Confidence Intervals & Designing Experiments Using Central Limit Theorem

LECTURE 23

CSCI 3022

Maribeth Oscamou

Content credit: Acknowledgments

Course Logistics: 10th and 11th Weeks At A Glance

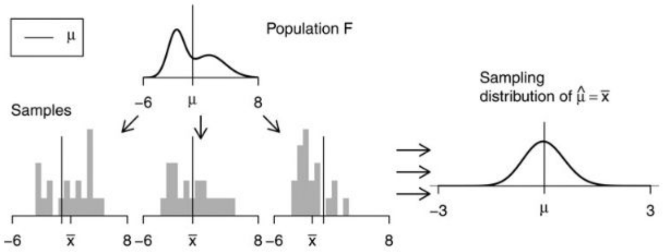
Mon 3/18	Tues 3/19	Wed 3/20	Thurs 3/21	Fri 3/22
Attend class via Zoom: Lesson 23 Confidence Intervals	TA NB Discussion 5pm-6pm via Zoom	Attend class via Zoom: Lesson 24	~~HW 9 Due~~	Class IN PERSON Quiz 6: Scope: L17-L21, HW 7, HW 8, TA discussion nb 8 and 9 HW 9 Due
		SPRING BREAK!		
Mon 4/1	Tues 4/2	Wed 4/3	Thurs 4/4	Fri 4/5
Attend & participate in class	TA NB Discussion 5pm-6pm via Zoom	Exam 2 In-Class Review Day		**Exam 2: SCOPE: Lessons 13-23 (including HW 6-9, Quiz 5, 6; TA Discussion NB 7-11)**

Today's Roadmap

- Using Central Limit Theorem to Calculate Confidence Intervals
  - Population Means
  - Population Proportions
- Using Confidence Intervals to Design Experiments

Recap: Confidence Intervals: Ideal World vs Bootstrap World

- We want to understand **variability** of our **estimate.**
  - **Need sampling distribution of sample statistic to do this.**
- Given the population, we could simulate:



Reality: We don't know the population distribution. All we have is a random sample.

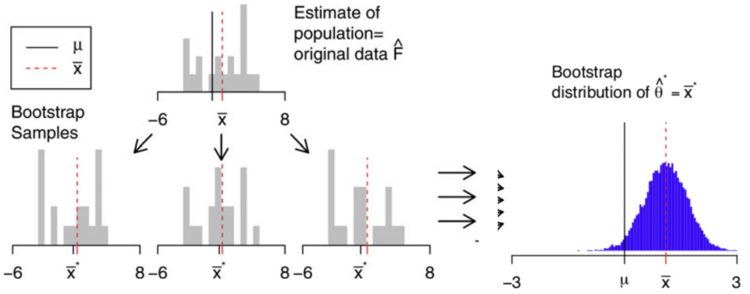
Bootstrap World:

Intuition: a random sample resembles the population, so a random resample resembles a random sample.

Method 1: Bootstrapping

- Treat our random sample as a "population", and resample from it **with replacement** computing the statistic of interest for each resample
- Create distribution of bootstrapped statistics
- Use the middle X% of this distribution to calculate the X% Confidence Interval for the population parameter

Note: The ***bootstrapped distribution is NOT centered at*** the actual population parameter.



Method 2: Use Central Limit Theorem (if it applies)

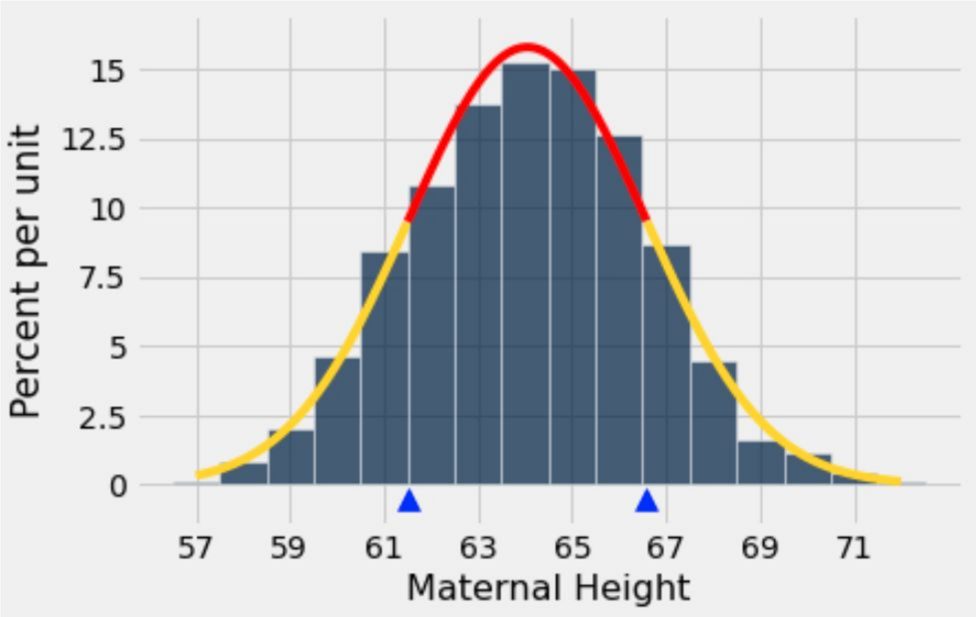
Today's Roadmap

- **Using Central Limit Theorem to Calculate Confidence Intervals**
  - Population Means
  - Population Proportions
- Using Confidence Intervals to Design Experiments

Review: SD and Bell-Shaped Curves

If a histogram is bell-shaped, then

- Where is the average?
- What about SD?



