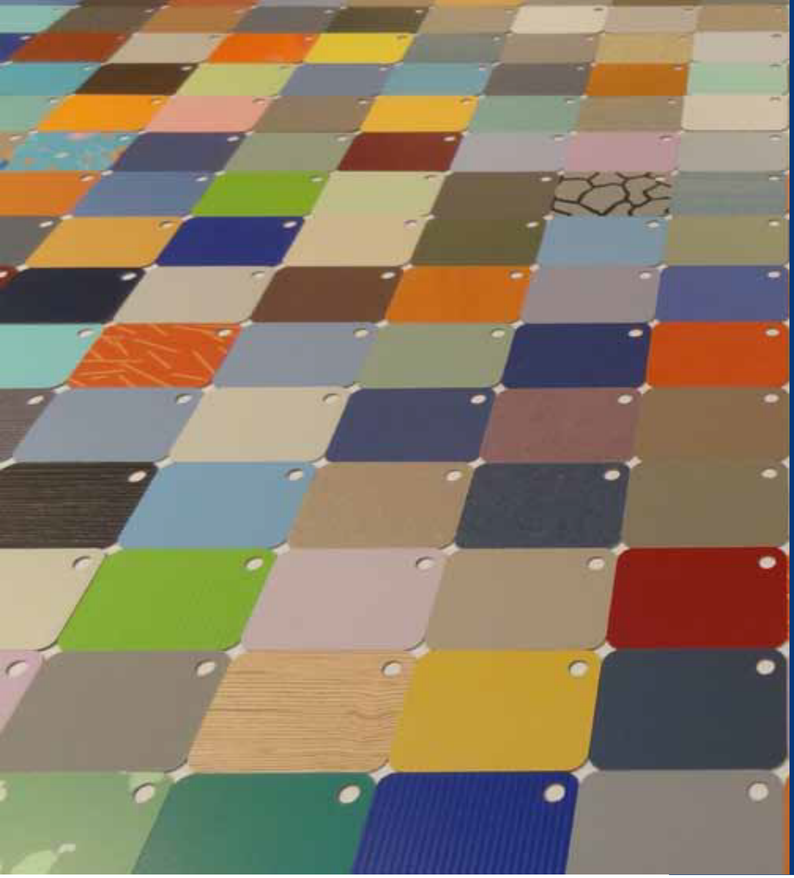


Chapter 16

Bootstrap Methods and Permutation Tests

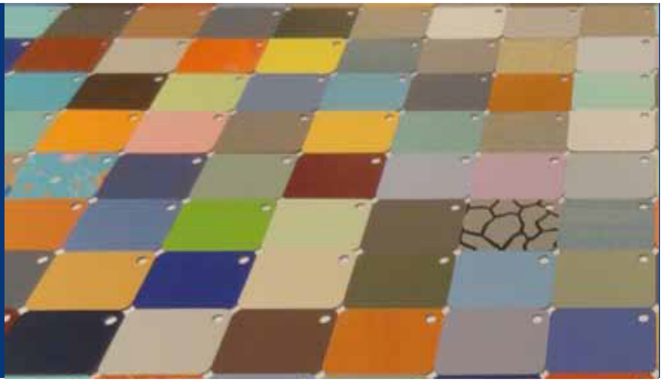
Introduction to the Practice of

STATISTICS EIGHTH EDITION

Moore / McCabe / Craig

Lecture Presentation Slides

Chapter 16 Bootstrap Methods and Permutation Tests



16.1: The Bootstrap Idea

16.2: First Steps in Using the Bootstrap

16.3: How Accurate Is a Bootstrap Distribution?*

16.4: Bootstrap Confidence Intervals

Why Resample?

- **Fewer assumptions:** Resampling methods do not require that distributions be Normal and are less reliant on large sample sizes.
- **Greater accuracy:** Permutation tests and some bootstrap methods are more accurate in practice than classical methods.
- **Generality:** Resampling methods are remarkably similar for a wide range of statistics and do not require new formulas for every statistic.
- **Intelligibility:** Bootstrap procedures promote intuition by providing concrete analogies to theoretical concepts.

Procedure for Bootstrapping

- **Step 1: Resampling.** Create hundreds of new samples, called bootstrap samples or **resamples**, by sampling *with replacement* from the original random sample. Each resample is the same size as the original random sample. **Sampling with replacement** means that, after we randomly draw an observation, we put it back in before drawing the next observation.

- **Step 2: Bootstrap Distribution.** Calculate the statistic for each resample. The distribution of these statistics is called a **bootstrap distribution**. The bootstrap distribution gives information about the shape, center, and spread of the sampling distribution of the statistic.

