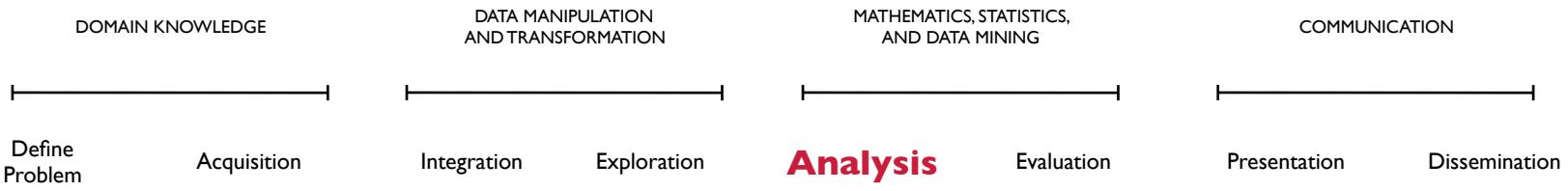
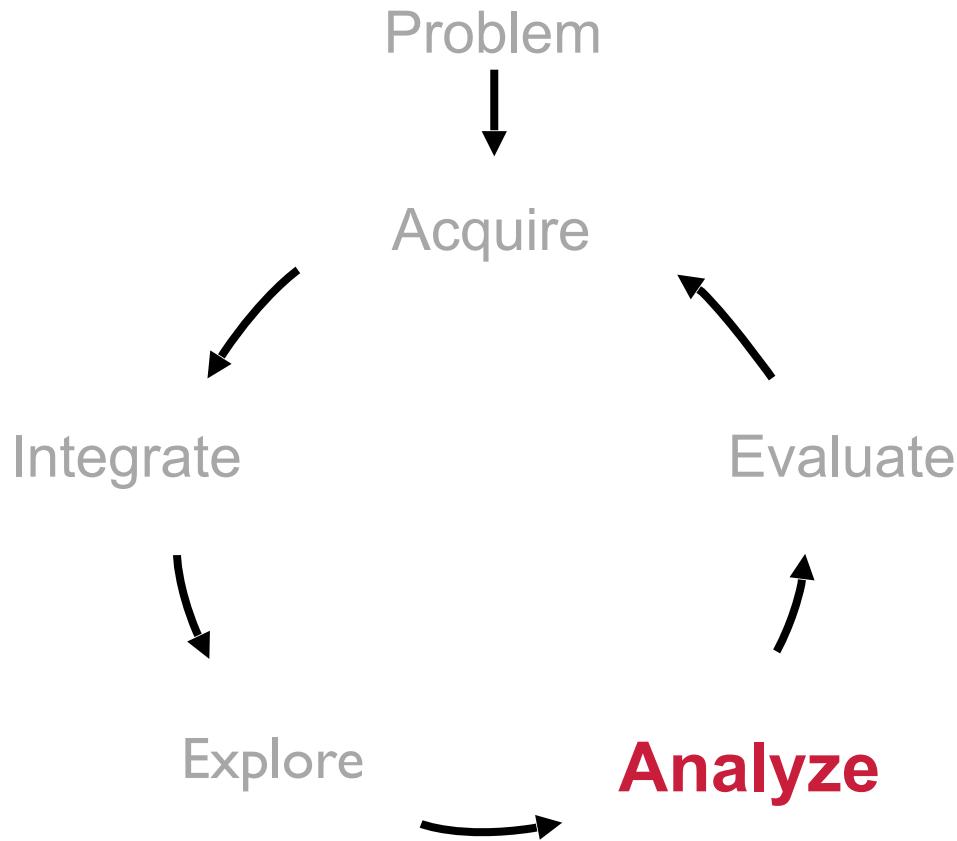


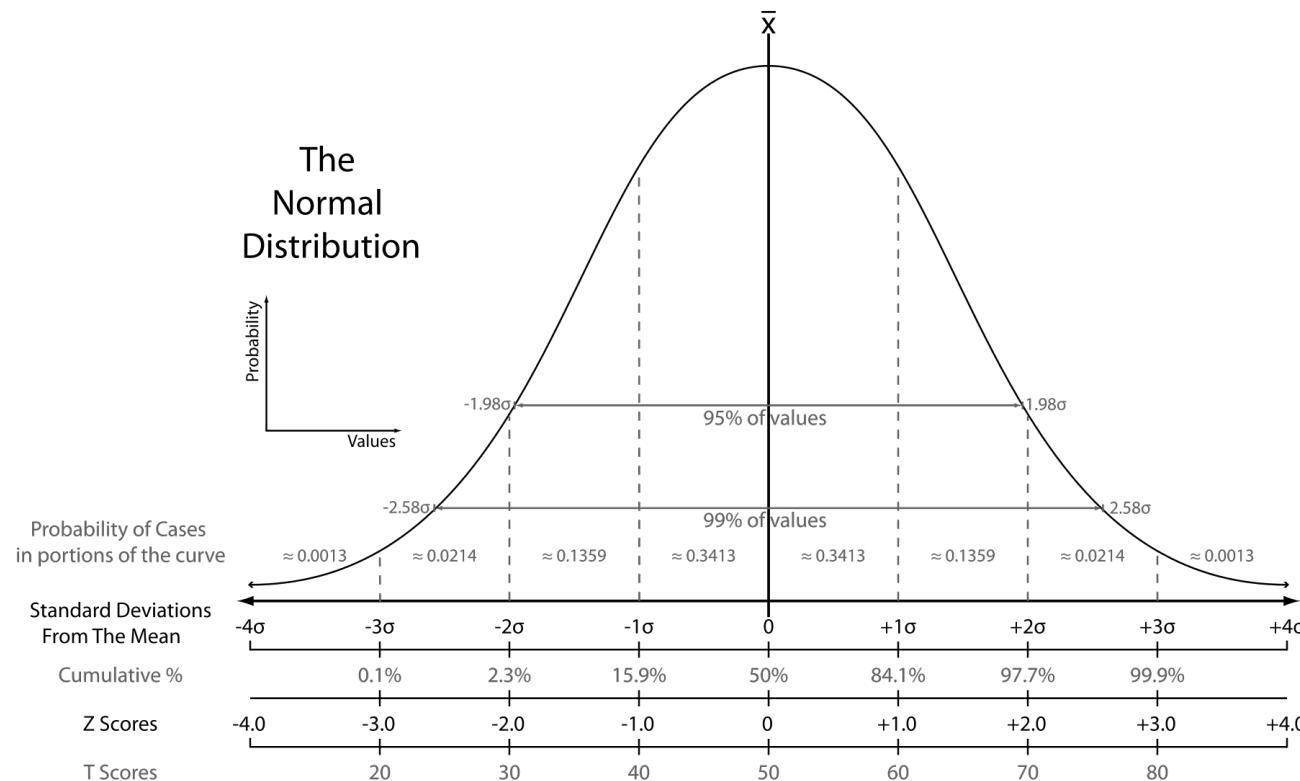
Process



Process + Iteration



What is Statistics?



What is Statistics?

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data.

- Wikipedia

What is Statistics?

Statistics is the activity of inferring results about a population given a sample.

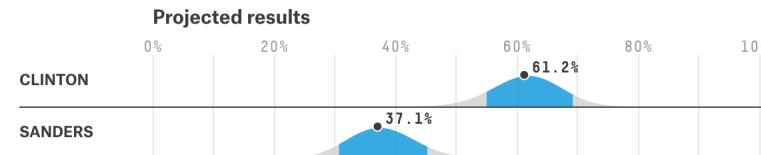
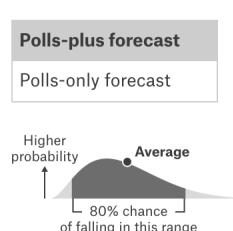
- Dennis Shasha & Manda Wilson, *Statistics is Easy!*

What is Statistics?

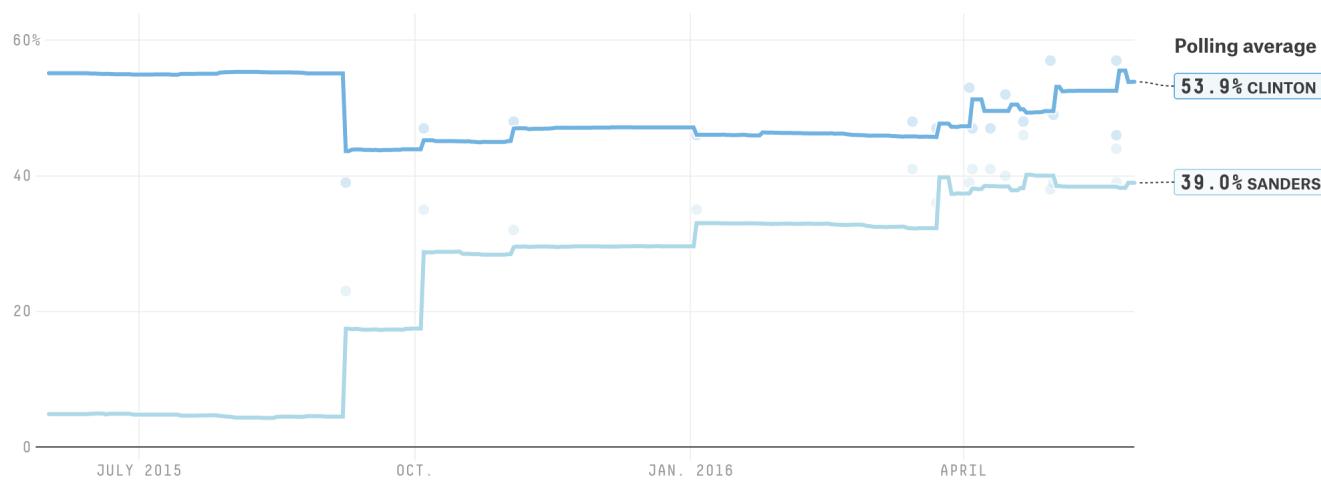
*Statistics is the activity of **inferring** results about a **population** given a **sample**.*

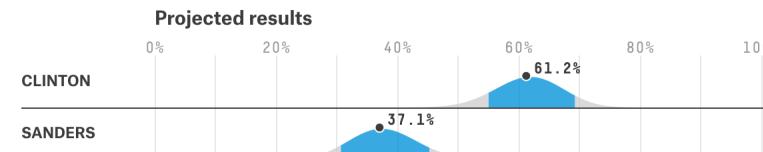
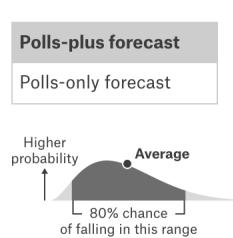
- Dennis Shasha & Manda Wilson, *Statistics is Easy!*

The Problem

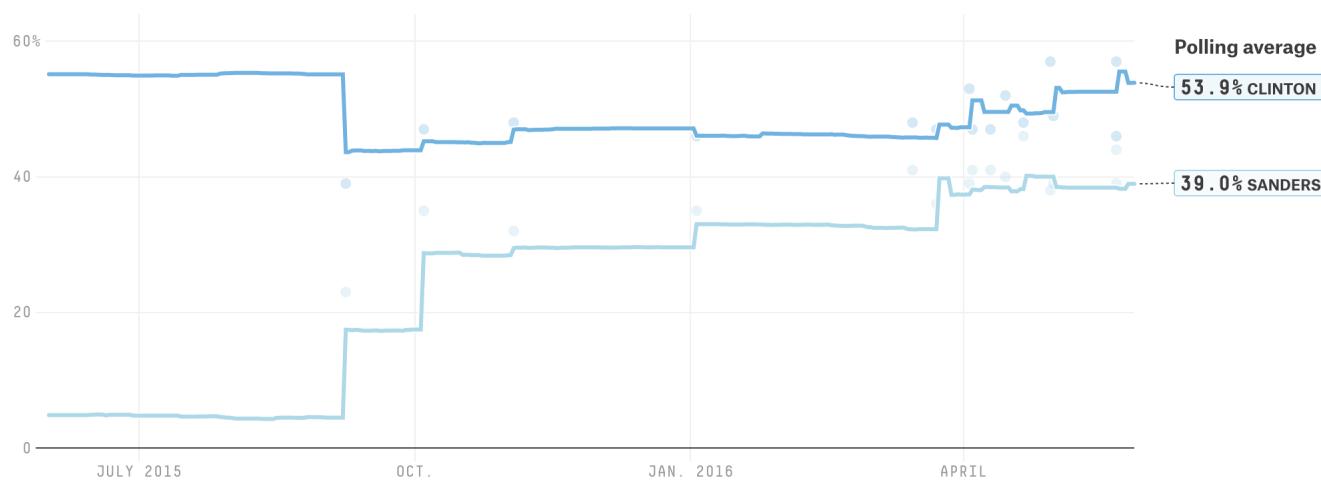


Not all polls ● are created equal, so our forecasts are calculated based on weighted polling averages ↗.





Not all polls ● are created equal, so our forecasts are calculated based on weighted polling averages ↗.



● = NEW A = ALL ADULTS RV = REGISTERED VOTERS LV = LIKELY VOTERS

POLLSTER		SAMPLE	WEIGHT	LEADER	CLINTON	SANDERS
● MAY 19-22	SurveyUSA	803 LV	1.48	Clinton +18	57%	39%
● MAY 13-22	PPIC	552 LV	0.34	Clinton +2	46%	44%
APR. 27-30	SurveyUSA	826 LV	0.05	Clinton +19	57%	38%
APR. 28-MAY 1	Sextant Strategies	1,617 RV	0.04	Clinton +10	49%	39%
APR. 18-21	Fox News	623 LV	0.01	Clinton +2	48%	46%
APR. 13-15	YouGov	1,123 LV	0.01	Clinton +12	52%	40%
APR. 7-10	Gravis Marketing	846 LV	0.00	Clinton +6	47%	41%
MAR. 24-APR. 4	Field Poll	584 LV	0.00	Clinton +6	47%	41%
MAR. 30-APR. 3	SurveyUSA	767 LV	0.00	Clinton +14	53%	39%
MAR. 16-23	USC Dornsife	363 LV	0.00	Clinton +11	47%	36%

Show more polls ▾

*Leader or runner-up is not in the race.

If you can't find a contest in the dropdown menu above, it's because there hasn't been enough polling in that state yet. We'll add new polling averages and forecasts as soon as the data is available. Notice any bugs or missing polls? [Send us an email](#).

By [Jay Boice](#), [Aaron Bycoffe](#), [Harry Enten](#), [Ritchie King](#), [Dhrumil Mehta](#), [Andrei Scheinkman](#) and [Nate Silver](#).

● = NEW A = ALL ADULTS RV = REGISTERED VOTERS LV = LIKELY VOTERS

POLLSTER

- MAY 19-22 SurveyUSA

- MAY 13-22 PPIC

APR. 27-30 SurveyUSA

APR. 28-MAY 1 Sextant Strategies

APR. 18-21 Fox News

APR. 13-15 YouGov

APR. 7-10 Gravis Marketing

MAR. 24-APR. 4 Field Poll

MAR. 30-APR. 3 SurveyUSA

MAR. 16-23 USC Dornsife

SAMPLE	WEIGHT	LEADER	CLINTON	SANDERS
803 LV	1.48	Clinton +18	57%	39%
552 LV	0.34	Clinton +2	46%	44%
826 LV	0.05	Clinton +19	57%	38%
1,617 RV	0.04	Clinton +10	49%	39%
623 LV	0.01	Clinton +2	48%	46%
1,123 LV	0.01	Clinton +12	52%	40%
846 LV	0.00	Clinton +6	47%	41%
584 LV	0.00	Clinton +6	47%	41%
767 LV	0.00	Clinton +14	53%	39%
363 LV	0.00	Clinton +11	47%	36%

Show more polls ▾

*Leader or runner-up is not in the race.

If you can't find a contest in the dropdown menu above, it's because there hasn't been enough polling in that state yet. We'll add new polling averages and forecasts as soon as the data is available. Notice any bugs or missing polls? [Send us an email](#).

By [Jay Boice](#), [Aaron Bycoffe](#), [Harry Enten](#), [Ritchie King](#), [Dhrumil Mehta](#), [Andrei Scheinkman](#) and [Nate Silver](#).

Reasoning in an Uncertain World

Questions

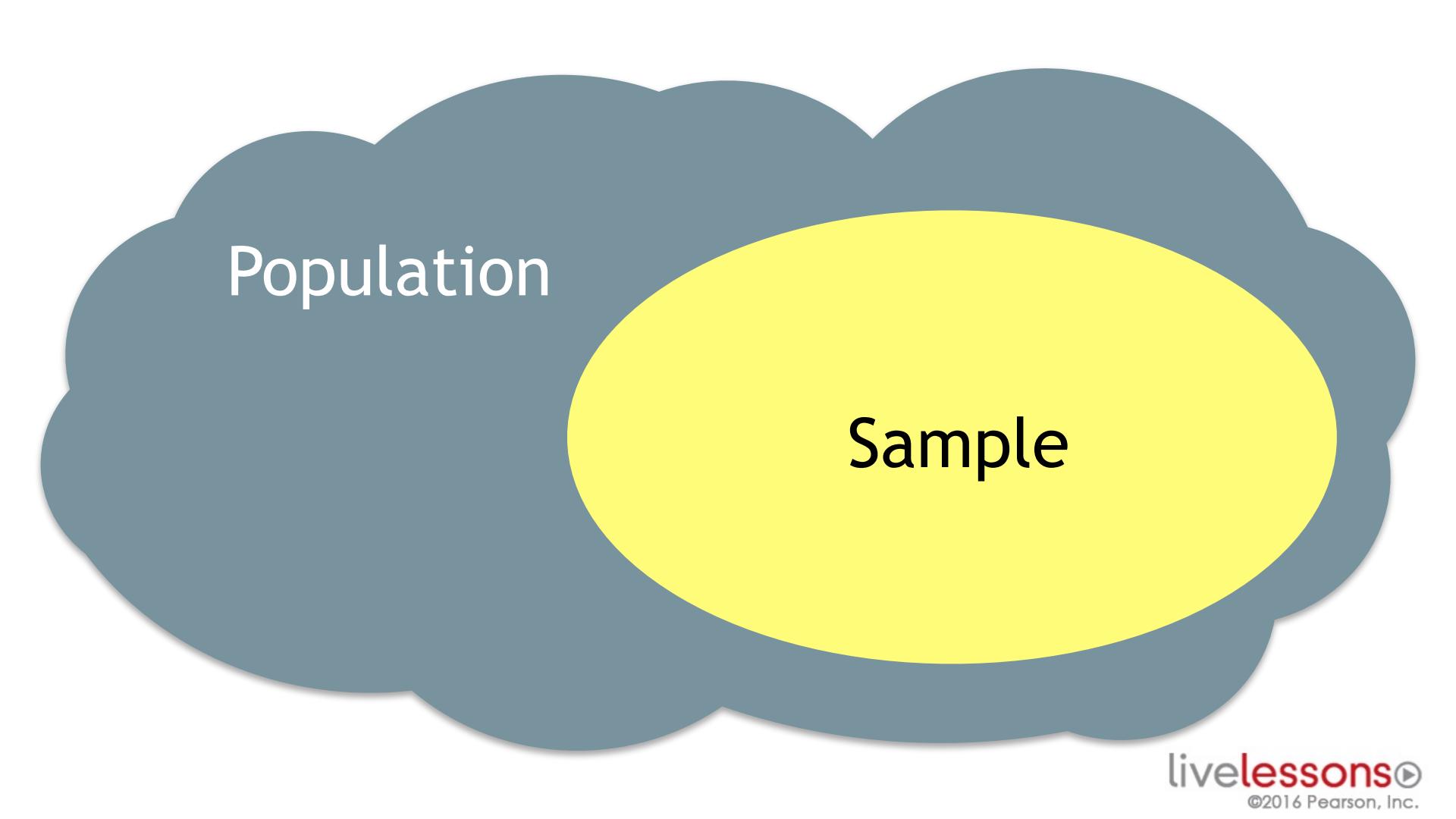
- Do US eligible voters favor Hillary Clinton over Bernie Sanders?
- Is drug A effective at treating a medical condition?
- Has the [Facebook News Feed](#) significantly increased the amount of time users spend on the platform?
- Is Uber making [NYC rush hour traffic](#) worse?
- Has an unregulated AirBnB marketplace led to an increasing rate of [evictions in San Francisco](#)?