No matter what population you are drawing from:

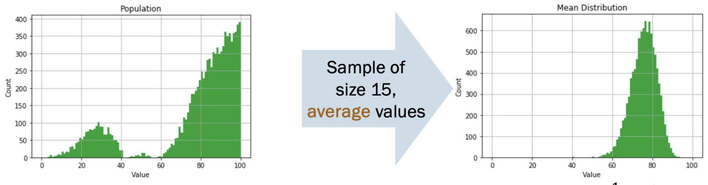
Let $X_1, X_2, \ldots, X_n$ iid, where $E[X_i] = \mu, \mathrm{Var}(X_i) = \sigma^2$. As $n \to \infty$:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Average of iid RVs (sample mean)

(so also works with sample proportions!)

Sampling Distribution of the Statistic:



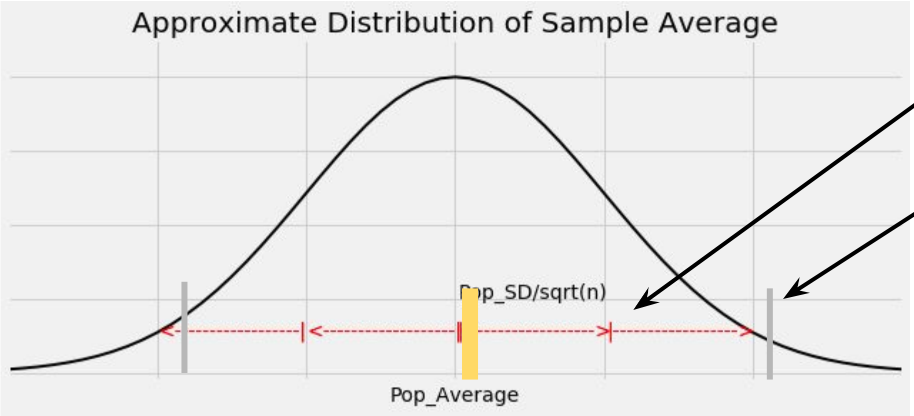
Distribution of X_i (sample mean) Distribution of $\frac{1}{15}\sum_{i=1}^{15} X_i$

Any theorem that provides the rough sampling distribution of a statistic and **doesn't need the distribution of the population** is valuable to data scientists because we rarely know a lot about the population!

Today’s Roadmap

- Using Central Limit Theorem to Calculate Confidence Intervals
  - **Population Means**
  - Population Proportions
- Using Confidence Intervals to Design Experiments

The Key to 95% Confidence



1 SD above the mean

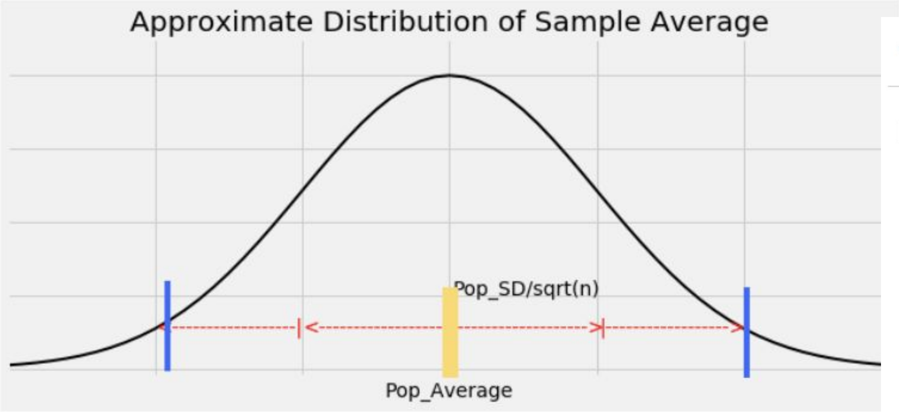
2 SDs above the mean

- **SE (Standard Error) of sample average** = SD of sample average =

$$\left(\frac{\text{Population SD}}{\sqrt{\text{Sample\_Size}}}\right)$$

- For about 95% of all samples, the sample average and population average are within **2 SEs** of each other.

