

Estimation & Sampling

# Sampling

DSI SEA5, [jf.omhover](http://jf.omhover), Sep 14 2016

*Drawing on the great slides of Ryan Henning*



## Estimation & Sampling

# Sampling

DSI SEA5, jf.omhover, Sep 14 2016

*Drawing on the great slides of Ryan Henning*

## STANDARDS

- Define the Central Limit Theorem
- Compute standard error
- Pseudocode Bootstrapping for a given sample of size  $N$
- Construct confidence interval for case where parametric construction does not work
- Discuss examples of times when you need bootstrapping.



## Estimation & Sampling

# Sampling

DSI SEA5, jf.omhover, Sep 14 2016

*Drawing on the great slides of Ryan Henning*

## OBJECTIVES

- **Discover** the CLT
- **Apply** the CLT to **construct** confidence intervals for the mean of a population
- **Use** bootstrapping to construct confidence intervals for any population statistics



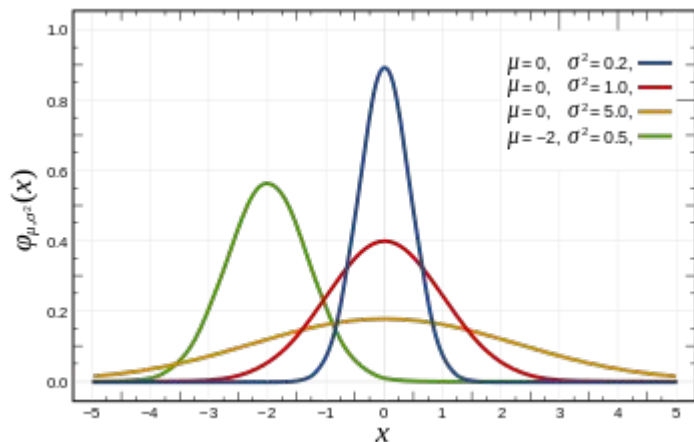
# Let's draw the “Normal Card” !

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



## PMF (DISCRETE)

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



## PARAMETERS

mu: float

sigma: float

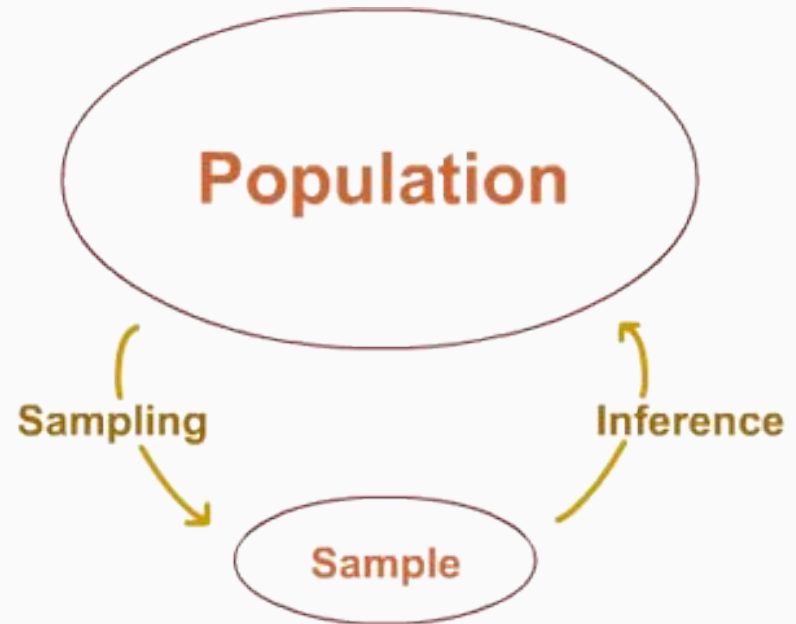
## USE CASES

...

## MOM

# Population Inference

- Start with a question/hypothesis
- Design an experiment
- Collect data
- Analyze
- Check the results
- Repeat? Redesign?

