# A STUDY ON RESAMPLING TECHNIQUES:BOOTSTRAP AND JACKKNIEF

# INTRODUCTION

In Statistical analysis we often concerned with making inferences on distributions and accuracy of statistics. when we collecting samples across large groups of people, objects or data, there are several ways to verify accuracy. One of the method that commonly used is resampling, where we take additional samples and observations to identify any bias or issues. Resampling is a series of techniques used in statistics to gather more information about a sample. This can include retaking a sample or estimating its accuracy. Resampling improves the overall accuracy and estimates any uncertainty within a population.
Resampling methods resamples from the obtained data and infer from the resampled data to estimate the underlying distributions and make estimations about accuracy of the statistic . The method of resampling does not involve the utilization of the generic distribution tables in order to compute the p value. The most widely used resampling methods are jackknife approach and bootstrap approach. These methods highly uses the technologies ,so it become easier to apply.

Resampling is a statistical approach that relies on empirical analysis, based on the observed data, instead of asymptotic and parametric theory .The goal of Parametric and re-sampling methods use different approaches are same; the defining difference is the type of sampling distribution used in relation to the relevant test statistic. A parametric method employs a theoretical sampling distribution to model sampling error probability. These distributions, such as the t or $\chi 2$, are mathematically derived, and are based on a set of assumptions. But, a re-sampling method employs an empirical sampling distribution to model sampling error probability. These distributions are created by the researcher from the particular unique set of observed data.

In statistics, we use samples to infer about the population given some set of conditions. The sample is not the complete representation of population. so we use the sample to make guesses. Because the sample is not the entire population. So there occurs standard deviations, standard errors, and confidence intervals with all statistics. If we took another sample, it would be slightly different from the original sample and would have slightly different statistics. Since standard errors of the statistics are calculated based on the sample, these estimates can be biased to the sample and have certain mathematical assumptions about the distribution.
Here in resampling we take randomly drawn (sub) samples of the sample and calculate the statistic from that (sub) sample. Do this enough times and you can get a distribution of statistic values that can provide an empirical measure of the accuracy of the test statistic, with less rigid assumptions. The most famous

resampling methods are randomization, Monte Carlo, bootstrap, jackknife, permutation testing, Cross-validation approaches.

In recent years many emerging statistical analytical tools, such as exploratory data analysis (EDA), data visualization, robust procedures, and resampling methods, have been gaining attention among psychological and educational researchers. The classical parametric tests compare observed statistics to theoretical sampling distributions, While the resampling is a revolutionary methodology because it departs from theoretical distributions. Rather, the inference is based upon repeated sampling within the same sample, and that is why this school is called resampling. The resampling methods; cross validation ,jackknife and bootstrap are very similar. But bootstrap is more flexible.

Here we discussing the four methods that are specified above. Among the four resampling methods the most widely used methods are bootstrap and jackknife approaches .Both are examples of nonparametric tests. The bootstrap and jackknife methods, are powerful statistical tools used to estimate the accuracy of sample statistics by sampling from the data repeatedly. These techniques help to assess the variability and reliability of estimates, particularly when the underlying distribution of the data is unknown or the sample size is small.

In 1949, Quenouille introduced jackknife method . Jackknife is used in statistical inference to estimate the bias and standard error of a test statistic by recomputing the estimate from subsamples of the available sample, leaving out one observation or a group of observations at a time from the sample.
The resampling method bootstrap was introduced by Efron. .Bootstrap is a statistical method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample, most often with the purpose of deriving robust estimates of standard errors and confidence intervals of a population parameter like a mean, median, and correlation coefficient.

In this project, we will explore these resampling techniques in detail, focusing on their applications and properties. Specifically, we will analyze the bias, standard error, variance, and confidence intervals of sample estimates using the bootstrap and jackknife methods. Additionally, we will apply these methods to linear regression analysis on the Gapminder dataset, which contains various socio-economic indicators for different countries .

OBJECTIVES

- Explore Various Resampling Methods: To study and understand different resampling techniques.
- Understand Bootstrap and Jackknife Techniques: To gain a comprehensive understanding of the bootstrap and jackknife methods, including their theoretical foundations and practical applications.
- Evaluate Statistical Properties: To evaluate the bias, standard error, variance, and confidence intervals of sample estimates using the bootstrap and jackknife methods.
- Apply to Linear Regression Models: To apply bootstrap and jackknife resampling techniques to linear regression models, analyzing their impact on regression coefficients and model performance.
- Analyze the 'Gapminder' Dataset: To conduct a detailed analysis of the Gapminder dataset using resampling techniques, highlighting the advantages and limitations of each method in the context of real-world data.
- Discuss Applications in Various Fields: To explore and discuss the application of resampling techniques in various fields, demonstrating their versatility and practical relevance.

## SCOPE OF THE STUDY

Resampling is a series of techniques used in statistics to gather information about a sample. The significant advantage of resampling is that you can repeatedly draw samples from the same population until your model achieves optimal performance. This approach saves time and resources by allowing the recycling of the same dataset, eliminating the need for new data collection. By resampling, we can improve overall accuracy and estimate uncertainty within a population.

Here this study will delve into the theoretical and practical aspects of bootstrap and jackknife resampling techniques, focusing on their mathematical formulations and underlying assumptions. By reviewing the literature and foundational concepts, we aim to build a robust theoretical understanding and compare the efficiency, accuracy, and computational complexity of both methods.

Empirically, we will apply both techniques to a dataset, analyzing bias, standard error, variance, and confidence intervals for various statistical estimates. This includes constructing confidence intervals and assessing reliability. We will also apply these methods to linear regression analysis to examine their impact on regression coefficients and model performance.

We will discuss the application of these techniques in fields such as finance, medicine, and social sciences, demonstrating their versatility and practical relevance. By identifying the most accurate resampling methods, this study aims to improve statistical analysis accuracy and reliability across various real-world applications.

# REVIEW OF LITERATURE

The development and application of resampling techniques, particularly the bootstrap and jackknife methods, have been extensively studied and documented in various research works. These studies have contributed significantly to the understanding and advancement of statistical estimation and inference.

In 1979, Bradley Efron introduced the bootstrap method in his seminal paper "Bootstrap Methods: Another Look at the Jackknife." Efron demonstrated that the bootstrap method works satisfactorily across various estimation problems and showed that the jackknife can be considered a linear approximation method for the bootstrap. Efron's exposition included examples such as the variance of the sample median, error rates in linear discriminant analysis, ratio estimation, and regression parameters, laying the groundwork for future research in resampling techniques.

Stuart Maxwell Angus's 1988 research paper, “A Comparison Of The Jackknife And Bootstrap Estimators In Linear Models With Reference To Production Models Used By Sasol,” conducted a detailed comparative analysis of the jackknife and bootstrap methods, focusing on their application in production models. This study provided insights into the practical implementation of these techniques and their relative performance.

Herwig Friedl and Erwin Stampfer's 2002 paper “Jackknife Resampling” offered a comprehensive explanation of the jackknife resampling method and its theoretical basis, contributing to a better understanding of its application in various statistical contexts.

Chong Ho Yu's 2002 article "Resampling Methods: Concepts, Applications, and Justification," provided a broad overview of resampling methods, including simple examples and software applications available for implementing these techniques. This work

highlighted the accessibility of resampling methods through standard statistical software and discussed the arguments for and against their use.

Anatoli Iambartsev's study on “Resampling Methods” focused mainly on the bootstrap and jackknife techniques, discussing their properties such as bias reduction, standard error, and hypothesis testing. This study emphasized the bootstrap method while providing valuable insights into the practical aspects of these resampling techniques.

James MacKinnon's 2007 work "Bootstrap Hypothesis Testing" discussed the basic ideas of bootstrap testing, its relationship with Monte Carlo testing, and factors affecting the performance of bootstrap tests under the null hypothesis. This work advanced the understanding of bootstrap methods in the context of hypothesis testing.

The 2007 article “Bootstrap and Jackknife Resampling Algorithms for Estimation of Regression Parameters” explored the application of these methods in regression analysis, providing practical examples and demonstrating their utility in improving regression parameter estimates.

The book "The Sage Handbook of Quantitative Methods in Psychology" by Beasley and Rodgers (2009) emphasized the flexibility and frequent use of the bootstrap method while also covering permutation tests and the jackknife. This text offered a detailed explanation of the concepts and mechanisms underlying resampling theory, making it a valuable resource for understanding these techniques.

Dale Berger's "A Gentle Introduction to Resampling Techniques" provided an accessible introduction to various resampling techniques, covering fundamental concepts, applications, and advantages. This work served as an entry point for those new to resampling methods, facilitating a deeper understanding.

Peter Young's work on "Jackknife and Bootstrap Resampling Methods in Statistical Analysis to Correct for Bias" at the University of California, Santa Cruz, focused on how these techniques address bias in statistical analysis, offering practical insights into their effectiveness.

Notes by G. Jogesh Babu from the Center for Astrostatistics provided a comprehensive overview of jackknife and bootstrap methods in astrostatistics, highlighting their application in analyzing astrophysical data.

The Spring 2014 publication "The Bootstrap Advanced Methods for Data Analysis" detailed the bootstrap resampling technique, particularly in constructing confidence intervals. This was further elaborated in the article "Bootstrap Confidence Intervals" by Thomas J. DiCiccio and Bradley Efron, which surveyed bootstrap methods aimed at producing accurate confidence intervals for complex problems.

The summer institute publication "The Bootstrap and Jackknife" provided educational resources on the properties and applications of these resampling techniques, contributing to the broader dissemination of knowledge in this field.

This comprehensive review of literature underscores the significant advancements in the study of bootstrap and jackknife resampling techniques. From foundational theoretical explorations to practical applications in various fields, these studies have collectively enhanced the understanding of resampling methods and their utility in improving statistical analysis.

# CHAPTER 1
# RESAMPLING METHODS

## 1.1 PERMUTATION AND RANDOMIZATION TESTS

The Permutation test, also known as the randomization technique, was developed by Ronald A. Fisher in the 1930s. This method involves performing the exact test multiple times under a null hypothesis, commonly used to assess whether the effect of treatment is zero. Instead of relying on a predefined form for the null distribution, analysts adopt a data-driven approach by using observed data to construct it.

In a permutation test, the process begins by iteratively shuffling the sample numerous times, creating new samples that systematically disrupt the relationships within the original sample. The statistic of interest is then calculated for each of these reshuffled samples. The final step involves comparing the estimate obtained from the original sample with the distribution of estimates derived from the reshuffled samples, providing insight into how distinct the observed estimate is compared to random shuffling.

If every possible reshuffling combination is calculated, this method is specifically termed a permutation test. Alternatively, if a large number of reshuffles are conducted, it is referred to as a randomization test. In both cases, the primary objective is to evaluate the degree of deviation of the observed estimate from what would be anticipated under the assumption of no underlying relationship in the data.

Permutation tests play a crucial role in automatically creating a sampling distribution within a population and conducting similar observations. This testing method helps establish the exchangeability of

different observations, assessing the likelihood of exchanging labels within a set.
How it work:

The Permutation tests, involve randomly redistributing observed scores into two groups based on the observed sample sizes ($N_1$ and $N_2$). A statistic of interest, such as the difference in means or medians, is then calculated for each random assignment. This process is repeated numerous times (e.g., 1000 or 10000) to create an empirical sampling distribution under the null hypothesis of no difference between the two populations.

For instance, if there are nine observed scores, they might be randomly divided into groups of four and five. By repeating this randomization process, a distribution of observed values for the chosen statistic is generated. This empirical sampling distribution is compared to the observed statistic to assess its likelihood under the assumption that the two population distributions are identical. If the observed statistic falls into an extreme portion of the distribution, it suggests that the probability of such an outcome is low.

The decision to reject the null hypothesis is based on comparing the observed statistic to the empirical sampling distribution. For example, if 32 out of 1000 samples are as extreme or more extreme than the observed sample, the conclusion might be that the probability of such an extreme outcome is approximately 0.032 (one-tailed), leading to the rejection of the hypothesis that the two populations are the same.

Randomization allows the generation of the sampling distribution for any statistic of interest without making assumptions about the shape or parameters of the population distributions. This empirical sampling distribution emerges from multiple randomizations of the observed

data, and the percentile location of the observed statistic on this distribution helps assess its likelihood under the null hypothesis.

In cases with small samples, it may be impractical to calculate the statistic for every possible order, so randomization becomes a more feasible method. For instance, with 10 cases per group, there are 20 choose 10 ($^{20}C_{10}$) = 184,756 possible randomizations, making randomization a more practical approach to generate the sampling distribution.

## 1.2. CROSS-VALIDATION

Cross-validation is a fundamental technique used to estimate the test error associated with a statistical learning model, serving both for performance evaluation and model flexibility selection. This approach involves dividing the dataset into two primary subsets: a training set and a validation (or hold-out) set. The model is trained on the training set, and its performance is assessed by making predictions on the validation set.

Several types of cross-validation techniques exist, each offering unique advantages:

Leave-One-Out Cross-Validation (LOOCV):

- In LOOCV, only one observation is used for validation at a time, while the rest constitute the training set. This process is repeated for each observation in the dataset, and the final performance metric is the average of the results for all observations. LOOCV is less biased than the validation set approach and tends to provide a more accurate estimate of the test error rate.

K-Fold Cross-Validation

- K-fold cross-validation involves randomly dividing the observations into k folds, treating one fold as the validation set, and fitting the model on the remaining folds. This process is repeated k times, with different

groups treated as the validation set in each iteration. The final performance metric is the average of the k results on new data.

LOOCV has advantages over the validation set approach. It has less bias since the statistical learning method is repeatedly fit using training sets that contain almost as many observations as the entire dataset. Additionally, LOOCV yields consistent results upon repeated applications, as there is no randomness in the training/validation set splits.

Cross-validation is widely used in predictive statistical models, allowing the separation of several pieces of data as the validating set within a sampling. The remaining observations constitute the training set, enabling predictions on the validating set. This iterative process helps in gathering accuracy means for predictions, preventing self-influence and providing a more accurate assessment of model performance.

Cross-validation, a powerful resampling technique, can be categorized into several types, each offering unique insights into model performance. Here, we explore simple cross-validation, double cross-validation, and multicross-validation, focusing on their application in regression analysis.

Simple Cross-Validation:

- In the context of regression, simple cross-validation involves dividing the dataset into sub-samples. The first sub-sample is used to derive the regression equation, while another sub-sample is employed to generate predicted scores using this equation. The cross-validity coefficient is then calculated by correlating the predicted and observed scores on the outcome variable. This method provides a straightforward evaluation of model performance.