At the end of the day we want to reason in an informed way to make decisions

(hypothetical) Decisions

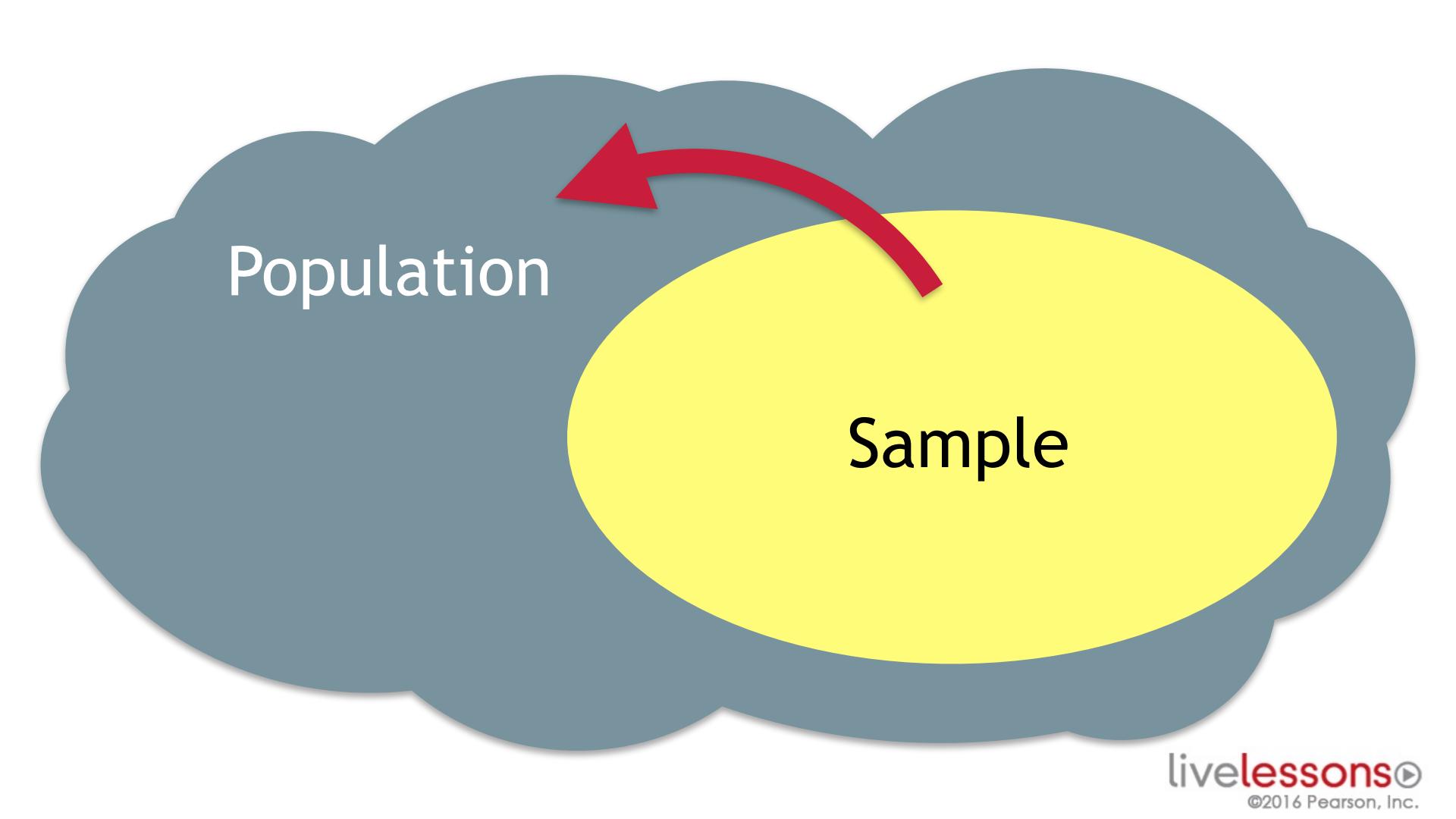
- Hillary Clinton should be the presumptive Democratic nominee
- Approve drug **A** for consumer use
- Roll out the Facebook News Feed interface to all users permanently
- Cap the number of Uber vehicles (and other services) in NYC
- Restrict private short term housing rentals in San Francisco to 75 nights per year

In Pictures



Population

Sample



Population

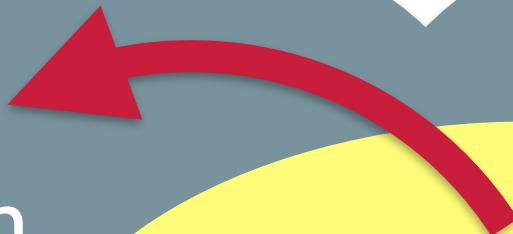
Sample

US Population
that is of
Voting Age

Individuals who
have answered
a Poll

US Population
that is of
Voting Age

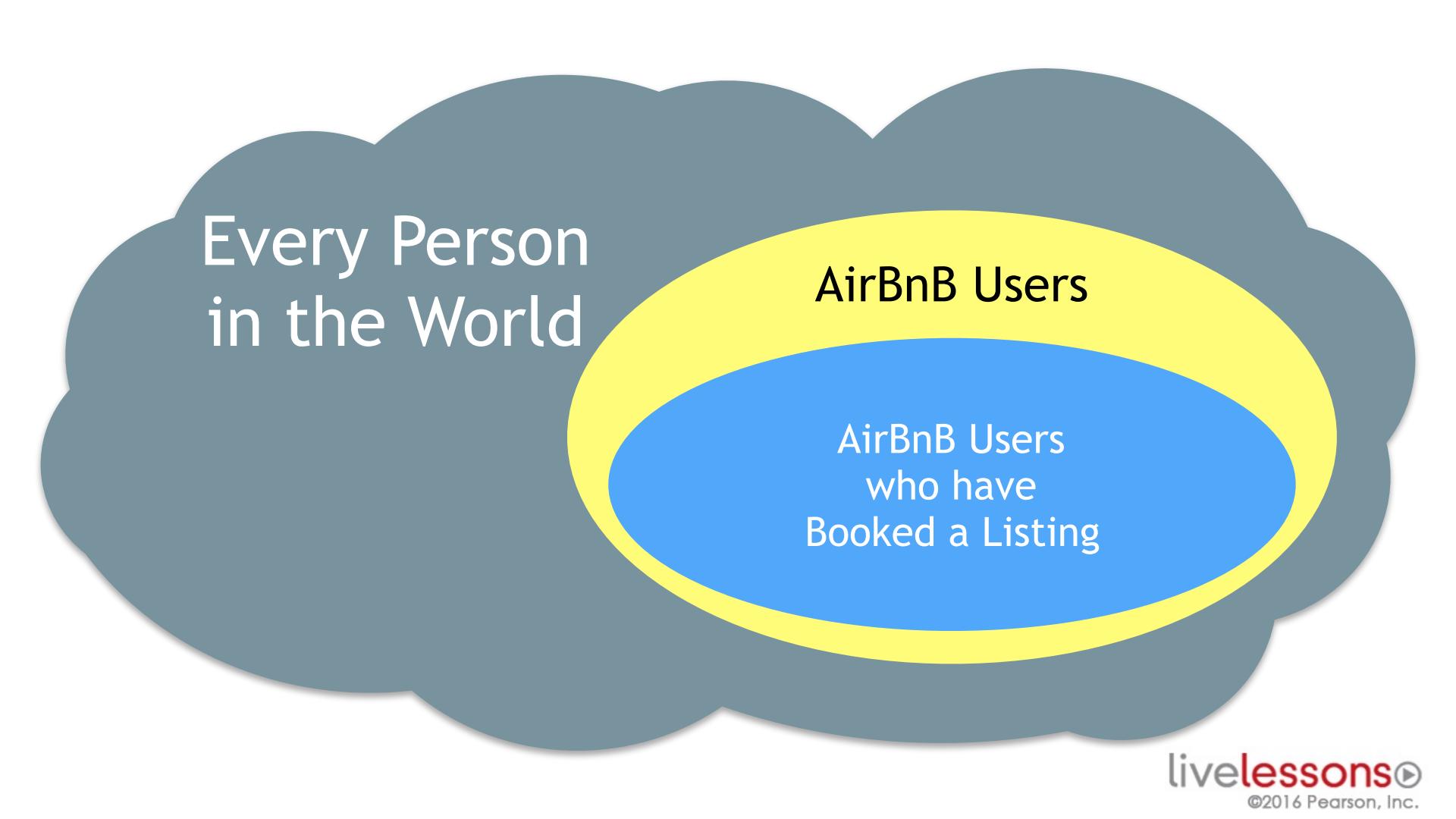
Individuals who
have answered
a Poll



A Venn diagram consisting of two overlapping circles. The larger circle is dark blue and labeled "Every Person in the World". The smaller circle inside it is yellow and labeled "AirBnB Users". The two circles overlap significantly.

