Day. 79, Feb 16, 2025 (Falgun 4, 2081)

↳ Asymptotics

Asymptotics is the term for the behaviour of statistics as the sample size (or some other relevant quantity) limits to infinity (or some other relevant number).

↳ Asymptotics are incredibly useful for simple statistical inference and approximations.

Contains the infinite Dev.

↳ Asymptotics form the basis for frequency interpretation of probabilities (the long run proportion of times an event occurs).

Limits of Random Variables

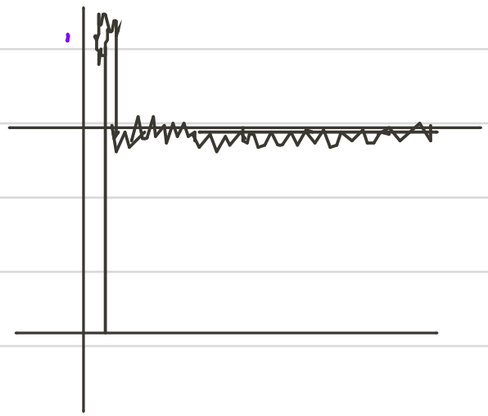
These results allow us to talk about the larger sample distribution of sample means of collection of iid observations.

The first of these results, the law of large numbers, we intuitively know

→ It says that the average limits to what it actually the population mean.

→ Example $\bar{X}_n$ could be the average of the result of n coin flips (i.e. the sample proportion of heads).

→ As we flip a fair coin over and over. if eventually converges to the true probability of a head.



Law of Large Numbers in Action:

```
n <- 1000
means <- cumsum(rnorm(n))/(1:n)
```

Mean → Cumulative sum

generate Random.

n of mean or ($\bar{x}$).

Law of large numbers in action, Bin flip

means <- cumsum(sample (0:1, n, replace = TRUE))/ (1:n)

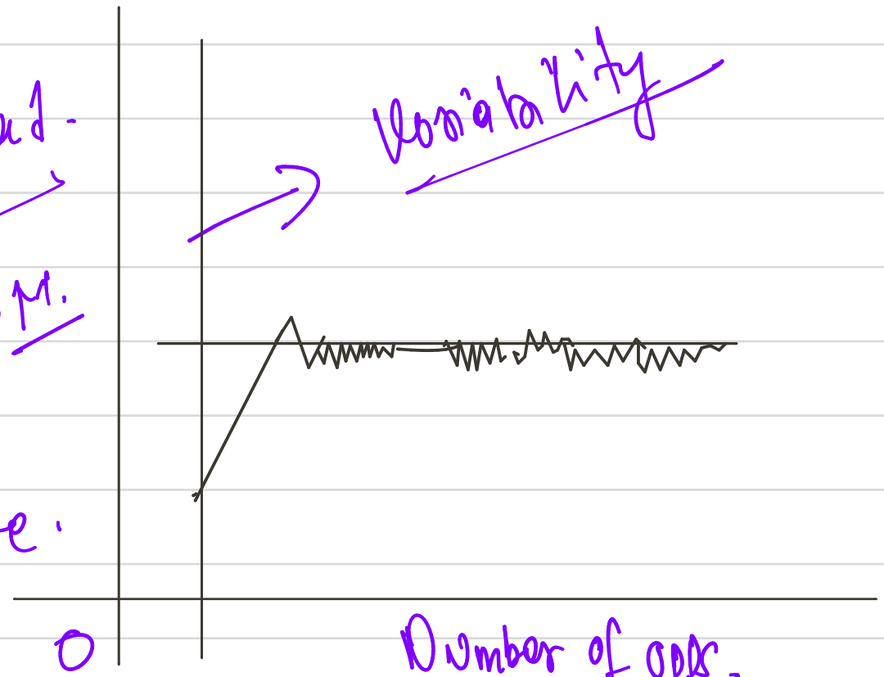
↳ 0's and 1's

↳ 1 is Head.

Discussion.

→ An estimator is Consistent if it CM.

Converges to what we want to estimate.



→ The LLN says that the sample mean of iid samples is Consistent for the population μ.

↳ Typically, good estimators are consistent, it's not too much to ask that if we go to the trouble of collecting an infinite amount of data that we get the Right answer.

⇒ The sample variance and the sample standard Deviation of iid random variables are consistent as well.

Central Limit Theorem

For our purposes, the CLT states that the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the sample size increases.

