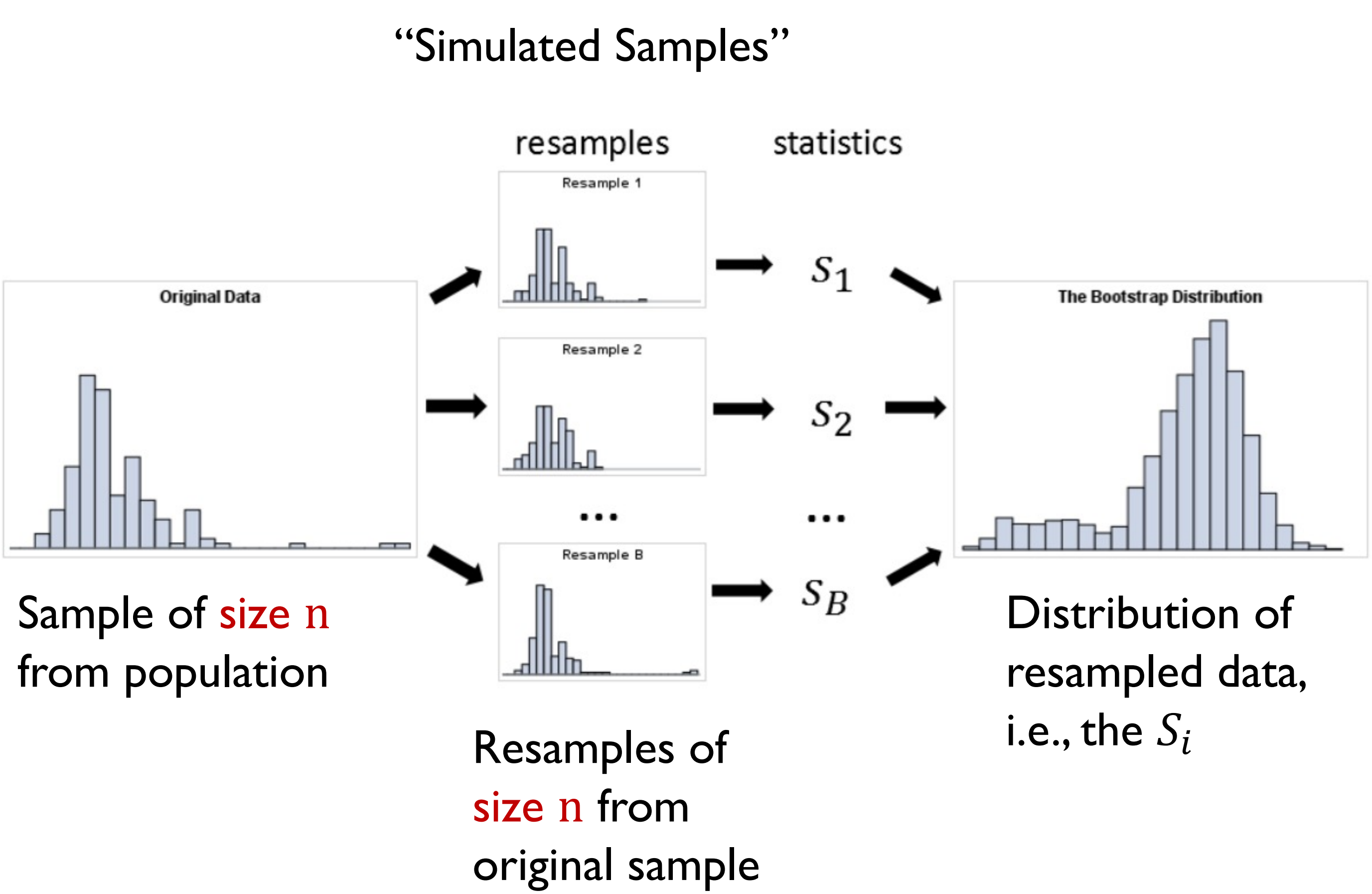
CHAPTER 18 & 23.3

Bootstrapping

What is bootstrapping?

- It is a resampling method that is accomplished by independently sampling, with replacement, from a single existing sample.

- Each resample has the same size n, and then inference is performed using the resampled data distribution.



“Simulated Samples”

resamples

statistics

Sample of size n from population

Resamples of size n from original sample

Distribution of resampled data, i.e., the S_i

The bootstrap method uses a very different approach to estimate sampling distributions. This method takes the sample data that a study obtains, and then resamples it over and over to create many simulated samples.