Every Person
in the World

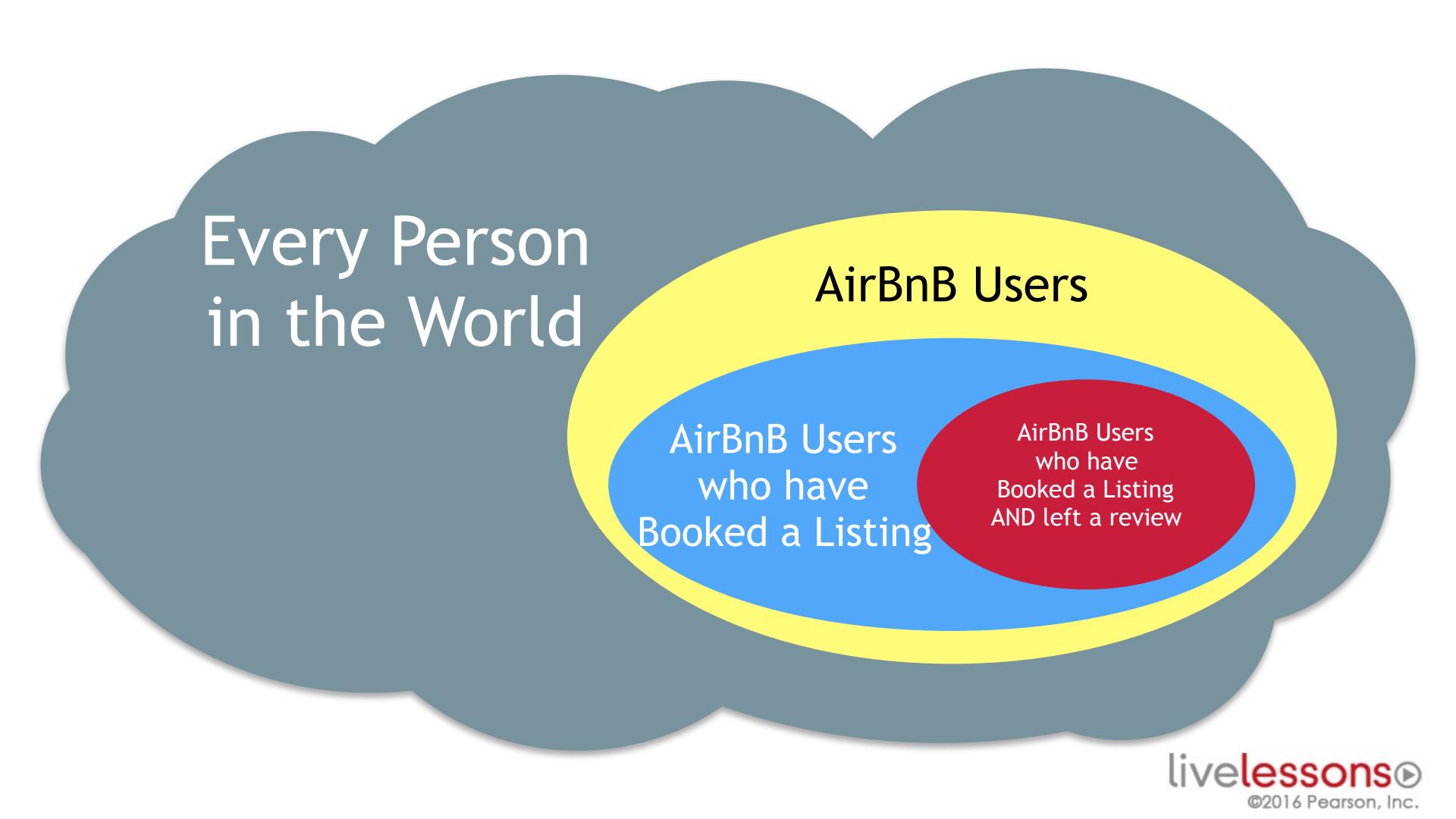
AirBnB Users



Every Person
in the World

AirBnB Users

AirBnB Users
who have
Booked a Listing

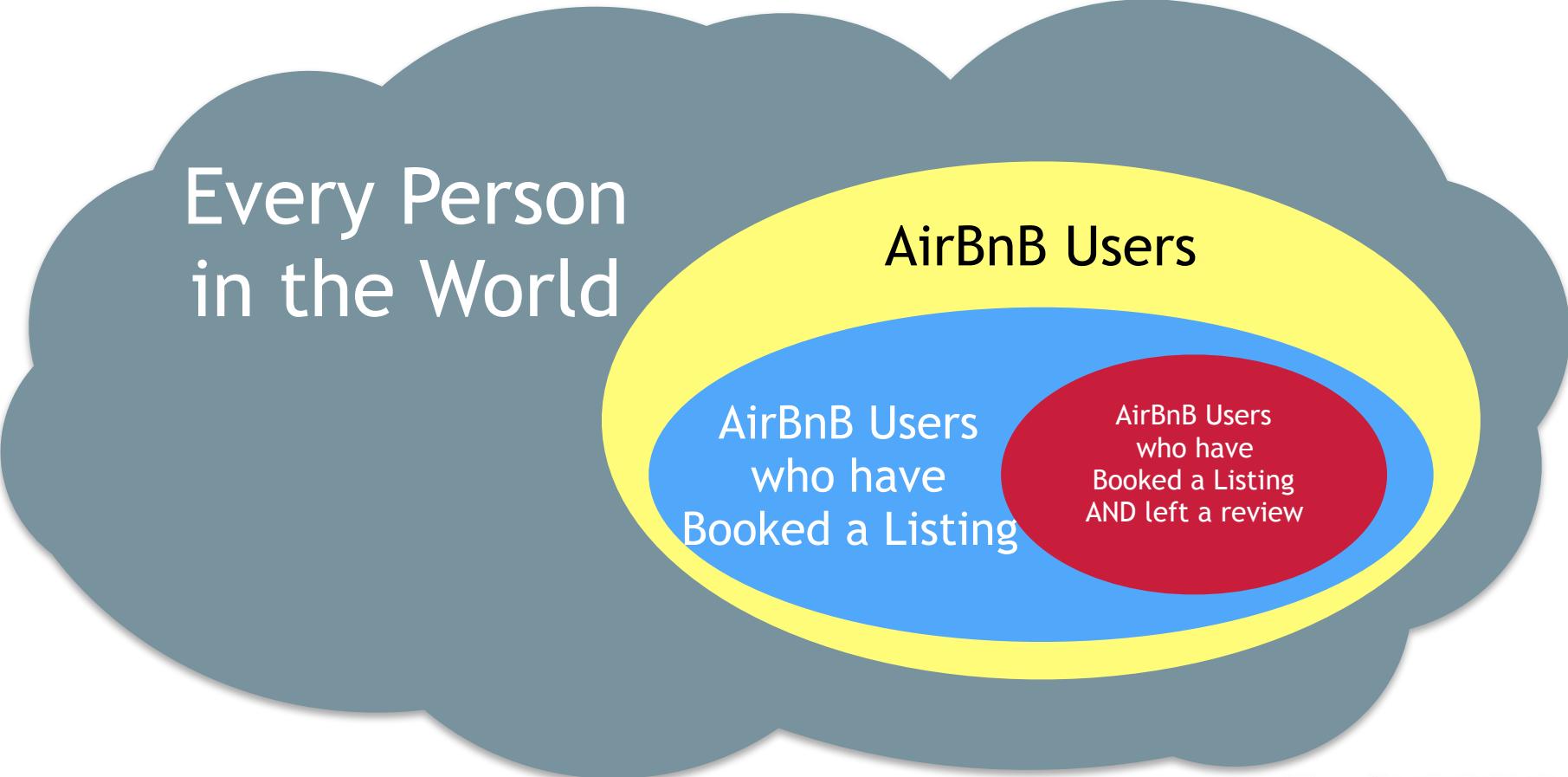


Every Person
in the World

AirBnB Users

AirBnB Users
who have
Booked a Listing

AirBnB Users
who have
Booked a Listing
AND left a review



Every Person
in the World

AirBnB Users

AirBnB Users
who have
Booked a Listing

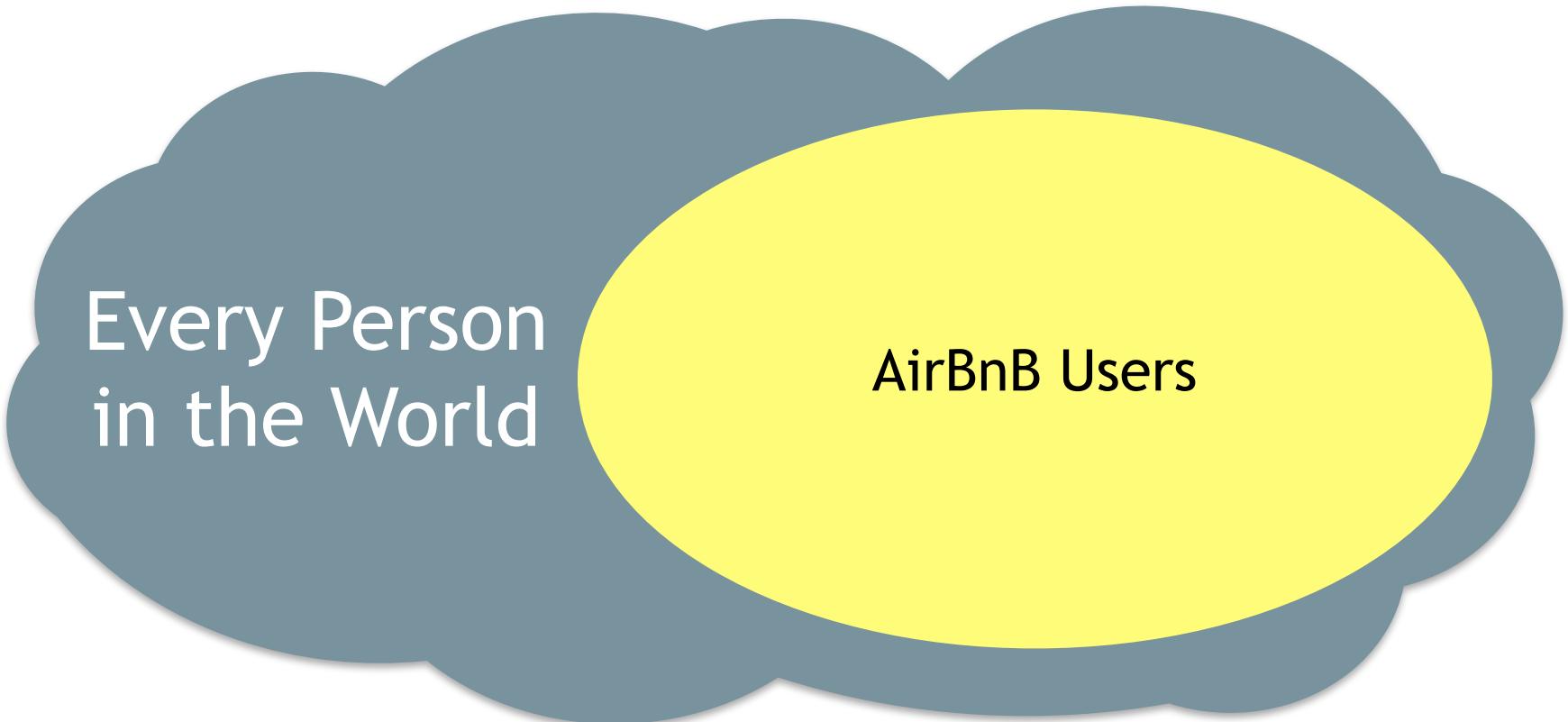
AirBnB Users
who have
Booked a Listing
AND left a review

* Not to Scale

We Want This

Every Person
in the World

Or Even This...



But Only Have This...

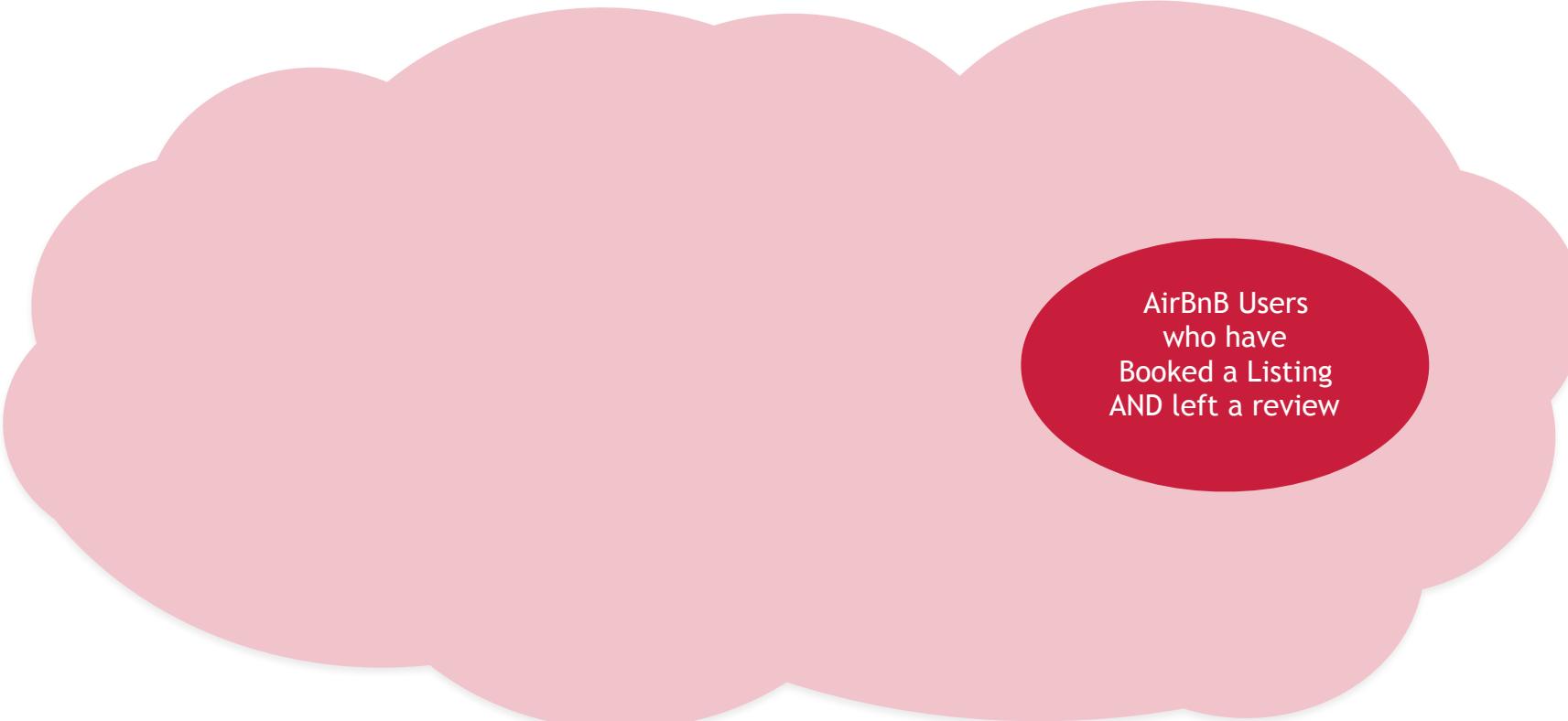


Every Person
in the World



AirBnB Users
who have
Booked a Listing
AND left a review

Uncertainty



AirBnB Users
who have
Booked a Listing
AND left a review

Goals

- Estimate Quantities of Interest (i.e. statistics/parameters)
- Quantify Uncertainty (i.e. confidence intervals)

Statistics

Descriptive

- Central Tendency (mean, median, mode, etc.)
- Dispersion (variance and standard deviation)
- Quantiles (min/max, median)
- Bivariate (correlation and covariance)

Inferential

- Estimation (MoM, MLE, MAP)
- Confidence Intervals
- Hypothesis Testing
- Bootstrap Methods
- Regression

Predictive

- Modeling (generative distributions)
- Machine Learning
 - Linear Regression
 - Logistic Regression
 - K Nearest Neighbors

Statistics

Descriptive

- Central Tendency (mean, median, mode, etc.)
- Dispersion (variance and standard deviation)
- Quantiles (min/max, median)
- Bivariate (**correlation** and **covariance**)

Inferential

- Estimation (MoM, MLE, MAP)
- Confidence Intervals
- Hypothesis Testing
- Bootstrap Methods
- Regression

Predictive

- Modeling (generative distributions)
- Machine Learning
 - **Linear Regression**
 - **Logistic Regression**
 - K Nearest Neighbors

Statistics

Descriptive

- Central Tendency (mean, median, mode, etc.)
- Dispersion (variance and standard deviation)
- Quantiles (min/max, median)
- Bivariate (correlation and covariance)

Inferential

- Estimation (MoM, MLE, MAP)
- Confidence Intervals
- Hypothesis Testing
- Bootstrap Methods
- Regression

Predictive

- Modeling (generative distributions)
- Machine Learning
 - Linear Regression
 - Logistic Regression
 - K Nearest Neighbors

Moments (mean, median, etc.)

Average

$$A = \frac{1}{n} \sum_{i=1}^n a_i$$

Expected Value

$$\mu = \sum xP(x)$$

Statistics

Descriptive

- Central Tendency (mean, median, mode, etc.)
- Dispersion (variance and standard deviation)
- Quantiles (min/max, median)
- Bivariate (correlation and covariance)

Inferential

- Estimation (MoM, MLE, MAP)
- Confidence Intervals
- Hypothesis Testing
- Bootstrap Methods
- Regression

Predictive

- Modeling (generative distributions)
- Machine Learning
 - Linear Regression
 - Logistic Regression
 - K Nearest Neighbors



Helpful... but not enough to answer our questions

Statistics

Understand

- Central Tendency (mean, median, mode, etc.)
- Dispersion (variance and standard deviation)
- Quantiles (min/max, median)
- Bivariate (correlation and covariance)

Decide

- Estimation (MoM, MLE, MAP)
- Confidence Intervals
- Hypothesis Testing
- Bootstrap Methods
- Regression

Forecast

- Modeling (generative distributions)
- Machine Learning
 - Linear Regression
 - Logistic Regression
 - K Nearest Neighbors

Statistics

Past

- Central Tendency (mean, median, mode, etc.)
- Dispersion (variance and standard deviation)
- Quantiles (min/max, median)
- Bivariate (correlation and covariance)

Present

- Estimation (MoM, MLE, MAP)
- Confidence Intervals
- Hypothesis Testing
- Bootstrap Methods
- Regression

Future

- Modeling (generative distributions)
- Machine Learning
 - Linear Regression
 - Logistic Regression
 - K Nearest Neighbors

Statistics

Lesson 7 & 8

- Central Tendency (mean, median, mode, etc.)
- Dispersion (variance and standard deviation)
- Quantiles (min/max, median)
- Bivariate (correlation and covariance)

Lesson 8 & 9

- Estimation (MoM, MLE, MAP)
- Confidence Intervals
- Hypothesis Testing
- Bootstrap Methods
- Regression

Lesson 9+

- Modeling (generative distributions)
- Machine Learning
 - Linear Regression
 - Logistic Regression
 - K Nearest Neighbors

Statistics

Lesson 7 & 8

- Central Tendency (mean, median, mode, etc.)
- Dispersion (variance and standard deviation)
- Quantiles (min/max, median)
- Bivariate (correlation and covariance)

Lesson 8 & 9

- Estimation (MoM, MLE, MAP)
- Confidence Intervals
- Hypothesis Testing
- Bootstrap Methods
- Regression

Lesson 9+

- Modeling (generative distributions)
- Machine Learning
 - Linear Regression
 - Logistic Regression
 - K Nearest Neighbors

Questions

AirBnB

- Does recommending listings increase the number of reservations that users make on the AirBnB platform?
- Do listings in neighborhoods with more restaurants and shops attract more visitors?

The Public

- Does the presence of AirBnB in a city cause rents to increase more than they would have otherwise?
- Has AirBnB contributed to falling hotel rates in NYC?

Questions

AirBnB

- Does recommending listings increase the number of reservations that users make on the AirBnB platform?
- Do listings in neighborhoods with more restaurants and shops attract more visitors?

The Public

- Does the presence of AirBnB in a city cause rents to increase more than they would have otherwise?
- Has AirBnB contributed to falling hotel rates in NYC?

Let's try to first solve a simpler problem

That doesn't involve a comparison

The simple, humble, mean

Questions

AirBnB

- Does recommending listings increase the number of reservations that users make on the AirBnB platform?
- What are the “best” restaurants in each neighborhood?

The Public



Simplify!

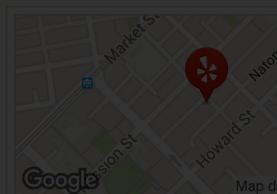
- Does the presence of AirBnB in a city cause rents to increase more than they would have otherwise?
- Has AirBnB contributed to falling hotel rates in NYC?

San Francisco > Financial District >

Yelp

4.5 stars 8381 reviews

Local Flavor, Mass Media



📍 140 New Montgomery St
San Francisco, CA 94105
b/t Natoma St & Minna St

Financial District, SoMa
Get Directions
(415) 908-3801

officialblog.yelp.com



"Locally, I enjoy being the way I do about writing reviews."



"Reviews galore, take it or leave it. I know, love, hate or whatever."



"Of course, all those reviews are great, I'm looking forward to attending more events."

Find tacos, cheap dinner, Max's

Near Noe Valley, San Francisco, CA

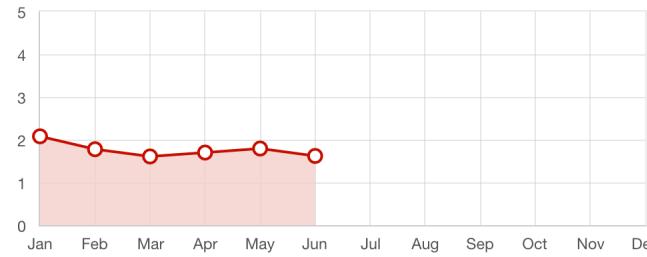


Home About Me Write a Review Find Friends Messages Talk Events

Rating Details

Monthly Trend

2016 2015 2014 2013 2012



Understand how a business' rating changes month-to-month. [Learn more.](#)

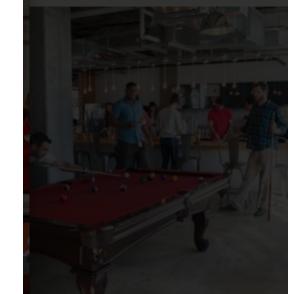
Overall Rating

Yelping since 2004 with 8381 reviews



We calculate the overall star rating using only reviews that our automated software currently recommends. [Learn more.](#)

Photo Share Bookmark



See all 1240

8:00 am - 6:00 pm Open now

8:00 am - 6:00 pm
8:00 am - 6:00 pm
8:00 am - 6:00 pm
8:00 am - 6:00 pm Open now
8:00 am - 6:00 pm
Closed
Closed