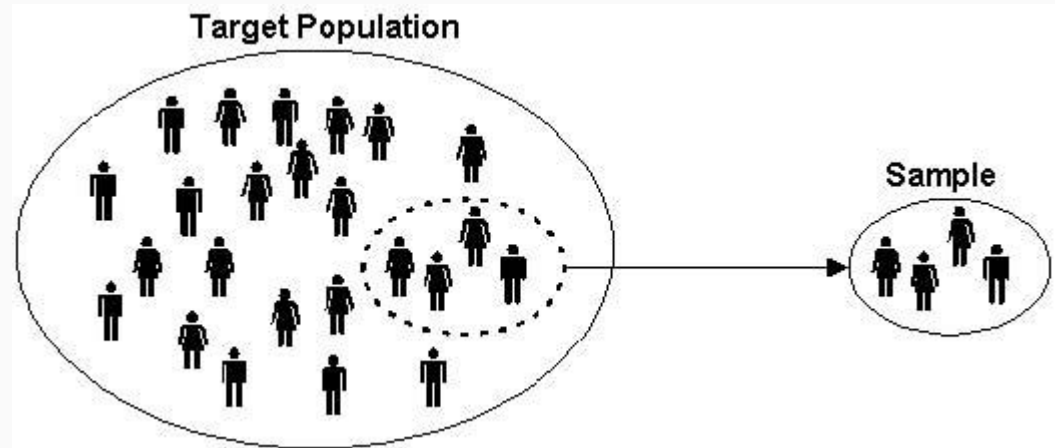


# Collecting data: Taking a sample

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.**





# 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.

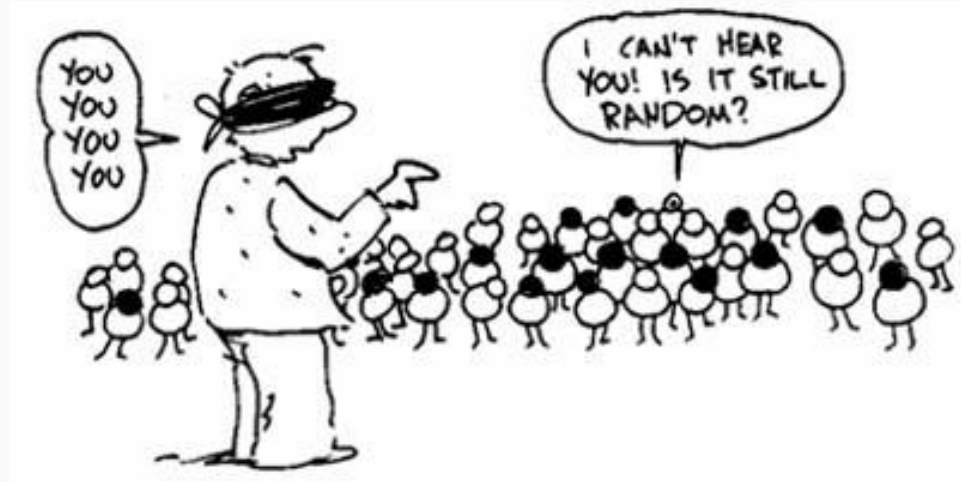
**Method 3:** Repeat  $n$  times: Pick a **random** neighborhood in Austin (weighted by census data per neighborhood), go to that neighborhood, ask **random** people you see if they own dogs until you get 1 response.