The Bootstrap Idea

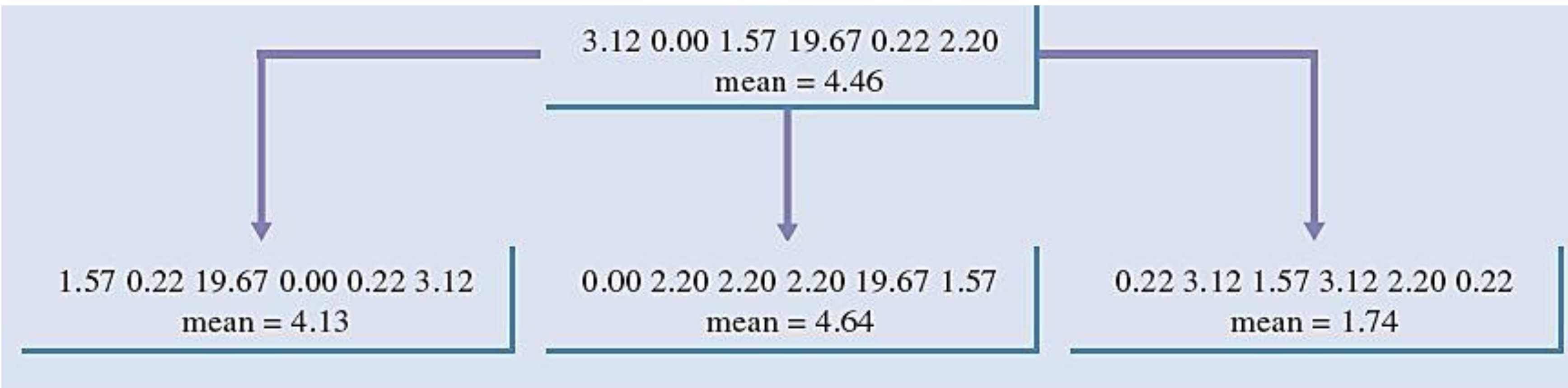
The Bootstrap Idea

The original sample represents the population from which it was drawn. Thus, resamples from this original sample represent what we would get if we took many samples from the population.

The bootstrap distribution of a statistic, based on many resamples, represents the sampling distribution of the statistic.

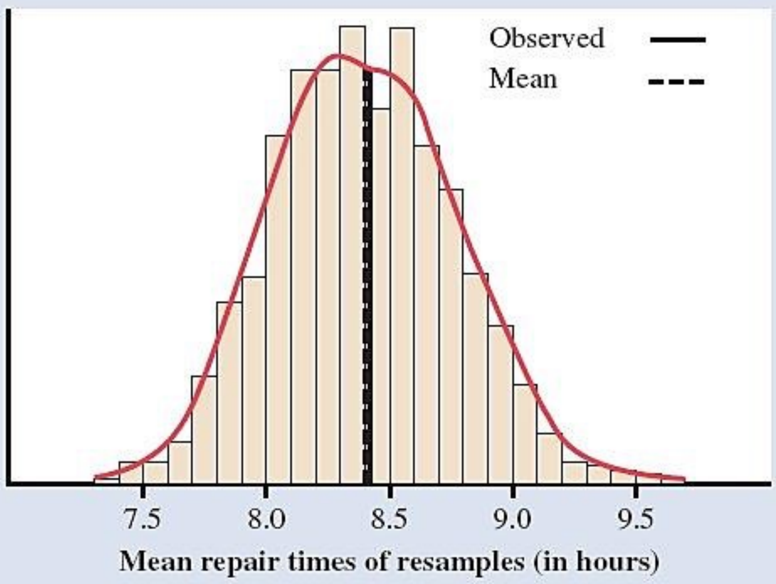
Example

Verizon is the primary local telephone company for a large area in the Eastern United States. It is responsible for providing repair services for the customers of other phone companies. We start with a random sample of 1664 repair times from Verizon's own customers.



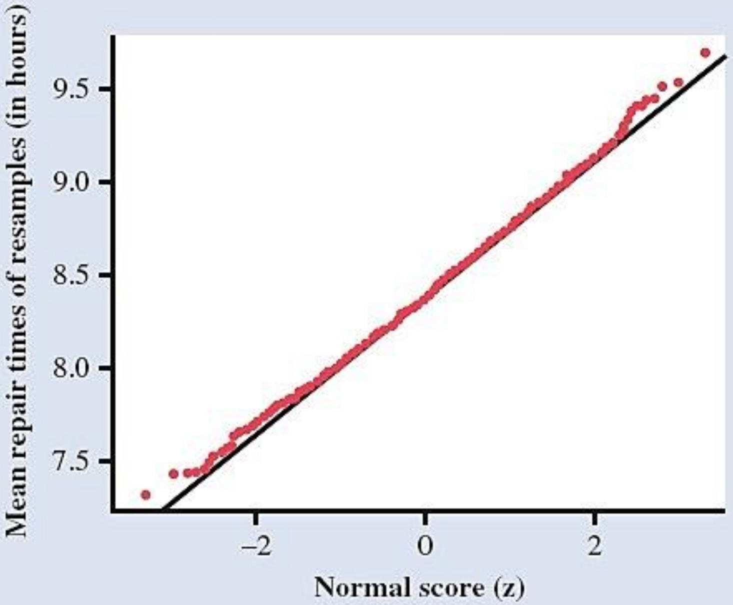
The resampling idea: The top box is a sample of size n = 6 from Verizon customers' repair times. The three lower boxes are three resamples from the sample in the top box. Some values from the original sample occur more than once in the resamples because each resample is formed by sampling *with replacement*. We calculate the statistic of interest—the sample mean in this example—for the original sample and each resample.