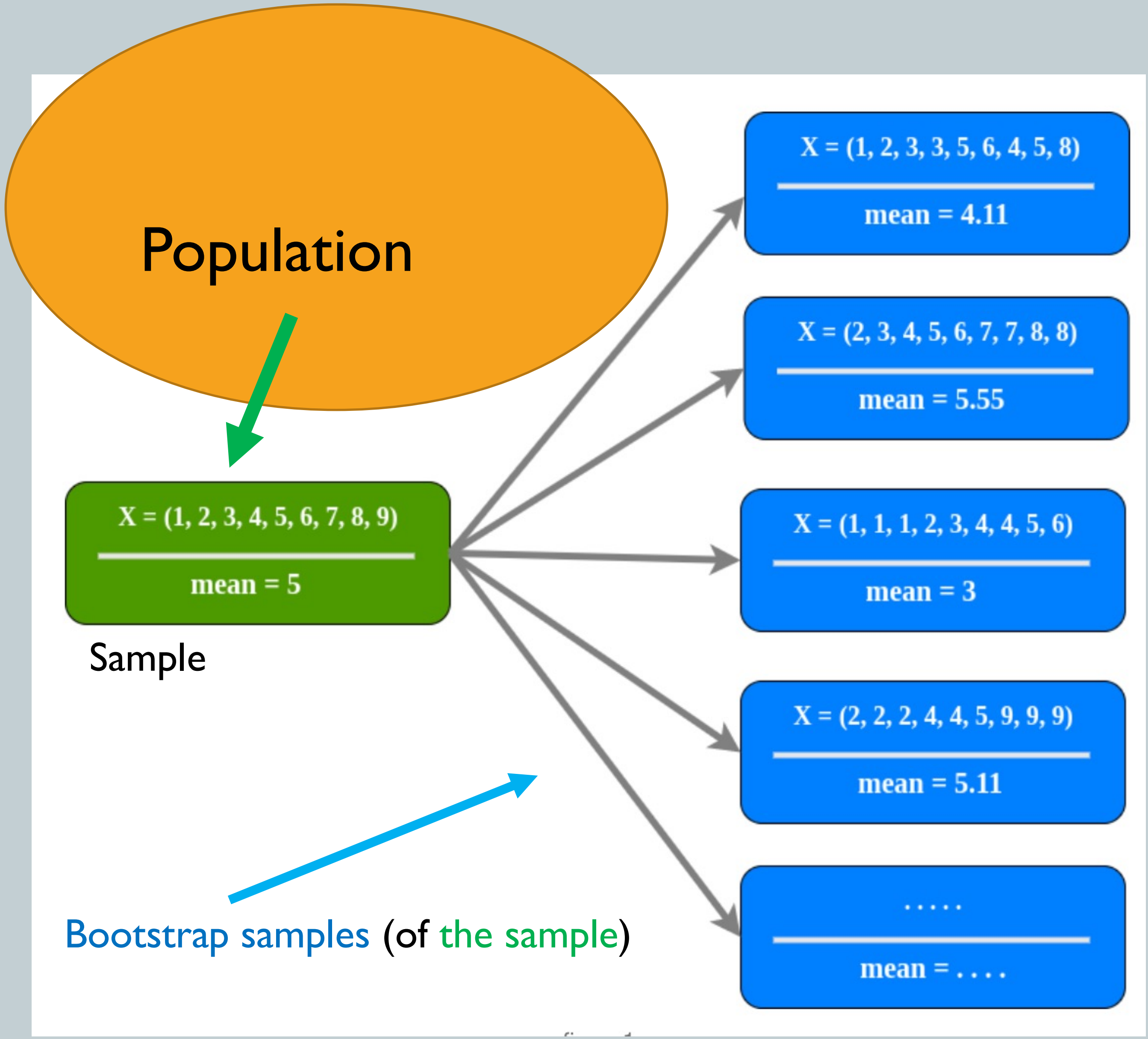
Each of these simulated samples has its own properties, such as the mean. When you graph the distribution of these means on a histogram, you can observe the sampling distribution of the mean. You don’t need to worry about test statistics, formulas, and assumptions.

The bootstrap procedure uses these sampling distributions as the foundation for confidence intervals and hypothesis testing.

Bootstrapping

Bootstrapping is a statistical procedure that resamples a single dataset to create many simulated samples.

These re-samples provide the basis for a distribution of the mean
(or some other statistic)



Population

X = (1, 2, 3, 4, 5, 6, 7, 8, 9)
mean = 5

Sample

X = (1, 2, 3, 3, 5, 6, 4, 5, 8)
mean = 4.11

X = (2, 3, 4, 5, 6, 7, 7, 8, 8)
mean = 5.55

X = (1, 1, 1, 2, 3, 4, 4, 5, 6)
mean = 3

X = (2, 2, 2, 4, 4, 5, 9, 9, 9)
mean = 5.11

. . . . .
mean = ...

