Resampling Methods: Bootstrap and Permutation

Applied Statistics

Fall 2025

目录

大纲 Outline

1. 引言：重抽样方法简介 Introduction to Resampling Methods
2. 自助法（Bootstrap）的基本思想 The Bootstrap Idea
3. 自助法分布及其性质 The Bootstrap Distribution and Its Properties
4. 自助法置信区间 Bootstrap Confidence Intervals
    - 自助法 t 置信区间 Bootstrap t Confidence Interval
    - 百分位数置信区间 Percentile Confidence Interval
    - 偏差校正与加速置信区间 (BC/BCa) Bias-Corrected and Accelerated (BC/BCa) Intervals
5. 自助法的应用 Bootstrap Applications
    - 相关系数的自助法推断 Bootstrap for Correlation
    - 两组比较的自助法 Bootstrapping to Compare Two Groups
6. 置换检验 Permutation Test
    - 基本思想与步骤 The Idea and Procedure
    - 与 t 检验的比较 Comparison with the t-test
    - 应用场景与限制 When to Use and Limitations
7. 总结与比较 Summary and Comparison

1 引言：重抽样方法简介 Introduction to Resampling Methods

1.1 计算与统计学的革命 The Revolution in Computing and Statistics

- 核心思想：应用大量计算来寻找统计量的抽样分布。
  The core idea: apply large amounts of computation to find sampling distributions of statistics.
- 概念简单：直接基于抽样分布的定义，通过模拟随机抽样过程来实现。
  Conceptually simple: directly based on the definition of the sampling distribution to mimic the random process of sampling.
- 抽样分布源于从总体中抽取的许多可能样本，因此我们通过从数据中抽取许多“重抽样”样本来“估计”它。
  The sampling distribution results from many possible samples from the population, so we “estimate” it using many “resamples” from our data.
- 在足够的计算能力（大量重复）下，重抽样方法比传统方法更准确。
  With sufficient computing power (lots of replications), resampling methods are more accurate than traditional methods.

1.2 适用场景与优势 When to Use and Advantages

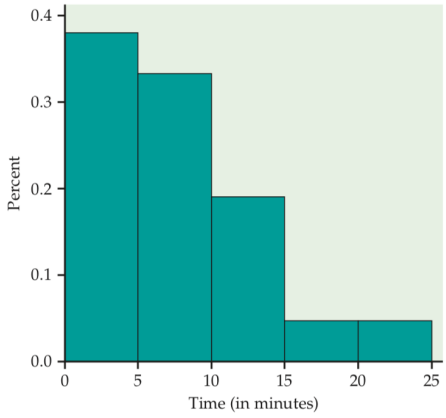
- **传统方法假设不满足时特别有用：**例如，对于中等样本量的强偏态数据，使用 t 检验可能不合适。
  Useful in settings where assumptions for traditional methods are not met, such as t test for strongly skewed data with modest sample size.
- **没有简单传统推断方法的统计量：**例如均值之比（如男性平均工资与女性平均工资之比）。
  Or for statistics where no simple traditional inference method exists like a ratio of means, such as the ratio of average men’s salary to average women’s salary.
- **对总体分布假设要求低：**重抽样方法要求数据是来自总体的随机样本，但不需要总体服从正态分布或其他分布假设。
  Resampling methods require random samples from the population. There is no need for Normality or other distributional assumptions for the population.
- **两种主要类型：**

- **自助法 Bootstrap:** 估计抽样分布。
  Bootstrap = estimate sampling distribution.
- **置换检验 Permutation test:** 在原假设允许的重新排列下检验假设。
  Permutation test = test hypotheses under rearrangements allowed by the null.

2 自助法（Bootstrap）的基本思想 The Bootstrap Idea

2.1 核心概念

- 抽样分布源于从总体中随机抽样的过程，因此我们使用“从数据中重抽样”来近似它。
  Sampling distribution comes from the process of random sampling from the population, so we use “resampling from data” to approximate it.
- **示例：**浏览 Facebook 个人资料的平均时间。数据右偏，且只有 21 个观测值，因此对使用 t 方法有顾虑。
  Example: Average time looking at a Facebook profile. The data are skewed to the right. With only 21 observations, concerns about using the t procedures.

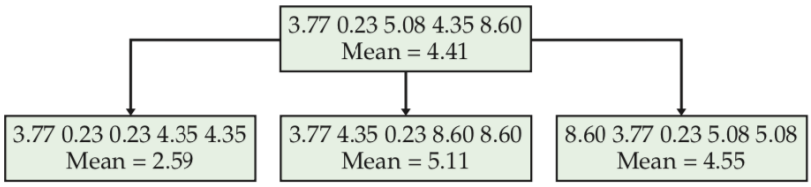


- **关键假设：**原始样本对其所来自的总体具有代表性。
  The original sample is representative of the population from which it was drawn.
- 从这个原始样本中重抽样，代表了如果我们从总体中抽取许多样本将会出现的变异模式。
  Resamples from this original sample represent the pattern of variation if we took many samples from the population.
- 因此，统计量的自助法分布代表了该统计量的抽样分布。
  Thus, the bootstrap distribution of a statistic represents the sampling distribution of the statistic.

2.2 重抽样步骤 The Resampling Step

- 自助法从我们已有的数据中重抽样许多样本开始。
  The bootstrap starts by resampling many samples from the data we have.

- **步骤 1：重抽样**。通过有放回地重复抽样来创建许多重抽样样本。每个重抽样样本的大小与原始随机样本相同。
  Step 1: Resampling. Create many resamples by repeatedly sampling **with replacement**. Each resample is the same size as the original random sample.



- **有放回抽样 Sampling with replacement:** 从原始样本中随机抽取一个观测值后，在抽取下一个观测值之前将其放回。
  After we randomly draw an observation from the original sample we put it back before drawing the next observation.

- 任何观测值都可能被多次抽取。
  Any number can be drawn more than once.

- 如果采用无放回抽样，我们将得到与开始时相同的一组数字，只是顺序不同（即只是一个排列）。
  If we sampled without replacement, we'd get the same set of numbers we started with, though in a different order (i.e., a permutation).

2.3 自助法分布 The Bootstrap Distribution

- 统计量的自助法分布总结了所有可能相同大小的重抽样样本中该统计量的值。
  The bootstrap distribution of a statistic summarizes the values of the statistic in all possible resamples of the same size.

- 基于大量重抽样（例如 3000 次），我们可以绘制自助法分布图，它提供了关于抽样分布的**形状和变异性**信息。
  Based on many resamples (e.g., 3000), we plot the bootstrap distribution, which gives information on the **shape and variability** of the sampling distribution.

- **形状 Shape:** 自助法分布通常接近正态（如果原始样本量足够大）。
  The bootstrap distribution is often nearly Normal.

- **中心 Center:** 自助法分布以原始样本统计量（例如 $\bar{x}$）为中心，而不是总体参数 μ。
  The bootstrap distribution is centered at the value of the original sample statistic (e.g., $\bar{x}$), not the population parameter μ.

- **变异性 Variability:** 自助法分布的标准差称为**自助法标准误 (bootstrap standard error)**。
  The standard deviation of the bootstrap distribution is called the **bootstrap standard error**.

2.4 与理论抽样分布的关系 Relationship to Theoretical Sampling Distribution

- 自助法分布的性质与统计理论所提示的抽样分布性质相匹配。
  The bootstrap distribution created by resampling matches the properties of the sampling distribution suggested by statistical theory.

- 例如，在 Facebook 时间示例中，$n = 21$, $s = 5.65$。
  For the Facebook time example, $n = 21$, $s = 5.65$.

