livelessons
video instruction from technology experts

Lesson 8: Making Inferences: Statistical Estimation and Evaluation

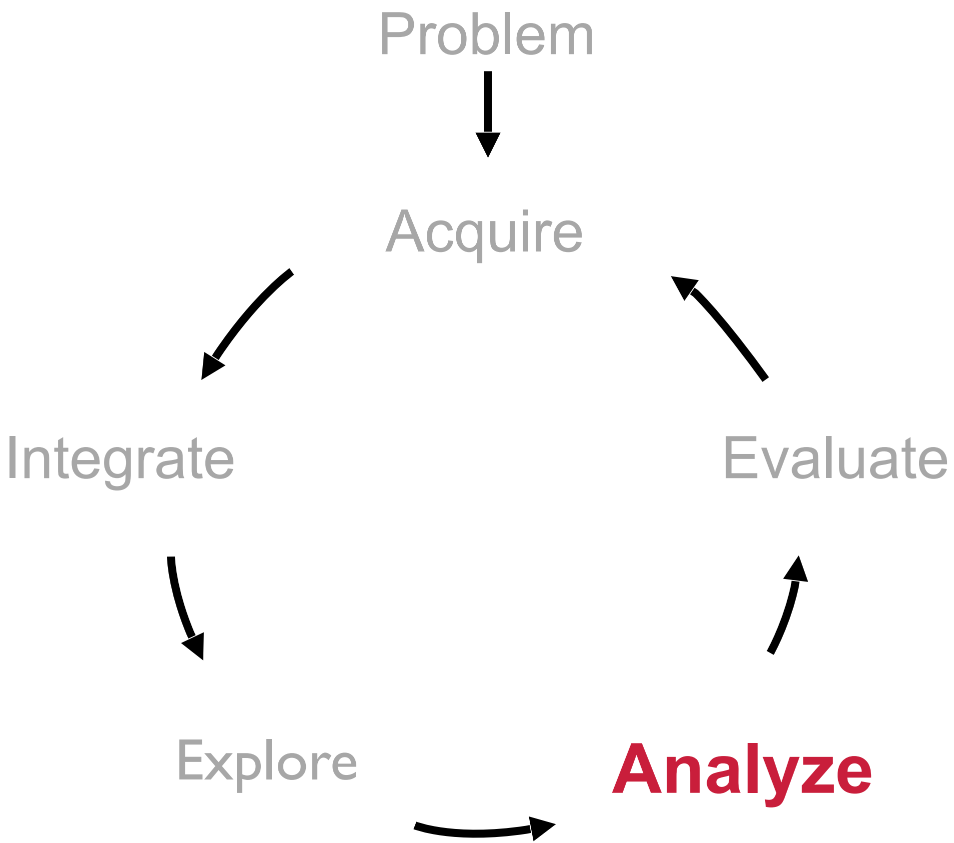
livelessons
©2016 Pearson, Inc.

Process

DOMAIN KNOWLEDGE		DATA MANIPULATION AND TRANSFORMATION		MATHEMATICS, STATISTICS, AND DATA MINING		COMMUNICATION	
Define Problem	Acquisition	Integration	Exploration	**Analysis**	Evaluation	Presentation	Dissemination

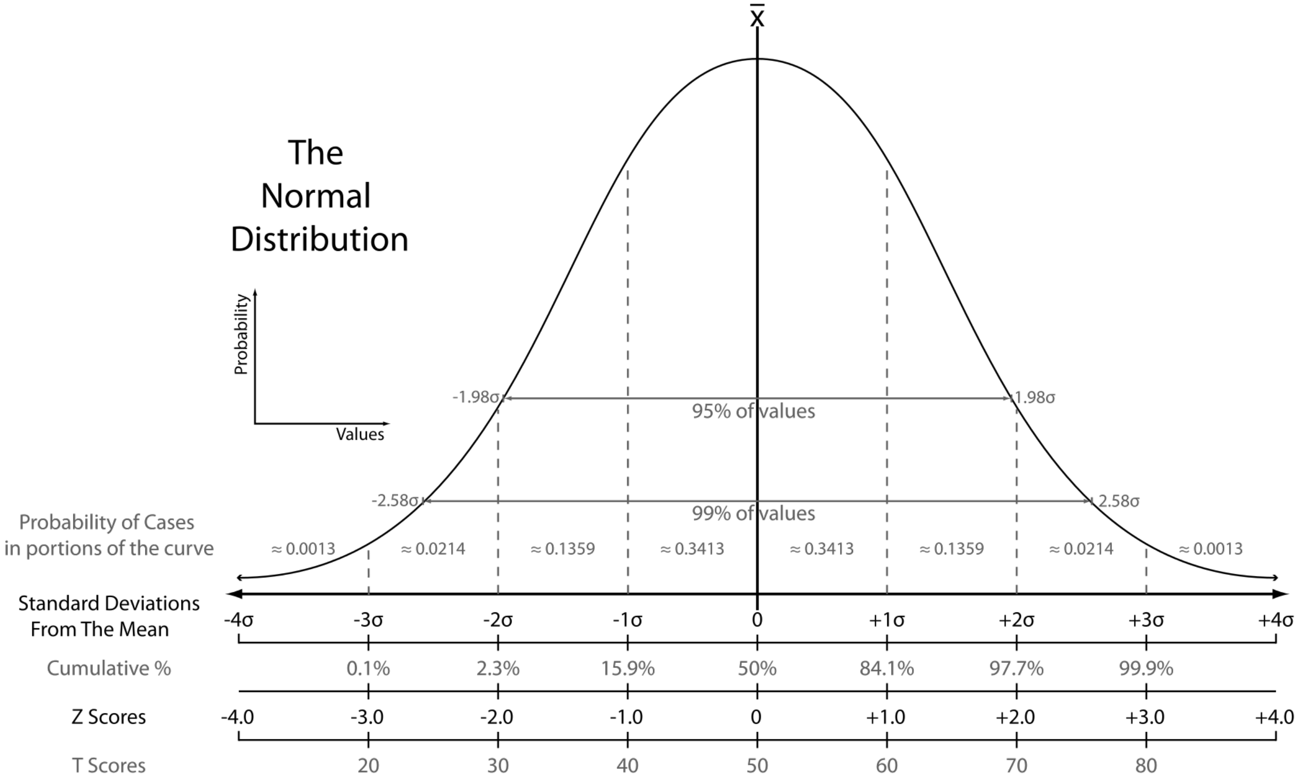
livelessons
©2016 Pearson, Inc.

Process + Iteration



©2016 Pearson, Inc.

What is Statistics?



source: Wikipedia

livelessons
©2016 Pearson, Inc.

What is Statistics?

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

- Wikipedia

livelessons
©2016 Pearson, Inc.

What is Statistics?

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

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

livelessons
©2016 Pearson, Inc.

What is Statistics?

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

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

livelessons
©2016 Pearson, Inc.

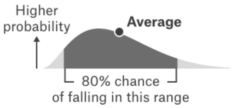
The Problem

livelessons
©2016 Pearson, Inc.

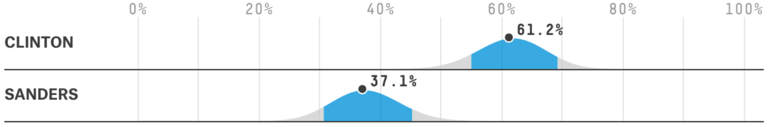
projects.fivethirtyeight.com/election-2016/primary-forecast/california-democratic/

Polls-plus forecast

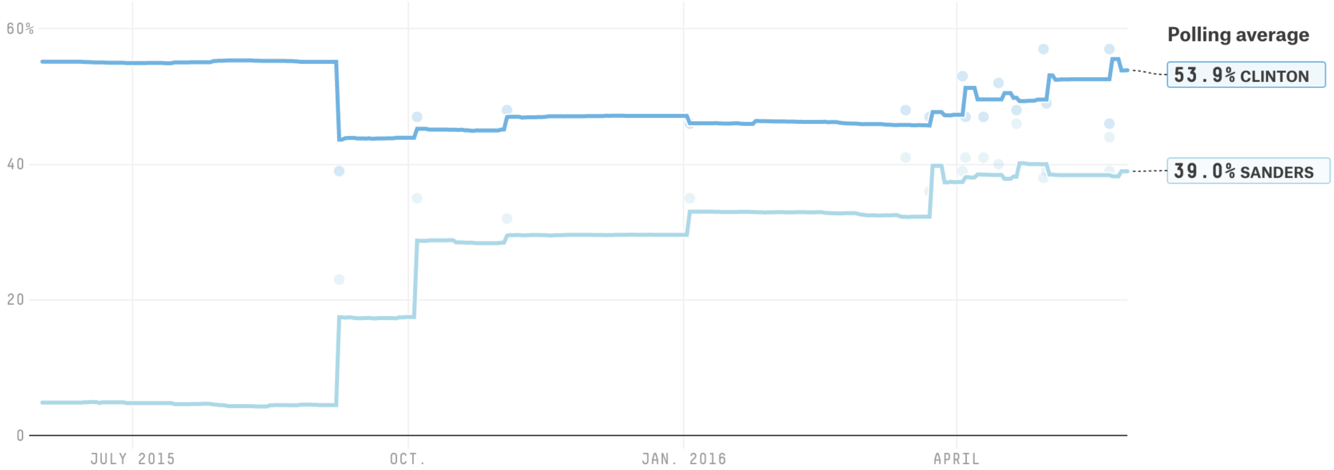
Polls-only forecast



Projected results

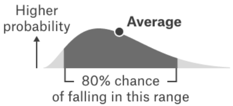


Not all **polls** are created equal, so our forecasts are calculated based on weighted **polling averages**.

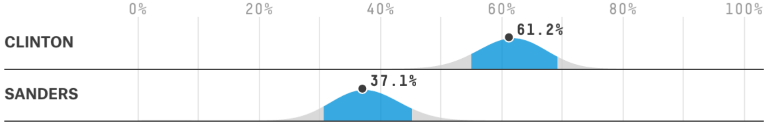


Polls-plus forecast

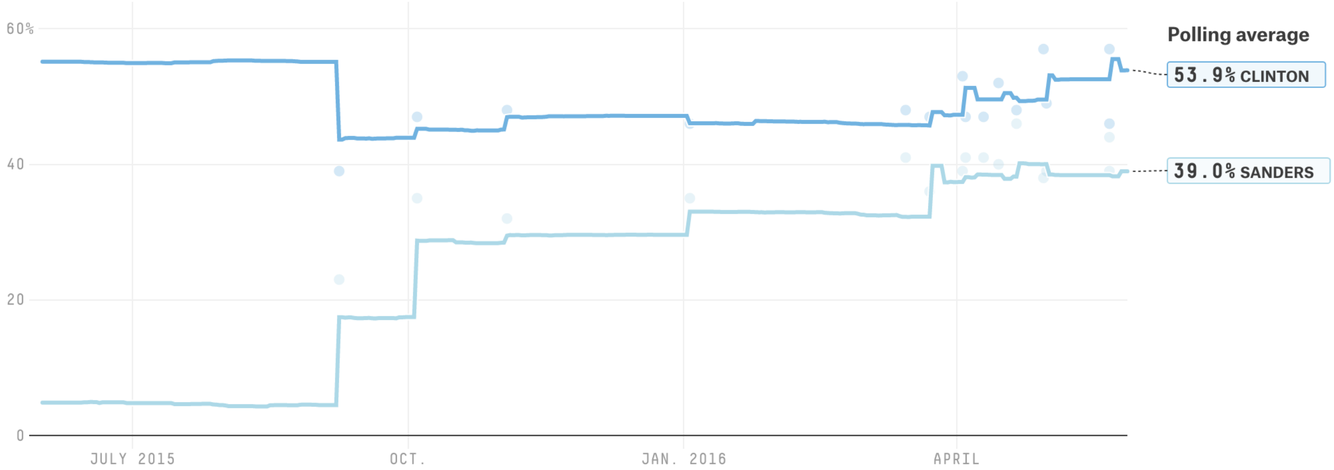
Polls-only forecast



Projected results



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



projects.fivethirtyeight.com/election-2016/primary-forecast/california-democratic/

● = 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	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.

Reasoning in an Uncertain World

livelessons
©2016 Pearson, Inc.

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?

livelessons
©2016 Pearson, Inc.

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

livelessons
©2016 Pearson, Inc.

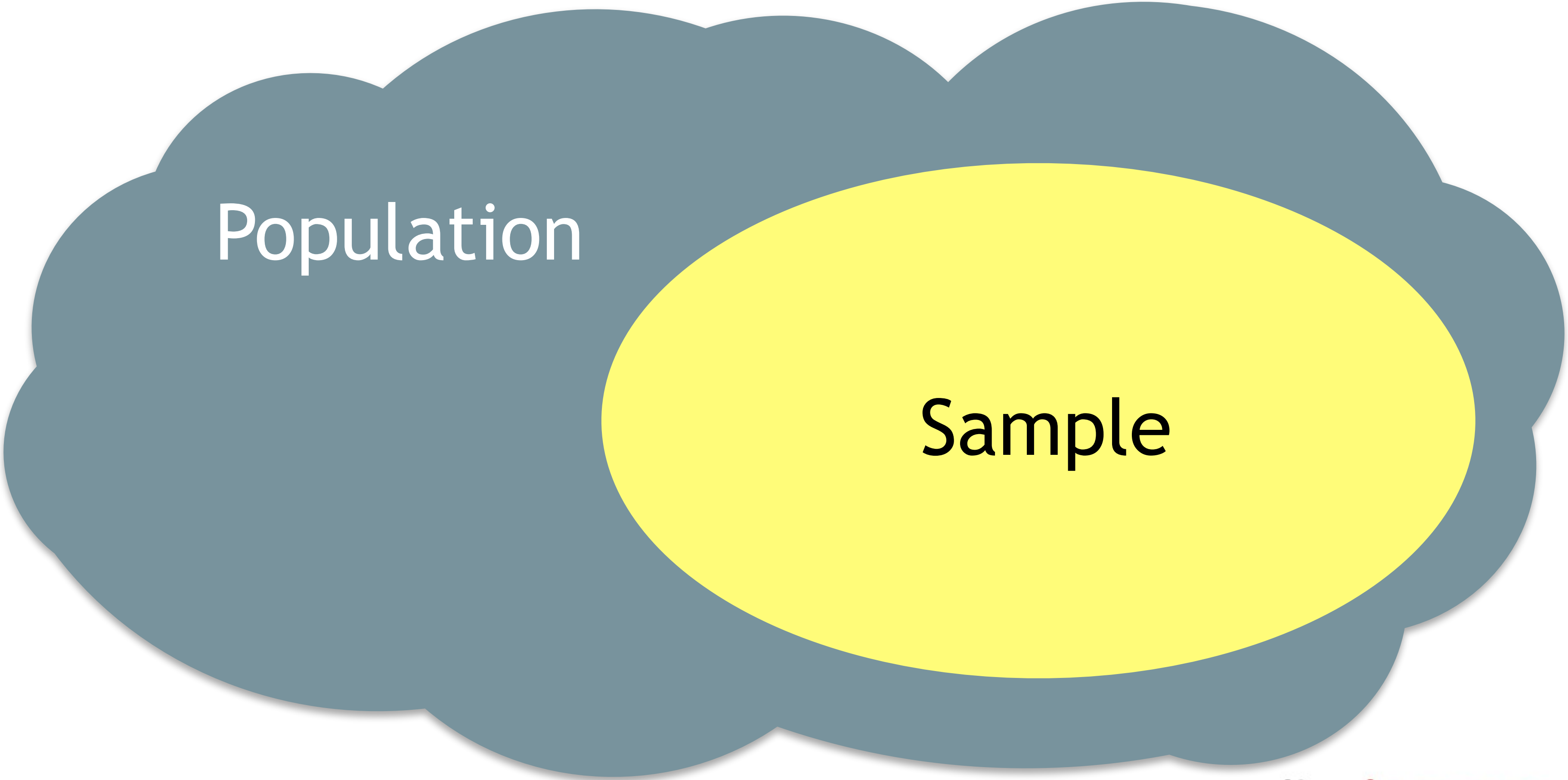
(hypothetical) Decisions

- 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

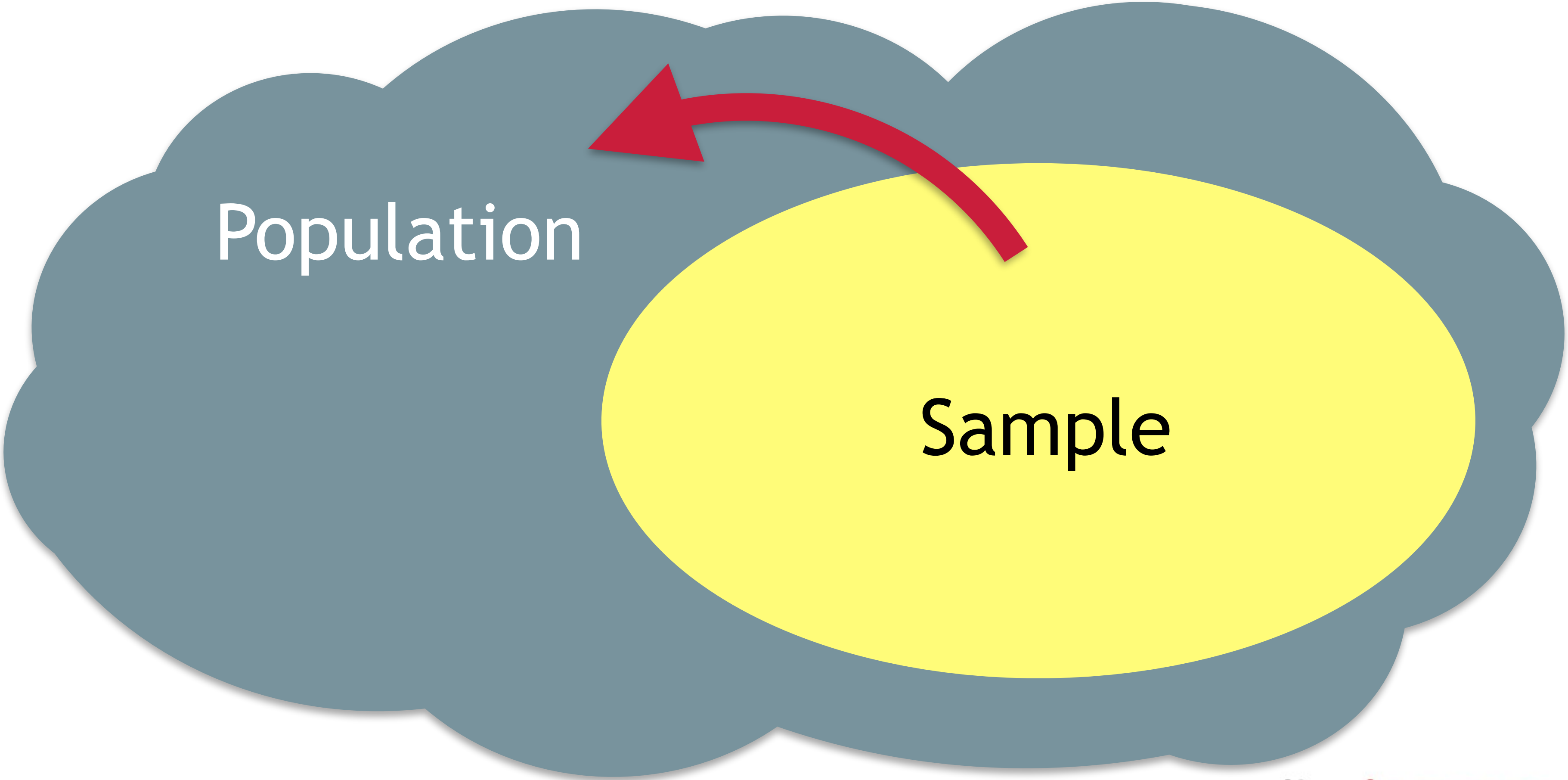
livelessons
©2016 Pearson, Inc.

