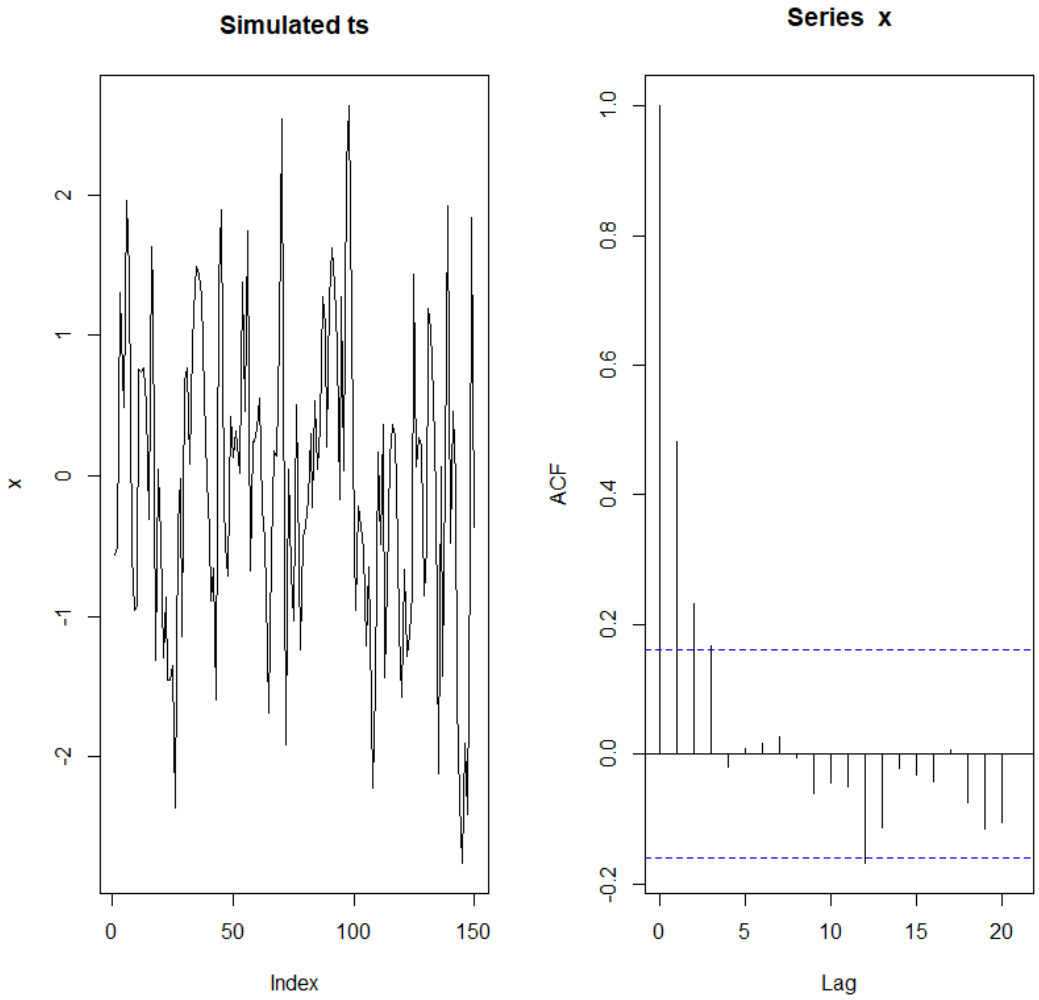
This bootstrapping method is particularly useful when the residual distribution is unknown or when it is difficult to resample observations directly.

Bootstrapping time series

Consider the $AR(1)$ process,

$$z_t = \beta z_{t-1} + \varepsilon_t$$

with $\varepsilon_t \sim N(0, 1)$.



Bootstrapping residuals (semiparametric)

Bootstrapping pairs is not an option in time series, since we would destroy the series structure.

It is possible to bootstrap (resample with replacement) the residuals if we can assume that they are i.i.d.