Bootstrap Distribution



(a)

The bootstrap distribution for 1000 resample means from a Verizon sample of repair times. The solid line in (a) marks the original sample mean, and the dashed line marks the average of the bootstrapped means.

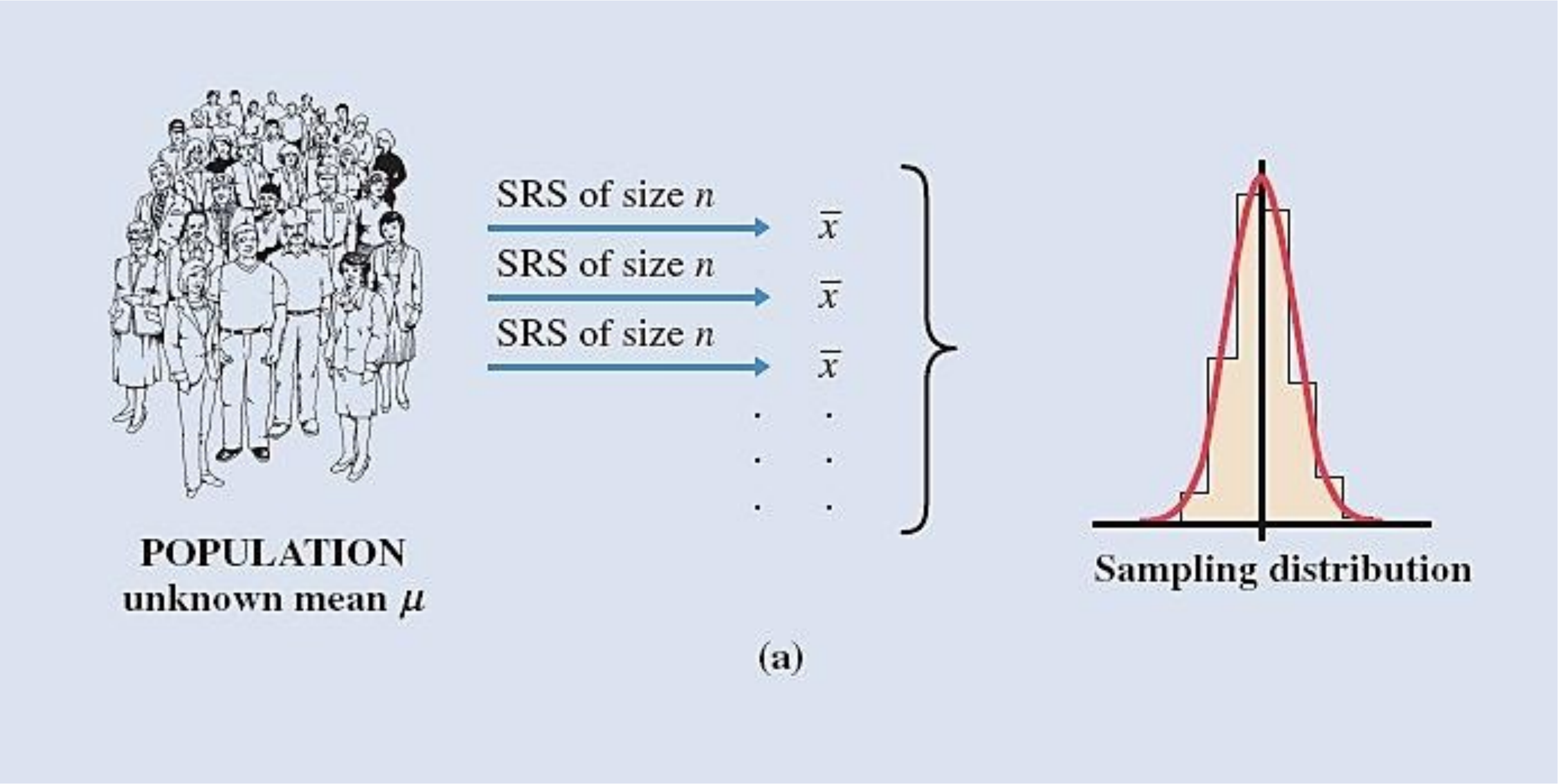


(b)

The Normal quantile plot confirms that the bootstrap distribution is nearly Normal in shape.