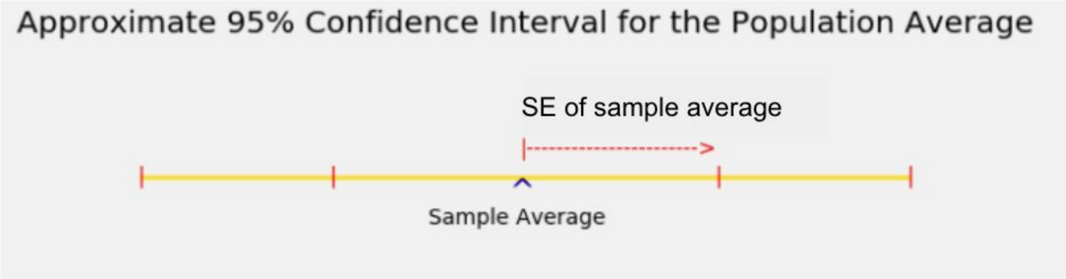
The Key to 95% Confidence



Constructing the Interval

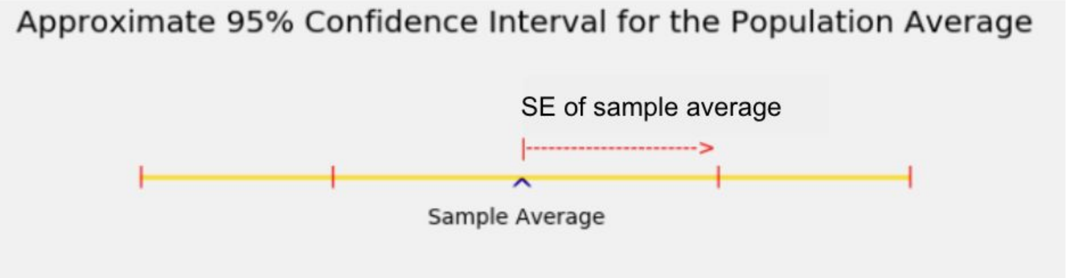
For 95% of all samples,

- If you stand at the population average and look two **SE**s on both sides, you will find the sample average.
- Distance is symmetric.
- So if you stand at the sample average and look two **SE**s on both sides, you will capture the population average.



Summarizing: Construction of 95% Confidence Intervals for the Population Mean

- 95% confidence interval for the population mean:



sample mean $\pm 2 \cdot$ (SE of sample mean)

$$= \text{sample mean } \pm 2 \cdot \left(\frac{\text{Population SD}}{\sqrt{\text{Sample\_Size}}}\right)$$

But we don't know the population SD!

Soln: Estimate it using the sample SD

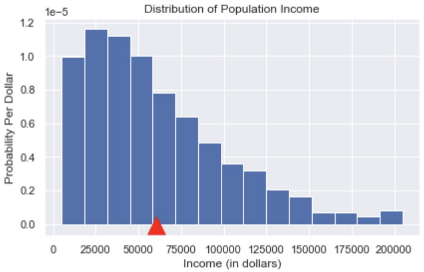
Sample variance: $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$ unbiased estimate of σ^2

Wait, why? See Appendix!

```
sample_SD = np.std(sample,ddof=1)
```

Three Different Distributions and 3 Different Standard Deviations (Recall HW 7)

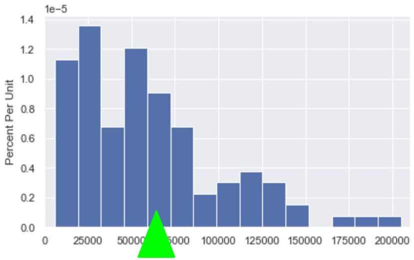
Pop Distribution



Population of Incomes:

- Population mean:
- **Population Income SD:** σ = $41,586

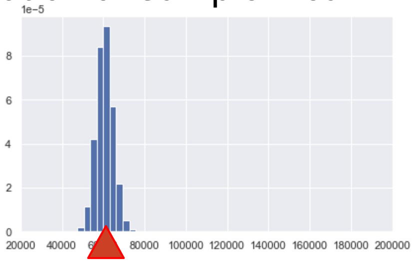
Sample Distribution



Random sample of 100 Incomes

- Sample mean: (estimate of)
- **Sample Income SD: s** = $42,342 (estimate of pop SD)

Sampling Distribution of Sample Mean



Sampling Distribution of Sample Means

- Mean of sample means
- **SD of Sample Means (also called Standard Error)** $\frac{\sigma}{\sqrt{n}}$

Bootstrapping vs CLT

Wait, if we can make 95% confidence interval in this way:

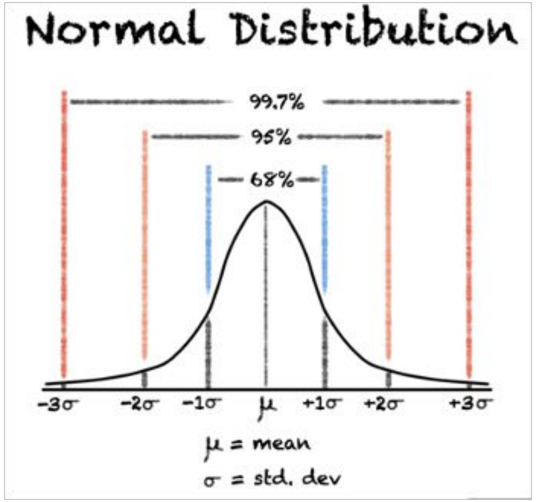
$$\text{sample mean} \pm 2 \cdot (\text{SE of sample mean})$$

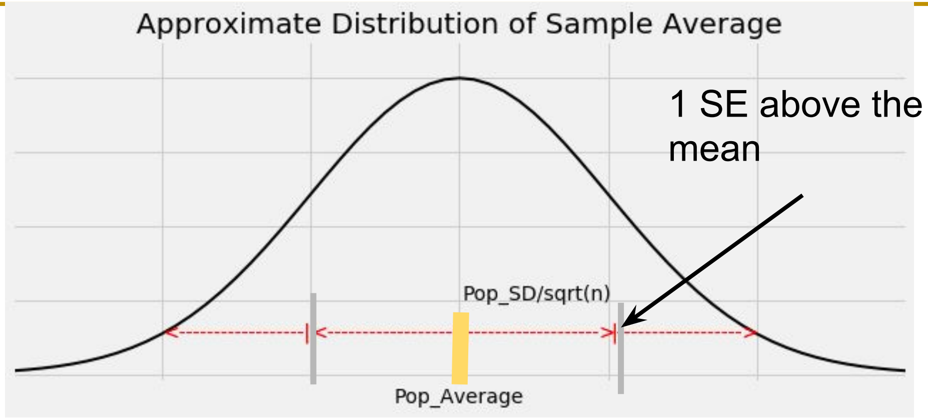
$$= \text{sample mean} \pm 2 \cdot \left(\frac{\text{Population SD}}{\sqrt{\text{Sample\_Size}}}\right)$$