Algorithm

1. Estimate parameters from the original time series.

2. Calculate the residuals for the given estimates of the parameters.

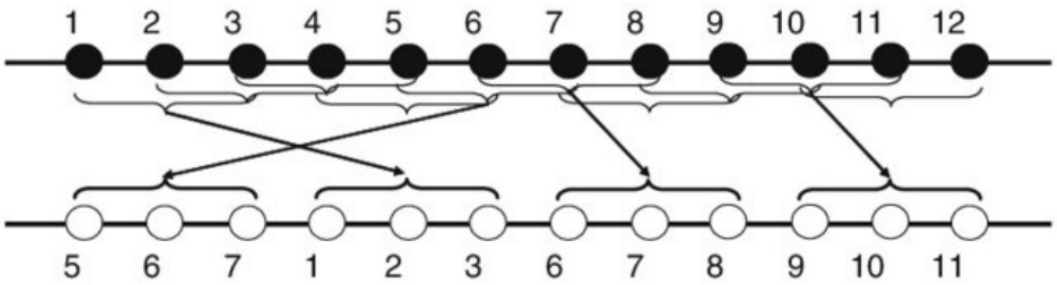
3. Sample with replacement n elements from the residuals and build a bootstrap replicate of the original time series with the estimated parameters and bootstrapped residuals.

4. Estimate the parameters of each bootstrap time series built in Step 3.

Moving blocks bootstrap

Assume that a short block of data already contains a characteristic pattern. If the length of the time series is n, for a fixed block length l, it is possible to sample with replacement n/l time points.

Then, consider a block of length l starting from each of them, and concatenate the blocks to form a bootstrap time series.



Confidence intervals

A confidence interval on a parameter θ with confidence level $1-\alpha$ is an interval computed from the statistics of the observed data in such a way that $(1-\alpha)\cdot 100$ out of 100 confidence intervals would be expected to contain the true θ.

If

$$t = \frac{\widehat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

is the test statistic for $H_0 : \theta = \theta_0$ and $P(t < t_{\alpha/2}) = \alpha/2$ under H_0, then

$$P\left(\widehat{\theta} - t_{1-\alpha/2}\sigma_{\hat{\theta}} < \widehat{\theta} < \hat{\theta} - t_{\alpha/2}\sigma_{\hat{\theta}}\right) = 1-\alpha$$

If the distribution of the test statistic under H_0 is symmetric (about 0), i.e. $t_{\alpha/2} = -t_{1-\alpha/2}$ it results on the well-know CI on θ

$$\widehat{\theta} \mp t_{\alpha/2}\sigma_{\hat{\theta}}$$

Together with the exact confidence interval, which makes use of the exact distribution of a statistic, we have asymptotic confidence intervals which are based on the asymptotic distribution of a statistic.

- **Exact CI** on μ (normal population)

$$\overline{x}_n \mp \frac{s_n}{\sqrt{n}} t_{1-\alpha/2}$$

- **Asymptotic CI** on μ (any population)

$$\overline{x}_n \mp \frac{s_n}{\sqrt{n}} z_{\alpha/2}$$

Example (exponential)

By assuming an exponential distribution, the exact confidence interval for the parameter (or for the mean) can be calculated. See:

en.wikipedia.org/wiki/Exponential_distribution

The $(1-\alpha)\cdot 100\%$ CI of λ from an exponential population is

$$\left[\frac{F^{-1}_{\chi^2_{2n}}(\alpha/2)}{2n\overline{x}_n}, \quad \frac{F^{-1}_{\chi^2_{2n}}(1-\alpha/2)}{2n\overline{x}_n}\right]$$

while the asymptotic CI of a population mean is

$$\overline{x}_n \mp \frac{s_n}{\sqrt{n}} z_{\alpha/2}$$

Basic bootstrap confidence interval

Consider a parameter θ of interest in a population and its estimator $\widehat{\theta}$ computed from a sample. The difference between the estimator and the true parameter is denoted by $\widehat{\delta} = \widehat{\theta} - \theta$. To apply the bootstrap method, we generate many bootstrap samples from the original data, compute the bootstrap estimate $\widehat{\theta}^*$ for each sample, and then calculate the bootstrap version of the statistic, $\widehat{\delta}^* = \widehat{\theta}^* - \widehat{\theta}$.

The essence of the bootstrap method to construct a confidence interval lies in using the empirical distribution of $\widehat{\delta}^*$ to approximate the sampling distribution of $\widehat{\delta}$. This is based on the pivotal quantity approach, where we find the quantiles of the bootstrap distribution that correspond to the desired confidence level $1 - \alpha$.

The basic bootstrap confidence interval then approximates the distribution of $(\widehat{\theta} - \theta)$ by $(\widehat{\theta}^* - \widehat{\theta})$.