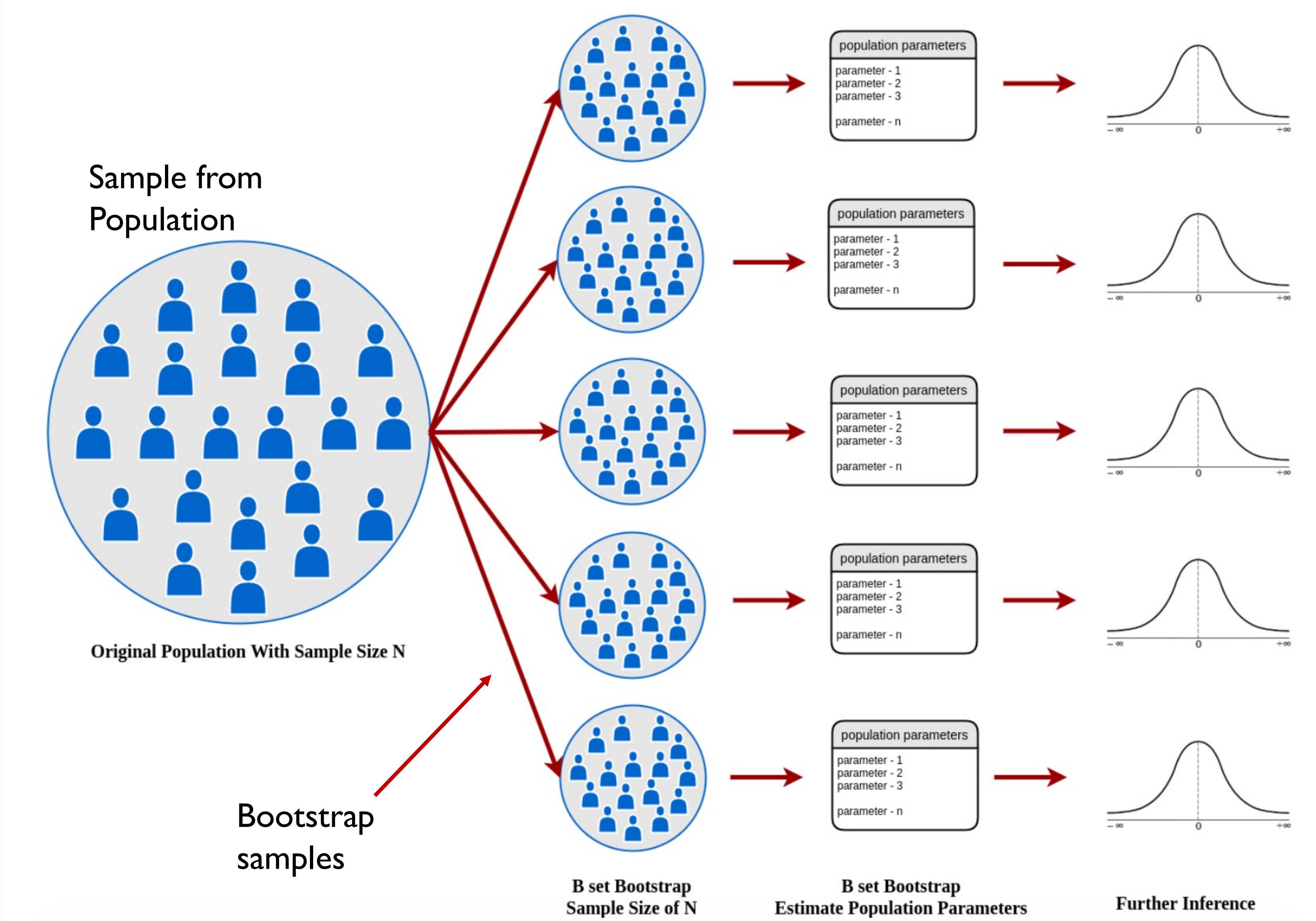
Bootstrap samples (of the sample)

Bootstrapping resamples the original dataset with replacement many thousands of times to create simulated datasets. This process involves drawing random samples from the original dataset.

Here's how it works:

The bootstrap method has an equal probability of randomly drawing each original data point for inclusion in the resampled datasets.

The procedure can select a data point more than once for a resampled dataset. This property is the "with replacement" aspect of the process. The procedure creates resampled datasets that are the same size as the original dataset.



Sample from Population

Bootstrap samples

- This process allows you to calculate standard errors, construct confidence intervals, and perform hypothesis testing for numerous types of sample statistics.

The process ends with your simulated datasets having many different combinations of the values that exist in the original dataset.

Each simulated dataset has its own set of sample statistics, such as the mean, median, and standard deviation.

Bootstrapping procedures use the distribution of the sample statistics across the simulated samples as the sampling distribution.

How well does bootstrapping work?

Resampling involves reusing your one dataset many times. It almost seems too good to be true! In fact, the term “bootstrapping” comes from the impossible phrase of pulling yourself up by your own bootstraps!

However, using the power of computers to randomly resample your one dataset to create thousands of simulated datasets produces meaningful results.

Here is a simple example that illustrates the properties of bootstrap samples.
The resampled datasets are the same size as the original dataset and only contain values that exist in the original set.

Population from which we draw the Original Sample	$\{\text{Some set of values}\}$	
Original Sample from the population:	$\{1, 2, 3, 4, 5\}$	$n = 5, \bar{x} = 3$
Bootstrap Sample #1 from the Original Sample:	$\{1, 1, 3, 3, 5\}$	$n = 5, \bar{x} = 2.6$
Bootstrap Sample #2 from the Original Sample:	$\{2, 3, 3, 3, 4\}$	$n = 5, \bar{x} = 3$
Bootstrap Sample #3 from the Original Sample:	$\{1, 2, 3, 5, 5\}$	$n = 5, \bar{x} = 3.2$
Bootstrap Sample #4 from the Original Sample:	$\{1, 1, 1, 4, 5\}$	$n = 5, \bar{x} = 2.4$

These values can appear more or less frequently in the resampled datasets than in the original dataset.

Of course, in a real study, you'd hope to have a larger sample size, and you'd create thousands of resampled datasets (not just 4).

Given the enormous number of resampled data sets,

you'll always use a computer to perform these analyses.