- Then why do we need to make confidence intervals using bootstraps?
  - A: This method only works for means and sums (as it is based on CLT) but bootstrap is a much more generalized approach which can work for other statistics like medians as well

Other Levels of Confidence

Recall:





What if we want to construct a 68% CI?

$$\text{sample mean} \pm 1 \cdot (\text{SE of sample mean})$$

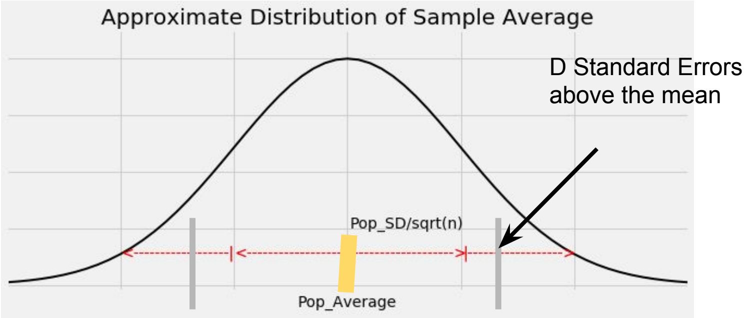
$$= \text{sample mean} \pm 1 \cdot \frac{\textit{Population SD}}{\sqrt{\textit{Sample Size}}}$$

For 68% of all samples,

- If you stand at the population average and look ____ SE on both sides, you will find the sample average.
- Distance is symmetric
- So if you stand at the sample average and look ____ SE on both sides, you will capture the population average.

Other Levels of Confidence

What if we want to construct an L% CI for the population mean?



$$\text{sample mean} \pm D \cdot (\text{SE of sample mean})$$
$$= \text{sample mean} \pm D \cdot \frac{\textit{Population SD}}{\sqrt{\textit{Sample Size}}}$$

```
D = -stats.norm.ppf(1/2*(100-L)/100)
```

For L% of all samples,

- If you stand at the population average and look D standard errors on both sides, you will find the sample average.
- Distance is symmetric
- So if you stand at the sample average and look D standard errors on both sides, you will capture the population average.

The acronym `ppf` stands for `probability point function`. It's the inverse of the `cdf`.

Specifically, `ppf(y)` returns the exact point where the probability of everything to the left is equal to y.

This can be thought of as the percentile function since the ppf tells us the value of a given percentile of the data.

To find the lower SE cut-off for a L% confidence interval, notice that we want the value on the x-axis of the standard normal distribution such that the area to left is equal to $\frac{1}{2}(\frac{100-L}{100})$

CI for Population Proportions

- Using Central Limit Theorem to Calculate Confidence Intervals
  - Population Means
  - **Population Proportions**
- Using Confidence Intervals to Design Experiments

CI for Proportions Using CLT

Ex: You randomly poll CU 400 students and ask if they think we should move the academic calendar to start (and end) the fall semester a week earlier. 192 students are in favor and the rest are opposed. Use the CLT to find a 95% CI for the proportion of all CU students who would be in favor of this change.

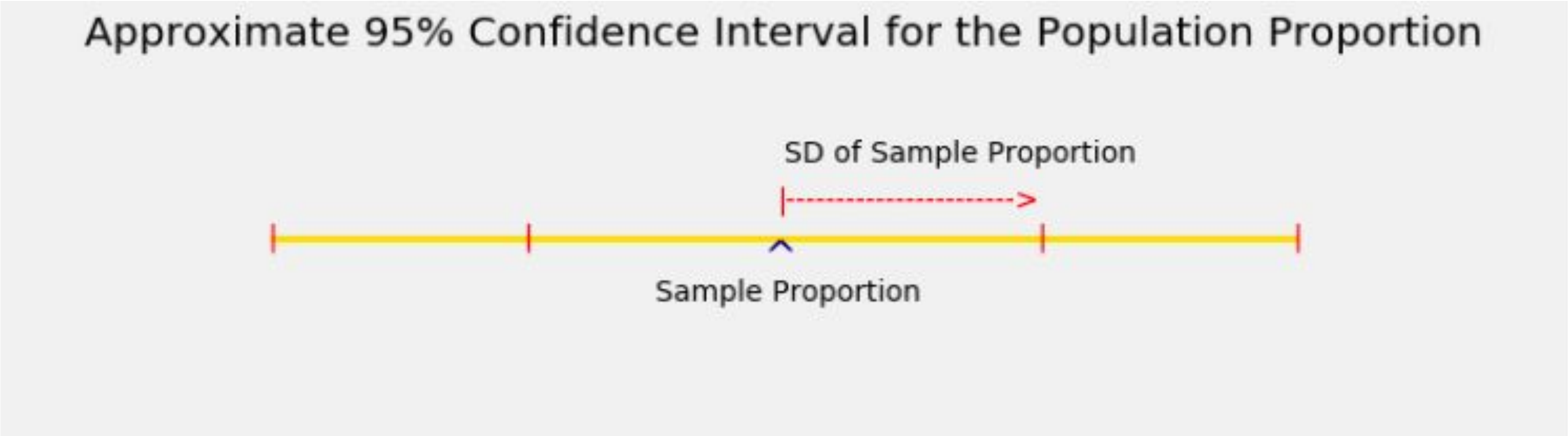
Proportions are Averages

- Data: 0 1 0 0 1 0 1 1 0 0 (10 entries)
- Sum = 4 = number of 1’s
- Average = 4/10 = 0.4 = proportion of 1’s

