Estimators and Sampling

Justin M. Rao, MSR-NYC

Outline

1) Basics of estimation
2) Samples vs. populations
3) Sampling theory, IID assumptions and departures
4) Regression: linear and non-parametric

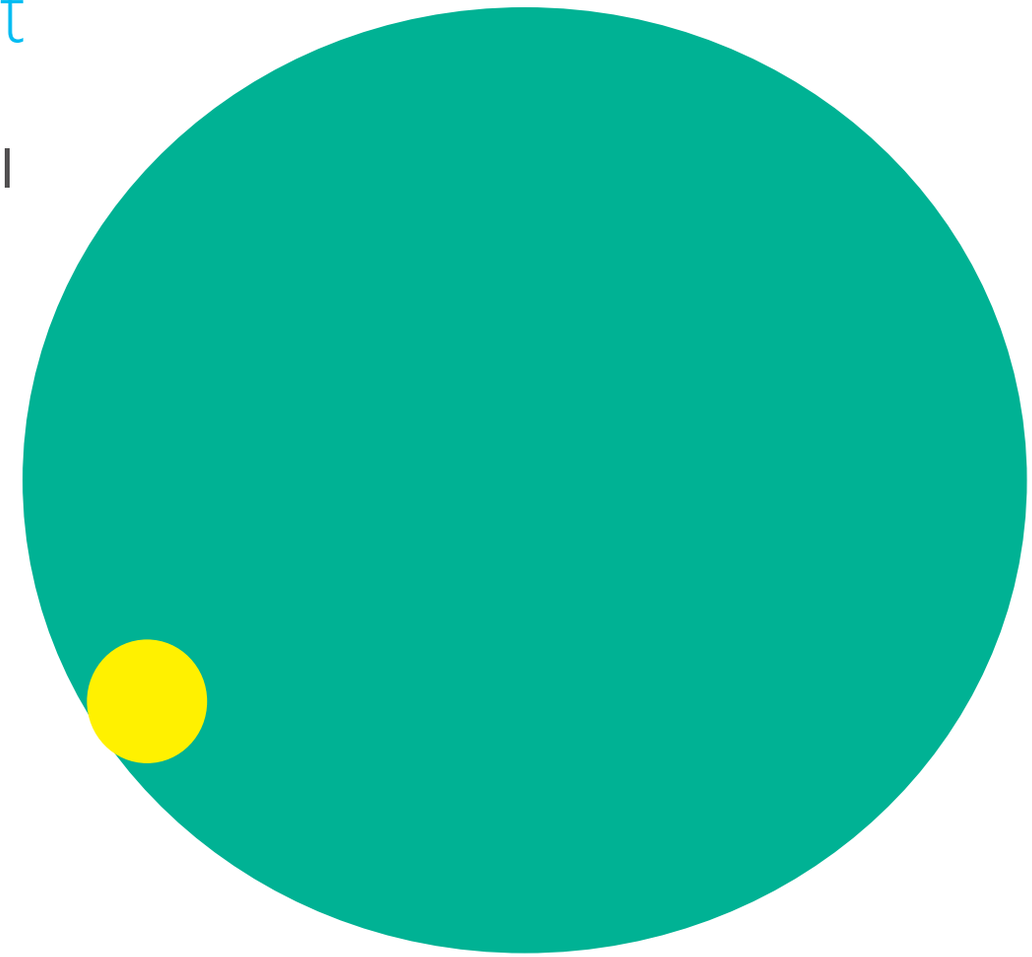
Population vs. samples

Population: every object in a specified set

Set must be well-defined.

