

Population Inference and Sampling

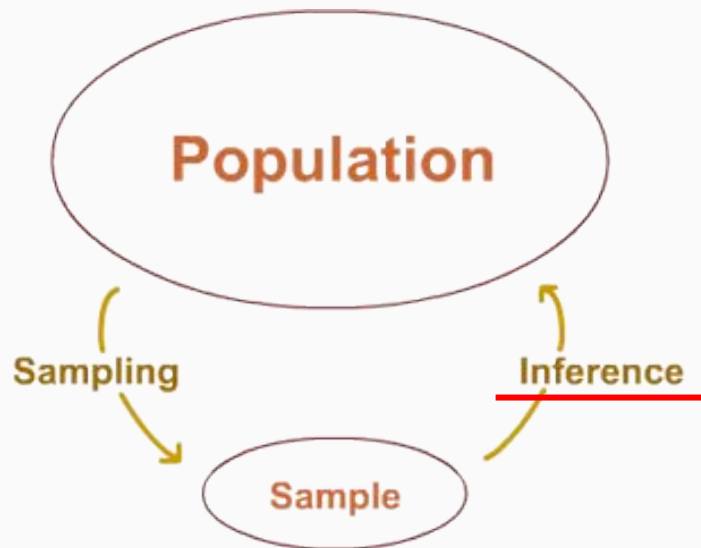
- Population Inference & Sampling
- Central Limit Theorem
- Confidence Intervals
- Bootstrapping

Population Inference & Sampling

Population Inference

You have a *sample* of a population.
What can you *infer* about the population from that sample?

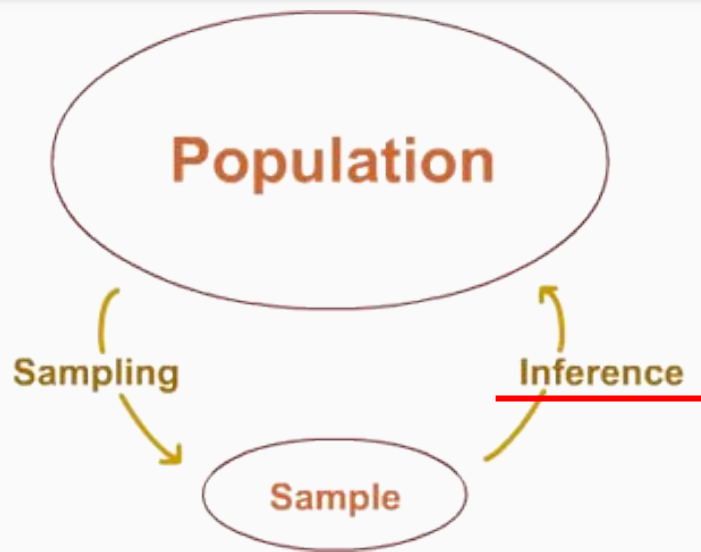
Example: Suppose we want to infer the percentage of voters in the United Kingdom who think the UK should leave the European Union based on a polled sample of the population.



Population Inference

You have a *sample* of a population.
What can you *infer* about the population from that sample?

Example: Suppose we want to infer the percentage of voters in the United Kingdom who think the UK should leave the European Union based on a polled sample of the population.



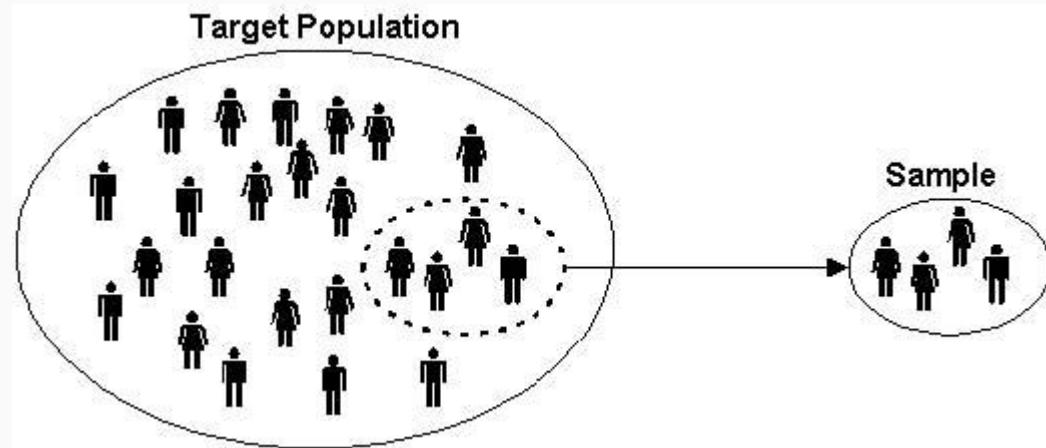
Food for thought: Why work with the sample? Why not the whole population?

Taking a sample

A sample should be representative of the population.

Random sampling is often the best way to achieve this. Ideally: **each subject has an equal chance of being in the sample.**

Simple conceptually, right?
And simple to implement, too, right?



Random sampling is surprisingly hard to do...

Scenario: You want to estimate the percentage of dog owners in Austin.

Method 1: Go to the nearby dog park and ask **random** people if they own dogs until you have n responses.

Method 2: Stand on 6th and Congress and ask **random** people if they own dogs until you have n responses.