Algorithm:

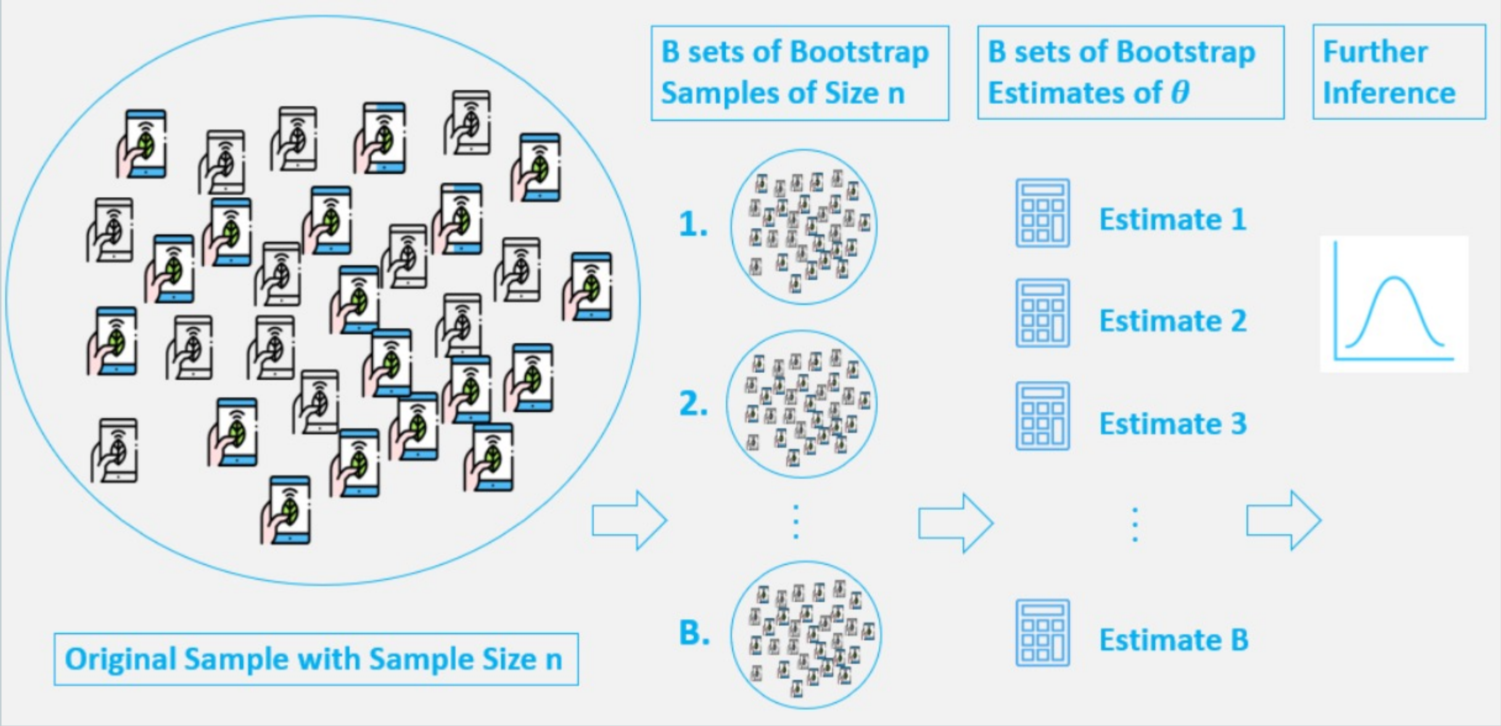
1] Draw a sample from the population of size n.

2] Draw a sample of size n from this sample, with replacement, and replicate B times.

3] Each re-sampled sample is called a bootstrap sample.

4] Evaluate the statistic of interest for each Bootstrap sample, and you will have a total of B estimates of said statistic.

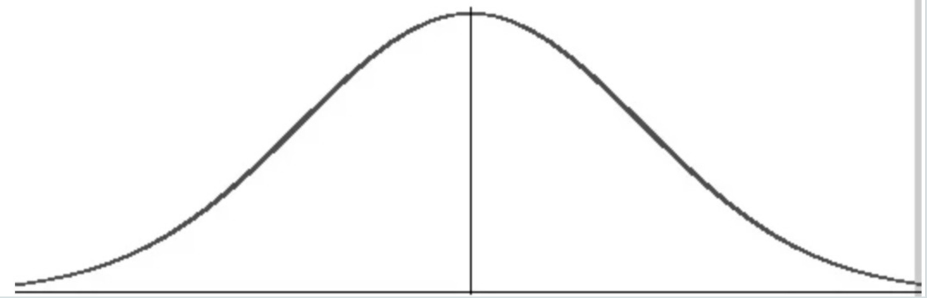
5] Construct a sampling distribution with these B Bootstrap statistics and use this distribution to make further statistical inference


B sets of Bootstrap Samples of Size n
B sets of Bootstrap Estimates of θ
Further Inference
1.
2.
B.
Estimate 1
Estimate 2
Estimate 3
Estimate B
Original Sample with Sample Size n

Pseudocode for a bootstrap 90% CI for the median.

```
medians = [ ]

given: original sample

For a large number of resamples:

    resample = random.choice(original sample, with replacement)
        # Make sure it is the same size as the original sample

    med_resample = np.median(resample)

    medians.append(med_resample)

CI = np.percentile(medians, [5, 95])

return CI
```

The return is a 90% CI for the median of the population.