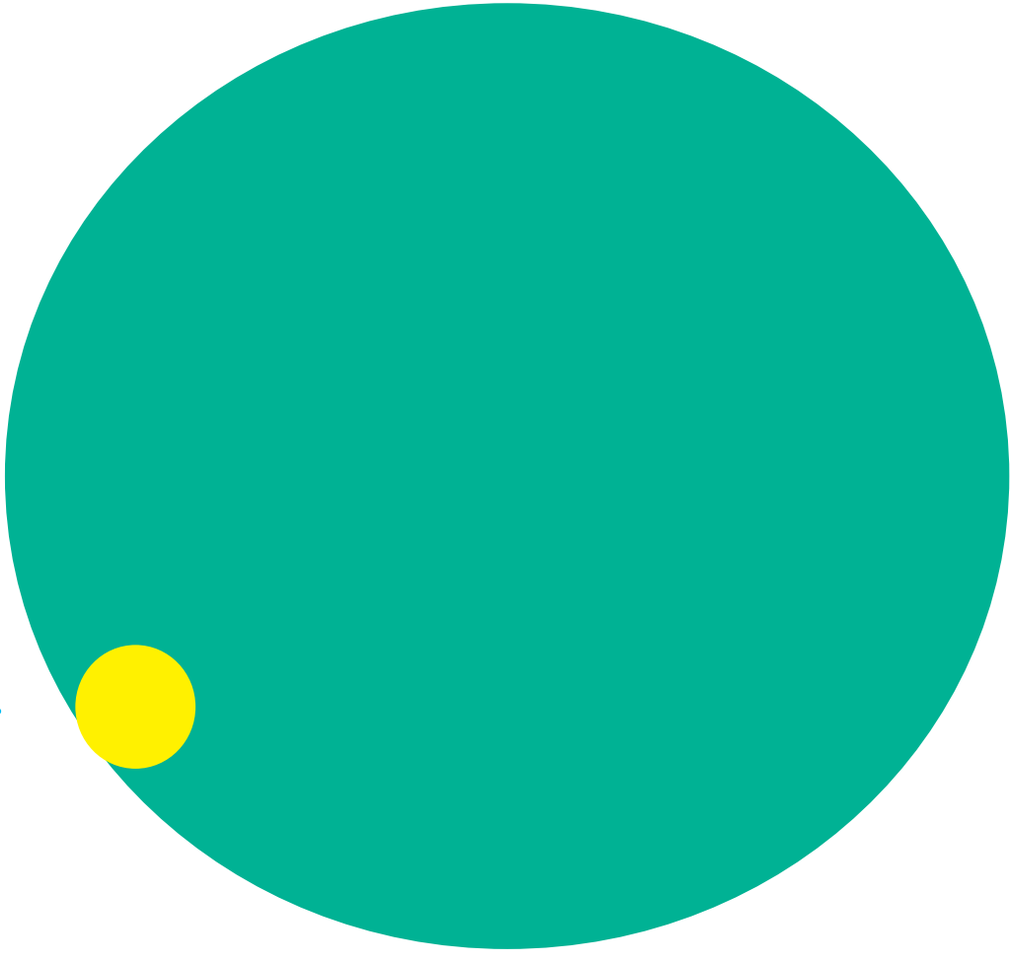
Ex. Americans, all internet users, all lab rats with lymphoma, all high school students, etc.

Sample: a subset of the population selected in *some way*



Ex. Population vs. samples

Population: heights of all US-located people age 18-65



Some samples:

1) Heights of the instructors
2) Heights of all the ppl in this room
3) Stop people on campus & measure.

How do these samples differ from the population?

Population mean vs. sample mean

Population: heights of all US-located people 18-65

Let's call the population mean of U.S. height μ (sometimes called "the truth")

At a given point in time, μ is *deterministic*. Let's assume it's 66.2 inches.

Sample: we'll draw a sample N people located in US

Let's call the sample mean of adult U.S. height $\overline{x_N}$ (pronounced x-bar)

Any given sample mean $\overline{x_N}$ is *random variable*

What is a “random variable”

Deterministic: not influenced by chance

In measurement: every time you measure, you get the same answer.