# 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;`

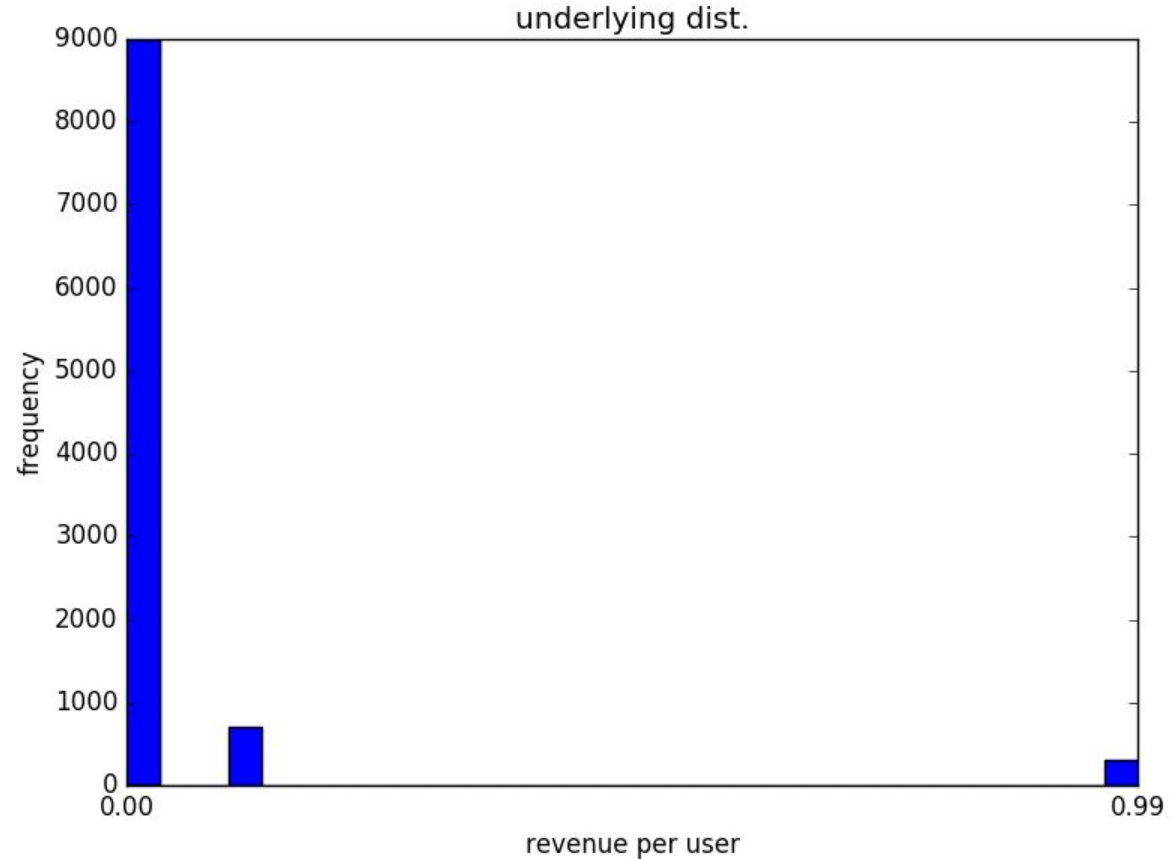


# Central Limit Theorem

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

## Underlying 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 