$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$\Rightarrow \frac{\sqrt{n}\,(\bar{X}_n - \mu)}{\sigma}$$

$$\Rightarrow \frac{\text{estimate} - \text{Mean of estimate}}{\text{std Error of estimate}}$$

has a distribution like that of a standard normal for large n.

The useful way to think about the CLT is that $\bar{X}_n$ is approximately $N(\mu, \sigma^2/n)$ -

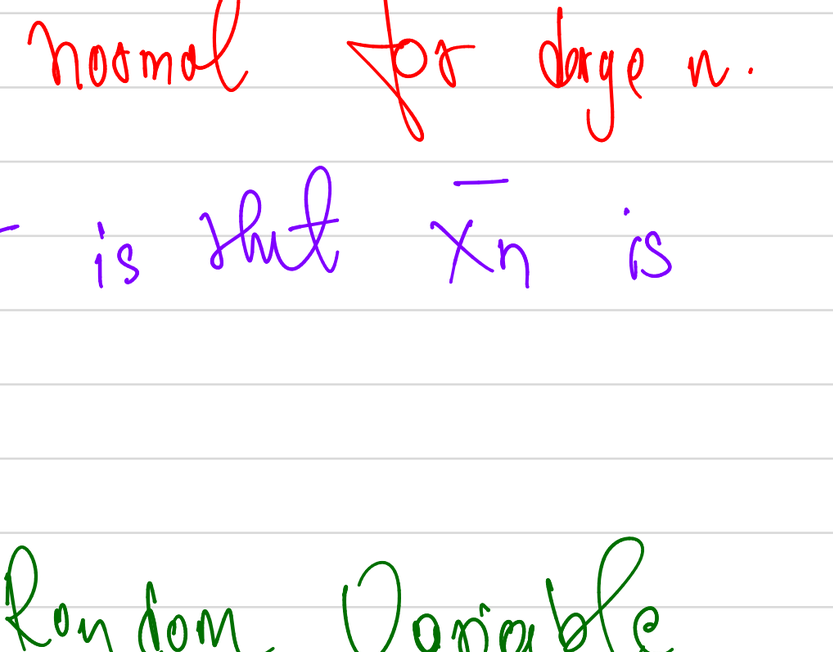
Example: Simulate a standard Normal Random Variable by rolling n (six sided).

Let X_i be the outcome for die i

then note that $\mu = E[X_i] = 3.5$

$Var(X_i) = 2.92$

SE $\sqrt{2.92/n} \Rightarrow 1.71/\sqrt{n}$



dots roll n dice, take their mean, subtract off 3.5 and divide by $1.71/\sqrt{n}$.

Coin CLT.

Let X_i be the 0 or 1 result of the i^{th} flip of a possibly unfair coin

$\rightarrow$ the sample proportion, say $\hat{p}$, is the average of the coin flips

$\rightarrow$ $E[X_i] = p$ and $Var(X_i) = p(1-p)$

$\rightarrow$ Standard error of the mean is $\sqrt{p(1-p)/n}$

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

→ $p = 1/2$

→ $p * (1-p) = 1/4$

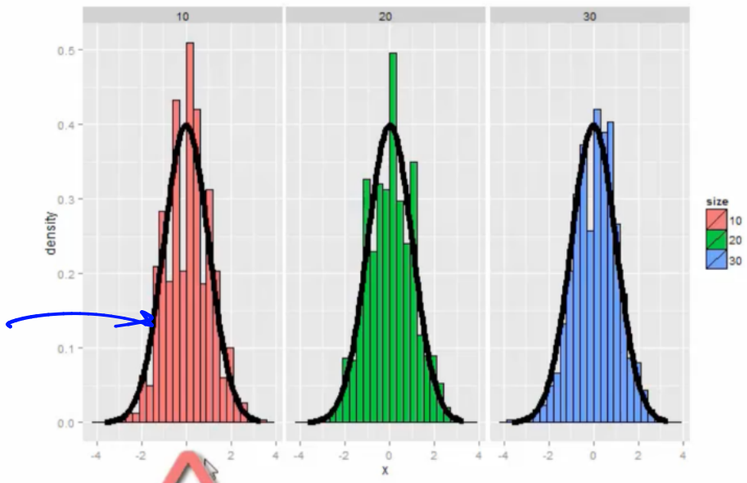
→ Sqrt $(p * (1-p)) = 1/2$

[p = 0.9] → Simulation Results.

View the Galton's Quincunx.

