Confidence Intervals & Designing Experiments

LECTURE 25

CSCI 3022

Maribeth Oscamou

Content credit: Acknowledgments

Announcements

- Hw 10 released tonight

Today's Roadmap

CSCI 3022

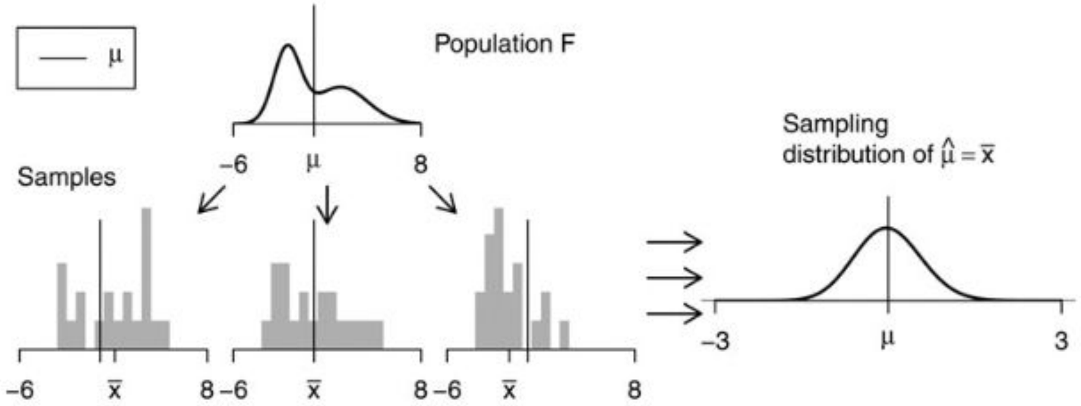
- Using Confidence Intervals for Hypothesis Testing
- Using Central Limit Theorem to Calculate Confidence Intervals
- Using Confidence Intervals to Design Experiments

Recap:
Calculating Confidence Intervals

The Sampling Distribution of a Statistic

"Ideal world":

- To determine the properties (i.e. shape and standard error) of the sampling distribution of an estimator, we'd need to have access to the population.
- Sampling distributions of a statistic are obtained by drawing repeated samples from the population, computing the statistic of interest for each and collecting (an infinite number of) those statistics as the sampling distribution.

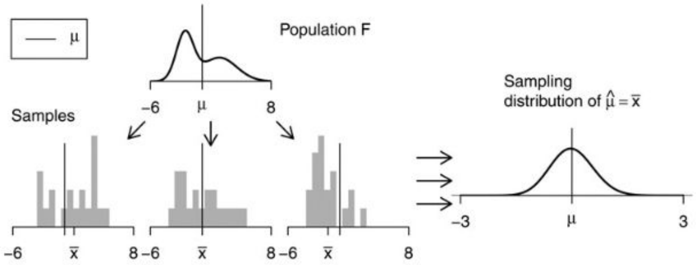


Calculating Confidence Intervals

- Method 1: Bootstrapping
- Method 2: Using the Central Limit Theorem (if it applies)

The Sampling Distribution of a Statistic

Ideal world:

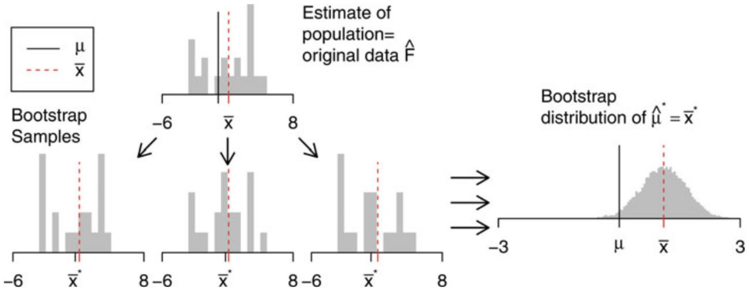


Reality: We don't know the population distribution.

Method 1: Treat our random sample as a "population", and resample from it.