  - 中心极限定理给出的标准误估计：$s/\sqrt{n} = 5.65/\sqrt{21} \approx 1.23$。
    The CLT-based estimated standard error: $s/\sqrt{n} = 5.65/\sqrt{21} \approx 1.23$.
  - 自助法标准误（基于 3000 次重抽样）计算为 1.22，两者非常接近。
    The bootstrap standard error (based on 3000 resamples) was 1.22, very close to 1.23.

- **重抽样思想的巨大优势：**它通常在理论不适用时仍然有效。
  The great advantage of the resampling idea is that it often works even when theory does not apply.

- 产生自助法分布所需的繁重计算，取代了繁重的理论（如中心极限定理）。
  The heavy computation needed to produce the bootstrap distribution replaces the heavy theory (central limit theorem).

- The bootstrap distribution of the resample means is used only to estimate how the sample mean would vary because of random sampling, i.e. the **bootstrap standard error**.
  重采样均值的 bootstrap 分布仅用于估计样本均值因随机抽样而产生的变化，即 **bootstrap 标准误**。

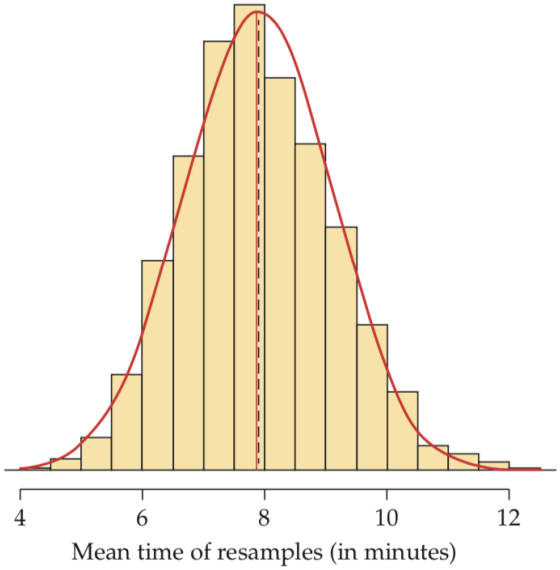


图 1: 自助法分布示例（基于 3000 次重抽样）Bootstrap Distribution Example (based on 3000 resamples)

- We also look at the bootstrap distribution to see if it is roughly Normal (or not).
  我们也会观察 bootstrap 分布是否大致服从正态分布（或不是）。
- Suppose that we take B resamples.
  假设我们进行 B 次重采样。

$$\text{mean}_{\text{boot}} = \frac{1}{B}\sum x^*$$

$$\text{Stan. Error}_{\text{boot}} = \sqrt{\frac{1}{B-1}\sum (x^* - \text{mean}_{\text{boot}})^2}$$

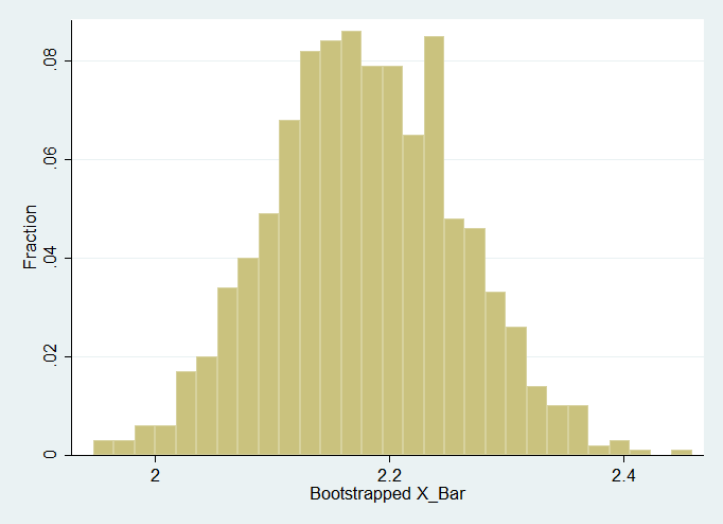
Recall our newspaper recycling example: we collect the weekly newspaper recycle weight in urban and suburbs. Below is a summary of the data.
回顾我们的报纸回收示例：我们收集了城市和郊区的周报纸回收重量。以下是数据摘要。

Variable 变量	Obs 观测数	Mean 均值	Std. Dev. 标准差	Min 最小值	Max 最大值
Newspaper	148	2.180405	.9811599	0	4.4
报纸回收量	148	2.180405	.9811599	0	4.4

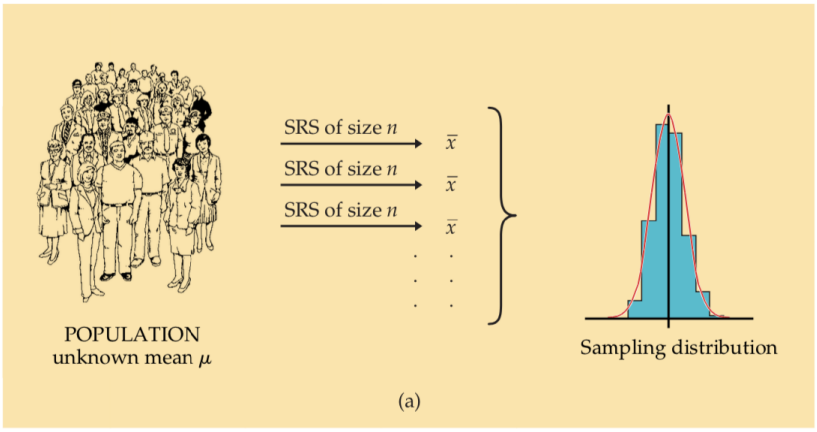
CLT suggests that the sampling distribution of the mean recycle weight is normal with standard error $\frac{s}{\sqrt{n}} = 0.0807$.
中心极限定理表明，平均回收重量的抽样分布是正态的，其标准误为 $\frac{s}{\sqrt{n}} = 0.0807$。

Resample 1000 times, plot the distribution of the 1000 means.
重采样 1000 次，绘制 1000 个均值的分布。

Variable 变量	Obs 观测数	Mean 均值	Std. Dev. 标准差	Min 最小值	Max 最大值
x_bar 样本均值	1,000 1,000	2.179164 2.179164	.0802463 .0802463	1.947297 1.947297	2.458784 2.458784

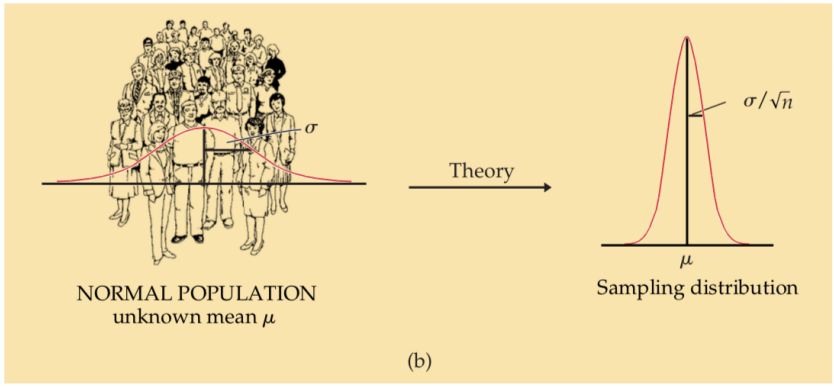


2.5 图示对比：三种“分布” Visual Comparison: Three ”Distributions”