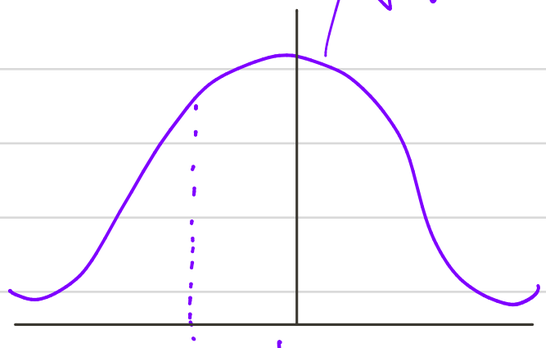
Result of our die rolling experiment

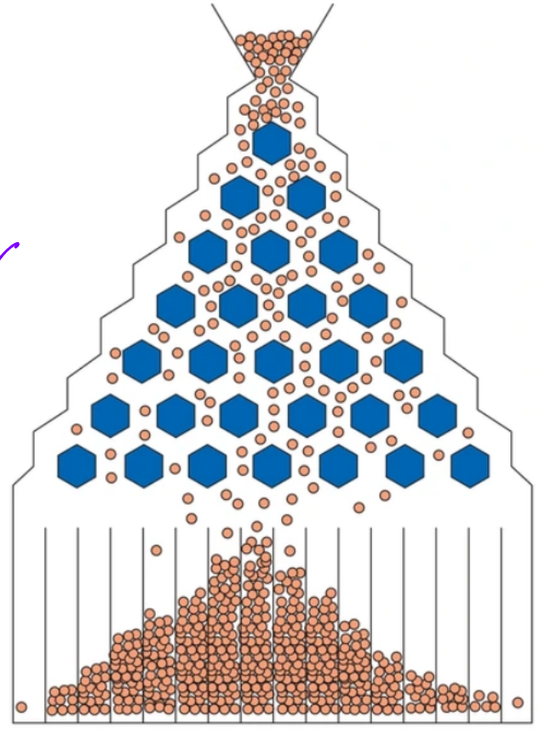
Density.



Confidence Intervals

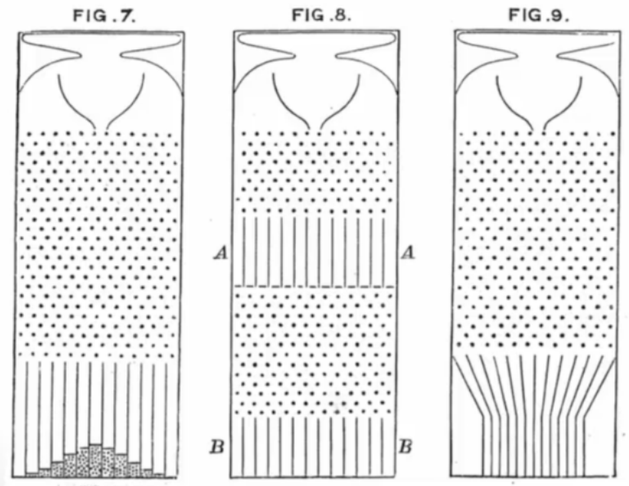
$\bar{X}$ is approximately normal with μ and SD. $\sigma/\sqrt{n}$.





Galton's quincunx

http://en.wikipedia.org/wiki/Bean_machine#mediaviewer/File:Quincunx_(Galton_Box)_-_Galton_1889_diagram.png



13/31

So the probability $\bar{X}$ is bigger than $\mu + 2\sigma/\sqrt{n}$ or smaller than $\mu - 2\sigma/\sqrt{n}$ is 5%.

$\bar{X} \pm 2\sigma/\sqrt{n}$ is called a 95% interval for μ.

Sample Proportions

- In the event that each Y_i is 0 or 1 with common success probability p then $\sigma^2 = p(1-p)$.
- The interval takes the form

