Statistical Learning and Data Mining

Module 6: Resampling Methods

Semester 2, 2017

Discipline of Business Analytics, The University of Sydney Business School

Module 6: Resampling Methods

1. Introduction
2. Resampling
3. The Bootstrap

Introduction

Model inference

Uncertainty quantification is an essential component of statistical analysis and decision making.

In this module we will study several tools for model inference. The next slides introduce and review some of the required concepts for this module.

Notation

We use the following notation throughout:

- Let $p(\boldsymbol{y}; \boldsymbol{\theta})$ denote a probability mass function or density with associated parameter vector $\boldsymbol{\theta}$.
- $Y_1, Y_2, \ldots, Y_n$ is random sample from this distribution. The random variables are independent.
- $\mathcal{D} = \{y_1, \ldots, y_n\}$ are the actual observed values.
- $\widehat{\boldsymbol{\theta}}(\mathcal{D})$ is an estimator of $\boldsymbol{\theta}$.
- $\widehat{\boldsymbol{\theta}}$ an estimator (as above) or estimate of $\boldsymbol{\theta}$ according to the context.

Empirical distribution

Given a set of data $\mathcal{D} = \{y_1, \ldots, y_n\}$, we define the **empirical distribution** as

$$p(A) = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}(A),$$

where $\delta_y(A)$ is the **Dirac measure**

$$\delta_y(A) = \begin{cases} 1 & \text{if } y \in A \\ 0 & \text{if } y \notin A \end{cases}.$$

The empirical distribution assigns probability $1/n$ to each point in the data, and probability zero to each point not in the data.

Percentiles and quantiles (key concept)

Consider a sample $\mathcal{D} = \{y_1, \ldots, y_n\}$.

We define the $k-$th **percentile** of the data as the number p_k such that $k\%$ of the sample falls below it. For example, sample median is the 50th percentile.

The **quantile** q_α is the the number such that a fraction $0 < \alpha < 1$ of the sample falls below it. For example, $q_{0.5}$ is the median.

Resampling

Resampling

Resampling is a nonparametric approach to statistical inference based that relies on computational methods instead of standard assumptions and large sample results.

Important examples are **permutation tests**, **bootstrapping**, and **cross validation**.

Resampling methods

There are several reasons for studying resampling methods:

1. Resampling methods are widely applicable, handling complex sampling and estimation algorithms. They allow us to conduct statistical inference when analytical results (formulas) are not available.

2. Resampling methods can lead to more accurate inference since they do not rely on distributional assumptions.

3. Resampling methods can provide a validity check for standard inference ("if $n \geq 30$...").

4. Resampling plays a role in machine learning algorithms such as bagging and random forests.

Resampling methods

We should also keep in mind the following pitfalls:

1. Resampling methods can fail in some cases.
2. Resampling methods rely on assumptions that may not be valid and are easy to overlook (such as independence of observations).
3. Resampling can be computationally expensive.

Permutation test

We introduce and illustrate the concept of resampling with the **permutation test**.

Permutation tests are widely used by companies such as Google for A/B testing.