In Pictures

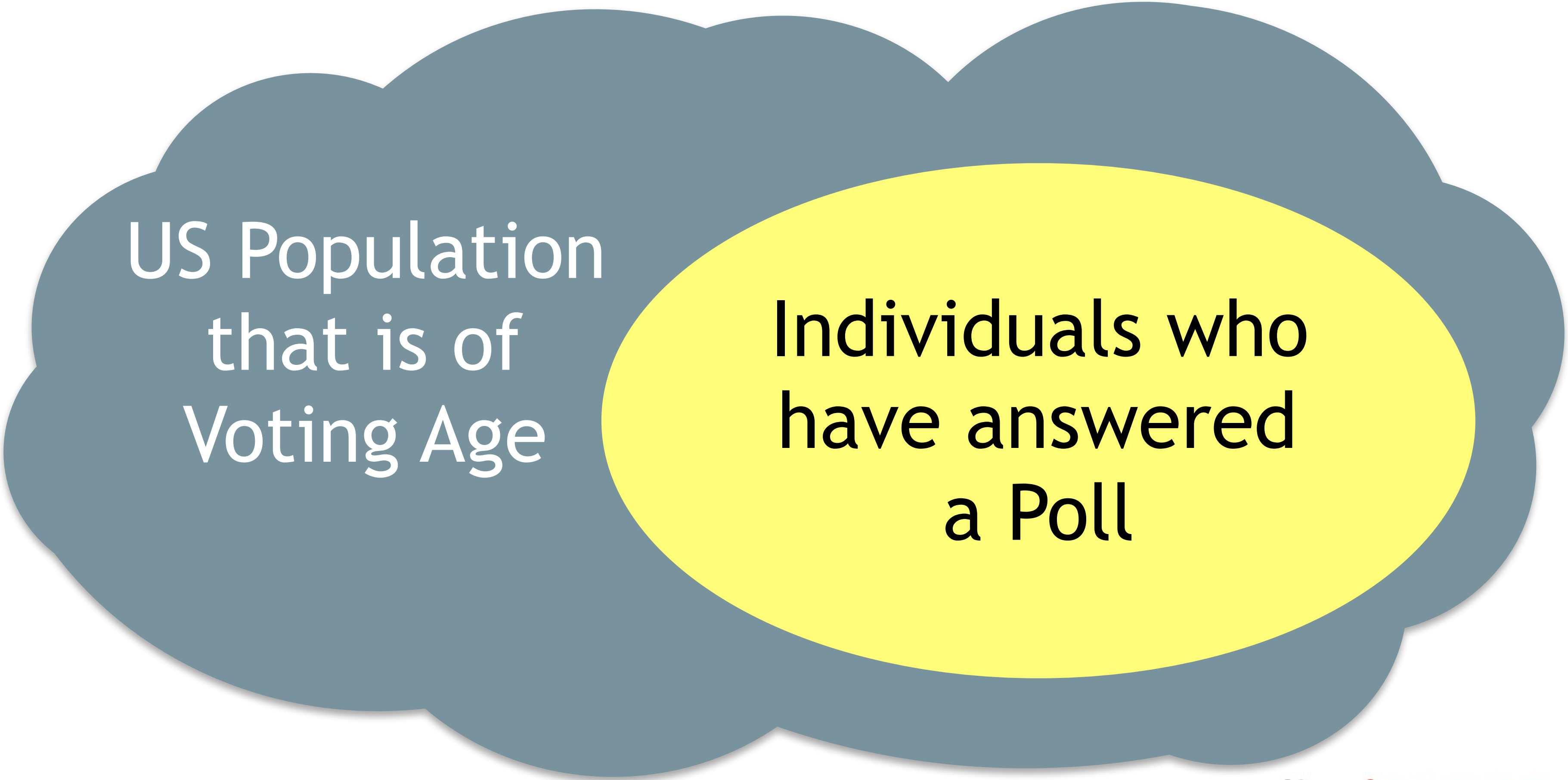
livelessons
©2016 Pearson, Inc.



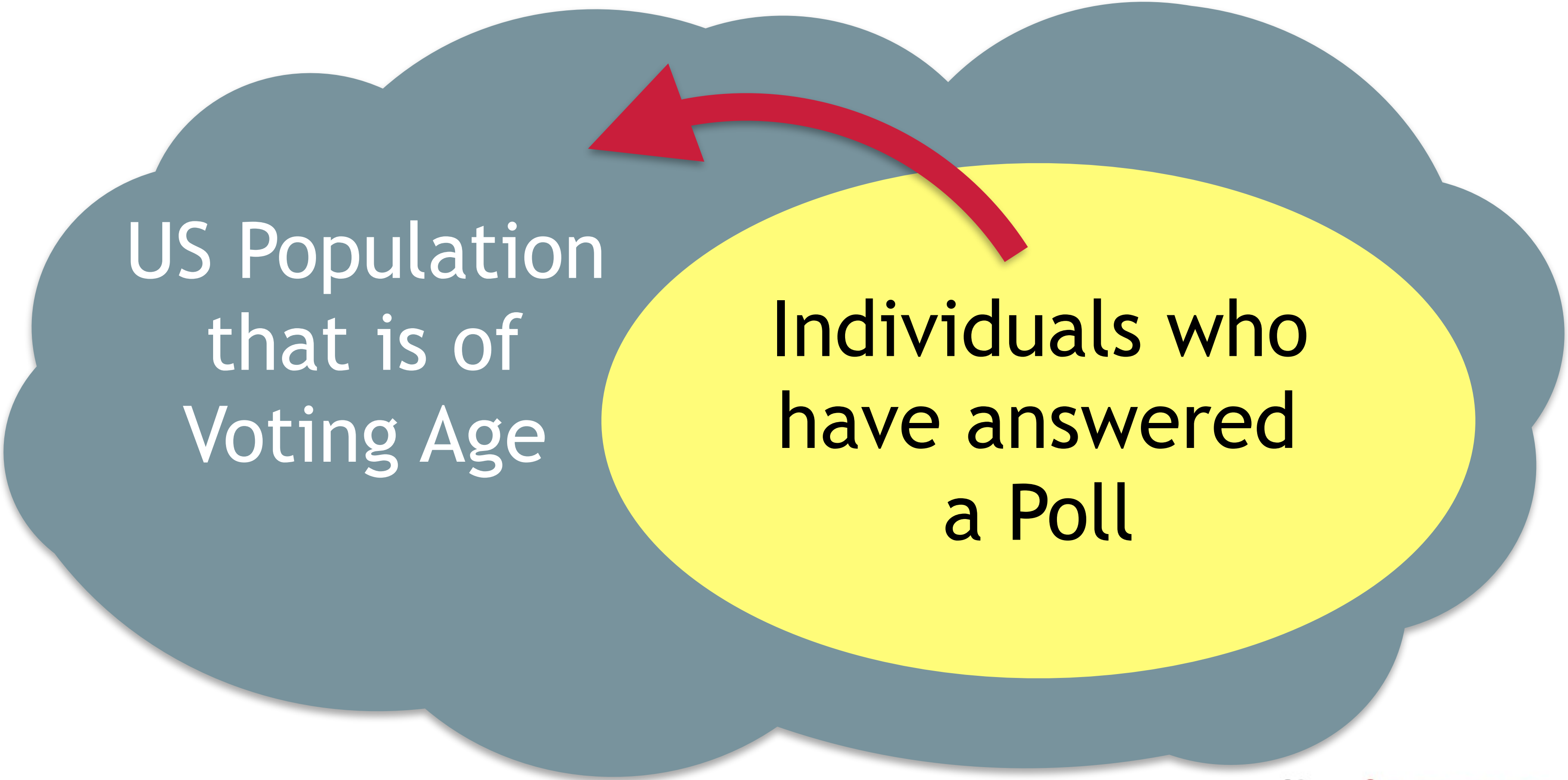
livelessons
©2016 Pearson, Inc.



livelessons
©2016 Pearson, Inc.

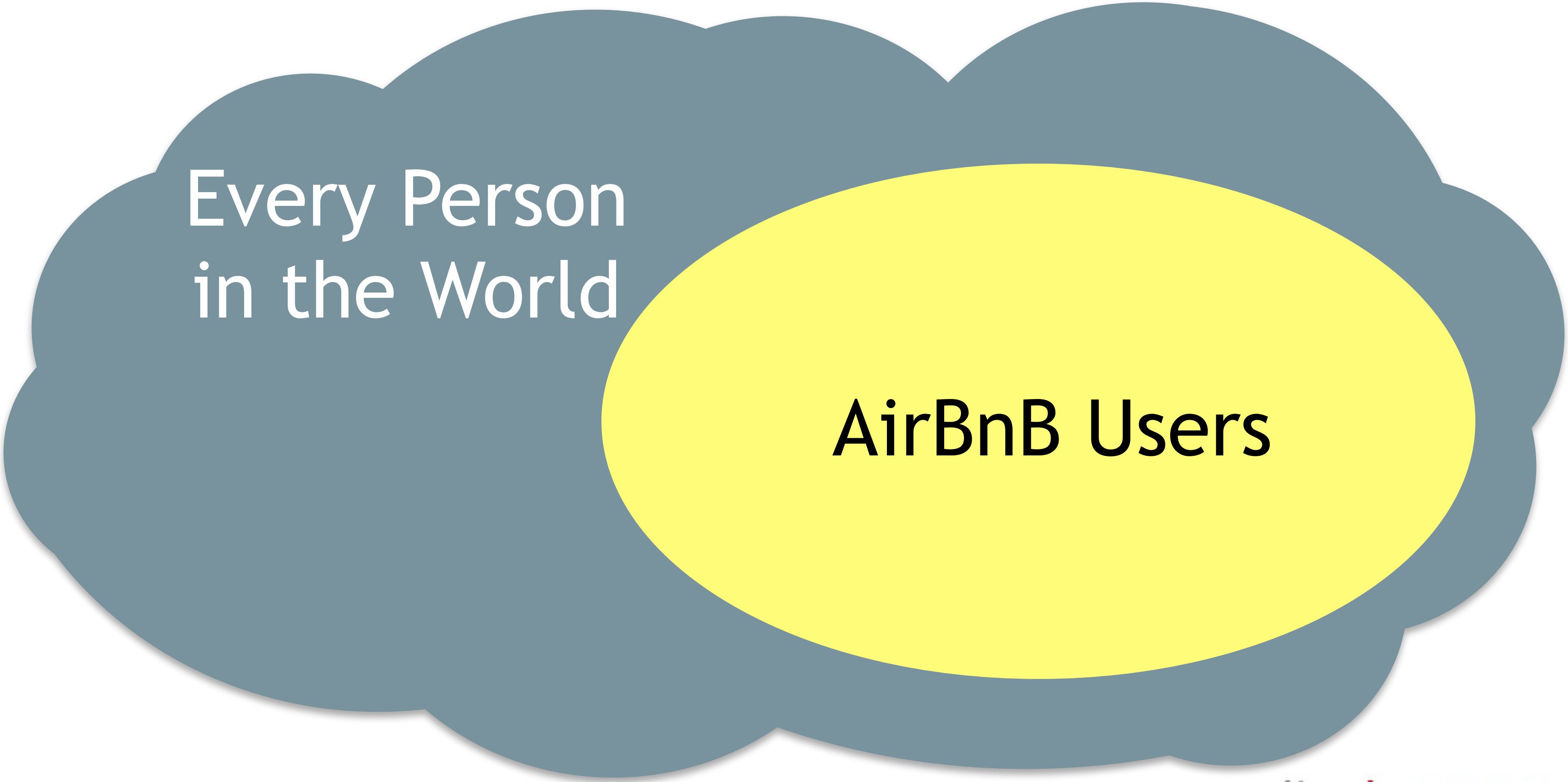


©2016 Pearson, Inc.

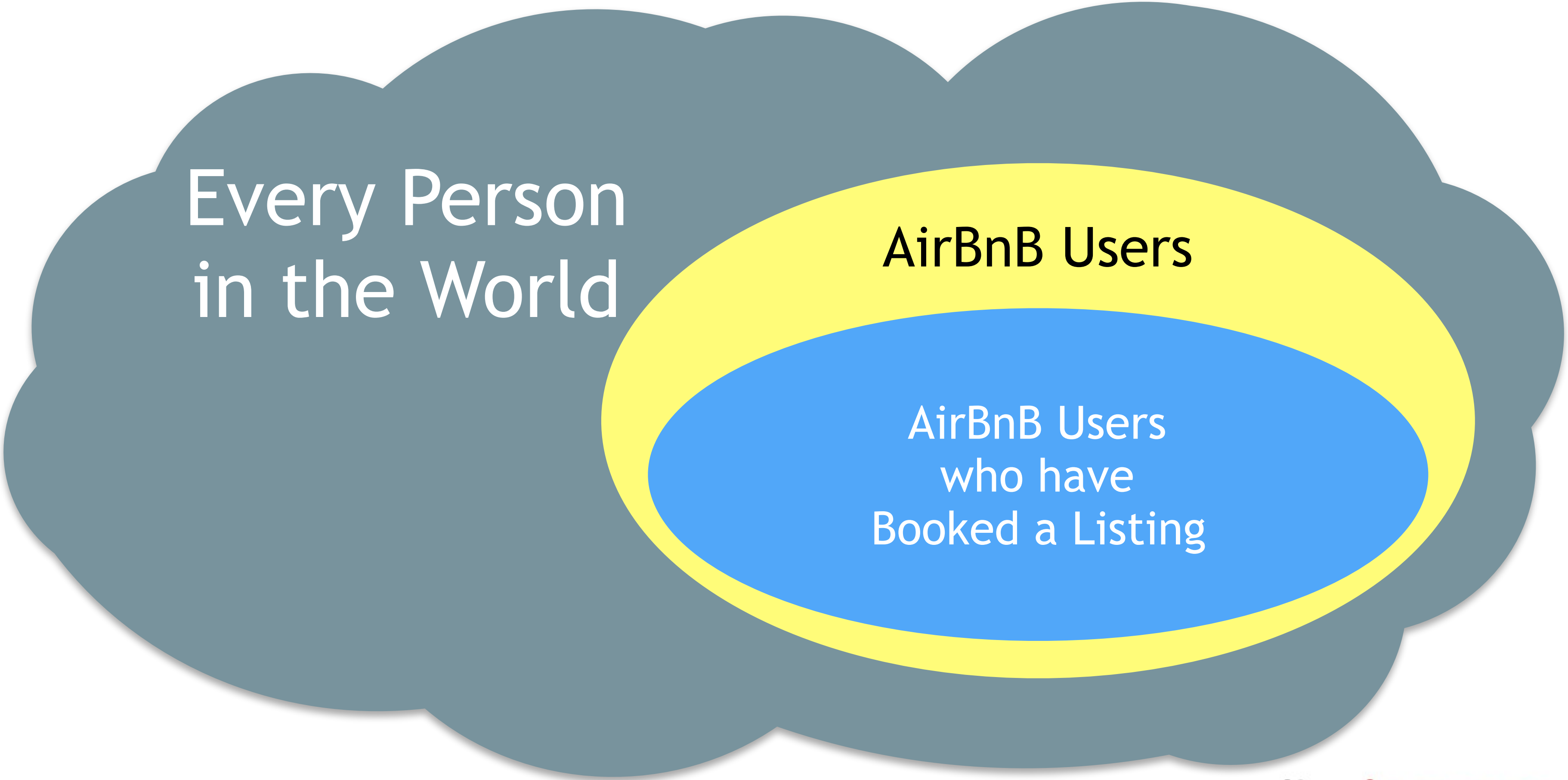


livelessons

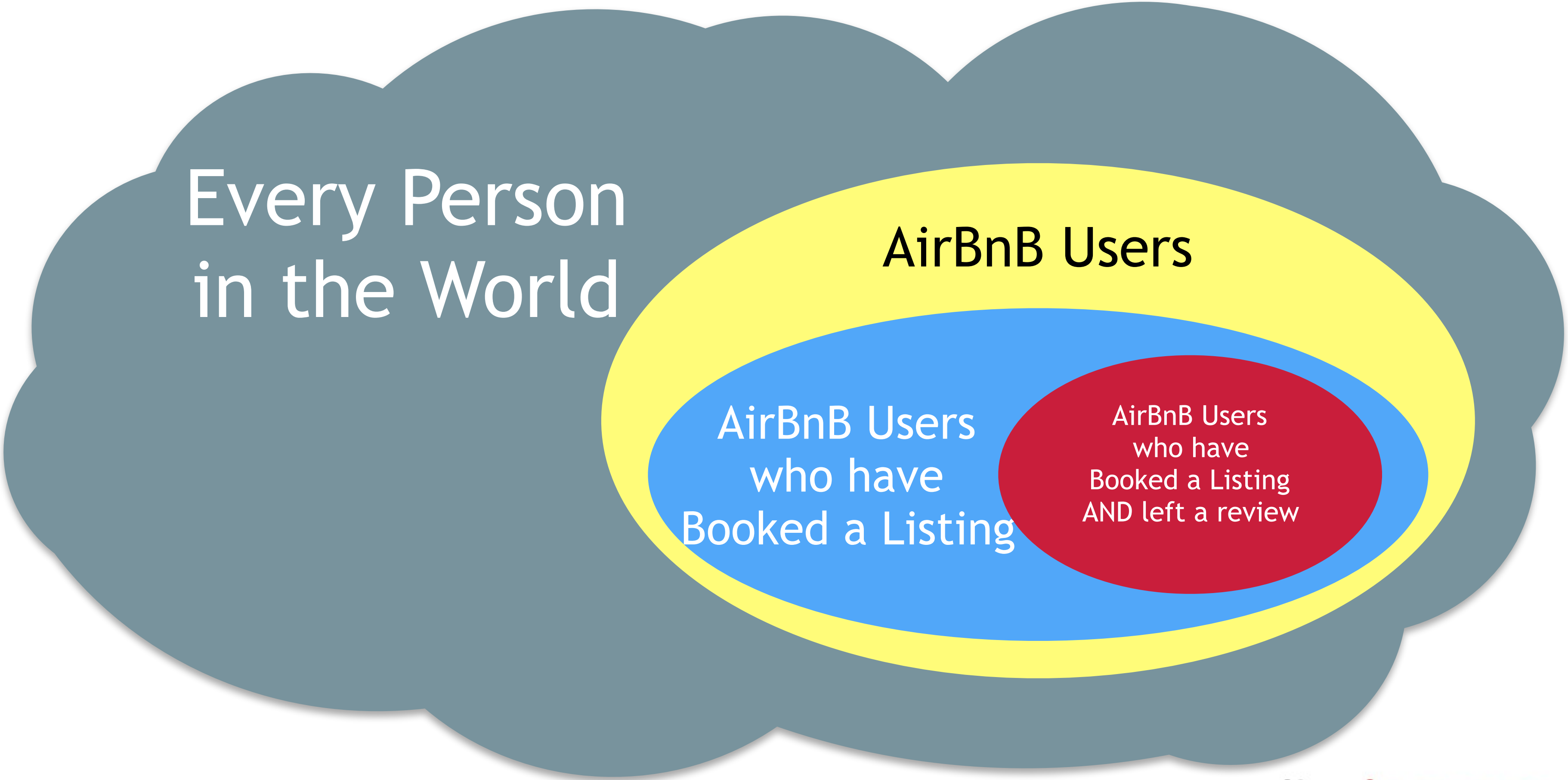
©2016 Pearson, Inc.