Double Cross-Validation:

- Building upon simple cross-validation, double cross-validation takes regression analysis a step further. Regression equations are generated

independently in both sub-samples. Both equations are then utilized to generate predicted scores and cross-validity coefficients. This more comprehensive approach aims to enhance the precision of performance evaluation.

Multicross-Validation:

Multicross-validation extends the principles of double cross-validation. In this form, the double cross-validation procedures are repeated multiple times by randomly selecting sub-samples from the dataset. This iterative process introduces additional variability, providing a more robust assessment of model performance.

In regression analysis, the beta weights computed in each sub-sample play a pivotal role in predicting the outcome variable within the corresponding sub-sample. The observed and predicted scores for the outcome variable in each sub-sample are then utilized to compute the cross-validated coefficient. This coefficient serves as a comprehensive metric, consolidating the model's effectiveness across diverse sub-samples.

Limitations and Considerations:

- Cross-validation shares a limitation with split-half reliability when dealing with small sample sizes. The act of dividing an already limited sample into two halves increases the risk that beta weights may be artifacts of the sub-sample, as emphasized by Ang (1998). Caution is essential, and results must be interpreted carefully, especially in situations involving smaller datasets.

In summary, the progression from simple to double and multicross-validation signifies a refinement in evaluating model performance, particularly in the context of regression analysis. While cross-validation provides valuable insights, researchers need to be mindful of its

limitations and exercise due diligence in experimental design, particularly when working with smaller sample sizes.

In the broader context of regression analysis, cross-validation is indispensable for determining the optimal number of predictor variables. Unlike traditional methods, the addition of predictors doesn't always result in a reduction of the residual sum of squares; instead, the cross-validated mean square error takes into account the impact of valuable predictors.

Overall, cross-validation stands out as a potent resampling technique, offering a robust estimate of model performance and facilitating informed decisions regarding model selection and generalization capabilities.

## 1.3 BOOTSTRAPPING

Statistical inference plays a pivotal role in drawing meaningful conclusions from data, and in this realm, the Bootstrap Resampling method emerges as a powerful and flexible tool.
Bootstrap resampling emerged as a breakthrough in statistical methodology, inspired by earlier work on the jackknife. Bradley Efron's initial work in 1979 laid the foundation, and subsequent developments, including bias-corrected and accelerated bootstrap, further enhanced its accuracy and applicability. The method's evolution reflects a commitment to refining its efficiency and expanding its scope.
Bootstrap resampling has become a cornerstone in statistical methodology, offering robust solutions in scenarios where traditional methods fall short.
The fundamental principle of bootstrap resampling lies in its ability to emulate the population distribution by resampling with replacement from the original data. This process creates a simulated sample of the same size as the initial dataset, allowing statisticians to calculate the

statistic of interest repeatedly. Notably, this sampling with replacement allows for the possibility of an observation being selected more than once in a particular trial.

The statistic of interest, be it a mean, median, or other parameter, is calculated for each bootstrapped sample. This process is iterated thousands of times, generating a distribution of the statistic. The mean of these bootstrapped values serves as the estimate of the parameter, and the resulting distribution is instrumental in constructing confidence intervals.

Bootstrap resampling offers several advantages that contribute to its widespread adoption in statistical practice:

- Non parametric approach
  Bootstrap does not rely on stringent distributional assumptions, making it applicable even when data deviate from normality. This flexibility is especially valuable in situations where traditional methods may be unreliable.
- Applicability to Complex Sampling Designs:
  Bootstrap adapts seamlessly to various sampling designs, including stratified and clustered samples. Its versatility extends to complex data-collection plans, allowing for straightforward application.

- Estimation of Confidence Intervals:
  One of the primary applications of bootstrap resampling is the construction of confidence intervals. The distribution of bootstrapped values is used to determine confidence limits, providing a robust measure of parameter uncertainty.
- Simple yet Powerful:
  Despite its sophistication, bootstrap resampling is remarkably simple in concept. It efficiently derives estimates of standard errors and confidence intervals for a diverse range of estimators, including percentiles, proportions, odds ratios, and correlation coefficients.
- No Distributional Assumptions:
  The method excels in scenarios with small sample sizes or poorly behaved data, as it does not rely on strict distributional assumptions.

Bootstrap extends its utility beyond estimation to hypothesis testing. Efron and Tibshirani proposed an algorithm for comparing means of two independent samples using bootstrap resampling. This approach involves generating resamples, calculating test statistics, and iteratively assessing the null hypothesis, providing a robust alternative to traditional hypothesis testing methods.
In comparison to the jackknife method, bootstrap resampling offers a more thorough approach. The ability to replicate the original sample extensively and employ sampling with replacement contributes to its accuracy in simulating chance. This fundamental difference positions bootstrap as a more versatile and powerful resampling technique.

The breadth of applications for bootstrap resampling is vast. It proves invaluable in estimating parameters, constructing confidence intervals, and conducting hypothesis tests across diverse fields such as economics, biology, physics, and genetics. Its adaptability to complex sampling designs and independence from distributional assumptions further extends its utility.

Limitations and Consideration:
Despite its strengths, bootstrapping has certain limitations and considerations:
Dependence on Estimator: The accuracy of bootstrapped estimates is contingent on the choice of the statistic or estimator.

- While less reliant on distributional assumptions, bootstrapping assumes the original sample is representative of the population.
- Conducting thousands of bootstrap iterations can be computationally demanding, particularly for large datasets.
- Bootstrapping may yield unreliable results with small sample sizes or highly skewed data.

Bootstrap resampling stands as a cornerstone in statistical inference, offering a flexible and robust approach to parameter estimation, hypothesis testing, and confidence interval construction. Its simplicity, versatility, and adaptability make it a preferred method in situations where traditional techniques may falter. As researchers continue to explore and refine statistical methodologies, bootstrap resampling

remains a beacon of innovation and reliability, shaping the landscape of modern statistical practice.

## 1.4 JACKKNIEF

The Jackknife resampling technique, pioneered by Maurice Quenouille in the 1950s, stands out as a robust statistical method for estimating the bias and variance of statistical estimators without imposing strict distributional assumptions.

Jackknife, a resampling method, is designed to gauge the robustness and variability of statistical estimators. Its nonparametric nature sets it apart, offering broad applicability across diverse settings. Similar to other resampling methods like bootstrap and cross-validation, Jackknife's computational demands can be substantial.

Fundamentally, Jackknife generates resamples by systematically excluding observations from the original dataset. The delete-one Jackknife involves sequentially removing single cases from the original sample, while the more generalized delete-d Jackknife leaves out multiple cases at once.

The Jackknife enables estimation of both bias and variance for a statistical estimator. The bias is determined as the average difference between the original estimator ($\widehat{\theta}$) and estimators derived from subsets. Variance is computed as the average squared difference between the estimator from each subset and the mean of all estimators.
Consider a scenario where our focus lies on a parameter (K), yet we possess solely an estimate (k), derived from a sample of n observations. In this context, the objective is to refine the estimation of K using the available statistic k. The approach involves generating a series of jackknifed samples, each comprising (n-1) observations. This process unfolds by iteratively excluding one observation from the original dataset. For each jackknifed sample, a corresponding statistic ($k_i$) is computed, denoting the estimate of K based on the remaining (n-1) observations. The pseudovalue $k_i$ is formulated as $k - (n - 1)(k_i - k)$. This procedure is reiterated for each possible omission, resulting in a set of n

pseudovalues . Subsequently, the jackknifed estimate of K is determined as the mean of these pseudovalues ($K = \bar{k}$). This process provides a refined estimation of the parameter K by systematically considering the exclusion of individual observations in the jackknife resampling technique.

Additionally, the standard error of the jackknifed estimate is determined by dividing the standard deviation of the pseudovalues by the square root of the sample size .

$SE_K = \frac{sd(k)}{\sqrt{n}}$

The confidence limits for K are established using the estimated K, $SE_K$, and the two-tailed critical t-score with (n-1) degrees of freedom and significance level $\frac{\alpha}{2}$.

$K \pm SE_K \, t_{(n-1),\alpha/2}$

This comprehensive approach allows for a robust assessment of the parameter K along with its associated standard error and confidence limits through the jackknife resampling technique.

Delete – d Jackknief:

To address limitations related to the smoothness of statistics, the delete-d Jackknife is introduced. Rather than excluding one observation at a time, d observations are omitted. The formula for the delete-d Jackknife estimate of the standard error is given by:

$SE = \sqrt{\frac{\sum_s (\hat{\theta}_s - \bar{\theta}_d)^2}{n-d}}$

Here, $\hat{\theta}_s$ signifies the estimate from a subset s of size n-d ,and $\bar{\theta}_d$is the mean of all estimates from subsets of size n-d.

<u>Pros of Jackknife Resampling:</u>

- Jackknife is a nonparametric method, making it robust and applicable in various statistical scenarios without strict distributional assumptions.
- The technique is versatile and adaptable to different data structures, allowing for its effective use in diverse settings.

- Jackknife is resilient to certain assumption violations, making it a reliable tool for estimating bias and variance.
- The leave-one-out procedure in Jackknife is effective for detecting outliers and influential cases in the data.
- The method's simplicity makes it straightforward to implement, providing a practical solution for various statistical problems.

Cons of Jackknife Resampling:

- Jackknife may perform inadequately for statistics that do not change smoothly across repetitions, leading to potential underestimation of standard errors.
- In small sample sizes, the number of repetitions or resamplings in Jackknife is limited, potentially impacting its effectiveness.
- The method may not perform well for non-smooth statistics, such as the median, resulting in underestimated standard errors.
- Like other resampling techniques, Jackknife can be computationally intensive, requiring significant computing resources for large datasets or complex analyses.
- While versatile, Jackknife may not be the most efficient tool for certain specific problems, necessitating consideration of alternative resampling methods based on the nature of the analysis.

The Jackknife resampling technique proves invaluable for statisticians and researchers seeking to assess estimator bias and variance without rigid distributional assumptions. Its simplicity, adaptability, and applicability across various data structures underscore its significance. However, users must remain mindful of limitations, especially in cases involving non-smooth or nonlinear statistics. The Jackknife, with its historical importance and continued relevance, remains a cornerstone in statistical resampling methods.

## 1.5. MONTE CARLO METHOD

The Monte Carlo method is a powerful simulation technique widely used to analyze and solve problems, particularly in situations where the parent distribution is either known or assumed. Originating from research conducted at Los Alamos in the 1940s and named after the Monte Carlo Casino in Monaco, this method involves creating statistical distribution functions by utilizing a series of random numbers.

Monte Carlo methods are often applied in simulations where the underlying distribution is either explicitly known or assumed. For example, in simulating the distribution of an F-statistic for two samples drawn from a normal distribution, the mean and variance of the normal distribution are specified. Through the generation of multiple samples and the calculation of the ratio of their variances, a frequency distribution of the F-statistic is produced. This technique is crucial for obtaining numerical results in scenarios where deterministic solutions are difficult or impossible.

Methodology and steps:

The Monte Carlo Method revolves around the principle of learning about a distribution, process, or system through random sampling. This involves the following steps:

- Identify the statistical properties of possible inputs relevant to the problem at hand.
- Create numerous sets of possible inputs based on the identified statistical properties.
- Conduct deterministic calculations or simulations using the generated sets of inputs.
- Analyze the results statistically, often by examining the distribution of outcomes and deriving relevant measures of interest.

Monte Carlo methods are particularly beneficial in three problem classes: optimization, numerical integration, and generating draws from a probability distribution. The use of random sampling, facilitated by extensive random number sequences, allows for addressing complex problems that may not have straightforward analytical solutions.

In the realm of statistical studies, Monte Carlo techniques offer a valuable approach for exploring the sensitivity of statistical tests and

estimators. By specifying populations with known characteristics and performing random sampling, researchers can generate sampling distributions for statistics of interest. For instance, Monte Carlo studies have revealed insights into the behavior of the t-test under various conditions, showcasing its robustness under certain scenarios but highlighting challenges with small, unequal sample sizes and differing population variances.

The Monte Carlo method stands as a versatile and robust approach for tackling problems across diverse domains. Its ability to leverage randomness for problem-solving, simulate complex scenarios, and provide valuable insights into statistical analyses makes it an indispensable tool in contemporary research and simulation studies.

# CHAPTER:2
# ESTIMATION USING BOOTSTRAP:

## 2.1 BIAS IN BOOTSTRAPPING:

In the context of bootstrapping, bias refers to the discrepancy between the expected value of a bootstrapped estimate and the true population parameter it aims to estimate. Bias can arise due to various factors, such as the specific resampling method used, the finite sample size, or the underlying distribution of the data.

To understand the concept of bias in bootstrapping, let's delve into the process of bootstrapping and its potential sources of bias:

- . Resampling Method: Bootstrapping typically involves resampling from the observed data with replacement to create multiple bootstrap samples. The resampling procedure itself may introduce bias if it does not accurately capture the underlying distribution of the data. For example, if the original sample is not representative of the population, resampling from it may perpetuate the biases present in the original sample.

- Finite Sample Size: In cases where the original sample size is small, the bootstrap estimates may be biased due to the limited amount of information available in the data. Small samples may not adequately represent the population distribution, leading to biased estimates of population parameters.

- Non-i.i.d. Data: Bootstrapping assumes that the data are independent and identically distributed (i.i.d.). If the data violate these assumptions, such as in the presence of autocorrelation or clustering, the bootstrap estimates may be biased. In such cases, alternative resampling methods, such as clustered bootstrapping or wild bootstrapping, may be more appropriate to account for the data structure.

- Model Misspecification: If the underlying statistical model used for bootstrapping is misspecified, the resulting estimates may be biased. For example, if the data exhibit heteroscedasticity or non-normality, but the bootstrapping assumes a parametric model without considering these issues, the estimates may be biased.

- Sample Variability: Due to the random nature of bootstrapping, the estimates obtained from different bootstrap samples may vary. This variability can contribute to bias if the average of the bootstrapped estimates deviates systematically from the true parameter value.

Bias in bootstrapping can be assessed by comparing the expected value of the bootstrapped estimates to the true parameter value or by examining the distribution of the bootstrap estimates. If the average of the bootstrap estimates is systematically higher or lower than the true parameter value, it indicates the presence of bias.