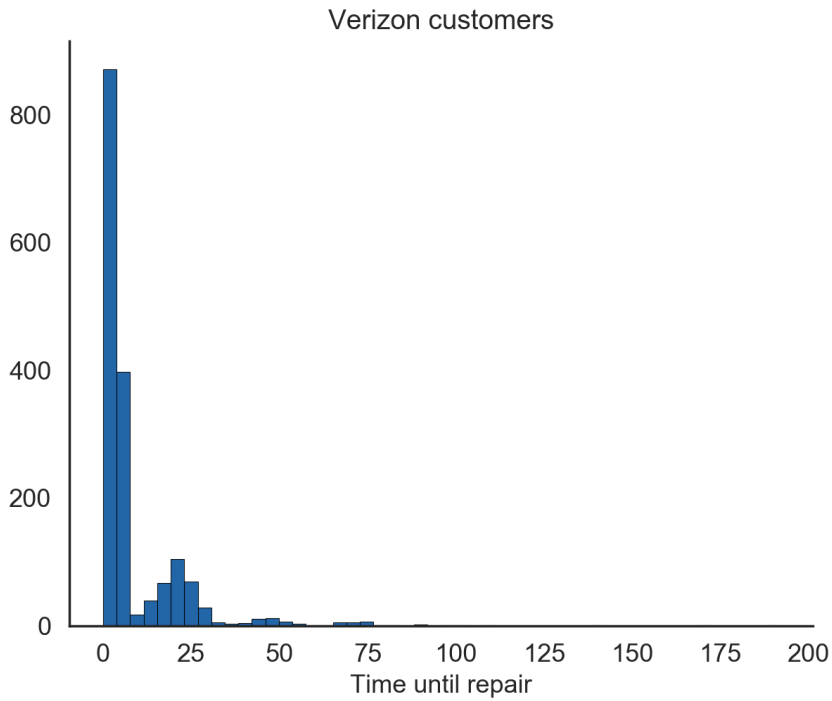
Data: Verizon repair times

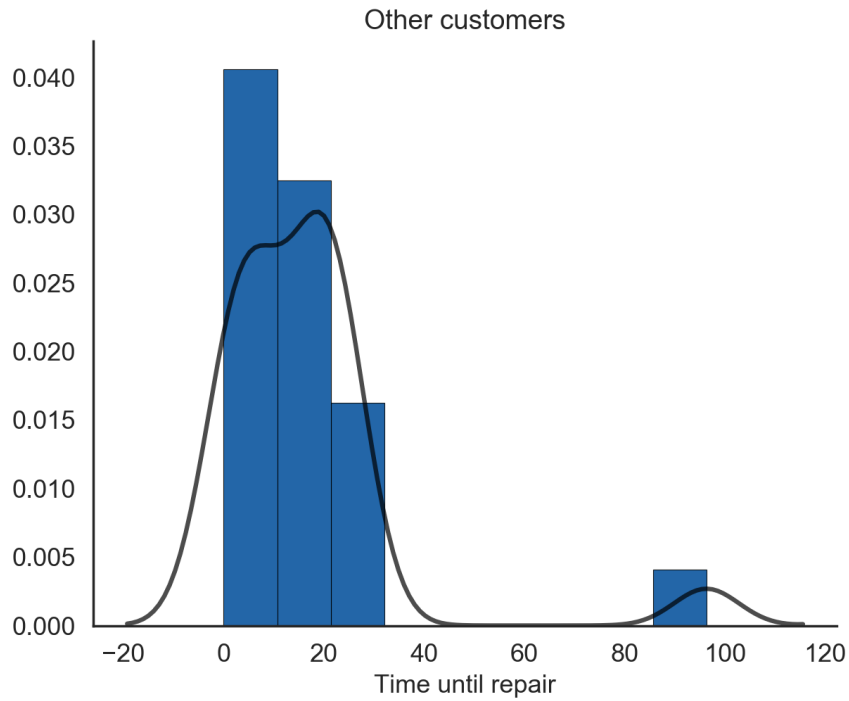
Business case. Verizon is the primary telephone company for a large area of the eastern US. As such, it is responsible for providing repair services for customers of competing companies. Verizon is subject to fines if the repair times are substantially worse for the customers of its competitors.

Data. Repair times for 1664 Verizon customers and 23 customers from competitors.

Reference: Chihara and Hesterberg (2011).

Data: Verizon repair times





Data: Verizon repair times

	Other customers	Verizon
count	23	1664
mean	16.51	8.41
std	19.50	14.69
min	0.00	0.00
25%	5.43	0.73
50%	14.33	3.59
75%	20.72	7.08
max	96.32	191.60

Two sample test

In the Verizon problem, we need to conduct two sample test for equal population means, against the alternative hypothesis that the repair times are longer on average for the population of non-Verizon customers.

$$H_0 : \ \mu_1 = \mu_2$$
$$H_1 : \ \mu_1 > \mu_2$$

The two pooled variance t-test leads to a p-value of 0.0045 for the Verizon data. We contrast this approach with a permutation test.

Permutation

In mathematics, a **permutation** is a rearrangement of an ordered list. For example, consider the list below.

$$(1,\ 2,\ 3,\ 4,\ 5,\ 6)$$

A permutation may be:

$$(2,\ 5,\ 3,\ 1,\ 4,\ 6)$$

Permutation test

Consider the following hypothetical situation.

Scenario. An educator wants to know whether participation in the Maths is Business program leads to better performance in statistical learning. However, we know that looking at averages does not work here since there are omitted variables.

Data. The educator recruits six student volunteers. Three are randomly allocated to take the Maths in Business classes (the treatment group), and three do not take these classes (the control group). The next slides show their marks for statistical learning.