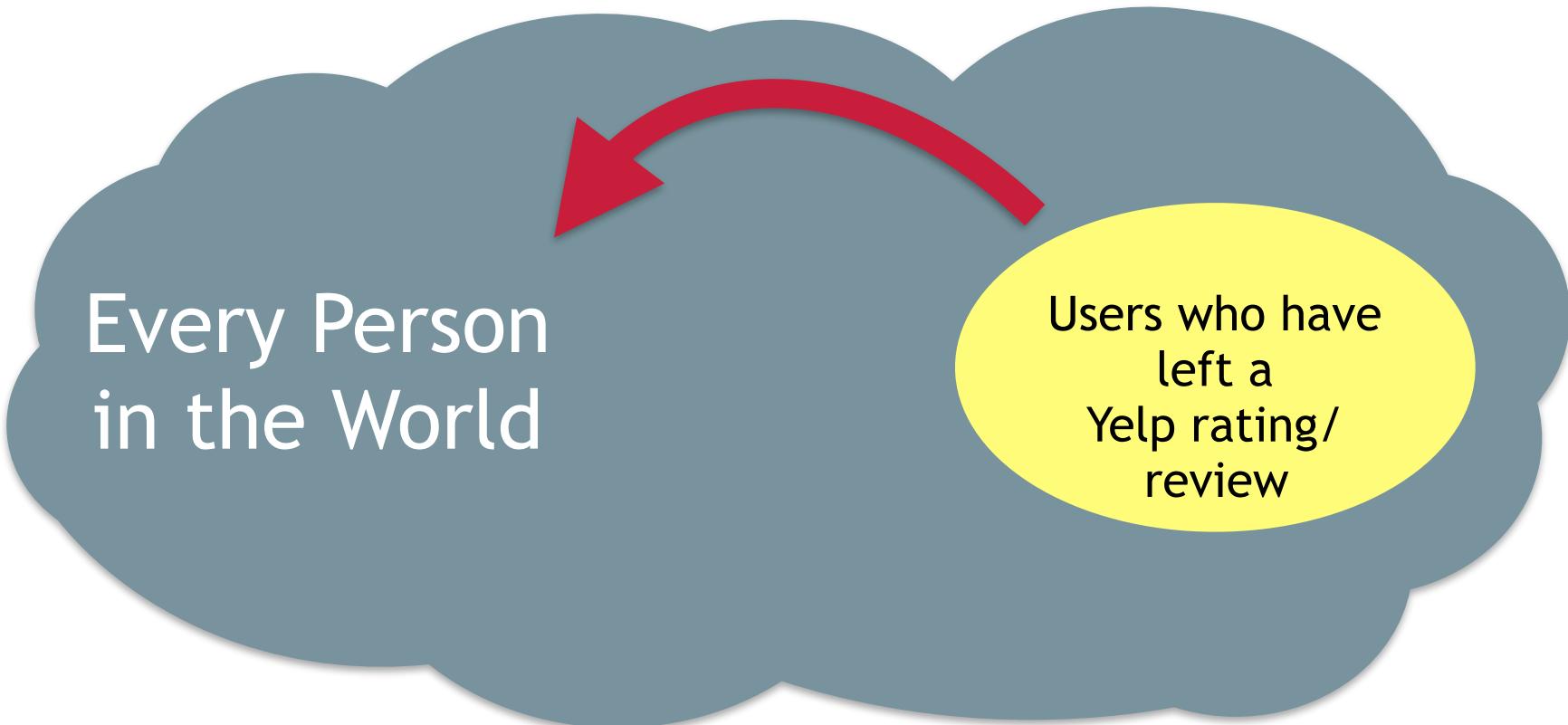
Sun

General Inference

Every Person
in the World

Users who have
left a
Yelp rating/
review

General Inference



Represent your Population by Your Sample

General Inference

Every Person
in the World

Users who have
left a
Yelp rating/
review

Sample your Sample

The Bootstrap

Every Person
in the World

Users who have
left a
Yelp rating/review

Bootstrap
Sample

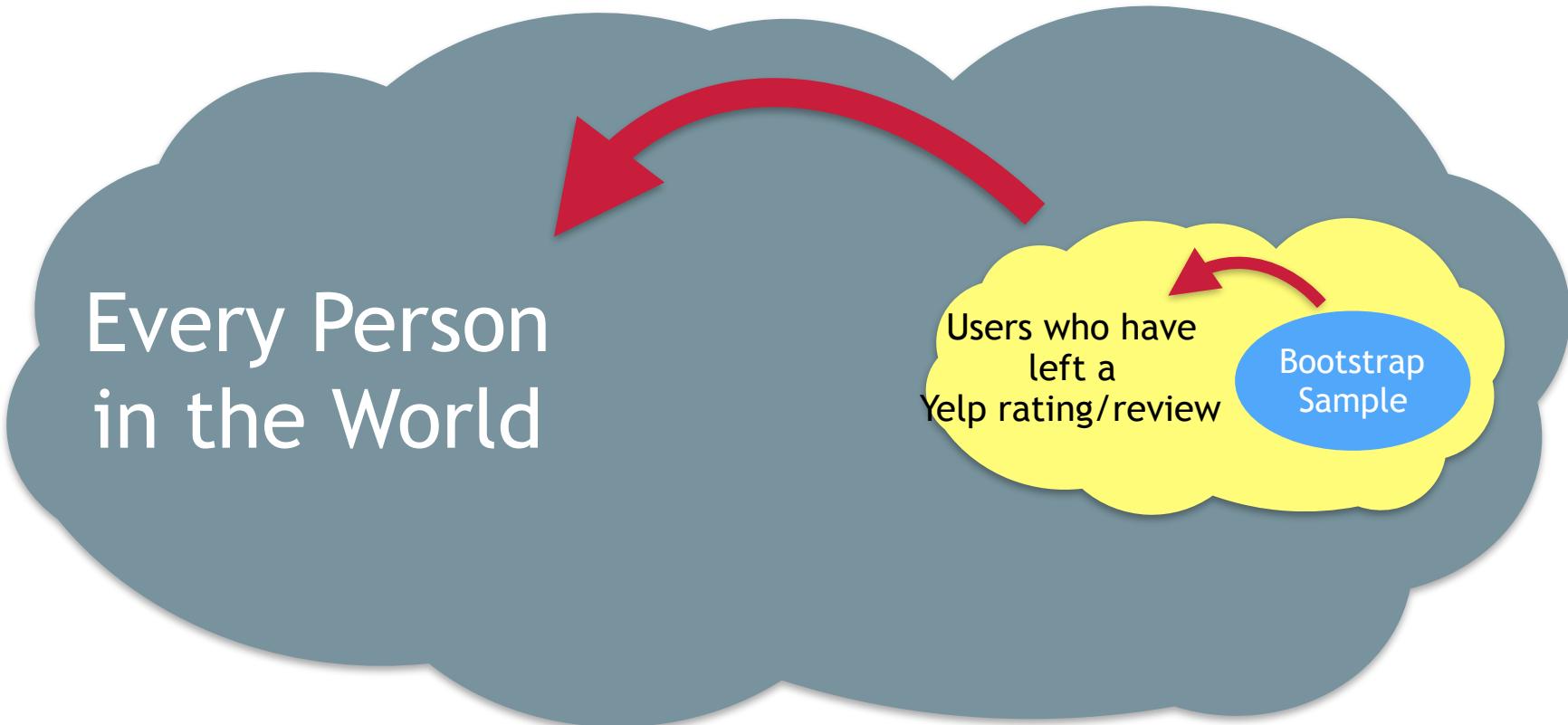
Sample → Population

Bootstrap sample -> Sample

Bootstrap sample → **Sample** → **Population**

* Works as long as the **sample** is somewhat representative of the population

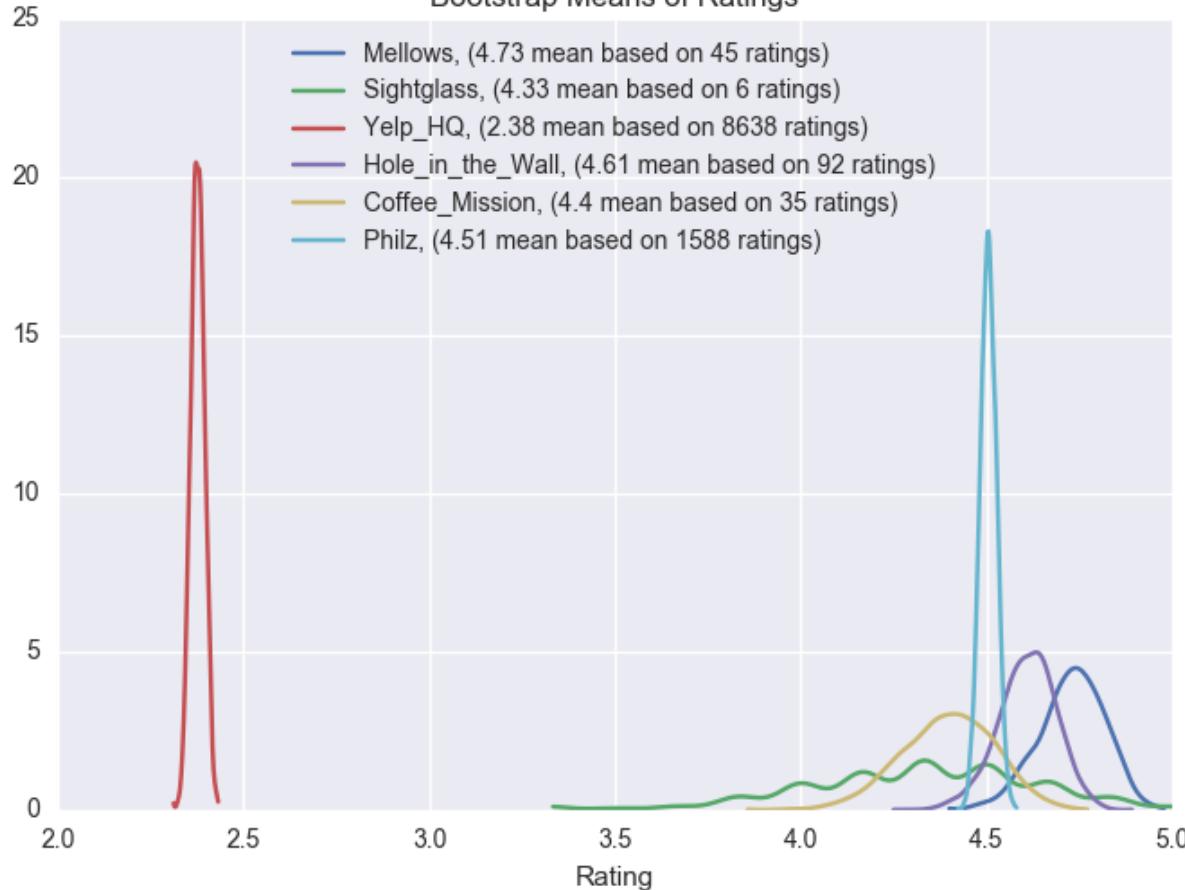
The Bootstrap



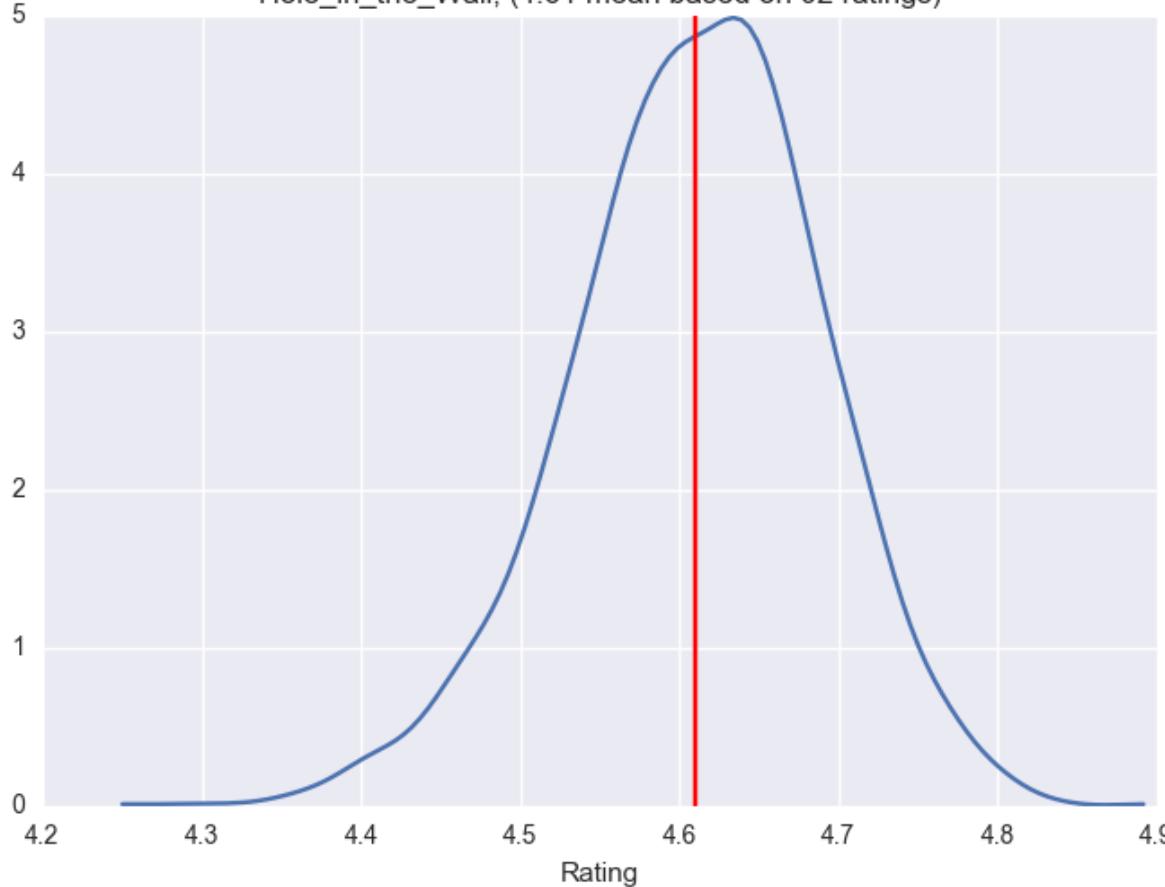
```
df = pd.read_csv('yelp_stars/' + filename)
num_samples = len(df)
boots = []

for i in range(2000):
    bootstrap = df.sample(num_samples, replace = True)
    boots.append(bootstrap.rating.mean())
```

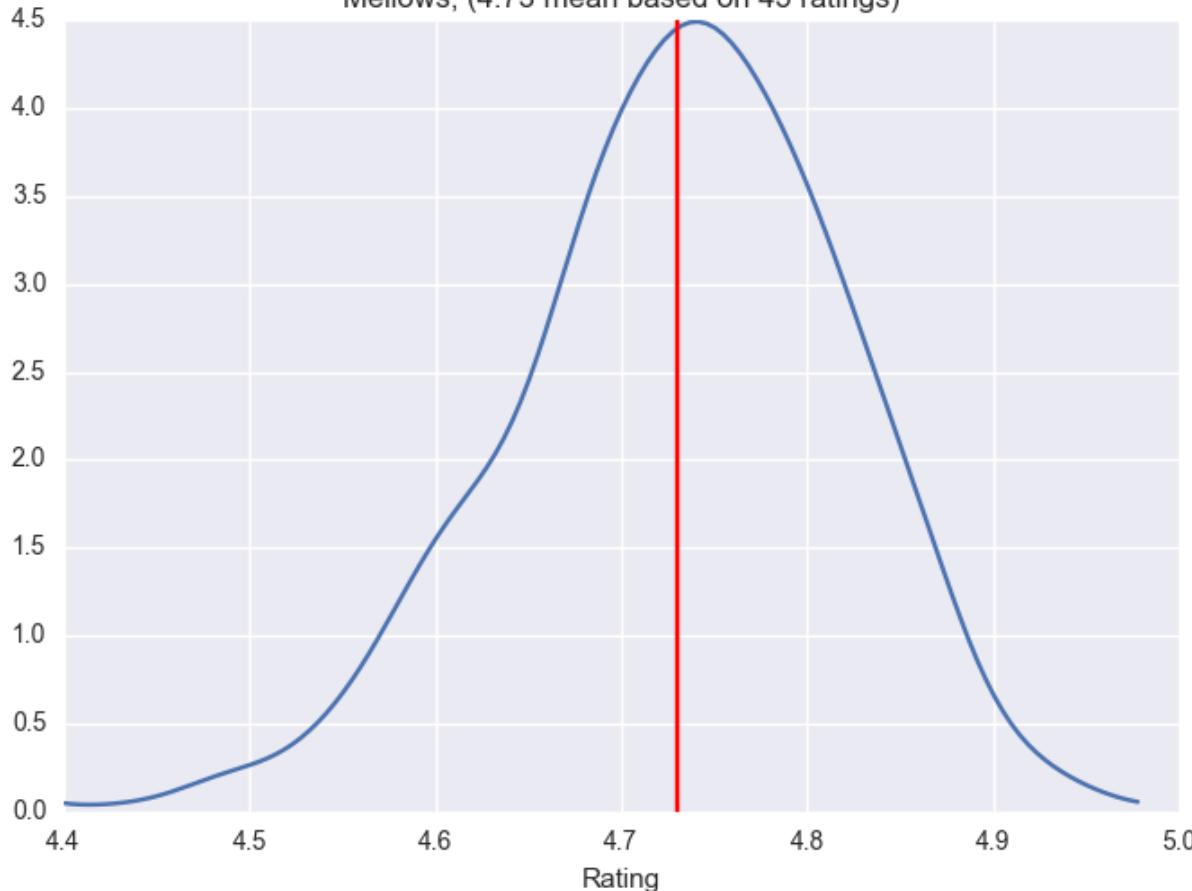
Bootstrap Means of Ratings



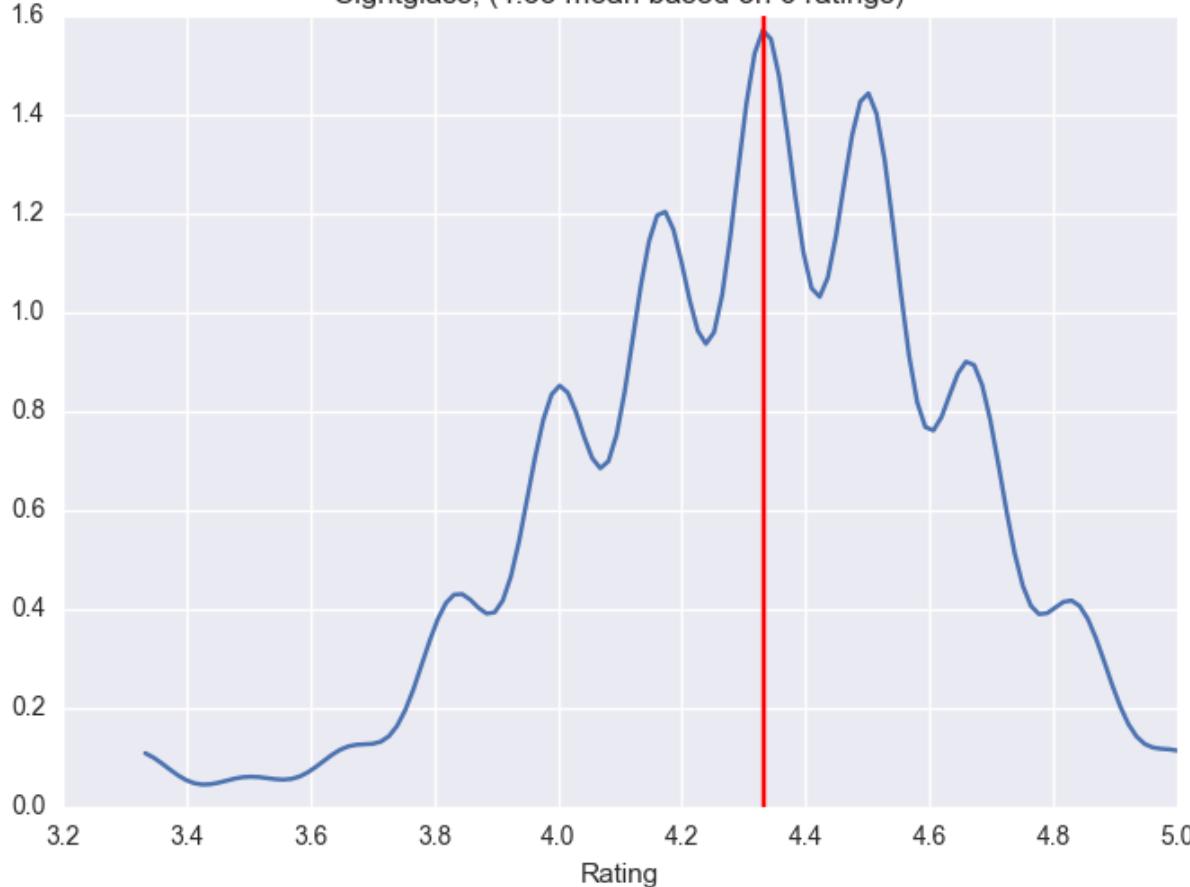
Hole_in_the_Wall, (4.61 mean based on 92 ratings)



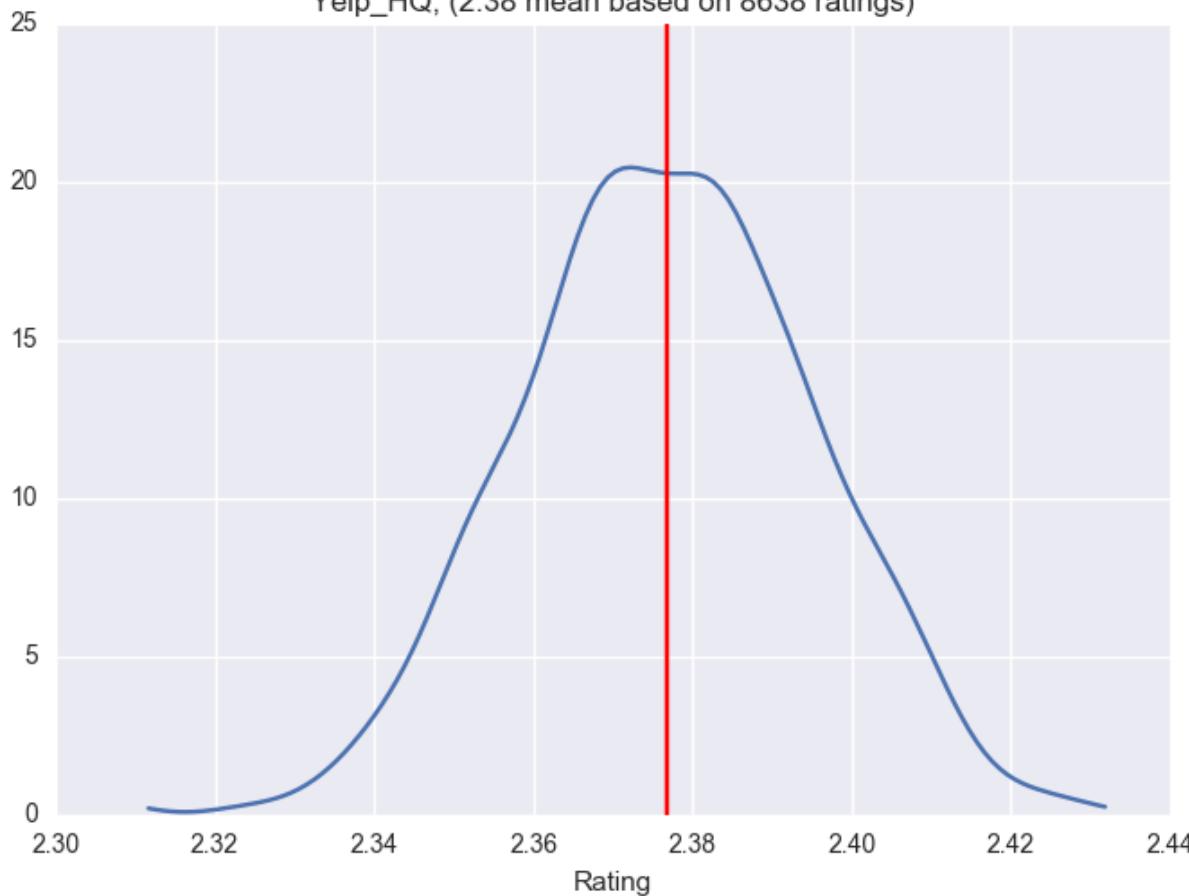
Mellows, (4.73 mean based on 45 ratings)



Sightglass, (4.33 mean based on 6 ratings)



Yelp_HQ, (2.38 mean based on 8638 ratings)



Given the ratings I have, there is a 95% chance that the "real" fraction of positive ratings is at least what?