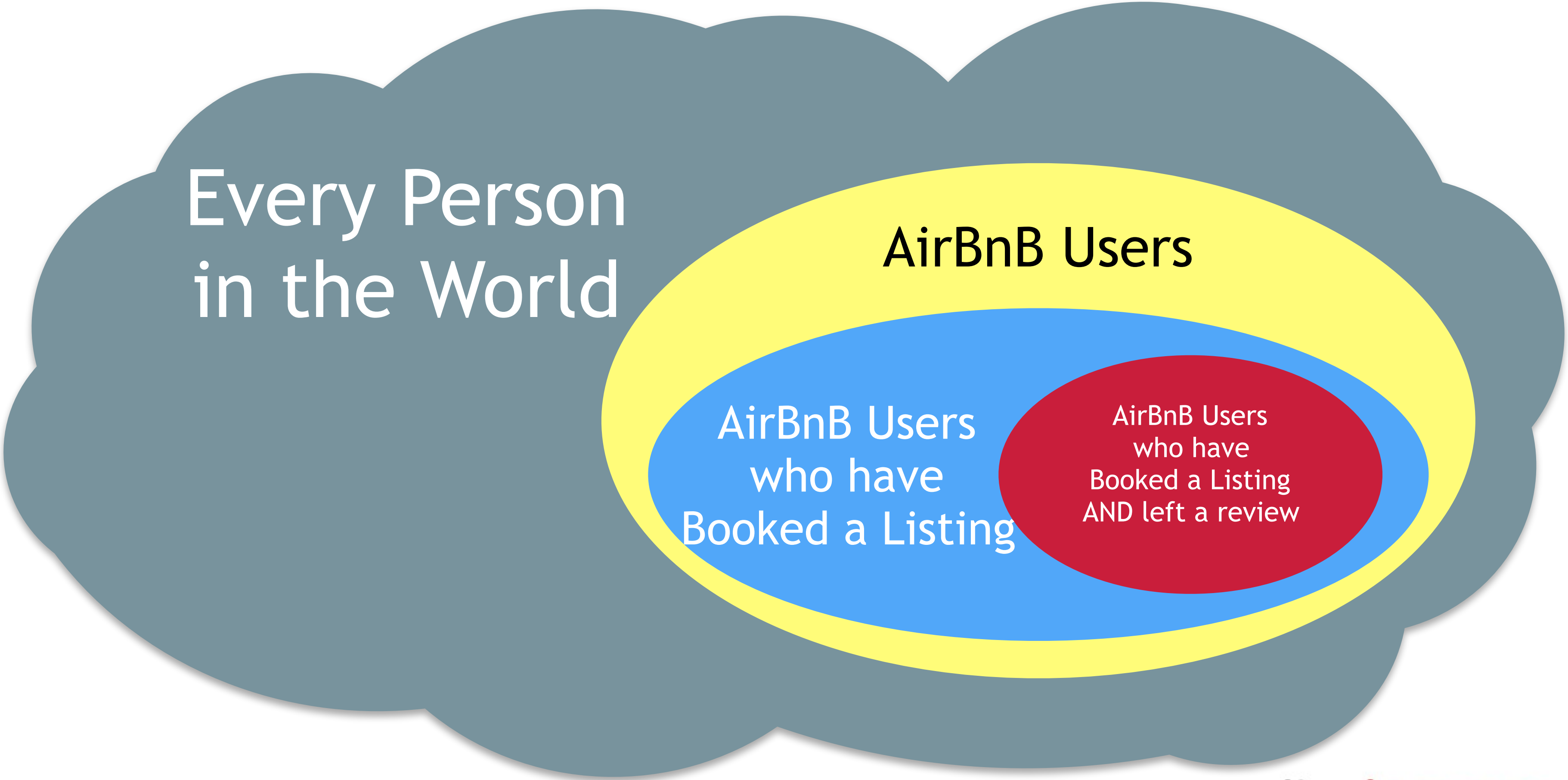
livelessons
©2016 Pearson, Inc.



©2016 Pearson, Inc.



livelessons
©2016 Pearson, Inc.



* Not to Scale

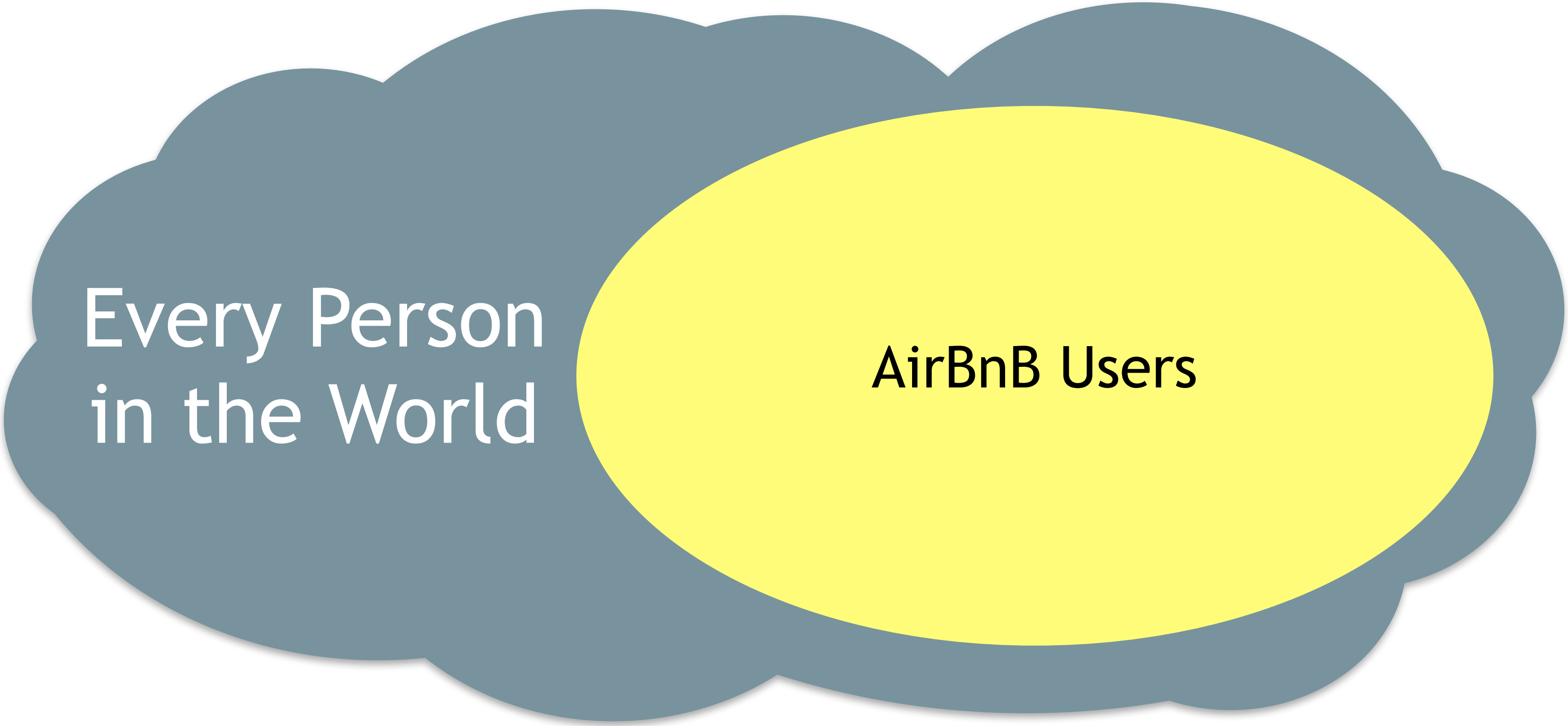
livelessons
©2016 Pearson, Inc.

We Want This

Every Person
in the World

livelessons
©2016 Pearson, Inc.

Or Even This...



livelessons
©2016 Pearson, Inc.

But Only Have This...

livelessons
©2016 Pearson, Inc.

Uncertainty

©2016 Pearson, Inc.

Goals

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

livelessons
©2016 Pearson, Inc.

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

livelessons
©2016 Pearson, Inc.

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

livelessons
©2016 Pearson, Inc.

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

livelessons
©2016 Pearson, Inc.

Moments (mean, median, etc.)

livelessons
©2016 Pearson, Inc.

Average

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

livelessons
©2016 Pearson, Inc.

Expected Value

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

livelessons
©2016 Pearson, Inc.

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

livelessons
©2016 Pearson, Inc.

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

livelessons
©2016 Pearson, Inc.

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

livelessons
©2016 Pearson, Inc.

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

livelessons
©2016 Pearson, Inc.

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

©2016 Pearson, Inc.

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?

livelessons
©2016 Pearson, Inc.

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?

livelessons
©2016 Pearson, Inc.

Let's try to first solve a simpler problem

livelessons
©2016 Pearson, Inc.

That doesn’t involve a comparison

livelessons
©2016 Pearson, Inc.

The simple, humble, mean

livelessons
©2016 Pearson, Inc.

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?

Simplify!

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?

livelessons
©2016 Pearson, Inc.

https://www.yelp.com/biz/yelp-san-francisco

Find tacos, cheap dinner, Max's
Near Noe Valley, San Francisco, CA

Home About Me Write a Review Find Friends Messages Talk Events

San Francisco > Financial District >

Yelp

8381 revie

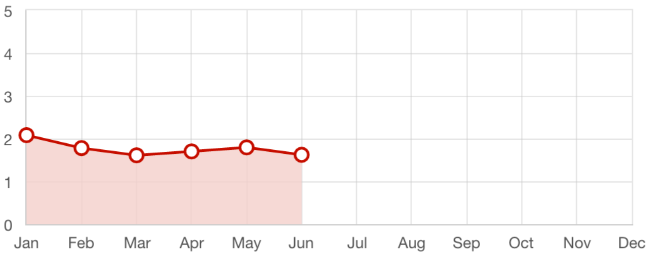
Local Flavor, Mass Media

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

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

5 stars	2007
4 stars	639
3 stars	527
2 stars	705
1 star	4507

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

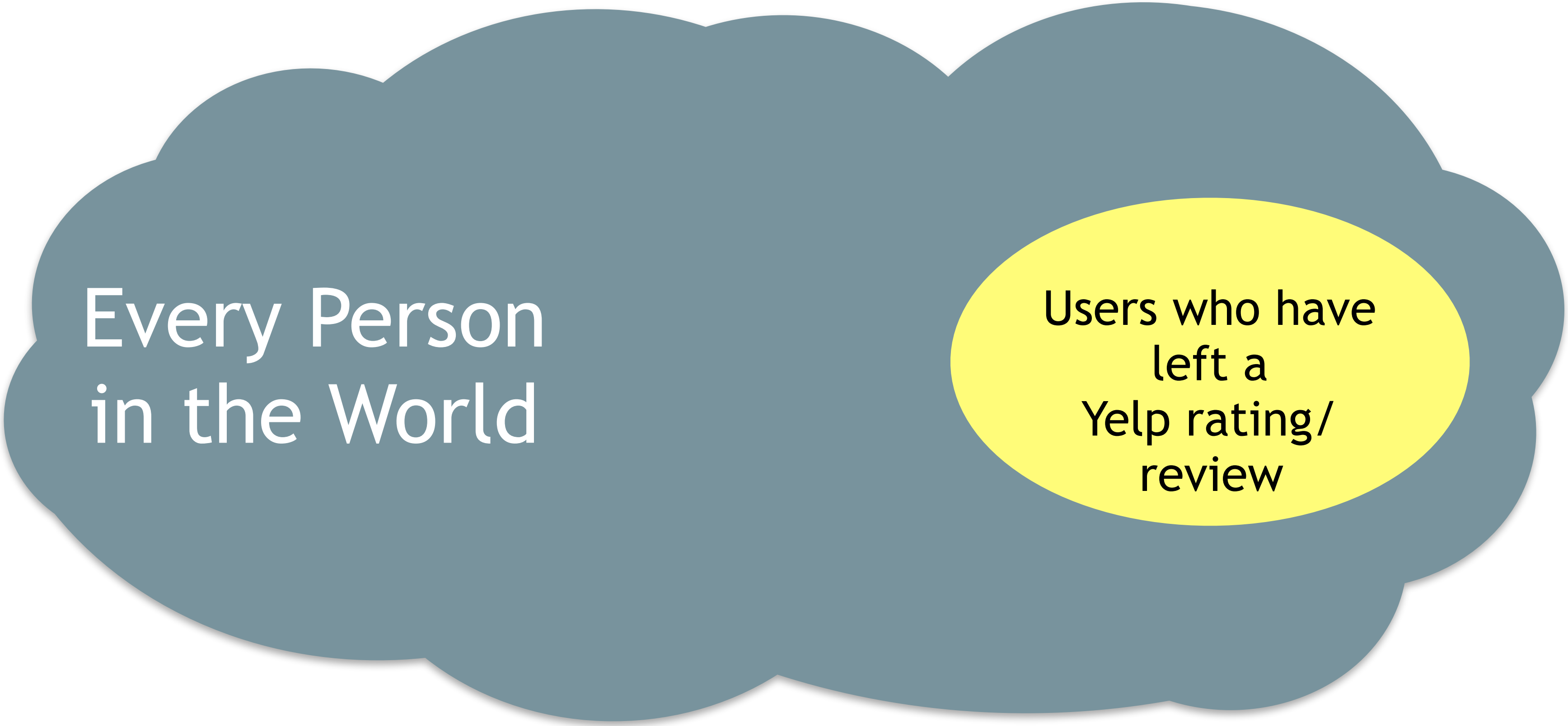
:00 am - 6:00 pm
:00 am - 6:00 pm
:00 am - 6:00 pm
:00 am - 6:00 pm Open now
:00 am - 6:00 pm
Closed
Sun Closed

"Locally, I enjoy bei
way I do about writ

"Reviews galore, ta
know, love, hate or

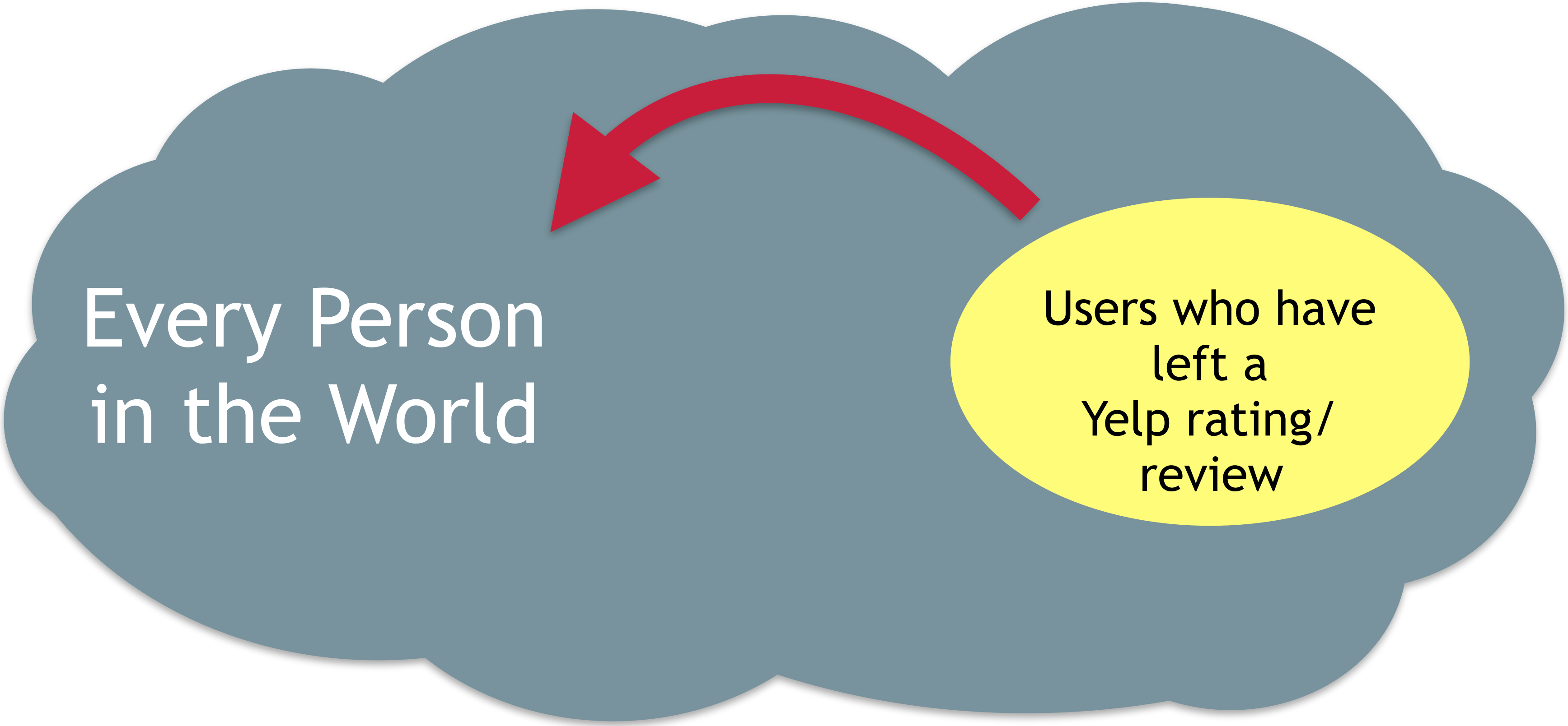
"Of course, all thos
forward to attend m

General Inference



livelessons
©2016 Pearson, Inc.

General Inference

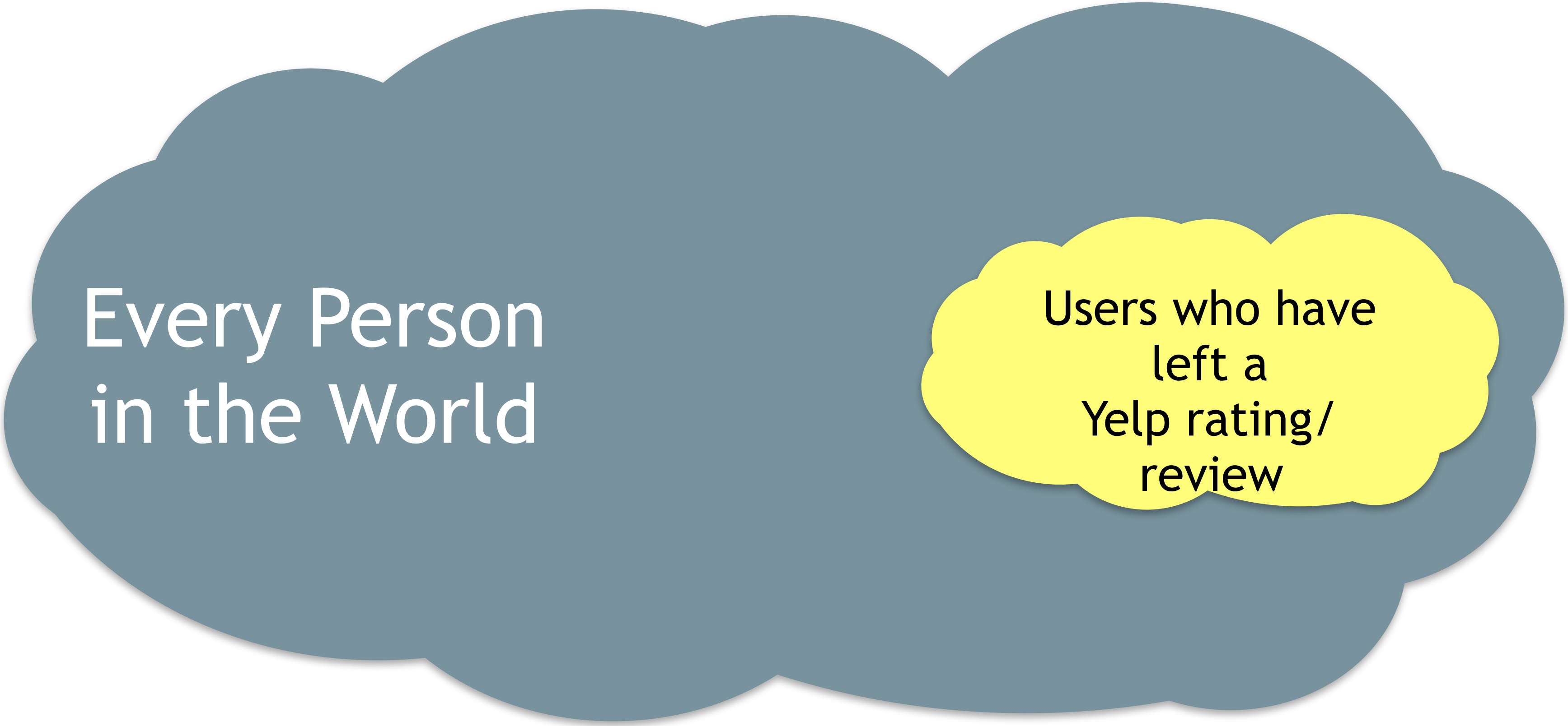


©2016 Pearson, Inc.