Permutation test

Treatment			Control		
85	**80**	**75**	73	76	77

The average marks were 80 and 75.33 for the treatment and control groups respectively, a difference of 4.67.

Can we reliably say that the difference is due to the program? Or could this result be just random variability? Maybe Maths in Business was "lucky" to get better students?

Permutation test

If the math classes have no benefit, then the split of marks between groups was essentially random (all the subjects come from the same population of students). An equally likely outcome would have been:

Treatment			**Control**		
85	**80**	73	**75**	76	77

Which yields a difference of 3.33.

Permutation test

Treatment			Control		
85	**80**	**75**	73	76	77

There are $\binom{6}{3} = 20$ possible ways to allocate 6 marks into the 3 treatment slots. Of the 20 possible differences in means, 3 are as large or larger than the observed 4.67.

Therefore, the probability that a result as favourable for Maths in Business would occur by chance is $3/20 = 0.15$.

Permutation test

Now, suppose that we have two samples with size m and n (we call them treatment and control respectively).

With larger samples, the number of differences becomes too large to compute exhaustively as we did the previous slide. For example, with $m = n = 20$ there are already more than a 100 billion possible permutation samples.

In this case, we use **Monte Carlo sampling**: we base our results on several random permutations of the data.

Permutation test

A **permutation resample** consists of sampling m observations *without replacement* from the pooled data of $m + n$ observations.

We label these m observations as the treatment group, and assign the remaining n to the control group.

Permutation test

Algorithm Two sample permutation test

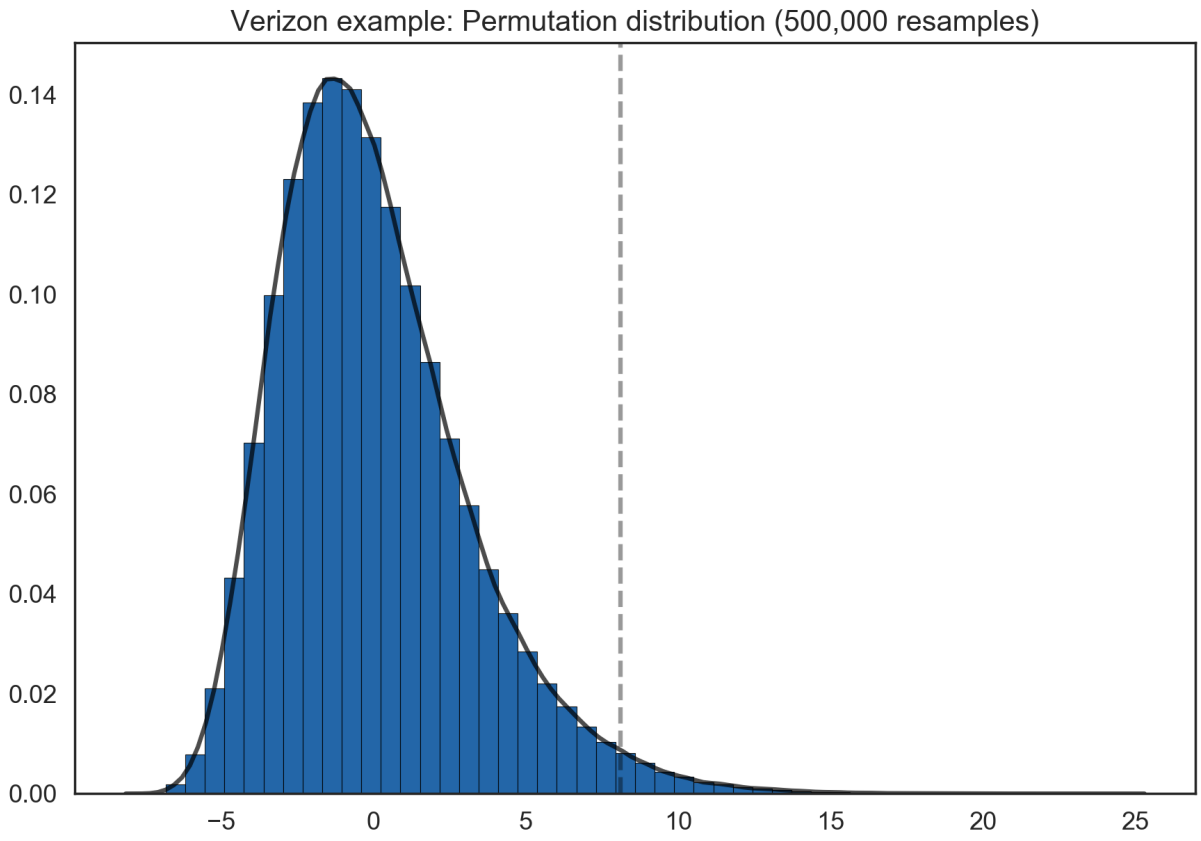
1: Pool the $m + n$ observations.
2: Set the number of replications S.
3: **for** s=1:S **do**
4: Draw a resample of size m without replacement.
5: Assign the remaining n observations to the other sample.
6: Calculate the difference in means (or another statistic).
7: **end for**
8: Compute the p-value as the fraction of resamples in which the random statistic exceeds the statistic calculated for the original sample. Multiply by two for a two-sided test.
9: Optionally, plot a histogram of the random statistics values.