To mitigate bias in bootstrapping, researchers can consider using alternative resampling methods, increasing the sample size, or assessing the sensitivity of the results to different modeling assumptions. Additionally, conducting sensitivity analyses and validating the results using alternative estimation techniques can help identify and address potential biases in bootstrapping.

Estimating Bias:

Bias in an estimator is a measure of systematic error, or the difference between the expected value of the estimator and the true value of the parameter it estimates. Let E(X) denote the expected value of random variable X. Consider $\hat{\theta}$ is an estimator of the parameter $\theta$,the bias of $\hat{\theta}$ is defined as: $b=E(\hat{\theta})-\theta$
An estimator is unbiased when $b$=0.

Bootstrap Method for Bias Estimation:
The bootstrap is a resampling technique that involves repeatedly drawing samples from a dataset with replacement and calculating the statistic of interest. This technique provides a way to estimate the sampling distribution of almost any statistic using only the data at hand.
Steps for estimating bias using the bootstrap method

- Generate Boostrap Samples:

From your original dataset, generate $B$ bootstrap samples. Each bootstrap sample is the same size as the original dataset and is created by randomly selecting observations from the original dataset with replacement.

- Compute the Estimator on Each Bootstrap Sample:

For each of the $B$ bootstrap samples, compute the estimator. This results in a series of bootstrap estimates: $\widehat{\theta_1^*}$, $\widehat{\theta_2^*}$,…, $\widehat{\theta_B^*}$ .

- Calculate the mean of bootstrap estimates
  Calculate the mean of these bootstrap estimates: $\overline{\theta^*}=\frac{1}{B}\sum_{i=1}^{B}\widehat{\theta_i^*}$
- Estimate the bias:
  The bias of the estimator can be estimated by comparing the bootstrap mean to the estimator calculated from the original dataset: $\hat{b}=\overline{\theta^*}-\hat{\theta}$ where $\hat{\theta}$ is the estimator computed from the original sample data.

## 2.2 ESTIMATING VARIANCE

The Bootstrap method, offers a robust and flexible approach for estimating the variance of statistical estimators. This method is particularly valuable in settings where traditional parametric assumptions are untenable or where the data exhibit unique characteristics that defy conventional analytical approaches. Its applicability spans a wide array of statistical challenges, from evaluating the variance of complex estimators to handling datasets with peculiar distributions.

The core principle behind the Bootstrap method is the simulation of the sampling distribution of a statistic by repeatedly resampling from the original dataset with replacement. Each resampling iteration creates a new "bootstrap sample," typically the same size as the original dataset, but composed such that some original observations may appear more than once or not at all. This sampling with replacement mimics the process of obtaining new sample sets from the same underlying population.

To estimate the variance of a statistic using the Bootstrap method, the procedure involves generating a substantial number of these bootstrap samples. For each sample, the statistic of interest is recalculated, resulting in an array of bootstrap replicates of the statistic. The variance among these replicates serves as an estimate of the variance of the statistic under the real-world condition of repeatedly sampling from the population. Mathematically, the variance of the bootstrap estimates $\widehat{\theta^*}$ is computed as:

$$\mathrm{Var}(\hat{\theta}) \approx \frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\theta_b^*} - \overline{\theta^*}\right)^2$$

where B is the number of bootstrap samples, $\widehat{\theta_b^*}$ is the estimate from the $b^{th}$ bootstrap sample, and $\overline{\theta^*}$ is the average of the bootstrap estimates.

The Bootstrap's major advantage lies in its minimal assumption requirement. Unlike methods that require normality or other specific distribution characteristics, the Bootstrap method can be applied to data with attributes like skewness, heavy tails, or significant bounds. These features make it exceptionally useful for empirical data analysis in real-world scenarios, where such conditions are frequently encountered.

Despite its flexibility and power, the Bootstrap method does come with limitations. One significant challenge is its computational intensity. Estimating the variance through the Bootstrap requires recalculating the statistic many times—often thousands—depending on the number of bootstrap samples. This can be computationally demanding, especially with large datasets or complex statistical models. Additionally, the Bootstrap may not provide accurate variance estimates for statistics that are highly sensitive to outliers or are not smooth functions of the data, such as quantiles.

Moreover, while the Bootstrap is less sensitive to the shape of the data distribution, it assumes that the sample represents the underlying population well. This assumption can lead to biased variance estimates if the original sample has peculiarities not representative of the population.

The Bootstrap method provides a practical and effective means for variance estimation, particularly suited to complex or uniquely distributed data. Its simplicity in implementation and broad applicability make it a valuable tool in the statistician's arsenal. However, researchers must consider its computational demands

and limitations in scope, particularly concerning the data's representativeness and the nature of the statistic being estimated.

## 2.3 ESTIMATING STANDARD ERROR

The bootstrap is a powerful statistical tool for estimating the standard error and bias of an estimator using the original sample data to generate new samples. This method is particularly useful when the underlying distribution of the data is unknown or when the sample size is not large enough to rely on asymptotic approximations.

We have, the bootstrap method is based on the idea of resampling from the observed data and recalculating the estimator to form a new distribution of the estimator. By doing so, it effectively simulates drawing additional samples from the same population from which the original sample was drawn.

Assume we have a sample $X=(X_1,X_2,\ldots,X_n)$ drawn from an unknown distribution F. Let $\hat{\theta}=S(\mathrm{X})$ be the statistic (or estimator) we are interested in, where S is some function of the sample.

Define $\hat{F}$ be the empirical distribution that assigns equal probability (1/n) to each data point in the sample.Then generate B bootstrap samples $X_1^*, X_2^*,\ldots., X_B^*$.where each $X_b^*$ is a sample of size n drawn from $\hat{F}$ with replacement.For each bootstrap sample $X_b^*$,calculate the bootstrap replicate of the statistic: $\widehat{\theta_b^*}$, $=S(X_b^*)$.

Compute the standard deviation of the bootstrap replicates

$$SE_{\hat{F}}(\hat{\theta}) \approx \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta_b^*} - \overline{\theta^*}\right)^2}$$

Here, $\overline{\theta^*}$ is the mean of the bootstrap replicates $\widehat{\theta_b^*}$

## 2.4 BOOTSTRAP CONFIDENCE INTERVAL

Confidence intervals are a fundamental concept in statistics, providing a range of values that are believed to contain the population parameter of interest, such as the mean or proportion, with a certain level of confidence. Typically expressed as a percentage, the confidence level (most commonly 95%) quantifies the degree of certainty in the interval estimate. The wider the interval, the higher the confidence level, reflecting greater certainty that the parameter lies within the interval, but also less precision in estimation.

Bootstrap confidence intervals offer a flexible alternative to traditional methods of interval estimation that rely on assumptions about the distribution of the sample statistic (e.g., normality). This method utilizes the original data set to generate numerous additional samples, known as bootstrap samples, which are created by sampling with replacement from the original data. Each sample is the same size as the original and is used to compute the statistic of interest. The distribution of these bootstrap statistics forms the basis for confidence interval estimation.

<u>The Standard Normal Bootstrap Confidence Interval</u>:

The Standard Normal Bootstrap Confidence Interval is a straightforward approach to estimating confidence intervals for a parameter $\theta$ ,but it is not without its limitations. This method typically hinges on the assumption that the estimator, $\hat{\theta}$ behaves in a manner that aligns with the Central Limit Theorem (CLT). Specifically, if $\hat{\theta}$ is the sample mean and the sample size is sufficiently large, the CLT suggests that the distribution of the standardized statistic,

$$Z=\frac{\hat{\theta}-\mathrm{E}[\hat{\theta}]}{SE(\hat{\theta})}$$

will approximate a standard normal distribution. In this framework, where $\hat{\theta}$ is assumed to be an unbiased estimator of , an approximate $100(1-\alpha)\%$confidence interval for $\theta$ can be constructed using the formula:

$\hat{\theta} \pm z_{\alpha/2} \cdot SE[\hat{\theta}]$

where $z_{\alpha/2}$ is the critical value from the standard normal distribution corresponding to $\alpha/2$.

The Standard Normal Bootstrap Confidence Interval simplifies the computation of confidence intervals but rests on a few crucial assumptions. Primarily, this method assumes that the distribution of the estimator $\hat{\theta}$ is normal. Typically, this assumption holds reasonably well when $\hat{\theta}$ is a sample mean, especially as the sample size becomes large, owing to the Central Limit Theorem (CLT). Furthermore, it is essential that $\hat{\theta}$ is an unbiased estimator of $\theta$.Should $\hat{\theta}$ exhibit any bias, the central point of the confidence interval may not accurately align with $\theta$ potentially leading to incorrect conclusions.

Another significant assumption involves the standard error, $SE[\hat{\theta}]$which the method assumes is either known or can be reliably estimated. Within the bootstrap framework, this standard error is typically estimated using the sample standard deviation of the bootstrap replicates. However, this introduces an additional layer of variability that can affect the accuracy and reliability of the confidence interval.

Addressing these assumptions in practical scenarios is crucial, especially where they may not hold. If there is bias in $\hat{\theta}$ ,it can be quantified and used to adjust the Z-statistic, potentially enhancing the accuracy of the confidence interval. However, because $\hat{\theta}$ and its standard error are themselves random variables, the transformation relying on their assumed normal distribution may lead to complications in interpreting and ensuring the accuracy of the interval.

Moreover, treating the standard error $SE[\hat{\theta}]$ as a known constant simplifies the calculations but ignores the inherent variability and uncertainty in estimating this parameter, especially when derived from bootstrap replicates. This oversight can lead to an underestimation of the confidence interval's width, thereby exaggerating the precision of the estimate.

While the standard normal bootstrap confidence interval is a convenient and computationally efficient approach for interval estimation, statisticians must remain aware of its foundational assumptions and inherent limitations. It performs optimally

under conditions of large sample sizes, presumed normality, and minimal bias. In scenarios where these conditions are unmet, alternative bootstrap methods such as the percentile method, the Bias-Corrected and Accelerated (BCa) method, or the studentized bootstrap might offer more robust and reliable confidence interval estimates.

i. Percentile Method
The percentile method is the simplest approach to deriving bootstrap confidence intervals. After resampling the original dataset with replacement many times and calculating the statistic of interest for each sample, the percentile confidence interval is determined directly from these bootstrap samples. For a 95% confidence interval, the 2.5th and 97.5th percentiles of the bootstrap statistics are used as the lower and upper bounds, respectively. This method assumes that the bootstrap distribution is a good approximation of the sampling distribution of the statistic, but it does not adjust for any bias or skewness in the distribution.

ii. Bias-Corrected and Accelerated (BCa)
The BCa method enhances the basic percentile method by correcting for both bias in the bootstrap distribution and its asymmetry (acceleration). This method involves calculating a bias-correction factor that adjusts the percentile values used to construct the confidence interval. It also considers the skewness of the bootstrap distribution to better align the confidence intervals with the true parameter values. The BCa method generally provides more accurate confidence intervals, especially when the distribution of the sample statistic is skewed or when bias is present in the bootstrap estimates.

iii. Studentized Bootstrap
The studentized bootstrap, also known as the bootstrap-t method, is a sophisticated approach to constructing confidence intervals that accounts for the variability in the standard error across bootstrap samples. This method recalculates both the statistic of interest and its standard error for each bootstrap sample, capturing potential

variations in estimation precision that simpler bootstrap methods, like the percentile method, might miss.

For each bootstrap sample, this method computes both a parameter estimate ($\widehat{\theta_n}$) and its standard error ($\widehat{\sigma_n}$). The bootstrap-t statistic is then calculated using the formula:

$$R_n = \frac{(\widehat{\theta_n^*} - \widehat{\theta_n})}{\widehat{\sigma_n^*}}$$

where $\widehat{\theta_n^*}$ and $\widehat{\sigma_n^*}$ are derived from the bootstrap sample. The distribution of $R_n$, known as $G_n$, is typically unknown and is estimated by the bootstrap distribution $G_{boot}$, which is expressed as:

$$G_{boot}\,(x) = P^*\left(\frac{(\widehat{\theta_n^*} - \widehat{\theta_n})}{\widehat{\sigma_n^*}} \leq x\right)$$

Here, $P^*$ refers to the probability under the bootstrap framework.

To estimate the confidence interval, particularly the lower confidence limit, the inverse of $G_{boot}$ is used. The formula for the lower bound at a confidence level of $100(1-\alpha)\%$

$$\theta_{BT} = \widehat{\theta_n} - \widehat{\sigma_n} G_{boot}^{-1}(1 - \alpha)$$

This involves accurately determining the $\alpha^{th}$ percentile of $G_{boot}$ using the bootstrap samples. If B bootstrap replicates are used, the $\alpha^{th}$ percentile is estimated by selecting the $k^{th}$ largest value from $R_n^*(b)$, where k is calculated as $[(B + 1)\alpha]$, and [] indicates rounding to the nearest integer.

The studentized bootstrap method is highly valued for its accuracy and robustness in adjusting for variations in the standard error of estimates across samples. While it requires more computational effort and advanced statistical methods, the precision and dependability it offers make it particularly useful in complex analyses where simpler methods might not provide reliable outcomes. This method is thus often preferred in statistical analyses where accuracy and reliability are crucial.

iv. Adjusted Bootstrap Percentile (ABC)
The adjusted bootstrap percentile method (ABC) modifies the simple percentile method by adjusting the endpoints of the confidence interval to account for observed bias in the bootstrap distribution. This adjustment is typically a simple shift based on the mean difference between the bootstrap estimates and the original statistic. It aims to improve the accuracy of the interval by correcting for bias, but it does not address skewness like the BCa method does.

The choice among these methods often depends on the size and nature of the data, as well as computational resources. The simple percentile method may suffice for large samples with a distribution that is close to symmetric. However, for skewed distributions or smaller samples, BCa or studentized methods offer better accuracy at the cost of additional computation. Each method has its strengths and is suited to different types of data and statistical requirements.

## 2.5 BOOTSTRAP HYPOTHESIS TESTING

Hypothesis testing is a foundational concept in statistics, used to decide whether there is enough evidence in a sample of data to support a particular belief about the population from which the sample was drawn. Traditional methods of hypothesis testing rely on the assumption that the distribution of the test statistic under the null hypothesis is known. However, in many practical scenarios, the exact distribution of the test statistic may be unknown or difficult to determine. This is where bootstrap methods come into play as a robust alternative.

The bootstrap method  is a powerful tool for assessing the distribution of a statistic based on random sampling with replacement from the data. This approach allows the estimation of the sampling distribution of almost any statistic using only the data at hand, without any strict assumptions about the form of the population distribution.

