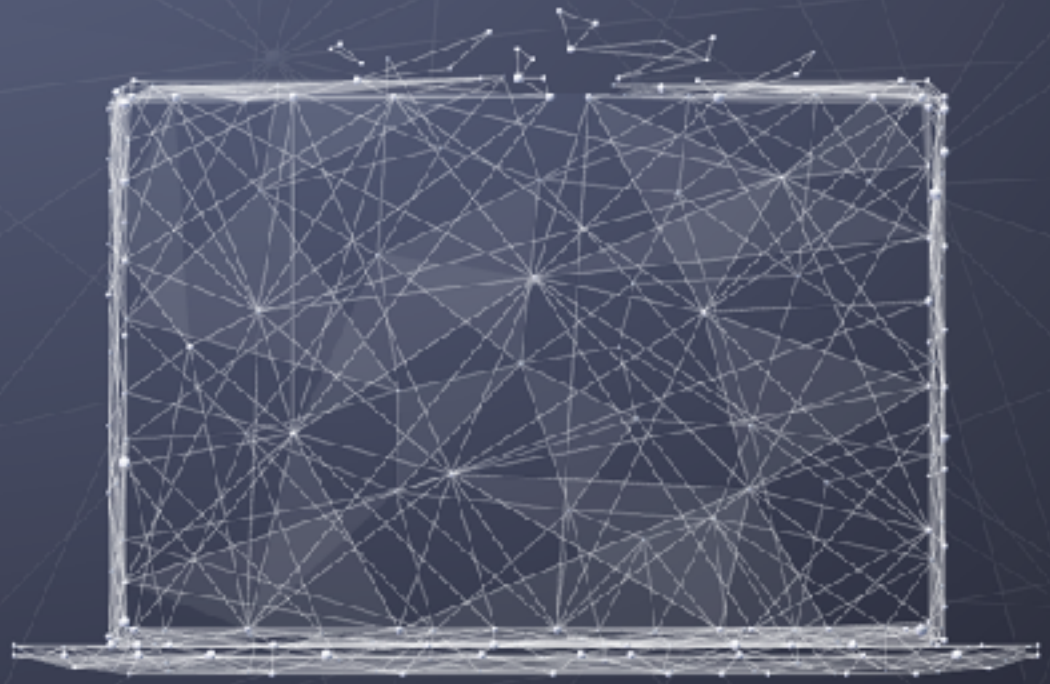
Data Science
Foundations of Decision Making

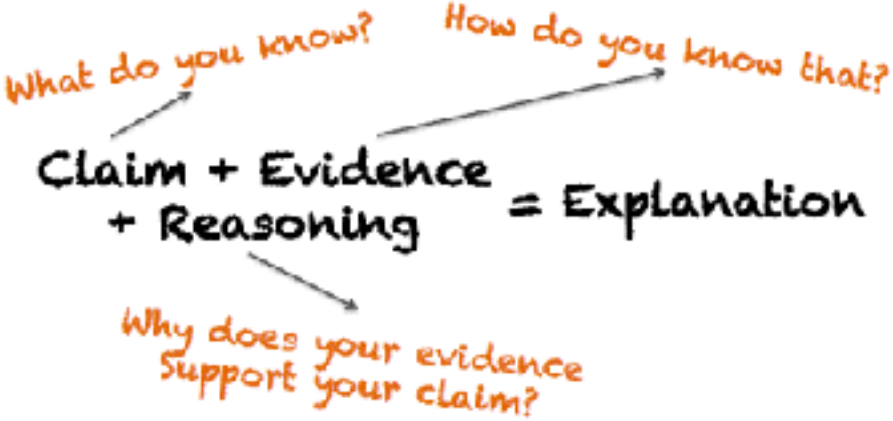
Claims, evidence, and data



PURDUE UNIVERSITY | College of Science

Explaining patterns in data

- A good explanation consists of:
  - Claim that answers a question/ problem
  - Evidence from data
  - Reasoning that connects evidence to claim, shows why data counts as evidence