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

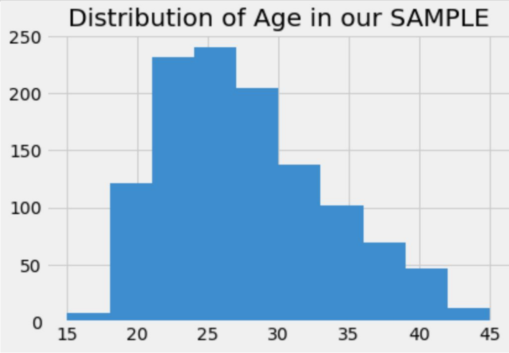
Bootstrap World:

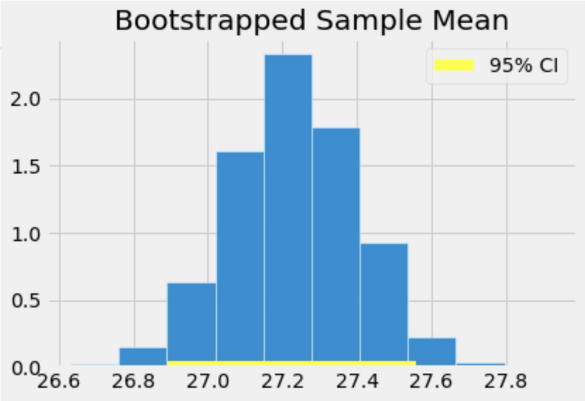


Bootstrap World: Draw repeated samples from an estimate of the population, computing the statistic of interest for each, and collecting those statistics. The ***bootstrapped distribution is centered at the observed statistic,*** not the actual population parameter.

Use the middle X% of this distribution to calculate X% Confidence Interval

Last Time:





95% Bootstrapped Confidence Interval for Average Age of Mothers in the Population:

[26.89267461669506, 27.55792163543441]

Using the Confidence Interval for Testing Hypotheses

Null: The average age of mothers in the population is 25 years; the random sample average is different due to chance.

Alternative: The average age of the mothers in the population is **not** 25 years.

Suppose you use the 5% cutoff for the p-value.

Poll: Based on the 95% Confidence Interval, what conclusion would you make?

A: Reject the null hypothesis

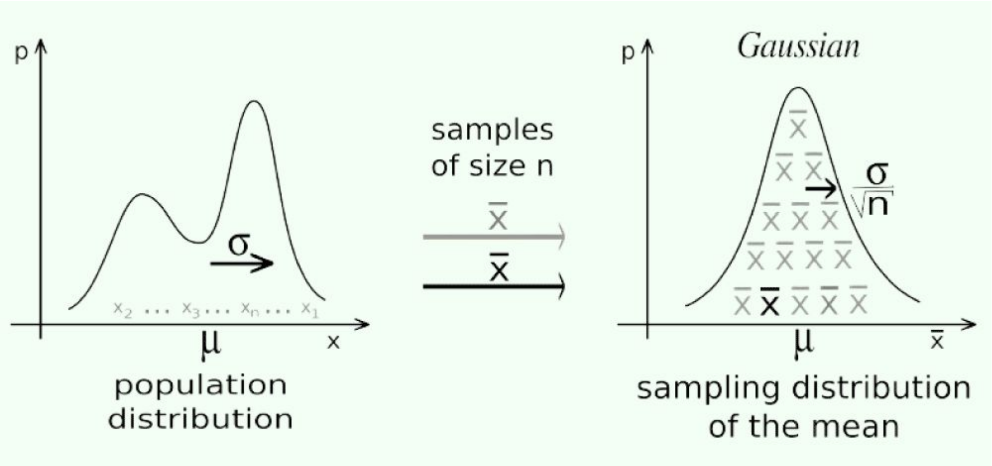
B: Fail to reject the null hypothesis

Using a CI for Testing

- **Null hypothesis**: **Population average = *x***
- **Alternative hypothesis**: **Population average ≠ *x***
- Cutoff for **p-value**: *p*%
- Method:
    - Construct a (100-*p*)% confidence interval for the population average
    - If *x* is not in the interval, reject the null
    - If *x* is in the interval, can't reject the null

(Demo)

Method 2: Using the Central Limit Theorem to Calculate Confidence Intervals



Recall: The Central Limit Theorem

No matter what population you are drawing from:

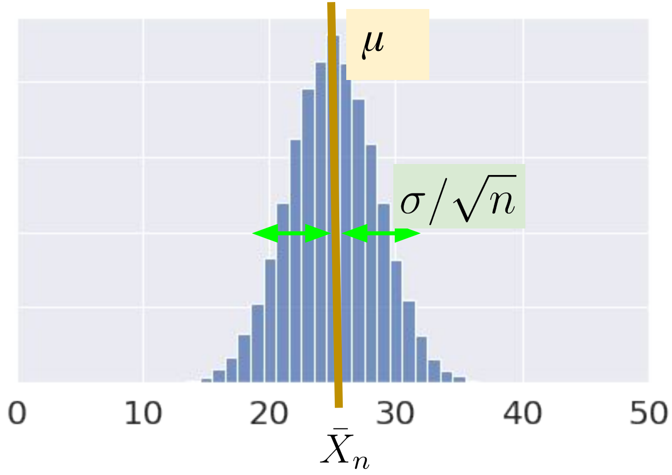
If an IID sample of size n is large,

the probability distribution of the **sample mean**

is **roughly normal** with

mean μ and SD (also called standard error SE) $\sigma/\sqrt{n}$

(where pop mean= μ, pop SD =σ)

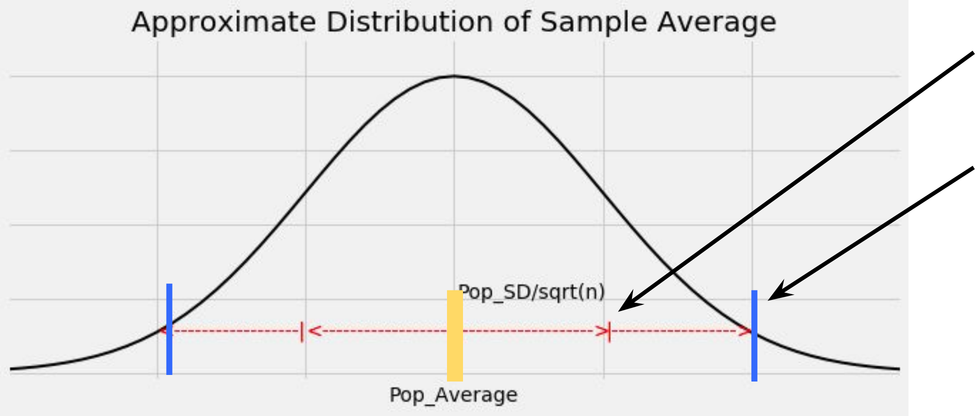


Any theorem that provides the rough distribution of a statistic
and **doesn't need the distribution of the population** is valuable to data scientists.

- Because we rarely know a lot about the population!

For a more in-depth demo: https://onlinestatbook.com/stat_sim/sampling_dist/

The Key to 95% Confidence



1 SE above the mean

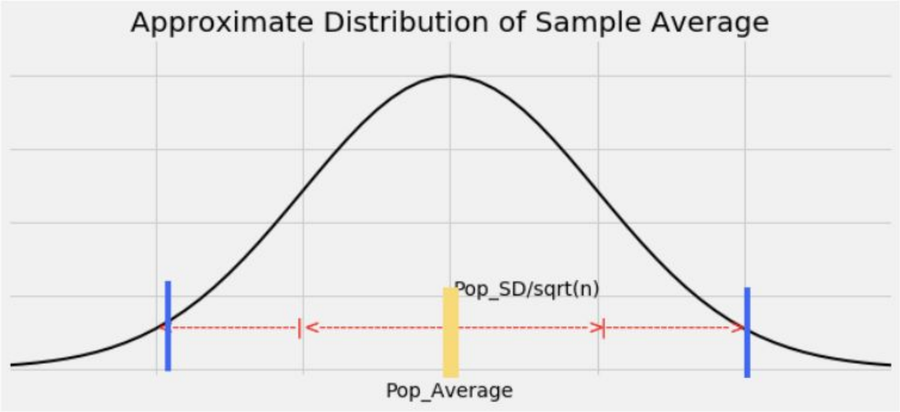
2 SEs above the mean

- **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.