To perform a bootstrap hypothesis test, one first generates a large number B of bootstrap samples from the original data. Each bootstrap sample is typically the same size as the original dataset and is created by sampling with replacement. This means each sample could have repeated values or might not include some values from the original data at all.

For each bootstrap sample, a test statistic (denoted as $\tau_j^*$) is computed in a similar manner to how the observed test statistic ($\hat{\tau}$) is calculated from the actual sample. It's essential to ensure these bootstrap samples are consistent with the null hypothesis, although this alignment isn't always straightforward.

The core idea in bootstrap hypothesis testing is to compare the observed test statistic to the distribution of the bootstrap test

statistics to calculate the p-value. The bootstrap p-value is defined as:

$\hat{p}^*(\hat{\tau})$ =1- $\hat{F}^*(\hat{\tau}) = \frac{1}{B}\sum_{j=1}^{B} I(\tau_j^* > \hat{\tau})$

Here, $\hat{F}^*$ denotes the empirical distribution function, or EDF, of the $\tau_j^*$ ,and I is the indicator function that is 1 if the condition inside is true and 0 otherwise. This formula calculates the proportion of the bootstrap test statistics that are more extreme than the observed statistic.

Example:

Imagine we want to test if the mean height of a group of individuals is equal to 170 cm. From our sample data:

Observed mean height, $\hat{\tau}$, is 172 cm.

We generate 1000 bootstrap samples from our data, recalculating the mean for each sample.

Let's say 250 out of these 1000 bootstrap sample means exceed 172 cm.

Thus, our bootstrap p-value is:

$\hat{p}^* = \frac{250}{1000} = 0.25$

If we set our significance level, $\alpha$, to 0.05, and since our p-value (0.25) is greater than $\alpha$, we do not reject the null hypothesis. This result implies that there is not enough evidence at the 5% significance level to conclude that the average height differs from 170 cm.

Bootstrap hypothesis testing offers a flexible and powerful alternative to traditional methods, particularly useful when the distribution of the test statistic under the null hypothesis is

unknown or difficult to estimate. This method not only provides a way to approximate the necessary distributions but does so with minimal assumptions about the underlying population. By leveraging the power of resampling, bootstrap methods can help make more informed and robust statistical decisions.

# CHAPTER 3
# ESTIMATION USING JACKKNIEF

## 3.1 BIAS REDUCTION USING JACKKNIFE

Bias estimation and reduction using the Jackknife method is a crucial technique in statistics for improving the accuracy of estimators derived from sample data. Introduced by Quenouille in 1949 and later refined, the Jackknife method is primarily used to estimate and reduce the bias of an estimator, as well as to estimate its variance. This method is especially valuable when dealing with estimators based on small or moderate sample sizes.

The principle behind the Jackknife method involves repeatedly recalculating the estimator by systematically leaving out each observation in the sample one at a time. For an estimator $\widehat{\theta_n}$based on n independent and identically distributed (i.i.d.) random vectors $X_1,\ldots,_n$ the Jackknife method re-computes the statistic $\hat{\theta}_{n,-i}$ by excluding the $i^{th}$observation from the data. Here, $\hat{\theta}_{n,-i}=f_{n-1}(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n)$where $f_{\text{n-1}}$ is a function analogous to $f_{\text{n}}$ but for n-1 observations.

The Jackknife estimator of bias Bias is calculated as follows:

$\text{Bias}_J = \frac{(n-1)}{n}\sum_{i=1}^{n}\ (\ \hat{\theta}_{n,-i} - \widehat{\theta_n})$

where the factor $(n-1)$accounts for the change in variance due to the reduction in sample size. This calculation results in an estimate of the bias of $\widehat{\theta_n}$

To correct the bias of $\widehat{\theta_n}$ the Jackknife bias-corrected estimator $\theta_{\text{J}}$ is computed:

$\theta_J=\widehat{\theta_n}-\text{Bias}_J = \frac{1}{n}\sum_{i=1}^{n}(n\widehat{\theta_n} - (n-1)\hat{\theta}_{n,-i})$

This formula essentially averages the so-called pseudo-values,

$\theta_{n,}= \widehat{\theta_n} -(n-1)\,\widehat{\theta_n}$, each of which is an individual adjustment reflecting the influence of each data point.

From a theoretical standpoint, if the expected value of $\widehat{\theta_n}$ deviates from $\theta$ due to terms involving 1/n or higher powers, the Jackknife method helps in reducing the bias significantly. For instance, if $\text{E}(\widehat{\theta_n})=\theta+\frac{a}{n}+\frac{b}{n^2}$ , the expected value of each pseudo-value $\theta_{(\text{n,i})}$ under the Jackknife method becomes $\theta-\frac{b}{n(n-1)}$ . This means the overall bias of $\theta_{\text{J}}$ converges to

zero at a rate of ($n^{-2}$) which is a faster rate than the ($n^{-1}$) rate of convergence for $\widehat{\theta_n}$ .
That is, the Jackknife method provides a practical and effective way to reduce the bias of estimators in statistical inference, making it particularly useful in scenarios where precision is critical and sample sizes are not very large. Its ability to also estimate the variance of the estimator further adds to its utility in statistical applications.

## 3.2 ESTIMATION OF VARIANCE

Estimating variance using the Jackknife method is a valuable statistical technique designed to provide robust estimates of the variability of a statistical estimator. The Jackknife method works by systematically recalculating the estimator after omitting one observation at a time from the dataset. This approach helps in assessing the influence of individual data points on the overall estimate, leading to a deeper understanding of the estimator's variability.
To begin with, consider an estimator $\widehat{\theta_n}$ , calculated from n observations. The Jackknife method involves creating n new datasets, each missing one of the original observations. For each new dataset, the estimator $\hat{\theta}_{n,-i}$ is recalculated. The influence of the omitted observation can then be assessed by examining how much the recalculated estimator deviates from the original estimator $\widehat{\theta_n}$.
The variance of the estimator using the Jackknife technique is given by:
$\text{var}(\widehat{\theta_n}) = \frac{(n-1)}{n} \sum_{i=1}^{n} (\hat{\theta}_{n,-i} - \overline{\theta_n})^2$
Here, $\overline{\theta_n}$ is the average of the n recalculated estimators $\hat{\theta}_{n,-i}$ . This formula acknowledges the fact that each Jackknife sample is slightly less varied due to the smaller sample size, and it scales up the variance estimate by n-1 to compensate for this effect.
One primary reason to use the Jackknife method is its simplicity and general applicability. It does not heavily depend on the underlying distribution of the data, making it versatile for a variety of statistical estimates, especially those based on means or other simple statistics that are smooth functions of the data.

Additionally, the Jackknife method is particularly useful for estimating the variance of estimators that are not straightforward, such as medians or other robust statistics, albeit with some limitations. For such estimators, the Jackknife can provide insight into the variability of the estimate, even if it may not always deliver consistency.
While the Jackknife is powerful, it's not infallible. It generally performs well with smooth statistics but can fail with non-smooth statistics like medians or quantiles, where other techniques like the bootstrap might be more appropriate. The Jackknife variance estimator might be biased in such non-smooth scenarios.
In cases where the statistic is heavily influenced by outliers or is very non-linear, the Jackknife method's assumptions about the influence of omitting a single observation might not hold, leading to inaccurate variance estimates. Additionally, for very large datasets, the computational burden of recalculating the estimator $n$ times might be impractical, although grouped or delete-d Jackknife methods can be used to reduce computational costs.
The Jackknife method provides a straightforward and often effective way to estimate the variance of statistical estimators. It works by assessing the impact of removing individual observations, thereby providing insights into the stability and reliability of the estimator. While it has limitations, particularly with non-smooth statistics, it remains a widely used tool in statistical analysis for its simplicity and general applicability to a range of statistical problems.

## 3.3 ESTIMATION OF STANDARD ERROR USING JACKKNIFE

The jackknife technique stands as one of the earliest methods used to derive reliable statistical estimators and is noted for requiring less computational power than many newer methodologies. Consider a dataset composed of n observations, denoted as ($x_1$, $x_2$,.., $x_n$), and suppose we have an estimator $\hat{\theta}$=s(x), where s is a function applied to the dataset x.

The jackknife procedure involves generating multiple subsets from the original dataset by omitting one observation at a time. These subsets are known as jackknife samples. For each index $i$ ranging from 1 to $n$ ,

the $i^{th}$ jackknife sample is formed by excluding the $i^{th}$ observation from the dataset, resulting in ($x_1$, $x_2$,…, $x_{i-1}$, $x_{i+1}$,…, $x_n$).

From each of these jackknife samples, the estimator $\hat{\theta}$ is recalculated, yielding what is referred to as the $i^{th}$ jackknife replication, $\hat{\theta}_{(i)}$ , of the estimator. The standard error of $\hat{\theta}$ using the jackknife method is then estimated by examining the variability of these replications around their average value. The formula for the jackknife estimate of the standard error, denoted $\widehat{SE}_J$ , is expressed as:

$$\widehat{SE}_J = \sqrt{\frac{(n-1)}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^2}$$

where $\hat{\theta}_{(.)}$represents the average of the jackknife estimates:

$$\hat{\theta}_{(.)} = \frac{\sum_{i=1}^{n}\hat{\theta}_{(i)}}{n}$$

This calculation shows how the jackknife technique helps estimate the precision of the estimator $\hat{\theta}$ by assessing how the removal of each data point influences the overall statistical estimate.

Standard error estimation in Deleted d-jackkniefs:
In the delete-d jackknife method, rather than excluding one observation at a time, d observations are omitted from the dataset. As a result, each delete-d jackknife sample consists of n-d observations, and the total number of these reduced-size samples corresponds to the number of combinations of n items taken d at a time, denoted as $\binom{n}{d}$.
Let $\hat{\theta}_{(s)}$represent the application of the estimator $\hat{\theta}$ to the dataset after removing subset s. The formula to estimate the standard error using this method is

$$\sqrt{\frac{(n-d)}{d\binom{n}{d}}\sum(\hat{\theta}_{(s)} - \hat{\theta}_{(.)})^2}$$

Where $\hat{\theta}_{(.)} = \frac{\sum \hat{\theta}_{(s)}}{\binom{n}{d}}$ ,denotes the average of the estimators computed across all subsets s that are of size $n$-d each formed by excluding d observations from the full dataset, and the sum is taken over all these subsets s.This methodology aggregates the variations in the estimator that result from each different configuration of the dataset with d observations omitted.

## 3.4 CONFIDENCE INTERVAL ESTIMATION

The Jackknife method is a robust resampling technique often used in statistics to estimate the variability and confidence intervals (CIs) of an estimator, especially when the underlying distribution of the dataset is unknown or when the dataset size is small. This method effectively assesses the stability and reliability of estimators by examining the impact of systematically excluding data points from the calculation.

We know that the jackknife process begins by generating multiple subsets of the original dataset, each time leaving out one observation. This results in n new datasets if the original dataset consists of n observations. These subsets are utilized to compute the estimator (e.g., mean, median) minus one observation at a time, which allows for the analysis of each observation's influence on the overall estimate.
Central to the jackknife method is the computation of pseudo-values. For each subset, a pseudo-value is calculated using:

$$S_i' = nS - (n-1)S_i$$

where S is the estimator using the full dataset, and $S_i$ is the estimator recalculated with the $i^{th}$ observation omitted. The pseudo-value $S_i'$ amplifies the effect of removing a single observation, thus mimicking the influence of each data point on the estimator's variability.
The mean of these pseudo-values, denoted as $\bar{S}'$ is computed as follows:

$$\bar{S}' = \frac{1}{N}\sum_{i=1}^{n} Si'$$

This mean acts as a robust estimate of the estimator's central tendency over the resampled datasets.

The standard error (SE) is a crucial component in constructing confidence intervals. It is calculated from the pseudo-values' standard deviation, s as follows: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Si' - \bar{S}')^2}$This standard deviation measures the spread of the pseudo-values around their mean, reflecting

the estimator's variability due to sampling. The jackknife CI at the $(1-\alpha)100\%$ level is then calculated using: $CI_J(1-\alpha) = \bar{S}' \pm t_{\alpha/2,n-1} \times \left(\frac{S}{\sqrt{n}}\right)$

Here, $t_{\alpha/2,-1}$ is the critical value from the t-distribution for $n$-1degrees of freedom at the desired confidence level. The use of the t-distribution is particularly important for small samples, providing a better estimate of variability where the normal distribution might not be sufficient.

The jackknife method's ability to transform the estimation of any population parameter into estimating a population mean through pseudo-values makes it a powerful tool in statistical analysis. By understanding each observation's impact and the overall estimator's sensitivity, the jackknife provides robust estimates of confidence intervals, making it particularly useful in cases with small sample sizes or non-standard data distributions.

# CHAPTER.4
# REGRESSION ANALYSIS USING RESAMPLING TECHNIQUES(BOOTSTRAP&JACKKNIFE)

Regression analysis is a statistical tool used to explore the relationship between a dependent variable and one or more independent variables. The primary objective is to develop a mathematical model that describes this relationship in a way that can be used for prediction and estimation. Particularly, the estimation of coefficients (represented as $\hat{\beta}$) and the assessment of their uncertainty—including variance, bias, and constructing confidence intervals—are crucial elements in regression analysis. Consider a standard linear regression model expressed as Y= X$\beta$ + $\varepsilon$, where Y represents the response vector, X denotes the matrix of regressors, and $\varepsilon$ is a vector of error terms with a mean of zero and constant variance $\sigma^2$.

Then the least squares estimator $\hat{\beta} = (X'X)^{-1}X'Y$ has variance-covariance matrix Var($\hat{\beta}$) = $\sigma^2$ $(X'X)^{-1}$and 100(1-α) % confidence intervals

$\hat{\beta}_j \pm t_{(\alpha/2,(n-p))}$SE($\hat{\beta}_j$) . Where SE($\hat{\beta}_j$) is the standard error of the coefficient $\hat{\beta}_j$ calculated as the square root of the diagonal elements of the variance-covariance matrix $\sigma^2$ $(X'X)^{-1}$ and $t_{(\alpha/2,(n-p))}$ is the critical value from the t-distribution with $n$-p degrees of freedom, where n is the number of observations and p is the number of parameters (including intercept),$\alpha$ is the significance level.

Traditional regression techniques, such as Ordinary Least Squares (OLS), heavily depend on assumptions like linearity, normality of errors, and homoscedasticity. These assumptions underpin the validity of the model estimates but are not always tenable in practical scenarios, potentially affecting the reliability of the statistical analysis. Various diagnostic and robust estimation methods have been developed to address and mitigate the effects of assumption violations.