Why [5, 95] ?

To create the bootstrapped confidence interval, we simply use percentiles.

For a 90% confidence interval, we need to identify the middle 90% of the distribution. To do that, we use the 95th percentile and the 5th percentile (95 – 5 = 90).

For a 95% confidence interval, we need to identify the middle 95% of the distribution. To do that, we use the 97.5th percentile and the 2.5th percentile (97.5 – 2.5 = 95).

In other words, if we order all sample means from low to high, and then chop off the lowest 2.5% and the highest 2.5% of the means, the middle 95% of the means remain.

That range is our bootstrapped confidence interval!

Bootstrap methods are alternative approaches to traditional hypothesis testing and are notable for being easier to understand and valid for more conditions.

The bootstrap method is not a way to reduce errors,
but only tries to estimate errors.

The bootstrap distribution usually estimates' the shape, spread, and bias of the actual sampling distribution.

The bootstrap process does not replace or add new data.

Bootstrapping cannot be done when:

…the data are so small that they do not approach values in the population.

…the data has many outliers.

…you have time series data.

The bootstrap is based on the assumption of data independence.

How does bootstrapping compare to traditional methods (what we've been doing?)

Both bootstrapping and traditional methods use samples to draw inferences about populations.

To accomplish this goal, these procedures treat the single sample that a study obtains as only one of many random samples that the study could have collected.

From a single sample, you can calculate a variety of sample statistics, such as the mean, median, and standard deviation.

For instance,

Suppose an analyst repeats their study many times. In this situation, the mean will vary from sample to sample and form a distribution of sample means. This distribution is called a 'sampling distribution.'

Sampling distributions are crucial because they place the value of your sample statistic into the broader context of many other possible values.

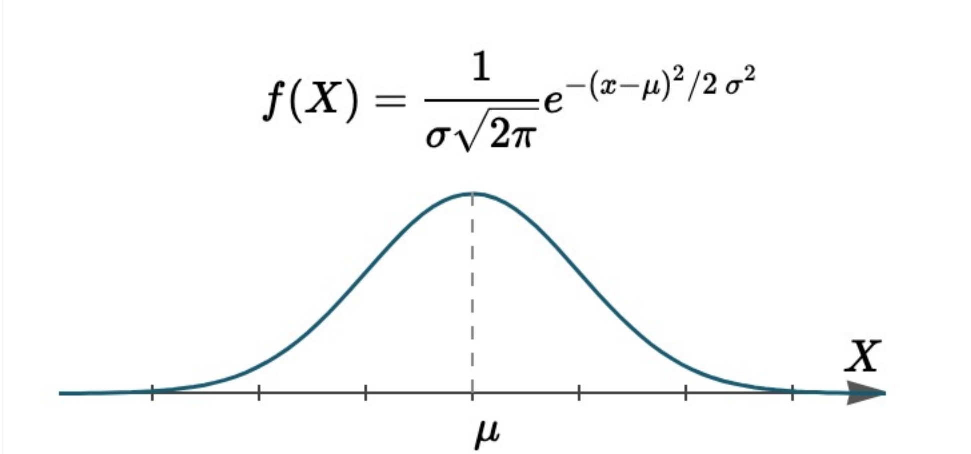
Since performing a study many times is infeasible,
both methods can estimate sampling distributions.

Using the larger context that sampling distributions provide, these procedures can construct confidence intervals and perform hypothesis testing.

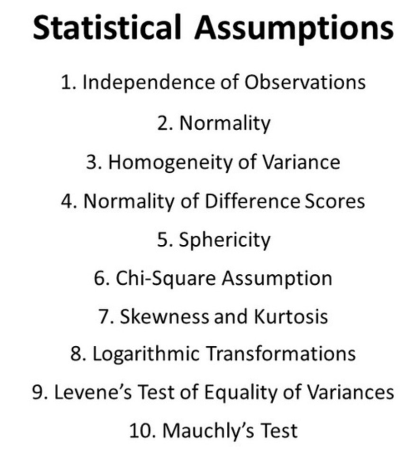
How is bootstrapping different?

A primary difference between bootstrapping and traditional statistics is how they estimate sampling distributions.

Traditional hypothesis testing procedures require equations that estimate sampling distributions using the properties of the sample data, the experimental design, and a test statistic.



$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\,\sigma^2}$$



Statistical Assumptions

1. Independence of Observations
2. Normality
3. Homogeneity of Variance
4. Normality of Difference Scores
5. Sphericity
6. Chi-Square Assumption
7. Skewness and Kurtosis
8. Logarithmic Transformations
9. Levene’s Test of Equality of Variances
10. Mauchly’s Test

$$\boxed{z \text{ or } t = \frac{\text{statistic - parameter}|H_0}{SE_{\textit{statistic}}}}$$

To obtain valid results, you’ll need to use the proper test statistic and satisfy the assumptions.

Traditional Method

Population has an unknown parameter, θ

We get a sample to produce a statistic, $\hat{\theta}$ to estimate θ.