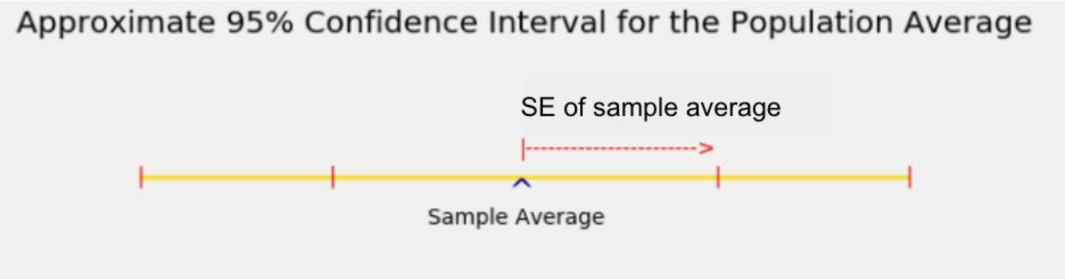
The Key to 95% Confidence



Constructing the Interval

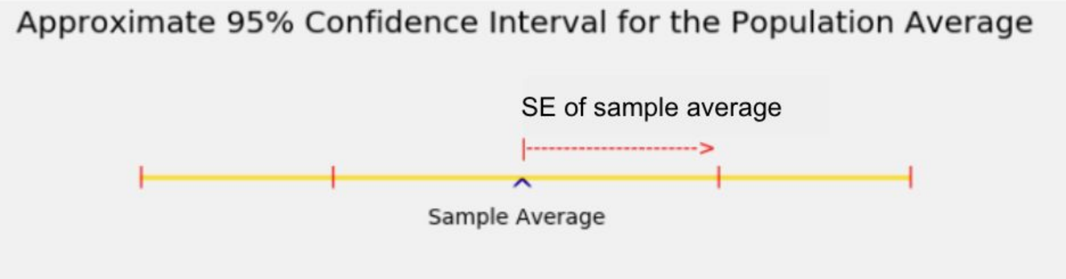
For 95% of all samples,

- 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 intervals

- 95% confidence interval for the sample mean:



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

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

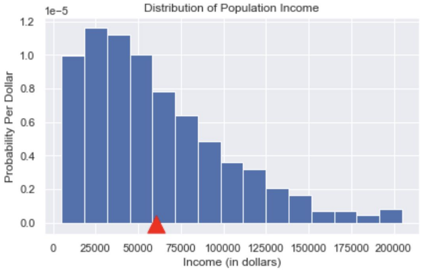
But we don't know the population SD!

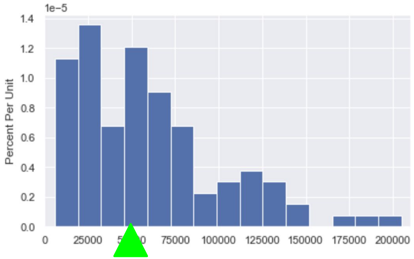
Soln: Estimate it using the sample SD

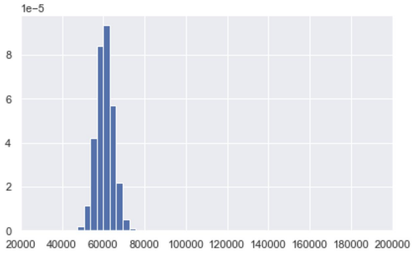
```
sample_SD = np.std(sample,ddof=1)
```

(Demo)

Three Different SDs: Recall HW 8







Population of Incomes:

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

Random sample of 100 Incomes

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

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

SE (SD of sample averages): $\frac{\sigma}{\sqrt{n}}$

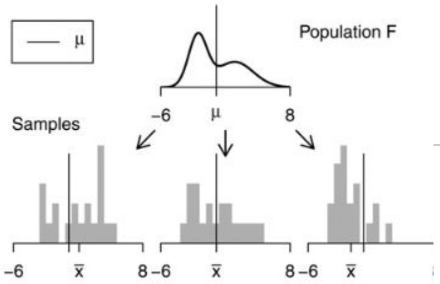
If we calculated from 10,000 samples, the SD of those 10,000 sample means would be approx $\frac{\sigma}{\sqrt{n}} =$

What if we want something other than 95% CI? Constructing a X% CI for the Population Average

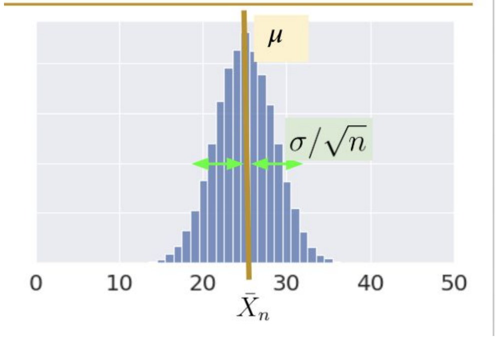
- Use sampling distribution of sample average to solve for D, where D is defined as follows:
  - For X% of all samples, if you stand at the **population average** and look a distance of D on both sides, you will find the **sample average**.

- But notice distance is symmetric!

- So if you stand at the ***sample average*** and look a distance of D on both sides, you will capture the **population average** X% of the time.