Example claim

More than 80% of dentists recommend Colgate

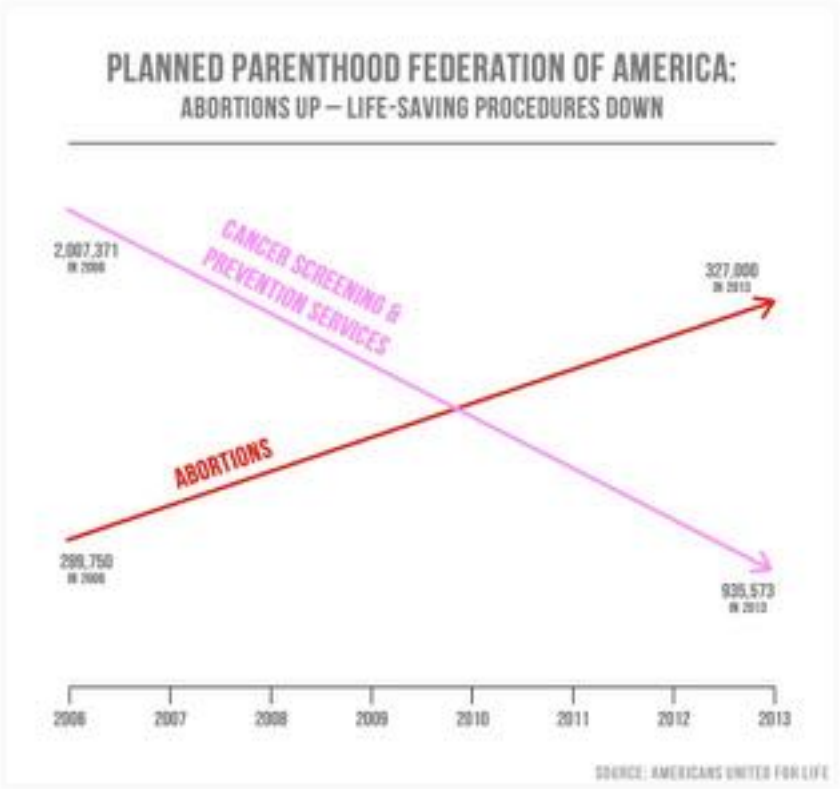
Example claim

Based on survey that let dentists select multiple brands of toothpaste.

In 2007, Colgate was ordered to abandon their claim.

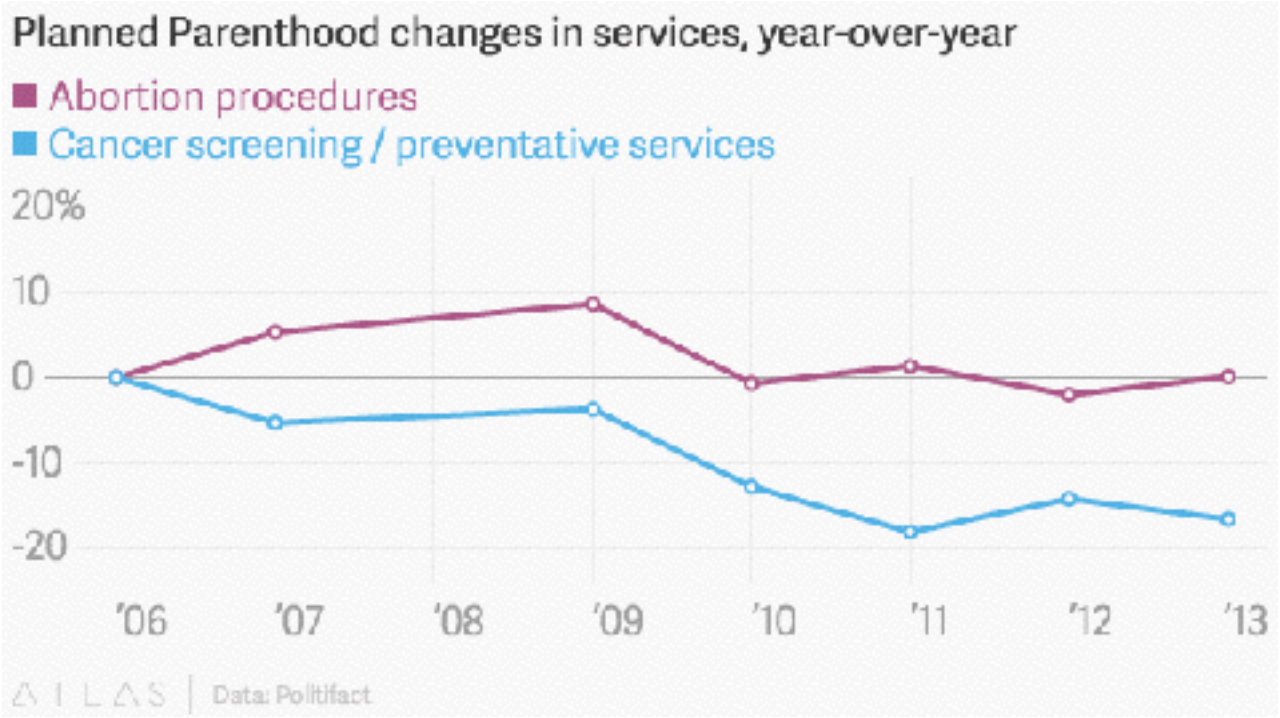
More than 80% of dentists recommend Colgate

Example claim



Planned Parenthood misappropriated its funds.
(e.g., abortions increased while screening for cancer decreased)