In models: for same input, you always get same output. Ex. Planetary motion

Stochastic: there is some element of *randomness*

In measurement: every time you “measure”, you *expect* a different outcome

In models: the same input leads to different output. Ex. Identical twins tend can have different medical outcomes due to differential gene expression and different enviroment

In both cases: you cannot perfectly predict what will happen

Examples: deterministic vs. stochastic

Baseball: outcome (hit or not)

Stochastic: batting avg. gives probability of a hit. Batting avg. can be modeled for specific conditions (e.g. facing pitcher X, in ballpark Y, with conditions Z).

Subway: trip time between Station X and Station Y

Stochastic: delays lead to small departures from schedule

Weight of a quarter

Technically stochastic: all quarters differ by microscopic amounts. Practically speaking it is deterministic

Length of a day

Deterministic: although days are getting slightly longer, this can be predicted perfectly

$\overline{x_N}$ as a random variable

Determine sample size, N

Determine sampling procedure

How are we going to get N observations?

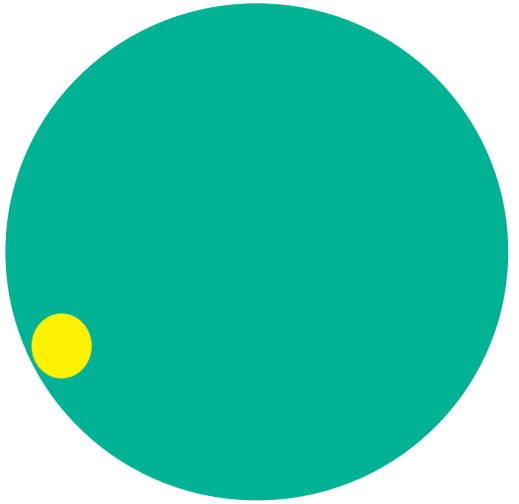
Provided N is less than total population, then each sample will tend to differ. Imagine we conduct two samples:

All *statistics* we measure will tend to differ between the two samples

Statistical theory will tell us the degree to which we should expect them to differ

$\bar{x}$ is an *estimator* for μ

Samples are used to estimate population parameters



Observe heights in the sample to make *inference* about heights in the population

Statistical properties of how the sample is drawn will determine how useful the estimator is

Properties of the sample mean

Assume $\boldsymbol{N}$ men are drawn from the population with *independent, random draws*

$$as\ N \rightarrow\ \infty$$

Law of large numbers: $\bar{x} \rightarrow \mu$

Central limit theorem: $(\bar{x} - \mu) \rightarrow Normal(0, \frac{\sigma}{\sqrt{N}})$

Where σ is the standard deviation of height in the population (note in practice σ must be estimated too)

Properties of the sample mean

Law of large numbers: $\bar{x} \rightarrow \mu$

Expected value of the estimator is equal to the population parameter we care about

LLN says: if we sample in the right way, as our data increases, we are more and more likely to estimate the "truth"

Example, we observe a baseball player at 3 times during the season:

Time 1: he has 10 "at bats" and 5 "hits". Success rate=0.500

Time 2: he has 100 at bats and 34 hits. Success rate=0.340

Time 3: he has 500 at bats and 150 hits: Success rate=0.300

LLN says we have the most confidence in the last estimate

Properties of the sample mean

Central limit theorem: $(\bar{x} - \mu) \rightarrow Normal(0, \frac{\sigma}{\sqrt{N}})$

The estimator is a *random variable* that is normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{N}}$

$\frac{\sigma}{\sqrt{N}}$ is known as the *standard error*

The standard error tells us how much we expect our sample statistics to differ from each other due to *chance alone*

Standard error vs. standard deviation

Standard deviation measures the dispersion in the *underlying population*

$$\sqrt{E(X-\mu)^2} = \sigma$$

Are far are data points from the population mean on average?

Ex. standard deviation of income tells us how different we expect a randomly selected to person's income to be from the overall mean.

Coin flipping: s.e. vs. s.d.

Standard deviation of a coin flip?

Outcome space: heads=1, tails=0

$\mu = 0.50 \text{ (fair coin)}$

$\sigma = \sqrt{.5 * (1 - .5)^2 + .5 * (0 - .5)^2}$

$= \sqrt{.5 * .25 + .5 * .25} = 0.5$

Each outcome is 0.50 from the mean!