Back to Non Parametrics

Just like non-parametric methods make no assumptions, I will make no assumptions of background or prerequisite knowledge either

But now we can use these ideas to understand uncertainty...

Goals

- Estimate Quantities of Interest (i.e. statistics/parameters)
- Quantify Uncertainty (i.e. confidence intervals)

Classic (frequentist) approach

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

Population Standard Deviation

Size of Sample

for a 95% CI

The diagram illustrates the formula for a 95% confidence interval: $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$. A red arrow points from the text "Population Standard Deviation" to the symbol σ . Another red arrow points from the text "Size of Sample" to the term \sqrt{n} . A third red arrow points from the text "for a 95% CI" to the multiplier 1.96 .

Why Bootstrap?

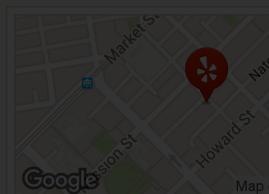
- Estimate arbitrary statistics (mean, percentiles, etc.)
- Distribution Agnostic (non-parametric)
- Natural Confidence Intervals

San Francisco > Financial District >

Yelp

★ ★ ★ ★ 8381 reviews

Local Flavor, Mass Media



📍 140 New Montgomery St
San Francisco, CA 94105

b/t Natoma St & Minna St
Financial District, SoMa

Get Directions

(415) 908-3801

officialblog.yelp.com



"Locally, I enjoy being the way I do about writing reviews."



"Reviews galore, take it or leave it. I know, love, hate or whatever."



"Of course, all those reviews are just what I forward to attend my next meeting."

Find tacos, cheap dinner, Max's

Near Noe Valley, San Francisco, CA

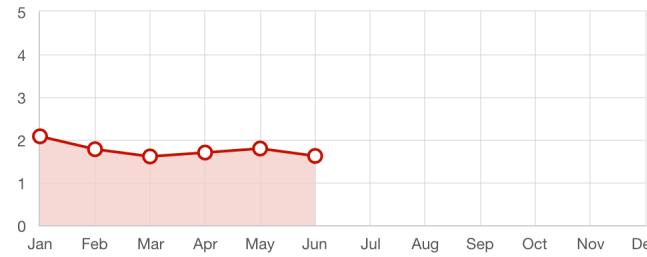


Home About Me Write a Review Find Friends Messages Talk Events

Rating Details

Monthly Trend

2016 2015 2014 2013 2012



Understand how a business' rating changes month-to-month. [Learn more.](#)

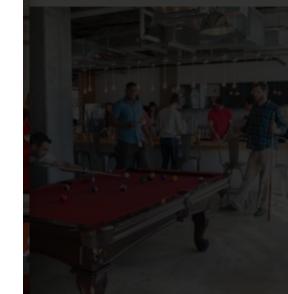
Overall Rating

Yelping since 2004 with 8381 reviews



We calculate the overall star rating using only reviews that our automated software currently recommends. [Learn more.](#)

Photo Share Bookmark



See all 1240

8:00 am - 6:00 pm Open now

:00 am - 6:00 pm

:00 am - 6:00 pm

:00 am - 6:00 pm

:00 am - 6:00 pm Open now

:00 am - 6:00 pm

Closed

Closed

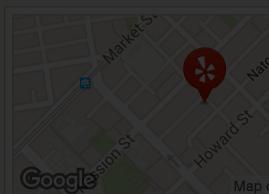
Sun

San Francisco > Financial District >

Yelp

4.5 stars 8381 reviews

Local Flavor, Mass Media



📍 140 New Montgomery St
San Francisco, CA 94105

b/t Natomas St & Minna St
Financial District, SoMa

Get Directions

(415) 908-3801

officialblog.yelp.com



"Locally, I enjoy being a part of the way I do about writing reviews."



"Reviews galore, take a look, love, hate or whatever."



"Of course, all those reviews forward to attend me."

Find tacos, cheap dinner, Max's

Near Noe Valley, San Francisco, CA

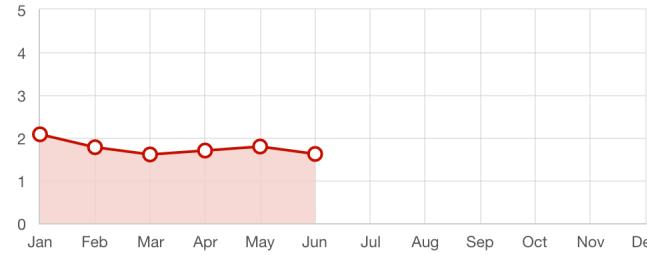


Home About Me Write a Review Find Friends Messages Talk Events

Rating Details

Monthly Trend

2016 2015 2014 2013 2012



Understand how a business' rating changes month-to-month. [Learn more.](#)

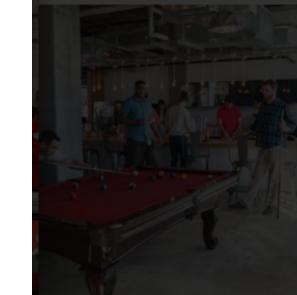
Overall Rating

Yelping since 2004 with 8381 reviews



We calculate the overall star rating using only reviews that our automated software currently recommends. [Learn more.](#)

Photo Share Bookmark



See all 1240 reviews

8:00 am - 6:00 pm Open now

8:00 am - 6:00 pm

8:00 am - 6:00 pm

8:00 am - 6:00 pm

8:00 am - 6:00 pm Open now

8:00 am - 6:00 pm

8:00 am - 6:00 pm

8:00 am - 6:00 pm

Sun Closed

Definitely not normal

The Bootstrap Formalized

The Bootstrap

Every Person
in the World

Users who have
left a
Yelp rating/review

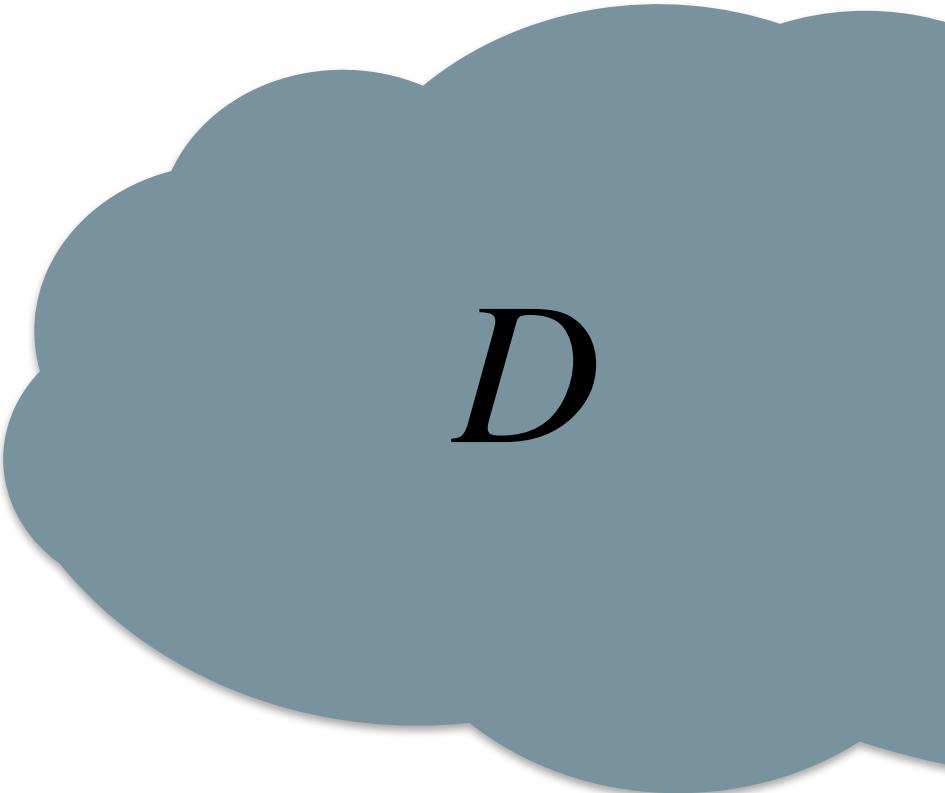
Bootstrap
Sample

The Bootstrap

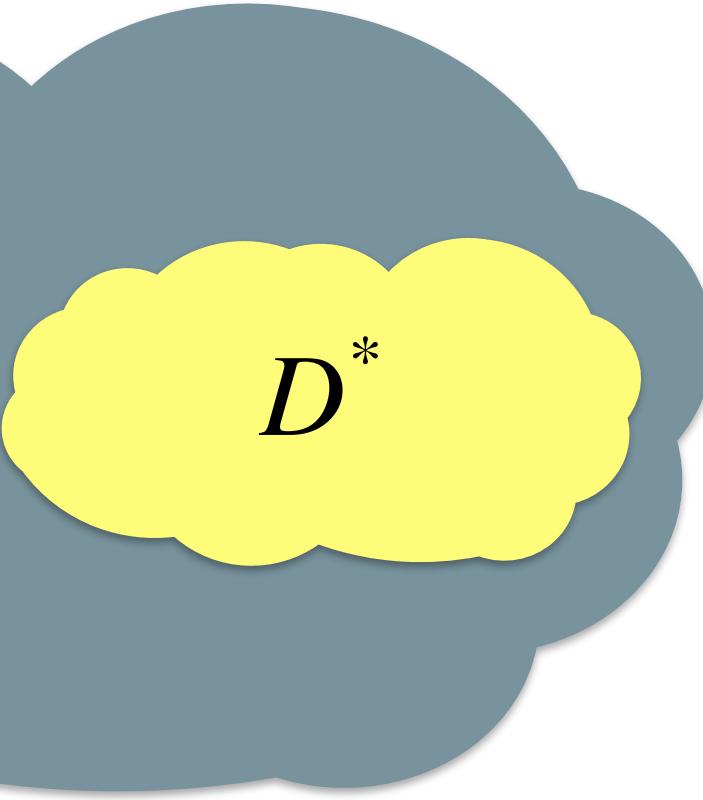
True Distribution

Empirical
Distribution

The Bootstrap



D



D^*

The Bootstrap



Why Bootstrap?

- Estimate arbitrary statistics (mean, percentiles, etc.)
- Distribution Agnostic (non-parametric)
- Natural Confidence Intervals

The Bootstrap Principle

Distribution Approximation

$$D^* \approx D$$

Statistic Approximation

For any statistic s , the (resampled) distribution of s^* approximates s

$$s^* \rightarrow s$$

**Just treat the confidence interval
as our statistic of interest**

Statistic of Interest

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

Difference in means
for one sample

Sample mean

True mean

The diagram illustrates the formula for the statistic of interest. It shows the equation $\theta = \bar{x} - \mu$. A red arrow points from the text "Difference in means for one sample" to the right side of the equation. Another red arrow points from the text "Sample mean" to the term \bar{x} . A third red arrow points from the text "True mean" to the term μ .

*For confidence intervals we are interested in
the **distribution** of the difference in means

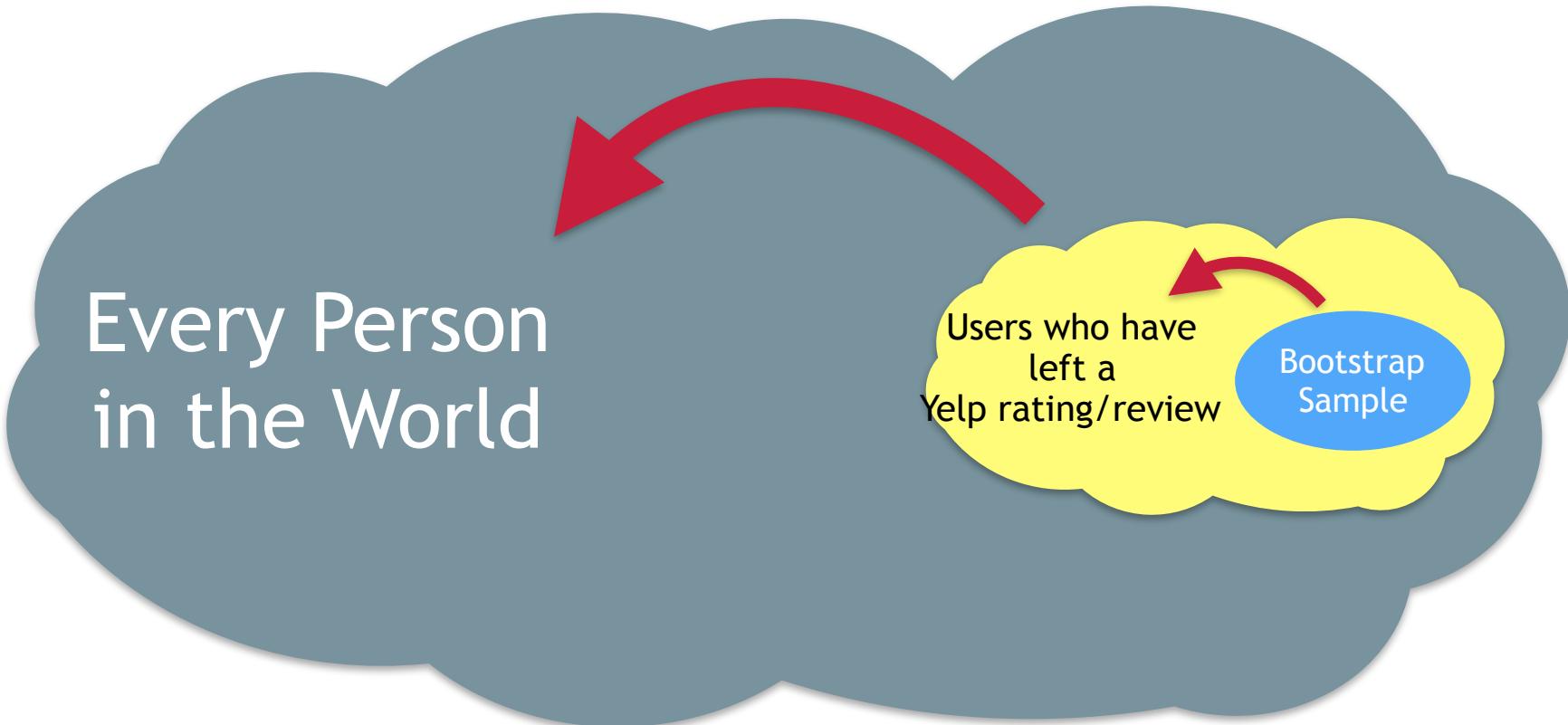
Bootstrap Approximation

For any statistic s , the (resampled) distribution of s^* approximates s

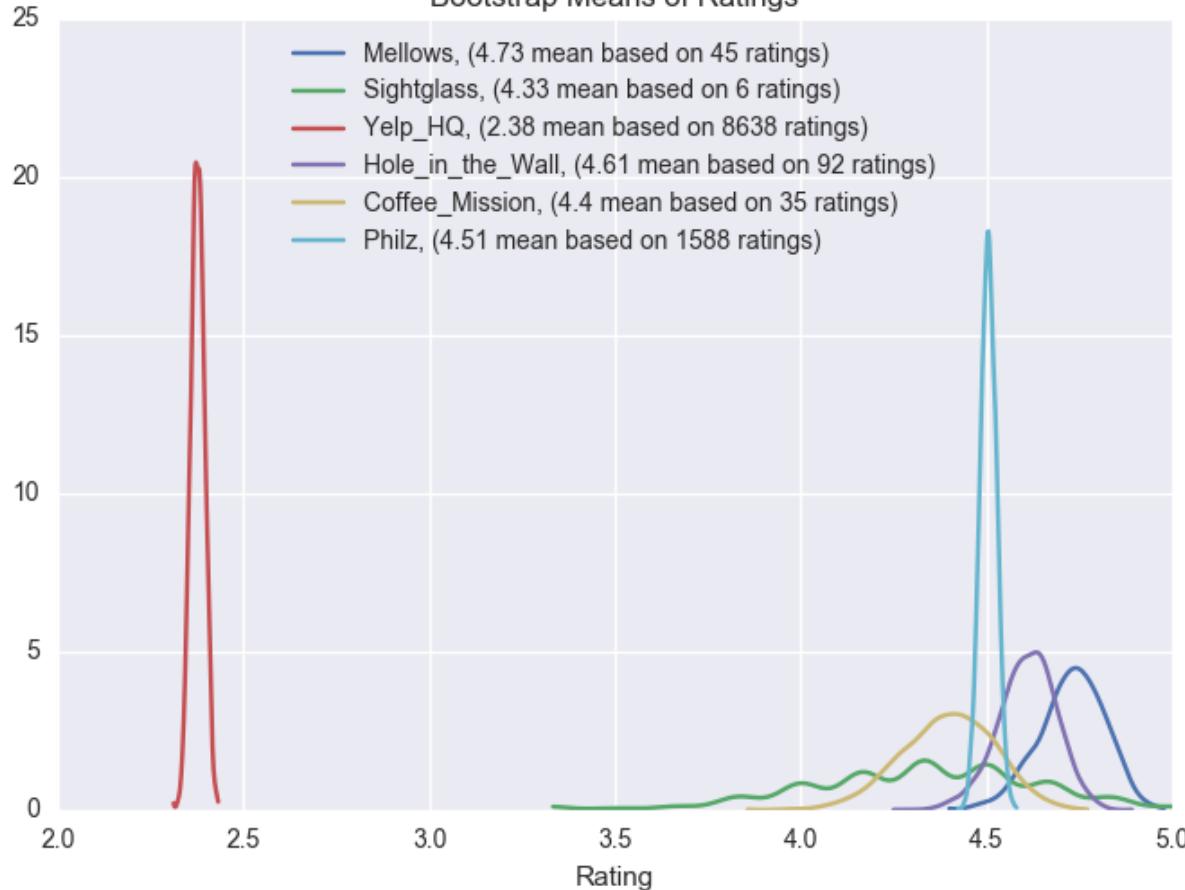
$$\theta^* = \bar{x}^* - \bar{x}$$

$$\theta^* \rightarrow \theta$$

The Bootstrap



Bootstrap Means of Ratings



```
df = pd.read_csv('yelp_stars/' + filename)
num_samples = len(df)
boots = []

for i in range(2000):
    bootstrap = df.sample(num_samples, replace = True)
    boots.append(bootstrap.rating.mean())
```