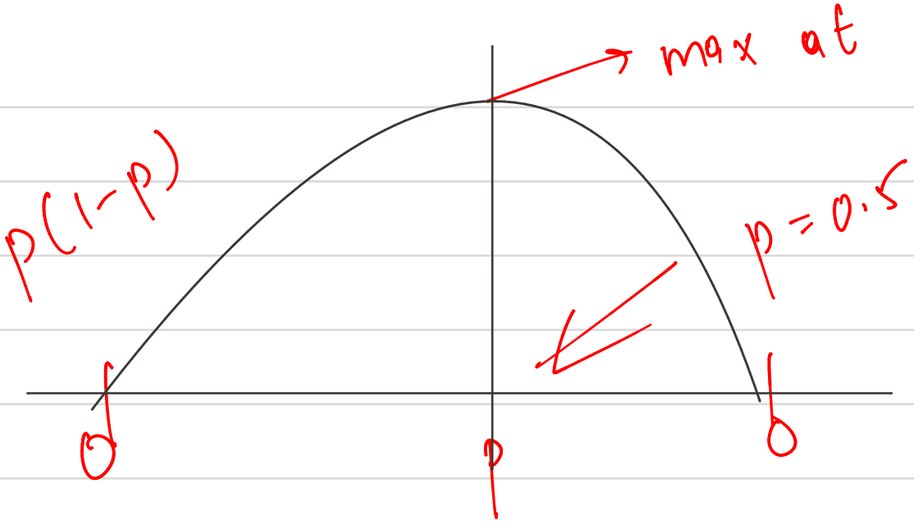
$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

- Replacing p by $\hat{p}$ in the standard error results in what is called a Wald Confidence interval for p.
- for 95% intervals

$$\hat{p} \pm \frac{1}{\sqrt{n}}$$

is a quick CI estimate for p.

$$\sqrt{\frac{p(1-p)}{n}} \leq \sqrt{\frac{\frac{1}{2}(1-\frac{1}{2})}{n}} = \frac{1}{2\sqrt{n}}$$



$$\text{or, } 2\sqrt{\frac{p(1-p)}{n}} \leq 2 \times \frac{1}{2\sqrt{n}} = \frac{1}{\sqrt{n}}$$

or, for 95%:

$$\hat{p} \pm \frac{1}{\sqrt{n}}$$

Simple Explanation on Confidence Intervals:

- population mean of penguin weight: 31 lbs
- population proportion of voters: 55%

Point estimate:

↳ uses a single value to estimate a population parameter.

Interval estimate:

↳ uses a range of values to estimate a population parameter.

↳ CI is also type of Interval Estimate.

Penguin data

- Population mean = 31 lbs.
- Sample mean = 28 lbs.
- Sample mean = 32 lbs.

Confidence Interval includes

- Sample Statistic
  - Margin of Error
  - Confidence Level

Interval

Sample statistic +/− margin of error. where

Margin of error

the maximum expected difference between a population parameter and a sample estimate

Interval : [28, 32]

30 + 2 = 32 , 30 − 2 = 28

[Margin of Error = Z-score * SE]

Confidence Level

Describe the likelihood that a particular sampling method will produce a confidence interval that includes the population parameter.

Example: 95% CI - (95 of 100 → contains actual population mean).
- 100 random samples

Common Confidence Levels

- 90%
- 95% → Mostly used choice –
- 99%

Sales Revenue:

- "I think we'll do $1,000,000 in sales."
- "Based on a 95% Confidence Interval, I estimate that our sales revenue 'll be between $ 950,000 and $ 1,050,000".

Components of a Confidence Interval

- Margin of Error:

The main components of a CI are a Sample statistic, margin of error, and confidence level. Confidence Interval help express the uncertainty of an estimate based on Sample data.

- Sample Statistic and Confidence Level are other Components of CI.

Interpret Confidence Intervals

Example: Estimating the mean height of all the red maple trees in the city -

- population: 10,000 trees
- Sample: 50 trees
- Sample mean = 50 ft
- Sample Standard deviation = 7.5 ft
- Confidence Interval: 95% of CI [48, 52]

Confidence Level:

↳ Expresses the uncertainty of the estimation process

↳ 95% of intervals Capture the population mean

↳ 5% of " donot Capture the population mean.

Repeated Random Sampling is often difficult, expensive and time Consuming.