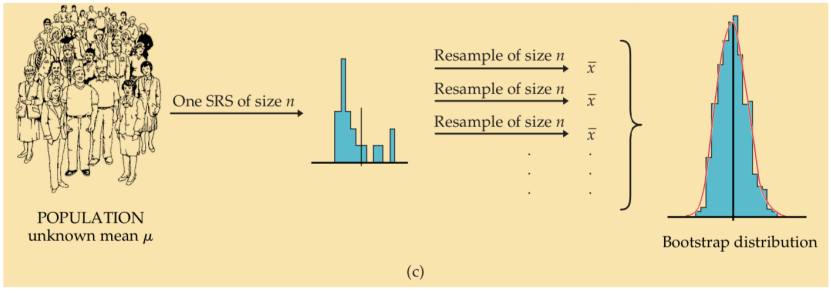
(a) 理论抽样分布

Theoretical Sampling Distribution



(b) 基于中心极限定理的近似

CLT-based Approximation



(c) 自助法抽样分布

Bootstrap Sampling Distribution

图 2: 三种获得抽样分布方式的对比

Comparison of three ways to obtain the sampling distribution

3 自助法分布的性质与可信度 Bootstrap Distribution Properties and Reliability

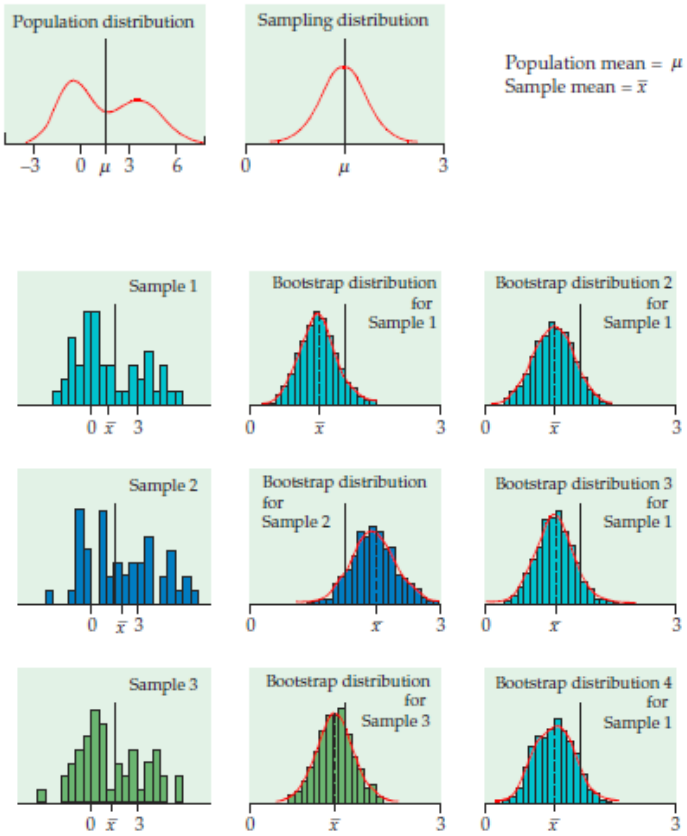
3.1 自助法分布中的随机变异来源 Sources of Random Variation

- 创建自助法分布时有两个随机变异来源：
  Two sources of random variation in creating a bootstrap distribution:

  1. 从总体中随机选择一个原始样本。
     Choosing an original sample at random from the population.
  2. 从原始样本中随机选择自助法重抽样样本。
     Choosing bootstrap resamples at random from the original sample.

- 我们希望自助法分布能够可靠地估计抽样分布。这取决于原始样本对总体的代表性以及重抽样的次数。
  We want the bootstrap distribution to reliably estimate the sampling distribution. This depends on how well the original sample represents the population and the number of resamples.

- 从以下总体中抽取样本量 50 的：



3.2 大样本下的自助法 Bootstrap with Large Samples

- 如果自助法分布基于一个来自总体的中等或大样本，那么它的形状和变异性不会严重依赖于原始样本，并且确实能够模拟抽样分布的形状和散布。
  If a bootstrap distribution is based on a moderately large sample from the population, its shape and variability don't depend heavily on the original sample and do mimic the shape and spread of the sampling distribution.

- 来自同一总体的不同大样本产生的自助法分布在形状、中心和散布上都非常相似。重抽样在大量重复的情况下引入的额外变异非常少。
  The bootstrap distributions from different large samples from the same population are very similar in shape, center, and spread. Resampling introduces very little additional variation with lots of reps.

- **安全使用条件：**中等或大样本量，以及大量的重抽样重复次数（通常 $B \geq 1000$）。
  Safe use for moderate sample size and lots of resampling replications (typically $B \geq 1000$).

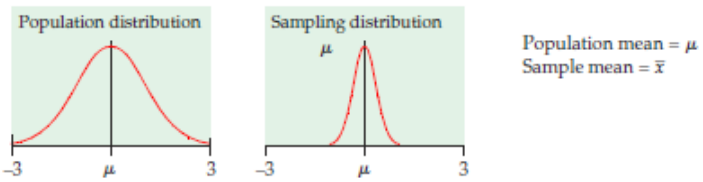
3.3 小样本下的自助法 Bootstrap with Small Samples

- 小样本在自助法情况下会给出更多变的结果。
  Small samples give more variable results in the case of bootstrap.

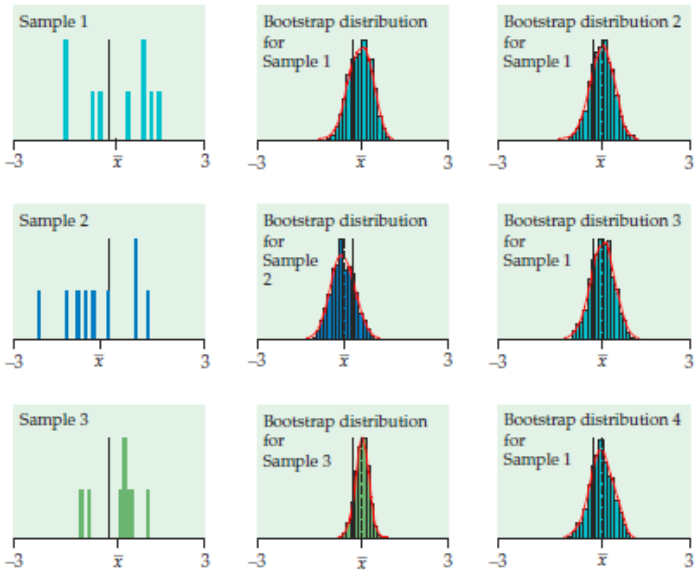
- 小样本对总体的代表性较差；反复进行自助法会放大它们的特殊性。
  Small samples represent the population poorly; bootstrapping repeatedly amplifies their quirks.

- 即使总体是正态的，基于不同小样本的自助法分布在形状和散布上也表现出更大的变异。
  Even with a Normal population, the bootstrap distributions based on different small samples show much more variation in shape and spread.

- 从以下总体中抽取样本量 9 的：



- **结论：**不能信任基于非常小样本的自助法分布能紧密模拟抽样分布的形状和散布。自助法不能克服小样本的弱点。对于来自小样本的任何推断（包括自助法推断）



都需要谨慎。

Can't trust a bootstrap distribution from a very small sample to closely mimic the shape and spread of the sampling distribution. Bootstrapping does not overcome the weakness of small samples. Use caution in any inference-including bootstrap inference-from a small sample.

3.4 偏差估计 Bias Estimation

- 自助法分布的中心接近其原始样本的 $\bar{x}$，而不是总体均值 μ。
  Each bootstrap distribution is centered close to the value of $\bar{x}$ of its original sample, not the population μ.

- 自助法中心与 $\bar{x}$ 之间的距离模拟了**偏差 (bias)**，即抽样分布中心与总体均值之间的差异 $E(\bar{X}) - \mu$。
  The distance between the bootstrap center and the $\bar{x}$ mimics **bias**, the difference between the center of the sampling distribution and the population mean $E(\bar{X}) - \mu$.