- General formula for binary outcomes, $\sigma = \sqrt{p * (1 - p)}$

Standard error depends on N

We estimate $\bar{x}$ and the standard deviation *of this estimate* is given by $\frac{.50}{\sqrt{N}}$

Standard error vs. standard deviation

Standard error measures the dispersion in *the estimator* (ex. sample mean) for a given sample size

The central limit theorem tells us how the variance in the population links up with the variance in our estimator: $\frac{\sigma}{\sqrt{N}}$

Ex. fair coin flip. Heads=1, Tails=0.

- μ=0.5 $\sigma = 0.5$
- s.e. = $0.5/\sqrt{N}$
  - N=4 → 0.25
  - N=25 → 0.1
  - N=100 → 0.05

s.e. vs. s.d.

Central limit theorem says the distribution *of the estimator* will be Guassian for *any* population distribution with finite mean and variance

The distribution of the sample mean is bell-shaped, “no matter what” the distribution of the underlying variable

As N increases, the distribution will “collapse” to the true population mean

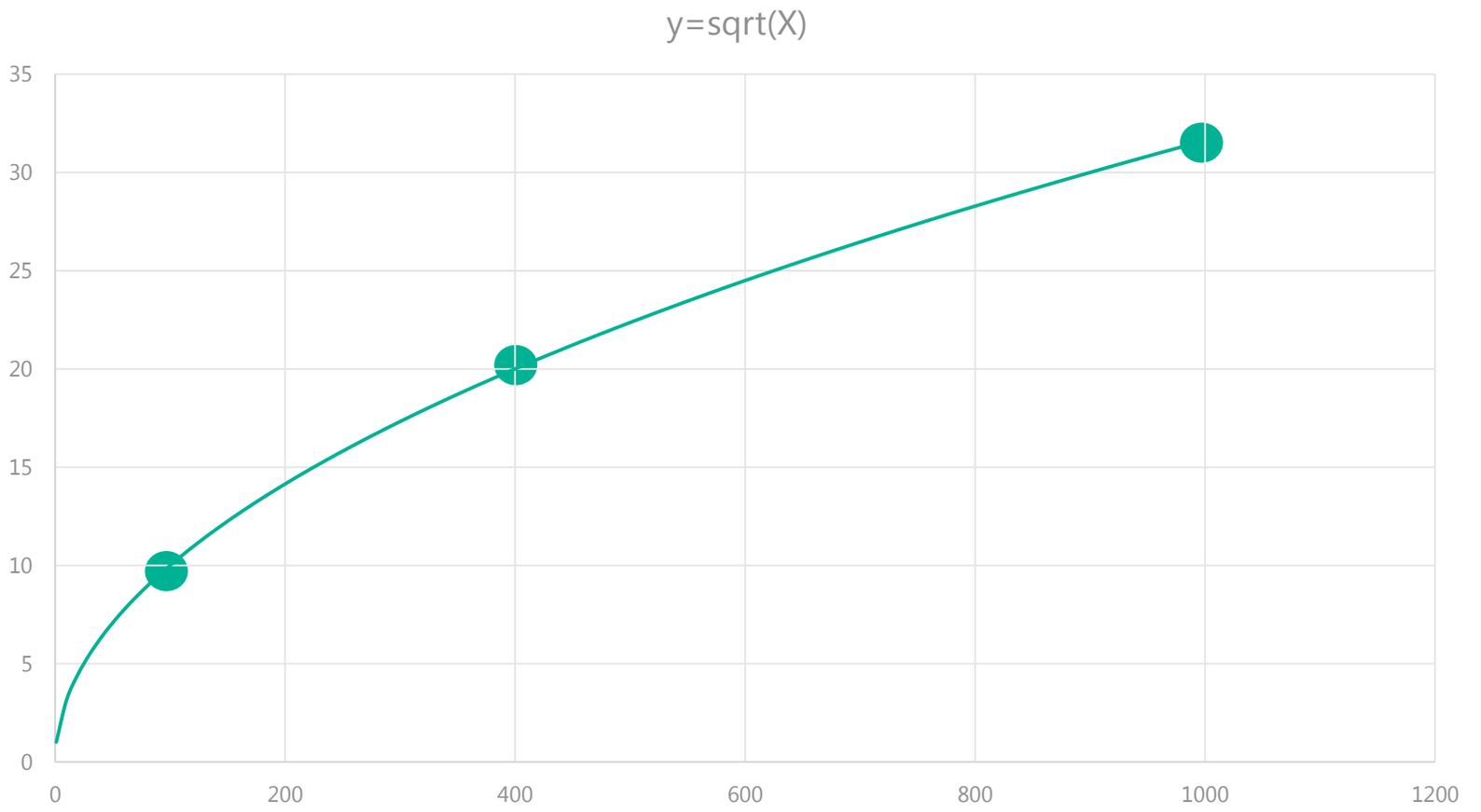
Statistical uncertainty and $\sqrt{N}$

Ex. fair coin flip. Heads=1, Tails=0.

- μ=0.5 $\sigma = 0.5$
- s.e. = $0.5/\sqrt{N}$

Ex. Points

$$\sqrt{100} = 10$$