Statistic of interest

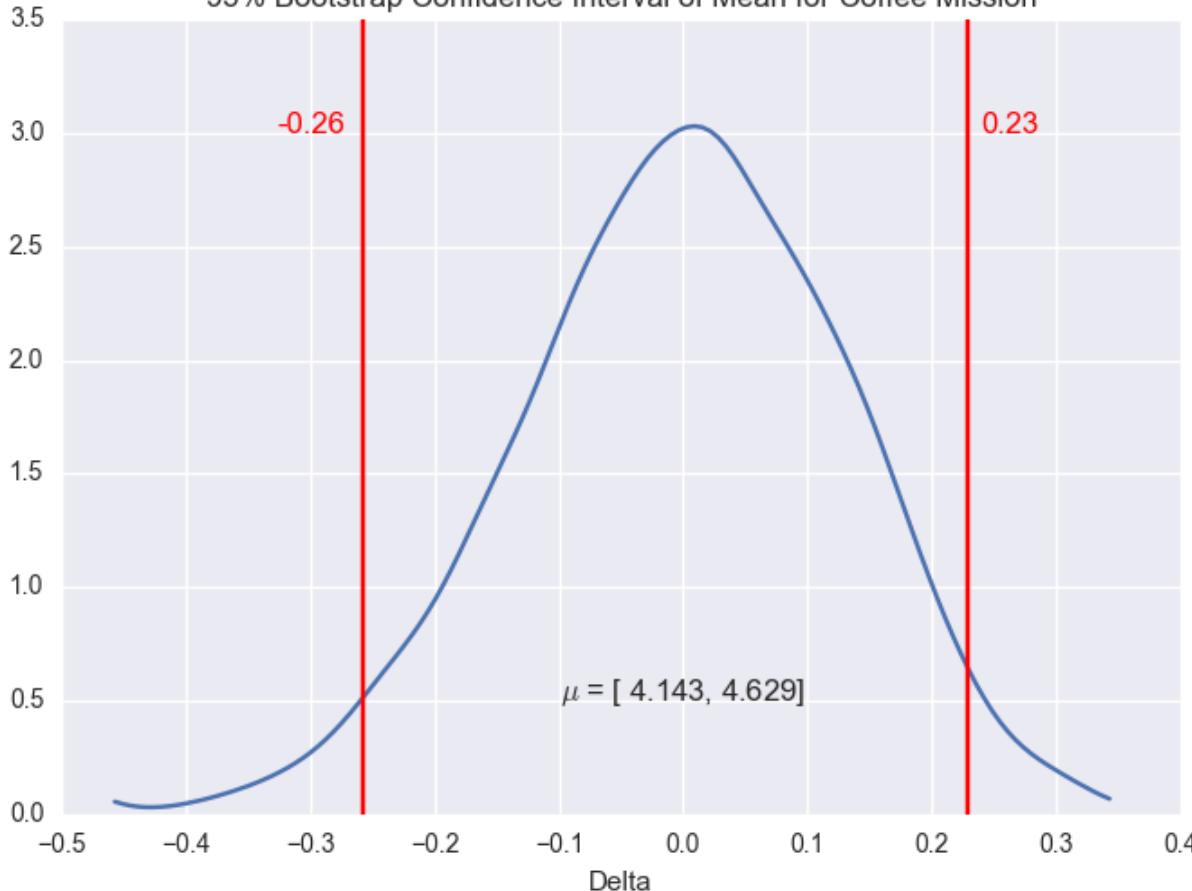
```
df = pd.read_csv('yelp_stars/' + filename)
num_samples = len(df)
boots = []
x_bar = df.rating.mean()

for i in range(2000):
    bootstrap = df.sample(num_samples, replace = True)
    boots.append(bootstrap.rating.mean() - x_bar)
```

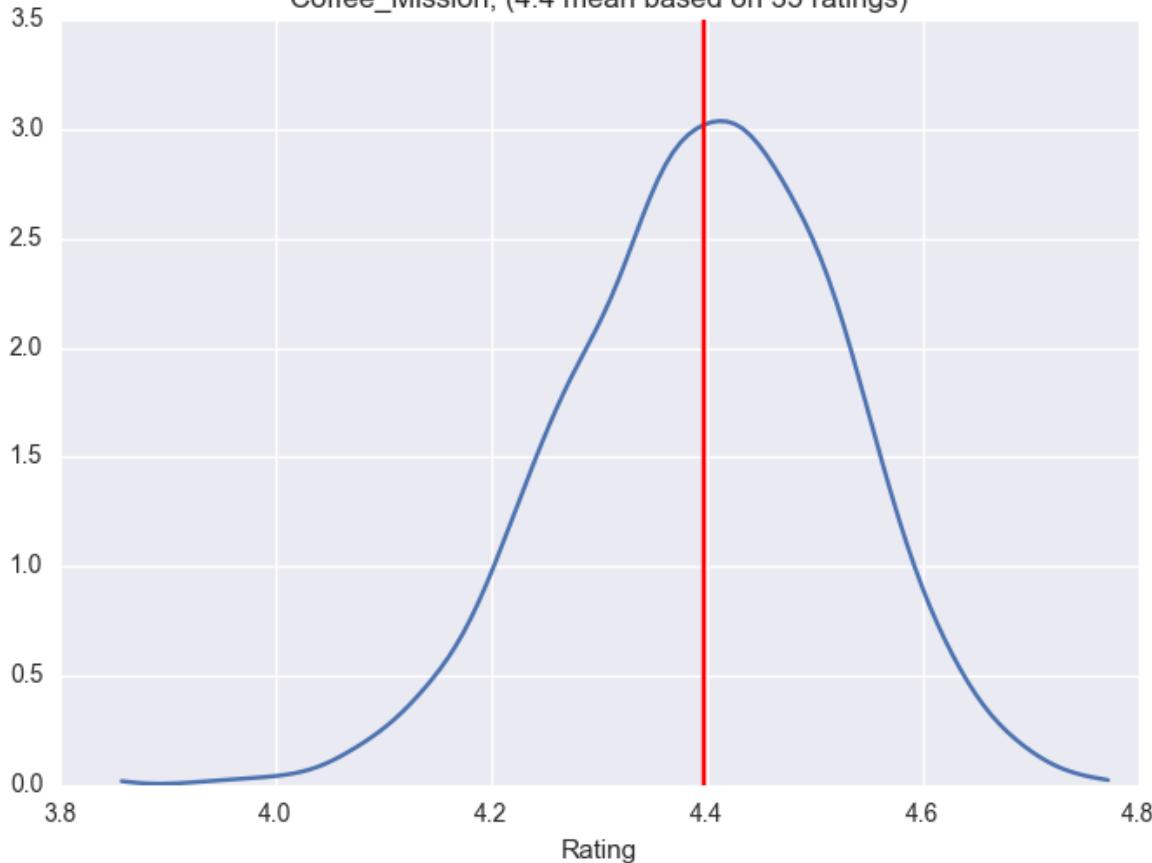


Statistic of interest

95% Bootstrap Confidence Interval of Mean for Coffee Mission



Coffee_Mission, (4.4 mean based on 35 ratings)



```
df = pd.read_csv('yelp_stars/Coffee_Mission.csv')
num_samples = len(df)
boots = []

for i in range(2000):
    bootstrap = df.sample(num_samples, replace = True)
    boots.append(bootstrap.rating.mean())

data_arr = np.array(boots)

>>>print("The 2.5th percentile is {0:.4f}".format(np.percentile(data_arr, 2.5)))
The 2.5th percentile is 4.1429

>>>print("The 97.5th percentile is {0:.4f}".format(np.percentile(data_arr, 97.5)))
The 97.5th percentile is 4.6286
```

Interpreting Confidence Intervals

an interactive visualization

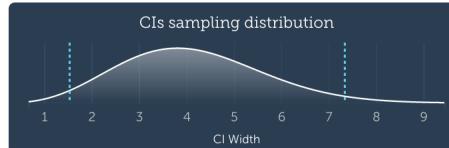
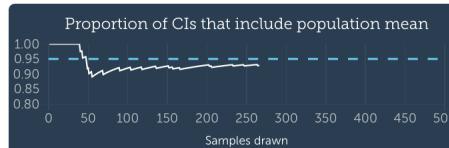
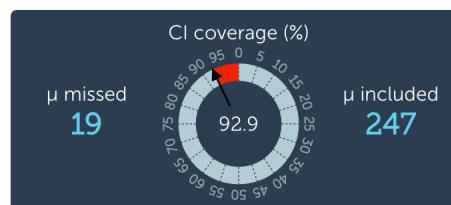
Created by Kristoffer Magnusson

Follow @krstoffer 1,298 follower Kristoffer's LinkedIn profile

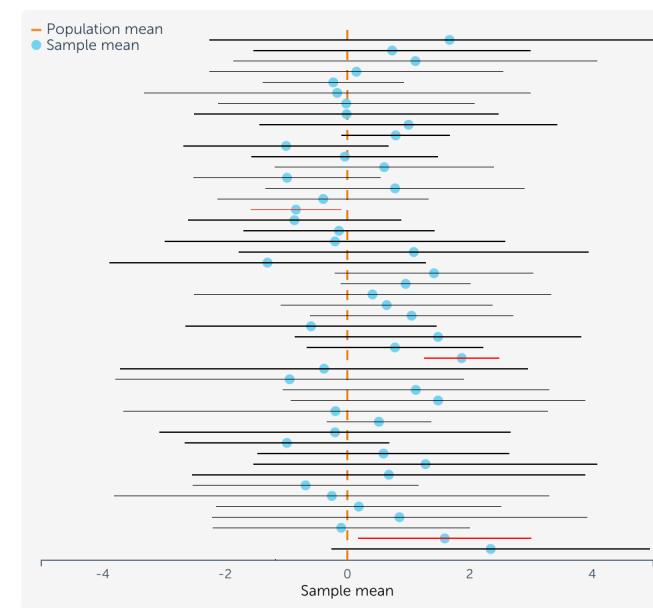
Tweet Gilla 226

Slide me

Simulation statistics



93% confidence intervals



Cheaper Beds, **Better** Breakfasts

Exploratory Question

What variables affect the price of a listing?

Hypothesis Statement

“Superhosts” have more expensive listings

Hypothesis Statement

“Superhosts” have more expensive listings

(implicit assumption: since they can attract more guests)

Confounding Variables

- Do they have the same volume of guests?
- Is it an equivalent price per # of guests or sq. footage?
- Or do they just happen to have bigger places?
- What else is correlated with price?

(correlation can be useful in determining causation)

Hypothesis Statement

More on this later...

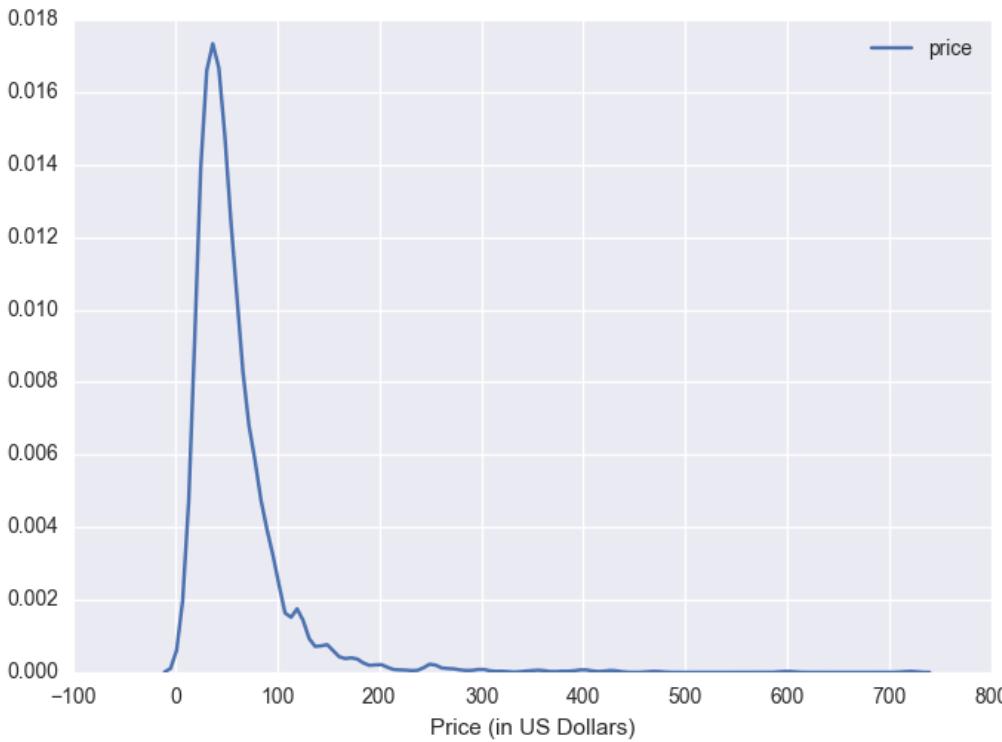
Question for Inference

What variables are correlated with the price of a listing?

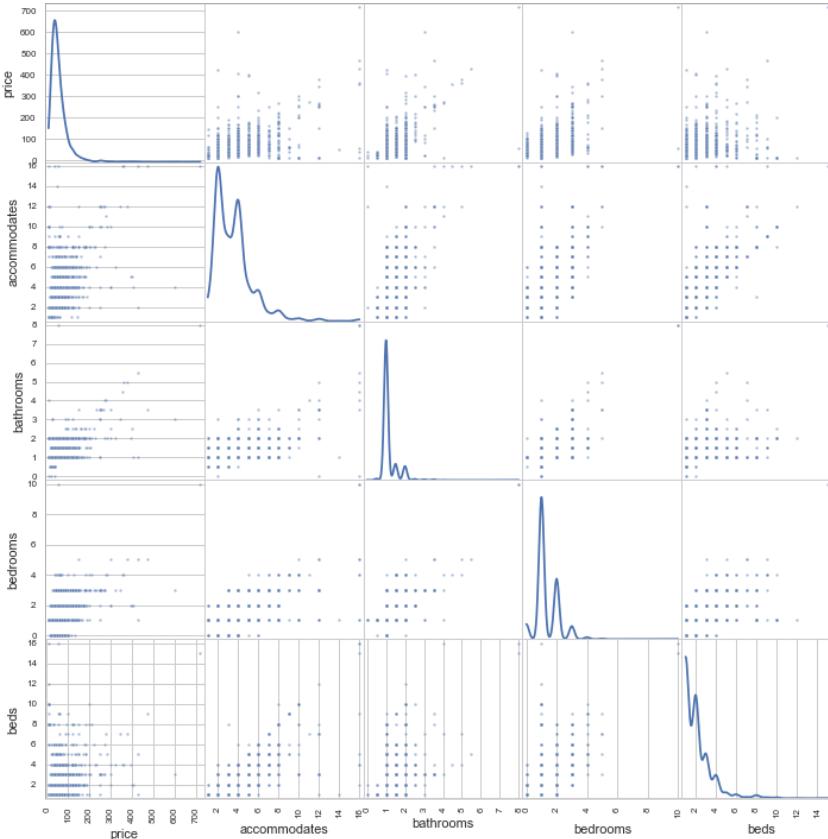
Correlation

Determining relationship between two variables

Price Distribution for Athens



Correlation: Visually



(Pearson's) Correlation Analytically

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Covariance* → (math)