- 我们可以将偏差估计为：
  We can estimate the bias as:

$$\widehat{\text{Bias}} = \text{mean}_{\text{boot}} - \bar{x}$$

  其中 $\text{mean}_{\text{boot}}$ 是自助法分布中统计量的平均值。
  where $\text{mean}_{\text{boot}}$ is the mean of the bootstrap distribution of the statistic.

- 如果偏差很小，表明原始样本统计量是总体参数的一个近似无偏估计。
  A small bias indicates that the original sample statistic is an approximately unbiased estimator of the population parameter.

4 自助法置信区间 Bootstrap Confidence Intervals

4.1 概述

- 获得自助法分布后，我们可以基于它进行推断。
  After we obtain the bootstrap distribution, we can base our inference on it.

- 主要用途：

  1. 检查抽样分布的正态性和偏差。
     Use the bootstrap distribution to check the Normality and bias of the sampling distribution.
  2. 结合正态分布或 t 分布使用自助法标准误：**自助法 t 置信区间**。
     The bootstrap standard error estimates in combination with a normal or t distribution: **Bootstrap t confidence interval**.
  3. 直接使用自助法分布进行推断：**百分位数置信区间，偏差校正 (BC) 和 偏差校正加速 (BCa) 置信区间**。
     Directly using the bootstrap distribution for inference: the **percentile confidence interval**, **bias-corrected (BC)** and **bias-corrected and accelerated (BCa) confidence intervals**.

4.2 自助法 t 置信区间 Bootstrap t Confidence Interval

- **定义**：通过使用自助法标准误和熟悉的 t 分布来构造参数的置信区间。
  A confidence interval for the parameter by using the bootstrap standard error and the familiar t distribution.

- **适用条件**：如果自助法分布呈现正态形状且偏差较小。
  Applicable if the bootstrap distribution shows a Normal shape and small bias.

- **公式**：类似于传统的 t 区间，但用自助法标准误 SE_{boot} 代替理论标准误 $s/\sqrt{n}$。
  Formula: similar to the traditional t interval, but using the bootstrap standard error SE_{boot} instead of the theoretical standard error $s/\sqrt{n}$.

$$\text{统计量} \pm t^*_{\alpha/2} \times SE_{\text{boot}}$$

其中 $t^*_{\alpha/2}$ 是自由度为 df 的 t 分布的临界值。通常，对于样本均值，$df = n - 1$。但对于其他统计量，自由度可能不明确。在实践中，如果样本量足够大，通常使用正态临界值 $z^*_{\alpha/2}$ 作为近似。
where $t^*_{\alpha/2}$ is the critical value from the t distribution with degrees of freedom df. Typically, for the sample mean, $df = n - 1$. But for other statistics, the degrees of

freedom may not be clear. In practice, if the sample size is sufficiently large, the normal critical value $z^*_{\alpha/2}$ is often used as an approximation.

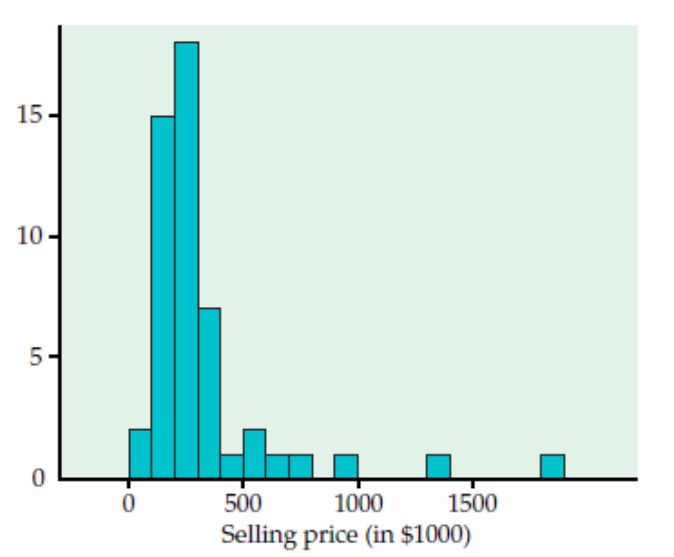
- 也可用于假设检验。
  Can also use the bootstrap standard error and t distribution for hypothesis testing.

4.2.1 示例：西雅图房地产价格（截尾均值）

西雅图房地产价格自助法 t 区间

背景：西雅图房地产价格数据强烈右偏，且有高异常值。需要一种对异常值更稳健的中心度量。选择 **25% 截尾均值**（即中间 50% 观测值的均值）。
Background: Real estate price in Seattle: strongly skewed, with high outliers. Need a measure of center that is more resistant to outliers. Choose the **25% trimmed mean** (the mean of the middle 50% observations).
为什么选择截尾均值而非中位数？因为通常的自助法程序对于中位数或其他百分位数效果不佳，除非原始样本量很大。
Why trimmed mean over median? Because the usual bootstrapping procedure doesn't work well for the median or other percentiles unless the original sample is quite large.
数据：原始样本量 $n = 100$，原始统计量（25% 截尾均值）$\bar{x}_{\text{trim}} = 244.00192$。
Data: Original sample size $n = 100$, original statistic (25% trimmed mean) $\bar{x}_{\text{trim}} = 244.00192$.
重抽样：进行 $B = 1000$ 次重抽样，得到自助法分布。
Resampling: Perform $B = 1000$ resamples, obtain the bootstrap distribution.

- 自助法分布均值：$\text{mean}_{\text{boot}} = 244.5123$
  Bootstrap distribution mean: $\text{mean}_{\text{boot}} = 244.5123$
- 自助法标准误：$SE_{\text{boot}} = 18.2454$
  Bootstrap standard error: $SE_{\text{boot}} = 18.2454$
- 偏差估计：$\widehat{\text{Bias}} = 244.5123 - 244.00192 \approx 0.51038$，相对较小。
  Bias estimate: $\widehat{\text{Bias}} = 244.5123 - 244.00192 \approx 0.51038$, relatively small.



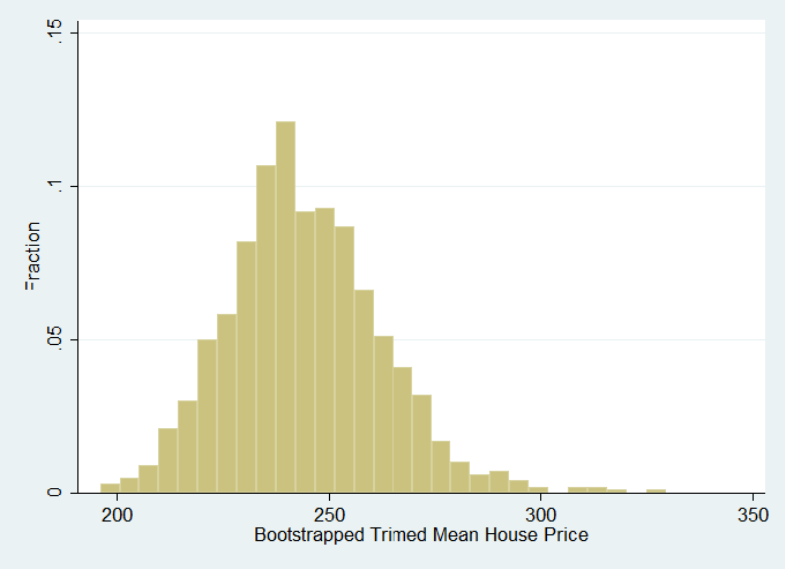


图 3: 截尾均值的自助法分布

Bootstrap Distribution of the Trimmed Mean

判断：自助法分布近似正态且偏差小，可以使用自助法 t 区间。

Judgment: The bootstrap distribution is approximately Normal and has a small bias, so we can use the bootstrap t interval.