$$\sqrt{400} = 20$$
$$\sqrt{1000} = 33$$



Standard error and sample size

Ex. fair coin flip. Heads=1, Tails=0.

- μ=0.5 $\sigma = 0.5$
- s.e. = $0.5/\sqrt{N}$

To reduce my standard error factor 2, I need to increase sample size by 2^2

Ex.
s.e.(N=40)=0.04
s.e.(N=160)=0.02
s.e.(N=640)=0.01

Height example cont.

Assume standard deviation is 2.5 inches

N=100, standard error = $\frac{2.5}{\sqrt{100}}$= 0.25

95% Confidence interval $\pm$ 1.96 $s.e. \rightarrow \bar{x} \pm$ 0.49 inches.

Says 95% of the time the true population mean will lie in this interval (if we have sampled correctly)

N=1,000, standard error = $\frac{2.5}{\sqrt{1000}}$=$\frac{2.5}{31.6}$= 0.08

Sample size increased by factor 10, s.e. only drops by factor $\sqrt{10}$

With a 1,000 *randomly selected* people we estimate within 0.2 inches of truth

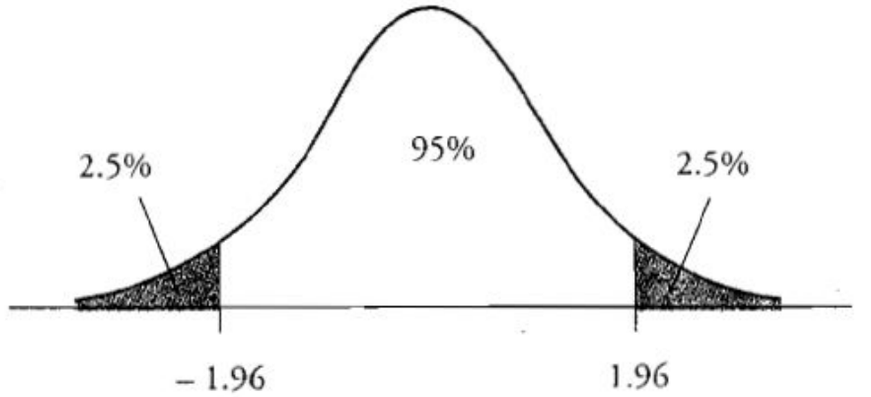
Constructing confidence intervals

1) Estimate sigma: $s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$

Intuition, use average differences from the sample mean.

2) Compute standard error= $\frac{s}{\sqrt{N}}$

3) +/- 1.96 captures 95% of the Normal distribution



What does this mean?