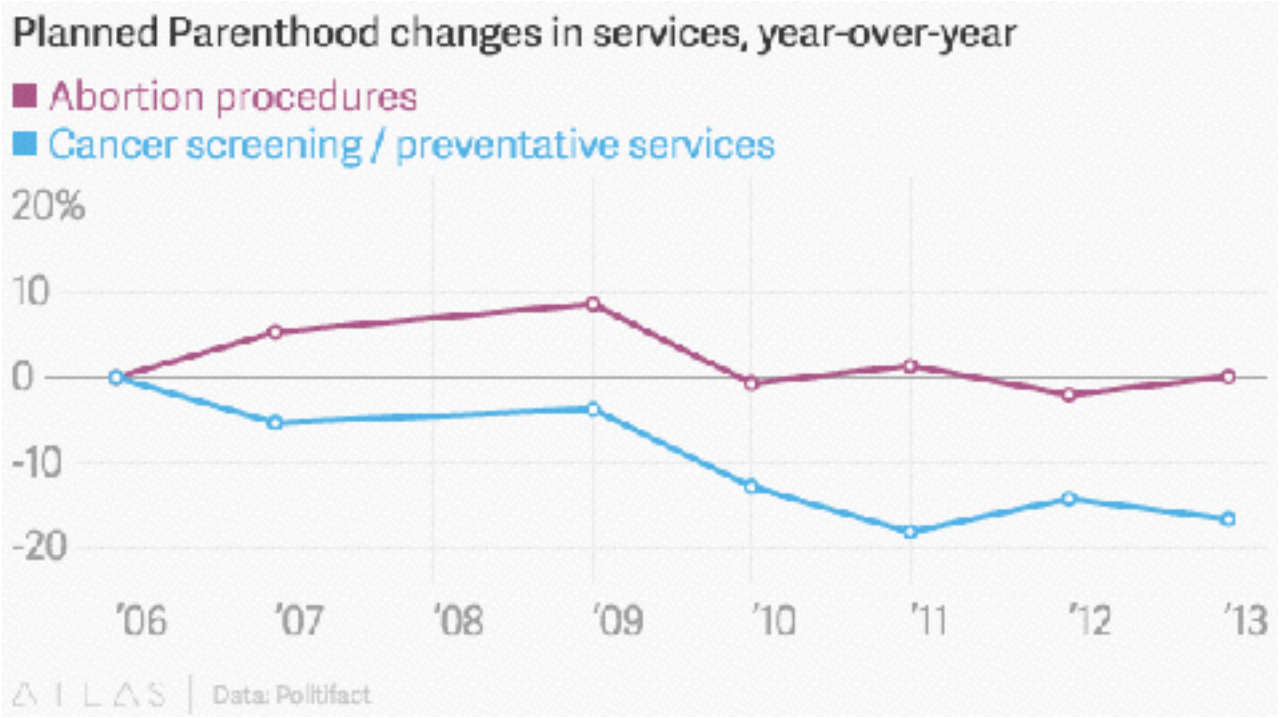
Example claim



Planned Parenthood misappropriated its funds.
(e.g., abortions increased while screening for cancer decreased)

Example claim

Given another view of the data, what part of the claim is true?



Planned Parenthood misappropriated its funds.
(e.g., abortions increased while screening for cancer decreased)

I had prostate cancer… My chance of surviving prostate cancer… in the United States? 82%

My chance of surviving prostate cancer in England? Only 44% under socialized medicine

Rudy Giuliani, 2007

Claim and evidence from data

Claim and evidence from data

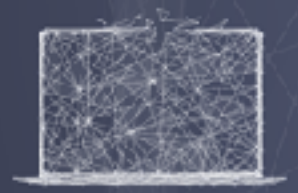
- Claim: Men are nearly twice as likely to survive in the United States (82%) as in England (44%), based on a five-year survival rate after detection.

Claim and evidence from data

- Claim: Men are nearly twice as likely to survive in the United States (82%) as in England (44%), based on a five-year survival rate after detection.
- Data: In a 2000 study, 49 British men per 100,000 were diagnosed with prostate cancer, of which 28 died within five years. Thus 21/49 survived, about 43%.

Claim and evidence from data

- Claim: Men are nearly twice as likely to survive in the United States (82%) as in England (44%), based on a five-year survival rate after detection.
- Data: In a 2000 study, 49 British men per 100,000 were diagnosed with prostate cancer, of which 28 died within five years. Thus 21/49 survived, about 43%.
- But screening for prostate cancer is different in England than in the United States. The USA relies heavily on the PSA test, which can sometimes detect cancer earlier. But the test is not widely used in England. Thus, the detection, and the five-year survival rate, spans a different time in the lives of patients in the two countries.



Does data support claim?

Does data support claim?

- Imagine a group of prostate cancer patients currently diagnosed at age 67, all of whom die at age 70. Each survived only three years, so the five-year survival of this group is 0 percent.

Does data support claim?