We compute the bootstrap statistics $\widehat{\theta}^{*(1)}, \ldots, \widehat{\theta}^{*(B)}$ and the quantiles $\alpha/2$ and $1-\alpha/2$. Call them $F^{-1}_{\widehat{\theta}^*}(\alpha/2)$ and $F^{-1}_{\widehat{\theta}^*}(1-\alpha/2)$, then:

$$\begin{aligned}
1-\alpha &= P\left(F^{-1}_{\widehat{\theta}^*}(\alpha/2) < \widehat{\theta}^* < F^{-1}_{\widehat{\theta}^*}(1-\alpha/2)\right) \\
&= P\left(F^{-1}_{\widehat{\theta}^*}(\alpha/2) - \widehat{\theta} < \widehat{\theta}^* - \widehat{\theta} < F^{-1}_{\widehat{\theta}^*}(1-\alpha/2) - \widehat{\theta}\right) \\
&\approx P\left(F^{-1}_{\widehat{\theta}^*}(\alpha/2) - \widehat{\theta} < \widehat{\theta} - \theta < F^{-1}_{\widehat{\theta}^*}(1-\alpha/2) - \widehat{\theta}\right) \\
&= P\left(F^{-1}_{\widehat{\theta}^*}(\alpha/2) - 2\widehat{\theta} < -\theta < F^{-1}_{\widehat{\theta}^*}(1-\alpha/2) - 2\widehat{\theta}\right) \\
&= P\left(2\widehat{\theta} - F^{-1}_{\widehat{\theta}^*}(1-\alpha/2) < \theta < 2\widehat{\theta} - F^{-1}_{\widehat{\theta}^*}(\alpha/2)\right)
\end{aligned}$$

So the confidence interval

$$\left[2\widehat{\theta} - F^{-1}_{\widehat{\theta}^*}(1-\alpha/2), \quad 2\widehat{\theta} - F^{-1}_{\widehat{\theta}^*}(\alpha/2)\right]$$

is an approximate $(1-\alpha)$ confidence interval for θ.

Bootstrap-t interval

If the distributions $(\widehat{\theta} - \theta)$ and $(\widehat{\theta}^* - \widehat{\theta})$ are not close, then the basic bootstrap confidence interval can be inaccurate.

Recall the traditional formula for a confidence interval (CI) based on a t-distribution:

$$\widehat{\theta} \mp t_{\alpha/2}\sigma_{\hat{\theta}},$$

where $\widehat{\theta}$ is the point estimate of θ, $t_{\alpha/2}$ is the critical value from the t distribution for a given significance level α, and $\sigma_{\widehat{\theta}}$ is the standard error of $\widehat{\theta}$. This formula assumes that the sampling distribution of the estimator is symmetric and follows a t distribution.

In the context of the bootstrap method, we do not directly use the quantiles of the t distribution. Instead, for each bootstrap sample indexed by b, we compute what are known as *approximate pivots*, defined as:

$$t^*(b) = \frac{\widehat{\theta}^*(b) - \widehat{\theta}}{\widehat{\sigma}^*_{\hat{\theta}}(b)},$$

where $\widehat{\theta}^*(b)$ is the bootstrap estimate of θ from the b-th bootstrap sample, and $\widehat{\sigma}^*_{\hat{\theta}}(b)$ is the bootstrap estimate of the standard error of $\widehat{\theta}$ from the b-th sample.

Using the distribution of t^*, which is derived from the bootstrap samples, we can then construct a confidence interval for θ without relying on the assumption of normality or symmetry.

The bootstrap-t confidence interval is given by:

$$\left[\widehat{\theta} - \widehat{\sigma}_{\hat{\theta}} F^{-1}_{t^*}(1-\alpha/2), \quad \widehat{\theta} - \widehat{\sigma}_{\hat{\theta}} F^{-1}_{t^*}(\alpha/2)\right],$$

where $F^{-1}_{t^*}$ is the inverse cumulative distribution function (quantile function) of the bootstrap t-statistics.

In many practical situations, an explicit expression for $\widehat{\sigma}^*_{\hat{\theta}}(b)$, the estimated standard error of the estimator for the b-th bootstrap sample, may not be available. This lack of an explicit formula imposes the use of an *iterated bootstrap* approach. Iterated bootstrap involves performing further bootstrap resampling within each primary bootstrap sample to estimate $\widehat{\sigma}^*_{\hat{\theta}}(b)$. This process is computationally intensive but allows for the approximation of the standard error of the bootstrap estimates in cases where analytical solutions are not feasible.

The iterated bootstrap method, while computationally demanding, enhances the accuracy of the bootstrap confidence intervals by providing a more reliable estimate of the standard error associated with each bootstrap sample. This approach is particularly useful in complex scenarios where the sampling distribution of the estimator is difficult to derive analytically.