We study a distribution for $\hat{\theta}$ to produce CI's.

Bootstrap Method

Population has an unknown parameter, θ

We get a sample to produce a statistic, $\hat{\theta}$ to estimate θ

We generate bootstrap samples, each with a $\hat{\theta}^*$.

We study a distribution of $\hat{\theta}^*$ to produce CI's.

Bootstrapping does not make assumptions about the distribution of your data.

You merely resample your data and use whatever sampling distribution emerges. Then, you work with that distribution, whatever it might be.

Conversely, the traditional methods often assume that the data follow the normal distribution or some other distribution.

For the normal distribution, the central limit theorem might let you bypass this assumption for sample sizes that are larger than ~30.

Consequently, you can use bootstrapping for a wider variety of distributions, unknown distributions, and smaller sample sizes. Sample sizes as small as 10 can be quite usable.

In this vein, all traditional methods use equations that estimate the sampling distribution for a specific sample statistic when the data follow a particular distribution.

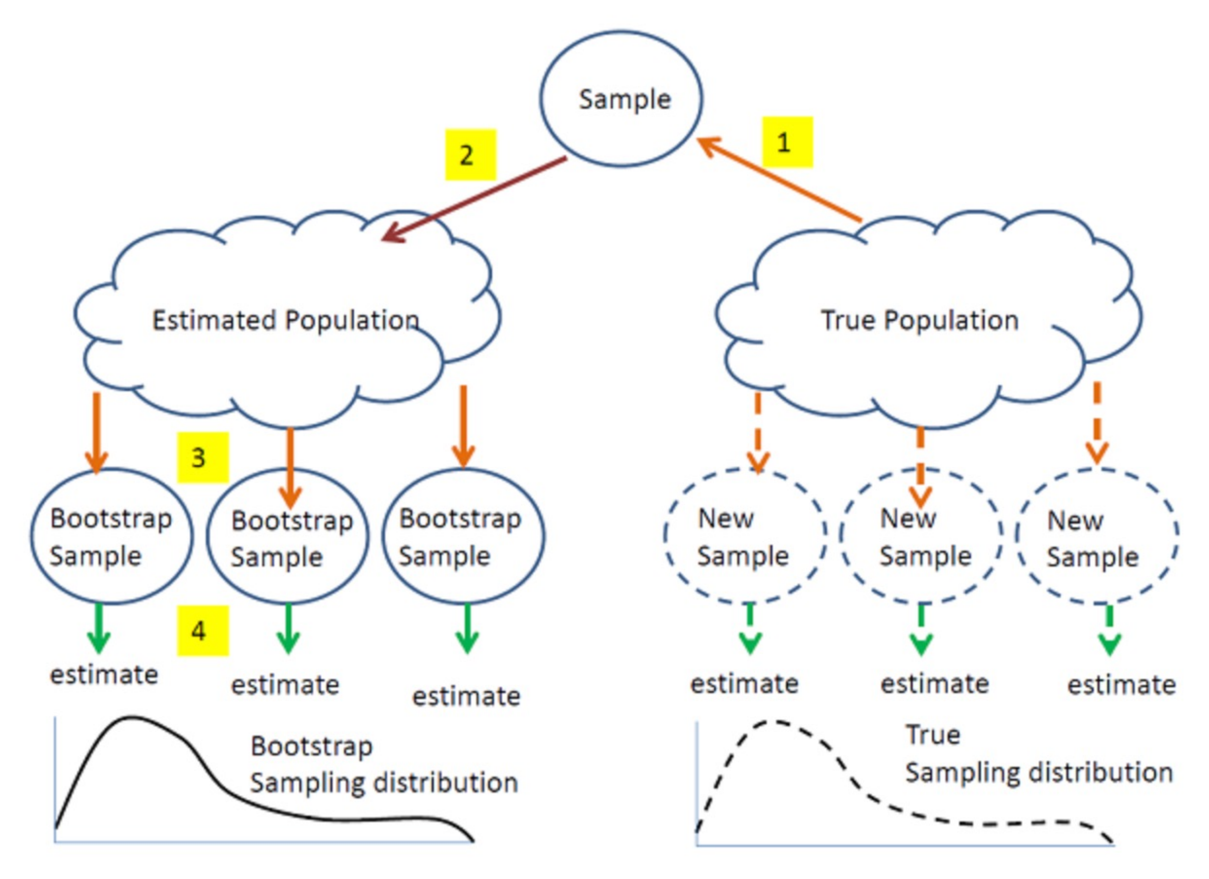
Unfortunately, formulas for all combinations of sample statistics and data distributions do not exist! For example, there is no known sampling distribution for medians, which makes bootstrapping the perfect analyses for the median.

Other analyses have assumptions such as equality of variances. However, none of these issues are problems for bootstrapping.

Observed quantities are denoted by solid curves.
Unobserved quantities by dashed curves.

The objective is to estimate the true sampling distribution of some quantity T.

The true sampling distribution is computed by taking new samples from the true population, computing T and then accumulating all of the values of T into the sampling distribution.