计算 95% 置信区间：

Compute 95% Confidence Interval:

- 对于截尾均值，有效的样本量减少。一种近似的自由度计算：$n_{\text{effective}} = n - 2\times$ (被截尾的观测数)。本例中，截尾 25% 意味着两端各去掉 25 个观测值，中间保留 50 个观测值。因此，有效样本量约为 50。更保守的做法是使用中间 50%

观测值的数量减 1。这里我们使用 t 临界值 $t_{0.025}(50-1) \approx t_{0.025}(49) \approx 2.009$。
For the trimmed mean, the effective sample size is reduced. An approximate degrees of freedom: $n_{\text{effective}} = n - 2 \times$ (number of observations trimmed). Here, trimming 25% means removing 25 observations from each end, leaving 50 in the middle. So effective sample size 50. More conservatively, use the number of observations in the middle minus 1. We use $t_{0.025}(50-1) \approx t_{0.025}(49) \approx 2.009$.

- 区间：$244.00192 \pm 2.009 \times 18.2454 \approx 244.00192 \pm 36.65 \approx (207.35, 280.65)$。
Interval: $244.00192 \pm 2.009 \times 18.2454 \approx 244.00192 \pm 36.65 \approx (207.35, 280.65)$.

4.3 百分位数置信区间 Percentile Confidence Interval

- **定义**：直接使用自助法分布的百分位数构造的置信区间。
The interval between the $\alpha/2$ and $1-\alpha/2$ percentiles of the bootstrap distribution.
- **95% 百分位数区间**：自助法分布的 2.5% 和 97.5% 百分位数之间的区间。
95% bootstrap percentile confidence interval: the interval between the 2.5% and 97.5% percentiles of the bootstrap distribution.
- **优点**：简单直观，无需假设分布形状或计算标准误。
Simple and intuitive, no need to assume a distribution shape or compute standard error.
- **缺点**：如果自助法分布有偏或不对称，区间可能不准确。
May be inaccurate if the bootstrap distribution is biased or asymmetric.
- **检验 t 区间合理性的方法**：比较自助法 t 区间和百分位数区间。如果两者不一致，则两种区间都不应被信任。
One way to check whether t intervals are reasonable is to compare them with the confidence interval directly based on the bootstrap distribution. When bootstrap t and bootstrap percentile intervals do not agree closely, neither type of interval should be used.

4.4 偏差校正与加速置信区间 (BC/BCa) Bias-Corrected and Accelerated (BC/BCa) Intervals

- **动机**：为了改进百分位数区间，考虑自助法分布的偏差和偏度。
To improve upon the percentile interval by accounting for bias and skewness in the bootstrap distribution.

- **偏差校正 (BC) 区间：**调整自助法分布的偏差。
  BC intervals adjust for bias in the bootstrap distribution.
- **偏差校正加速 (BCa) 区间：**同时考虑偏差和偏度（通过“加速”参数）。
  The BCa interval further takes into account both bias and skewness (via an "acceleration" parameter).
- **原理：**
  - 偏差校正因子与自助法估计值中小于观测统计量的比例有关。
    The bias correction factor is related to the proportion of bootstrap estimates that are less than the observed statistic.
  - 加速参数与自助法分布的偏度成比例。Stata 使用刀切法 (jackknife) 来估计加速参数。
    The acceleration parameter is proportional to the skewness of the bootstrap distribution. Stata uses the jackknife method to estimate the acceleration parameter.
  - 如果偏差和加速项为零，这些区间就简化为百分位数区间。
    If the bias and acceleration terms are zero, these intervals reduce to the percentile confidence interval.
- **何时使用 BC/BCa 区间？**
  When to use BC/BCa intervals?
  - 统计量存在强烈偏差（由自助法偏差估计表明）。
    The statistic is strongly biased, as indicated by the bootstrap estimate of bias.
  - 统计量的抽样分布明显偏斜（由自助法分布以及比较 t 区间和百分位数区间表明）。
    The sampling distribution of the statistic is clearly skewed, as indicated by the bootstrap distribution and by comparing the t and percentile intervals.
- **推荐：**通常，BCa 是推荐的自助法置信区间方法。它是“二阶精确”的，意味着它以更快的速度收敛到正确的覆盖率。然而，偏差和加速项估计的准确性需要大量的自助法样本，并且在使用刀切法获取加速参数时，计算量可能很大。
  In general, BCa is the recommended bootstrap CI method. It is "second-order accurate", meaning that it converges faster to the correct coverage. However, the accuracy of the bias and acceleration terms requires a large number of bootstrap samples and, especially when using the jackknife to get the acceleration parameter, this can be computationally intensive.

4.4.1 软件输出示例 (Stata)

Stata 自助法置信区间输出

```
Bootstrap results

command: trimmean HousePrice, p(25)
trimmean: r(tmean25)

------------------------------------------------------------------------------
|   Observed   Bootstrap         Normal-based
|      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
trimmean   |   244.00192   16.623825    14.68   0.000     211.4198    276.584 (N)
|                                                       214.1135   277.9211 (P)
|                                                       213.5865   277.2635 (BC)
|                                                       213.9154   277.5635 (BCa)
------------------------------------------------------------------------------
(N) normal confidence interval
(P) percentile confidence interval
(BC) bias-corrected confidence interval
(BCa) bias-corrected and accelerated confidence interval
```

注意： Stata 在自助法 t 区间中使用正态分布（z）而非 t 分布。

Note: Stata uses normal distribution in place of t distribution for the bootstrap t CI. https://www.stata.com/manuals13/rbootstrap.pdf#rbootstrap

5 自助法的应用 Bootstrap Applications

5.1 相关系数的自助法推断 Bootstrap for Correlation

- **场景：**涉及两个相关变量的统计量。
  Use of the bootstrap for a statistic that depends on two related variables.
- **示例：** MLB 球员的薪水与表现评分数据。我们有 50 名随机选择的 MLB 球员的姓名、薪水和表现评分。
  Example: we have data on the names, salaries, and performance ratings for 50 randomly selected MLB players.
- **关键：**每个观测值由一名球员的表现评分和薪水组成。我们重抽样球员，并将这对信息一起放入我们的重抽样样本中。
  Because each observation consists of the performance rating and salary for one player, we resample **players** and get the pair of information into our resamples together.

- **不能单独重抽样：**不能分别重抽样表现评分和薪水，否则会失去球员表现与其薪水之间的联系。
  We don’t resample performance ratings and salaries separately: this would lose the tie between a player’s performance and his salary.

- **步骤：**

  1. 从原始数据集中有放回地随机抽取 n 个观测值（即球员），构成一个自助法样本。
     Draw a bootstrap sample of n observations (i.e., players) with replacement from the original dataset.
  2. 计算该自助法样本的相关系数 r^*。
     Compute the correlation coefficient r^* for this bootstrap sample.
  3. 重复多次（如 1000 次），得到相关系数的自助法分布。
     Repeat many times (e.g., 1000) to get the bootstrap distribution of the correlation.
  4. 基于该分布计算置信区间（如百分位数区间或 BCa 区间）。
     Compute a confidence interval based on this distribution (e.g., percentile or BCa interval).