The Bootstrap-t interval method offers several advantages over traditional parametric methods for constructing confidence intervals, especially in the following situations:

- **Non-normal Data:** When the underlying distribution of the data is unknown or significantly non-normal, traditional methods that rely on normality assumptions may not be appropriate. The Bootstrap-t method does not require these assumptions, making it more versatile.

- **Small Sample Sizes:** For small sample sizes, the sampling distribution of the estimator may not well approximate a normal distribution, even if the underlying population is normal. The Bootstrap-t method can provide more accurate confidence intervals in these cases.
- **Complex Statistics:** When dealing with complex statistics for which the distribution is difficult to determine, the Bootstrap-t method allows for the estimation of confidence intervals without needing an explicit formula for the standard error.

Nevertheless, it also has limitations:

- **Computationally Intensive:** Especially when using iterated bootstrap for estimating the standard error, the computational burden can be significant. This may limit its use in very large datasets or real-time analysis scenarios.
- **Dependence on Resampling:** The accuracy of the Bootstrap-t interval depends on the quality of the bootstrap samples. If the original sample does not represent the population well, the bootstrap samples may also be misleading.
- **Edge Cases:** In some edge cases, such as when dealing with heavily skewed data or data with outliers, the Bootstrap-t method may still produce biased intervals.

Implementing Bootstrap-t Method

1. **Generate Bootstrap Samples:** From the original dataset of size n, create B bootstrap samples.
2. **Calculate Bootstrap Estimates:** For each bootstrap sample, calculate the estimate of the parameter of interest, $\widehat{\theta}^*(b)$, and the corresponding standard error, $\widehat{\sigma}^*_{\widehat{\theta}}(b)$. If the standard error cannot be directly calculated, use **iterated bootstrap** to estimate it:
    - For each of the B bootstrap samples created in step 1, perform another round of bootstrap sampling. This means creating, for each first-level bootstrap sample, a number B' of second-level bootstrap samples.
    - Calculate the statistic of interest for each second-level bootstrap sample, and then estimate the standard error for each of the B sets of second-level bootstrap samples.
3. **Compute Bootstrap t-Statistics:** For each bootstrap sample, compute the t-statistic using the formula:

$$t^*(b) = \frac{\widehat{\theta}^*(b) - \widehat{\theta}}{\widehat{\sigma}^*_{\widehat{\theta}}(b)},$$

where $\widehat{\theta}$ is the estimate from the original sample.

4. **Determine Quantiles:** Calculate the empirical quantiles of the bootstrap t-statistics, corresponding to $1-\alpha/2$ and $\alpha/2$, to obtain the critical values for the confidence interval.

5. **Construct Confidence Interval:** Use the quantiles from the bootstrap t-statistics to construct the confidence interval for θ as:

$$\left[\widehat{\theta} - \widehat{\sigma}_{\widehat{\theta}} F_{t^*}^{-1}(1-\alpha/2), \quad \widehat{\theta} - \widehat{\sigma}_{\widehat{\theta}} F_{t^*}^{-1}(\alpha/2)\right].$$

When applying the Bootstrap-t method, it should be kept in mind that the choice of B affects the accuracy and computational cost of the method. A larger B provides a more accurate estimation of the confidence interval but at the expense of higher computational time. The common choices for B range from 1000 to 10000.

Percentile-Type Intervals

Another type of interval that can be considered is based on *percentiles*.

Suppose that a sample $\mathbf{x}^*$ is obtained from $\hat{P} \to \mathbf{x}^*$, and then the statistic $\hat{\theta}^* = s(\mathbf{x}^*)$ is computed.

Let $\hat{G}$ denote the cumulative distribution function of $\hat{\theta}^*$.

Then, the percentile interval at level $1-\alpha$ is defined as the percentiles $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ of $\hat{G}$:

$$\left[\hat{\theta}_{\%low};\ \hat{\theta}_{\%up}\right] = \left[\hat{G}^{-1}\left(\frac{\alpha}{2}\right);\ \hat{G}^{-1}\left(1-\frac{\alpha}{2}\right)\right]$$

By definition, since $\hat{G}^{-1}(\alpha) = \hat{\theta}^*_{(\alpha)}$, the interval can also be rewritten as:

$$\left[\hat{\theta}_{\%low};\ \hat{\theta}_{\%up}\right] = \left[\hat{\theta}^*_{(\frac{\alpha}{2})};\ \hat{\theta}^*_{(1-\frac{\alpha}{2})}\right]$$

The above expression refers to an ideal case where there are infinitely many bootstrap replications. In practice, a finite number B of such replications is considered.