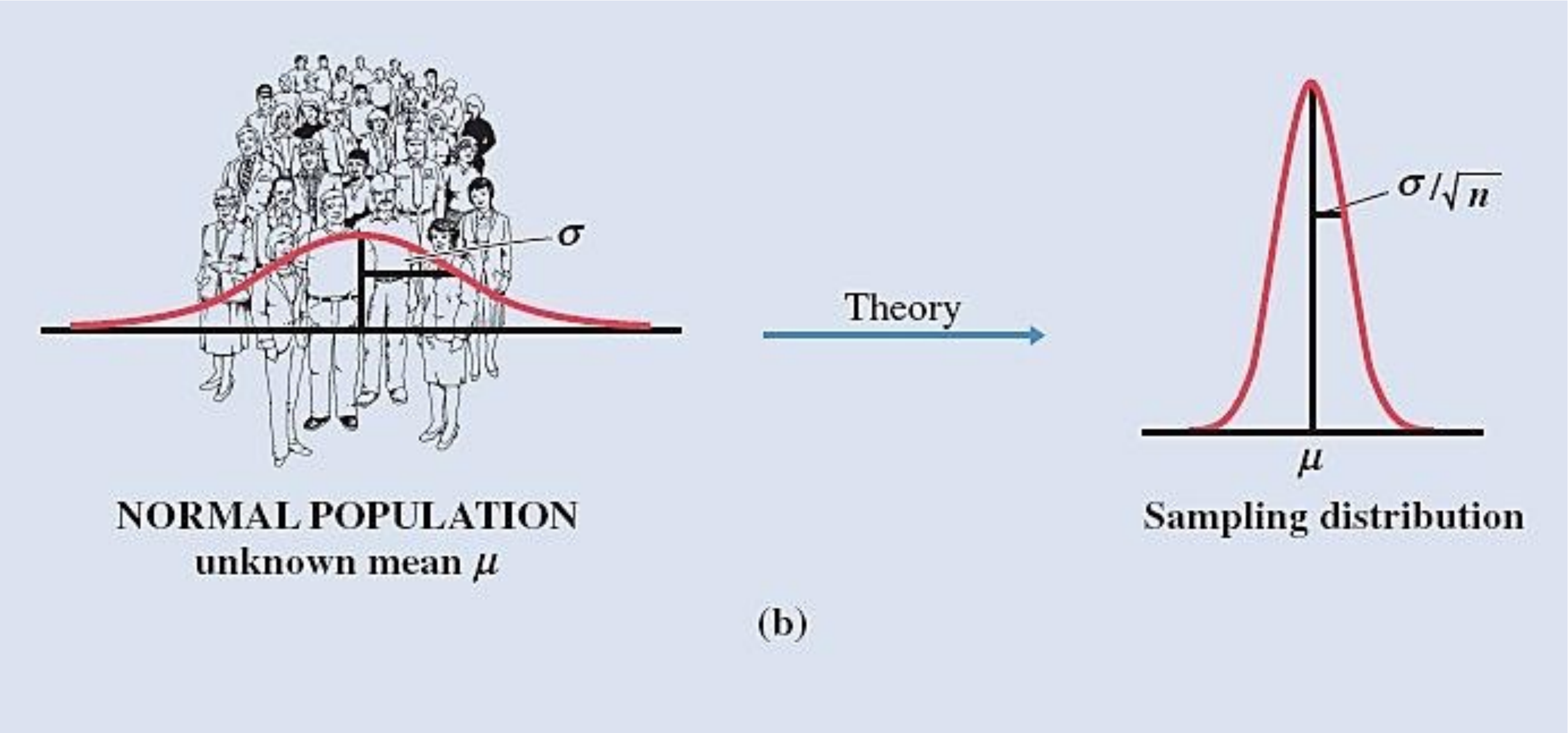
The Bootstrap Intuition

The idea of the sampling distribution of the sample mean *x*-bar: Take many samples from the population, calculate *x*-bar for each sample, and look at the distribution of all the *x*-bar values.