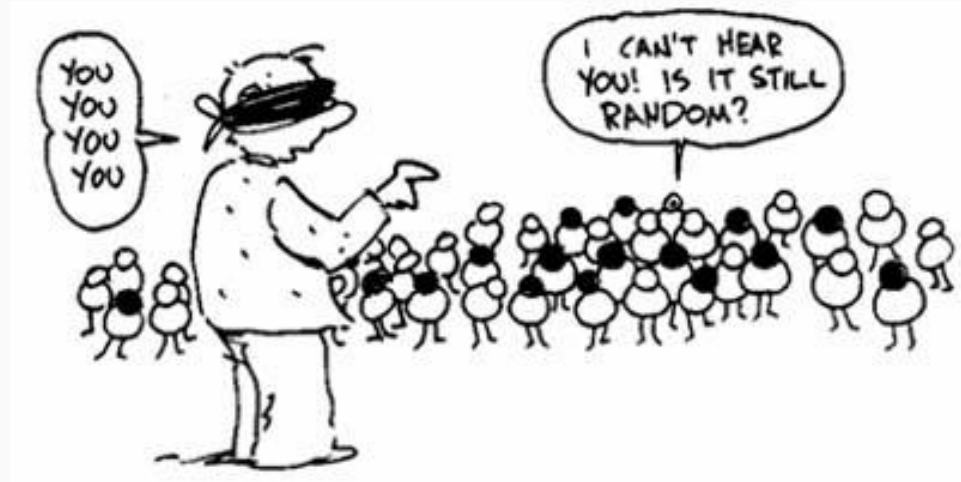
Method 3: Repeat n times: Pick a **random** neighborhood in Austin (weighted by census data per neighborhood), go to that neighborhood, ask **random** people you see if they own dogs until you get a response. Repeat.

Random sampling... just do the best you can.

Often it's impossible to do *perfect* random sampling.

So...

1. do the best you can,
2. call out possible objections, and
3. make a case for why you think your results are valid.



Random sampling in the digital age...

You might think that random sampling in a digital context is easier, and you're right!* But there are still gotchas.

Scenario: *Slack* is testing a new features ("channel polling", a way to survey people in a channel). They'd like to test the feature on only a subset of their users (n), then draw inference about their entire userbase.

Method 1: `SELECT user_id FROM users LIMIT n;`

Method 2: `SELECT user_id FROM users ORDER BY RAND() LIMIT n;`

Central Limit Theorem

(an example linking sample statistics to population statistics)

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The Central Limit Theorem

“...the central limit theorem (CLT) states that ... the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined (finite) expected value and finite variance, will be approximately normally distributed, regardless of the underlying distribution.”

- Wikipedia

Though very specific, through a little extrapolation can be applied to many practical problems...

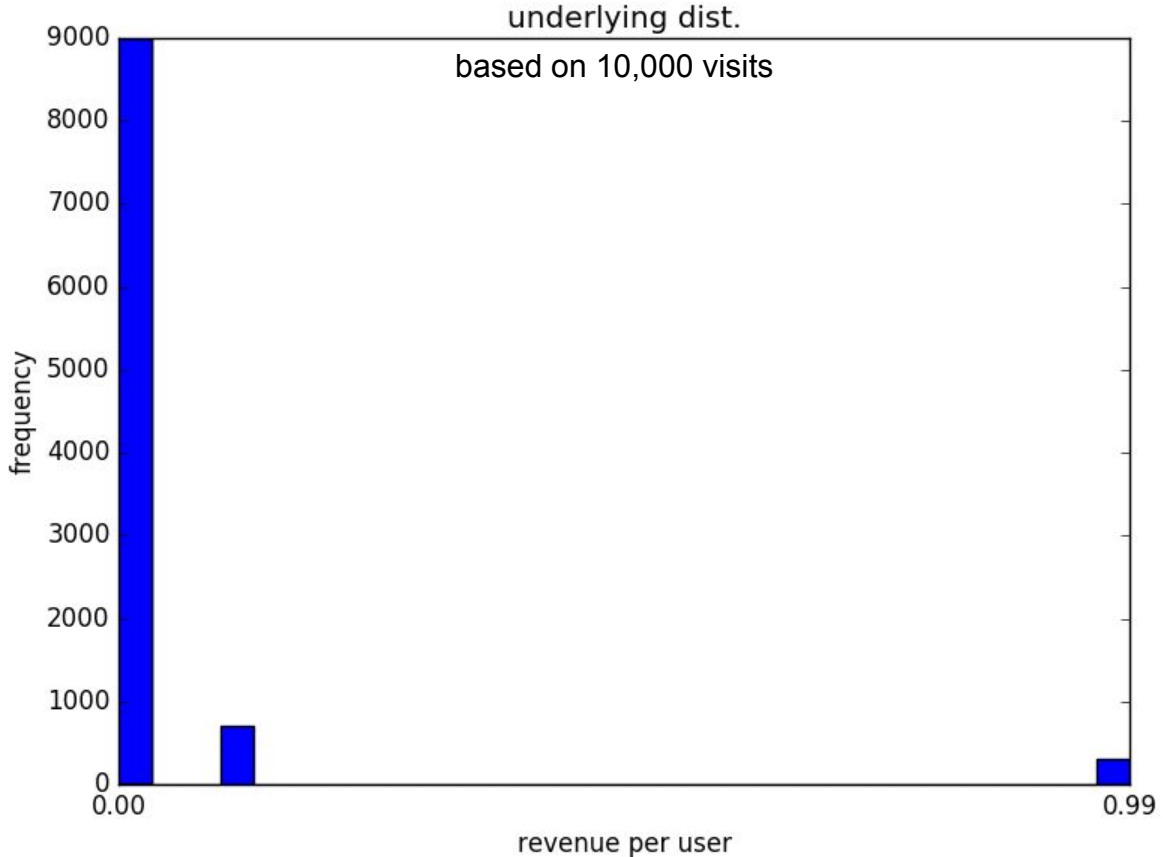
The Central Limit Theorem - one application, calculating solar flux on the receiver



Demonstration of the CLT: Distribution of website revenue per visitor

Underlying Distribution:
(assume this is the
population distribution)