A total of B independent bootstrap datasets $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$ are generated, and the corresponding estimators are computed as:

$$\hat{\theta}^*(b) = s(\mathbf{x}^{*b}),$$

for $b = 1, \ldots, B$.

The notation $\hat{\theta}^*_{B(\alpha)}$ is used to denote the α-th percentile of the values $\hat{\theta}^*(b)$, which corresponds to the $(B \times \alpha)$-th value in the ordered list of the B bootstrap replications of $\hat{\theta}^*$.

For example, if $B = 2000$ and $\alpha = 0.05$, then $\hat{\theta}^*_{B(\alpha)}$ is the 100th value in the ordered list of replications.

Similarly, the same reasoning applies to $\hat{\theta}^*_{B(1-\alpha)}$, which represents the $(1-\alpha)$-th percentile of the values.

Thus, the percentile interval at level $1-\alpha$ is approximated by:

$$\left[\hat{\theta}_{\%low};\ \hat{\theta}_{\%up}\right] \approx \left[\hat{\theta}^*_{B(\frac{\alpha}{2})};\ \hat{\theta}^*_{B(1-\frac{\alpha}{2})}\right]$$

If the distribution of $\hat{\theta}^*$ is approximately normal, then the percentile interval and the standard normal interval coincide.

Bias corrected accelerated (BCa) bootstrap interval

The Bias-Corrected and Accelerated (BCa) bootstrap method provides a way to construct confidence intervals (CIs) that correct for both bias and skewness in the bootstrap distribution.

This method is particularly useful when the distribution of the estimator is not symmetric or when standard bootstrap methods give biased estimates.

Given a statistic or parameter estimate $\widehat{\theta}$, we consider its bootstrap distribution $F_{\hat{\theta}^*}$, where $\widehat{\theta}^*$ represents the bootstrap estimates of $\widehat{\theta}$.

Assumptions and Definitions

Assume there exists a function g that is monotone increasing and transforms $\widehat{\theta}$ such that:

$$g(\widehat{\theta}) \sim N\left(g(\theta) - z_0,\ 1 + ag(\theta)\right)$$

where z_0 is the bias correction term and a is the acceleration factor. These components are crucial for adjusting the bootstrap distribution to account for bias and skewness.

The BCa confidence interval is given by:

$$\left[F^{-1}_{\hat{\theta}^*}(\alpha_1), \quad F^{-1}_{\hat{\theta}^*}(\alpha_2)\right]$$

The adjustment factors α_1 and α_2 for the lower and upper bounds of the confidence interval, respectively, are determined by:

$$\alpha_1 = \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + \Phi^{-1}(\alpha/2)}{1 - \widehat{a}\left(\widehat{z}_0 + \Phi^{-1}(\alpha/2)\right)}\right)$$

$$\alpha_2 = \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + \Phi^{-1}(1-\alpha/2)}{1 - \widehat{a}\left(\widehat{z}_0 + \Phi^{-1}(1-\alpha/2)\right)}\right),$$

where Φ represents the cumulative distribution function (cdf) of a standard normal distribution, and Φ^{-1} is its quantile function (inverse cdf).

The bias correction factor $\widehat{z}_0$ and the acceleration factor $\widehat{a}$ are calculated as follows:

$$\widehat{z}_0 = \Phi^{-1}\left(\frac{\#\{\widehat{\theta}^*(b) - \widehat{\theta}\}}{B}\right)$$

$$\widehat{a} = \frac{\sum_{i=1}^{n}\left(\widehat{\theta}_{(\cdot)} - \widehat{\theta}_{(i)}\right)^3}{6\left(\sum_{i=1}^{n}\left(\widehat{\theta}_{(\cdot)} - \widehat{\theta}_{(i)}\right)^2\right)^{3/2}}$$

where $\widehat{\theta}_{(i)}$ is the estimate of θ holding out the i-th observation, $\widehat{\theta}_{(\cdot)}$ is the average of the jackknife estimates, and B is the number of bootstrap samples.

The BCa method for constructing confidence intervals provides a more accurate representation of the underlying parameter's uncertainty by correcting for both bias and skewness in the bootstrap distribution. This method relies on the calculation of the bias correction and acceleration factors, which adjust the confidence interval bounds to more accurately reflect the distribution of the estimator.

Properties of bootstrap CIs

A key property of some CI construction methods is being **transformation respecting**. This means that if you have a parameter θ and you apply a monotone function g to it, creating $g(\theta)$, then the CI for $g(\theta)$ can be directly derived by applying g to the endpoints of the CI for θ. This property is essential for ensuring that the CIs are consistent across transformations, maintaining their interpretability and utility regardless of the scale or transformation applied to the parameter.