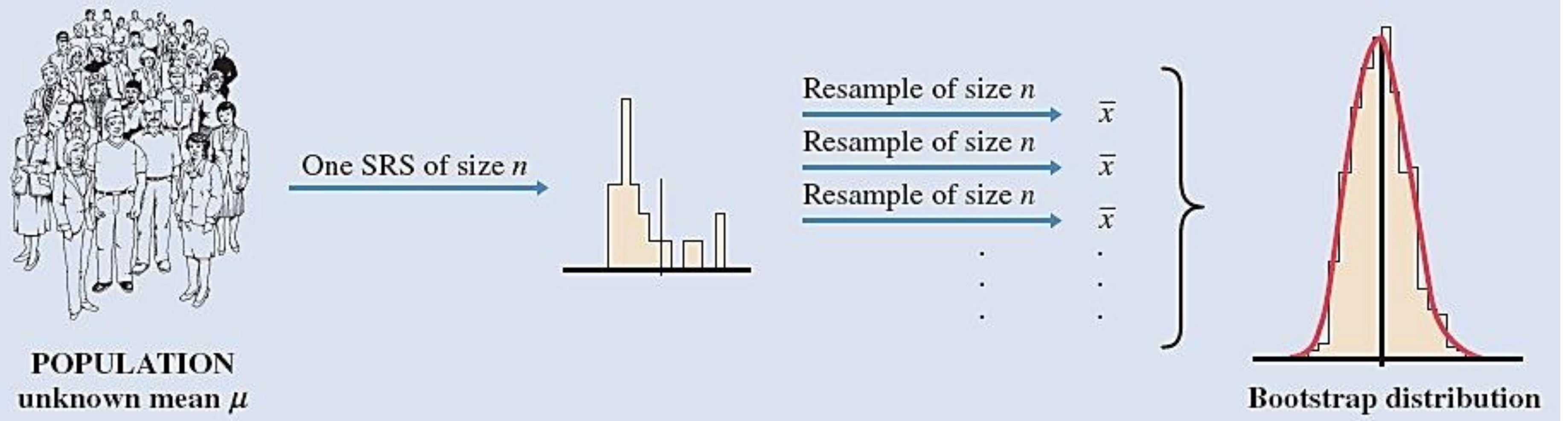
The Bootstrap Intuition

The probability theory shortcut: If we know that the population values follow a Normal distribution, theory tells us that the sampling distribution of *x*-bar is also Normal.



The Bootstrap Intuition

The bootstrap idea: When the theory fails and we can afford only one sample, that sample stands in for the population and the distribution of *x*-bar in many resamples stands in for the sampling distribution.



(c)

Thinking About the Bootstrap Idea

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

- The bootstrap mean and standard error based on *B* resamples are:

$$\text{mean}_{\text{boot}} = \frac{1}{B}\sum_{i=1}^{B} \bar{X}_i^{*}$$

$$\text{SE}_{\text{boot}} = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\bar{X}_i^{*} - \text{mean}_{\text{boot}})^2}$$

The Plug-In Principle

To estimate a parameter, a quantity that describes the population, use the statistic that is the corresponding quantity for the sample.

16.3 How Accurate is a Bootstrap Distribution?*

- Sources of variation among bootstrap distributions
- Bootstrapping small samples
- Bootstrapping a sample median

How Accurate is a Bootstrap Distribution?

Sources of Variation Among Bootstrap Distributions

Bootstrap distributions and conclusions based on them include two sources of random variation:

1. Choosing an original sample from the population.

2. Choosing bootstrap resamples at random from the original sample.

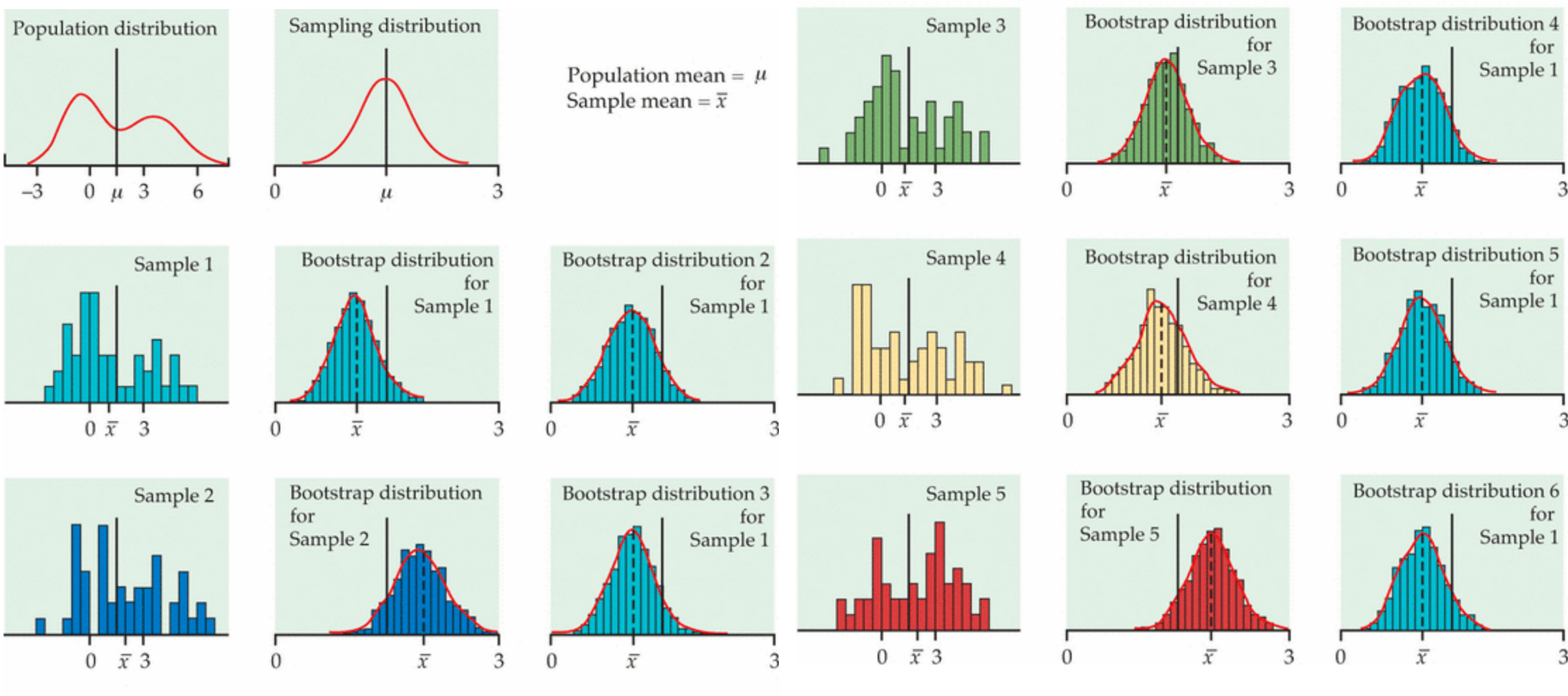
Dealing with Variation in Bootstrap Distributions