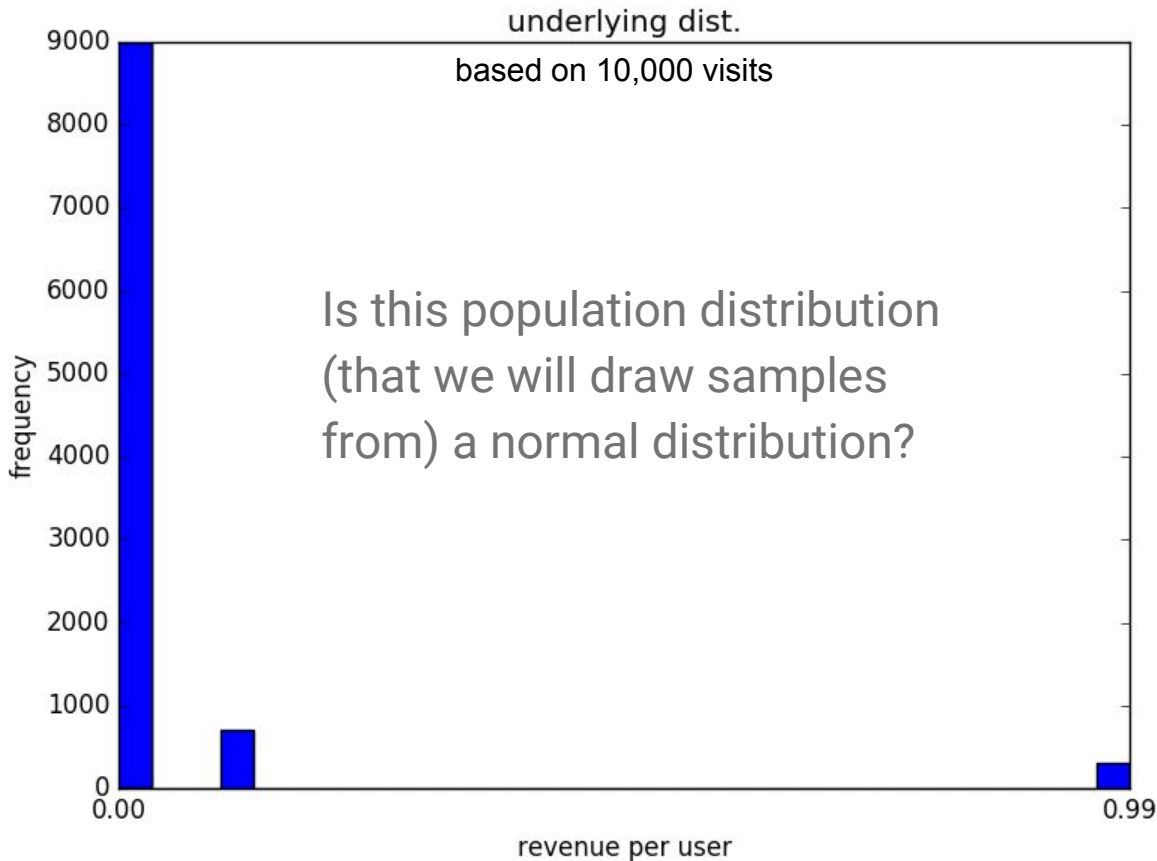
Random variable: <i>X</i> = revenue per visitor	P(<i>X</i>):
<i>X</i> = \$0.00 (no revenue)	90%
<i>X</i> = \$0.10 (ad-click)	7%
<i>X</i> = \$0.99 (app purchase)	3%



Distribution of website revenue per visitor

Underlying Distribution: (assume this is the population distribution)

Random variable: <i>X = revenue per visitor</i>	P(X):
X = \$0.00 (no revenue)	90%
X = \$0.10 (ad-click)	7%
X = \$0.99 (app purchase)	3%



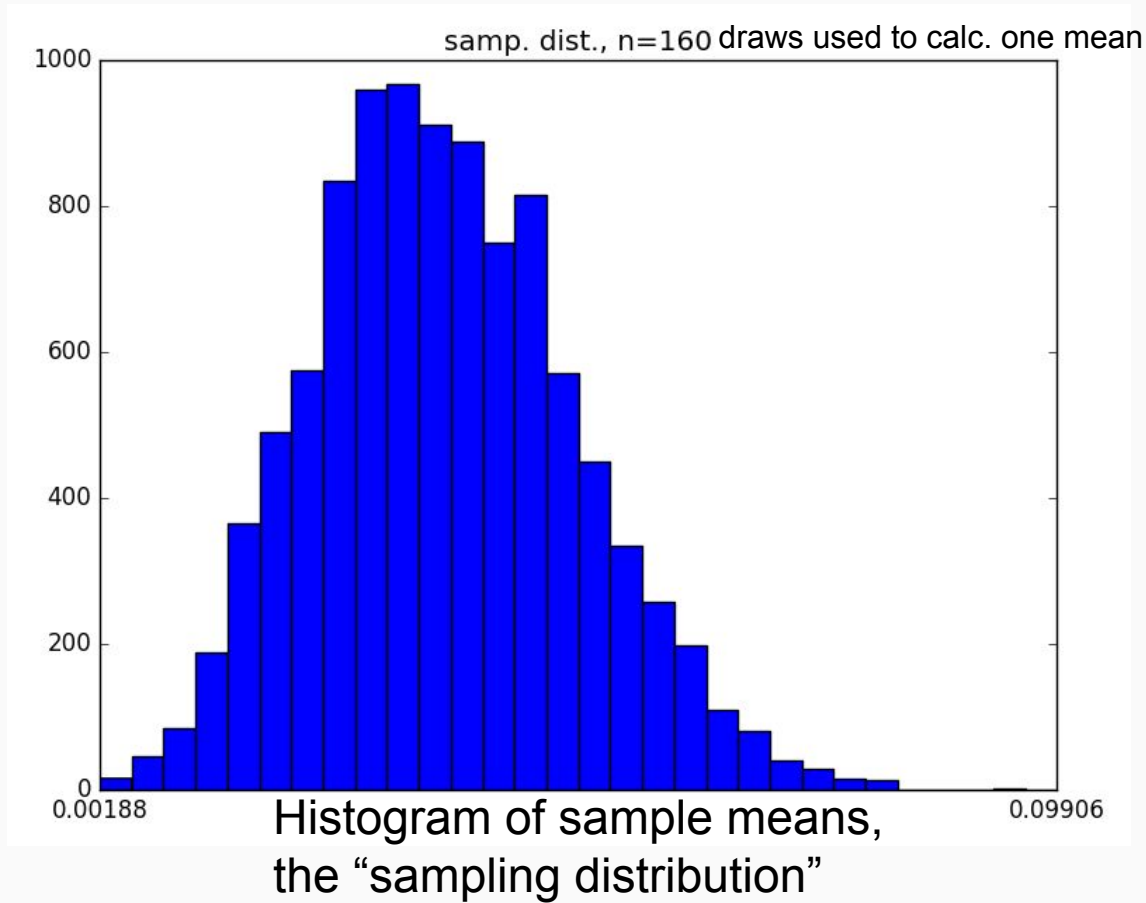
Random sampling from the population: the distribution of sample means