If the population consists of 1’s and 0’s (yes/no answers to a question), then:

- the population average is the proportion of 1’s in the population
- the sample average is the proportion of 1’s in the sample

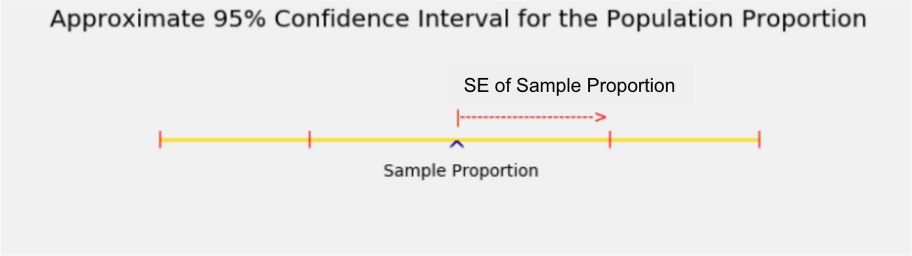
Confidence Interval for Population Proportions Using CLT



Controlling the Width

- Total width of an approximate 95% confidence interval for a population proportion

$$= \quad 2 * 2* \underbrace{\left(\frac{\text{SD of 0/1 population}}{\sqrt{\text{Sample\_Size}}}\right)}_{\text{SE of sample proportion}}$$



- The narrower the interval, the more precise your estimate.
- Wait, what is the SD of a 0/1 population? We've done this!

Recall: Standard Deviation of a Bernoulli RV

$$\mathbb{E}[X] = \sum_x x P(X = x)$$

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2\end{aligned}$$

Definitions

Let X be a **Bernoulli**(p) random variable.

- Takes on value <u>1 with probability p</u>, and <u>0 with probability 1 - p</u>.
- AKA the "indicator" random variable.

Variance =

Standard Deviation=

Standard Deviation of Bernoulli RV

$$\mathbb{E}[X] = \sum_x x P(X = x)$$

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$
$$= \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Definitions

Let X be a **Bernoulli**(p) random variable.

- Takes on value 1 with probability p, and 0 with probability 1 - p.
- AKA the "indicator" random variable.

$$\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1 - p) = p$$

We will get an average value of p across many, many samples

Variance:

Standard Deviation of Bernoulli RV

$$\mathbb{E}[X] = \sum_x x P(X = x)$$

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2\end{aligned}$$

Definitions

Let X be a **Bernoulli**(p) random variable.

- Takes on value <u>1 with probability p</u>, and <u>0 with probability 1 - p</u>.
- AKA the “indicator” random variable.

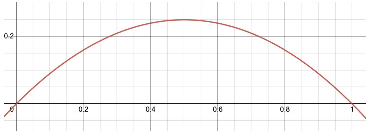
$$\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1 - p) = p$$

We will get an average value of p across many, many samples

Variance:

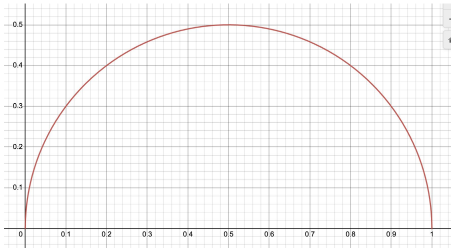
$$\mathbb{E}[X^2] = 1^2 \cdot p + 0 \cdot (1 - p) = p$$

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= p - p^2 = p(1 - p)\end{aligned}$$



Standard Deviation of 0/1’s:

$$\sqrt{p(1 - p)}$$



CI for Proportions Using CLT

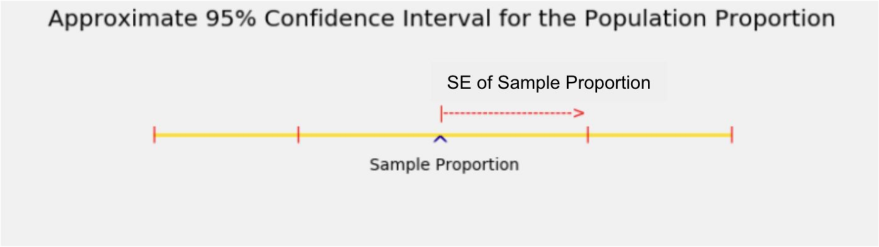
Ex: You randomly poll CU 400 students and ask if they think we should move the academic calendar to start (and end) the fall semester a week earlier.   192 students are in favor and the rest are opposed. Use the CLT to find a 95% CI for the proportion of all CU students who would be in favor of this change.