sample means

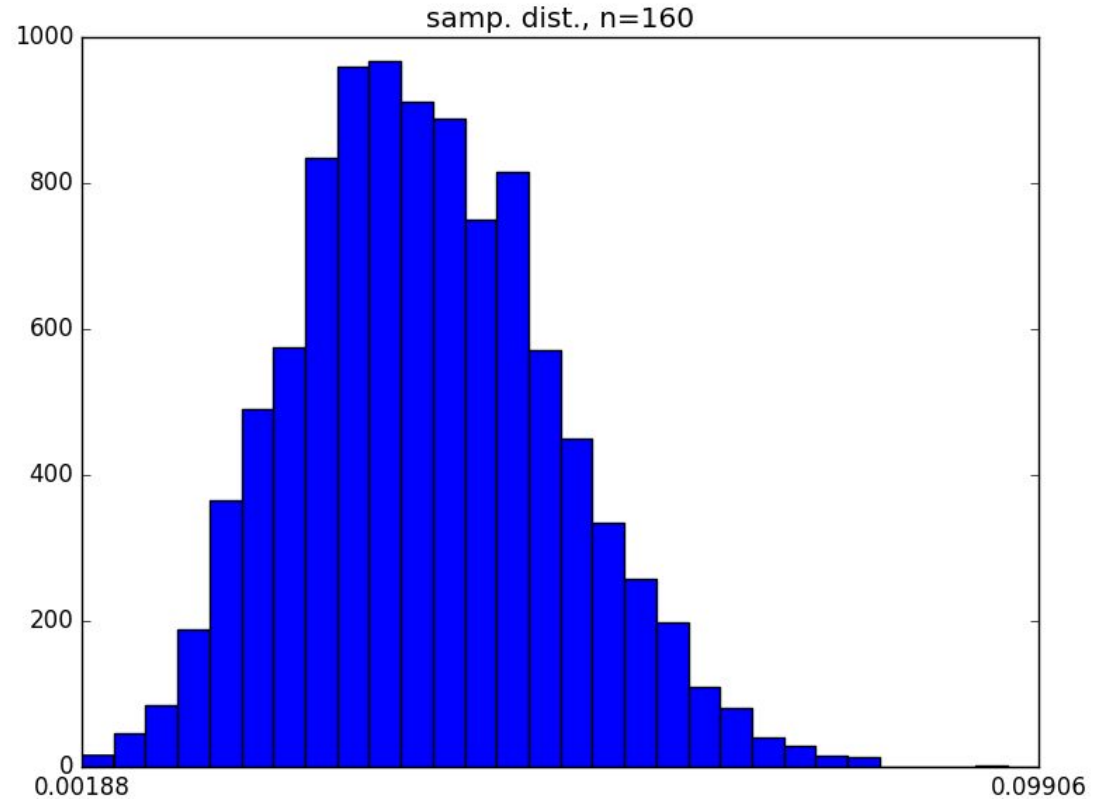


Collect  $n$  samples from the website revenue distribution, calculate the sample mean  $\bar{x}$

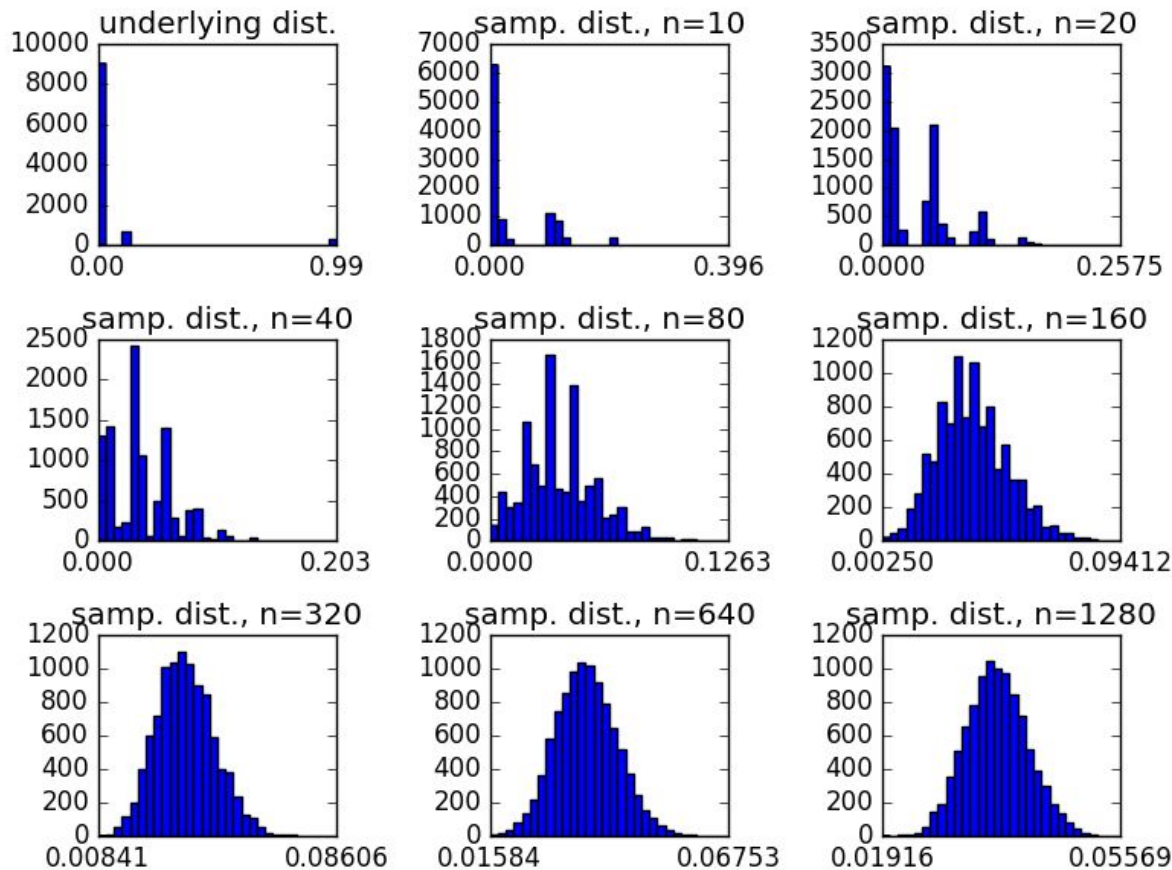
Repeat 10,000 times, we get:

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

Plot all 10,000 sample means.



# Central Limit Theorem

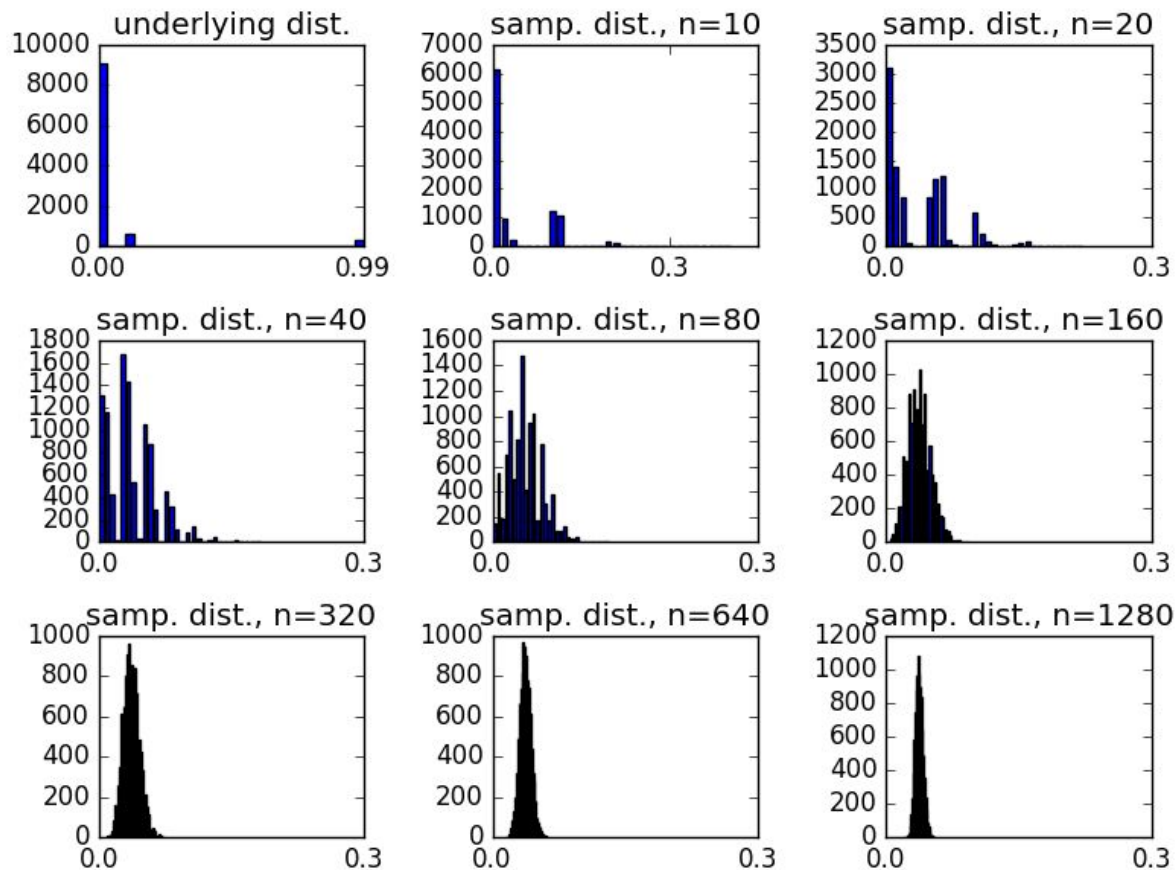


$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The distribution of sample means (aka, the “sampling distribution”) is normally distributed. \*

\* Under certain conditions; e.g. sufficiently large sample sizes, and i.i.d. r.v.