For most statistics, almost all the variation in bootstrap distributions comes from the selection of the original sample from the population. You can reduce this variation by using a larger original sample.

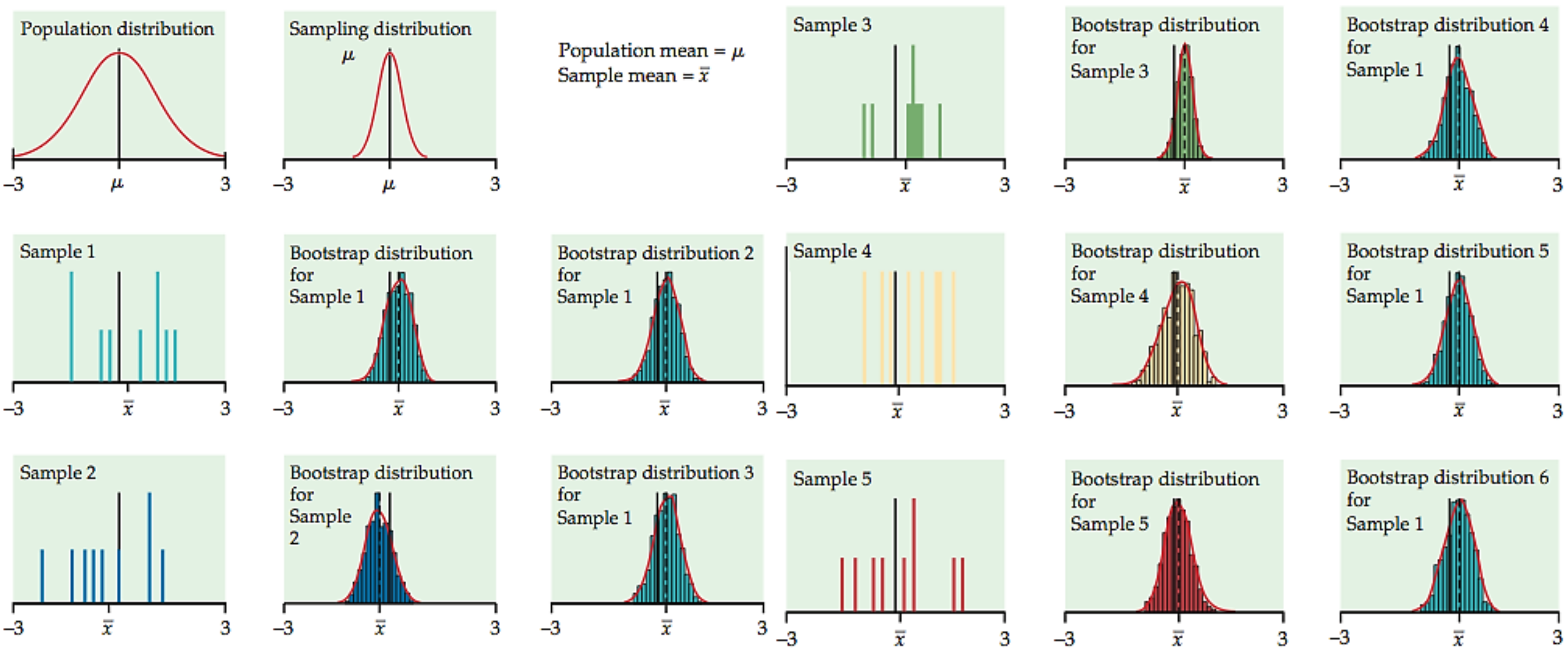
Bootstrapping does not overcome the weakness of small samples as a basis for inference. Use caution in any inference—including bootstrap inference—from a small sample.

The bootstrap resampling process using 1000 or more resamples introduces very little additional variation.