However, taking new samples is expensive, so
1] we take a single sample
2] use the sample to estimate the population
3] Then take samples on the computer from the estimated population
4] compute T from each computer sample
5] accumulate all of the values of T into an estimate of the sampling distribution.
6] From this estimated sampling distribution we can estimate the desired features of the sampling distribution.

keep in mind that bootstrapping does not create new data. Instead, it treats the original sample as a proxy for the real population and then draws random samples from it.

Consequently, the central assumption for bootstrapping is that the original sample accurately represents the actual population.

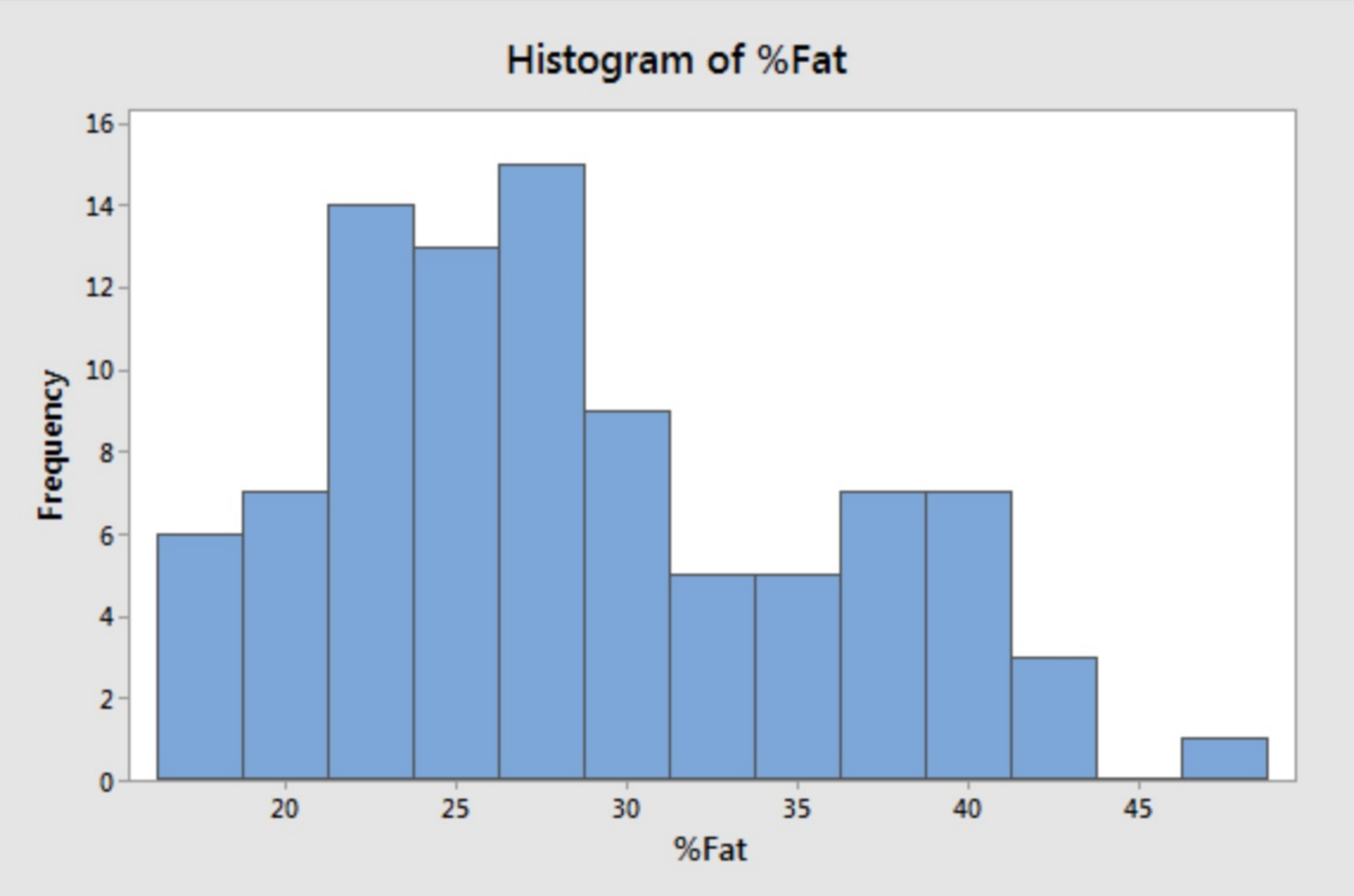
The resampling process creates many possible samples that a study could have drawn. The various combinations of values in the simulated samples collectively provide an estimate of the variability between random samples drawn from the same population

The range of these potential samples allows the procedure to construct confidence intervals and perform hypothesis testing. Importantly, as the sample size increases, bootstrapping converges on the correct sampling distribution under most conditions.