Collect n samples from the website revenue distribution, calculate the sample mean \bar{x}

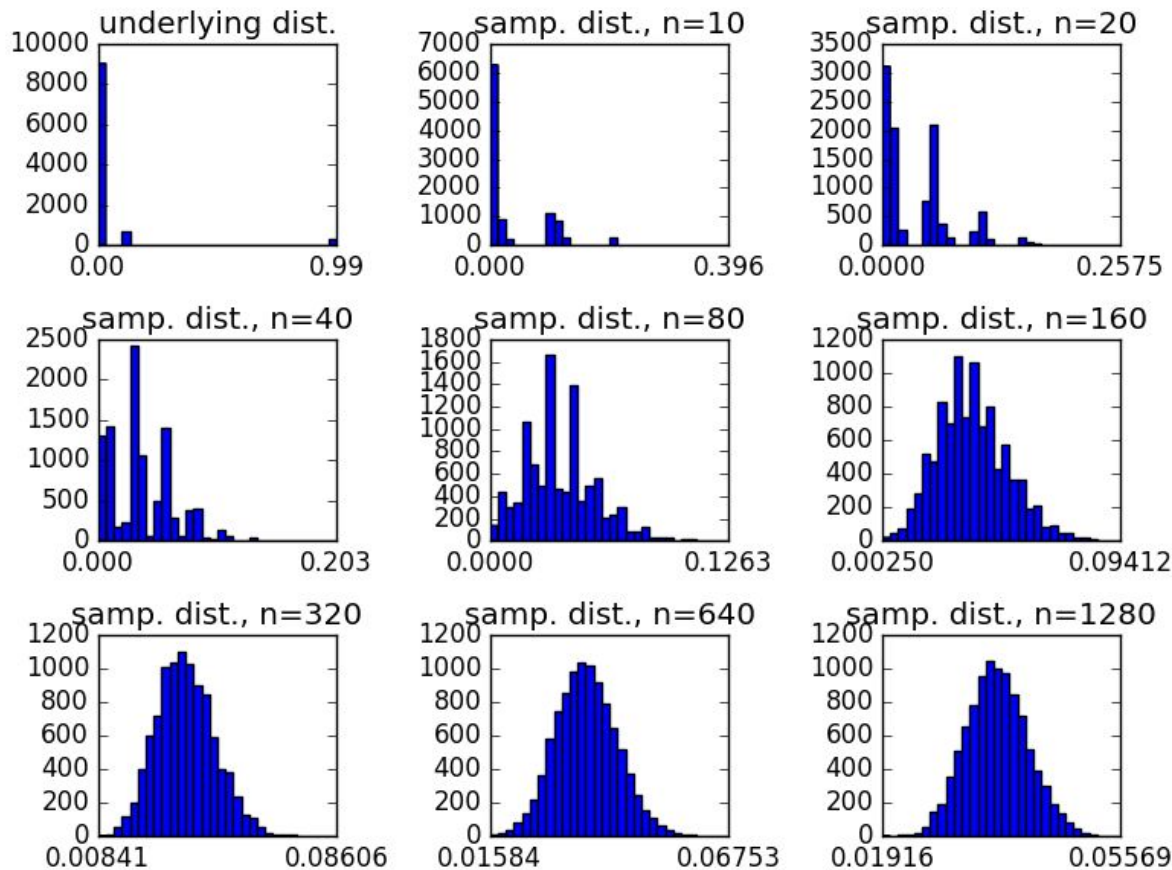
Repeat 10,000 times, we get:

$\bar{x}_0, \bar{x}_1, \dots, \bar{x}_{9999}$

Plot all 10,000 sample means.