The table below summarizes whether various bootstrap CI methods respect this property and their computational speed, which is often inversely related to the number of bootstrap samples n required for accurate estimation.

Method/Properties	**Transformation Respecting**	**Speed**
BC_a	Yes	$1/n$
Bootstrap t	No	$1/n$
Basic	No	$1/\sqrt{n}$
Percentile	Yes	$1/\sqrt{n}$

Notes:

- **BC_a method:** This method adjusts for bias and skewness in the bootstrap distribution and respects transformations. It is computationally intensive, with its speed inversely proportional

to the sample size n.

- **Bootstrap t method:** Unlike BC_a, it does not respect transformations, making it less versatile for analyses involving transformed parameters.

- **Basic method:** This method is faster than BC_a and bootstrap t in terms of computational speed ($1/\sqrt{n}$), but it does not respect transformations.

- **Percentile method:** It respects transformations and has the same computational speed as the basic method. This makes it a valuable method for cases where transformations are applied.

Permutation tests

Permutation tests are a type of nonparametric statistical significance test that can be used to compare two or more samples. They are particularly useful when the assumption of normality in the underlying populations cannot be satisfied, which is a common requirement for traditional parametric tests such as the t-test. The fundamental principle behind permutation tests is to evaluate the probability of observing the effect measured in the samples if the null hypothesis were true. This is achieved by calculating all possible values of the test statistic under rearrangements of the observed data points.

Procedure

- **Samples:** Consider two samples drawn from two populations. These populations may be the same or different. Let's denote the first sample as $z = (z_1, \ldots, z_n)$ with population cumulative distribution function (cdf) F, and the second sample as $y = (y_1, \ldots, y_m)$ with population cdf G.

- **Null Hypothesis:** The null hypothesis for a permutation test is typically formulated as $H_0 : F = G$. This implies that there is no difference between the distribution functions of the two populations from which the samples were drawn.

- **Merging Samples:** To perform the permutation test, the two samples, $\mathbf{z}$ and $\mathbf{y}$, are merged together into a single combined sample. This combined sample is then used to generate multiple independent subsamples, $\mathbf{z}'$ and $\mathbf{y}'$, of respective sizes m and n, taken **without** replacement. This process simulates the idea of randomly allocating the observations to one of the two groups, under the assumption that the null hypothesis is true.

- **Test Statistic:** A test statistic is calculated for each permutation of the data. Common choices for the test statistic include the difference in means or medians between the two groups. The idea is to measure how much the groups differ under each permutation.

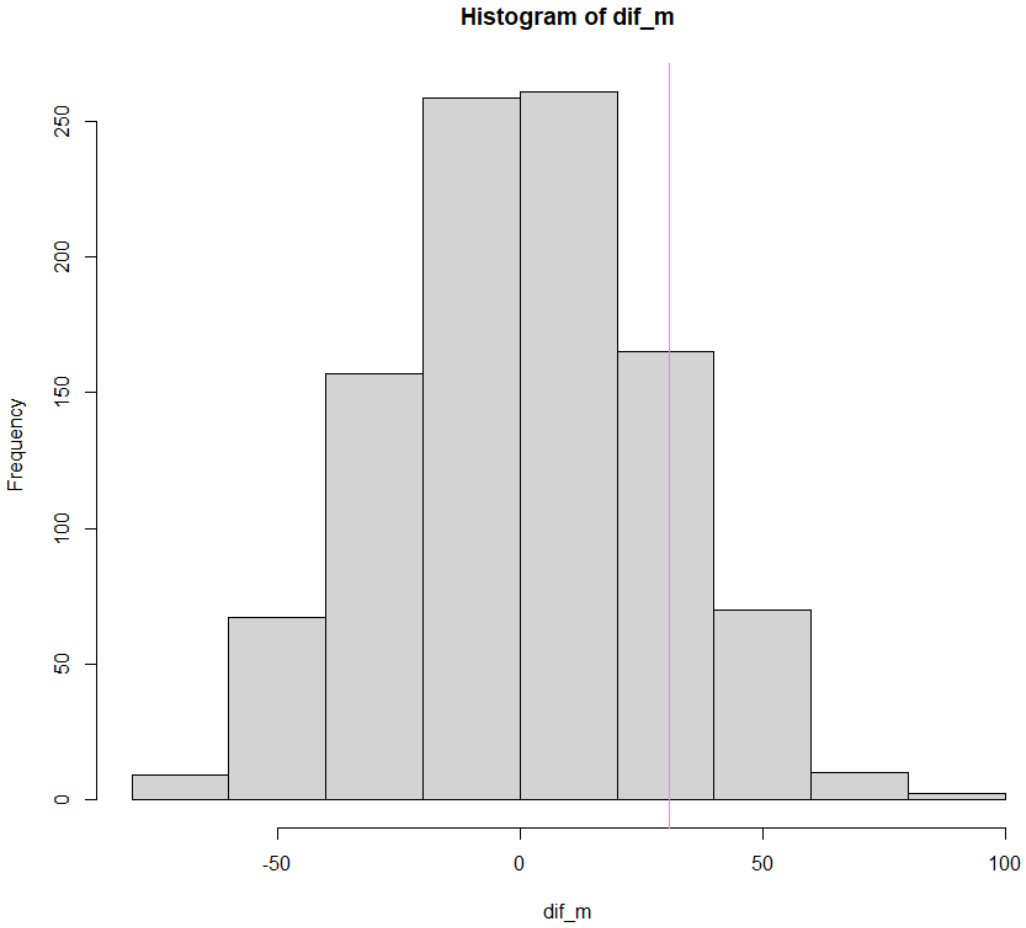
- **Approximate Significance Level (ASL):** The approximate significance level (ASL) is then calculated as the fraction of the B subsamples that are more consistent with the alternative hypothesis than the original grouping of $\mathbf{z}$ and $\mathbf{y}$. The ASL represents the p-value of the test, providing a measure of how extreme the observed statistic is under the null hypothesis.