Unlike conventional methods that require strict assumptions about the distribution of estimators, resampling techniques offer a flexible approach by treating the observed dataset as a representative mini-population. Techniques like the bootstrap and jackknife allow for the estimation of standard errors and confidence intervals for regression coefficients without relying on parametric assumptions. These methods are particularly valuable in situations where traditional parametric inference is challenging or the distributional assumptions are questionable.

Both the bootstrap and jackknife are nonparametric resampling methods that provide mechanisms to derive robust estimates of standard errors and confidence intervals. The bootstrap involves repeatedly resampling with replacement from the original dataset and recalculating estimates across these samples to form an empirical distribution of the estimator. This distribution can then be used to approximate the variability and bias of the estimator.

On the other hand, the jackknife method systematically removes one or more observations from the dataset to assess the influence of individual data points on the estimation. By recalculating the estimates for each reduced dataset, the jackknife can provide insights into the stability and reliability of the parameter estimates.

The integration of bootstrap and jackknife techniques into regression analysis represents a significant advancement in statistical methodologies, allowing for more reliable and accurate estimation under a broader range of conditions. By addressing the limitations of traditional regression methods, these resampling techniques enhance the capability of statistical analysis to generate meaningful insights, even when the typical assumptions of parametric models are not met.

## 4.1 BOOTSTRAP APPROACH:

Bootstrap method involve drawing repeated samples from the observed data and recalculating statistics or models across these samples to assess variability, bias, or other characteristics of the statistical measures.

Consider a dataset with n observations, where each observation $w_i$ consists of a response variable $Y_i$ and a vector of k predictors $X_{ji}$. This can be compactly represented as: $w_i=(Y_i,X_{ji})'$for $i=1,2,...,n$ and $j=1,2,...,k$. The entire dataset can thus be described as consisting of responses $Y=(y_1,y_2,...,y_n)'$and predictors arranged in a $n\times k$ matrix $X=(x_{j1},x_{j2},...,x_{jn})'$.

Bootstrap Methods in Regression:

In regression analysis, the bootstrap can be applied in two distinct ways, depending on whether the regressors are considered fixed or random:

i. Bootstrap Based on Resampling Observations(Random Regressors):

It is more common in observational studies where both the response variables and predictors are considered to have been randomly sampled from broader populations. Consider the (k+1)x1 vector $w_i=(y_i,x_{ji}')'$that represents the values linked to the $i^{th}$ observation, where the collection of observations includes vectors $w_1,w_2,\ldots,w_n$ .The bootstrap technique based on resampling observations proceeds as follows:

- Generate a bootstrap sample $(w_1^{(b)},w_2^{(b)},\ldots,w_n^{(b)})$of size n by sampling from the observations with replacement, each $w_i$ having an equal probability of $1/n$. For each vector, label its components as $w_i^{(b)}=(y_i^{(b)},x_{ji}^{(b)})'$,where $j=1,2,\ldots,k$ and $i=1,2,\ldots,n$. Form the response vector $Y^{(b)}=(y_1^{(b)},y_2^{(b)},\ldots,y_n^{(b)})'$and the matrix $X_{ji}^{(b)}=(x_{j1}^{(b)},x_{j2}^{(b)},\ldots,x_{jn}^{(b)})'$ from these components.

- Compute the ordinary least squares (OLS) coefficients from the bootstrap sample using: $\hat{\beta}^{(b)}=(X^{(b)\prime}X^{(b)})^{-1}X^{(b)\prime}Y^{(b)}$
- Repeat the above steps for $r=1,2,\ldots,B$ where $B$ is the number of bootstrap repetitions.
- Assess the probability distribution $F(\hat{\beta}^{(b)})$ of the bootstrap estimates $\hat{\beta}^{(1)}$, $\hat{\beta}^{(2)}$,… $\hat{\beta}^{(B)}$. Utilize this distribution to calculate regression coefficients, variances, and confidence intervals. The mean of the bootstrap estimate of the regression coefficient is calculated as: $\hat{\beta}^{(b)}=\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}^{(b)}$
- The resultant bootstrap regression equation is expressed as: $Y^{(b)}=X\hat{\beta}^{(b)}+\varepsilon$ where $\hat{\beta}^{(b)}$is an unbiased estimator of $\beta$.

ii. Bootstrap Based on Resampling Errors(Fixed Regressors):

This approach is used when the predictor variables $X$are considered fixed—typical in experimental designs or when the data come from a controlled study where X does not vary.The procedures are given bellow

- Fit the full sample to a least squares regression to get $\hat{Y}_i$
- Calculate residuals $e_i=Y_i-\hat{Y}_i$
- Draw a bootstrap sample $e_1^{(b)},e_2^{(b)},\ldots,e_n^{(b)}$ from these residuals with probability 1/n each.
- Generate bootstrap responses by adding the resampled residuals back to the fitted values $Y^{(b)}=X\hat{\beta}+e^{(b)}$ .
- Obtain OLS estimates from each bootstrap sample using, $\hat{\beta}^{(b)}=(X'X)^{-1}X'Y^{(b)}$.
- Repeat the above steps for $r=1,2,\ldots,B$ , and compute the bias, variance, and confidence intervals from these bootstrap replications.

The bootstrap bias, variance, confidence and percentile interval

Bias and variance of bootstrap estimates are calculated as follows:

- Bias: bias($\hat{\beta}^{(b)}$)=$\hat{\beta}^{(b)} - \hat{\beta}$
- Variance: var($\hat{\beta}^{(b)}$)=$\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}^{(b)} - \hat{\beta}\right)^2$
- Confidence intervals are estimated using the normal approximation for large samples, or a t-distribution for smaller samples, as well as a nonparametric approach known as the percentile method, which uses quantiles from the bootstrap distribution. The confidence interval using the normal approach with the bootstrap method is calculated as:

$\hat{\beta}^{(b)} \pm t_{(\alpha/2,(n-p))}\text{SE}(\hat{\beta}^{(b)})$

where $t_{(\alpha/2,(n-p))}$ is the critical t-value for $\frac{\alpha}{2}$ probability and (n-p) degrees of freedom. The term $\text{SE}(\hat{\beta}^{(b)})$ represents the standard error of the bootstrap estimate of the regression coefficient. If the sample size n is 30 or larger, Z-distribution values are used instead of t-values for estimating the confidence intervals.

Alternatively, a nonparametric confidence interval, known as the percentile interval, can be constructed from the bootstrap sampling distribution of $\hat{\beta}^{(b)}$. This interval is defined by the quantiles:

$\hat{\beta}^{(lower)} < \beta < \hat{\beta}^{(upper)}$

where $\hat{\beta}^{(lower)}$ and $\hat{\beta}^{(upper)}$ are the bootstrap estimates at positions $\alpha/2 \times B$ and (1-$\alpha$/2) $\times B$ in the ordered list of bootstrap estimates, respectively. These quantile indices provide the bounds of the confidence interval directly from the empirical distribution of the bootstrap estimates.

## 4.2 JACKKNIEF APPROACH

The Jackknifing algorithm in regression models is typically used when the predictors in the models are fixed. This method involves two types of jackknife resampling: the delete-one jackknife, which removes one case at a time, and the delete-d jackknife, which removes multiple cases sequentially. These approaches are based on seminal works by authors such as Efron and Gong (1983), Wu (1986), and Shao and Tu (1995). Below, the delete-one jackknife approach for regression analysis is outlined:

i. Setup: Initially, a sample of size n is randomly drawn from the population, and the data for each observation is represented by the vector $w_i = (y_i, x_{ji})'$ for i=1,2,...,n , where j ranges from 1 to k. This vector forms the vector $Y = (y_1, y_2, ..., y_n)'$ and the matrix $X = (x_{j1}, x_{j2}, ..., x_{jn})'$ .

ii. . Delete-One Jackknife Sampling: For each observation i , the $i^{th}$ row $w_i$ is removed from the dataset. The remaining dataset forms the new sample $w_i^{(J)}$, which consists of $Y^{(J)} = (y_1, ..., y_{(i-1)}, y_{(i+1)}, ..., y_n)'$ and $X^{(J)} = x_{j1}, ..., x_{j(i-1)}, x_{j(i+1)}, ..., x_{jn})'$ . The ordinary least squares (OLS) regression coefficients $\hat{\beta}_i^{(J)}$ are estimated using this reduced dataset.

iii. Estimation of Jackknife Estimates: The jackknife estimates $\hat{\beta}_i^{(J)}$ are calculated for each dataset obtained by omitting each observation one at a time. This forms the set $\hat{\beta}_1^{(J)}, \hat{\beta}_2^{(J)}, ..., \hat{\beta}_n^{(J)}$.

iv. Average Jackknife Estimate: The overall jackknife estimate of the regression coefficient is computed as the mean of all jackknife estimates:

$\hat{\beta}^{(J)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\beta}_i^{(J)}$

This mean serves as the jackknife estimate for the regression coefficient.

v. . Jackknife Regression Equation: The final regression equation using the delete-one jackknife estimate is:

$Y = X\hat{\beta}^{(J)} + \varepsilon$

where $\hat{\beta}^{(J)}$ is the jackknife estimate of the regression coefficient.

This jackknife procedure is a robust method for assessing the stability and reliability of regression estimates, particularly useful in scenarios where the assumptions of more complex resampling methods like the bootstrap may not hold.

The Algorithms for Delete-d Jackknife Regression

The delete-d jackknife method in regression analysis involves systematic removal of multiple observations from the dataset to estimate the variability and bias of regression coefficients. Here is an outlined step-by-step procedure for the delete-d jackknife regression:

i. Sample Division : Start by drawing a sample of size n from the population, consisting of observations $w_1, w_2, \ldots, w_n$ . Divide this sample into s independent groups, each containing d observations.
ii. First Deletion : Remove the first set of d observations from the full sample. Use the remaining ( n-d ) observations to estimate the Ordinary Least Squares (OLS) coefficients, denoted as $\hat{\beta}_1^{(J)}$ , from this reduced sample, referred to as the delete-d jackknife sample.
iii. Sequential Deletions : Continue by sequentially omitting each subsequent group of d observations and estimate the

OLS coefficients, $\hat{\beta}_2^{(J)}$, $\hat{\beta}_3^{(J)}$, …, for each remaining set of n-d observations.

iv. Estimation for All Groups : Repeat the deletion and estimation process for each of the s groups, thereby computing a set of jackknife regression coefficients $\hat{\beta}_k^{(J)}$ for each deletion, where k = 1, 2,…, s .

v. Distribution of Estimates : Compile the jackknife estimates $\hat{\beta}_1^{(J)}, \hat{\beta}_2^{(J)}, \ldots., \hat{\beta}_s^{(J)}$ to assess the probability distribution F($\hat{\beta}^{(J)}$) of these estimates.

vi. Average Coefficient Estimate : Calculate the average of these jackknife estimates to obtain the overall jackknife regression coefficient, given by:

$$\hat{\beta}^{(J)} = \frac{1}{s}\sum_{k=1}^{s} \hat{\beta}_k^{(J)}$$

This mean serves as the jackknife estimate for the regression coefficients.

vii. Jackknife Regression Equation : Formulate the final regression equation using the jackknife estimated coefficients:

$$Y = X\hat{\beta}^{(J)} + \varepsilon$$

where $\hat{\beta}^{(J)}$ represents the average jackknife estimate of the regression coefficients.

This delete-d jackknife approach provides a robust mechanism for estimating regression coefficients, particularly useful for assessing the influence of groups of data points on the regression model and for estimating the variability and bias of the regression estimates.

<u>Jackknife bias, variance, confidence and percentile interval</u>

In statistical analysis using the jackknife resampling technique, the bias, variance, confidence intervals, and percentile intervals of the regression coefficients are estimated from the distribution of

the jackknife estimates $\hat{\beta}^{(J)}$. Below, the formulas and methods for calculating these metrics are delineated:

Jackknife Bias:

The bias in the jackknife estimator is calculated by comparing the jackknife estimate to the original estimate, adjusted by a factor related to the number of observations, given by:

$$\text{bias}(\hat{\beta}^{(J)}) = (n - 1)(\hat{\beta}^{(J)} - \hat{\beta})$$

Here, n is the number of observations, $\hat{\beta}$ is the estimate from the full dataset, and $\hat{\beta}^{(J)}$ is the mean of the jackknife estimates.

Jackknife Variance:

The variance of the jackknife estimates is computed as:

$$\text{Var}(\hat{\beta}^{(J)}) = \frac{(n-1)}{n}\sum_{i=1}^{n}(\hat{\beta}_i^{(J)} - \hat{\beta}^{(J)})^2$$

where $\hat{\beta}_i^{(J)}$ is the estimate obtained from the dataset with the $i^{th}$ observation (or group) removed.

Jackknife Confidence Interval:

The confidence interval for the jackknife estimates is constructed using the standard error and the critical t-values:

$$\hat{\beta}^{(J)} \pm t_{(\alpha/2,(n-p))}\text{SE}(\hat{\beta}^{(J)})$$

where $t_{(\alpha/2,(n-p))}$ is the critical value of the t-distribution with n-p degrees of freedom at $\alpha/2$, and $\text{SE}(\hat{\beta}^{(J)})$ is the standard error of $\hat{\beta}^{(J)}$.

Jackknife Percentile Interval:

The percentile interval, a nonparametric confidence interval, is based on the quantiles of the jackknife estimates. It is defined as:

$$\hat{\beta}^{(J\ lower)} < \beta < \hat{\beta}^{(J\ upper)}$$

where $\hat{\beta}^{(J\ lower)}$ and $\hat{\beta}^{(J\ upper)}$ are the jackknife estimates at positions $\frac{\alpha}{2} \times n$ and $(1\text{-}\alpha/2) \times n$ in the ordered list of jackknife estimates.

These methods enable robust estimation of regression coefficients by accounting for potential biases and variability introduced by influential observations. Each metric provides a different insight into the stability and reliability of the estimator under the jackknife resampling scheme.

# CHAPTER:5

# APPLICATIONS

## 5.1 DEPENDENT DATA SET

### Bootstrap In Dependent Dataset

Bootstrap methods offer a powerful approach for estimating parameters and making inferences from data, particularly when dealing with dependent or autocorrelated datasets where traditional statistical methods may falter. Specialized techniques, such as the Block Bootstrap and the Moving Block Bootstrap, are tailored to address the challenges posed by dependence structures inherent in such datasets.