1) Your estimate is your best guess of the truth

2) 95% confidence interval: 95% of the time the truth is in this range

3) If wide, indicates lots of uncertainty in best guess

4) In practice, you only have 1 sample. Statistical theory *quantifies your uncertainty in this estimate*

Visualizing the CLT

N=4

Fair coin

http://blog.vctr.me/posts/central-limit-theorem.html

What does standard error capture?

Standard error captures how much our estimator will move around due the chance involved in sampling

This is known as **sampling variation**, it captures the degree to which small samples are different from the population.

It does not capture *model uncertainty or bias in the sampling procedure*

What if I have a non-random sample?

The “margin for error” you see on polls usually assumes they have a representative sample, and thus only reports sampling variation

Sampling and estimation

Sampling is the way we generate “observations”

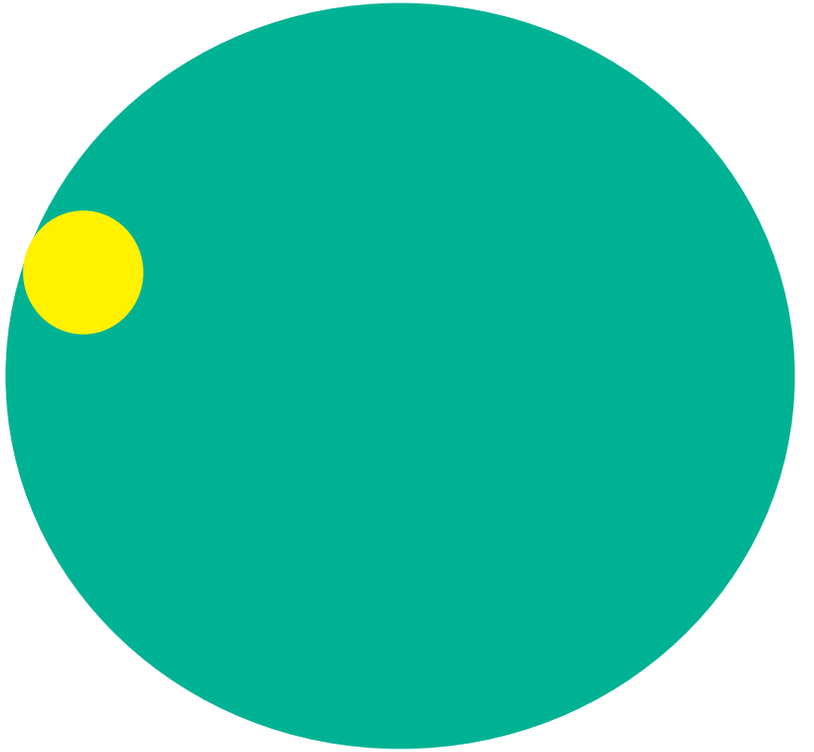
I.e. it is the way we construct datasets

Any dataset that is not a “complete census” is a sample (and even a complete census only represents one point in time, so can be view as a sample too)

Returning to our example

Some samples:

1) Heights of the course instructors
2) Heights of all the people in the room
3) All MSFT employees
4) MSFT employees that will stop and talk to us