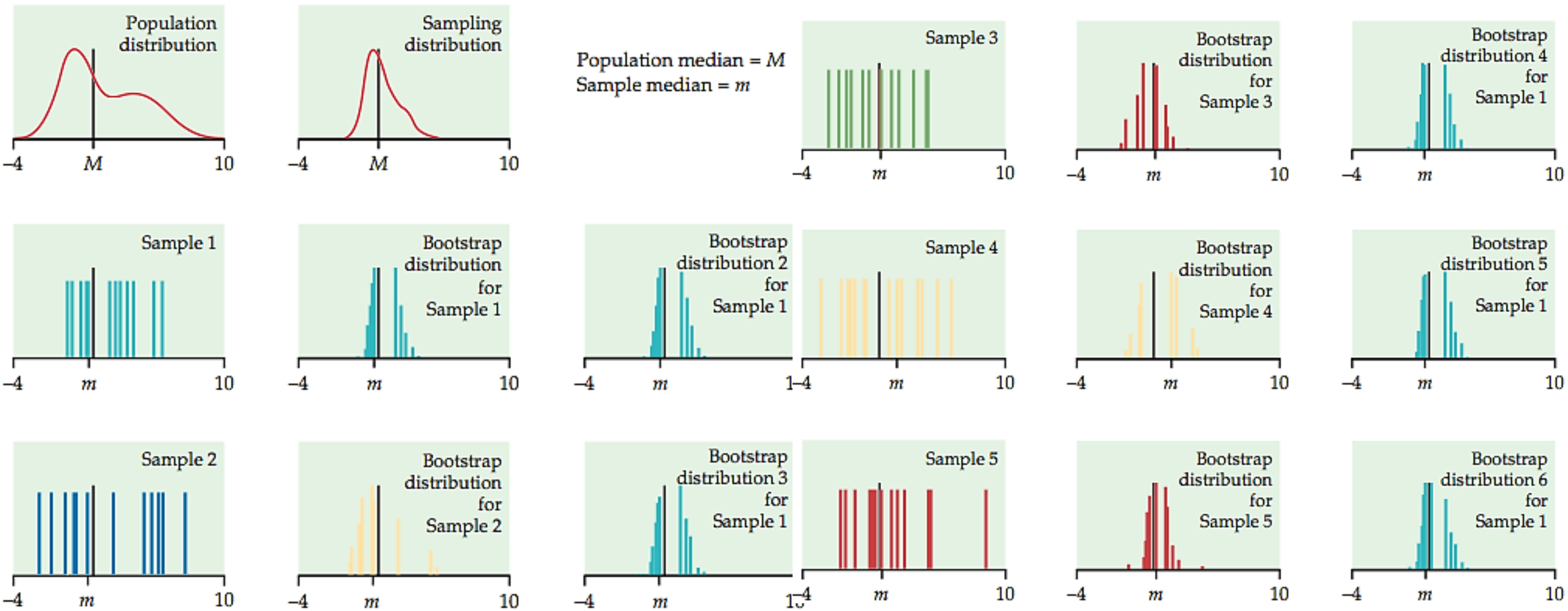
Bootstrapping Large Samples



Bootstrapping Small Samples



Bootstrapping a Sample Median



16.2 Confidence intervals Using the Bootstrap

- Bootstrap *t* confidence intervals
- Bootstrap percentile confidence interval

Bootstrap Distributions and Bias

Bias

A statistic used to estimate a parameter is **biased** when its sampling distribution is not centered at the true value of the parameter. The bias of a statistic is the mean of the sampling distribution minus the parameter.

The bootstrap method allows us to check for bias by seeing whether the bootstrap distribution of a statistic is centered at the statistic of the original random sample. The bootstrap estimate of bias is the mean of the bootstrap distribution minus the statistic for the original data.

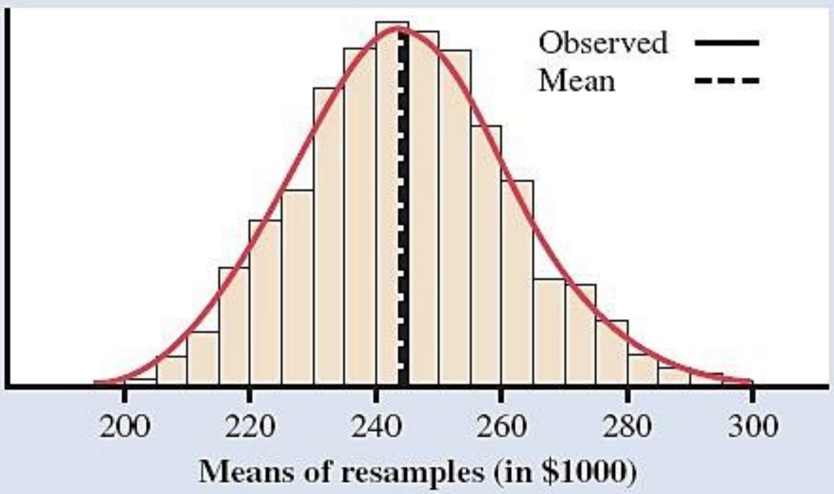
For most statistics, bootstrap distributions approximate the shape, spread, and bias of the actual sampling distribution. One statistic we may consider is:

Trimmed Mean

A **trimmed mean** is the mean of only the center observations in a data set. In particular, the 25% trimmed mean $\bar{x}_{25\%}$ ignores the smallest 25% and the largest 25% of the observations. It is the mean of the middle 50% of the observations.