Determining Sample Sizes

- Using Central Limit Theorem to Calculate Confidence Intervals
- **Using Confidence Intervals to Design Experiments**

Determining The Sample Size for a Given Width

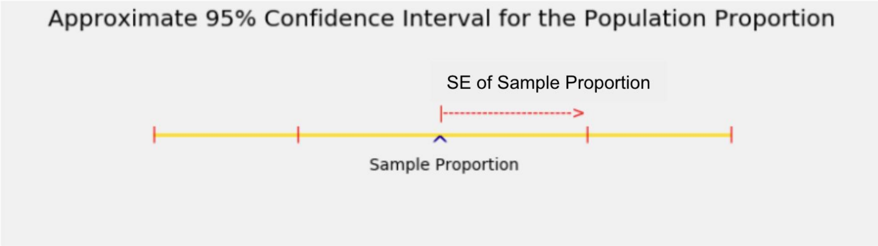
Ex: Suppose you want the total width of the 95% CI interval for a proportion to be no more than 1%. What sample size should you use?



Determining The Sample Size for a Given Width

Ex: Suppose you want the total width of the 95% CI interval for a proportion to be no more than 1%. What sample size should you use?

$$0.01 = 2*2*\left(\frac{\text{SD of 0/1 population}}{\sqrt{\text{Sample\_Size}}}\right)$$

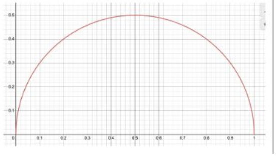


Left side:
the max total width that you'll accept

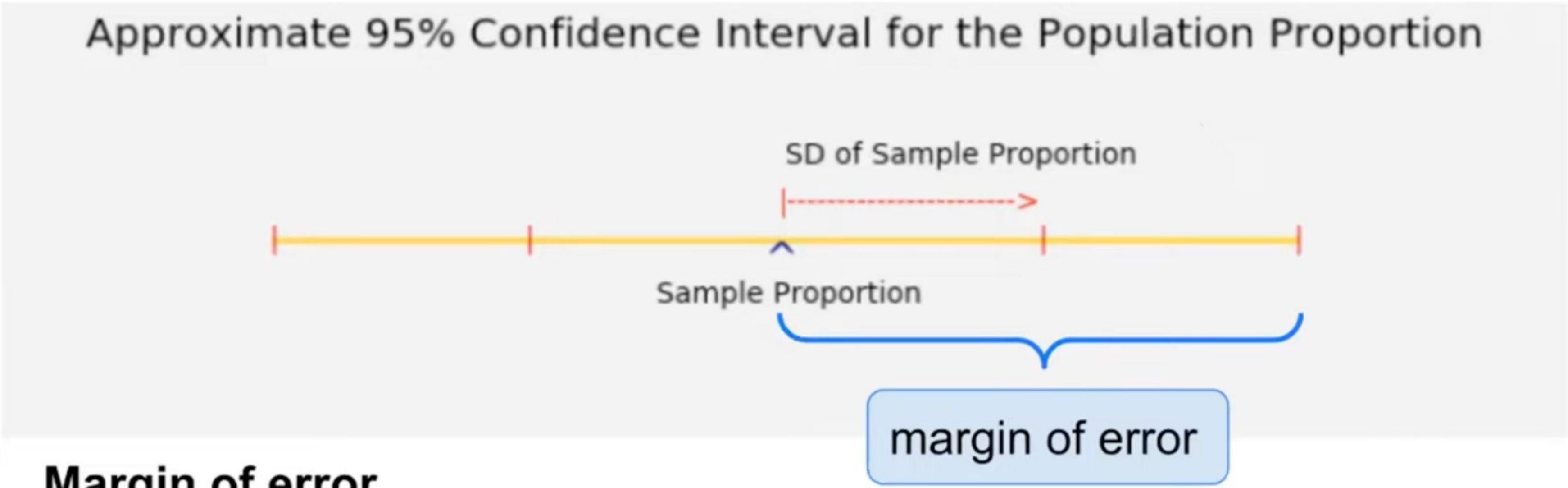
Right side:
formula for the total width

Standard Deviation of 0/1's:

$\sqrt{p(1-p)}$



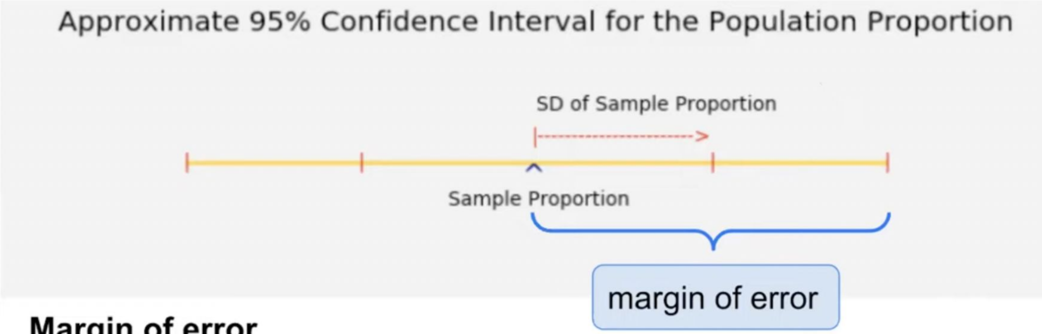
Margin of Error in Polls



Margin of error

- Distance from the center to an end
- Half the width of the interval

Margin of Error in Polls



Margin of error

- Distance from the center to an end
- Half the width of the interval
- 2 * SD of sample proportion

Warm-Up:

How many Americans would you have to randomly poll (about whether or not they'll vote for a particular candidate) to get a 95% CI with a margin of error less than or equal to 3%? Choose the smallest number that is applicable.