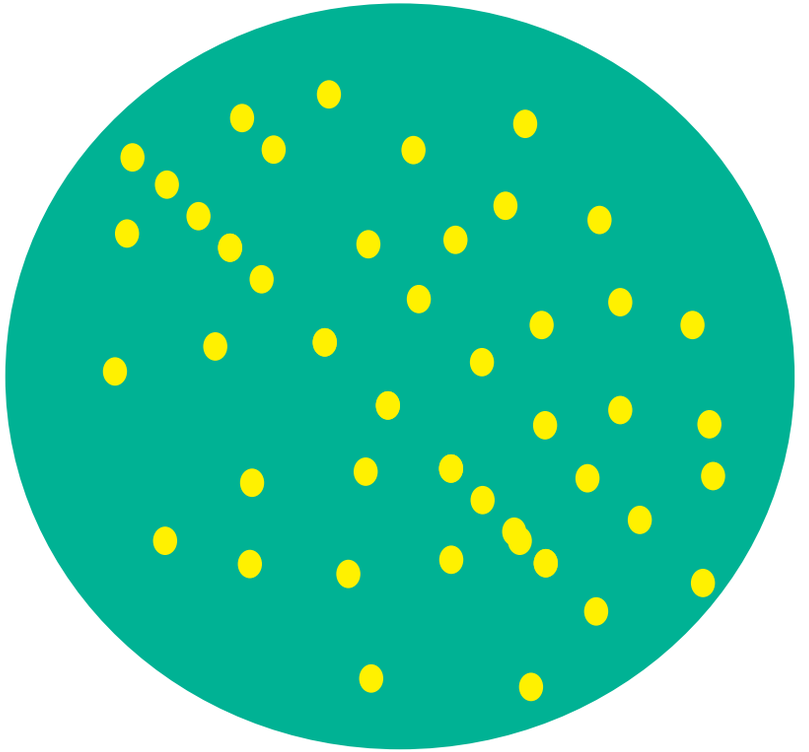
The samples are all biased, (3) will have the lowest s.e.

Random sampling for unbiasedness

s.e.'s are not very meaningful if the estimator is *biased*

Random sampling ensures that the sample mean will equal the population mean in expectation

In big data applications, bias is much more important than sampling variation!



Example: polls

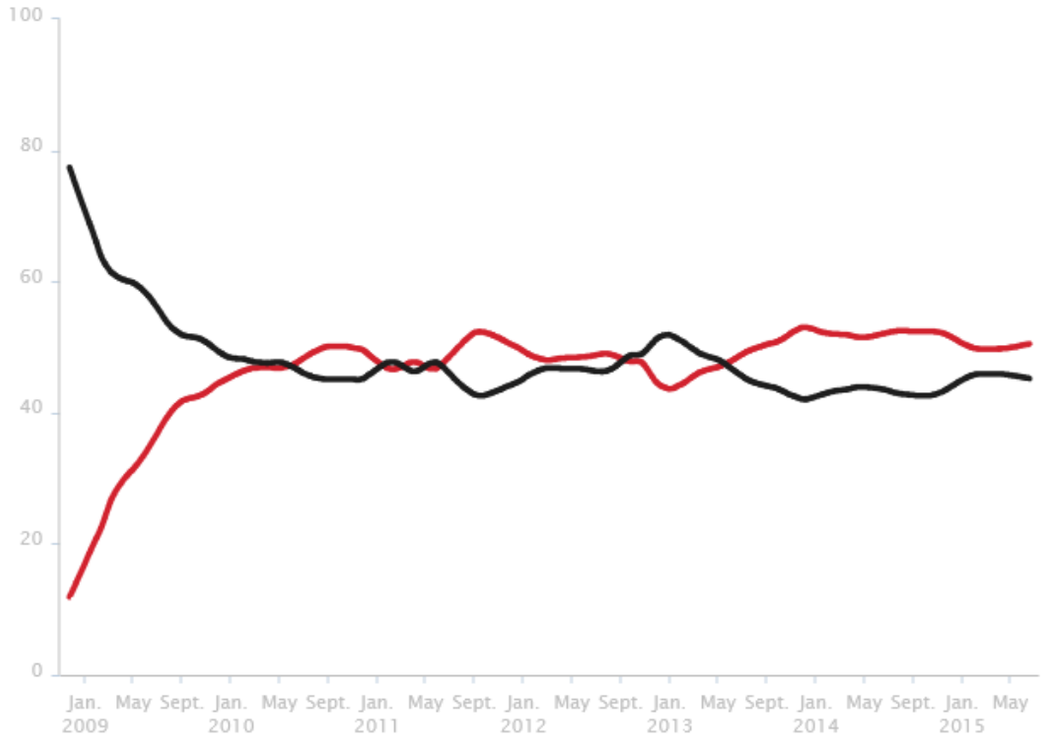
POLL CHART

Obama Job Approval

Currently tracking 3,004 polls from 96 pollsters Updated 3 days ago

HuffPost Model Create Your Own

This chart combines the latest opinion polls and is updated whenever a new poll is released.