The Block Bootstrap divides the data into contiguous blocks or segments, preserving the temporal or spatial dependence structure present in the original data. This method ensures that resampling is conducted at the block level, maintaining the integrity of the dependence relationships within the dataset. Careful selection of block lengths is crucial to accurately capture the underlying dependence.

Similarly, the Moving Block Bootstrap employs a moving window to create overlapping blocks instead of fixed blocks. This adaptive approach improves the capture of short-term dependencies in the data, enhancing the accuracy of resampling.

By utilizing these specialized techniques, bootstrap methods can effectively handle dependence structures in datasets, ensuring that resampled data maintain essential characteristics such as autocorrelation or spatial dependency. This, in turn, leads to more reliable estimation and inference, making bootstrap methods invaluable tools for analyzing dependent data sets.

- Time Series Analysis:

  In time series analysis, observations are often correlated with their past values. Bootstrap methods can effectively handle this dependence by resampling blocks of data, such as consecutive time

points, while preserving the autocorrelation structure.

For example, in financial markets, where stock prices are typically autocorrelated, bootstrap resampling can be used to estimate parameters like volatility, correlation coefficients, and Value at Risk (VaR) with better accuracy than traditional methods.

- Spatial Data Analysis:

  Spatial data, such as geographical data or data collected from sensor networks, often exhibit spatial autocorrelation, where nearby observations are more similar to each other than those farther apart. Bootstrap methods can account for this spatial dependence by resampling spatial blocks or neighborhoods of data points.

  For instance, in environmental studies, bootstrap techniques can be used to estimate parameters like soil contamination levels or air pollutant concentrations across geographic regions while accounting for spatial autocorrelation.

- Longitudinal Studies:

  In longitudinal or panel data analysis, observations are collected from the same individuals or entities over time, leading to dependence between observations within the same subject. Bootstrap resampling can address this dependence by resampling entire subjects or clusters of observations.

  For example, in medical research involving longitudinal clinical trials, bootstrap methods can be employed to assess treatment effects or predict outcomes while properly accounting for the correlation between repeated measures from the same patients.

- Spatial and Temporal Interpolation:

  In geostatistics or climate modeling, where data is collected at irregularly spaced locations or time intervals, bootstrap

techniques can be used for spatial or temporal interpolation. By resampling observed data points and estimating missing values, bootstrap methods can generate spatially or temporally continuous fields while preserving the underlying correlation structure.

In summary, bootstrap methods offer a flexible and robust approach to handling dependent data sets across various fields, allowing researchers to obtain reliable estimates and make valid inferences even in the presence of complex correlation structures.

## Jackknief In Dependent Dataset

The jackknife method, although originally developed for independent data sets, can also be adapted and applied to dependent data sets, albeit with some modifications.

In specific types of dependent data sets, such as time series, spatial data, and longitudinal data, the jackknife method can be tailored to account for their unique characteristics:

- Time Series Data:

  For time series data, where observations are ordered chronologically and exhibit autocorrelation, the jackknife can be adapted to leave out contiguous blocks of data rather than individual observations. This approach preserves the temporal dependence structure.

  _Modified jackknife techniques, like the moving block jackknife, systematically remove consecutive segments of the time series to assess parameter estimates, model performance, or bias correction.

- Spatial Data:

  Spatial data often exhibit spatial autocorrelation, where nearby observations are correlated. In this case, the jackknife can be adjusted to leave out spatial blocks or clusters of observations instead of individual data points.

Techniques such as spatial leave-one-out cross-validation or the spatial block jackknife remove spatially contiguous groups of observations to evaluate model performance, stability, or bias correction.

- Longitudinal Data:

  Longitudinal data involve repeated measurements on the same subjects over time, leading to within-subject correlation. The jackknife can be modified to leave out entire subjects or clusters of observations at a time while retaining the temporal dependence structure.

  Longitudinal leave-one-out cross-validation or the longitudinal block jackknife systematically remove subjects or clusters of observations to assess parameter estimates, model performance, or bias correction in longitudinal analyses.

  In each of these cases, adapting the jackknife method to the specific dependence structure of the data allows for more accurate inference, validation, and model selection. These modified jackknife techniques ensure that the inherent dependencies within the data are appropriately accounted for, leading to more reliable results in time series, spatial, and longitudinal analyses.

## 5.2.APPLICATION IN SAMPLE SURVEYS

**Bootstrap in Sample Surveys**

- Estimation of Population Parameters:

  In sample surveys, the goal is to estimate population parameters (e.g., mean, total, proportion) based on data collected from a sample of the population. Bootstrap resampling can be employed to generate multiple bootstrap samples from the original survey

sample, allowing for the estimation of the sampling distribution of the parameter of interest.

By calculating the parameter estimate for each bootstrap sample, confidence intervals can be constructed, providing a measure of uncertainty around the point estimate. Bootstrap confidence intervals are particularly useful when the population distribution is unknown or when traditional asymptotic methods are not applicable due to small sample sizes or complex survey designs.

- Variance Estimation:

Bootstrap methods offer a robust approach to estimating the variance of survey estimators. Traditional variance estimation techniques, such as the Taylor series linearization or the jackknife, rely on assumptions like linearity and finite population correction, which may not hold in all survey settings.

Bootstrap variance estimation involves resampling with replacement from the survey data to create multiple bootstrap samples. Variance estimates are then calculated for each bootstrap sample, and the variability across these estimates provides an approximation of the variance of the survey estimator.

- Complex Survey Designs:

Sample surveys often employ complex sampling designs, such as stratification, clustering, and unequal probabilities of selection, to improve efficiency and reduce costs. Bootstrap methods can accommodate these complexities by resampling within defined sampling units (e.g., strata, clusters) while preserving the original survey design.

Bootstrap techniques can be adapted to account for stratification, clustering, and unequal probabilities of selection, ensuring that the resampled data reflect the survey's complex structure. This

enables more accurate variance estimation and inference under complex survey designs.

## Jackknief In Sample Surveys

- Estimation of Population Parameters:By systematically excluding each observation, the jackknife method provides robust estimates of population parameters, assessing the influence of individual data points.

- Variance EstimationThe jackknife method estimates the variance of survey estimators without relying on traditional assumptions, making it suitable for complex survey settings.
- Complex Survey Designs: The method adapts to complex sampling designs, including stratification, clustering, and unequal probabilities of selection, ensuring accurate variance estimation and inference.

  That is the jackknife technique is a versatile tool in sample surveys, offering robust estimation and variance calculation capabilities, particularly in the context of complex survey designs. This ensures that survey results are reliable and reflective of the true variability in the population, leading to more accurate and meaningful conclusions.

## 5.3.NONLINEAR MODELS

Bootstrap and jackknife methods can be applied in the context of nonlinear models to assess parameter estimation accuracy, model performance, and prediction uncertainty. Here's a detailed explanation of their applications:

## Bootstrap in Nonlinear Models

- Parameter Estimation Uncertainty:

  Nonlinear models often involve estimating parameters from observed data, and the accuracy of these estimates is crucial for model interpretation and inference. Bootstrap resampling can be

used to quantify parameter estimation uncertainty by generating bootstrap samples from the original dataset.

For each bootstrap sample, the nonlinear model is fitted, and parameter estimates are obtained. By repeating this process multiple times, a distribution of parameter estimates is obtained, allowing for the calculation of confidence intervals and hypothesis tests for model parameters.

- Model Validation and Performance:

  Bootstrap methods are valuable for assessing the performance and predictive accuracy of nonlinear models. Resampling from the original dataset allows for the generation of bootstrap samples, which can be used for model validation by comparing predicted outcomes to observed outcomes.

- Confidence Intervals for Predictions:

  In nonlinear regression or curve fitting, researchers often need to make predictions for new data points and quantify prediction uncertainty. Bootstrap methods can be employed to construct confidence intervals for predicted outcomes by resampling from the original dataset and fitting the nonlinear model to each bootstrap sample.

  Prediction intervals can be derived from the distribution of predicted outcomes across bootstrap samples, providing a measure of uncertainty around the predictions made by the nonlinear model. This allows researchers to make informed decisions based on the reliability of model predictions.

  Bootstrap and jackknife methods find extensive applications in nonparametric models, which are flexible statistical techniques that do not rely on assumptions about the underlying data distribution. Here's a detailed explanation of their applications in nonparametric modeling:

**Jackknife in Nonlinear Models:**

- Parameter Estimation and Bias Correction:

  The jackknife method can be used to assess parameter estimation accuracy and correct for bias in nonlinear models. By systematically leaving out one observation at a time and re-estimating the model parameters, researchers can obtain jackknife estimates of model parameters.

  Jackknife estimates provide insights into the sensitivity of parameter estimates to individual observations and can help identify influential data points that disproportionately impact the parameter estimates. Jackknife bias correction techniques can be applied to improve the accuracy of parameter estimates by adjusting for bias introduced by the dependence between model parameters and the entire dataset.

- Model Stability Assessment:

  The jackknife is valuable for assessing the stability of parameter estimates and model selection procedures in nonlinear models. By iteratively leaving out one observation at a time and re-fitting the model, researchers can evaluate the robustness of parameter estimates to individual observations.

  Jackknife-based stability metrics can help identify robust model specifications or features that are consistently selected across different subsets of the data, aiding in more stable and reliable model inference and selection in nonlinear settings.

- Leave-One-Out Cross-Validation (LOOCV):

  The jackknife is commonly used in leave-one-out cross-validation (LOOCV) for model validation and performance evaluation in nonlinear models. By systematically leaving out one observation at a time and re-fitting the model, researchers can assess the

predictive accuracy and generalization performance of the nonlinear model.

LOOCV provides an unbiased estimate of prediction error and helps identify overfitting or underfitting issues in nonlinear models. Jackknife estimates of prediction error are less biased and more efficient compared to traditional k-fold cross-validation methods, particularly in small-sample settings.

That is the two methods offer versatile tools for assessing parameter estimation uncertainty, model performance, and prediction uncertainty in nonlinear models. These resampling techniques enhance the reliability and validity of inference and prediction in nonlinear regression, curve fitting, and other nonlinear modeling applications.

## 5.4.NONPARAMETRIC MODELS

### Bootstrap in Nonparametric Models

- Estimation of Distribution Functions:

  Nonparametric models, such as kernel density estimation or empirical distribution functions, are often used to estimate the underlying distribution of data. Bootstrap resampling can be employed to quantify uncertainty in the estimated distribution by generating bootstrap samples from the observed data.

  For each bootstrap sample, the nonparametric model is fitted, and the distribution function or density estimate is computed. By repeating this process multiple times, confidence intervals can be constructed for the estimated distribution, providing insights into the variability and uncertainty of the nonparametric model.

- Confidence Intervals for Quantiles:

  Nonparametric models are frequently used to estimate quantiles or percentiles of a distribution. Bootstrap methods can be applied

to construct confidence intervals for quantiles by resampling from the original dataset and computing quantile estimates for each bootstrap sample.

Quantile intervals derived from the distribution of bootstrap quantile estimates provide a measure of uncertainty around the estimated quantiles, allowing researchers to make reliable inferences about population percentiles based on the nonparametric model.

- Regression and Smoothing Techniques:

  Nonparametric regression and smoothing techniques, such as local regression (LOESS) or spline smoothing, are used to model the relationship between variables without assuming a specific functional form. Bootstrap resampling can aid in assessing the uncertainty of nonparametric regression estimates and smoothing parameters.

  Bootstrap methods allow researchers to generate bootstrap samples, fit nonparametric regression models to each sample, and examine the variability in regression estimates across samples. This facilitates the construction of confidence intervals for regression curves and prediction intervals for individual observations.

**Jackknife in Nonparametric Models**

- Leave-One-Out Estimation:

  The jackknife method is widely used in nonparametric modeling for leave-one-out estimation, where one observation is systematically left out at a time, and the model is fitted to the remaining data. Jackknife estimates of model parameters or predictions are obtained by averaging over the leave-one-out estimates.

Leave-one-out jackknife estimates provide insights into the stability and bias of nonparametric models, helping researchers identify influential data points and assess the robustness of model inference.

- Bias Correction and Validation:

  Jackknife methods can be applied to estimate and correct for bias in nonparametric models. By comparing leave-one-out estimates with estimates obtained from the full dataset, researchers can assess the bias introduced by individual observations and adjust model parameters or predictions accordingly.

  Jackknife bias correction techniques help improve the accuracy and reliability of nonparametric model estimates by accounting for bias introduced by the dependence between model parameters and the entire dataset.

- Model Selection and Validation:

  The jackknife is valuable for model selection and validation in nonparametric modeling. By systematically leaving out one observation at a time and re-fitting the model, researchers can assess the stability and generalization performance of nonparametric models.

  Jackknife-based model selection procedures, such as leave-one-out cross-validation (LOOCV), help identify the optimal model complexity or smoothing parameters by minimizing prediction error. Jackknife estimates of prediction error provide unbiased assessments of model performance and aid in selecting the most appropriate nonparametric model.

  Both bootstrap and jackknife methods offer versatile tools for assessing uncertainty, bias, and model performance in nonparametric models. These resampling techniques enhance the reliability and validity of inference, prediction, and model

selection in settings where parametric assumptions may not be appropriate.

## 5.5 BOOTSTRAP IN QUALITY CONTROL:

In quality control, bootstrap methods serve various purposes, aiding in the assessment of process variability, determination of confidence intervals for key quality metrics, and evaluation of sampling strategies. Here's how bootstrap is applied in quality control:

- Estimation of Process Variability:

  In quality control, understanding process variability is crucial for ensuring consistent product quality. Bootstrap resampling can be used to estimate parameters like the process standard deviation or variance, especially when the underlying distribution of process data is unknown or non-normal.

  By resampling the observed data with replacement, bootstrap methods provide an empirical estimate of the sampling distribution of the process variability metrics, enabling more accurate characterization of process performance.

- Confidence Interval Estimation:

  Quality control practitioners often need to estimate confidence intervals for key quality metrics, such as the mean, median, or proportion of non-conforming units in a production process. Bootstrap techniques can be employed to construct confidence intervals for these metrics, even when the underlying distribution is unknown or skewed.

  Bootstrap confidence intervals are robust against distributional assumptions and can provide accurate coverage probabilities, making them particularly useful when traditional methods based on normality assumptions may be unreliable.

- Process Capability Analysis:

Assessing process capability involves evaluating whether a manufacturing process meets specified quality standards and can consistently produce products within tolerance limits. Bootstrap methods can aid in process capability analysis by estimating process capability indices.