Represent your Population by Your Sample

livelessons
©2016 Pearson, Inc.

General Inference

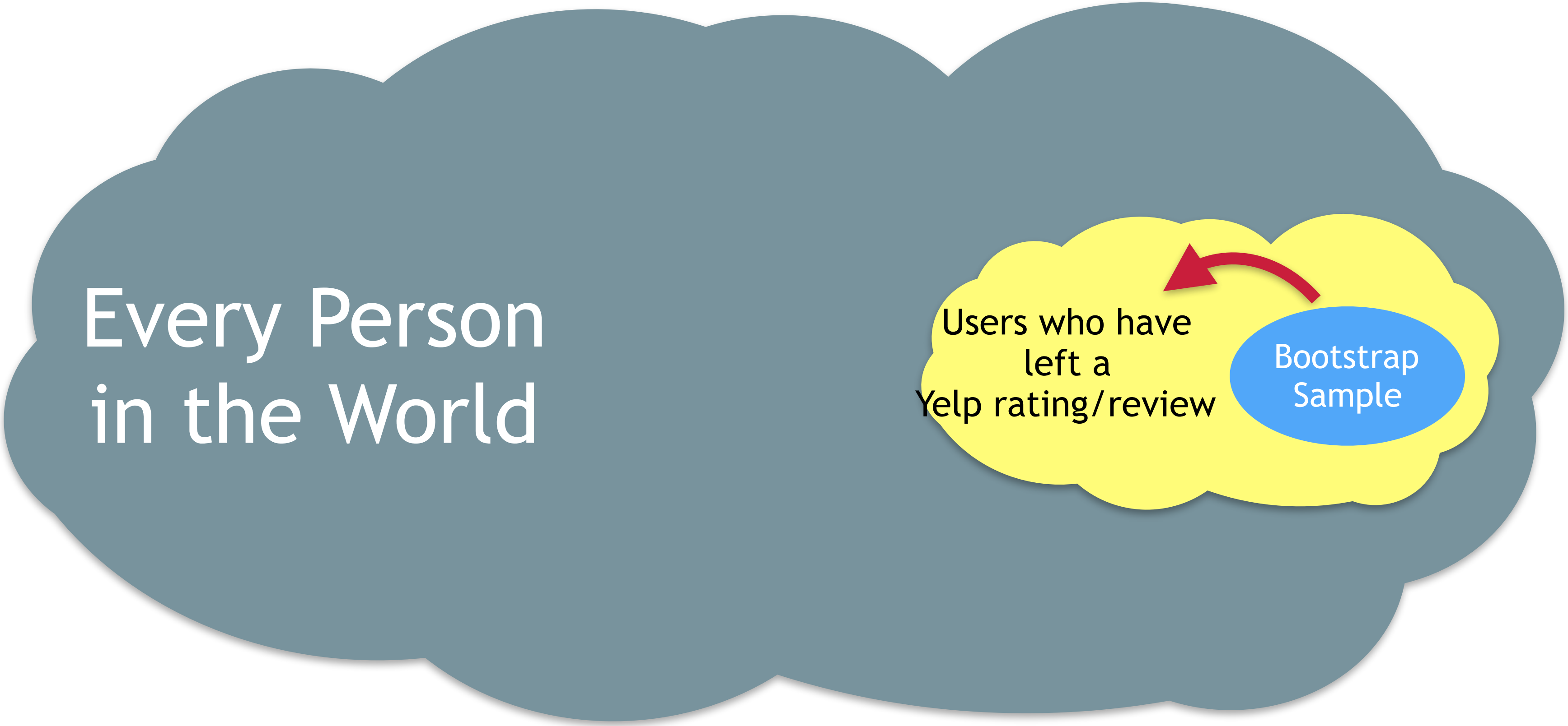


©2016 Pearson, Inc.

Sample your Sample

©2016 Pearson, Inc.

The Bootstrap



livelessons
©2016 Pearson, Inc.

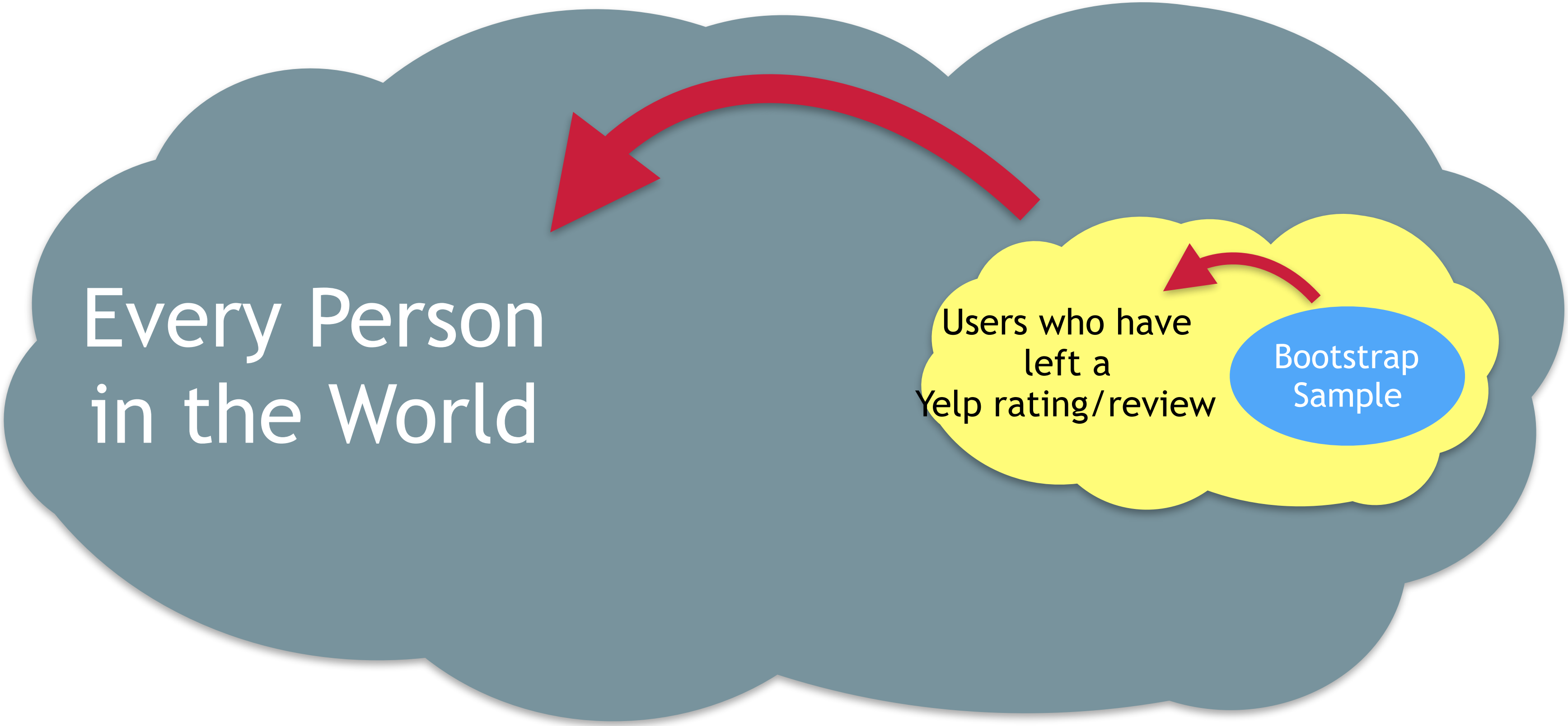
Sample -> Population

Bootstrap sample -> Sample

Bootstrap sample -> Sample -> Population

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

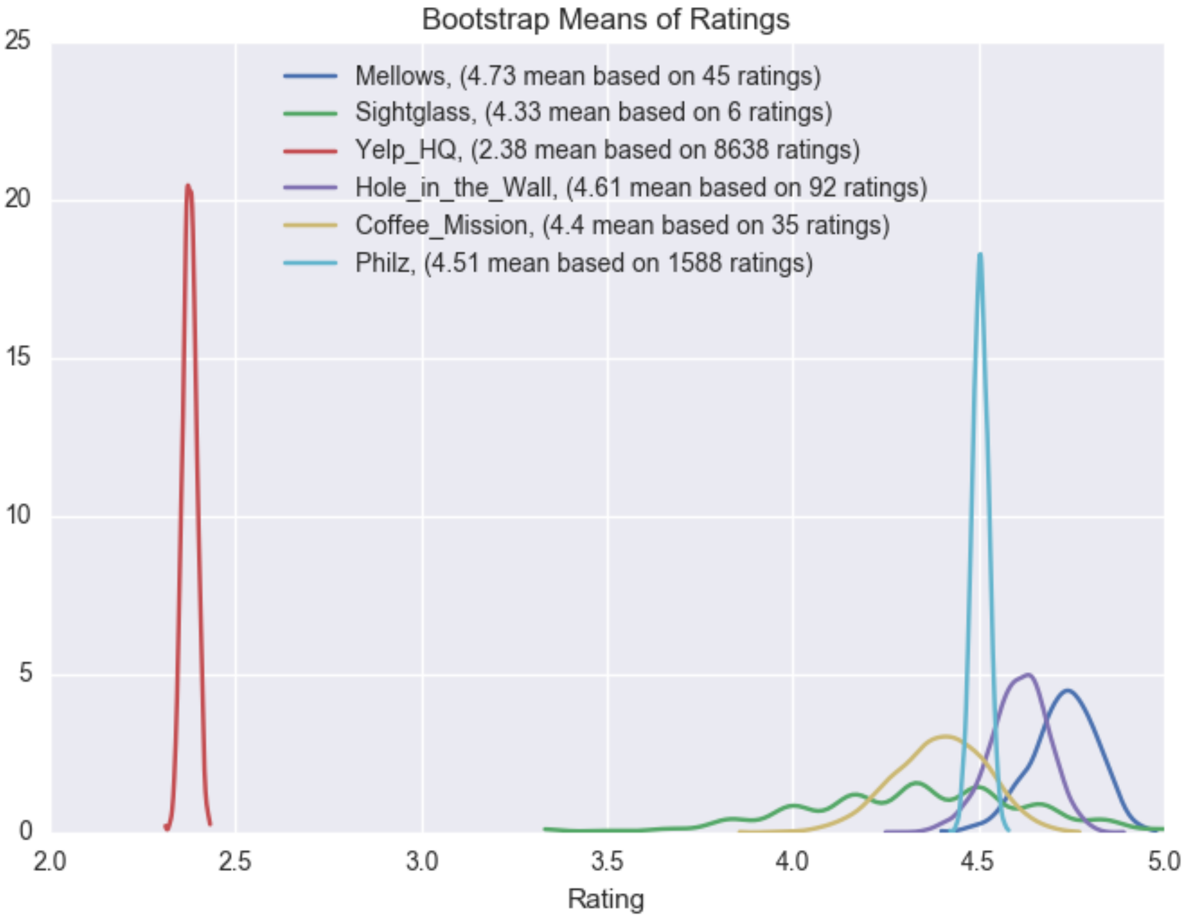
The Bootstrap



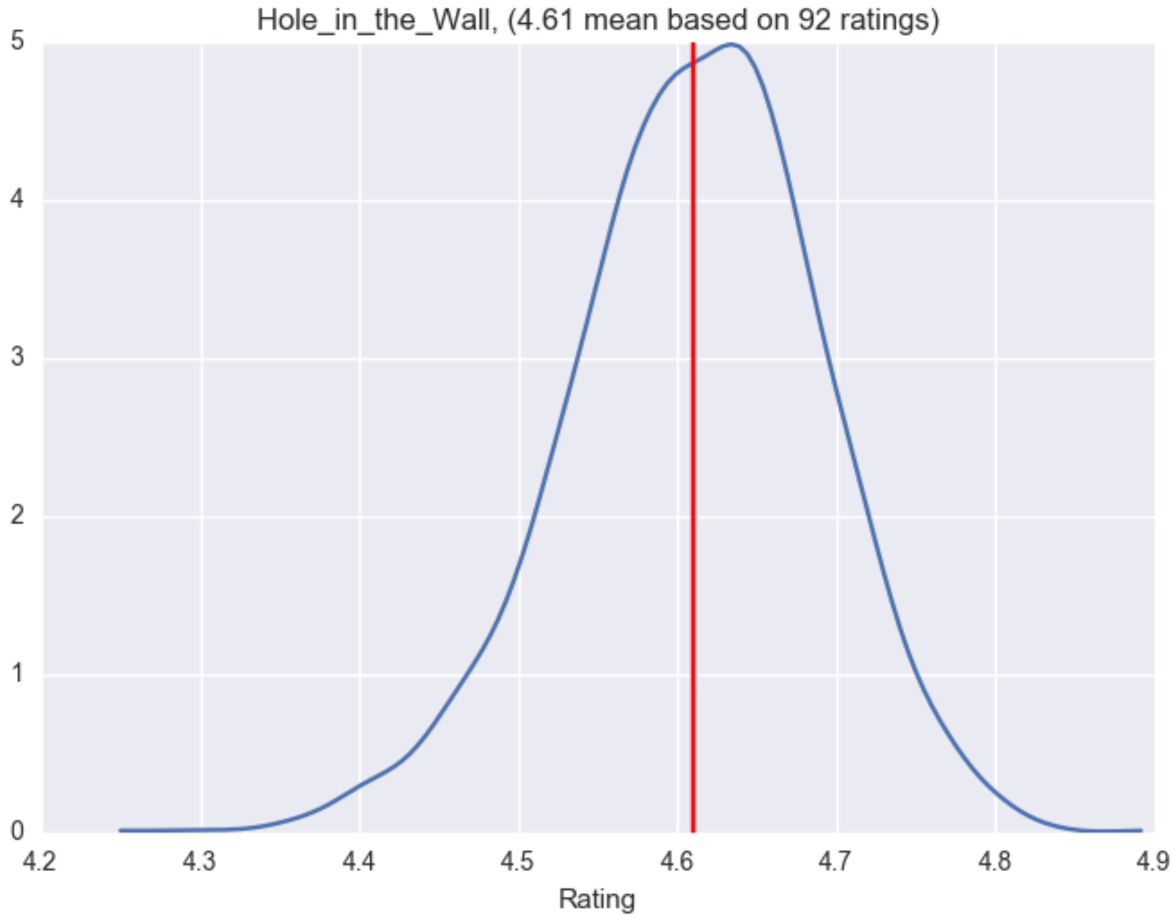
livelessons
©2016 Pearson, Inc.

```python
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())
```

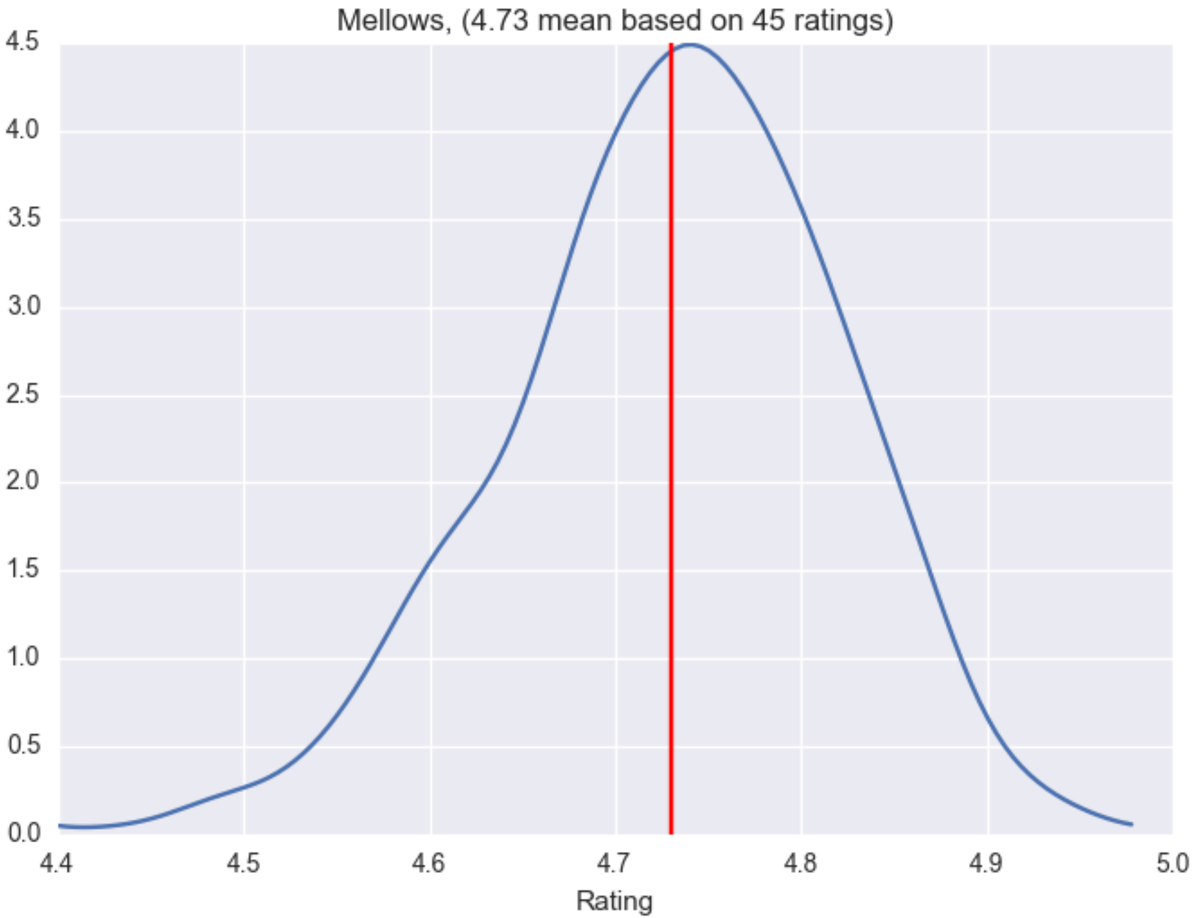
livelessons
©2016 Pearson, Inc.



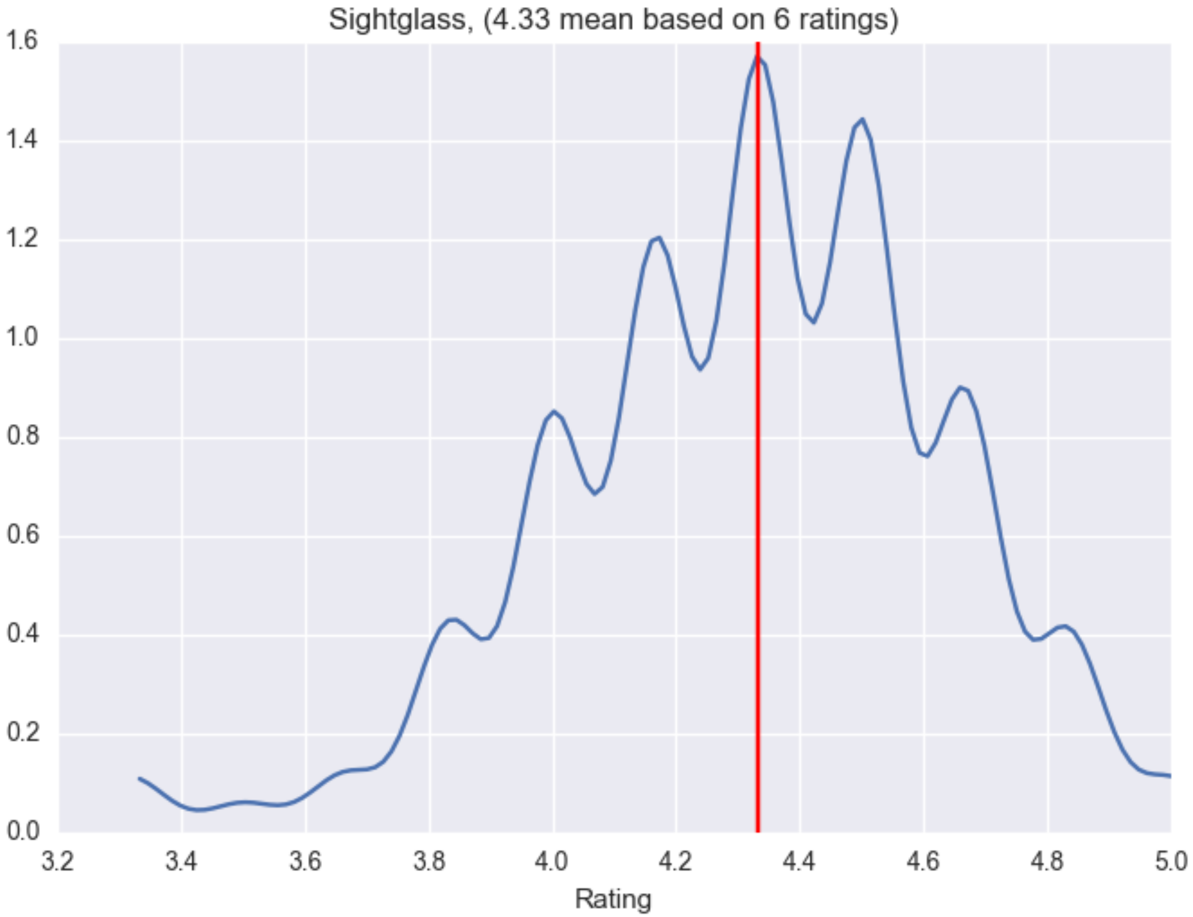
livelessons
©2016 Pearson, Inc.