# 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!

Now we can see: **What happens when the sample size increase?**

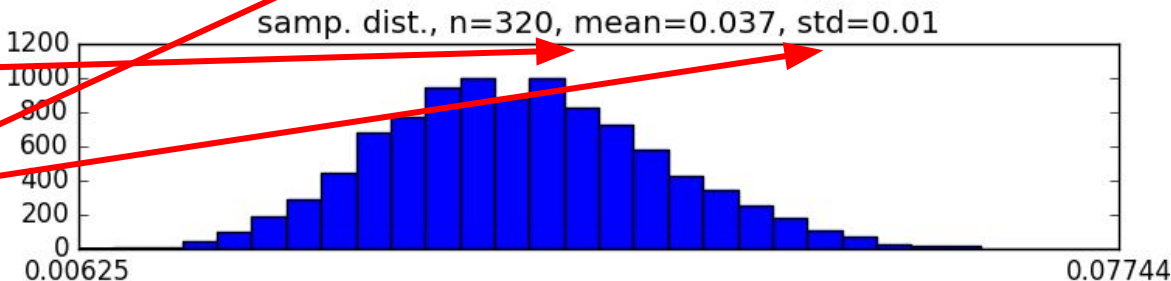
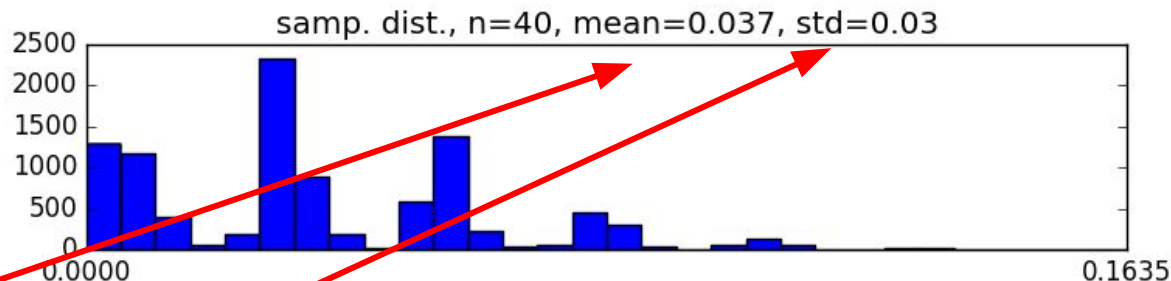
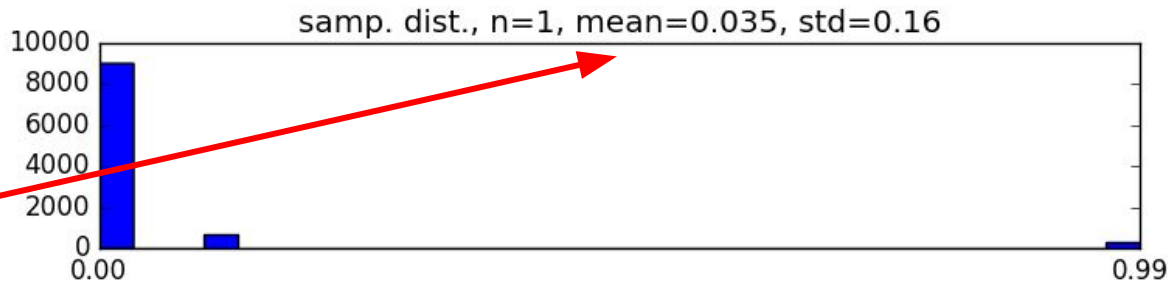
Let the underlying distribution have mean and std. dev.