The ASL is estimated as the proportion of shuffled datasets that produce a test statistic as extreme or more extreme than the actual observed test statistic. This provides an estimate of the probability of observing a test result as extreme as the actual result if the null hypothesis were true.

Comparison with the t-test

To illustrate the application of permutation tests, consider comparing two means under the null hypothesis $H_0 : \mu_t = \mu_c$ versus the alternative hypothesis $H_1 : \mu_t > \mu_c$. In this context, a permutation test can be applied by calculating the difference in means between the two samples for each possible permutation of the data. This approach is contrasted with the traditional t-test, which assumes that the samples are drawn from normally distributed populations with equal variances.

The advantage of the permutation test over the t-test lies in its flexibility and fewer assumptions about the distribution of the data. While the t-test relies on theoretical distribution properties, permutation tests use the actual data to generate the distribution of the test statistic under the null hypothesis. This makes permutation tests more robust to deviations from normality and applicable to a wider range of data types and sample sizes.



One-sample randomization test

The one-sample randomization test is a non-parametric statistical method used to test hypotheses about the central tendency (e.g., median) of a single sample without assuming a specific distribution for the data. This test is particularly useful when the normality assumption is questionable.

- The one-sample randomization test allows us to make inferences about the central location of the data, which can include the median, quantiles, or any other location parameter, under some additional assumptions. This flexibility makes it a valuable tool in exploratory data analysis and hypothesis testing when the underlying distribution of the data is unknown or non-normal.

- The hypotheses for this test can be set up as follows:

  Test the null hypothesis H_0 : median $= m_0$ against the alternative hypothesis H_1 : median $\neq m_0$. This format implies a two-sided test. However, the method can also be adapted for one-sided tests without significant complication.

For a one-sided test, the hypotheses would be adjusted to either H_1 : median $> m_0$ or H_1 : median $< m_0$, depending on the research question.

- Consider an example where we want to test whether the median of a dataset related to air-conditioning, perhaps measuring some performance or satisfaction metric, is equal to $m_0 = 6$. This value, m_0, is the hypothesized median that we aim to test against the observed data.

- In our specific dataset, there are two observations below 6 out of a total of $n = 12$ observations. This information is crucial as it helps us to calculate the test statistic under the null hypothesis.

- Under the null hypothesis H_0, which assumes the median is equal to 6, the distribution of the test statistic X, defined as the number of observations below 6, can be modeled by a binomial distribution. Specifically, X follows a binomial distribution with parameters $n = 12$ and $p = 0.5$, denoted as $X \sim Bin(n = 12, p = 0.5)$. This is because, under the null hypothesis, each observation has a 50% chance of being below or above the median due to the symmetry around the median.

- The probability of observing two or fewer observations below the hypothesized median (or equivalently, ten or more observations above the hypothesized median) under the null hypothesis is calculated as $P(X \leq 2) = 0.019287$. This probability serves as the p-value for our test. The p-value measures the strength of evidence against the null hypothesis. In this context, it quantifies the likelihood of observing a sample as extreme as, or more extreme than, our actual sample if the null hypothesis were true.