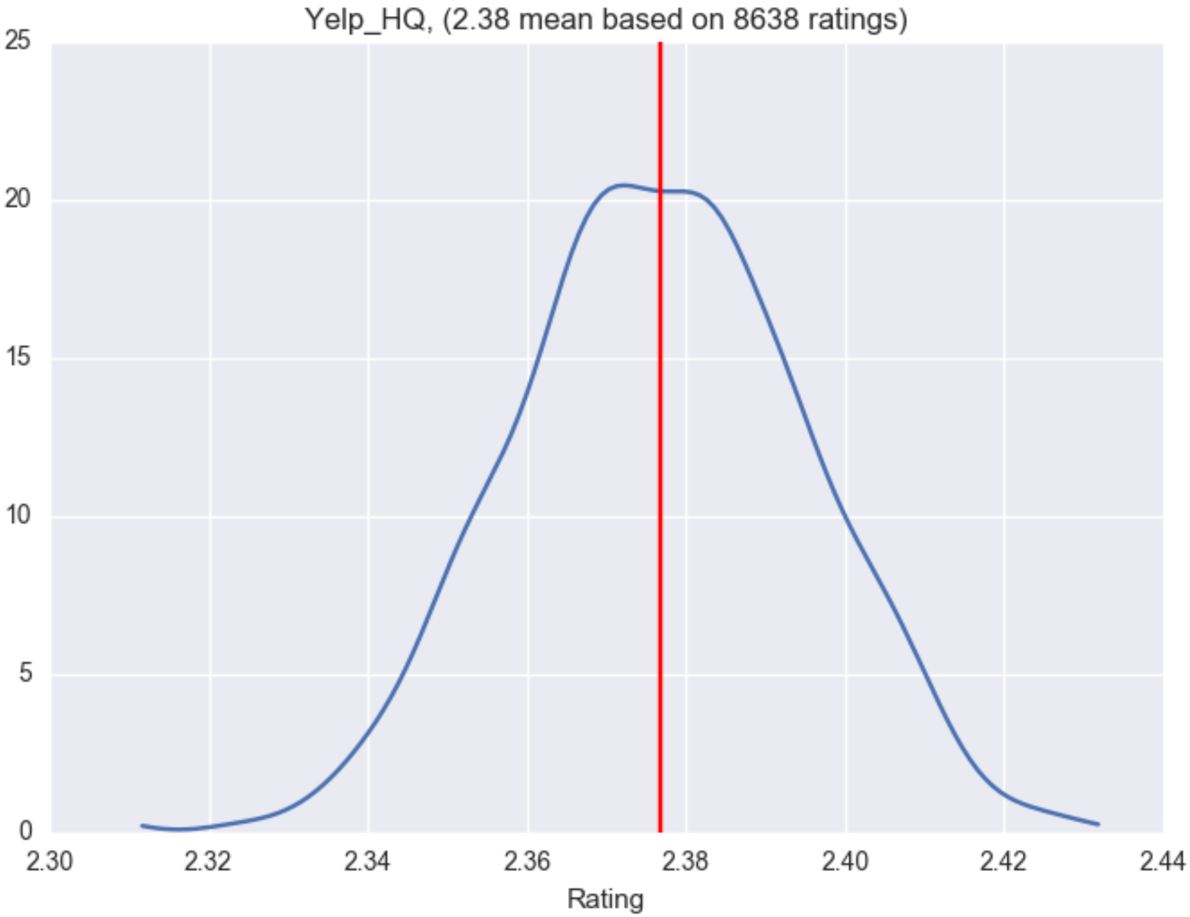
livelessons
©2016 Pearson, Inc.



livelessons
©2016 Pearson, Inc.



livelessons
©2016 Pearson, Inc.



livelessons

©2016 Pearson, Inc.

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

livelessons
©2016 Pearson, Inc.

Back to Non Parametrics

livelessons
©2016 Pearson, Inc.

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

livelessons
©2016 Pearson, Inc.

But now we can use these ideas to understand uncertainty…

livelessons
©2016 Pearson, Inc.

Goals

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

livelessons
©2016 Pearson, Inc.

Classic (frequentist) approach

Population Standard Deviation

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

Size of Sample

for a 95% CI

livelessons
©2016 Pearson, Inc.

Why Bootstrap?

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

livelessons
©2016 Pearson, Inc.

https://www.yelp.com/biz/yelp-san-francisco

Find tacos, cheap dinner, Max's | Near Noe Valley, San Francisco, CA

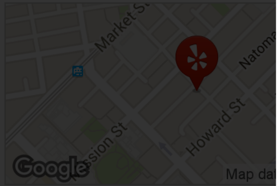
Home About Me Write a Review Find Friends Messages Talk Events

San Francisco > Financial District >

Yelp

8381 revie

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

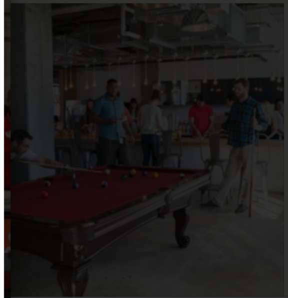
officialblog.yelp.com

"Locally, I enjoy bei
way I do about writ

"Reviews galore, ta
know, love, hate or

"Of course, all thos
forward to attend m

Photo Share Bookmark



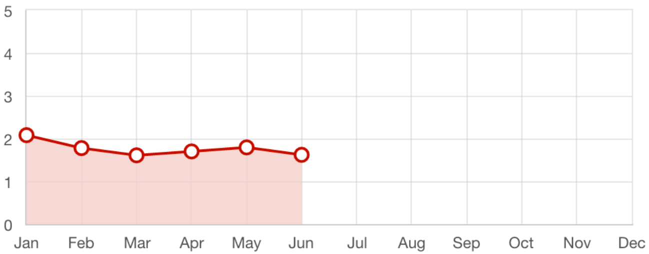
See all 1240

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
Sun Closed

Rating Details

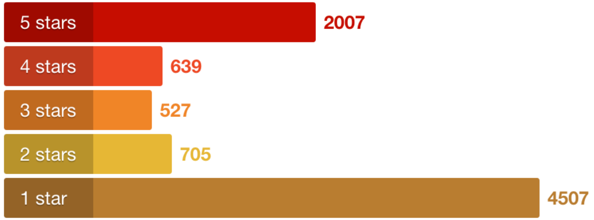
Monthly Trend 2016 2015 2014 2013 2012



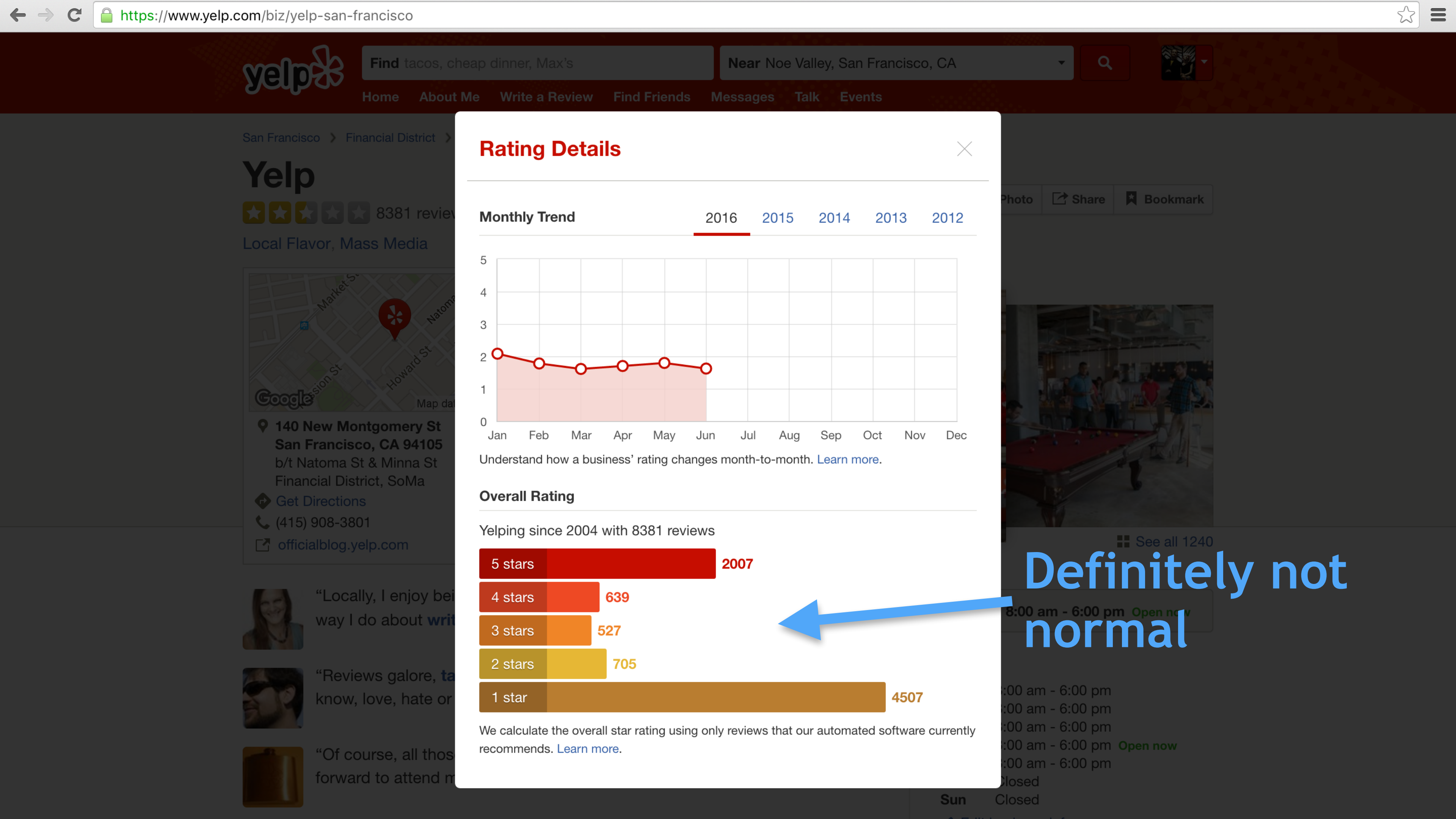
Understand how a business' rating changes month-to-month. Learn more.

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.


https://www.yelp.com/biz/yelp-san-francisco
Find tacos, cheap dinner, Max's
Near Noe Valley, San Francisco, CA
Home About Me Write a Review Find Friends Messages Talk Events
San Francisco > Financial District >
Yelp
8381 revie
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
Rating Details
Monthly Trend
2016 2015 2014 2013 2012
5
4
3
2
1
0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Understand how a business' rating changes month-to-month. Learn more.
Overall Rating
Yelping since 2004 with 8381 reviews
5 stars 2007
4 stars 639
3 stars 527
2 stars 705
1 star 4507
We calculate the overall star rating using only reviews that our automated software currently recommends. Learn more.
Photo
Share
Bookmark
See all 1240
Definitely not normal
Sun Closed

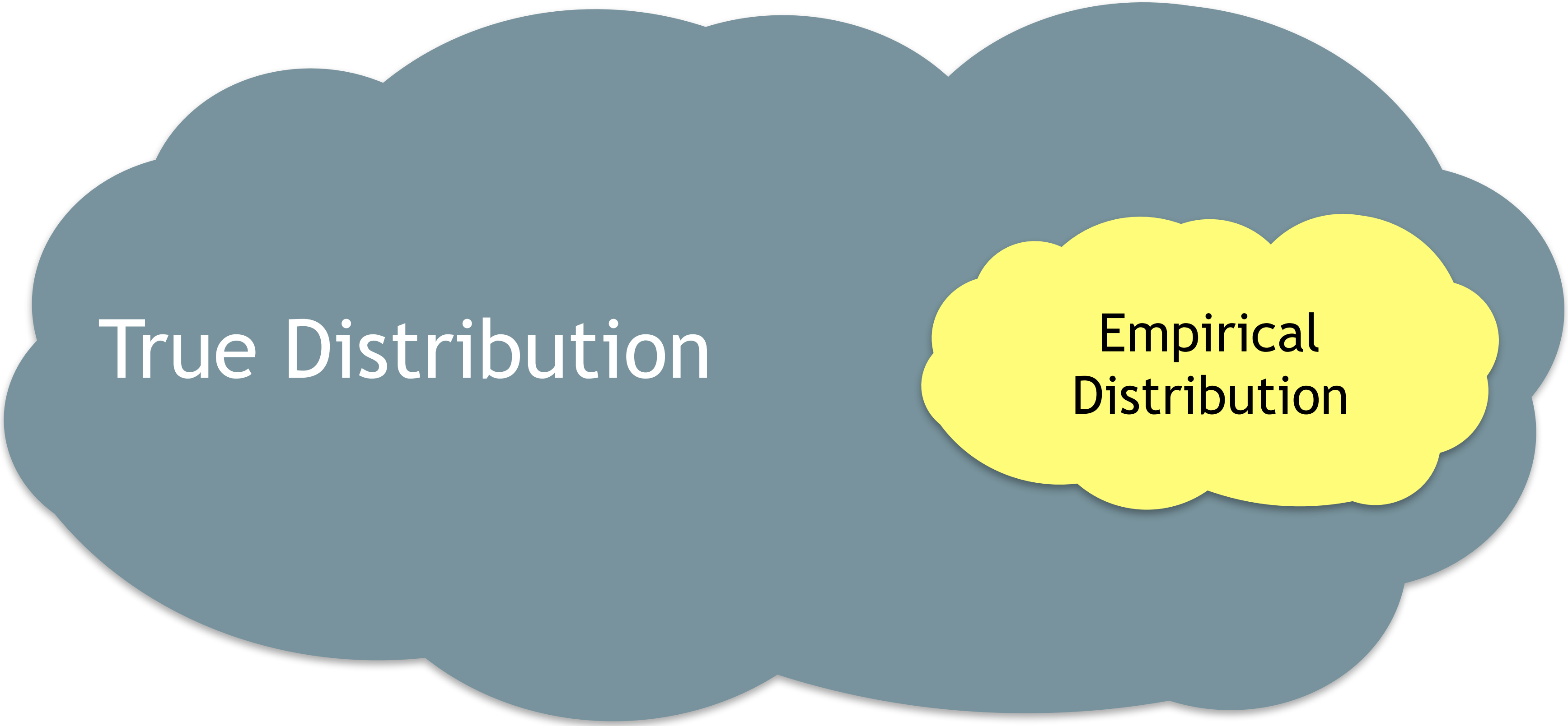
The Bootstrap Formalized

livelessons
©2016 Pearson, Inc.

The Bootstrap

livelessons
©2016 Pearson, Inc.

The Bootstrap

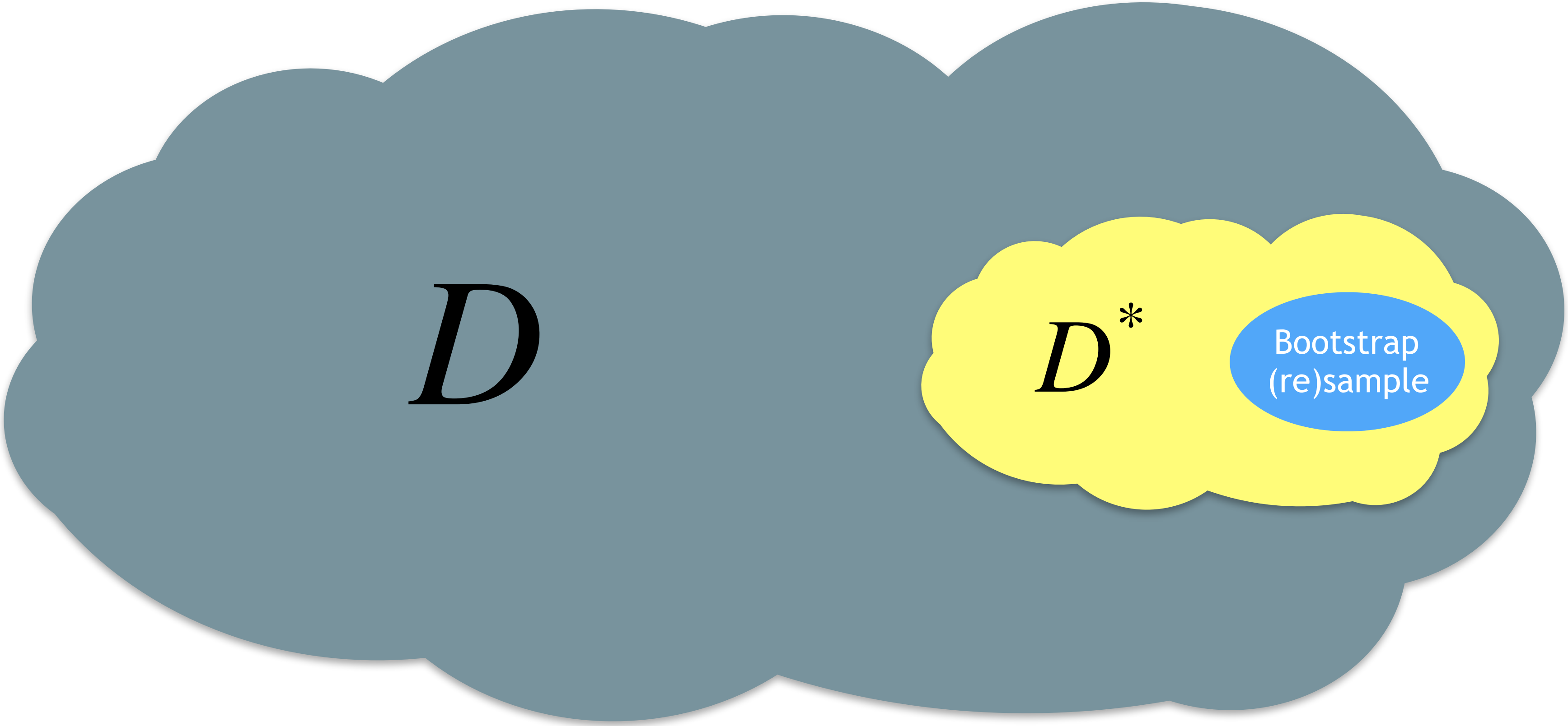


©2016 Pearson, Inc.

The Bootstrap

livelessons
©2016 Pearson, Inc.

The Bootstrap



livelessons
©2016 Pearson, Inc.

Why Bootstrap?

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

livelessons
©2016 Pearson, Inc.

The Bootstrap Principle

Distribution Approximation

$$D^* \approx D$$

Statistic Approximation

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

$$s^* \rightarrow s$$

livelessons
©2016 Pearson, Inc.

Just treat the confidence interval as our statistic of interest

livelessons
©2016 Pearson, Inc.

Statistic of Interest

Sample mean

True mean

Difference in means
for one sample

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

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

livelessons
©2016 Pearson, Inc.