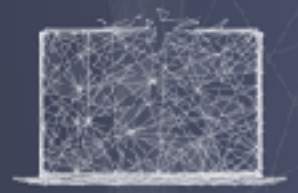
- Imagine a group of prostate cancer patients currently diagnosed at age 67, all of whom die at age 70. Each survived only three years, so the five-year survival of this group is 0 percent.
- Now imagine that the same group is diagnosed with prostate cancer by PSA tests earlier, at age 60, but they all still die at age 70.

Does data support claim?

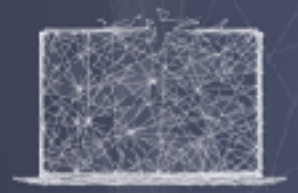
- Imagine a group of prostate cancer patients currently diagnosed at age 67, all of whom die at age 70. Each survived only three years, so the five-year survival of this group is 0 percent.
- Now imagine that the same group is diagnosed with prostate cancer by PSA tests earlier, at age 60, but they all still die at age 70.
- If the patients in the second group lived to the age of 65, their five-year survival rate would be 100 percent, although they all died by age 70. Even though the survival rate has changed dramatically, nothing has changed about the time of death.

Does data support claim?

- Imagine a group of prostate cancer patients currently diagnosed at age 67, all of whom die at age 70. Each survived only three years, so the five-year survival of this group is 0 percent.
- Now imagine that the same group is diagnosed with prostate cancer by PSA tests earlier, at age 60, but they all still die at age 70.
- If the patients in the second group lived to the age of 65, their five-year survival rate would be 100 percent, although they all died by age 70. Even though the survival rate has changed dramatically, nothing has changed about the time of death.
- Are American men half as likely to die from prostate cancer as British men are? No. The risk is about the same: About 26 prostate cancer deaths per 100,000 American men versus 27 per 100,000 in Britain.



How to evaluate whether data supports a claim?

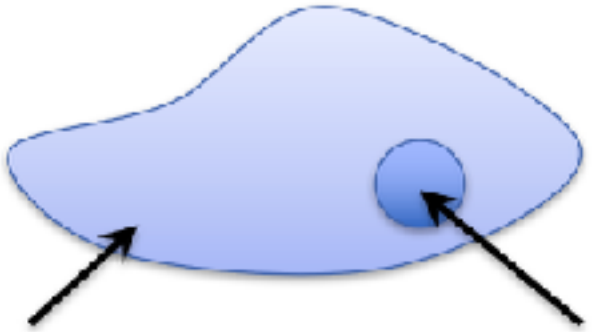


Asking scientific questions

- Suppose you work for a company that is considering a redesign of their website; does their new design (design B) offer any statistical advantage to their current design (design A)?
- In linear regression, does a certain variable impact the response? (E.g. does energy consumption depend on whether or not a day is a weekday or weekend?)
- In both settings, we are concerned with making actual statements about the nature of the world

Sample statistics

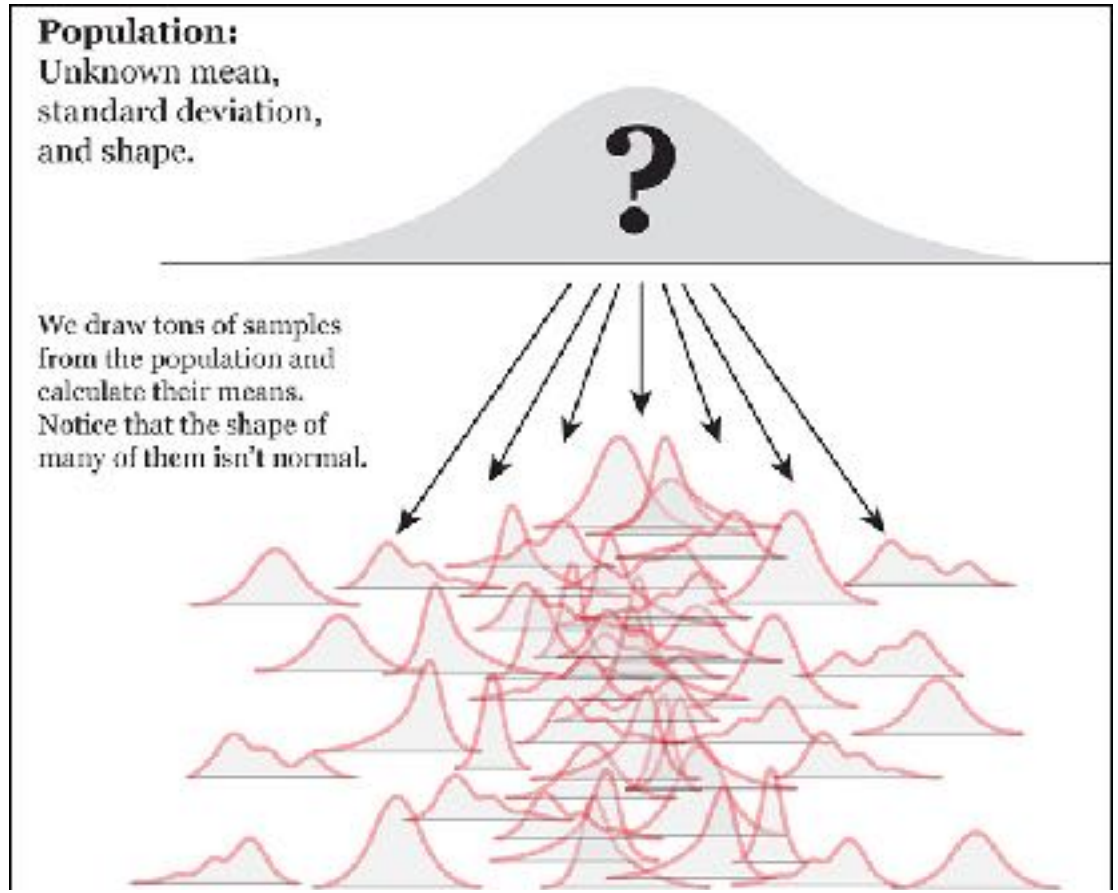
- To be a more consistent with standard statistics notation, we need to differentiate between a population and a sample