Central Limit Theorem



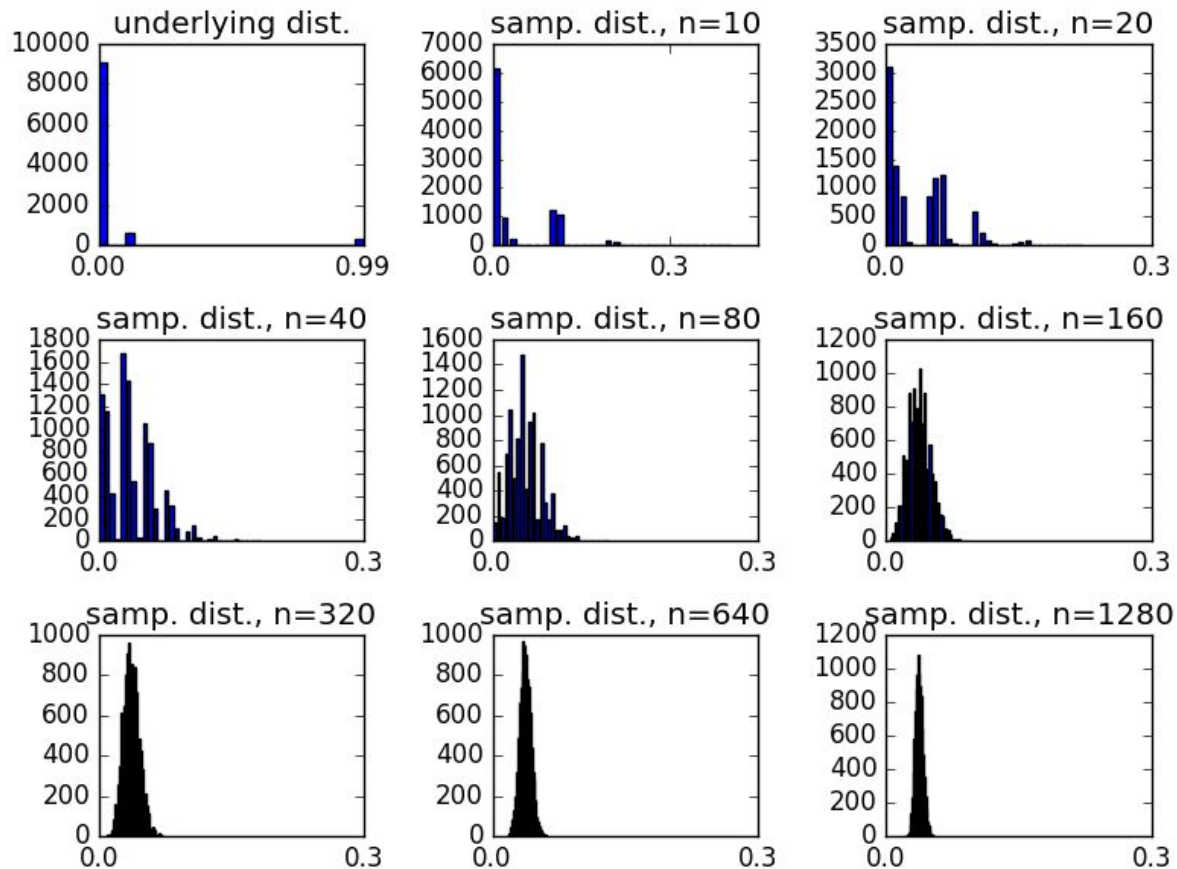
$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

An arrow points from the \bar{X} term in the equation to the text below.

The distribution of sample means (aka, the “sampling distribution”) is normally distributed*, even though the underlying distribution isn’t.

* Under certain conditions; e.g. sufficiently large sample sizes, independent and identically distributed random values (i.i.d. r.v.), finite variance.

Central Limit Theorem: What happens when the sample size increases?



Same charts as the previous slide, but now the scale of each x-axis is the same!

The standard deviation of the sampling distribution decreases as n increases. The **standard error** decreases.

Central Limit Theorem: Std. Dev precise relationship to sample mean

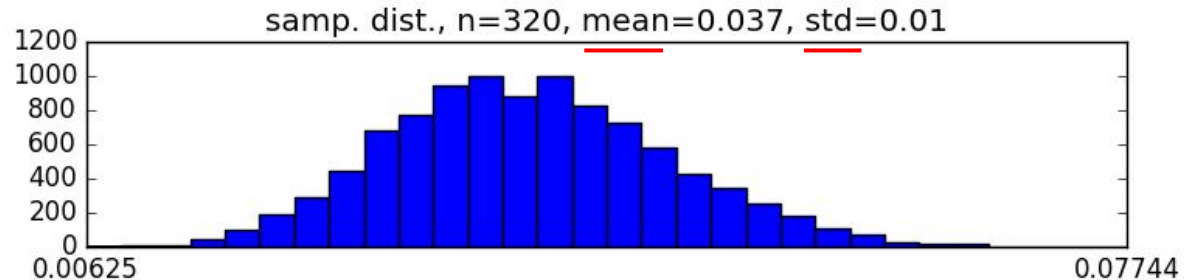
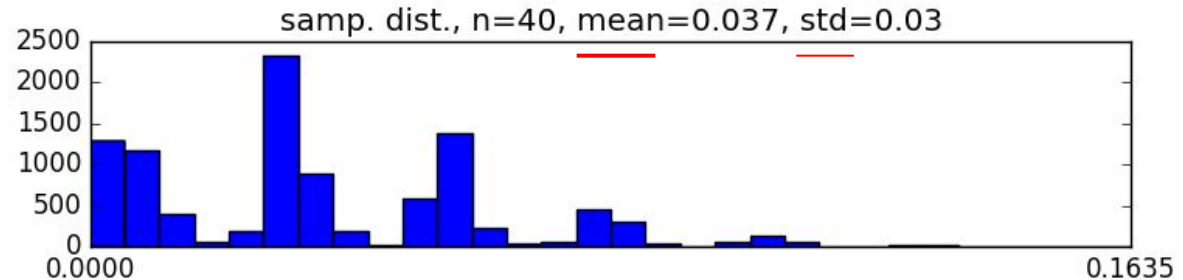
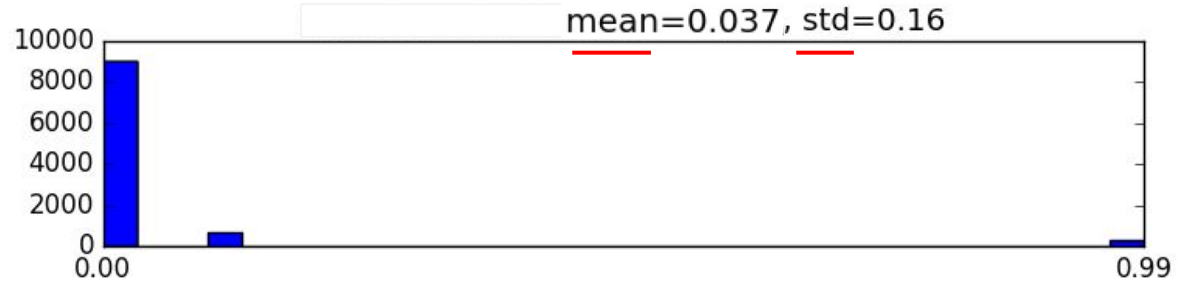
Let the underlying distribution have mean and std. dev.