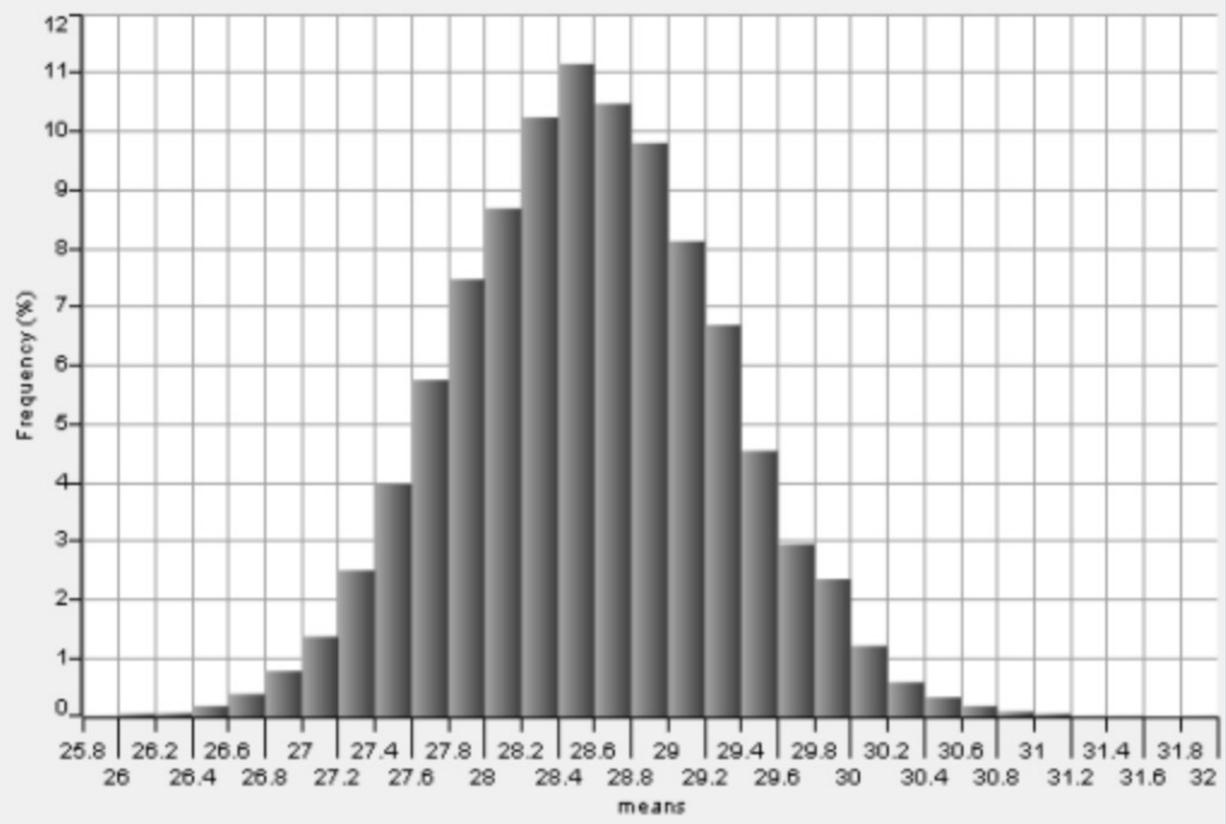
Example: Consider the body fat percentage of 92 individuals.

Notice, the distribution is not exactly normal.

Because it does not meet the normality assumption of traditional statistics, it's a good candidate for bootstrapping (Large sample size might let us bypass this assumption.)



Take the original dataset and resample it with replacement 500,000 times.
This process produces bootstrapped samples with 92 observations in each.
Your code then calculates each sample's mean and plots the distribution of the 500,000 means in a histogram.



The histogram is our sampling distribution of means. Bootstrapping methods create these distributions using resampling, while traditional methods use equations for probability distributions.

Notice how the sampling distribution in the histogram approximates a normal distribution even though the underlying data distribution is skewed. This approximation occurs thanks to the central limit theorem. As the sample size increases, the sampling distribution converges on a normal distribution regardless of the underlying data distribution.

For the body fat data, the code/program calculates a 95% bootstrapped confidence interval of the mean [27.16 30.01].

We can be 95% confident that the population mean falls within this range.

For the 95% confidence interval, we identify the middle 95% of the distribution.
To do that, we use the 97.5th percentile and the 2.5th percentile (97.5 – 2.5 = 95).

This interval has the same width as the traditional confidence interval for these data, and it is different by only several percentage points. The two methods are very close.

Suppose you are working with Bob's Buffalo Ranch to determine the best feed for buffalo weight gain. You are studying two different diets: Food A and Food B.

Y = Weight	X = Food	BootSmp1	BootSmp2	...	BootSmpB
325	A	315	380	...	380
257	A	380	315	...	380
303	A	315	257	...	257
315	A	325	315	...	325
380	A	315	257	...	303
368	B	379	390	...	390
390	B	260	260	...	379
379	B	390	390	...	390
260	B	379	368	...	379

$Est_1 = (\bar{Y}_B - \bar{Y}_A) = (349.25 - 316) = 33.25$ (Studying the difference in means)

$Est_1(B_1) = (352 - 330) = 22$ (These are for building a distribution for a CI)

$Est_1(B_2) = (352 - 304.8) = 47.2$

$\vdots \qquad \vdots \qquad \vdots$

When can I use the bootstrap method?

You may choose to bootstrap when you have a small sample

Or when large sample assumptions are not met

Or when the standard error is difficult to estimate

Or when you don’t know the shape of the sampling distribution

Next Time: Small sample hypothesis testing

Hypothesis testing assumes the null hypothesis is true and the null value is the focal point.

Confidence intervals are centered around the estimate; the sample estimate.