A) 1,112

B) 10,112

C) 50,112

D) 100,112

E) None of the above

Margin of Error in Polls

Subscribe SCIENTIFIC AMERICAN Cart 0 Sign In | Stay Informed

THE SCIENCES MIND HEALTH TECH SUSTAINABILITY EDUCATION VIDEO PODCASTS BLOGS PUBLICATIONS

THE SCIENCES

How can a poll of only 1,004 Americans represent 260 million people with only a 3 percent margin of error?

Note that the 3 percent margin of error is an understatement because opinions change.
A poll is a snapshot, not a forecast.

https://www.scientificamerican.com/article/howcan-a-poll-of-only-100/

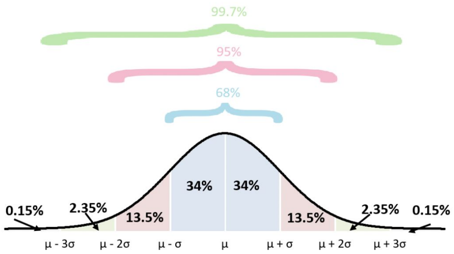
Discussion Question

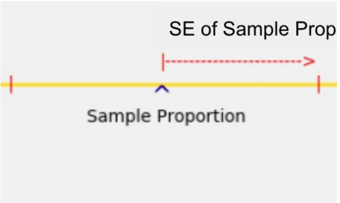
- I am going to use a 68% confidence interval to estimate a population proportion.
- I want the total width of my interval to be no more than 2.5%.
- How large must my random sample be?

Discussion Question

- I am going to use a 68% confidence interval to estimate a population proportion.
- I want the total width of my interval to be no more than 2.5%.
- How large must my random sample be?

The following picture depicts a much-often spouted fact in statistics classes that roughly 68% of the probability for a normal distribution falls within 1 standard deviation of the mean, roughly 95% falls within two standard deviations of the mean, etc:





Discussion Question

- A researcher is estimating a population proportion based on a random sample of size 10,000.

Fill in the blank with a decimal:

- With chance at least 95%, the estimate will be correct to within ______________.

Discussion Question

- With chance at least 95%, the estimate will be correct to within **0.01**.

width = $4 * (0.5) / \sqrt{10000}$

width = 0.02, so margin of error = 0.01

Appendix

Intuition about formula for unbiased estimator of variance

Estimating Population Variance

If we knew the entire population $(x_1, x_2, \ldots, x_N)$:

population mean

population variance $$\sigma^2 = E[(X-\mu)^2] = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

If we only have a sample, $(X_1, X_2, \ldots, X_n)$:

sample mean

sample variance $$\boxed{S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Video: https://www.youtube.com/watch?v=sHRBg6BhKjI

Estimating Population Variance

If we only have a sample, $(X_1, X_2, \ldots, X_n)$:

The best estimate of σ^2 is the **sample variance**: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

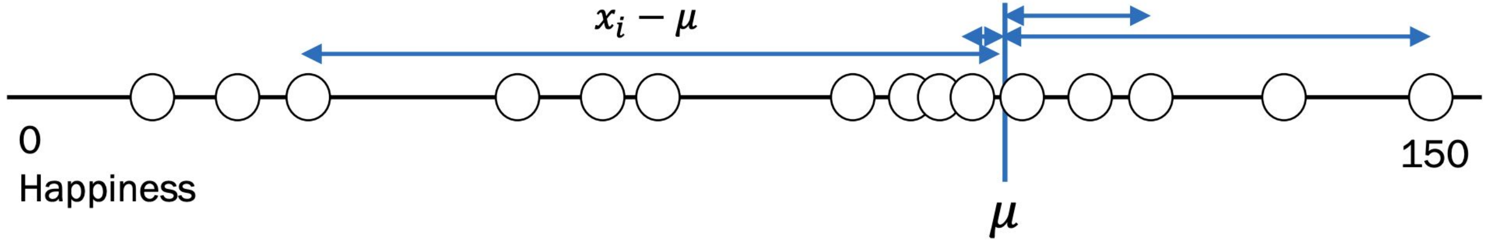
S^2 is an **unbiased estimator** of the population variance, σ^2. $E[S^2] = \sigma^2$

Intuition about the sample variance, S^2

Actual, σ^2

population mean

population variance $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$



Population size, N

Calculating population statistics <u>exactly</u> requires us knowing all N datapoints.