Permutation test

Assumption: The two sample permutation test assumes that the two groups have the same distribution under the null hypothesis.

The test is usually robust to deviations from this assumption.

Verizon data: permutation distribution



Verizon example: conclusion

- The p-value for the permutation test is 0.018, four times as high as the large sample test based on the t distribution.
- The reason for this discrepancy is the skewness in the data. The sample sizes are not sufficient for a large sample approximation to be reliable.
- In contrast, the permutation distribution reflects the impact of skewness.

The Bootstrap

The Bootstrap

The **Bootstrap** is computational method for approximating the sampling distribution of a statistic by resampling the data and recomputing the statistic several times.

The name comes from the expression "pull yourself up by your bootstraps", which means to succeed on one's own effort or skills.

Sampling distribution (review)

- Draw many different datasets $\mathcal{D}_s$ $(s = 1, \ldots, S)$ of size n from the population $p(\boldsymbol{y}; \boldsymbol{\theta})$.
- Apply the estimator $\widehat{\boldsymbol{\theta}}(\cdot)$ to each dataset and obtain a set of estimates $\{\widehat{\boldsymbol{\theta}}_s\}_{s=1}^{S}$.
- The sampling distribution of $\widehat{\boldsymbol{\theta}}(\mathcal{D})$ is the induced distribution on $\widehat{\boldsymbol{\theta}}_s$ as $S \to \infty$.

The Bootstrap idea (key concept)

The idea behind bootstrapping is that the sample approximates the population from which it was drawn.

Resampling from the data approximates drawing samples from the population. Therefore, the sampling distribution of a statistic over the bootstrap samples approximates the sampling distribution of the statistic.

By analogy, **the population is to the sample as the sample is to the bootstrap samples**.

Bootstrap distribution (key concept)

Algorithm Bootstrap distribution

1: Set the number of replications S.

2: **for** $s = 1 : S$ **do**

3: Sample n observations with replacement from the original data $\mathcal{D}$ to obtain the bootstrap sample $\mathcal{D}_s^*$.

4: Compute the estimate $\widehat{\boldsymbol{\theta}}_s^* = \widehat{\boldsymbol{\theta}}(\mathcal{D}_s^*)$ (or another statistic of interest).

5: **end for**

6: The bootstrap distribution is the empirical distribution of $\{\widehat{\boldsymbol{\theta}}_s^*\}_{s=1}^S$ (or the empirical distribution of the statistic of the interest).

The Bootstrap

To estimate the distribution of $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$, we use the distribution of $\widehat{\boldsymbol{\theta}}^*_s - \widehat{\boldsymbol{\theta}}$.

The Bootstrap

The **bootstrap standard error** of a statistic is the standard deviation of the bootstrap distribution of that statistic.

$$\mathsf{SE}_{\mathsf{boot}}(\widehat{\boldsymbol{\theta}}) = \sqrt{\frac{\sum_{s=1}^{S}(\widehat{\boldsymbol{\theta}}_s^* - \overline{\boldsymbol{\theta}^*})^2}{S-1}},$$

where

$$\overline{\boldsymbol{\theta}^*} = \frac{\sum_{s=1}^{S}\widehat{\boldsymbol{\theta}}_s^*}{S}.$$

The **bootstrap bias estimate** is the mean of the bootstrap distribution minus the original sample statistic.

$$\widehat{\mathsf{bias}}_{\mathsf{boot}}(\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{\theta}} - \overline{\boldsymbol{\theta}^*}.$$

How useful is the bootstrap distribution? (key concept)

Center. The bootstrap distribution is *not* an accurate approximation to the center of the sampling distribution. The center of the bootstrap distribution is approximately the original value of the statistic.