Bootstrap Approximation

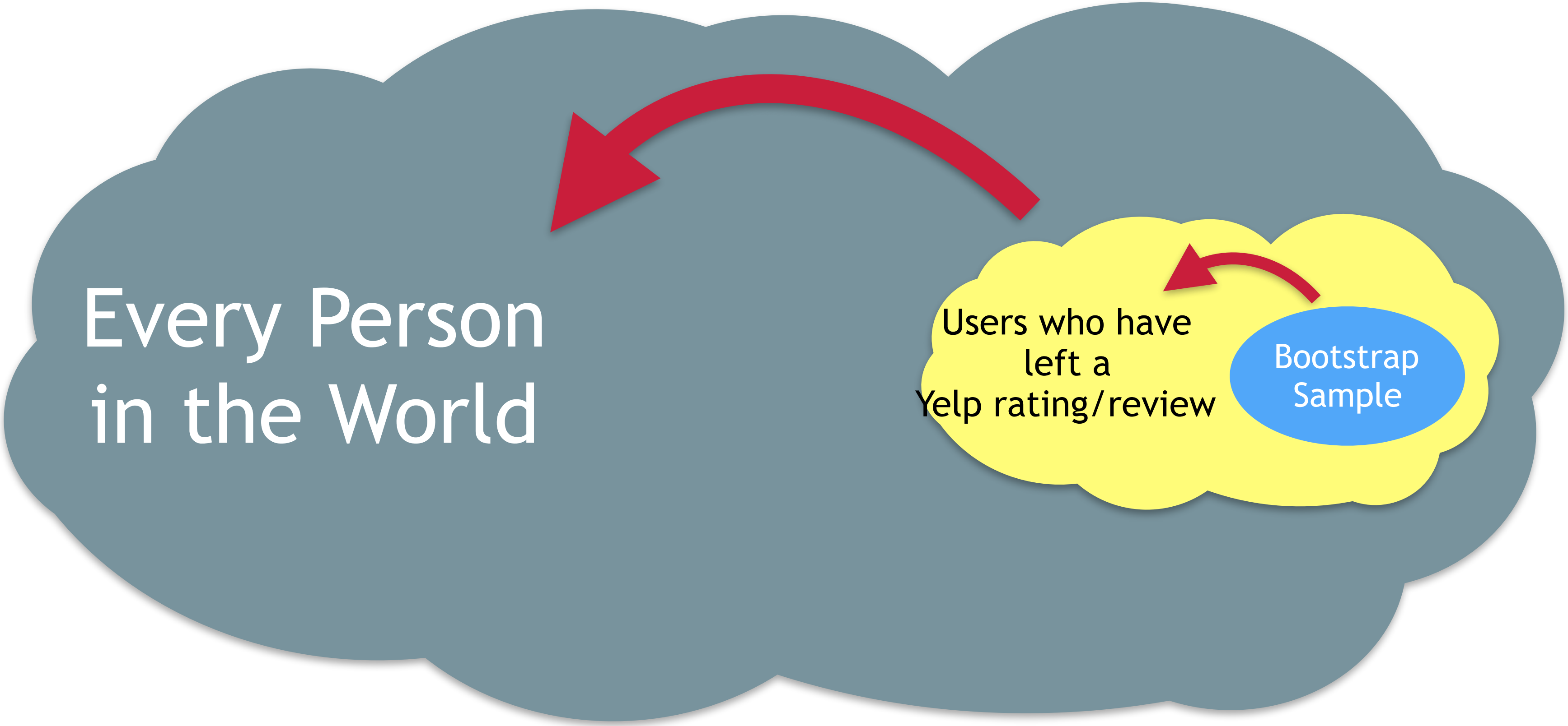
For any statistic ***s***, the (resampled) distribution of ***s**** approximates ***s***

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

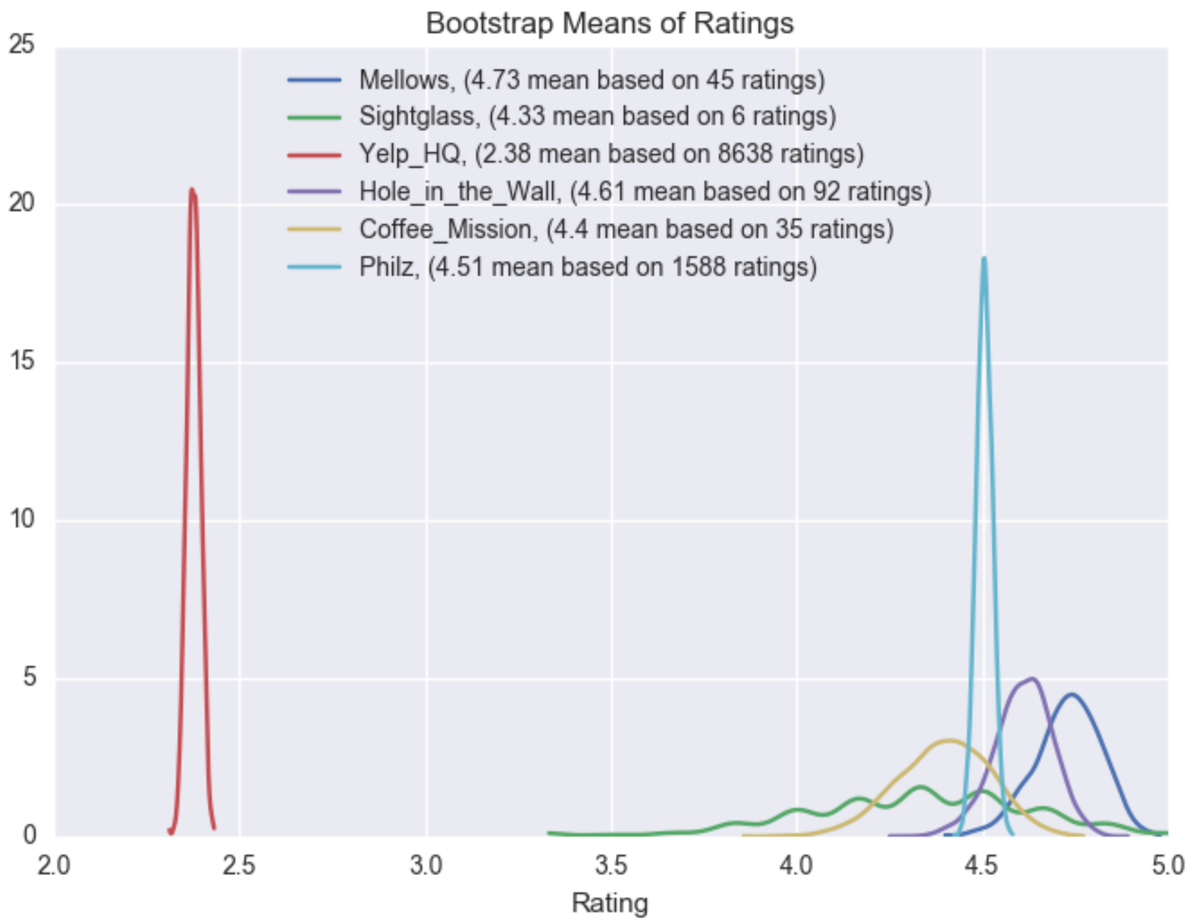
$$\theta^* \rightarrow \theta$$

livelessons
©2016 Pearson, Inc.

The Bootstrap



livelessons
©2016 Pearson, Inc.



livelessons
©2016 Pearson, Inc.

```
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

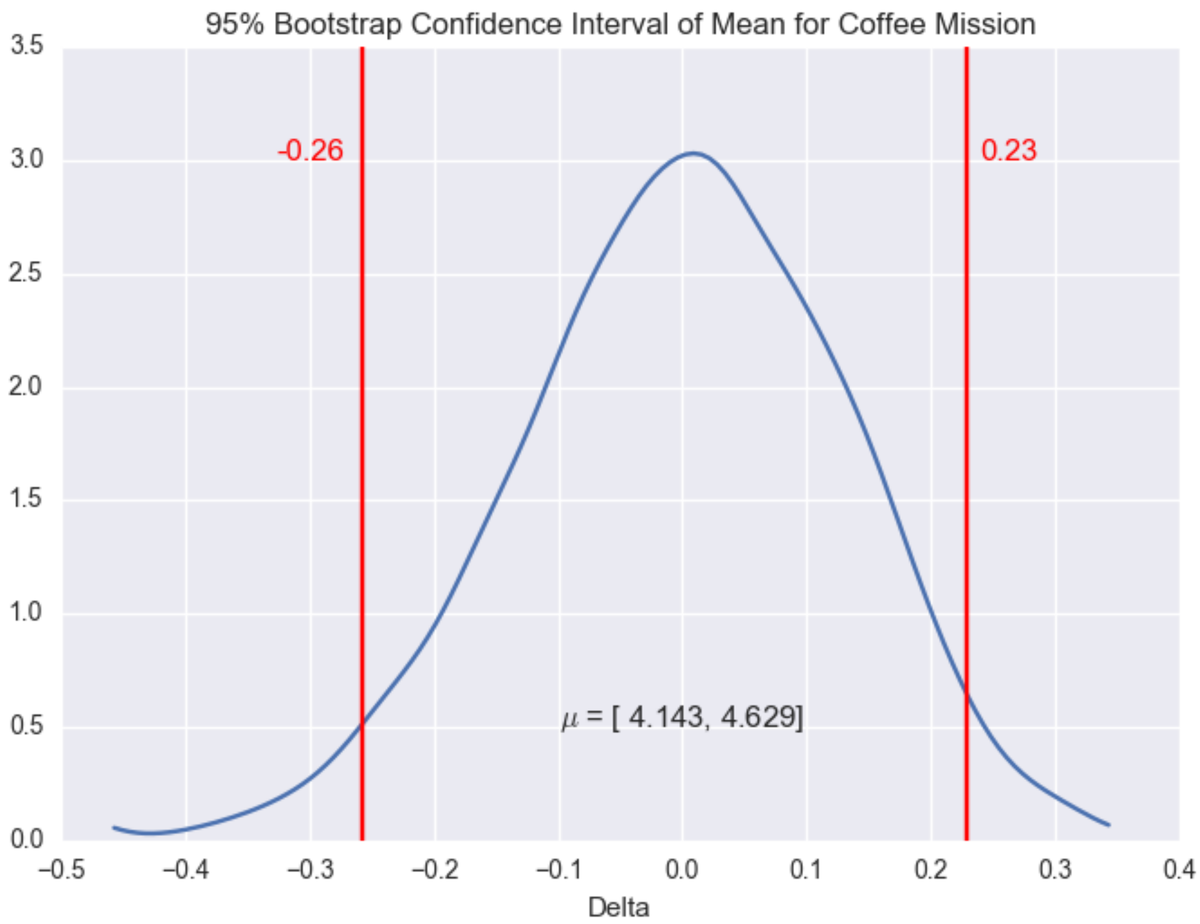
livelessons

©2016 Pearson, Inc.

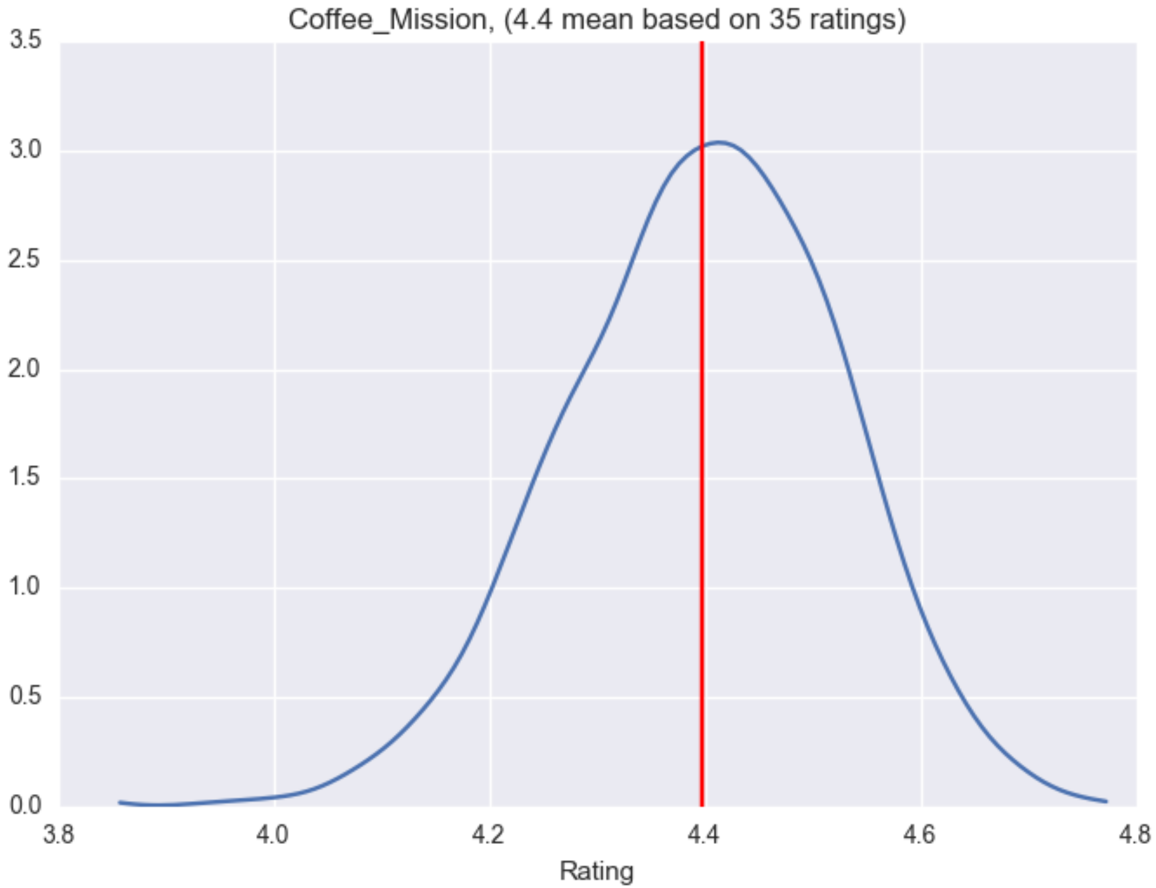
```
df = pd.read_csv('yelp_stars/' + filename)
num_samples = len(df)
boots = []
x_bar = df.rating.mean()

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

Statistic of interest

livelessons
©2016 Pearson, Inc.



livelessons
©2016 Pearson, Inc.



livelessons
©2016 Pearson, Inc.

```
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
```

livelessons
©2016 Pearson, Inc.

rpsychologist.com/d3/CI/

Interpreting Confidence Intervals

an interactive visualization

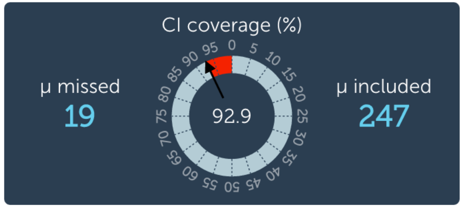
Created by Kristoffer Magnusson

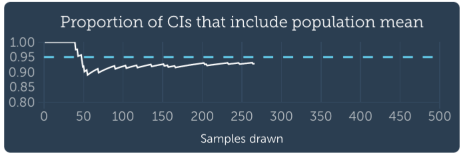
Follow @krstoffr | 1,298 follower | Kristoffer's LinkedIn profile

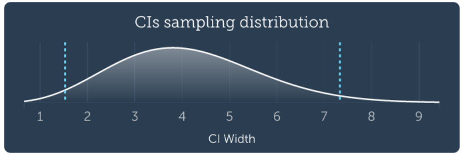
Tweet | Gilla | Dela | 225

Slide me

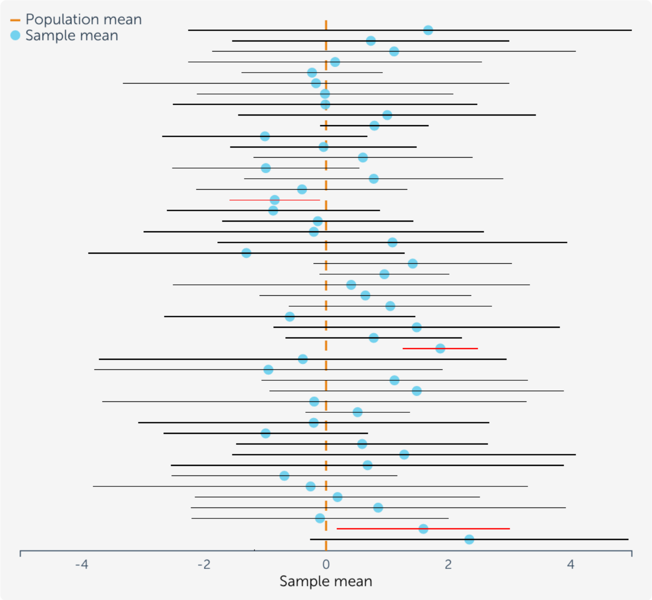
Simulation statistics







93% confidence intervals



Cheaper Beds, **Better** Breakfasts

livelessons
©2016 Pearson, Inc.

Exploratory Question

What variables affect the price of a listing?

livelessons
©2016 Pearson, Inc.

Hypothesis Statement

“Superhosts” have more expensive listings

livelessons
©2016 Pearson, Inc.

Hypothesis Statement

“Superhosts” have more expensive listings

(implicit assumption: since they can attract more guests)

livelessons
©2016 Pearson, Inc.

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)

livelessons
©2016 Pearson, Inc.

Hypothesis Statement

More on this later…

livelessons
©2016 Pearson, Inc.

Question for Inference

What variables are correlated with the price of a listing?

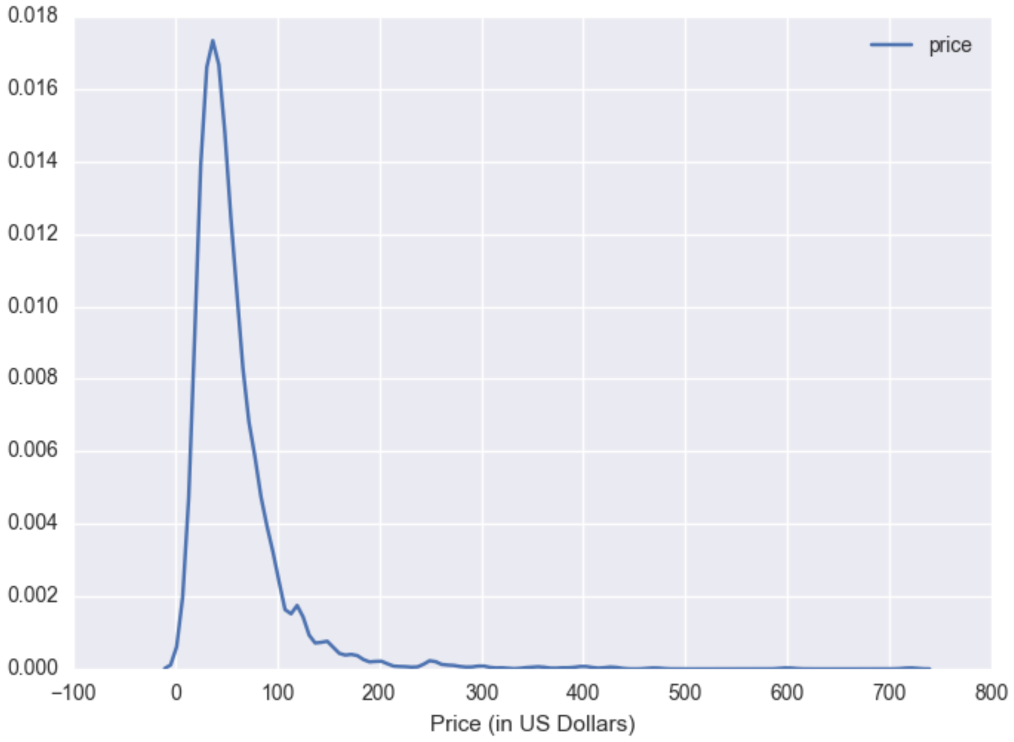
livelessons
©2016 Pearson, Inc.