μ and σ

The sampling distribution's mean and std. dev. will equal:

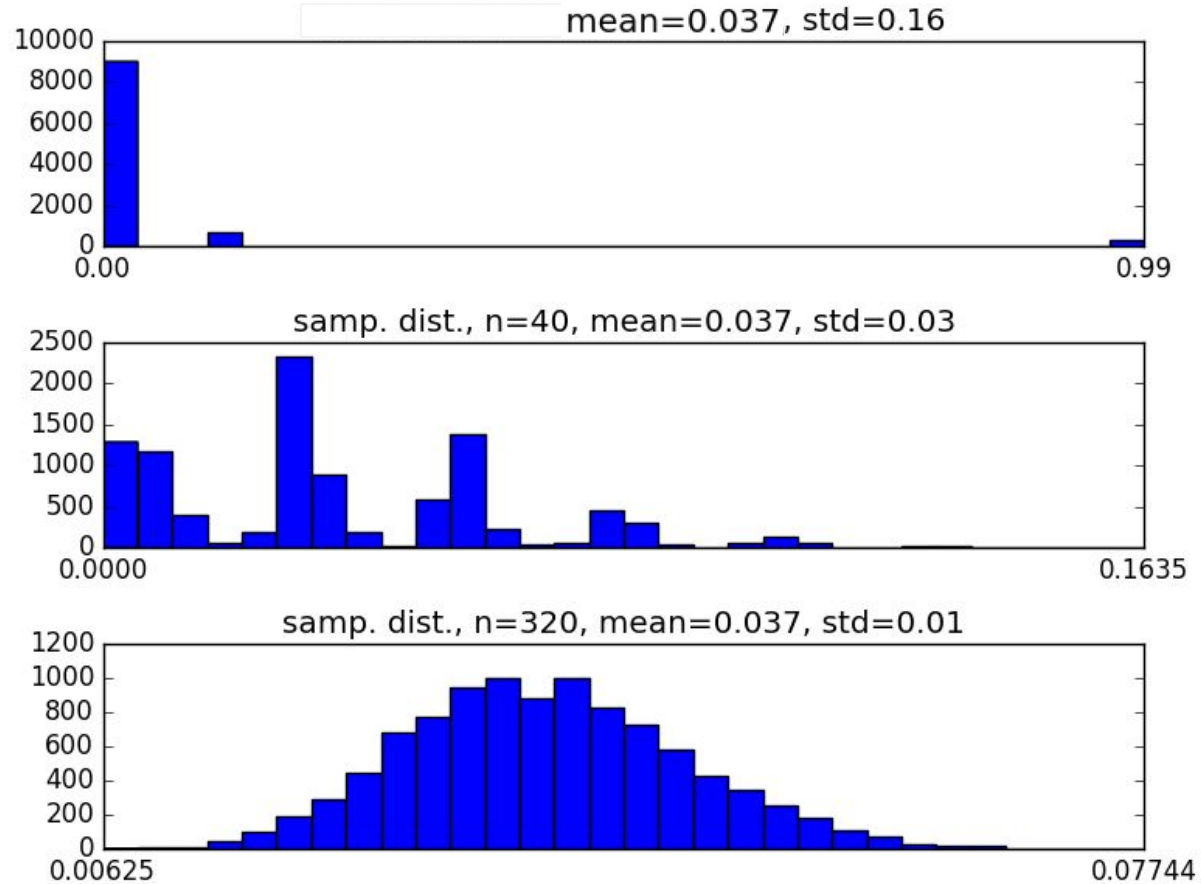
$$\mu' = \mu$$

$$\sigma' = \sigma / \sqrt{n}$$



Intuitively, does the Central Limit Theorem make sense?

Intuitively, why does the std. dev. decrease as the sample size increases?



Confidence Intervals

(given sample statistics, can we give a reliable *interval* that contains the true population statistics?)

Confidence Interval

A *confidence interval* (CI) is an interval estimate of a population parameter.

How can we do this?

- 1) Use a known relationship between sampling statistics and their distributions to calculate population statistics. Example?

Confidence Interval

A *confidence interval* (CI) is an interval estimate of a population parameter.

How can we do this?

- 1) Use a known relationship between sampling statistics and their distributions to calculate population statistics. Example?
- 2) Bootstrapping

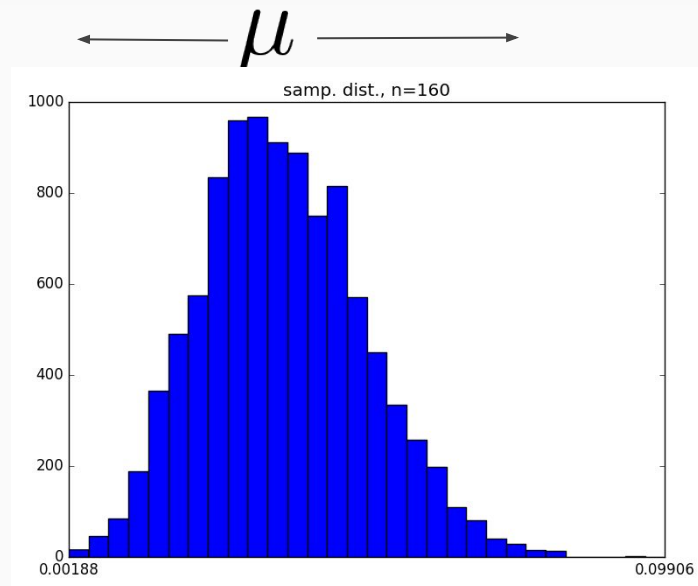
Confidence interval for a population mean using the CLT

From CLT we know that, given enough samples:

$$\mu = \bar{x}$$

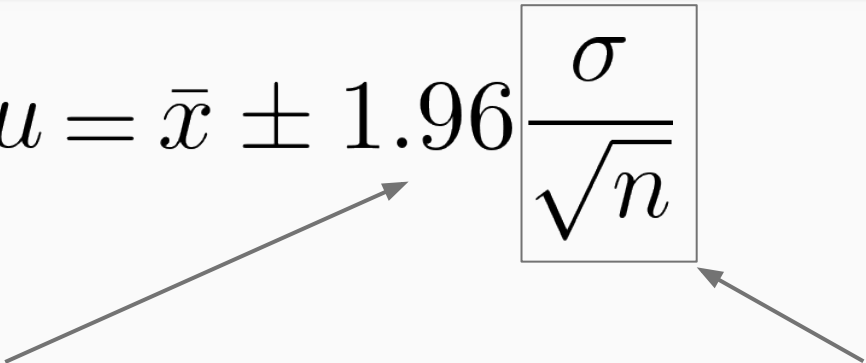
But different samples could be drawn from the population such that the sample means will vary. Therefore the sample mean will have a distribution.