Spread. The spread of bootstrap distribution reflects the spread the sampling distribution.

Skewness. The skewness of bootstrap distribution reflects the spread the sampling distribution.

Bias. The bootstrap bias estimate reflects the bias of the sampling distribution (but it is typically a noisy estimate).

Bootstrap percentile interval (key concept)

The interval between the $q_{\alpha/2}$ and $q_{1-\alpha/2}$ quantiles of the bootstrap distribution of a statistic is a $100 \times (1-\alpha)\%$ **bootstrap percentile confidence interval** for that statistic.

Bootstrap t confidence interval (key concept)

For each bootstrap sample, calculate the t statistic

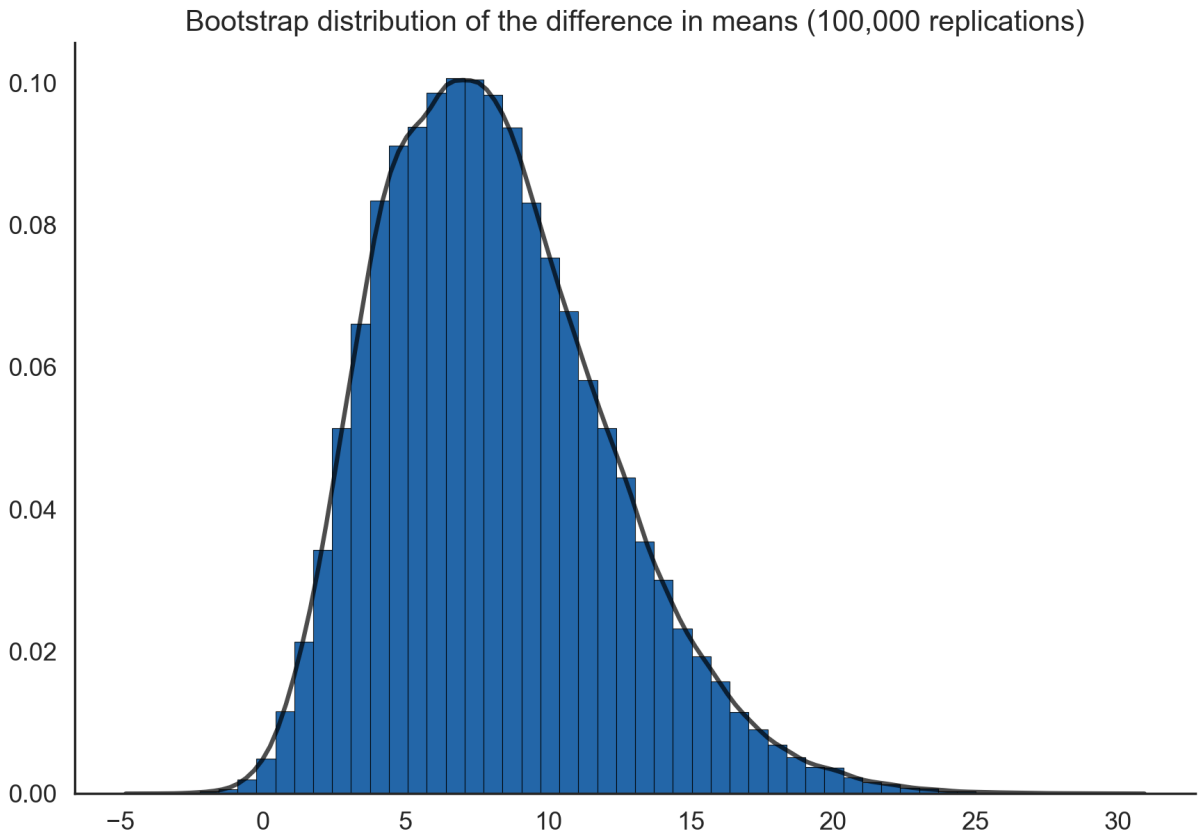
$$t^* = \frac{\widehat{\theta}^* - \widehat{\theta}}{\mathsf{SE}(\widehat{\theta}^*)}$$

Let Q_1^* and Q_2^* be $\alpha/2$ and $(1-\alpha/2)$ empirical quantiles. The **bootstrap t confidence interval** is

$$\left(\widehat{\theta} - Q_2^* \times \mathsf{SE}(\widehat{\theta}),\, \widehat{\theta} - Q_1^* \times \mathsf{SE}(\widehat{\theta})\right)$$

Note that $\mathsf{SE}(\widehat{\theta})$ refers to a (formula based) standard error for the original sample.