$\mu$  and  $\sigma$

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

$$\mu' = \mu$$

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

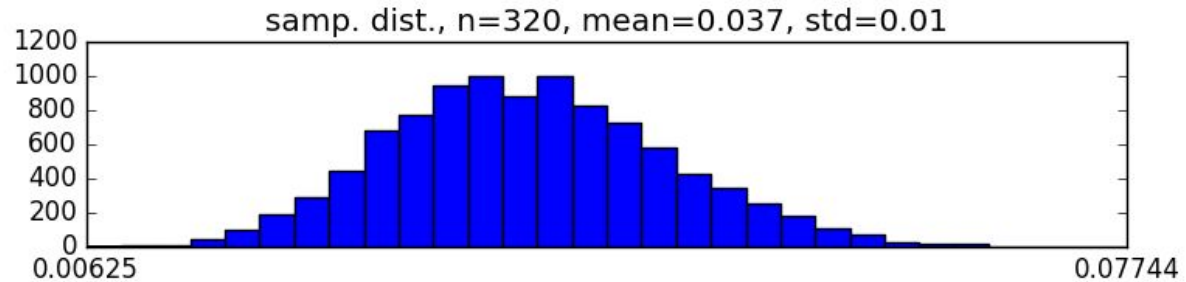
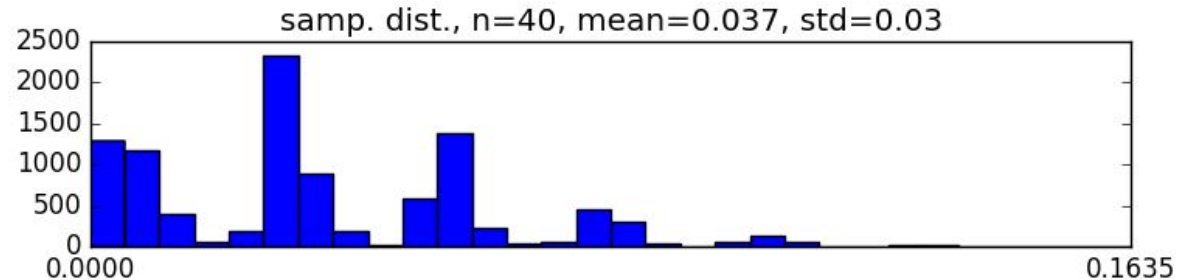
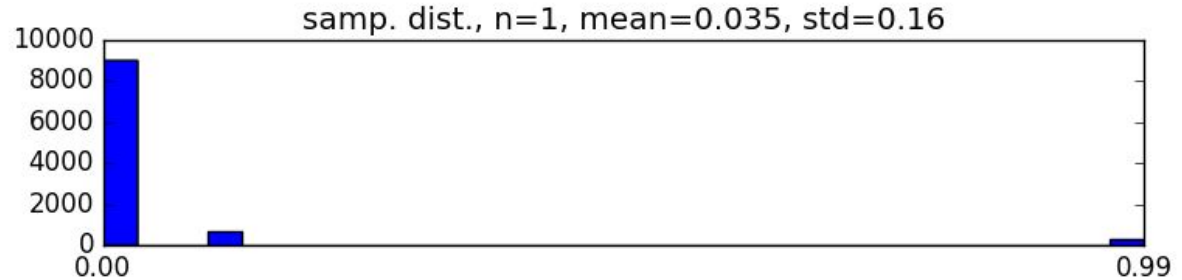


Intuitively, does the Central Limit Theorem make sense?



Intuitively, why does the mean stay the same in each histogram?

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







# Confidence Intervals

# Confidence Interval

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

The typical level of confidence is 95%, but they can be calculated for any level.

For example, a 95% CI for the population mean is given by:

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

Grr. We  
don't know  
sigma...

Why assuming the  
normal distribution here?

Where does the sqrt(n)  
come from?

```
conf = 0.95
```


```
scipy.stats.norm.ppf((1 + conf) / 2.) # <-- handling two-sided-ness
```

# Confidence Interval (con't)

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

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

Btw, this value is the  
“standard error of the  
mean (SEM)”



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

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




# Bootstrapping



# Bootstrap Sampling

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

Recall, how is the “sampling distribution” distributed when the estimator is the sample average?