Pollster Trend

Disapprove	50.5%
Approve	45.2%
Undecided	

INDIVIDUAL POLL RESULT

POLL AVERAGE

Why do these numbers differ?

Latest Polls

POLLSTER	DATES	POP.	APPROVE	DISAPPROVE	UNDECIDED	MARGIN
Gallup	6/16 - 6/18	1,500 A	44	50	-	Disapprove +6
Rasmussen	6/16 - 6/18	1,500 LV	45	54	-	Disapprove +9
Gallup	6/13 - 6/15	1,500 A	46	48	-	Disapprove +2
YouGov/Economist	6/13 - 6/15	1,000 A	46	52	2	Disapprove +6
Rasmussen	6/11 - 6/15	1,500 LV	47	52	-	Disapprove +5
PPP (D)	6/11 - 6/14	1,129 LV	45	50	5	Disapprove +5
Gallup	6/10 - 6/12	1,500 A	45	49	-	Disapprove +4
Rasmussen	6/8 - 6/10	1,500 LV	47	52	-	Disapprove +5
Ipsos/Reuters	6/6 - 6/10	2,949 A	42	52	6	Disapprove +10
Gallup	6/7 - 6/9	1,500 A	45	49	-	Disapprove +4

SHOW MORE ▼

RSS | CSV | API

POLL UPDATE

Obama Job Approval - Disapprove 50%, Approve 44% (Gallup 6/16-6/18)

Population	1,500 Adults
Margin of Error	±3.0 percentage points
Polling Method	Live Phone
Source	Gallup

This poll asked respondents 1 question tracked by HuffPost Pollster.

POLL UPDATE

Obama Job Approval - Disapprove 52%, Approve 47% (Rasmussen Reports 6/11-6/15)

Population	1,500 Likely Voters
Margin of Error	±3.0 percentage points
Polling Method	Automated Phone
Source	Rasmussen

This poll asked respondents 1 question tracked by HuffPost Pollster.

POLL UPDATE

Obama Job Approval - Disapprove 52%, Approve 42% (Ipsos/Reuters (Web) 6/6-6/10)

Population	2,949 Adults
Margin of Error	±2.1 percentage points
Polling Method	Internet
Source	Ipsos/Reuters

This poll asked respondents 3 questions tracked by HuffPost Pollster.

Different target populations

Different methods → different biases in sample selection

Different times → different "populations"

Different random samples→ sampling variation

Two key properties of an estimator

Bias: if $E[\hat{\beta}] = \beta$ then the estimator is said to be unbiased.
"with a very large sample we expect to uncover the truth"

Variance: the degree to which the estimator jumps around

In more complicated settings, different estimators of the same underlying quantity can have different s.e.'s

Returning to the height example

A fully random sample of 25 U.S. adults will be unbiased but have relatively high variance

A large sample of all MSFT employees will have a very low variance, but is likely biased

Later in the course, we will see how methods trade-off bias vs. variance

What determines properties of an estimator?

The estimation method or algorithm

The sample mean is an estimator that is unbiased if the sampling is done correctly

Other more complicated estimators have different properties, we will see this in later lectures

How the sample was drawn

Even “good estimators” are lousy with poor samples.

Hypothesis testing

H_o: $null\ hypothesis\ \longrightarrow population\ parameter = \mu_0$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{N}}$$

s =sample standard deviation

$\bar{x}$=sample mean or anything that "behaves like a sample mean"

Hypothesis testing: p values

p-value: probability we observe an $\bar{x} - \mu_0$ difference due to chance driven by sampling, assuming null hypothesis is true

If p-value is low, it is unlikely to observe the estimator value due to sampling chance, thus we are inclined to reject the null hypothesis.

In other words, the evidence we observe would be very unlikely if the null hypothesis is true

Hypothesis testing

H_o: *null hypothesis* $\longrightarrow$ *population parameter* $= \mu_0$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{N}}$$