Example: Verizon data



Example: Verizon data

95% confidence intervals:

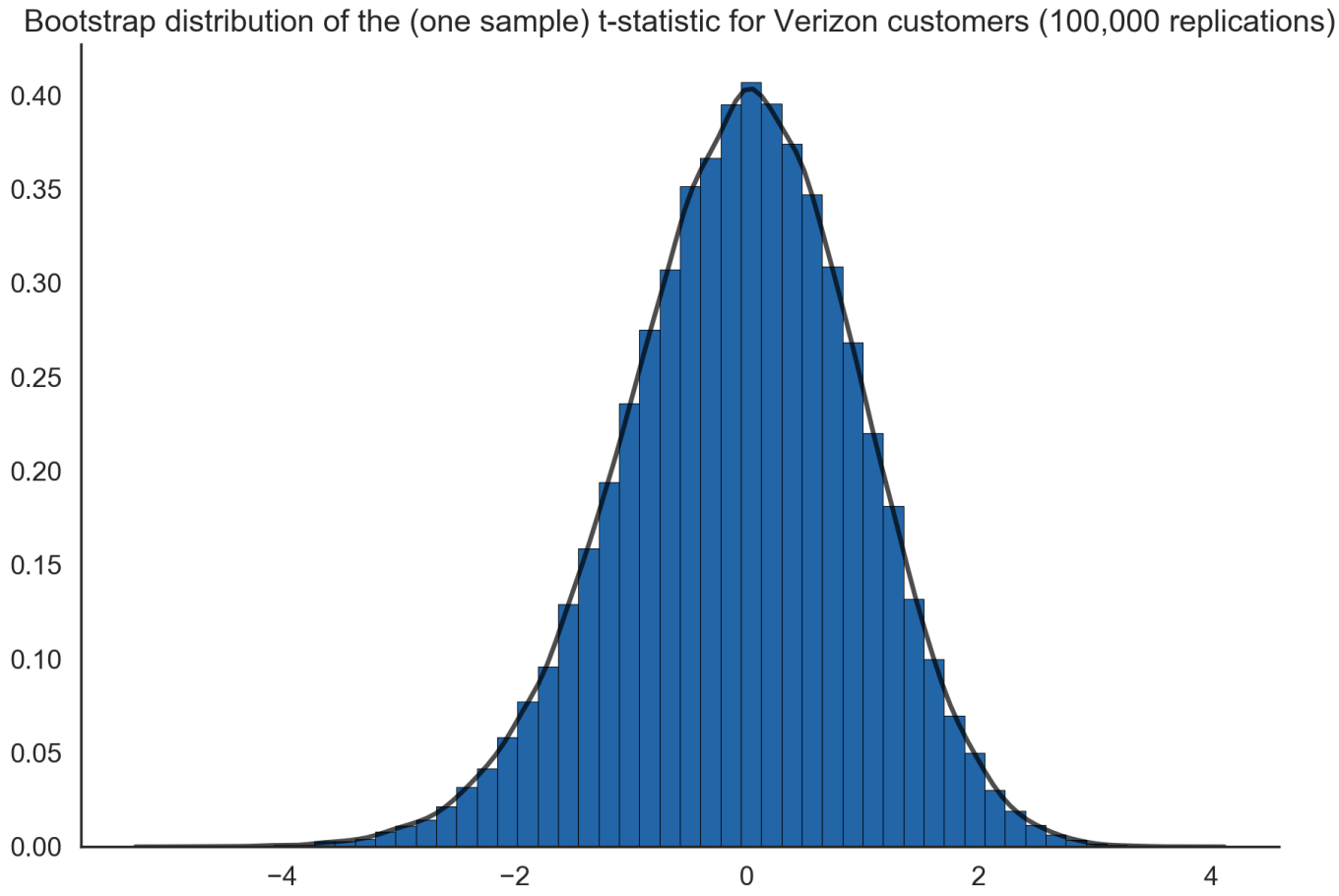
- Formula interval: $(0.4, 16.6)$
- Bootstrap percentile interval: $(1.7, 17.0)$
- Bootstrap t interval: $(2.2, 22.4)$

Example: Verizon data

In the two-sample test, we may blame the low number of observations in the sample of other customers ($n = 23$) for the failure of the formula interval.