Bootstrap resampling allows practitioners to generate multiple samples from the observed data, enabling the estimation of variability in process performance metrics and providing insights into the stability and capability of the production process.

- Sampling Plan Evaluation:

  In quality control, designing an effective sampling plan is essential for efficiently monitoring and controlling product quality. Bootstrap methods can be used to evaluate different sampling strategies by resampling subsets of data and assessing the variability in quality metrics under different sampling scenarios.

  By simulating the sampling process, bootstrap techniques help in optimizing sample sizes, sampling frequencies, and sampling locations to ensure adequate quality control while minimizing costs and resources.

- Nonparametric Control Charts:

  Traditional control charts, such as the Shewhart chart or the X-bar chart, assume normality and independence of process data. However, in practice, these assumptions may not hold. Bootstrap methods can be employed to construct nonparametric control charts that are robust to deviations from normality and independence.

  Bootstrap-based control charts, such as the bootstrap control chart or the wild bootstrap control chart, utilize resampling techniques to estimate control limits and monitor process stability effectively, even when data distribution is unknown or dependent.

By leveraging the flexibility and robustness of bootstrap resampling, quality control practitioners can enhance the effectiveness of process monitoring, improve decision-making, and ensure the consistent delivery of high-quality products to customers.

## 5.6 OTHER APPLICATIONS

i. Finance and Economics:

*Risk Management*: Estimating Value at Risk (VaR) and other risk measures for portfolios by generating numerous resamples of historical data.

*Economic Forecasting*: Constructing confidence intervals for economic indicators, such as GDP growth rates, to better understand their variability.

ii. Medical Research:

*Clinical Trials*: Estimating the reliability of treatment effects and survival rates by resampling patient data to construct confidence intervals for medical statistics.

*Genomics*: Assessing the stability of gene expression levels and identifying significant genes by resampling genomic data.

iii. Engineering:

*Reliability Engineering*: Estimating the lifetime of components and systems by resampling failure time data to determine the distribution of lifetimes.

*Quality Control*: Constructing control charts and process capability indices to monitor and improve manufacturing processes.

iv. Environmental Science:

*Climate Modeling*: Evaluating the uncertainty in climate model predictions by resampling historical climate data.

*Ecology*: Estimating species diversity and population parameters by resampling ecological survey data.

v. Marketing and Business Analytics:

*Customer Analytics*: Estimating customer lifetime value and retention rates by resampling customer data to predict future behaviors.

*A/B Testing*: Assessing the effectiveness of different marketing strategies by resampling test results to determine the robustness of observed differences.

vi. Biostatistics:

*Estimating Bias and Variance*: In clinical research, using the jackknife to estimate the bias and variance of estimators such as mean survival time.

*Phylogenetics*: In evolutionary biology, assessing the robustness of phylogenetic trees by repeatedly leaving out one species and recalculating the tree.

vii. Survey Methodology:

*Error Estimation*: Estimating standard errors for complex survey data, particularly when dealing with small sample sizes or non-standard sampling designs.

*Weight Adjustment*: Adjusting weights in survey analysis to improve estimates and reduce bias.

viii. Economics:

*Regression Analysis*: Estimating the stability of regression coefficients by systematically leaving out one observation at a time.

*Time Series Analysis*: Assessing the robustness of time series models by jackknifing individual time points.

ix. Machine Learning:

*Model Validation*: Using the jackknife to validate the stability and reliability of machine learning models by systematically leaving out individual data points or subsets of data.

*Feature Selection*: Evaluating the importance of features in predictive models by analyzing how the exclusion of each feature affects model performance.

x. Astronomy:

*Estimating Distances*: In astronomical surveys, using jackknife techniques to estimate distances to celestial objects and their uncertainties.

*Cosmology*: Assessing the robustness of cosmological parameters estimated from large-scale surveys by jackknifing subsets of the data.

xi. Sports Analytics:

Player Performance: Using bootstrap methods to estimate the variability in player performance metrics and to assess the impact of different factors on performance.

*Team Strategies*: Evaluating the effectiveness of team strategies by resampling game data to determine the robustness of observed strategies.

xii. Social Sciences:

*Behavioral Studies*: Estimating the variability and reliability of behavioral metrics by resampling survey or experimental data.

*Policy Analysis*: Assessing the impact of policy changes by bootstrapping historical policy outcome data to provide confidence intervals for predictions.

xiii. Education:

*Standardized Testing*: Estimating the reliability and validity of test scores by resampling test data to evaluate different metrics.

*Program Evaluation*: Assessing the effectiveness of educational programs by resampling participant data to construct confidence intervals around program impact estimates.

These examples illustrate the versatility and utility of bootstrap and jackknife methods across a wide array of fields, helping practitioners make more informed decisions by understanding the uncertainty and variability inherent in their data.

## CHAPTER 6:

# ANALYZING THE GAPMINDER DATASET USING BOOTSTRAP AND JACKKNIFE RESAMPLING TECHNIQUES

The 'gapminder' dataset provides valuable insights into global development trends by tracking various socio-economic indicators for countries over time. The dataset includes data on:

country: The country name.

year: The year of the observation.

lifeExp: Life expectancy at birth, in years.

gdpPercap: Gross Domestic Product per capita.

Here we try to analyze the relationship between life expectancy (lifeExp) and GDP per capita (gdpPercap) for the year 2007 using bootstrap and jackknife resampling techniques. We will focus on parameter estimation, bias, standard error, variance, confidence intervals, hypothesis testing, and linear regression.

For this analysis, we will filter the dataset to include only the data for the year 2007.

We know that bootstrap resampling involves repeatedly sampling with replacement from the dataset to estimate the distribution of a statistic. This method helps estimate bias, standard error, variance, and confidence intervals.

Jackknife resampling systematically leaves out one observation at a time from the dataset and calculates the statistic for each subset. This method is useful for estimating bias, standard error, and confidence intervals.

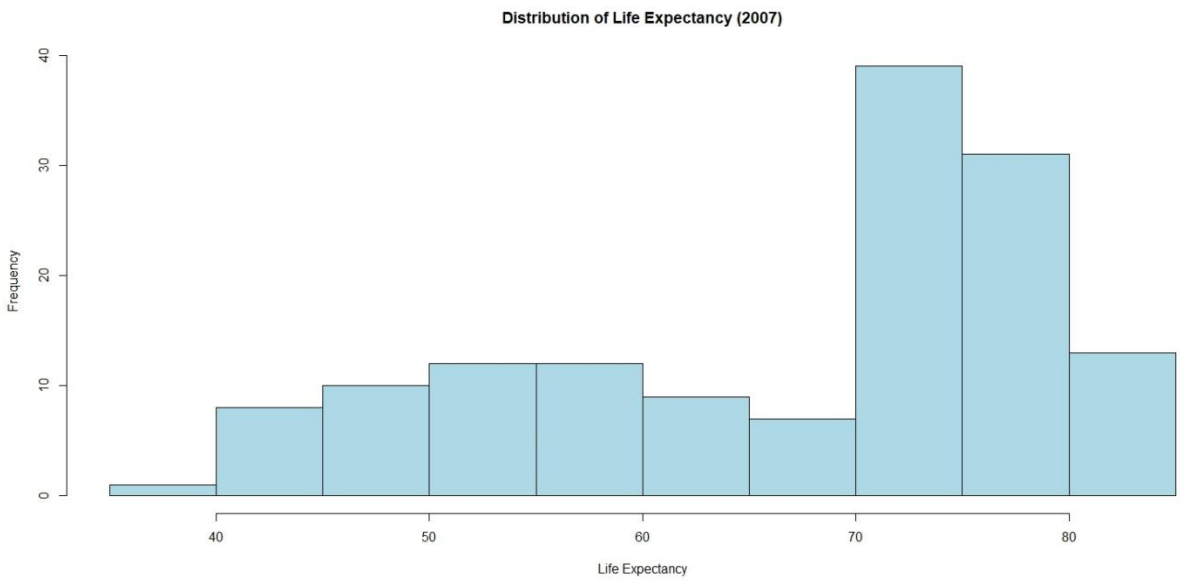
**Analysis:**

**Estimation of Mean Life Expectancy Using Bootstrap and Jackknife**

**Bootstrap**

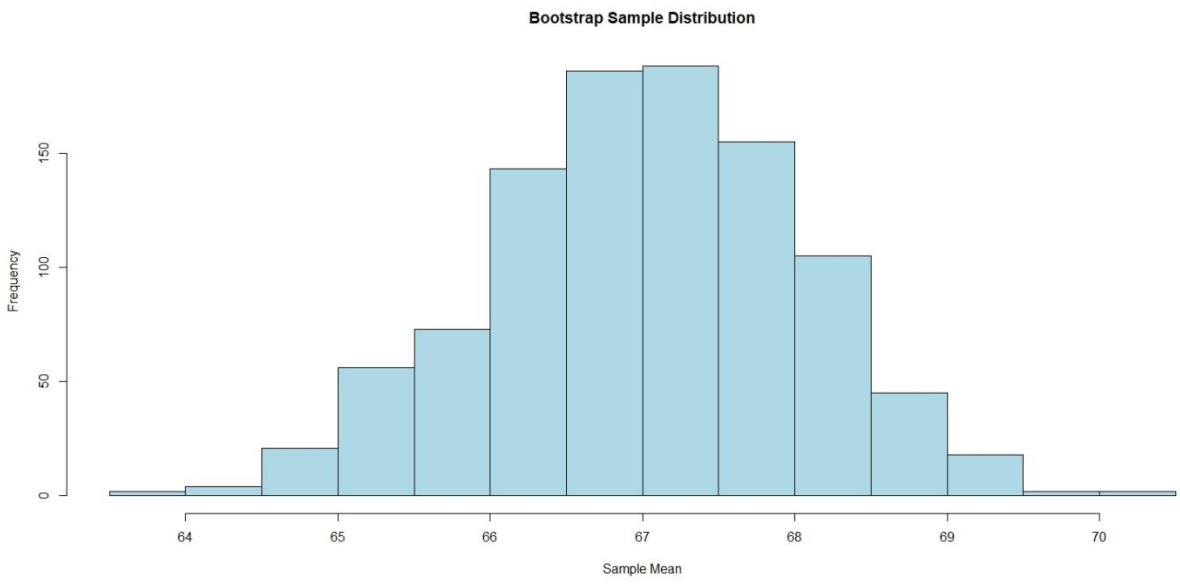
```
 # Load necessary packages
 install.packages("gapminder")
ibrary(gapminder)
# Load the gapminder dataset
 data(gapminder)
 head(gapminder)
# A tibble: 6 x 6
  country     continent  year lifeExp      pop gdpPercap
  <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
1 Afghanistan Asia       1952    28.8  8425333      779.
2 Afghanistan Asia       1957    30.3  9240934      821.
3 Afghanistan Asia       1962    32.0 10267083      853.
4 Afghanistan Asia       1967    34.0 11537966      836.
5 Afghanistan Asia       1972    36.1 13079460      740.
6 Afghanistan Asia       1977    38.4 14880372      786.
# Filter data for the year 2007
```

```
 gapminder_2007 <- subset(gapminder, year == 2007)

head(gapminder_2007)

# A tibble: 6 x 6

  country     continent  year lifeExp      pop gdpPercap

  <fct>       <fct>     <int>   <dbl>    <int>     <dbl>

1 Afghanistan Asia       2007    43.8 31889923      975.

2 Albania     Europe     2007    76.4  3600523     5937.

3 Algeria     Africa     2007    72.3 33333216     6223.

4 Angola      Africa     2007    42.7 12420476     4797.

5 Argentina   Americas   2007    75.3 40301927    12779.

6 Australia   Oceania    2007    81.2 20434176    34435.

 # Plot histogram for life expectancy

 hist(gapminder_2007$lifeExp, main = "Distribution of Life Expectancy
(2007)",

     xlab = "Life Expectancy", ylab = "Frequency", col = "lightblue")
```




```
# Set the number of bootstrap samples
n_bootstrap <- 1000
# Initialize a vector to store bootstrap sample means
bootstrap_means <- numeric(n_bootstrap)
# Perform bootstrap resampling
set.seed(123)  # For reproducibility
for (i in 1:n_bootstrap) {
  bootstrap_sample <-
gapminder_2007[sample(1:nrow(gapminder_2007), replace = TRUE), ]
  bootstrap_means[i] <- mean(bootstrap_sample$lifeExp)
}
# Calculate bootstrap estimates
bootstrap_mean_estimate <- mean(bootstrap_means)
```

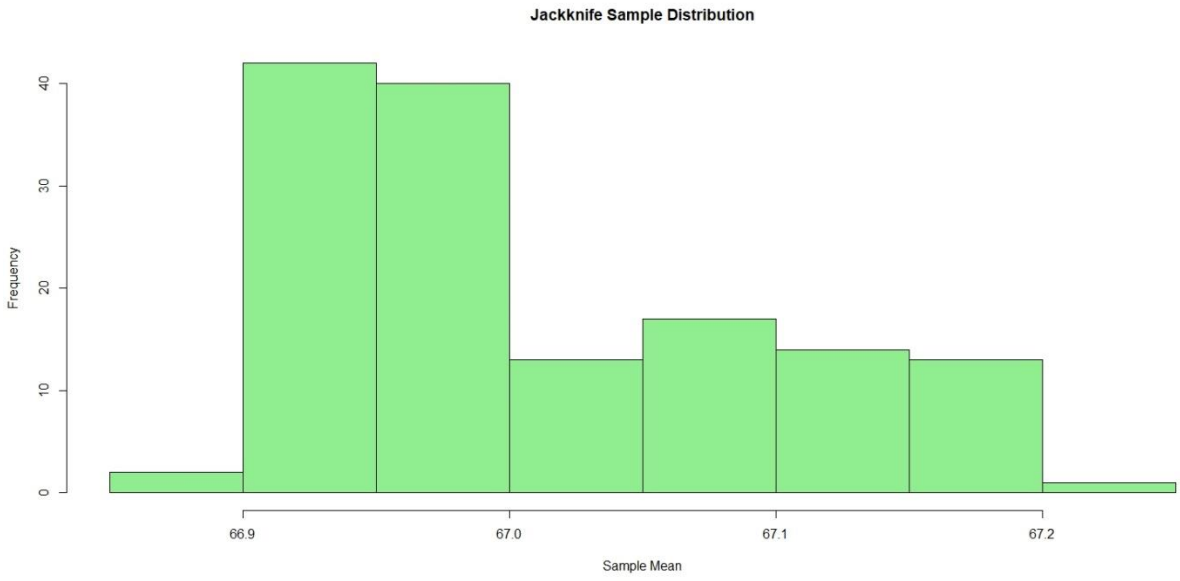
```
 bootstrap_bias <- bootstrap_mean_estimate -
mean(gapminder_2007$lifeExp)

 bootstrap_variance <- var(bootstrap_means)

 bootstrap_se <- sd(bootstrap_means)

 bootstrap_ci <- quantile(bootstrap_means, c(0.025, 0.975))

 # Output results

 bootstrap_mean_estimate

[1] 67.01898

 bootstrap_bias

[1] 0.01155942

bootstrap_variance

[1] 1.037006

 bootstrap_se

[1] 1.018335

 bootstrap_ci

    2.5%    97.5%

64.94169 68.93795

 # Plotting Bootstrap Distribution

 hist(bootstrap_means, main = "Bootstrap Sample Distribution", xlab =
"Sample Mean", ylab = "Frequency", col = "lightblue")
```