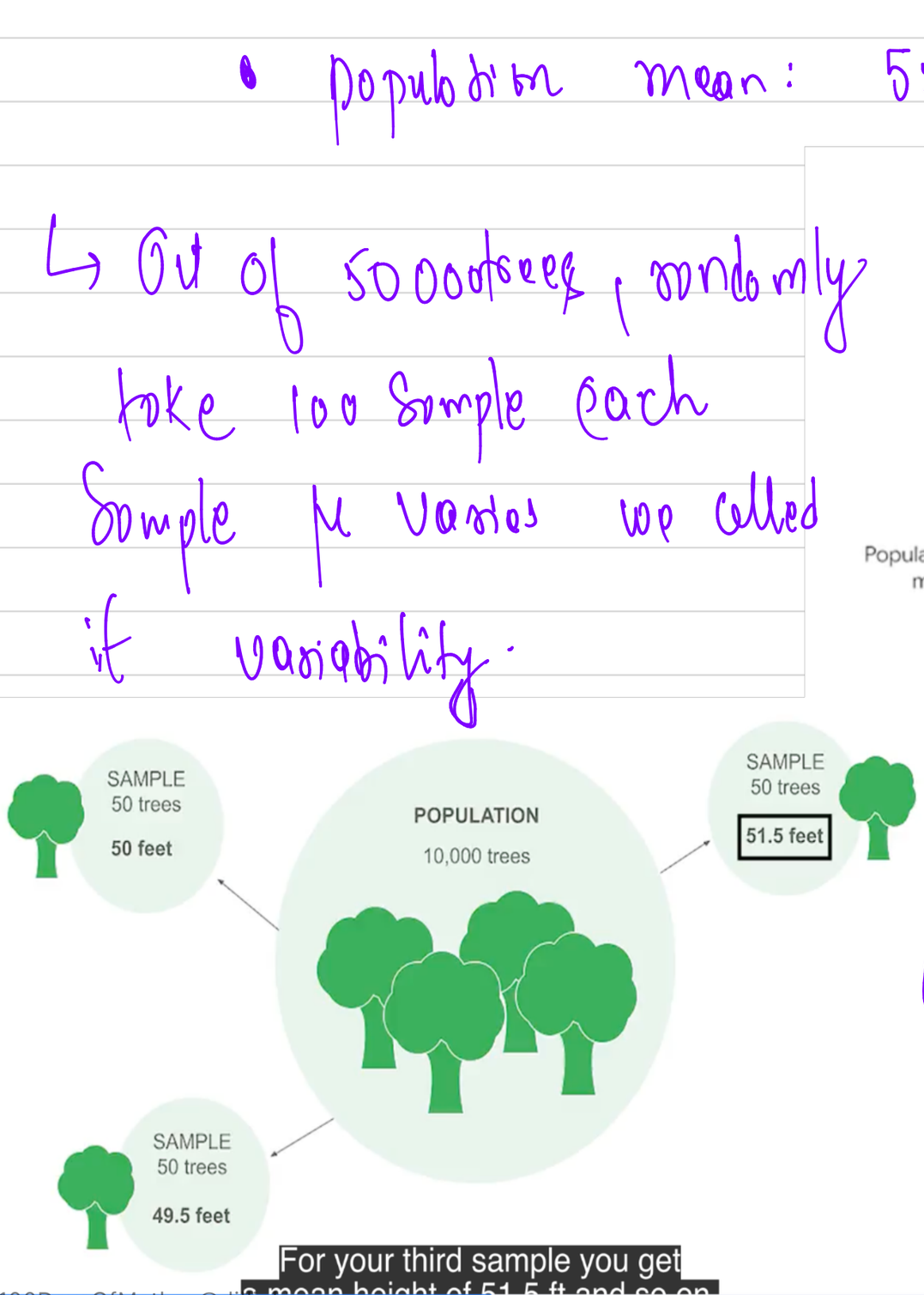
↳ Confidence Intervals give data professionals a way to quantify the uncertainty due to random Sampling.