Ex: Sampling distribution of sample average



Ex: Suppose pop average = 25

You draw one sample with sample ave =29

X% Conf Int = [29 -D, 29+D]

Question

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

Confidence Intervals for Sample Proportions

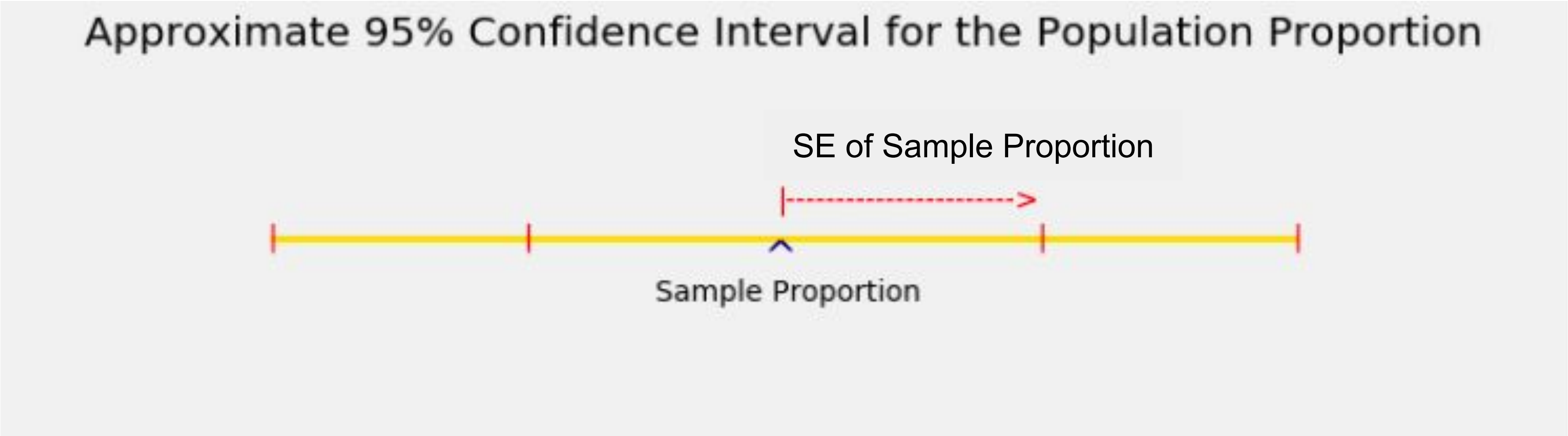
Proportions are Averages

- 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

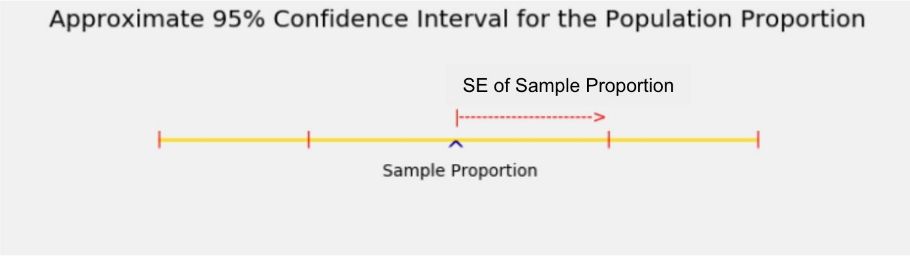
Confidence Interval



Controlling the Width

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

$$= \quad 4 * \underbrace{\left(\frac{\text{SD of 0/1 population}}{\sqrt{\text{Sample\_Size}}}\right)}_{\text{SE of sample proportion}}$$



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

Recall: Lesson 14: Standard Deviation of a Bernoulli RV

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

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

Definitions

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

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

Variance =

Standard Deviation=

Standard Deviation of Bernoulli RV

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

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

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

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

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

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.

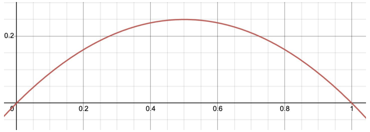
$$\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1 - p) = p$$

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

Variance:

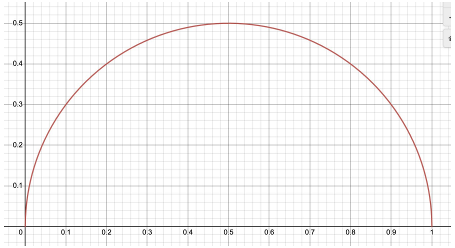
$$\mathbb{E}[X^2] = 1^2 \cdot p + 0^{\cdot}(1 - p) = p$$