Example: Real Estate Data. The Selling prices (in $1000) for an SRS of 50 Seattle real estate sales in 2002 are given below.

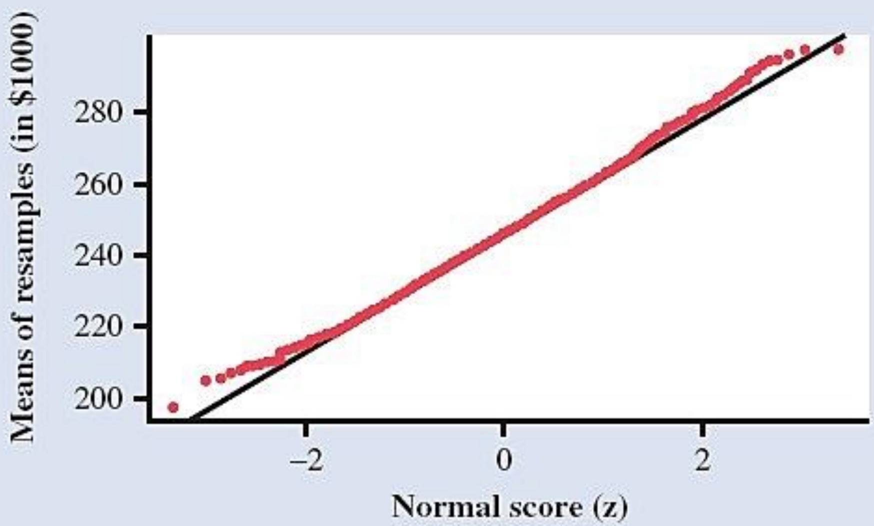
142	232	132.5	200	362	244.95	335	324.5	222	225
175	50	215	260	307	210.95	1370	215.5	179.8	217
197.5	146.5	116.7	449.9	266	265	256	684.5	257	570
149.4	155	244.9	66.407	166	296	148.5	270	252.95	507
705	1850	290	164.95	375	335	987.5	330	149.95	190



(a)

The bootstrap distribution of 25% trimmed means from 1000 resamples. ***Note:*** *the distribution is roughly Normal*.

Below is the Normal quantile plot of the 25% trimmed means of 1000 resamples from the above data.



(b)

Bootstrap *t* Confidence Interval

Bootstrap *t* Confidence Interval

Suppose that the bootstrap distribution of a statistic from an SRS of size n is approximately Normal and that the bootstrap estimate of bias is small. By the plug-in principle, an approximate level C confidence interval for the parameter that corresponds to this statistic is:

$$\text{statistic} \pm t^* \, SE_{\text{boot}}$$

where SE_{boot} is the bootstrap standard error for this statistic and t^* is the critical value of the $t(n-1)$ distribution with area C between $-t^*$ and t^*.

Example

We want to estimate the 25% trimmed mean of the population of all Seattle real estate sales in 2002. We have an SRS of n = 50. Software output shows that the trimmed mean is 244 and the bootstrap standard error is 16.83. A 95% C.I. for the trimmed mean is therefore:

$$\bar{x}_{25\%} \pm t^{*} SE_{boot} = 244 + (2.009)(16.83)$$
$$= 244 + 33.81$$
$$= (210.19,\ 277.81)$$

We are 95% confident that the 25% trimmed mean for the population of all Seattle real estate sales in 2002 is between \$210,190 and \$277,810.

(Use t^* = 2.009 with 50 degrees of freedom since Table D does not have 49 degrees of freedom.)

Bootstrap Confidence Intervals

We have met one type of inference procedure based on resampling: the bootstrap *t* confidence interval. We can check these *t* confidence intervals for accuracy using bootstrap percentiles.

Bootstrap Percentile Confidence Intervals

The interval between the 2.5 and 97.5 percentiles of the bootstrap distribution of a statistic is a 95% **bootstrap percentile confidence interval** for the corresponding parameter.

If the bias of the bootstrap distribution is small and the distribution is close to Normal, the bootstrap *t* and percentile confidence intervals will agree closely. If they do not agree, this is evidence that the Normality and bias conditions are not met. Neither type of interval should be used if this is the case.

Example: The 95% bootstrap *t* C.I. for the 25% trimmed mean of Seattle real estate sales was (210.19, 277.81). Using software, the percentile C.I. is (212.3, 280.3). The two intervals are quite close. This suggests that both intervals are reasonably accurate.

Cautions for the Bootstrap

1. When the bootstrap distribution is non-Normal, we can't trust the bootstrap *t* confidence interval.

2. We can't trust a bootstrap distribution from a very small sample to closely mimic the shape and spread of the actual sampling distribution.

3. Unless you have expert advice or undertake further study, avoid bootstrapping the median and quartiles unless your sample is rather large.

4. When bootstrap *t* and bootstrap percentile intervals do not agree closely, neither type of interval should be used.