Correlation

Determining relationship between two variables

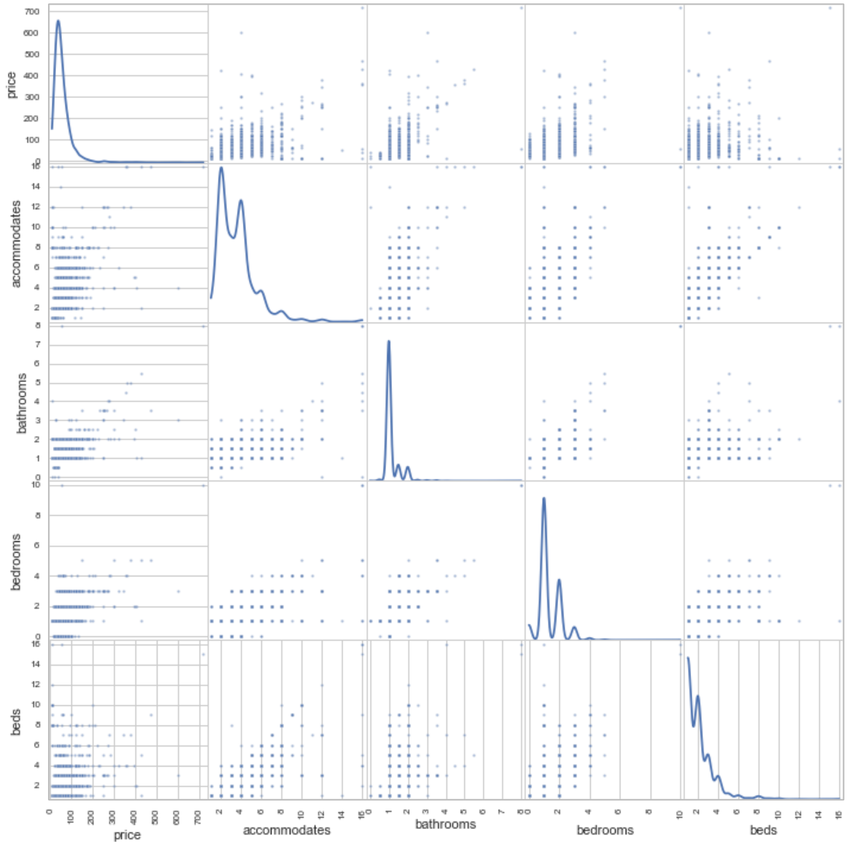
livelessons
©2016 Pearson, Inc.

Price Distribution for Athens



livelessons
©2016 Pearson, Inc.

Correlation: Visually



livelessons
©2016 Pearson, Inc.

(Pearson's) Correlation Analytically

Covariance*

Standard Deviation* of x

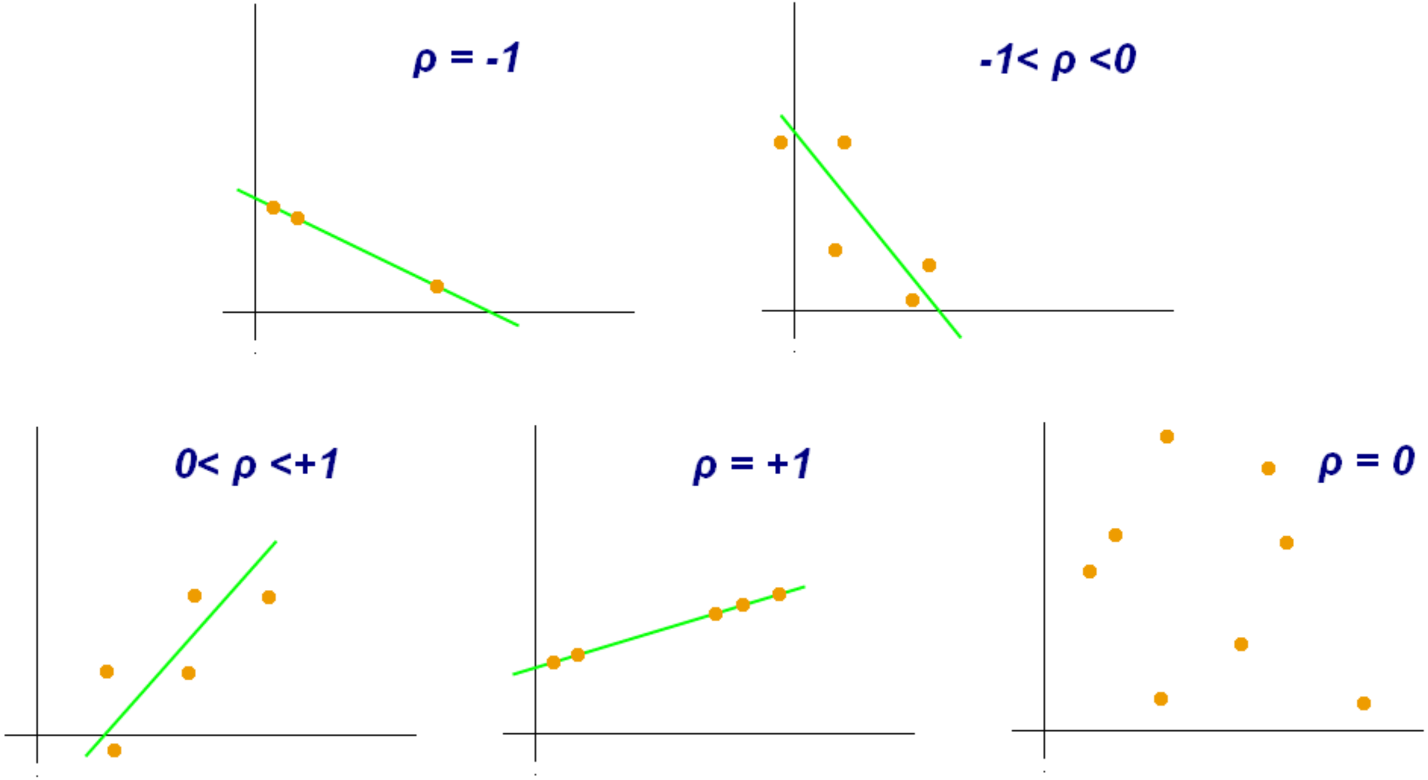
Standard Deviation* of y

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

*without normalizing 1/N factors

livelessons
©2016 Pearson, Inc.

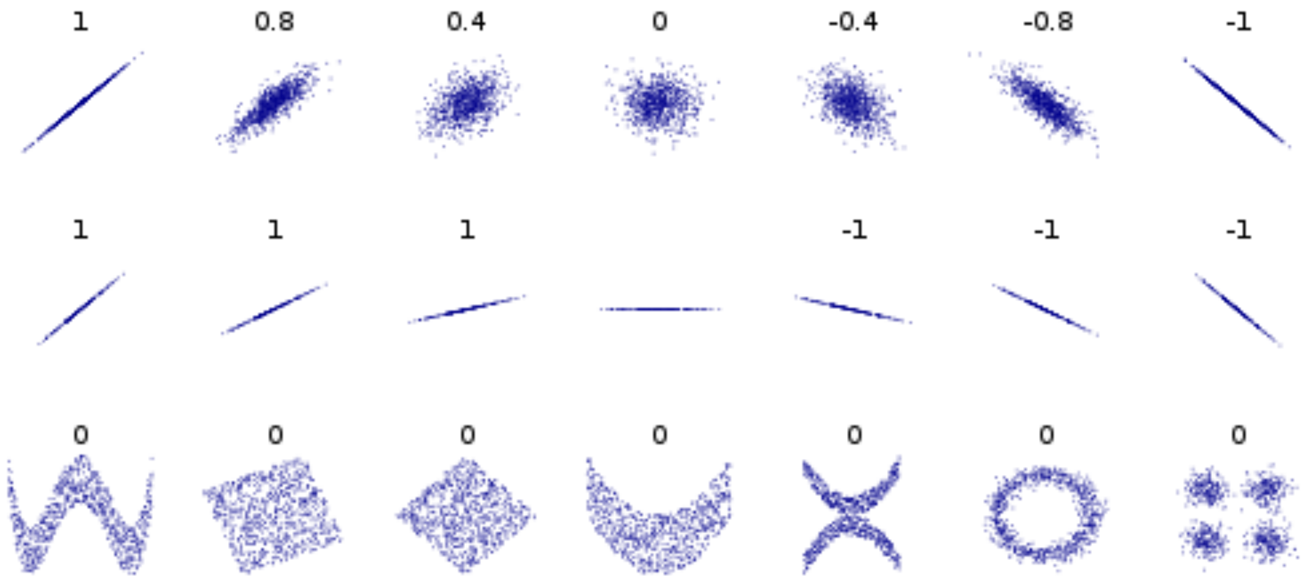
Correlation: Visually



source: Wikipedia

livelessons
©2016 Pearson, Inc.

Correlation: Visually



source: Wikipedia

livelessons
©2016 Pearson, Inc.

Correlation Confidence Intervals

Live Code

livelessons
©2016 Pearson, Inc.

Correlation Confidence Intervals

Problematic to calculate analytically…

livelessons
©2016 Pearson, Inc.

Correlation Confidence Intervals

Sampling distribution is not normally distributed…

livelessons
©2016 Pearson, Inc.

Correlation Confidence Intervals

To the bootstrap!

livelessons
©2016 Pearson, Inc.

Correlation Confidence Intervals

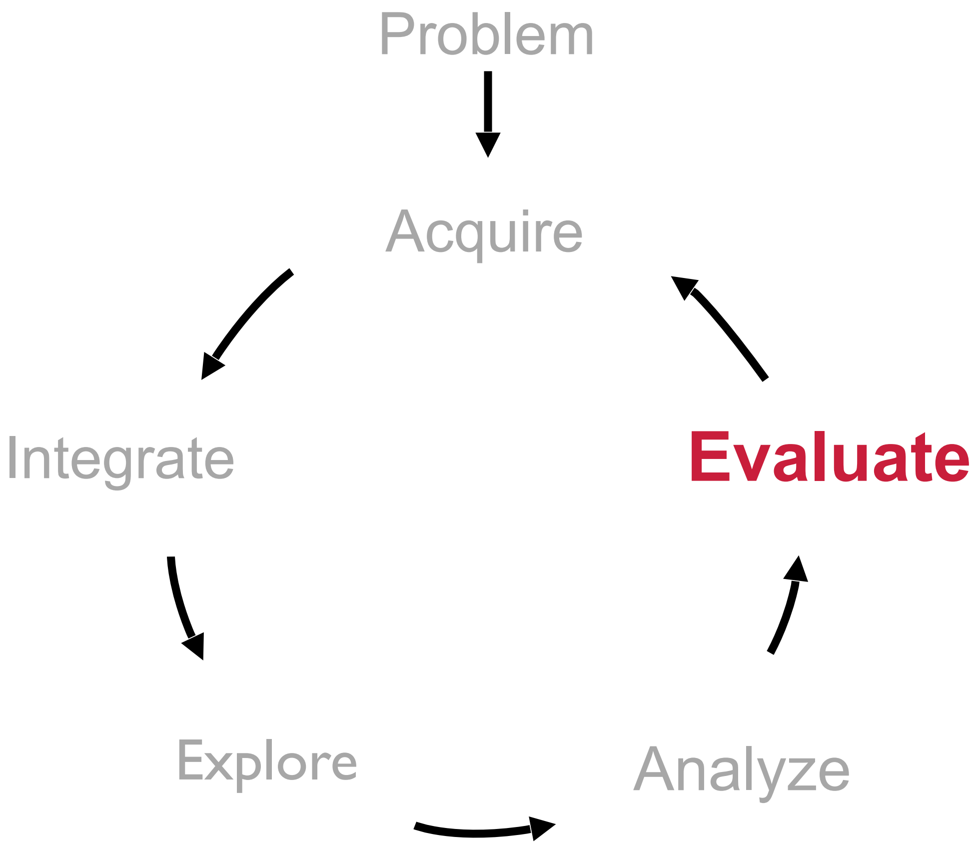
Live Code

livelessons
©2016 Pearson, Inc.

Hypothesis Testing

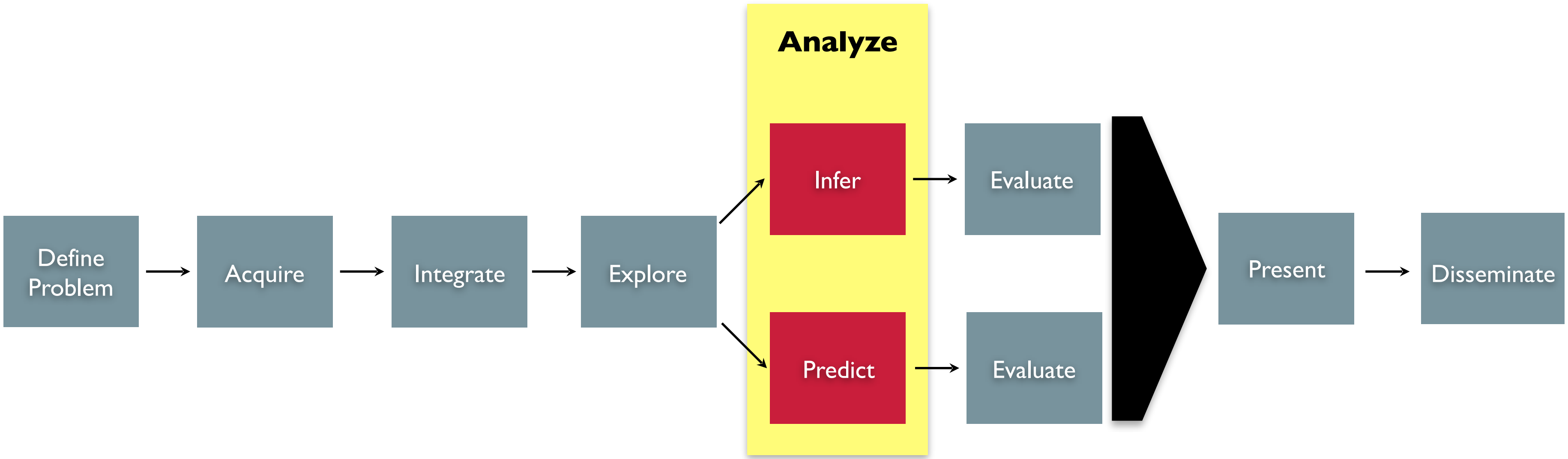
livelessons
©2016 Pearson, Inc.

Process + Iteration



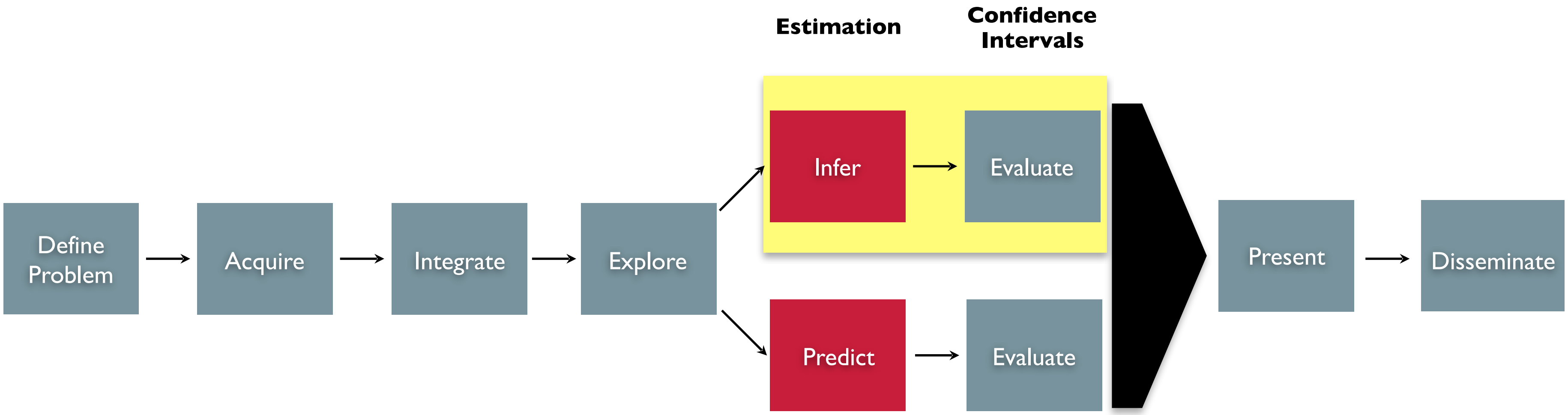
livelessons
©2016 Pearson, Inc.

Process + Statistics



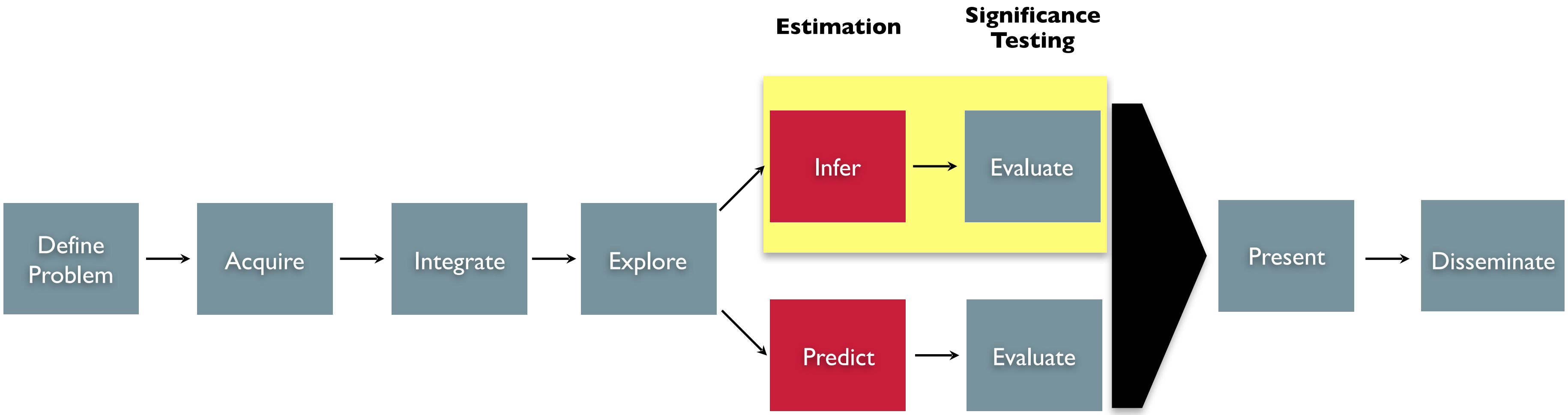
livelessons
©2016 Pearson, Inc.

Process + Statistics



livelessons
©2016 Pearson, Inc.

Process + Statistics



livelessons
©2016 Pearson, Inc.

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

livelessons
©2016 Pearson, Inc.

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

livelessons
©2016 Pearson, Inc.

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.

livelessons
©2016 Pearson, Inc.

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`

livelessons
©2016 Pearson, Inc.

Hypothesis Testing: Outcomes

Reject the null, accept the alternative

— OR —

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

livelessons
©2016 Pearson, Inc.

Hypothesis Testing: Outcomes

Instant booking increases the amount of guests a listing receives

— OR —

“Not Guilty”

livelessons
©2016 Pearson, Inc.

Hypothesis Testing: Mechanics

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

livelessons
©2016 Pearson, Inc.

Hypothesis Testing: Mechanics

But this time two different samples…

livelessons
©2016 Pearson, Inc.

Hypothesis Testing: Mechanics

Live Code

livelessons
©2016 Pearson, Inc.