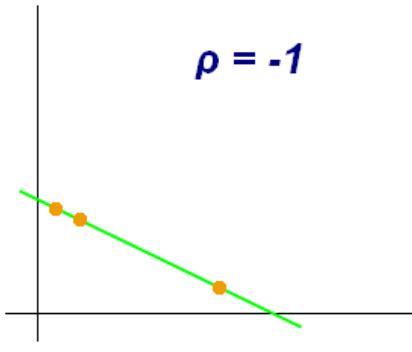
← **Standard Deviation* of x**

← **Standard Deviation* of y**

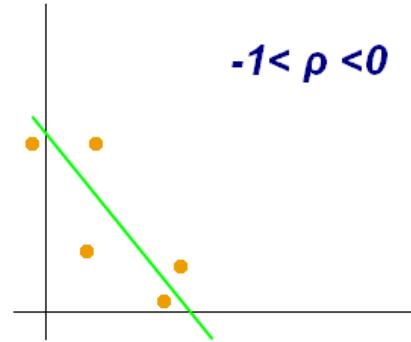
*without normalizing 1/N factors

Correlation: Visually

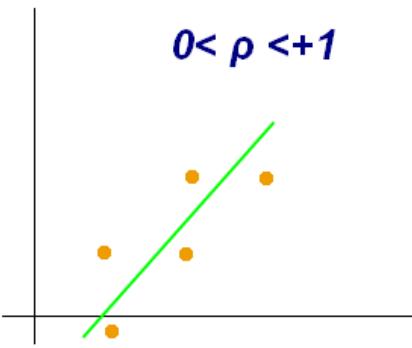
$$\rho = -1$$



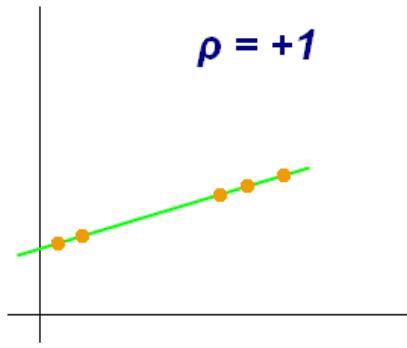
$$-1 < \rho < 0$$



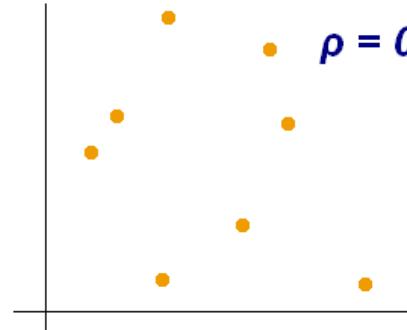
$$0 < \rho < +1$$



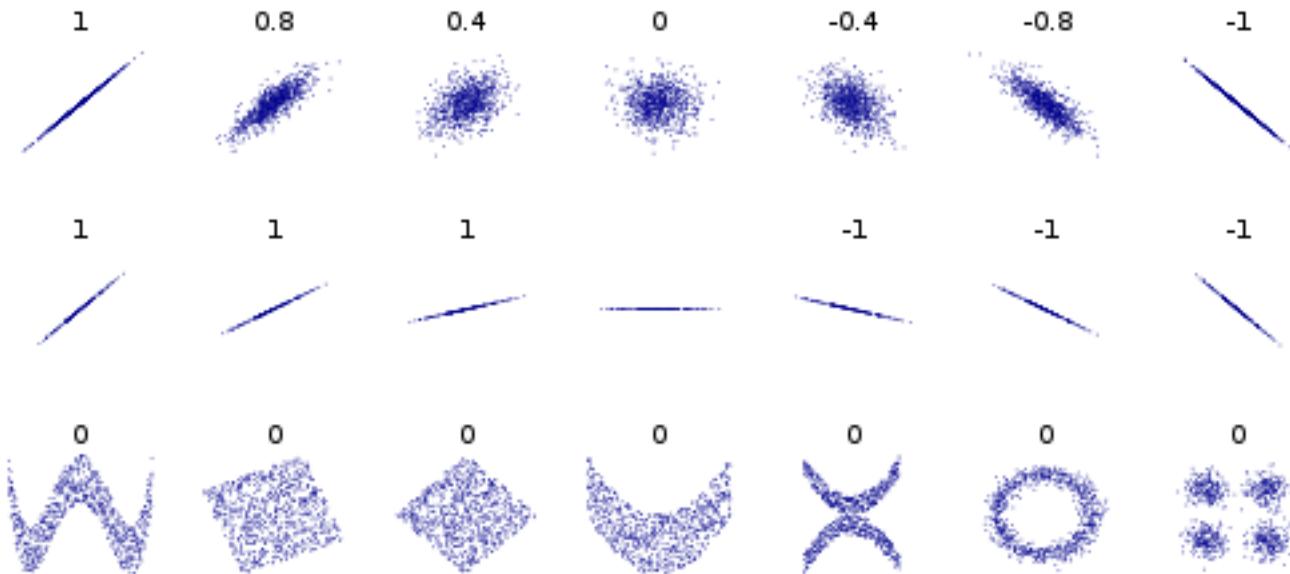
$$\rho = +1$$



$$\rho = 0$$



Correlation: Visually



Correlation Confidence Intervals

Problematic to calculate analytically...

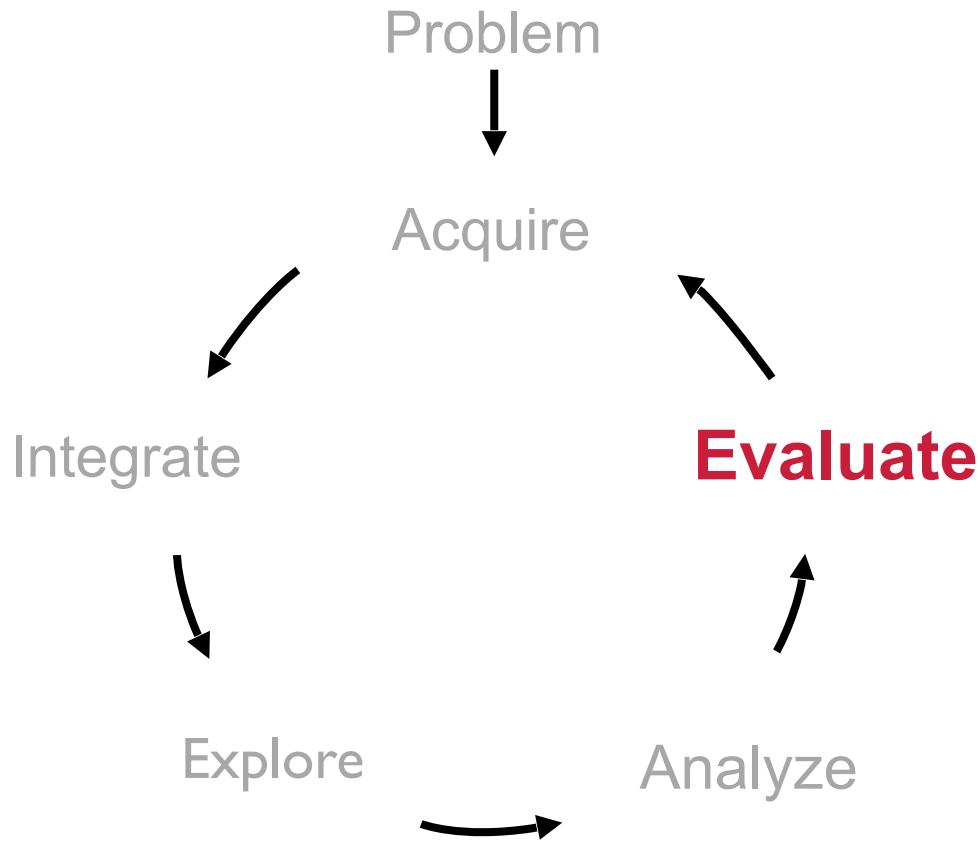
Correlation Confidence Intervals

Sampling distribution is not normally distributed...

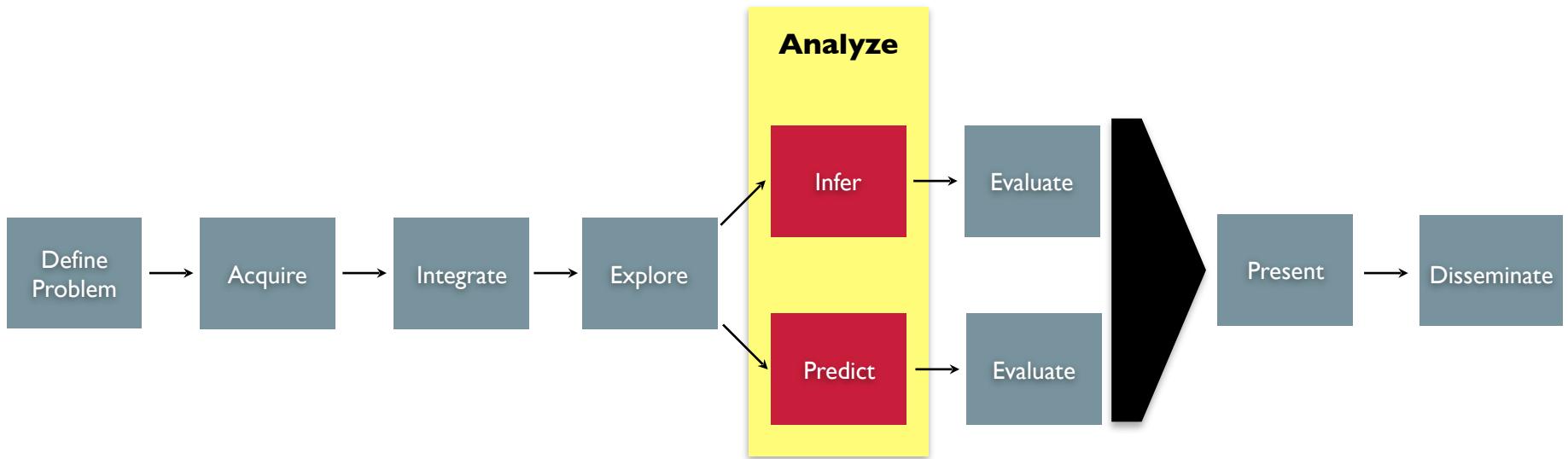
Correlation Confidence Intervals

To the bootstrap!

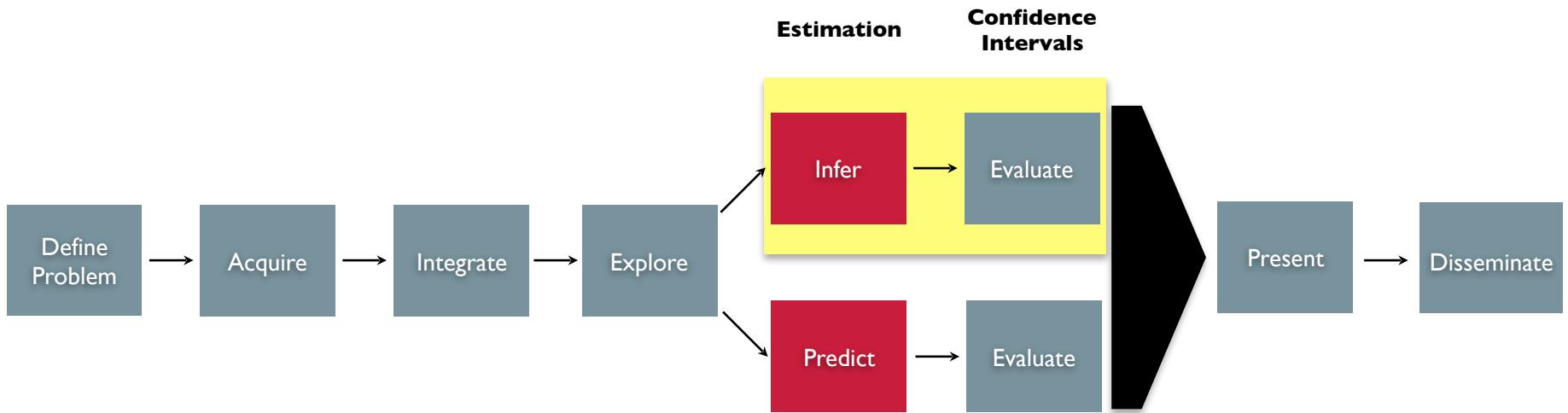
Process + Iteration



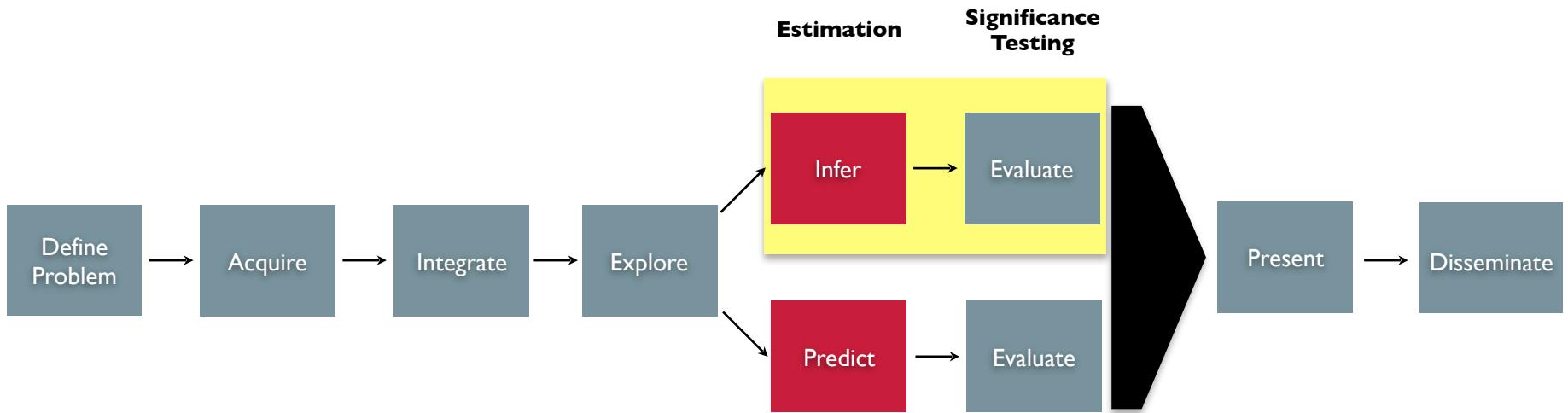
Process + Statistics



Process + Statistics



Process + Statistics



Hypothesis Testing

Experimental Study

- Design experiment a priori
- Manipulate explanatory variable (control versus treatment)
- More robust to confounding factors
- Can demonstrate causal effects (double blind randomized trial)

Observational Study

- Leverage existing data
- No manipulation of subjects
- Used when experiment would be infeasible or unethical
- Analysis post hoc
- Only possible to show correlation

Hypothesis Testing

Experimental Study

- Ex: Showing photos above the fold results in higher rates of requests for a listing

Observational Study

- Ex: Hosts who make less than \$30,000 a year accept a higher proportion of requests

Observational Study Setup

Null Hypothesis (H_0)

Instantly bookable listings **do not** have a **higher rate** (# reviews per month) of guests.

Alternative

Instantly bookable listings **have** a **higher rate** (# reviews per month) of guests.

Observational Study Setup

Explanatory Variable (used to segment groups)

Whether or not a given listing can be booked instantly as determined by the `instant_bookable` variable.

Response Variable (property of interest)

Throughput of a listing as determined by `reviews_per_month` variable.

Statistic

Mean of `reviews_per_month`

Hypothesis Testing: Outcomes

Reject the null, accept the alternative

— OR —

**Fail to reject the null, conclude there is
not enough evidence**

Hypothesis Testing: Outcomes

**Instant booking increases the amount
of guests a listing receives**

— OR —

“Not Guilty”

Hypothesis Testing: Mechanics

**Confidence interval of the difference of
means (or your test statistic)**

Hypothesis Testing: Mechanics

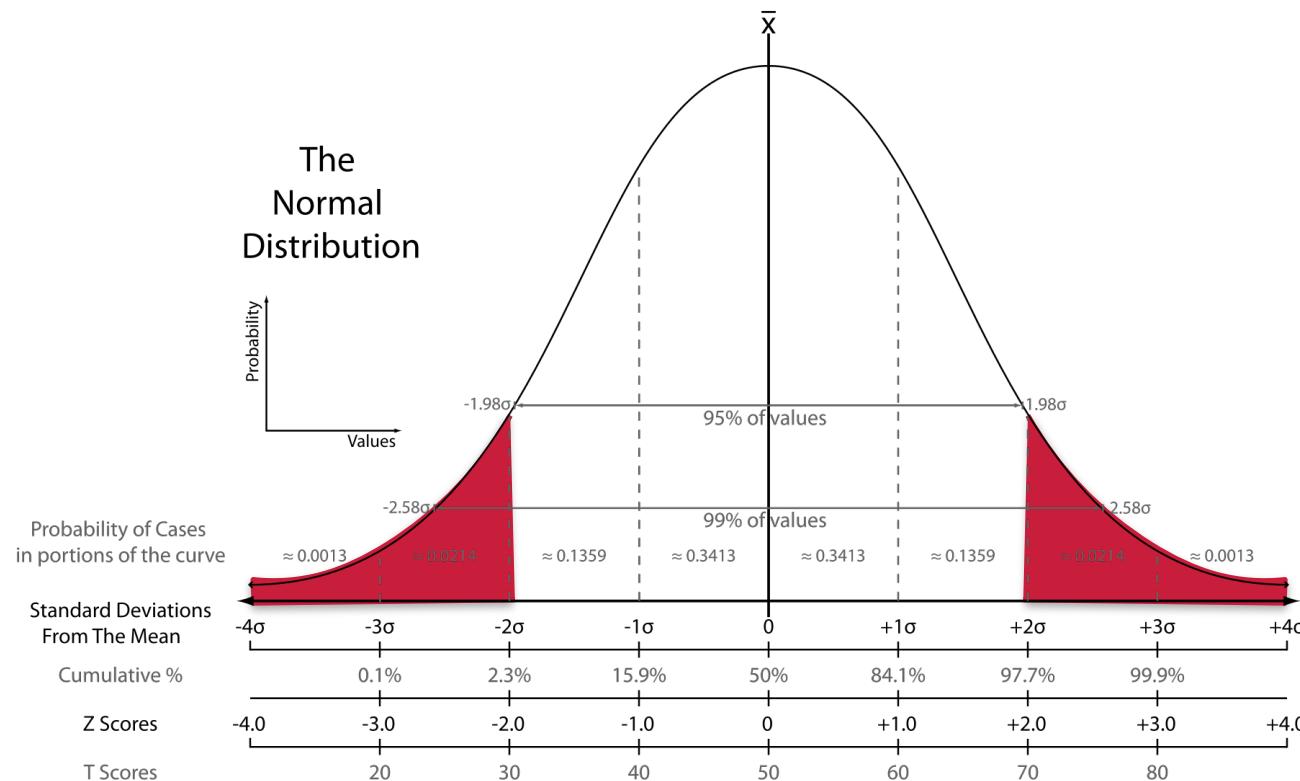
But this time two different samples...