The population mean is somewhere in the sample mean distribution.



Histogram of sample means

Confidence interval for a population mean using the CLT

$$\mu = \bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$
The diagram shows the formula $\mu = \bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$. An arrow points from the text 'Number of non-dimensional steps away from mean allowed according to our desired confidence level.' to the '1.96' in the formula. Another arrow points from the text 'standard deviation of sample mean a.k.a. standard error of the mean a.k.a step size' to the fraction $\frac{\sigma}{\sqrt{n}}$, which is enclosed in a box.

Number of non-dimensional steps away from mean allowed according to our desired confidence level.

For 95% confidence two-sided normal,
`scs.norm.ppf(0.975) = 1.96`

standard deviation of sample mean
a.k.a. standard error of the mean
a.k.a step size

Confidence interval for a population mean using the CLT

$$\mu = \bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

← But we don't know population's standard deviation!

That's ok, just approximate it as the sample's standard deviation, s .

Number of non-dimensional steps away from mean allowed according to our desired confidence level.

For 95% confidence two-sided normal,
`scs.norm.ppf(0.975) = 1.96`

Standard deviation of sample mean
Standard error of the mean
Step size

Confidence Interval (con't)

Since we don't know the population sigma, we can substitute sample s for it:

$$\mu = \bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$$

When n is small (<30), we should use the t-distribution instead of the normal:

$$\mu = \bar{x} \pm t_{(\alpha/2, n-1)} \frac{s}{\sqrt{n}}$$

Bootstrapping

Bootstrap Sampling

Estimates the **sampling distribution** of an estimator by sampling with replacement from the original sample.

Advantages:

- Completely automatic
- Available regardless of how complicated the estimator may be
- Doesn't rely on a predefined relationship between sample and population parameters.

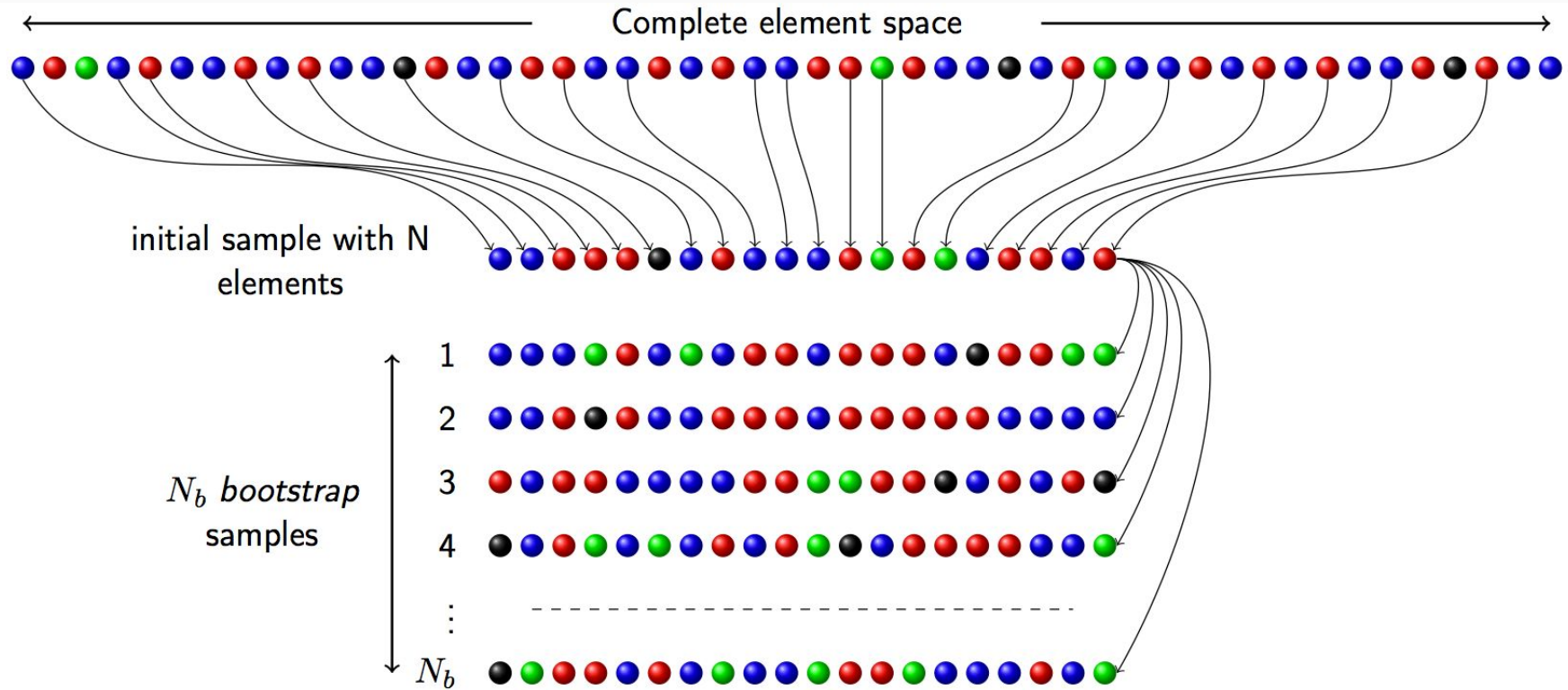
Often used to estimate the standard errors and confidence intervals of an unknown population parameter.