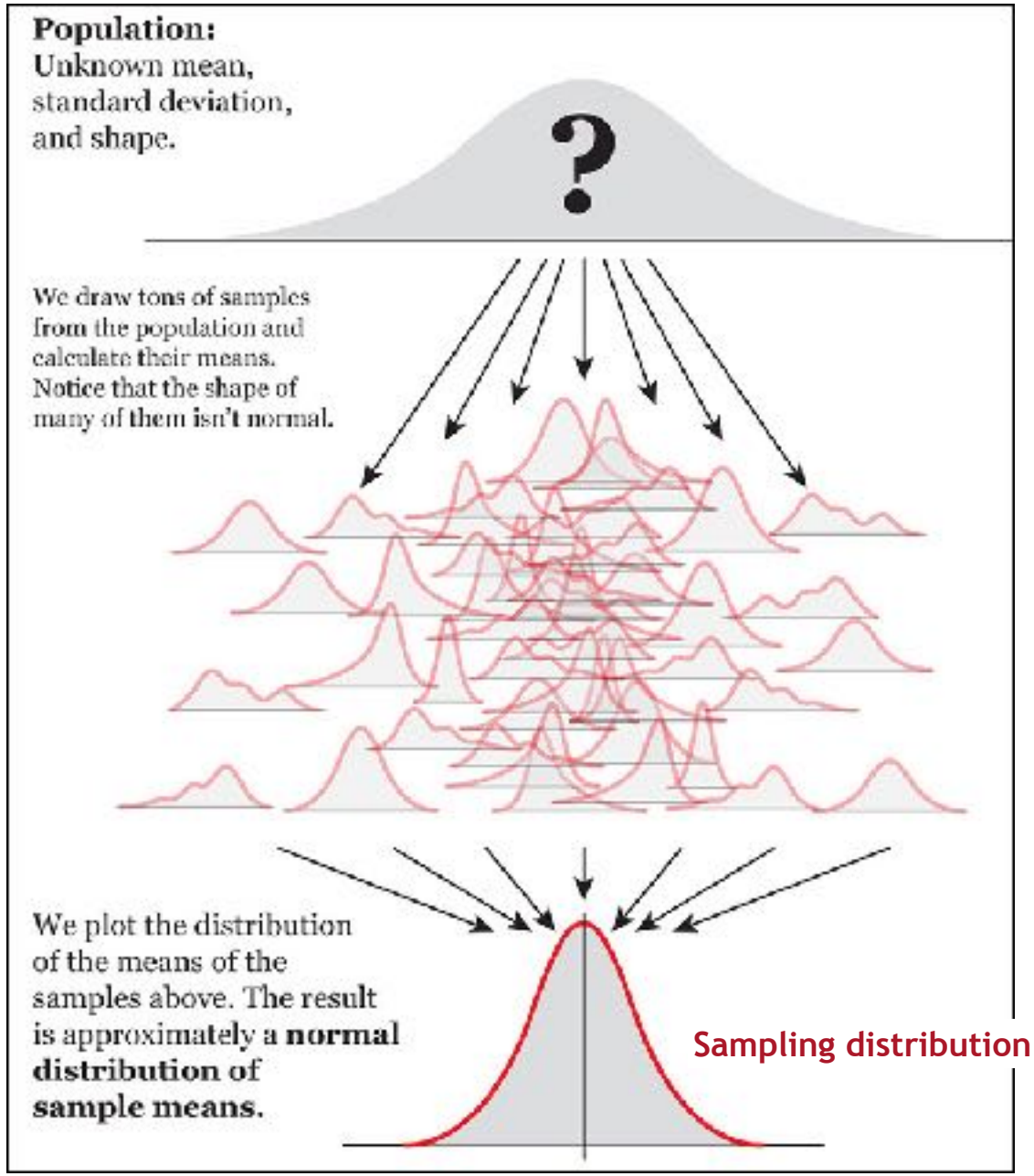


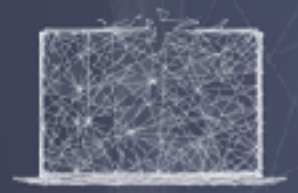
Every sample has some inherent error, called the standard error (standard deviation of sampling distribution)