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



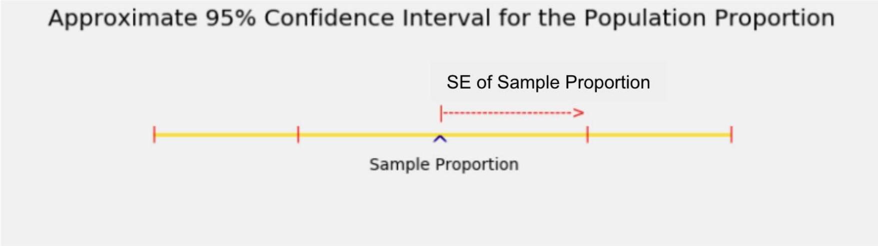
Standard Deviation of 0/1's:

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



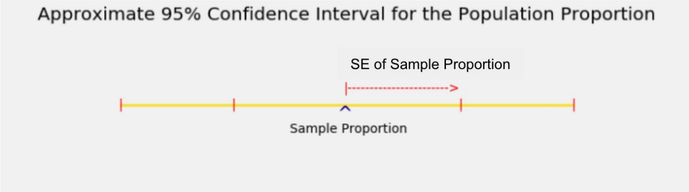
Controlling the Width With Sample Size

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?



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

$$0.01 = 4 * \left(\frac{\text{SD of 0/1 population}}{\sqrt{\text{Sample\_Size}}} \right)$$

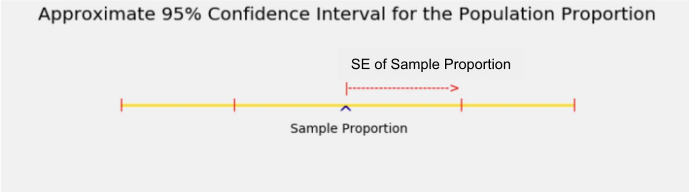


Left side:
the max total width that you'll accept

Right side:
formula for the total width

Solving for sample size:

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



$$0.01 = 4 * \left(\frac{\text{SD of 0/1 population}}{\sqrt{\text{Sample\_Size}}} \right)$$

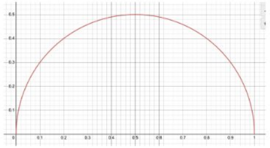
Left side:
the max total width that you'll accept

Right side:
formula for the total width

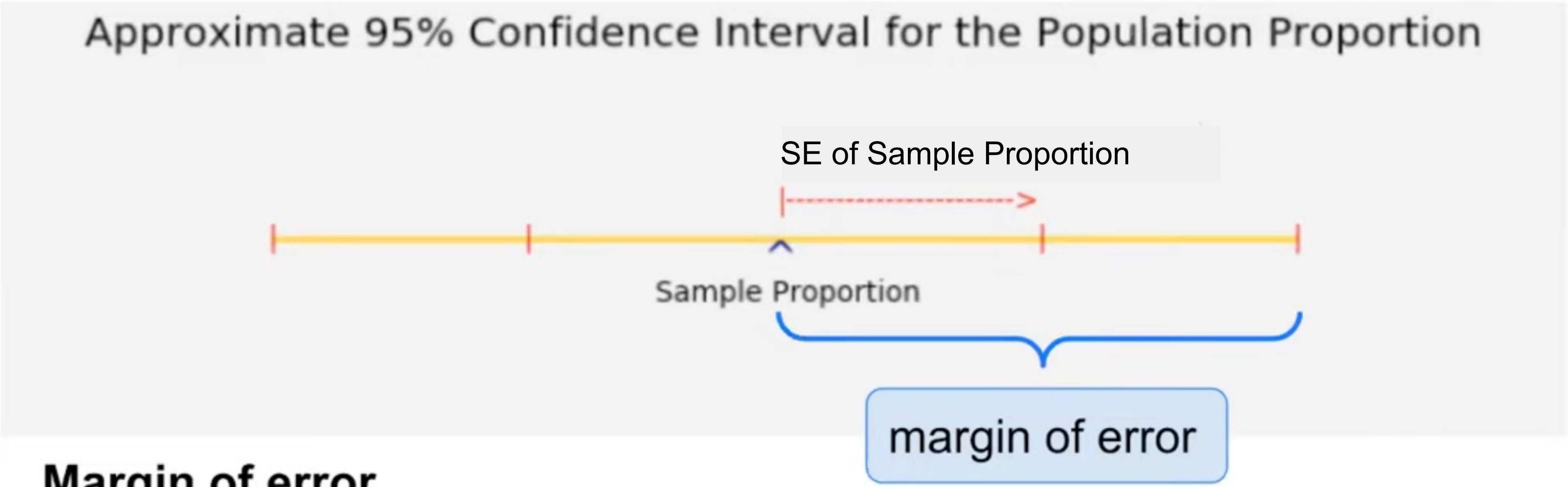
Solving for sample size:

Standard Deviation of 0/1's:

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



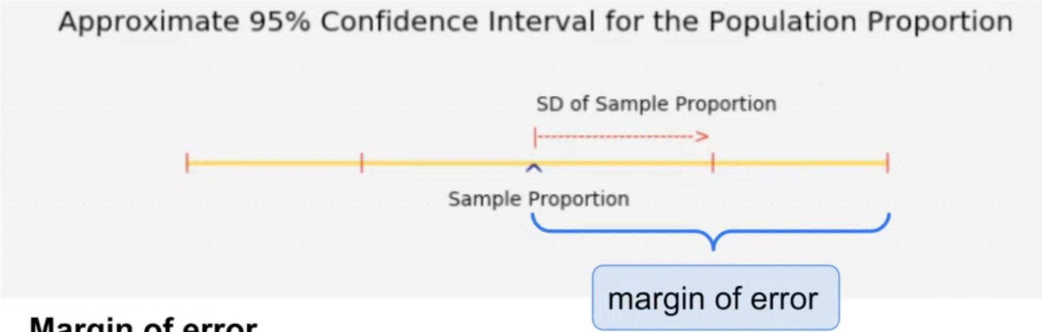
Margin of Error in Polls



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

Poll:

How many Americans would you have to randomly poll (about whether or not they'll vote for a particular candidate) to get 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

Discussion Question


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?

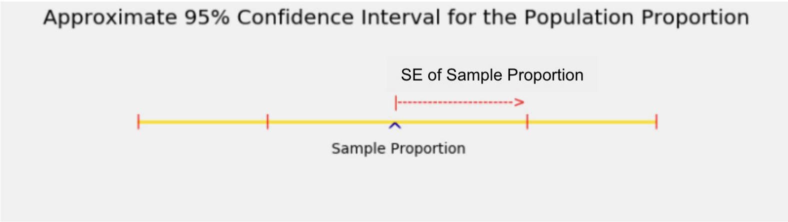
Finally, the 3 percent margin of error is an understatement because opinions change. A poll is a snapshot, not a forecast.

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

Discussion Question

THE SCIENCES

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



- 3% margin of error means **width of** ______

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

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

- How large must my random sample be?

$0.025 = \mathbf{2} * (0.5) / \sqrt{\text{sample size}}$

$\sqrt{\text{sample size}} = 2 * (0.5) / 0.025$

sample size = 40**2 = 1600

Extra Info about Sample Variance vs 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 $$\boxed{S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

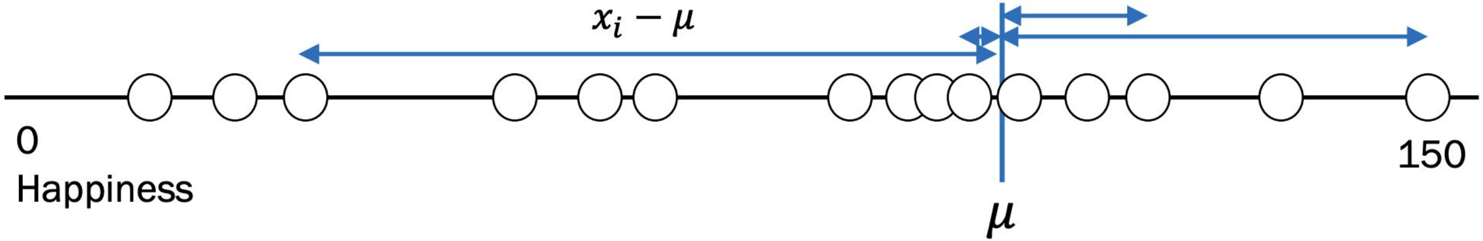
Video: https://www.youtube.com/watch?v=sHRBg6BhKjI