5.1.1 示例：棒球薪水与表现

棒球薪水与表现相关性自助法

背景：散点图显示薪水与击球率之间的关系较弱。样本相关系数 $r = 0.107$。
Background: The scatterplot suggests that the relationship between salary and batting average is weak. The sample correlation coefficient is $r = 0.107$.
问题：计算总体相关系数的 95% 置信区间，并查看它是否包含 0。
Question: Calculate a 95% confidence interval for the population correlation and see whether or not it covers 0.
Stata 输出：
Stata Output:

```
Bootstrap results

command: correlate Salary Average
r: r(rho)

------------------------------------------------------------------------------
|   Observed   Bootstrap         Normal-based
|      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
r |   .10675751   .12728372     0.84   0.402    -.142714    .356229 (N)
```

```
|                                                          -.1301489   .3627903 (P)
|                                                          -.1295977   .3642086 (BC)
-------------------------------------------------------------------------------
(N) normal confidence interval
(P) percentile confidence interval
(BC) bias-corrected confidence interval
```

解读：

- 观测相关系数 $r = 0.1068$。
  Observed correlation $r = 0.1068$.
- 自助法标准误 $SE_{\text{boot}} = 0.1273$。
  Bootstrap standard error $SE_{\text{boot}} = 0.1273$.
- 偏差估计 $\widehat{\text{Bias}} = 0.000563$，非常小。
  Bias estimate $\widehat{\text{Bias}} = 0.000563$, very small.
- 三种 95% 置信区间都包含 0，表明在 0.05 水平上相关性不显著。
  All three 95% confidence intervals include 0, indicating that the correlation is not significant at the 0.05 level.
- 例如，百分位数区间为 $(-0.1301, 0.3628)$，包含 0。
  For example, the percentile interval is $(-0.1301, 0.3628)$, which includes 0.

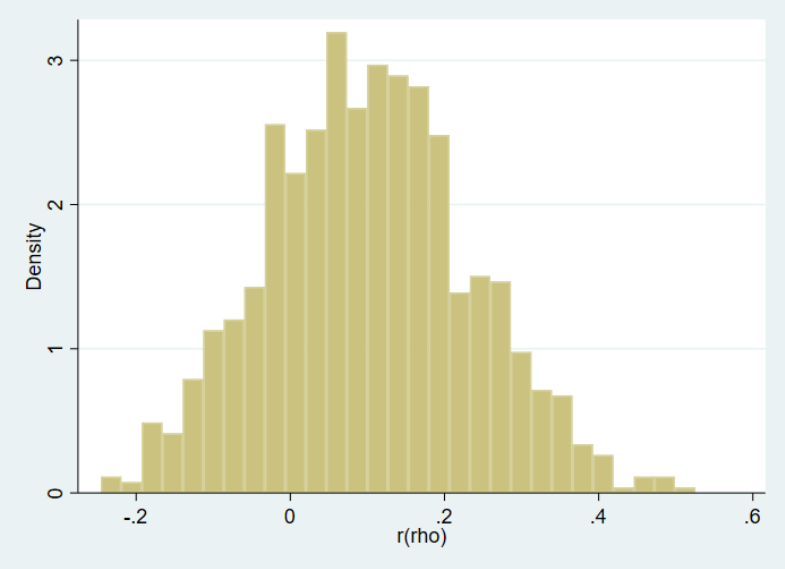


图 4: 相关系数的自助法分布
Bootstrap Distribution of the Correlation

5.2 两组比较的自助法 Bootstrapping to Compare Two Groups

- **场景**：给定来自两个总体的独立随机样本，样本量分别为 n_1 和 n_2。
  Given independent random samples of sizes n_1 and n_2 from two populations.
- **步骤**：
  1. 分别在两个样本内重抽样：从第一个样本中有放回地抽取大小为 n_1 的重抽样样本，从第二个样本中有放回地抽取大小为 n_2 的独立重抽样样本。
     Resample **separately** within two samples: draw a resample of size n_1 with replacement from the first sample and a separate resample of size n_2 from the second sample.
  2. 计算感兴趣的统计量，例如两个样本均值的差 $\bar{x}_1^* - \bar{x}_2^*$。
     Compute the statistic, such as the difference between the two sample means.
  3. 重复多次，得到统计量（如均值差）的自助法分布。
     Repeat many times to get the bootstrap distribution of the statistic (e.g., difference in means).
  4. 基于该分布进行推断（如计算置信区间）。
     Make inference based on this distribution (e.g., compute a confidence interval).
- **注意**：必须分别重抽样，以保持两组之间的独立性假设。
  Must resample separately to maintain the assumption of independence between the two groups.

6 置换检验 Permutation Test

6.1 基本思想与步骤

- **定义**：一种基于重抽样的假设检验方法，在原假设下通过重新排列数据来构建抽样分布。
  A hypothesis testing method based on resampling that constructs the sampling distribution under the null hypothesis by rearranging the data.
- **与自助法的区别**：假设检验是在原假设下进行的，因此我们必须以符合原假设的方式重抽样。这与自助法的重抽样方式不同：不能简单地从观测样本中重抽样。更接近于我们在秩检验中的抽样方式。
  The hypothesis testing operates under the null hypothesis, so we have to resample in a way that is consistent with the null hypothesis. It is different from the bootstrap way of resampling: cannot simply resample from the observed sample. Closer to the way we sample in rank tests.

- **示例**：定向阅读活动是否能提高阅读能力？
  Example: Do directed reading activities improve reading abilities?
  - 处理组（21 名学生）接受新教学法，对照组（23 名学生）接受传统教学法。
    The Treatment group (21 students) is assigned to the new teaching method and the control group (23 students) is taught using the traditional method.
  - 原假设：教学方法不影响阅读分数的分布。
    The null hypothesis for the resampling test is that the teaching method does not affect the distribution of reading scores.
  - 在原假设下，学生无论被分配到哪组，都应有相似的分数。
    Under the null hypothesis, students should have similar scores no matter which group he is assigned to.

- **置换检验的思想：**
  The idea of permutation tests:
  - 如果原假设成立，我们可以随机将学生分配到处理组和对照组，并计算平均分数的差异。
    If the null hypothesis is true, we can then randomly assign students to treatment and control groups and find the difference in average scores.
  - 通过多次随机分配（置换），我们可以得到在原假设下统计量（如均值差）的分布，即**置换分布 (permutation distribution)**。
    By repeating the random assignment many times (permutations), we obtain the distribution of the statistic (e.g., difference in means) under the null hypothesis, called the **permutation distribution**.
  - 将实际观测到的统计量定位在置换分布上，计算 P 值：出现至少与实际观测值一样极端的情况的比例。
    Locate the actual value of the statistic on the permutation distribution to get the P-value: the proportion of at least as extreme cases.

6.1.1 置换检验步骤 General Procedure

1. 计算原始数据的统计量（如两组均值差）。
   Compute the statistic for the original data.

2. 以符合原假设的方式从数据中选择置换重抽样样本。从大量重抽样中构建置换分布。
   Choose permutation resamples from the data in a way that is consistent with the

null hypothesis. Construct the permutation distribution from a large number of resamples.

3. 通过将原始统计量定位在置换分布上来找到 P 值：置换重抽样统计量中“更极端”的比例。
   Find the P-value by locating the original statistic on the permutation distribution: the proportion of the permutation resample statistics that are “more extreme”.