Advantages:

- Completely automatic
- Available regardless of how complicated the estimator may be

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

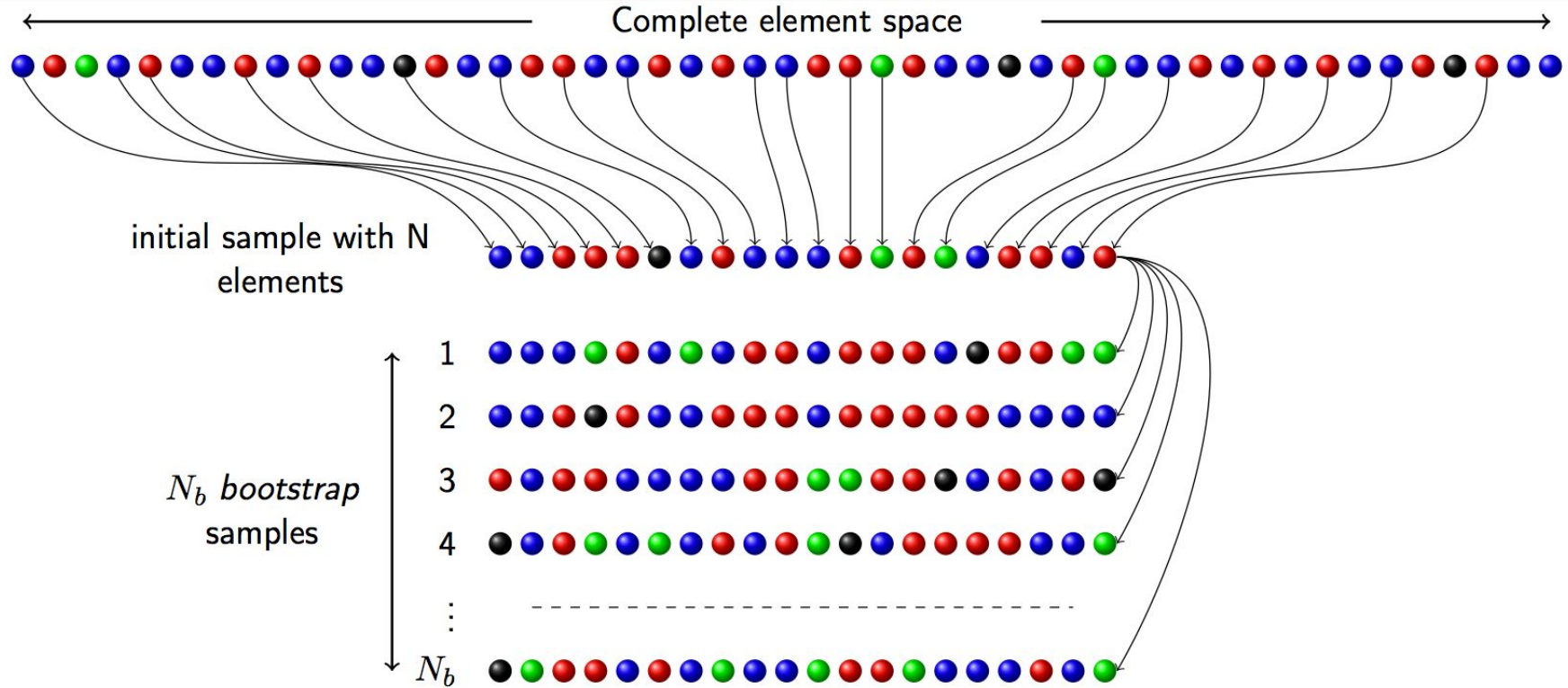


# Bootstrap Sampling

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





1. Draw a bootstrap sample:

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

2. Calculate bootstrap estimate of your parameter (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:

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



# Bootstrap Confidence Intervals

Percentile method:

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

Interval assuming approximately *normal* bootstrap sampling distribution:

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



# 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.