Intuition about the sample variance, S^2

Actual, σ^2

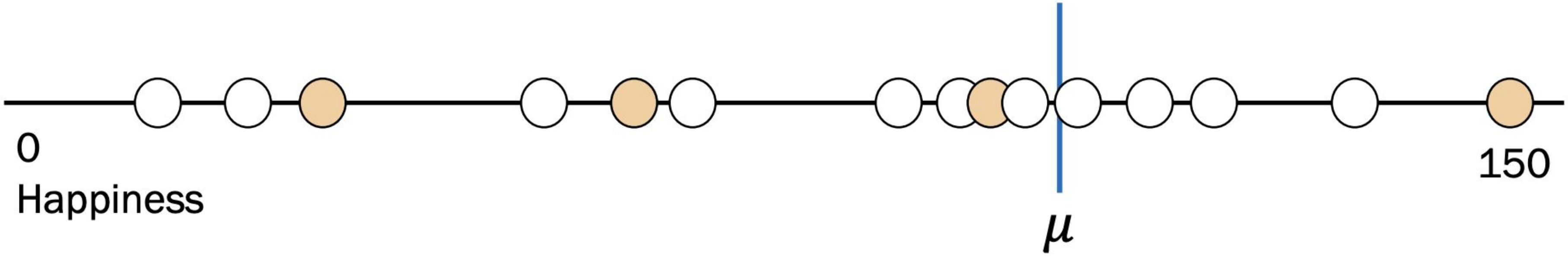
population mean

population variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

Estimate, S^2

sample mean

sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$



Population size, N

Intuition about the sample variance, S^2

<u>Actual, σ^2</u>

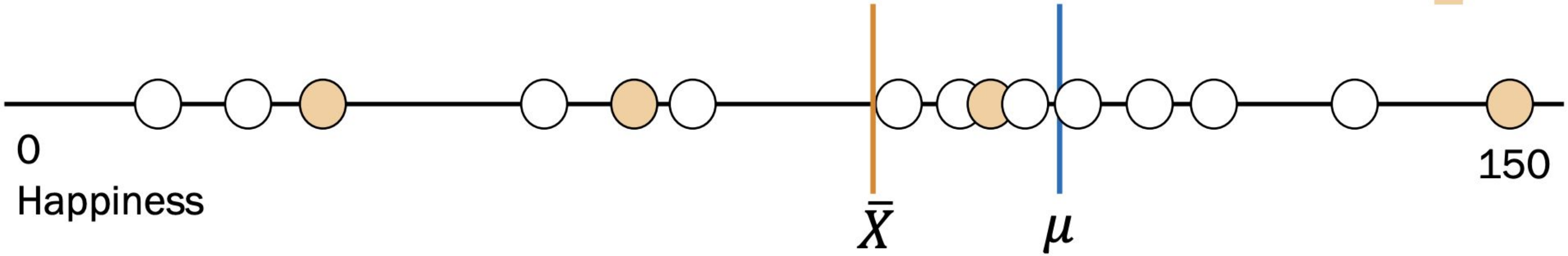
population mean

population variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

<u>Estimate, S^2</u>

sample mean

sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$



Population size, N

Intuition about the sample variance, S^2

Actual, σ^2

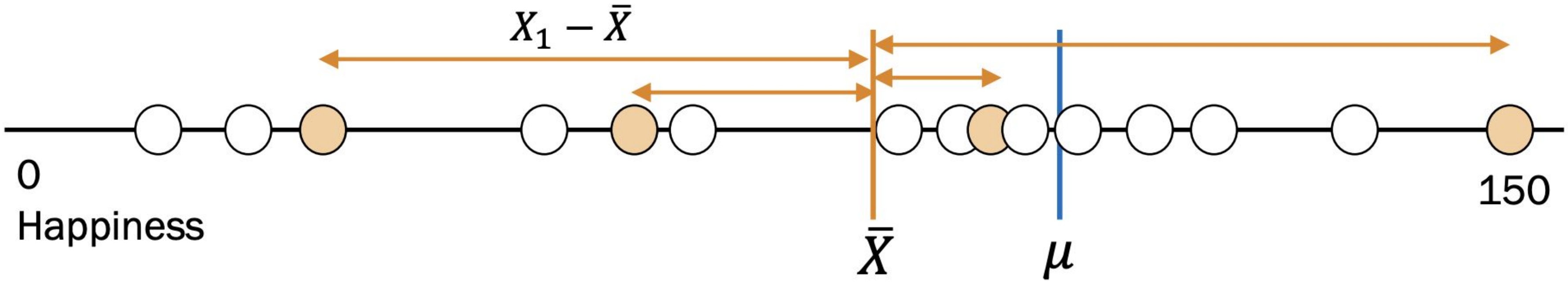
population variance — population mean

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Estimate, S^2

sample variance — sample mean

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\underbrace{(X_i - \bar{X})}^2$$



Population size, N

Sample variance is an estimate using an estimate, so it needs additional scaling.

Proof that S^2 is unbiased (just for reference)

$$E[S^2] = \sigma^2$$

$$E[S^2] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \quad \Rightarrow \quad (n-1)E[S^2] = E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

$$(n-1)E[S^2] = E\left[\sum_{i=1}^{n}\big((X_i - \mu) + (\mu - \bar{X})\big)^2\right]$$

(introduce $\mu - \mu$)

$$= E\left[\sum_{i=1}^{n}(X_i - \mu)^2 + \sum_{i=1}^{n}(\mu - \bar{X})^2 + 2\sum_{i=1}^{n}(X_i - \mu)(\mu - \bar{X})\right]$$

$$2(\mu - \bar{X})\sum_{i=1}^{n}(X_i - \mu)$$

$$2(\mu - \bar{X})\left(\sum_{i=1}^{n}X_i - n\mu\right)$$

$$2(\mu - \bar{X})n(\bar{X} - \mu)$$

$$-2n(\mu - \bar{X})^2$$

$$= E\left[\sum_{i=1}^{n}(X_i - \mu)^2 + n(\mu - \bar{X})^2 - 2n(\mu - \bar{X})^2\right]$$

$$= E\left[\sum_{i=1}^{n}(X_i - \mu)^2 - n(\mu - \bar{X})^2\right] = \sum_{i=1}^{n}E[(X_i - \mu)]^2 - nE[(\bar{X} - \mu)^2]$$

$$= n\sigma^2 - n\mathrm{Var}(\bar{X}) = n\sigma^2 - n\frac{\sigma^2}{n} = n\sigma^2 - n\sigma^2 = (n-1)\sigma^2$$

Therefore $E[S^2] = \sigma^2$