## Jackknief

```
# Initialize a vector to store jackknife sample means

n <- nrow(gapminder_2007)

jackknife_means <- numeric(n)

# Perform jackknife resampling

for (i in 1:n) {

jackknife_sample <- gapminder_2007[-i, ]

  jackknife_means[i] <- mean(jackknife_sample$lifeExp)

}

# Calculate jackknife estimates

jackknife_mean_estimate <- mean(jackknife_means)

jackknife_bias <- (n - 1) * (jackknife_mean_estimate -
mean(gapminder_2007$lifeExp))
```

```
jackknife_variance <- (n - 1) * mean((jackknife_means - mean(jackknife_means))^2)

jackknife_se <- sqrt(jackknife_variance / n)

alpha <- 0.05

t_value <- qt(1 - alpha/2, df = n - 1)

jackknife_ci <- c(jackknife_mean_estimate - t_value * jackknife_se,

                  jackknife_mean_estimate + t_value * jackknife_se)

# Output results

jackknife_mean_estimate

[1] 67.00742

jackknife_bias

[1] 0

jackknife_variance

[1] 1.026464

jackknife_se

[1] 0.08502127

jackknife_ci

[1] 66.83934 67.17550

# Plotting Jackknife Distribution

> hist(jackknife_means, main = "Jackknife Sample Distribution", xlab = "Sample Mean", ylab = "Frequency", col = "lightgreen")
```




To test the hypothesis that the mean life expectancy is different from a given value (e.g., 65 years), we can use the confidence intervals obtained from bootstrap and jackknife methods. If the given value falls outside the confidence interval, we reject the null hypothesis.

## Linear Regression Analysis Using Bootstrap and Jackknife

### Bootstrap in regression

```
# Set the number of bootstrap samples

n_bootstrap <- 1000

# Initialize matrices to store bootstrap sample coefficients

bootstrap_coefficients <- matrix(NA, nrow = n_bootstrap, ncol = 2)
```

```
 # Perform bootstrap resampling

 set.seed(123)  # For reproducibility

 for (i in 1:n_bootstrap) {

   bootstrap_sample <-
gapminder_2007[sample(1:nrow(gapminder_2007), replace = TRUE), ]

   lm_bootstrap <- lm(lifeExp ~ gdpPercap, data = bootstrap_sample)

   bootstrap_coefficients[i, ] <- coef(lm_bootstrap)

 }

 # Calculate bootstrap estimates and confidence intervals

 bootstrap_mean_coefficients <- colMeans(bootstrap_coefficients)

 bootstrap_ci <- apply(bootstrap_coefficients, 2, function(x) quantile(x,
c(0.025, 0.975)))

 # Output results

 bootstrap_mean_coefficients

[1] 5.953887e+01 6.409877e-04

# Formatting the values

 intercept <- format(bootstrap_mean_coefficients[1], scientific = FALSE)

slope <- format(bootstrap_mean_coefficients[2], scientific = FALSE)

intercept

[1] "59.53887"

 slope

[1] "0.0006409877"
```

```
 bootstrap_ci
          [,1]         [,2]
2.5%  57.30683 0.0005463233
97.5% 61.60131 0.0007520042
```

**Jackknief in regression**

```
# Initialize matrices to store jackknife sample coefficients
 jackknife_coefficients <- matrix(NA, nrow = n, ncol = 2)
 # Perform jackknife resampling
 for (i in 1:n) {
   jackknife_sample <- gapminder_2007[-i, ]
   lm_jackknife <- lm(lifeExp ~ gdpPercap, data = jackknife_sample)
   jackknife_coefficients[i, ] <- coef(lm_jackknife)
 }
 # Calculate jackknife estimates and standard errors
 jackknife_mean_coefficients <- colMeans(jackknife_coefficients)
 jackknife_se <- sqrt((n - 1) / n * colSums((jackknife_coefficients -
jackknife_mean_coefficients)^2))
 # Calculate the 95% confidence intervals using t-distribution
```

```
 jackknife_ci <- t(sapply(1:2, function(i) {
   jackknife_mean_coefficients[i] + c(-1, 1) * t_value * jackknife_se[i]
}))
# Output results
 jackknife_mean_coefficients
[1] 5.956553e+01 6.371609e-04
intercept <- format(jackknife_mean_coefficients[1], scientific = FALSE)
 slope <- format(jackknife_mean_coefficients[2], scientific = FALSE)
 intercept
[1] "59.56553"
 slope
[1] "0.0006371609"
 jackknife_ci
          [,1]     [,2]
[1,] -929.0843 1048.215
[2,] -988.7267  988.728
```

**Results:**

Bootstrap Resampling

Mean Estimate: 67.01898

Bias: 0.01155942

Variance: 1.037006

Standard Error: 1.018335

95% Confidence Interval: [64.94169 ,68.93795 ]

The mean estimate of life expectancy in 2007, derived from the bootstrap resampling method, is approximately 67.019 years. The bias is relatively small, indicating that the bootstrap method's estimate is close to the actual sample mean. The variance and standard error suggest moderate variability in the resampled means. The 95% confidence interval, ranging from 64.942 to 68.938 years, provides a reasonable estimate of the population mean life expectancy.

The symmetric distribution observed in the bootstrap resampling reflects its ability to capture the dataset's variability comprehensively, including the presence of outliers in multiple samples.

## Jackknife Resampling

Mean Estimate: 67.00742

Bias: 0

Variance: 1.026464

Standard Error: 0.08502127

95% Confidence Interval: [66.83934 , 67.17550]

The mean estimate of life expectancy, derived from the jackknife resampling method, is approximately 67.007 years. The absence of bias confirms that the jackknife method's estimate aligns perfectly with the sample mean. The variance and standard error are lower than those from the bootstrap method, indicating less variability in the jackknife estimates. The 95% confidence interval, ranging from 66.839 to 67.176 years, is narrower, suggesting a more precise estimate of the population mean life expectancy.

The skewed distribution observed in jackknife resampling may be attributed to its exclusion of observations one at a time, leading to biased estimations and skewed distributions, especially in datasets with outliers.

## Linear Regression Using Bootstrap

Intercept: 59.5388

Slope (gdpPercap)**: 0.0006409877

95% Confidence Intervals**:

Intercept: [57.30683 ,61.60131]

Slope (gdpPercap): [0.0005463233 , 0.0007520042]

So, the linear regression equation based on these bootstrap estimates would be:

$$lifeExp=59.53887+0.0006409877\times gdpPercap$$

This equation indicates that, according to the bootstrap resampling analysis, for every unit increase in GDP per capita, the life expectancy is expected to increase by approximately 0.0006409877 years. The confidence intervals for both the intercept and slope are relatively narrow, suggesting precise estimates.

## Linear Regression Using Jackknife

Intercept: 59.56553

Slope (gdpPercap): 0.0006371609

95% Confidence Intervals:

Intercept: [-929.0843 ,1048.215]

Slope (gdpPercap): [-988.7267 , 988.728]

Based on the jackknife mean coefficients, the regression model is:

$$\text{lifeExp} = 59.56553 + 0.0006371609 \times \text{gdpPercap}$$

The estimates of the intercept and slope from the jackknife resampling are similar to those obtained from the bootstrap method. However, the confidence intervals are extremely wide and unrealistic. This discrepancy suggests that the jackknife method may not be appropriate for estimating the confidence intervals of the regression coefficients in this context, possibly due to the small sample size or the influence of outliers.

The bootstrap and jackknife methods provide similar point estimates for the mean life expectancy and the coefficients of the regression model. The slight differences in the mean estimates and regression coefficients are within acceptable limits.

However, the precision of the estimates differs significantly between the two methods. The bootstrap method provides narrower and more realistic confidence intervals for the regression coefficients, indicating that it may be more reliable for this dataset. The wide confidence intervals from the jackknife method suggest that it might not be suitable for this type of analysis, possibly due to the variability and outliers in the dataset.

The analysis of the 'gapminder' dataset using bootstrap and jackknife resampling techniques offers valuable insights into life expectancy and its relationship with GDP per capita. The findings indicate that higher GDP per capita is associated with increased life expectancy, as evidenced by the positive slope in the regression models.

Here Bootstrap method provides reliable estimates with reasonable confidence intervals. It is effective for this dataset, offering a clear understanding of the relationship between life expectancy and GDP per capita.

While the Jackknief offers similar point estimates, the wide confidence intervals highlight its limitations in this context. This suggests the need for caution when interpreting the jackknife results, especially for regression analysis.

In summary, the bootstrap method appears to be the preferred technique for analyzing the 'gapminder' dataset, providing robust and precise estimates for both the mean life expectancy and the relationship between life expectancy and GDP per capita.

## CONCLUSION

In conclusion, resampling methods are essential statistical techniques for assessing model accuracy by repeatedly drawing samples from a dataset. These methods are particularly valuable for limited data or complex models, aiding in model selection and performance evaluation. Thus, they are vital tools for data analysts and machine learning practitioners.

This study explored and compared two fundamental resampling techniques: the bootstrap and the jackknife. These methods are crucial for estimating estimator properties such as bias, confidence intervals (CI), standard error, and variance, as well as for hypothesis testing and regression analysis.

Both the bootstrap and the jackknife effectively estimated estimator bias, with the bootstrap often providing more accurate estimates due to extensive resampling. Variance estimation was also more precise with the bootstrap. The jackknife, while simpler and less computationally intensive, provided reliable variance estimates for small samples.

In terms of confidence intervals, the jackknife was reliable but less flexible than the bootstrap, especially with small or non-normal datasets. Bootstrap estimates of standard error were consistently accurate across various scenarios due to its adaptability, while the jackknife's standard error estimation, although less flexible, was adequate for larger samples.

Bootstrap hypothesis testing excelled in non-parametric situations, offering robust alternatives to traditional tests with accurate p-value estimation without strict distribution assumptions. In regression analysis, both methods enhanced parameter estimate robustness. The bootstrap provided detailed insights into regression coefficient variability and stability, while the jackknife offered valuable insights into individual observation influence.

Using the Gapminder dataset, we applied both techniques to analyze life expectancy, GDP per capita, and population growth. The bootstrap demonstrated superior flexibility with diverse, skewed data, providing

detailed confidence intervals and bias estimates, thus enhancing the reliability of our conclusions. The jackknife, simpler and less computationally demanding, offered valuable preliminary insights and robustness in variance and standard error estimation, proving useful despite some limitations compared to the bootstrap.

Beyond these applications, we discussed the broader applicability of these resampling methods in various contexts, including dependent datasets, sample surveys, non-parametric models, and non-linear models. The bootstrap's resampling technique can be adapted for dependent datasets, providing more accurate estimates. In sample surveys, both methods help understand survey estimate variability and reliability. Non-parametric models benefit significantly from the bootstrap's flexibility, and both techniques assist in estimating variability and bias in non-linear models.

Both the bootstrap and jackknife are powerful tools, each with unique strengths. The bootstrap stands out for its flexibility and accuracy, making it ideal for complex, non-parametric data analyses. The jackknife, with its simplicity and efficiency, remains valuable for initial analyses and when computational resources are limited.

Overall, this study highlights the importance of selecting the appropriate resampling technique based on the analysis context and requirements, leading to more informed and accurate statistical inferences. As data complexity grows, resampling techniques like the bootstrap and jackknife will become increasingly vital for robust and reliable statistical analyses.

# REFERENCES


1. Yu, C. H. (2002). Resampling methods: Concepts, applications, and justification. *Practical Assessment, Research, and Evaluation, 8*(19).

2. Sahinler, S. (n.d.). Bootstrap and jackknife resampling algorithms for estimation of regression parameters. Biometry and Genetics Unit, Department of Animal Science, Agriculture Faculty, University of Mustafa Kemal, Hatay, Turkey.

3. Efron, B. (n.d.). Bootstrap methods: Another look at the jackknife. Stanford University.

4. MacKinnon, J. G. (2007). Bootstrap hypothesis testing (Queen's Economics Department Working Paper No. 1127). Queen's University, Department of Economics.

5. The Bootstrap: Advanced methods for data analysis (36-402/36-608) Spring 2014. (2014).

6. DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science, 11*(3), 189-212.

7. Berger, D. (n.d.). A gentle introduction to resampling techniques. Claremont Graduate University.

8. Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. Stanford University, Department of Statistics.

9. Friedl, H., & Stampfer, E. (2002). Jackknife resampling. In A. H. El-Shaarawi & W. W. Piegorsch (Eds.), *Encyclopedia of Environmetrics* (Vol. 2, pp. 1089–1098). Chichester: John Wiley & Sons, Ltd.

10. Aguirre, J. F. R. (n.d.). Revisiting jackknife confidence intervals. Departamento de Estatística, ICEx, Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte, MG, Brazil.

11. Resampling Methods: The Jackknife

12.The Bootstrap and Jackknief ,Summer Institute

13. Şahinler, S., & Topuz, D. (2007, January 1). Bootstrap and jackknife resampling algorithms for estimation of regression parameters. Journal of Applied Quantitative Methods, 2.

14. Babu, G. J. (n.d.). Jackknife and bootstrap [Notes]. Center for Astrostatistics, The Pennsylvania State University.

15. Berger, D. (n.d.). A gentle introduction to resampling techniques [Handout]. Claremont Graduate University.

16.Resampling Methods

17. Phillip I. Good Second Edition, Introduction to Statistics through Resampling methods and R, John Willey and Sons Publication