Intuition about the sample variance, S^2

<u>Actual, σ^2</u>

population mean

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



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

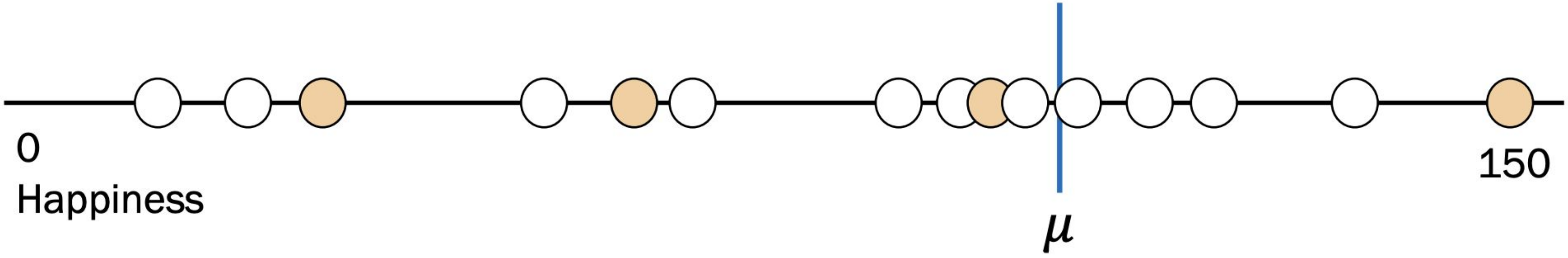
population mean

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

Actual, σ^2

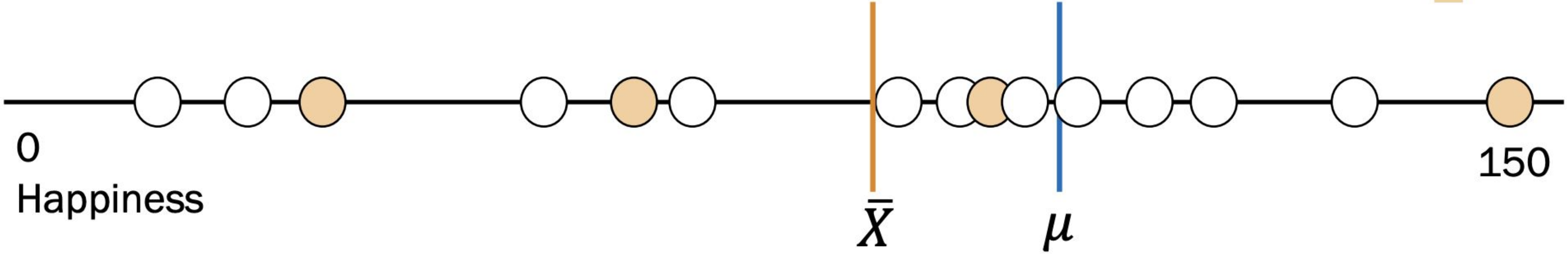
population mean

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

Actual, σ^2

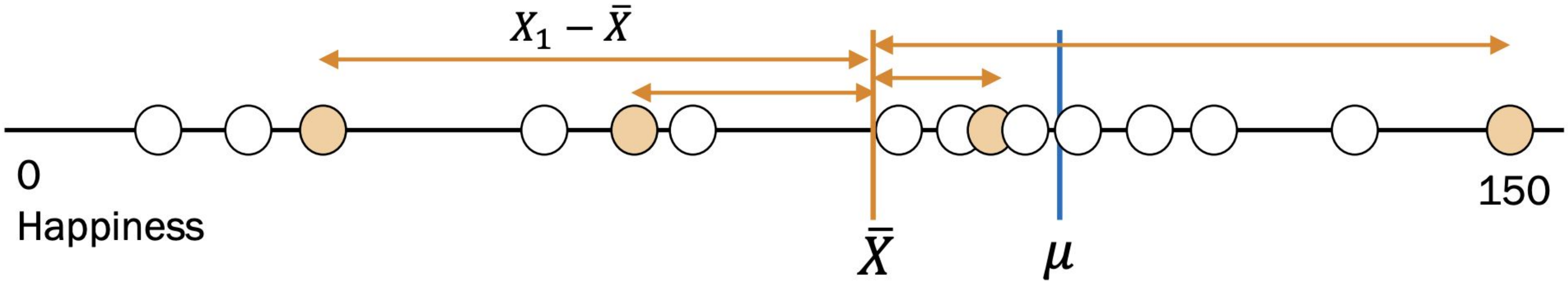
population mean

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

Estimate, S^2

sample mean

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



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