	Population	Sample
Mean	$\mu = \mathbf{E}[X]$	$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$
Variance	$\sigma = \mathbf{E}[(X-\mu)^2]$	$s^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x^{(i)}-\bar{x})^2$


Population:
Unknown mean,
standard deviation,
and shape.
?

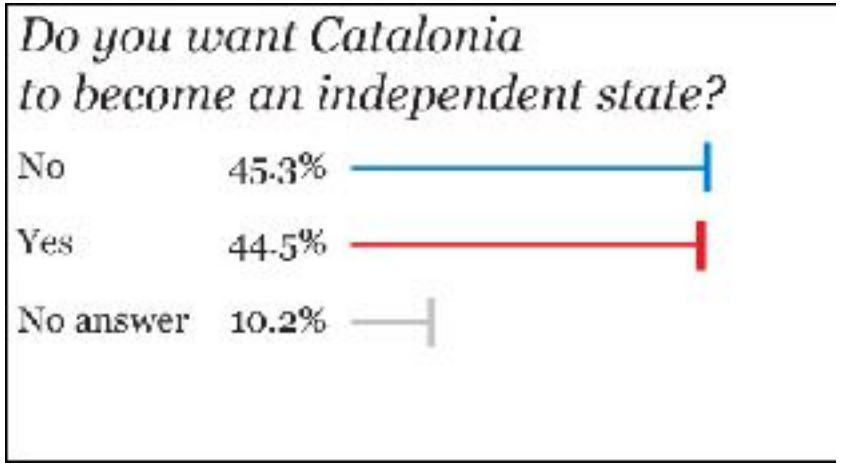

Population:
Unknown mean,
standard deviation,
and shape.
?
We draw tons of samples
from the population and
calculate their means.
Notice that the shape of
many of them isn't normal.


Population:
Unknown mean,
standard deviation,
and shape.
?
We draw tons of samples
from the population and
calculate their means.
Notice that the shape of
many of them isn't normal.
We plot the distribution
of the means of the
samples above. The result
is approximately a normal
distribution of
sample means.
Sampling distribution