- Therefore, the two-tailed p-value for this test is 0.03857, obtained by doubling the one-tailed probability ($2 \times P(X \leq 2)$) because we are considering both tails of the distribution for the two-sided alternative hypothesis. This step is necessary in a two-sided test to account for the possibility of the observed median being either significantly higher or significantly lower than the hypothesized median.

In conclusion, the one-sample randomization test provides a robust method for testing hypotheses about the median of a dataset without relying on normality assumptions. By leveraging the binomial distribution under the null hypothesis, this test offers a straightforward way to assess the likelihood of observing our sample data if the null hypothesis were true, thereby facilitating evidence-based decision making in statistical analysis.

Bootstrap hypothesis tests

In hypothesis testing, we often compare a null hypothesis H_0 against an alternative hypothesis H_a. A common framework is to test $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$, where θ represents a population parameter, and θ_0 is a specific value of θ under the null hypothesis.

The decision to reject H_0 is based on whether θ_0 falls within a $(1-\alpha)100\%$ confidence interval on θ. If θ_0 does not belong to this confidence interval, H_0 is rejected. The significance level α represents the probability of rejecting H_0 when it is true (Type I error).

Key Terms:

- *Confidence Interval (CI)*: A range of values within which we expect θ to fall, with a certain degree of confidence (e.g., 95% confidence).
- *Significance Level (α)*: The probability of rejecting the null hypothesis when it is actually true.
- *ASL (Achieved Significance Level)*: The observed significance level of the test, representing the probability of obtaining a test statistic at least as extreme as the one observed, assuming the null hypothesis is true. It must be computed specifically if required.

The Two-Sample Bootstrap Test Statistic

When comparing two samples, we may be interested in testing whether they come from the same distribution. This scenario is addressed by the two-sample bootstrap test.

Procedure:

1. Consider two samples: $z = (z_1, \ldots, z_n)$ from the population with cumulative distribution function (cdf) F, and $y = (y_1, \ldots, y_m)$ from population with cdf G.
2. The null hypothesis is $H_0 : F = G$, suggesting that there is no difference between the two population distributions.
3. Generate B independent bootstrap samples $\mathbf{z}'$ (size n) and $\mathbf{y}'$ (size m) from the combined samples $(\mathbf{z}, \mathbf{y})$ with replacement. This step simulates drawing samples from the population under the assumption that $F = G$.
4. Calculate a test statistic (e.g. difference in means, medians) for each pair of bootstrap samples $(\mathbf{z}', \mathbf{y}')$.

5. Approximate the Achieved Significance Level (ASL) as the fraction of bootstrap statistics that exceed the observed test statistic calculated from the original samples $(\mathbf{z}, \mathbf{y})$. This fraction provides an estimate of the p value for the test.

This methodology allows for the estimation of the p-value without making strict assumptions about the distribution of the test statistic under the null hypothesis, thereby providing a flexible approach to hypothesis testing.

Bootstrap t-tests

Bootstrap t-tests are a resampling method used to assess the significance of the difference between two sample means when the underlying population distribution is unknown or when the sample size is small.

- The **null hypothesis** for the bootstrap t test is denoted as $H_0 : F = G$, where F and G represent the cumulative distribution functions of the two populations from which the samples $\mathbf{z}$ and $\mathbf{y}$ are drawn, respectively. The null hypothesis states that there is no difference between the two population means.

- To perform the test, we first take B independent bootstrap samples $\mathbf{z}'$ (size n) and $\mathbf{y}'$ (size m) from the combined samples $(\mathbf{z}, \mathbf{y})$ *with replacement*. This process involves randomly selecting observations from the original samples to create new samples of the same sizes.

- For each of bootstrap sample pairs, we calculate the bootstrap statistic for the difference in means (or any other parameter of interest) and its standard error. If the standard error cannot be derived analytically, it can be estimated through the bootstrap method itself by taking the standard deviation of the bootstrap statistics across all replications.

- The *approximate significance level (ASL)* is then estimated. This is done by comparing the observed test statistic (the difference in means divided by its standard error in the original sample) to the distribution of the bootstrap test statistics (differences in means divided by their standard errors across all bootstrap samples).

  The ASL is the fraction of times the bootstrap test statistics exceed (or are less than, depending on the hypothesis) the observed test statistic. This fraction gives an estimate of the p value, which is used to decide whether to reject the null hypothesis.