The next slide shows what we find when we bootstrap the t statistic for the sample of Verizon customers ($n = 1664$).

Example: Verizon data



Often, we should not trust large sample approximations with skewed data.

Bootstrap confidence intervals

- Bootstrap percentile intervals tend to be too narrow, particularly in small samples. They tend to be less accurate than common t intervals for small samples, though more accurate in larger samples.

- Bootstrap t intervals (if available) are the most accurate among the basic bootstrap intervals, and more accurate than standard intervals.

- These intervals may perform poorly for statistics that depend on a small number of observations, such as the sample median and other quantiles.

- More complicated adjustments to bootstrap intervals are available and improve accuracy.

Bootstrapping regression models (key concept)

There are two ways of bootstrapping regression models.

Bootstrapping the observations. Directly resampling the observations follows the same procedure as before. In this design, we treat the predictors as *random*.

Bootstrapping the residuals. In this case, we treat the predictors as fixed (as in standard inference) and resample only the residulals.

Bootstrapping regression models

Bootstrapping the residuals works as follows:

1. Estimate the regression coefficients on $\mathcal{D}$ to obtain the fitted values $\widehat{y}_i = \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_{ij}$ and residuals $e_i = y_i - \widehat{y}_i =$ for $i = 1, \ldots, n$.

2. For each bootstrap replication s, sample n values e^*_{si} with replacement from the original residuals. Construct the response values for the bootstrap sample as $y^*_{si} = \widehat{y}_i + e^*_{si}$, $i = 1, \ldots, n$.

3. Regress the constructed response values $\boldsymbol{y}^*_s = (y^*_{s1}, \ldots, y^*_{sn})$ on the fixed design matrix $\boldsymbol{X}$.

4. The coefficients $\widehat{\boldsymbol{\beta}}^*_s$ are the estimates from this regression.

How many bootstrap replications?

- The results from bootstrapping are subject to **Monte Carlo error**.
- The larger the number of replications S, the better. But we need to balance this against the computational cost.
- Typically, at least 10^4 bootstrap samples are required for good accuracy. Fewer samples provide rough estimates.

Review questions

- What is the k-th percentile of a sample?
- What is sampling with and without replacement?
- What is resampling?
- What is a permutation test?
- What is the key idea behind the bootstrap method?
- What is a bootstrap sample?
- How do we obtain bootstrap confidence intervals?
- What are the advantages and disadvantages of the bootstrap relative to the large sample approximations?

The Jackknife

The **Jackknife** is a resampling method that allows us to estimate the bias and variance of an estimator by averaging over leave-one-out samples. We introduce it here as it is very useful in some settings.

1. For each observation $i = 1, \ldots, n$, compute the estimate $\widehat{\theta}_{-i}$ based on the leave-one-out sample $y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n$.

2. Compute $\widehat{\theta}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\theta}_{-i}$, the sample average of the leave one out replicates.

3. The Jackknife standard error is

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

The Jackknife

1. For each observation $i = 1, \ldots, n$, compute the estimate $\widehat{\theta}_{-i}$ based on the leave-one-out sample $y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n$.

2. Compute $\widehat{\theta}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\theta}_{-i}$, the sample average of the leave one out replicates.

3. The Jackknife standard error is

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

The Jackknife

When discussing model assessment, we introduced the MSE as a measure of performance:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2,$$

where $\widehat{y}_i$ are the predictions. In addition, we often want to estimate the variability of the MSE.

The Jackknife

Let $z_i = (y_i - \widehat{y}_i)^2$, so that MSE $= \overline{z}$. It follows from basic statistics that

$$\text{SE(MSE)} = \frac{s_z}{\sqrt{n}},$$

where s_z is the sample standard deviation

$$s_z = \sqrt{\frac{\sum_{i=1}^{n}(z_i - \overline{z})}{n-1}}.$$

The Jackknife: example

Now, suppose that we want to report the RMSE:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}$$

The result from basic statistics does not apply in this case.

However, we can use the Jackknife estimate

$$\text{SE}(\widehat{\theta})_{\text{jack}} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\text{RMSE}_{-i} - \text{RMSE}_{(\cdot)})^2}$$