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

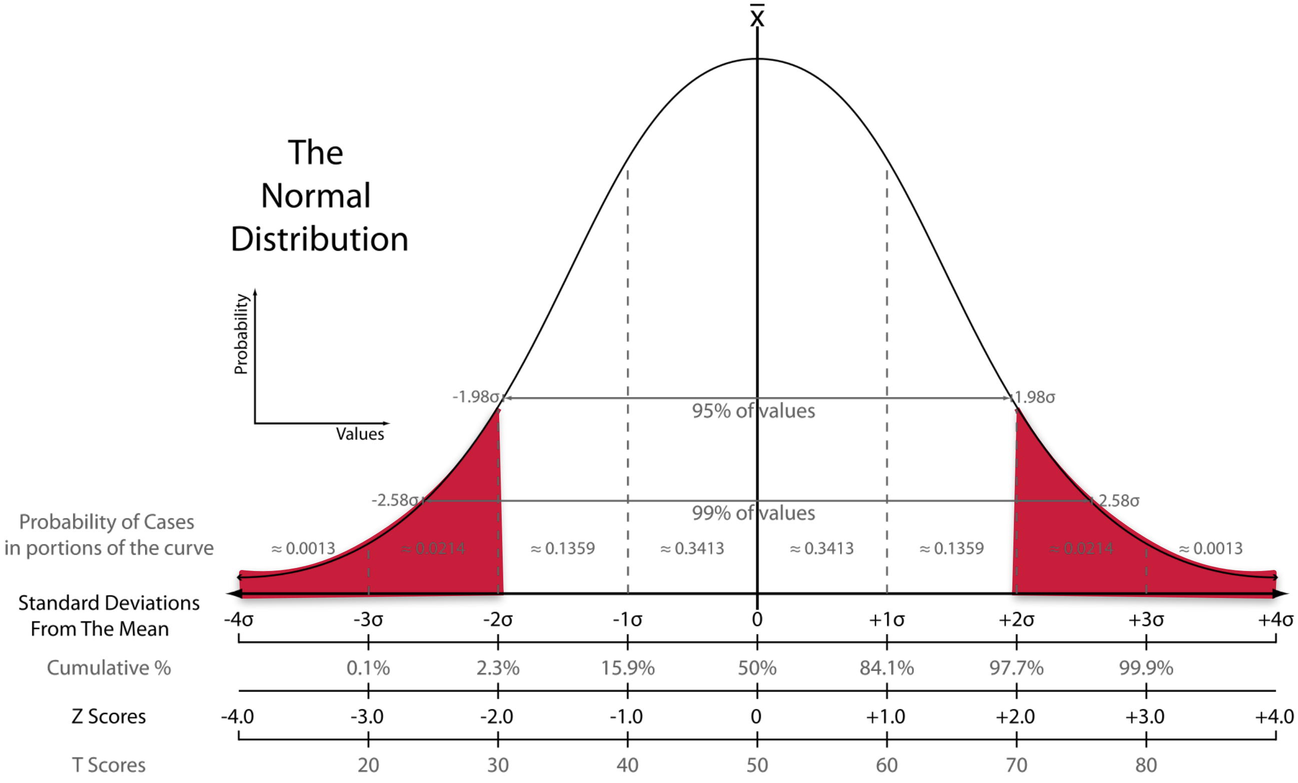
True Condition (H_0)

Predicted Condition (H_0)

livelessons

©2016 Pearson, Inc.

Our Familiar Friend



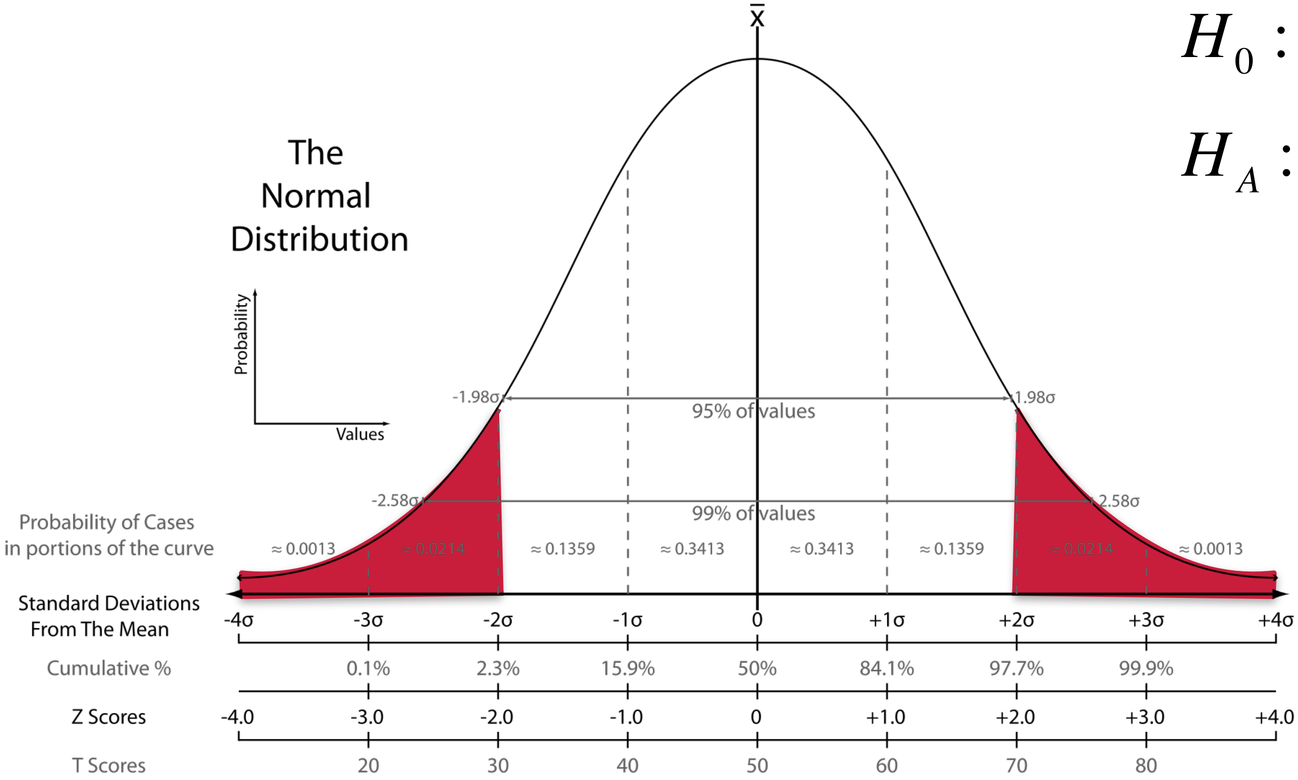
source: Wikipedia

livelessons
©2016 Pearson, Inc.

Two Sided

$$H_0 : \delta = 0$$

$$H_A : \delta \neq 0$$



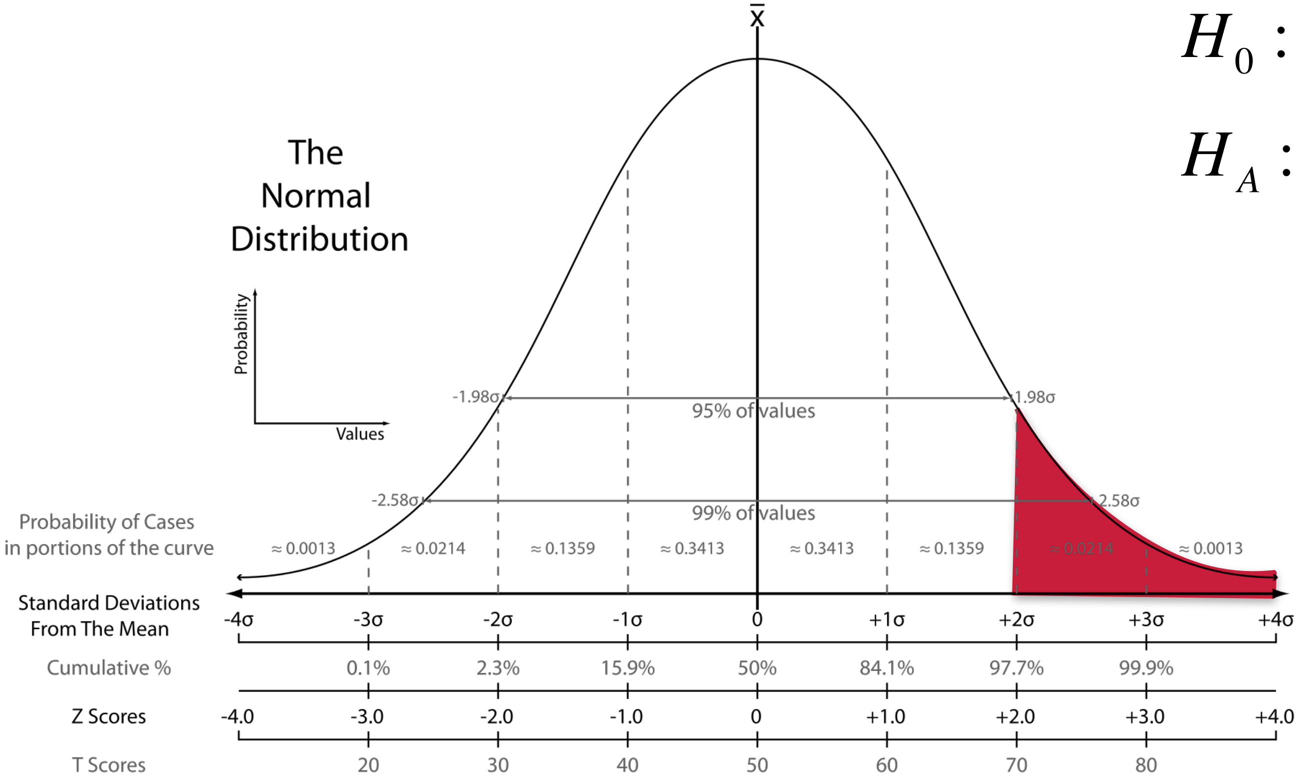
source: Wikipedia

livelessons
©2016 Pearson, Inc.

One Sided

$$H_0 : \delta \leq 0$$

$$H_A : \delta > 0$$



source: Wikipedia

livelessons
©2016 Pearson, Inc.

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**

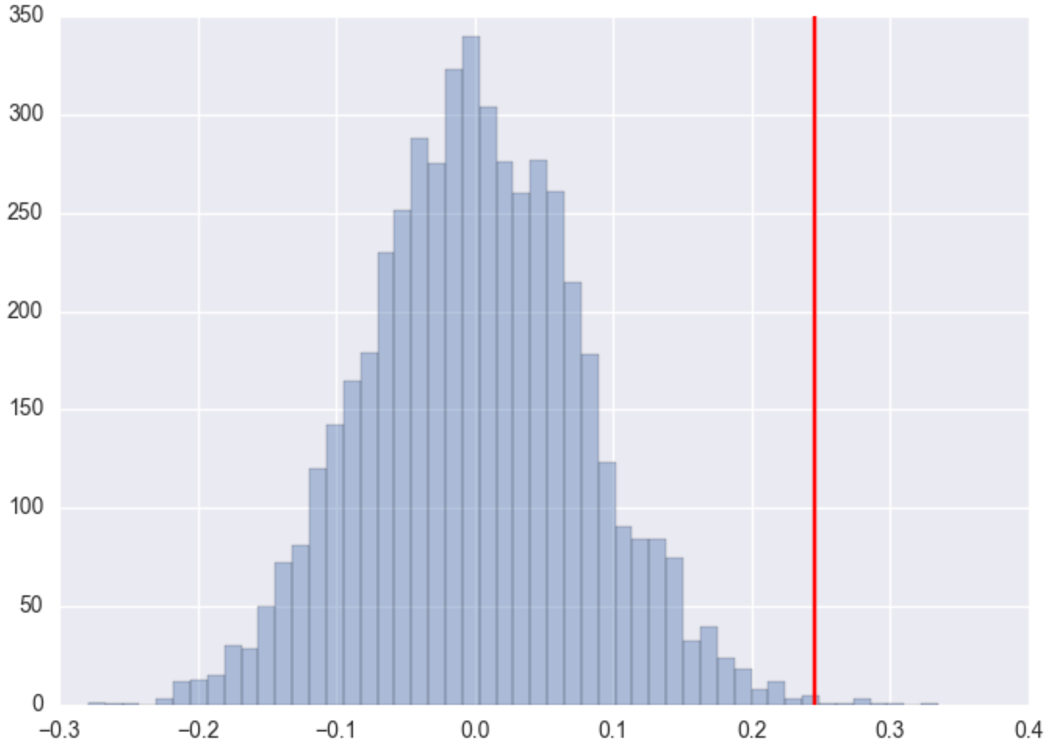
p-value

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

source: Wikipedia

livelessons
©2016 Pearson, Inc.

Sampling Distribution of Instant Bookable



©2016 Pearson, Inc.

Hypothesis Testing Terminology

Power of the Test

The probability that the test correctly rejects the null hypothesis (H0) when the alternative hypothesis (H1) 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.

source: Wikipedia

livelessons
©2016 Pearson, Inc.

rpsychologist.com/d3/NHST/

Settings

Solve for? ◉ Power ○ Alpha ○ n ○ d

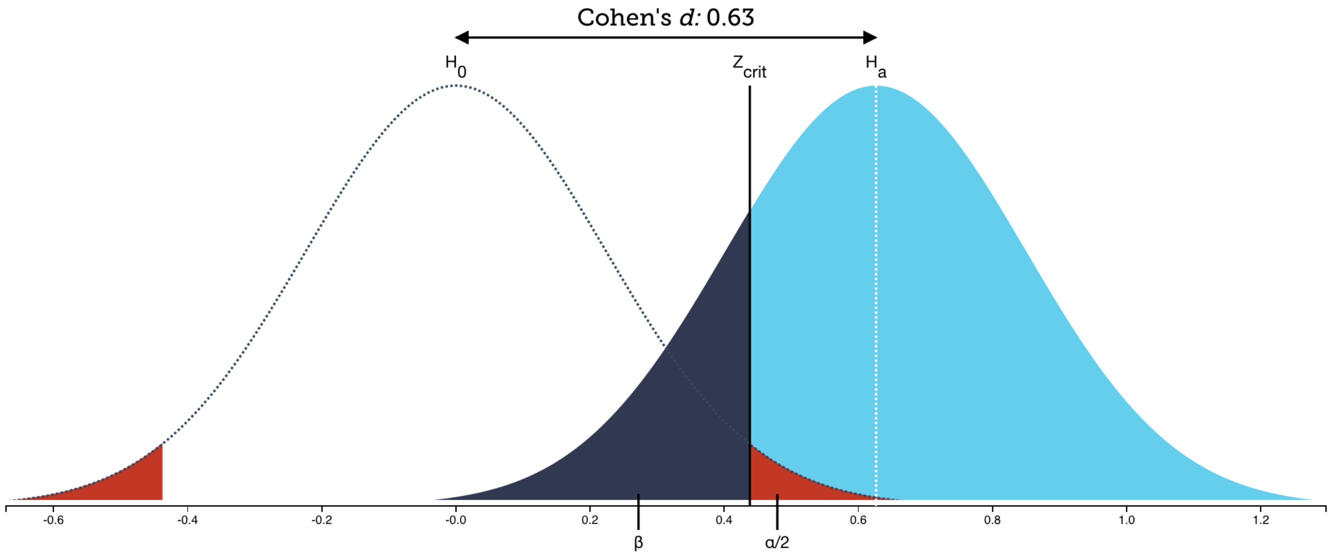
Significance level (α = 0.05)

Sample size (n = 20)

Effect size (d = 0.63)

One-tailed | Two-tailed

Reset zoom



5 %	20 %	80 %	20
Type I error	**Type II error**	**Power**	**Sample size**

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

livelessons
©2016 Pearson, Inc.

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.

livelessons
©2016 Pearson, Inc.

Hypothesis Testing: Outcomes

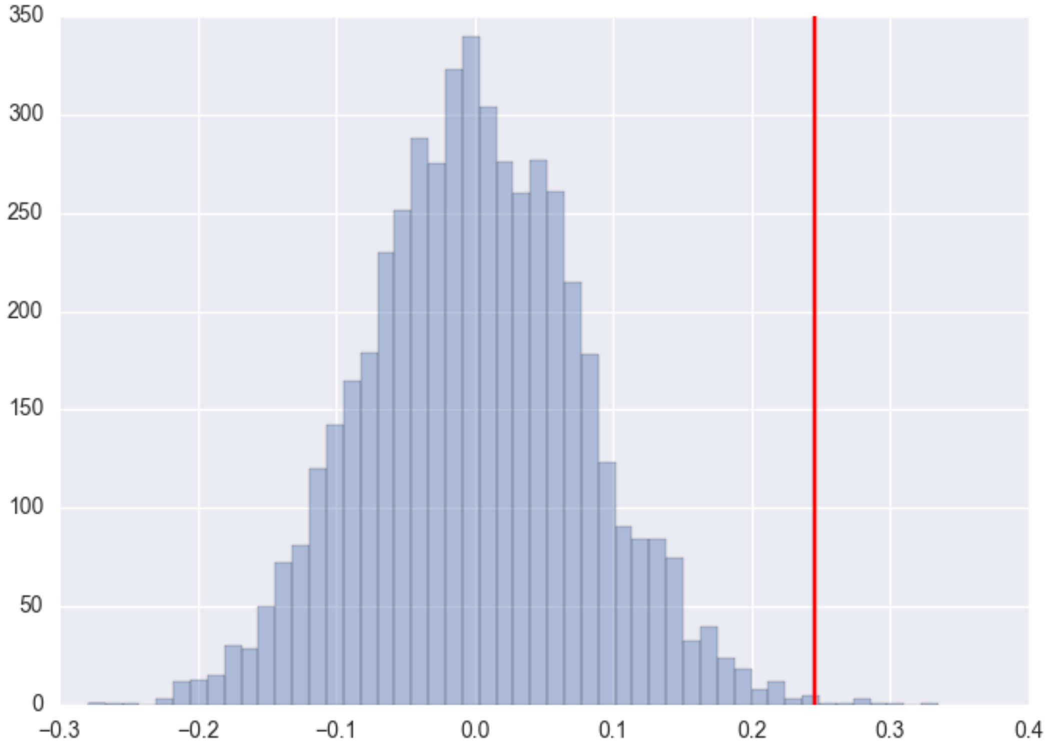
Instant booking increases the amount of guests a listing receives

— OR —

“Not Guilty”

livelessons
©2016 Pearson, Inc.

Sampling Distribution of Instant Bookable



livelessons
©2016 Pearson, Inc.

Hypothesis Testing: Outcomes

Instant booking increases the amount of guests a listing receives

— OR —

“Not Guilty”

livelessons
©2016 Pearson, Inc.

Business Decisions and Actions

What **action** to take and **when** to take it

livelessons
©2016 Pearson, Inc.

Business Decisions and Actions

- Provide this **insight/statistic** when new users are **creating their listing**
- Suggest changing their listing to **enable instant-bookable** for hosts who are **struggling to attract guests**.
- Suggest changing their listing to **disable instant-bookable** for hosts who are **receiving too many guests**.

livelessons
©2016 Pearson, Inc.

Business Decisions and Actions

Context and goal dependent….

livelessons
©2016 Pearson, Inc.

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

livelessons
©2016 Pearson, Inc.

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

livelessons
©2016 Pearson, Inc.

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

livelessons
©2016 Pearson, Inc.

Spectrum of Statistics

Non-parametric Parametric Bayesian

Less Assumptions ⟷ More (statistical) Power & Specificity

livelessons
©2016 Pearson, Inc.

Spectrum of Statistics

Non-parametric Parametric Bayesian

More Objective ⟷ More Subjective

livelessons
©2016 Pearson, Inc.

Spectrum of Statistics

Non-parametric | Parametric | Bayesian

Easier to Learn?
(and apply)

More Subjective

livelessons
©2016 Pearson, Inc.

Spectrum of Statistics

Non-parametric

Parametric

Bayesian

Less Assumptions

More (statistical) Power & Specificity

Middle Ground

livelessons
©2016 Pearson, Inc.

Spectrum of Statistics

Non-parametric | Parametric | Bayesian

Less Assumptions ⟷ More (statistical) Power & Specificity

Tends to be most popular/common

livelessons
©2016 Pearson, Inc.

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

livelessons©
©2016 Pearson, Inc.

Spectrum of Statistics

Non-parametric

Parametric

Bayesian

Less Assumptions

More
(statistical) Power

Most Differences are cosmetic and philosophical

Tends to be most
popular/common

livelessons©
©2016 Pearson, Inc.

Next Up: Modeling and Machine Learning

livelessons
©2016 Pearson, Inc.