Bootstrap Sampling

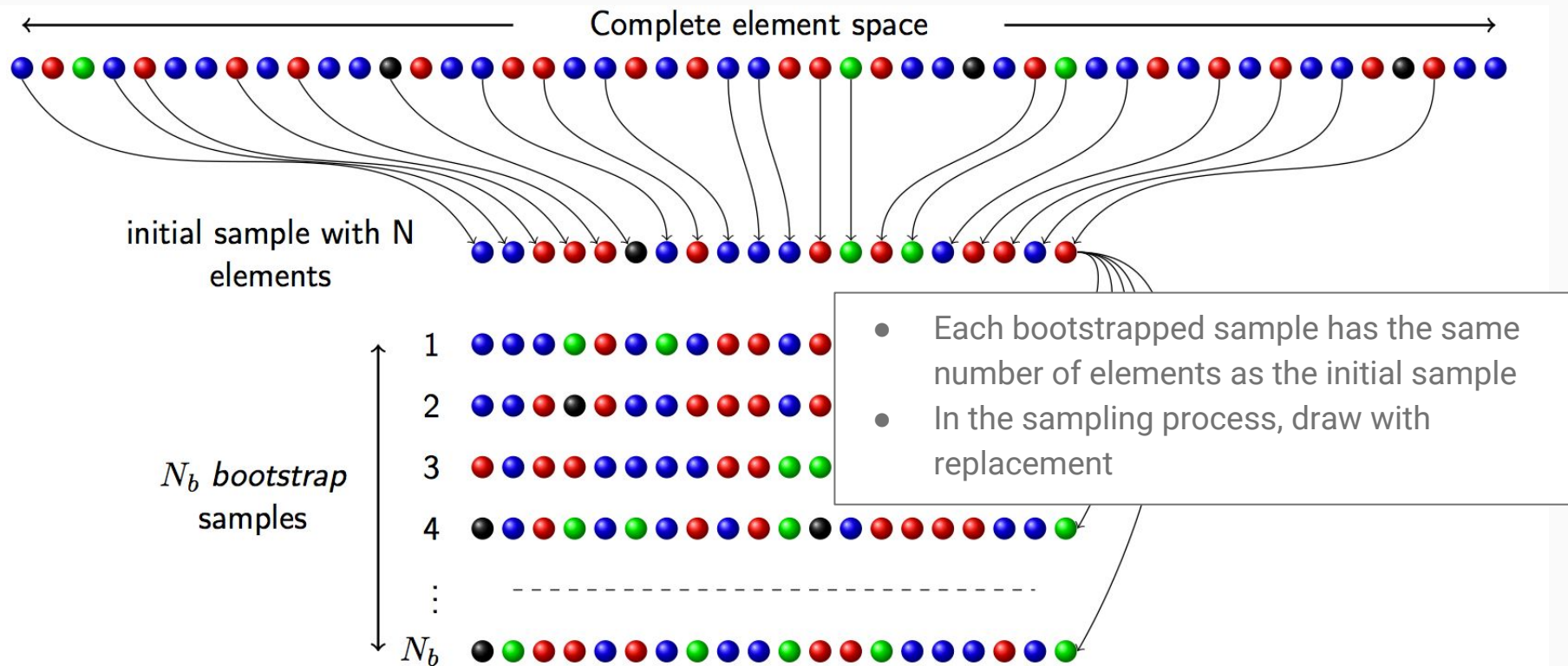
Method:

1. Start with your dataset of size n
2. Sample from your dataset with replacement to create 1 bootstrap sample of size n
3. Repeat B times
4. Each bootstrap sample can then be used as a separate dataset for estimation and model fitting

Bootstrap Sampling



Bootstrap Sampling



Bootstrap Mean - Group exercise

Consider this list of samples from a population: [2, 4, 3, 5, 3, 6]

1. Generate five bootstrapped samples from this list
2. For each sample, calculate the mean.
3. What is the mean of the means?
4. Do you think the mean of means is a more reliable estimate of the population mean than any one mean? Why or why not?

Bootstrap process - estimate the variance of a test statistic

1. Draw a bootstrap sample:

$$X_1^*, X_2^*, \dots, X_n^*$$

2. Calculate bootstrap estimate of your statistic (the parameter you're interested in):

$$\hat{\theta}^* = t(X_1^*, X_2^*, \dots, X_n^*)$$

3. Repeat steps 1 and 2, B times to get:

$$\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$$

4. Calculate the bootstrapped variance (one way to do it):

$$s_{\text{boot}}^2 = \frac{1}{B} \sum_{b=1}^B (\hat{\theta}_b^* - \bar{\theta}^*)^2 \quad \text{where} \quad \bar{\theta}^* = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b^*$$

Bootstrap Confidence Intervals

Percentile method:

$$(\hat{\theta}_{\alpha/2}^*, \hat{\theta}_{1-\alpha/2}^*)$$

`numpy.percentile(array_of_values, [2.5, 97.5])`

Interval assuming approximately *normal* bootstrap sampling distribution:

$$\bar{\theta}^* \pm 1.96 s_{\text{boot}}$$

Bootstrap Confidence Intervals

Percentile method:

$$(\hat{\theta}_{\alpha/2}^*, \hat{\theta}_{1-\alpha/2}^*)$$

`numpy.percentile(array_of_values, [2.5, 97.5])`

Interval assuming approximately *normal* bootstrap sampling distribution:

~~$$\bar{\theta}^* \pm 1.96 s_{\text{boot}}$$~~

But why restrict ourselves by assuming a normal sampling distribution of the statistic????

When to Bootstrap

When the theoretical distribution of the statistic (parameter) is complicated or unknown. (E.g. Median or Correlation)

When the sample size is too small for traditional methods.

Favor accuracy over computational cost.