6.1.2 示例说明

阅读能力置换检验示例

假设我们有 4 个学生的分数：24, 61, 42, 33, 46, 37。原假设下，组别标签不影响分数。

- 随机将学生分配到两组（每组 2 人），计算均值差。
  1. 分配：组 1: (24, 61)，组 2: (42, 33, 46, 37)
     均值差：$\bar{x}_1 - \bar{x}_2 = 42.5 - 39.5 = 3.0$
  2. 分配：组 1: (33, 46)，组 2: (24, 61, 42, 37)
     均值差：$39.5 - 41 = -1.5$
  3. 分配：组 1: (33, 61)，组 2: (24, 42, 46, 37)
     均值差：$47 - 37.25 = 9.75$
  4. 分配：组 1: (37, 42)，组 2: (24, 61, 33, 46)
     均值差：$39.5 - 41 = -1.5$
- 重复所有可能的分配（或大量随机分配），得到均值差的置换分布。
- 假设观测到的均值差为 10.125。在置换分布中，计算出现绝对值大于等于 10.125 的比例，即为 P 值。

Stata 输出示例 (蒙特卡洛置换)：
Stata Output Example (Monte Carlo permutation):

```
Monte Carlo permutation results

command:  ttest Reading, by(Group)
diff: r(mu_2)-r(mu_1)
permute var: Group

------------------------------------------------------------------------------
|     T       T(obs)       c       n      p=c/n   SE(p)  [95% Conf. Interval]
-----------+------------------------------------------------------------------
diff |         10.125        29      1000     0.0290  0.0053   .0195059   .0413847
```

```
------------------------------------------------------------------------------
Note: Confidence interval is with respect to p=c/n.
Note: c = #{|T| >= |T(obs)|}
```

解读：观测到的均值差 $T(\text{obs}) = 10.125$。在 1000 次置换中，有 29 次得到的 $|T| \geq 10.125$，因此 P 值 $= 29/1000 = 0.0290$。

Interpretation: The observed difference $T(\text{obs}) = 10.125$. Out of 1000 permutations, 29 had $|T| \geq 10.125$, so P-value $= 29/1000 = 0.0290$.

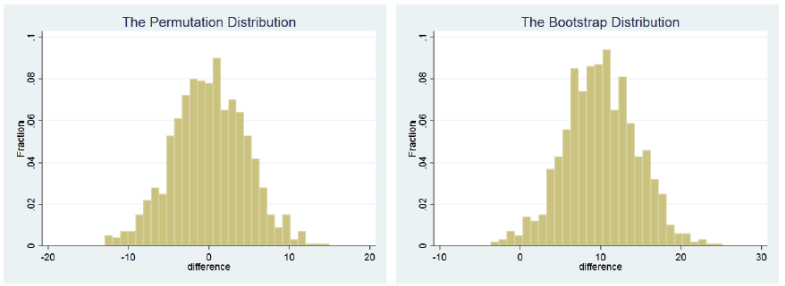


图 5: 置换分布与观测值

Permutation Distribution and the Observed Statistic

6.2 置换检验与 t 检验的比较 Comparison with the t-test

6.2.1 置换检验的优势 Advantages of Permutation Tests

- **假设更一般：**原假设是“两组分布相同”，而不仅仅是均值相等。
  The permutation test hypotheses are more general. The null hypothesis is “same distribution in both groups”.

- **可以检验更灵活的效应度量：**例如截尾均值、均值比等。t 检验仅限于比较样本均值，因为我们需要推导出在原假设下服从 t 分布的统计量。置换检验用计算替代了寻找抽样分布的“理论搜索”。
  Can use more flexible measures of the treatment effect such as trimmed means, and ratio of means. t test is limited to comparing sample means because we have to work out that the statistic has a t distribution when H_0 is true. Permutation replaces the “theoretical searching” for sampling distributions.

- **即使抽样分布不接近正态，也能给出准确的 P 值。**
  The permutation test gives accurate P-values even when the sampling distribution is not close to Normal.

- **可以检查 t 检验的可靠性：**如果 t 检验的 P 值与置换检验的 P 值差异很大，则两样本 t 检验的条件可能不成立。

Can check the reliability of t-tests: if the P-value from t test differs considerably, the conditions for the two-sample t may not hold.

6.2.2 置换检验的局限性 Limitations of Permutation Tests

- **更强的原假设：**置换检验要求当原假设成立时，两个总体完全相同（不仅均值相同，而且方差和形状也相同）。而两样本 t 检验允许方差异质（如果使用不等方差版本）。
  However, it requires that the two populations are identical when the null hypothesis is true, not only are their means the same but so are their variations and shapes. In contrast, the two-sample t allows for unequal variance.
- **计算量可能很大：**特别是当样本量较大时，所有可能的排列数量巨大，通常使用蒙特卡洛方法（随机抽取一部分排列）来近似。
  Can be computationally intensive, especially for large sample sizes. Often use Monte Carlo methods (randomly sample a subset of permutations) to approximate.

6.2.3 示例比较：Verizon 维修时间

Verizon 维修时间：t 检验 vs 置换检验

背景：比较 Verizon 客户（1664 个）与竞争对手客户（23 个）的维修时间。数据严重偏斜且方差异质。

Background: Comparing repair times for 1664 Verizon customers with 23 customers of competing companies. Data are strongly skewed with unequal variances.

不同检验的 P 值比较：

Comparison of P-values from different tests:

检验统计量	t 检验 P 值	置换检验 P 值
$\bar{x}_1 - \bar{x}_2$ (原始均值差)	0.0045	0.0183
合并 t 统计量 Pooled t statistic	0.0044	0.0183
修正 t 统计量 Modified t statistic	0.0044	0.0195

观察：置换检验给出的 P 值大约比 t 检验大 4 倍。这是因为数据严重偏离 t 检验的正态性和方差齐性假设，导致 t 检验的 P 值偏小（更容易犯第一类错误）。置换检验在这种情况下更可靠。

Observation: The permutation test gives P-values about 4 times larger than the t-test. This is because the data severely violate the normality and equal variance assumptions of the t-test, causing the t-test P-values to be too small (more likely to commit Type I error). The permutation test is more reliable in this situation.

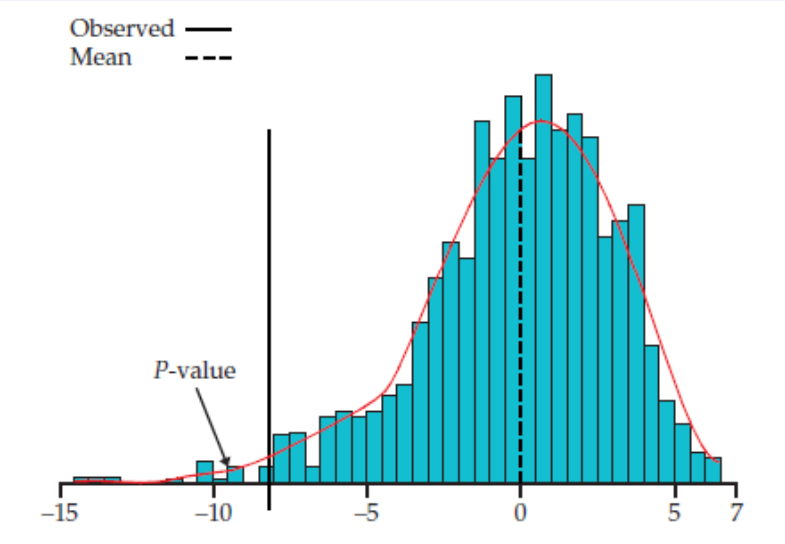


图 6: Verizon 数据均值差的置换分布

Permutation Distribution of Difference in Means for Verizon Data

6.3 置换检验的应用场景与限制 When to Use and Limitations of Permutation Tests

6.3.1 适用场景 When can we use permutation tests?

- **两样本问题：**比较总体均值、比例、标准差或其他统计量。
  Two-sample problems: to compare population means, proportions, standard deviations, or other statistics.

- **配对设计：**例如，研究满月是否影响行为。记录每个患者在“满月日”和“其他日”的攻击行为。原假设：满月无影响，则每个患者的两个分数来自同一分布。重抽样时，随机分配每个患者的两个分数之一给“满月”组，另一个给“其他”组。不混合不同患者的结果，因为原始数据是配对的。
  Matched pairs: Example, can the full moon influence behavior? Record aggressive behaviors on “moon day” and “other day” for each patient. The null hypothesis: the full moon has no effect on behavior. If this is true, the two entries for each patient are from the same distribution. Resampling randomly assigns one of each patient’s two scores to “moon” and the other to “other.” We don’t mix results for different subjects because the original data are paired.