	Fail to Reject	Reject
TRUE	Correct Inference (True Negative)	Type I Error (False Positive)
FALSE	Type II Error (False Negative)	Correct Inference (True Positive)

Predicted Condition (H_0)

True Condition (H_0)

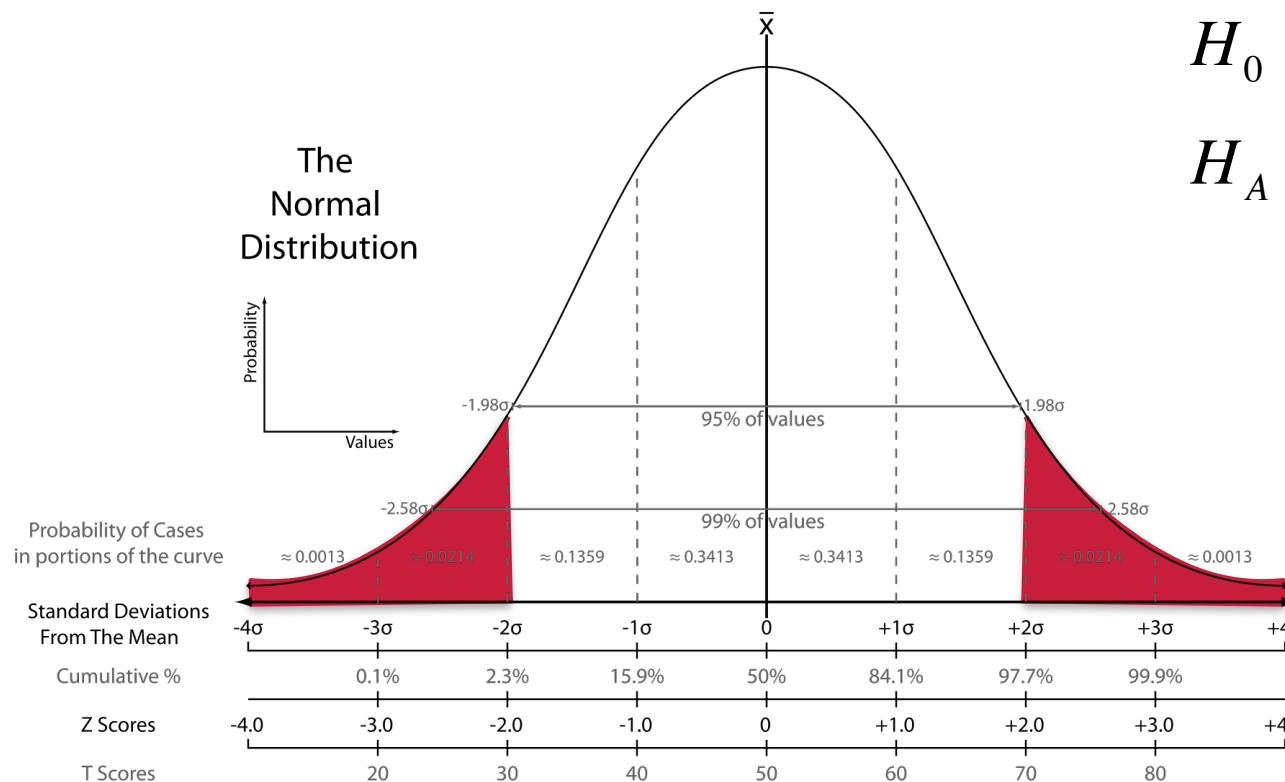
Our Familiar Friend



Two Sided

$$H_0 : \delta = 0$$

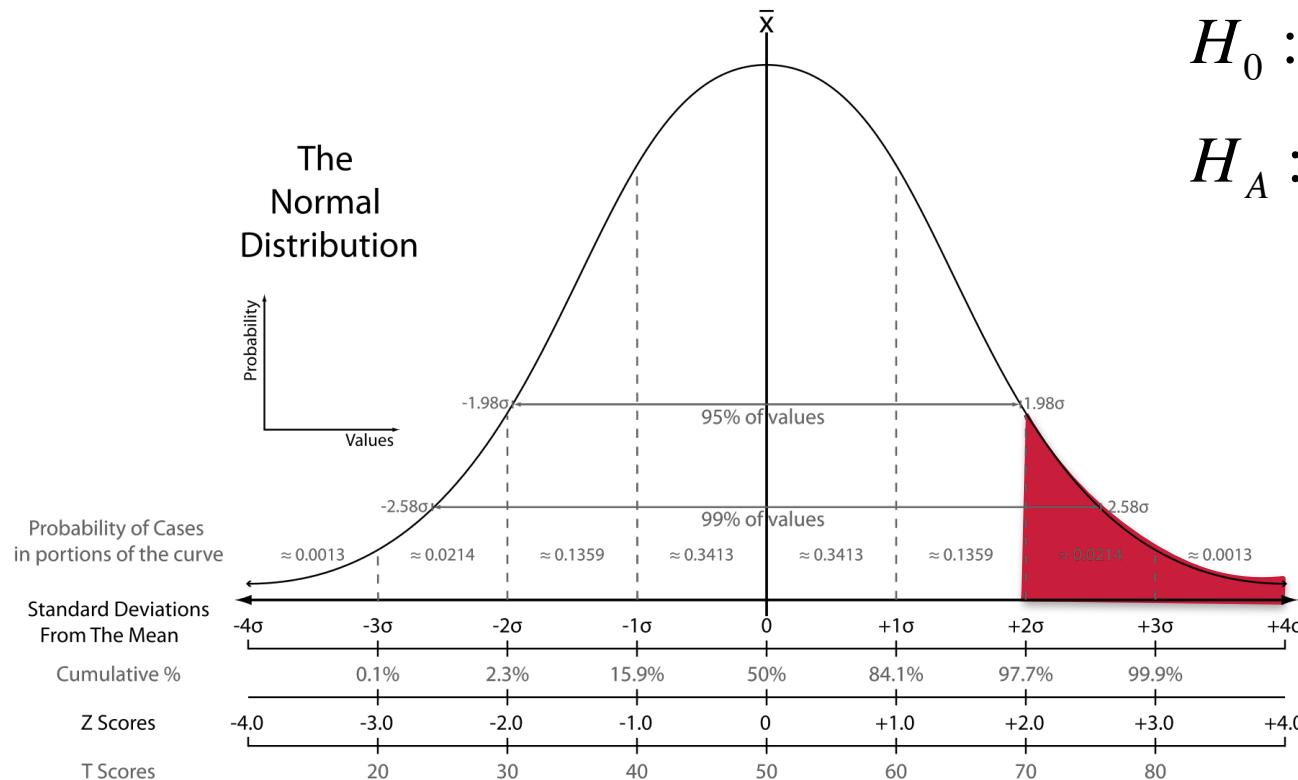
$$H_A : \delta \neq 0$$



One Sided

$$H_0 : \delta \leq 0$$

$$H_A : \delta > 0$$



Hypothesis Testing Terminology

Null Hypothesis

A default position that there is **no relationship** between two measured phenomena (or no association among groups)

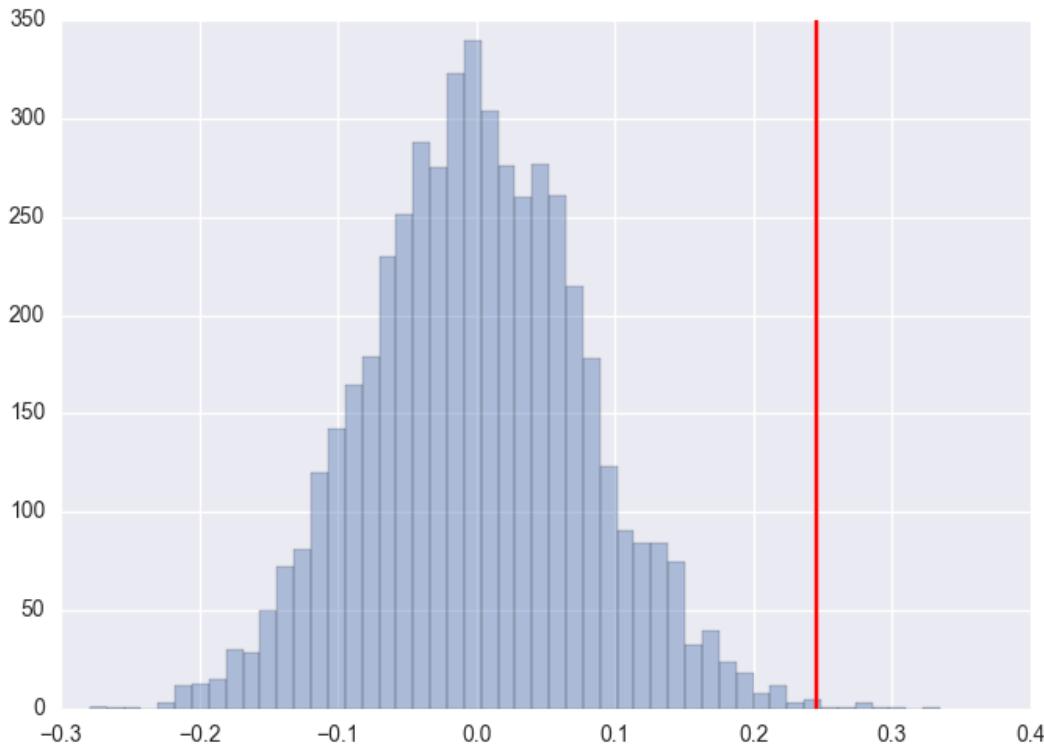
Z-score

Number of standard deviations away from the mean of the **sampling distribution**

p-value

The probability of obtaining a result equal to or "more extreme" than what was actually observed, when the null hypothesis is true

Sampling Distribution of Instant Bookable



Hypothesis Testing Terminology

Power of the Test

The probability that the test correctly rejects the null hypothesis (H_0) when the alternative hypothesis (H_1) is true.

Effect Size

They estimate the strength (magnitude) of, for example, an apparent relationship, rather than assigning a significance level reflecting whether the magnitude of the relationship observed could be due to chance.

Settings

Solve for? Power Alpha n d

Significance level ($\alpha = 0.05$)



Sample size (n = 20)

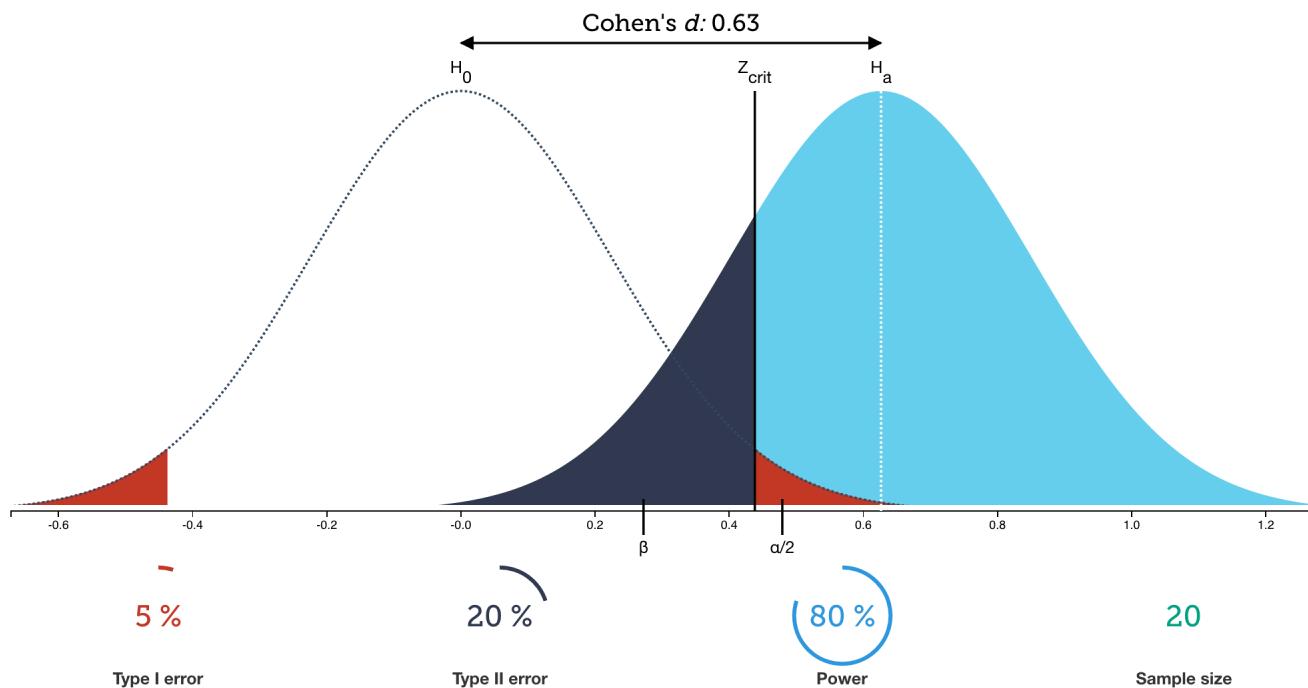


Effect size (d = 0.63)



One-tailed Two-tailed

Reset zoom



Hypothesis Testing Errors

- **Multiple Testing:** Bonferroni correction (or similar corrections)
- **Stopping Early:** Use a calculated sample size or Bayesian methods
- **Large N:** Use an appropriate effect size

Flavors of Statistics

Parametric

- Assume underlying generating distribution
- More accurate and precise (when they work)
- More informative (though can be harder to interpret)
- Simpler & easier to compute (can be done by hand)
- Generative (i.e. can *create* data)

Non-Parametric

- Makes little to no assumptions about data (distribution free)
- Requires no *a priori* decisions (i.e. probability distributions)
- More general and robust
- Needs larger sample sizes
- Often requires simulation/computation

Flavors of Statistics

Parametric

- Assume underlying generating distribution
- More accurate and precise (when they work)
- More informative (though can be harder to interpret)
- Simpler & easier to compute (can be done by hand)
- Generative (i.e. can *create* data)

Non-Parametric

- Makes little to no assumptions about data (distribution free)
- Requires no *a priori* decisions (i.e. probability distributions)
- More general and robust
- Needs larger sample sizes
- Often requires simulation/computation

Examples

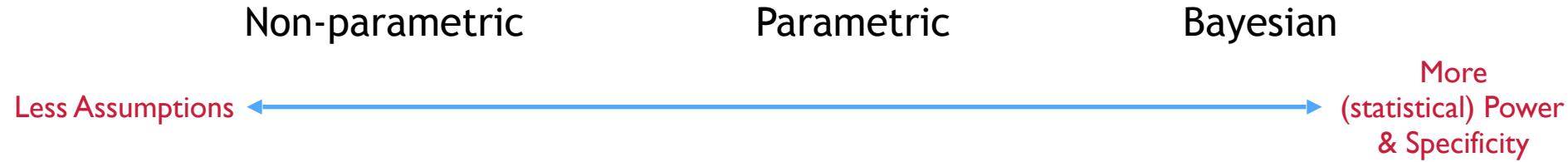
Parametric

- **Descriptive:** Mean & Standard Deviation
- **Inferential:** Hypothesis Tests & MLE
- **Predictive:** Linear Regression (OLS)

Non-Parametric

- **Descriptive:** Histograms & KDE
- **Inferential:** Bootstrap Methods
- **Predictive:** K Nearest Neighbors

Spectrum of Statistics



Spectrum of Statistics

Non-parametric

Parametric

Bayesian

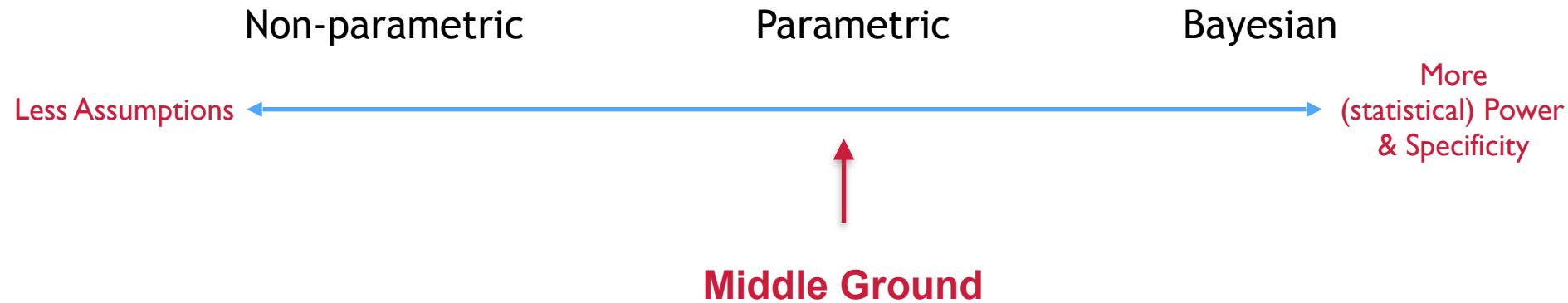
More Objective

More Subjective

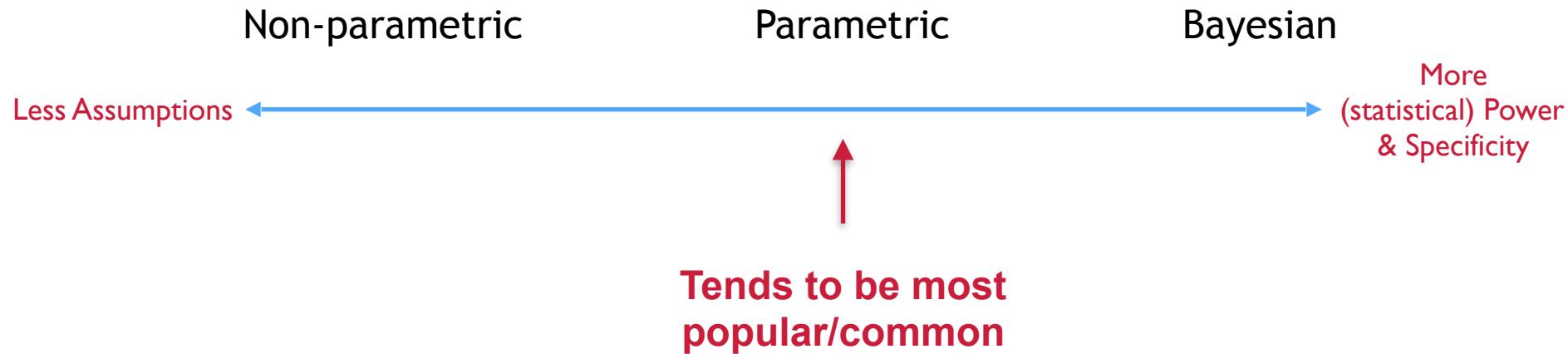
Spectrum of Statistics



Spectrum of Statistics



Spectrum of Statistics



Spectrum of Statistics

Non-parametric

Parametric

Bayesian

Less Assumptions

More
(statistical) Power
& Specificity

But at the End of the Day...

Tends to be most
popular/common

Spectrum of Statistics

Non-parametric

Parametric

Bayesian

More

(statistical) Power

Less Assumptions

Most Differences are cosmetic and philosophical

Tends to be most
popular/common

Next Up: Modeling and Machine Learning