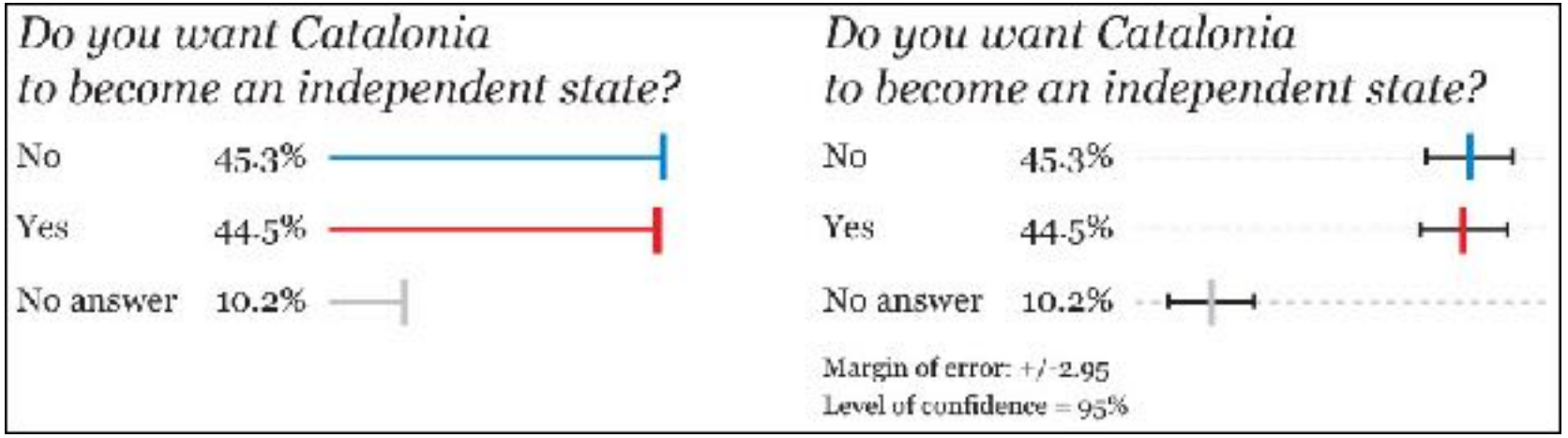
Expressing uncertainty due to sampling

Expressing uncertainty: Margin of error



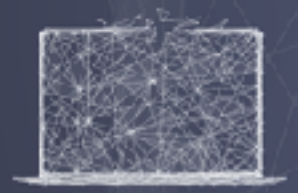
No uncertainty quantification

Expressing uncertainty: Margin of error



No uncertainty quantification

With uncertainty quantification



Measures of uncertainty

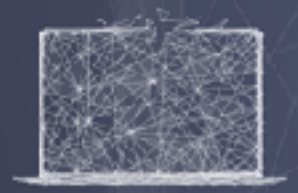
Measures of uncertainty

- Sampling distribution: probability distribution of a statistic (e.g., mean) when calculated from a random sample of size n

Measures of uncertainty

- Sampling distribution: probability distribution of a statistic (e.g., mean) when calculated from a random sample of size n
- Standard error: standard deviation of the sampling distribution

$$SE = \frac{\sigma}{\sqrt{n}}$$



Measures of uncertainty

- Sampling distribution: probability distribution of a statistic (e.g., mean) when calculated from a random sample of size n
- Standard error: standard deviation of the sampling distribution

$$SE = \frac{\sigma}{\sqrt{n}}$$

- Margin of error: confidence interval of sampling distribution

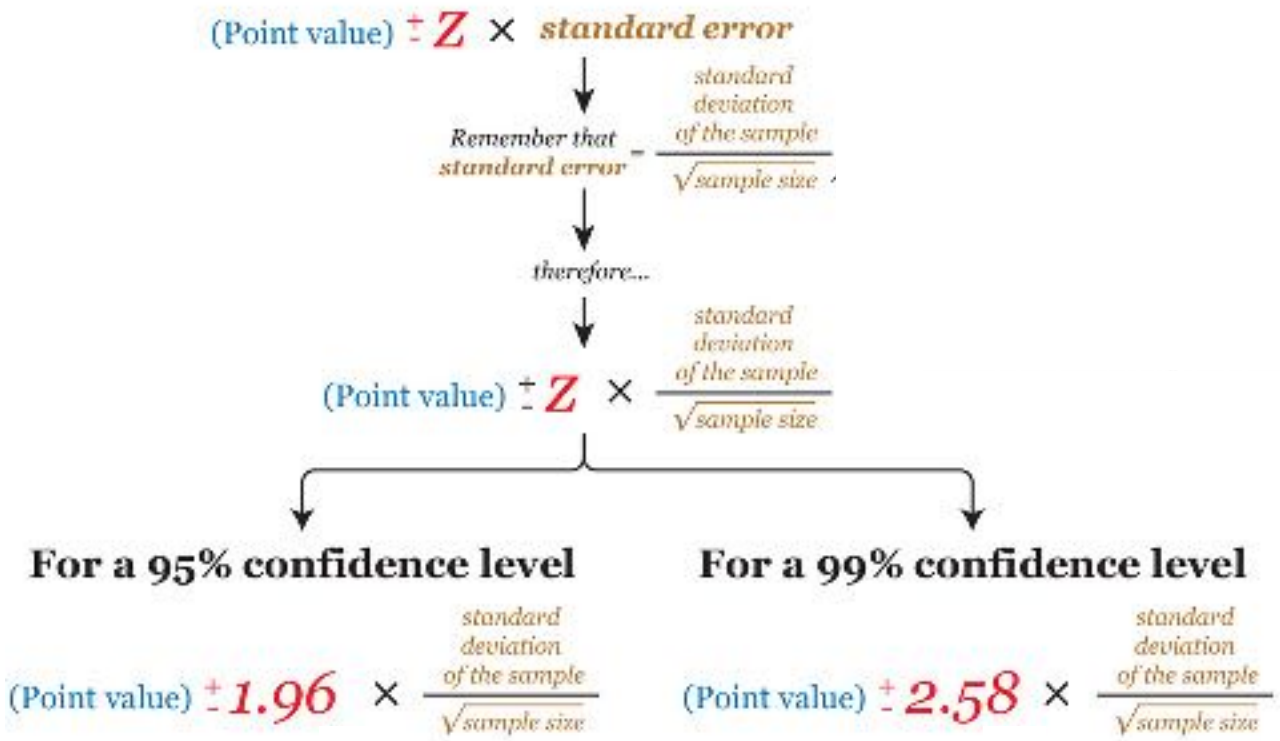
Measures of uncertainty

- Sampling distribution: probability distribution of a statistic (e.g., mean) when calculated from a random sample of size n
- Standard error: standard deviation of the sampling distribution

$$SE = \frac{\sigma}{\sqrt{n}}$$

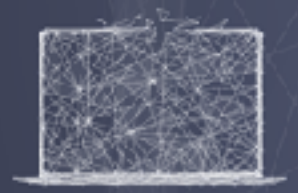
- Margin of error: confidence interval of sampling distribution
- Confidence interval: expression of uncertainty wrt to estimated statistic
  E.g., for 95% CI: Given 100 samples of the same size, 95 of the estimated CIs would contain the true value of the statistic

Calculating the confidence interval of a mean



In a standard normal distribution, 95% of the scores lie between –1.96 and 1.96 standard deviations from the mean. 99% of the scores lie between –2.58 and 2.58 standard deviations from the mean.