- **两个变量的相关性：**例如，棒球球员表现与薪水的关系。原假设：没有关系。重抽样时，将观察到的薪水在球员中随机分配。
  Correlation between two variables: For example, the relationship between baseball

players' performance ratings and salaries. The null hypothesis is that there is no relationship. Resample by assigning the observed salaries among the players at random.

6.3.2 不适用场景 When can't we use permutation tests?

- 当我们不知道如何以符合原假设的方式重抽样时。
  When we don't know how to resample in a way that matches the null hypothesis.
- **具体包括：**
  - 关于单个总体的假设检验。
    Testing hypotheses about a single population.
  - 比较在原假设下也不同的总体（例如，检验两个总体的方差是否相等，但即使方差相等，分布也可能不同）。
    Comparing populations that differ even under the null hypothesis (e.g., testing if two populations have equal variances, but even if variances are equal, distributions could be different).
  - 检验一般关系（如回归斜率是否为零）。
    Testing general relationships (e.g., whether a regression slope is zero).
- **替代方法：**当我们无法进行置换检验时，通常可以计算自助法置信区间。如果置信区间未包含原假设值，那么我们就在相应的显著性水平上拒绝 H_0。这不如置换检验准确，但置信区间既能估计效应大小，也能提供一些关于统计显著性的信息。
  When we can't do a permutation test, we can often calculate a bootstrap confidence interval. If the confidence interval fails to include the null hypothesis value, then we reject H_0 at the corresponding significance level. This is not as accurate as doing a permutation test, but a confidence interval estimates the size of an effect as well as giving some information about its statistical significance.

总结与比较 Summary and Comparison

自助法 vs 置换检验 Bootstrap vs Permutation Test

方面 Aspect	自助法 Bootstrap	置换检验 Permutation Test
主要目的	估计统计量的抽样分布，构建置信区间。	在原假设下检验假设，计算 P 值。
重抽样方式	从原始样本中有放回地抽样。	以符合原假设的方式重新排列数据（通常是无放回地重新分配组别标签）。
对原假设的依赖	不依赖于原假设；基于“原始样本代表总体”的假设。	完全在原假设下进行；重抽样必须模拟 H_0 为真时的数据生成过程。
输出结果	标准误、偏差估计、置信区间。	P 值。
优势	适用性广，可处理各种统计量；对总体分布假设要求低。	提供精确的 P 值（当所有排列被枚举时）；不依赖于正态性等假设。
局限性 Limitations	小样本时不可靠；计算量大。	仅适用于特定假设检验场景；原假设要求强（如两总体完全相同）；计算量可能非常大.
典型应用 Typical Applications	均值、中位数、比例、相关系数、回归系数等的置信区间。	两样本位置检验、相关性检验、配对差异检验。

表 1: 自助法与置换检验的比较
Comparison of Bootstrap and Permutation Test

重抽样方法与传统方法的比较 Resampling vs Traditional Methods

- **共同优点：**减少对严格分布假设（如正态性）的依赖；借助现代计算能力，概念直观。
  Common advantages: Reduce reliance on strict distributional assumptions (e.g., normality); conceptually intuitive with modern computing power.
- **传统方法的优势：**有成熟的理论基础；计算快速；对于符合假设的数据，通常功效更高；可以扩展到复杂的模型（如 ANOVA，多元回归）。
  Advantages of traditional methods: Well-established theoretical foundation; fast computation; typically more powerful when assumptions hold; extendable to complex models (e.g., ANOVA, multiple regression).

- **选择建议：**
  - 当数据明显违背传统方法的假设（如严重偏斜、异常值）且样本量不大时，考虑使用重抽样方法。
    Consider resampling when data clearly violate assumptions of traditional methods (e.g., severe skewness, outliers) and sample size is not large.
  - 对于没有简单理论分布的统计量（如比率、截尾均值），自助法是构建置信区间的有力工具。
    For statistics with no simple theoretical distribution (e.g., ratio, trimmed mean), bootstrap is a powerful tool for constructing confidence intervals.
  - 当想检验更一般的假设（如分布是否相同）或检查传统检验的稳健性时，置换检验很有用。
    Permutation test is useful when you want to test more general hypotheses (e.g., whether distributions are identical) or check the robustness of traditional tests.
  - 在报告结果时，可以同时给出传统方法和重抽样方法的结果，以增强结论的可信度。
    When reporting results, consider presenting both traditional and resampling results to strengthen the credibility of conclusions.

实践建议 Practical Recommendations

1. **样本量：**确保样本量足够大，以使重抽样方法可靠。对于自助法，通常 $n \geq 30$ 可提供合理的结果；对于置换检验，每组至少 10-20 个观测值。
   Ensure sample size is large enough for resampling methods to be reliable. For bootstrap, typically $n \geq 30$ gives reasonable results; for permutation tests, at least 10-20 observations per group.

2. **重抽样次数：**对于自助法，通常 $B = 1000$ 次或更多以获得稳定的标准误和百分位数；对于置换检验，如果无法枚举所有排列，使用 $B = 1000$ 或更多的蒙特卡洛重复。
   Number of resamples: For bootstrap, typically $B = 1000$ or more for stable standard errors and percentiles; for permutation tests, if enumerating all permutations is infeasible, use $B = 1000$ or more Monte Carlo replicates.

3. **检查自助法分布：**始终绘制自助法分布图，检查其形状、偏差和离散度。如果分布严重偏斜或有明显偏差，考虑使用 BCa 区间。
   Always plot the bootstrap distribution to examine its shape, bias, and spread. If the distribution is severely skewed or has notable bias, consider using BCa intervals.

4. **比较不同区间：**如果可能，计算并比较自助法 t 区间、百分位数区间和 BCa 区间。如果它们差异很大，则需谨慎解释，并考虑使用最保守的区间或寻找其他方法。
   Compare different intervals if possible: bootstrap t, percentile, and BCa. If they differ substantially, interpret with caution and consider using the most conservative interval or seeking alternative methods.

5. **软件实现：**大多数统计软件（如 R, Stata, Python 的 statsmodels 和 scikit-learn）都提供了重抽样方法的实现。熟悉你所用软件的相关函数。
   Most statistical software (e.g., R, Stata, Python's statsmodels and scikit-learn) provide implementations of resampling methods. Familiarize yourself with relevant functions in your software.

本章核心要点 Core Takeaways

- **重抽样方法的核心思想是用计算代替理论，通过从数据中重复抽样来近似抽样分布。**
  The core idea of resampling methods is to replace theory with computation, approximating the sampling distribution by repeatedly sampling from the data.

- **自助法主要用于估计抽样分布和构建置信区间，对总体分布假设要求低，但小样本时需谨慎。**
  Bootstrap is mainly used to estimate sampling distributions and construct confidence intervals, with minimal distributional assumptions, but caution is needed with small samples.

- **置换检验是一种非参数假设检验方法，在原假设下通过重新排列数据来获得 P 值，适用于两样本比较、相关性检验等场景。**
  Permutation test is a nonparametric hypothesis testing method that obtains P-values by rearranging data under the null hypothesis, suitable for two-sample comparisons, correlation tests, etc.

- **BCa 置信区间通常比简单的百分位数区间更准确，因为它考虑了偏差和偏度。**
  The BCa confidence interval is generally more accurate than the simple percentile interval because it accounts for bias and skewness.

- **当传统参数检验的假设可能被违反时，重抽样方法提供了稳健的替代方案。在应用研究中，将它们与传统方法结合使用可以增强结论的说服力。**
  When assumptions of traditional parametric tests are likely violated, resampling methods provide robust alternatives. In applied research, using them

alongside traditional methods can strengthen the persuasiveness of conclusions.