PURDUE UNIVERSITY | College of Science



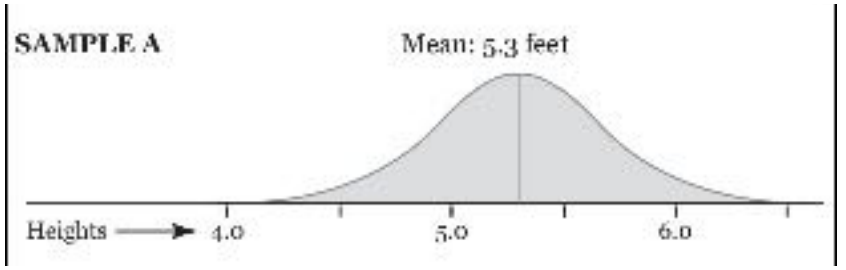
Example confidence interval

Example confidence interval

- Suppose you wish to estimate the heights of 12-year-old girls in your town.

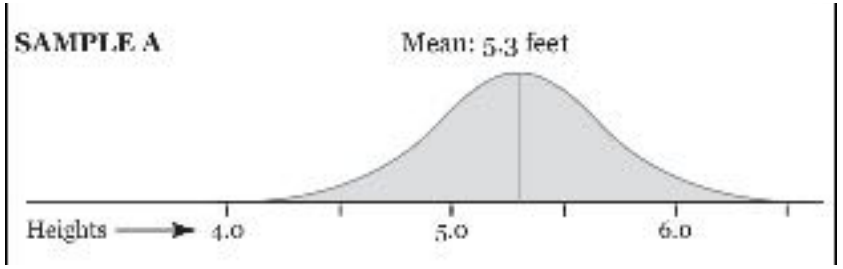
Example confidence interval

- Suppose you wish to estimate the heights of 12-year-old girls in your town.
- You randomly choose a sample of 40 girls, measure them and get a distribution with mean of 5.3 feet and a standard deviation of 0.5



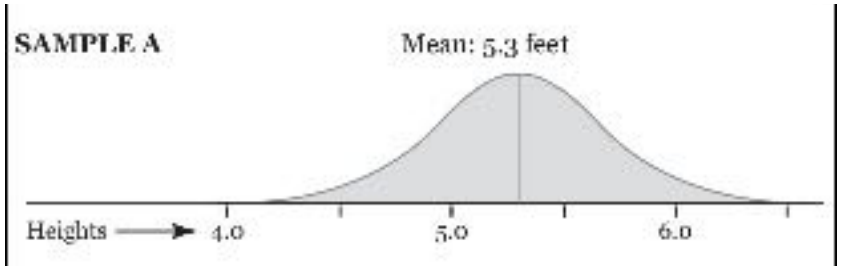
Example confidence interval

- Suppose you wish to estimate the heights of 12-year-old girls in your town.
- You randomly choose a sample of 40 girls, measure them and get a distribution with mean of 5.3 feet and a standard deviation of 0.5
- Therefore: StdErr = 0.5 / √40 = 0.5/6.32 = 0.08



Example confidence interval

- Suppose you wish to estimate the heights of 12-year-old girls in your town.
- You randomly choose a sample of 40 girls, measure them and get a distribution with mean of 5.3 feet and a standard deviation of 0.5
- Therefore: StdErr = 0.5 / √40 = 0.5/6.32 = 0.08
- Thus the confidence intervals are:



For a 95% confidence level

$$5.3 \text{ feet} \pm 1.96 \times \frac{0.5 \text{ feet}}{\sqrt{40 \text{ girls}}}$$

$$5.3 \text{ feet} \pm 0.16$$

For a 99% confidence level

$$5.3 \text{ feet} \pm 2.58 \times \frac{0.5 \text{ feet}}{\sqrt{40 \text{ girls}}}$$

$$5.3 \text{ feet} \pm 0.21$$

Disclosing uncertainty in plots

- Confidence intervals are a means to display the uncertainty in the data, this is crucial information to include in visualizations
- Some types of visualizations naturally encode the uncertainty by summarizing the distribution
- Note: there is a difference between the distribution of the statistic (e.g., girls heights) and the sampling distribution (e.g., average height)