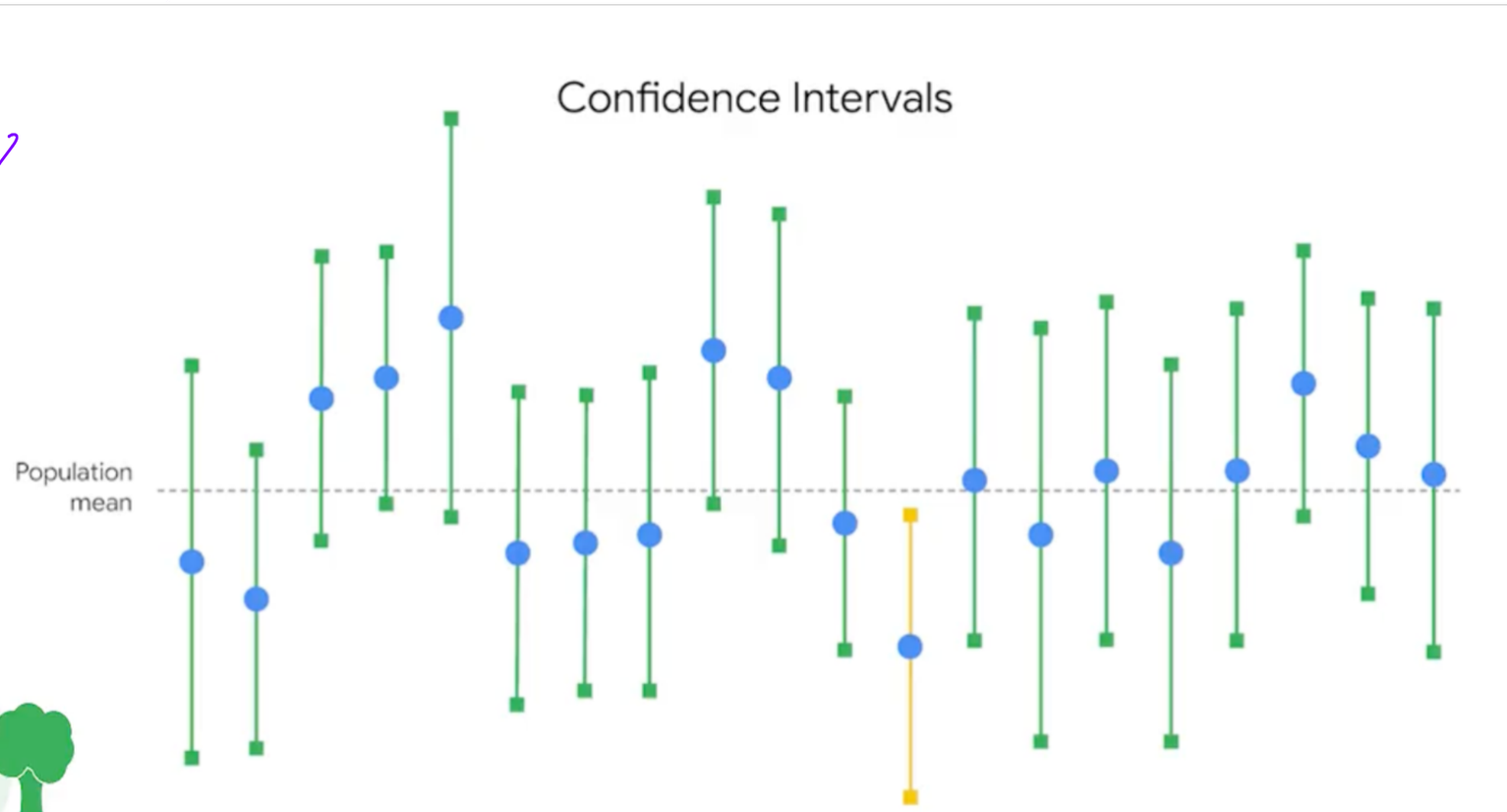
Red Maple Tree Data:

- Confidence Interval: 95% of CI [48, 52]

- Population mean: 52 ft.

↳ Out of 500 trees, randomly take 100 sample each Sample μ varies we called it variability.





↳ Because of sample variability any given sample will not necessarily be equal to the actual population mean.

① A 95% CI means that you can expect that 19 out of 20 intervals or 95% of the total will capture the population μ.

② 95% refers to the success rate of that process or we can expect 95% of the random intervals generate to capture the population parameter.

Construct a Confidence Interval for a Proportion

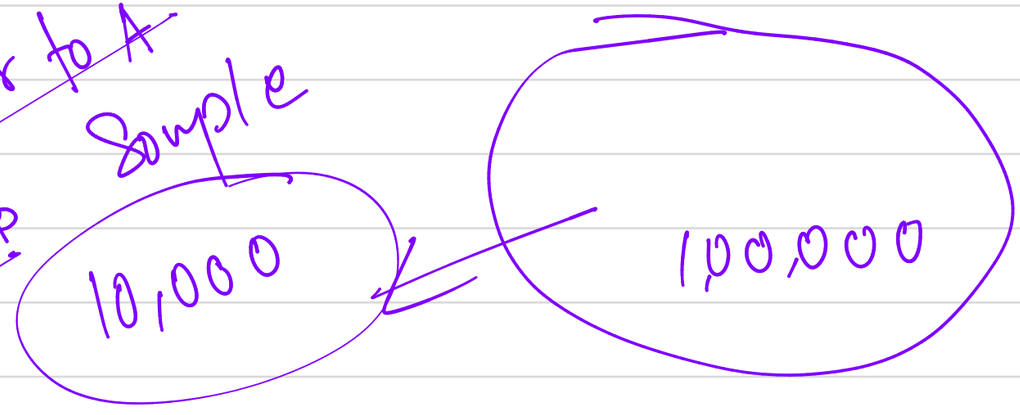
Elective A

Elective B

↳ 55% favor to A

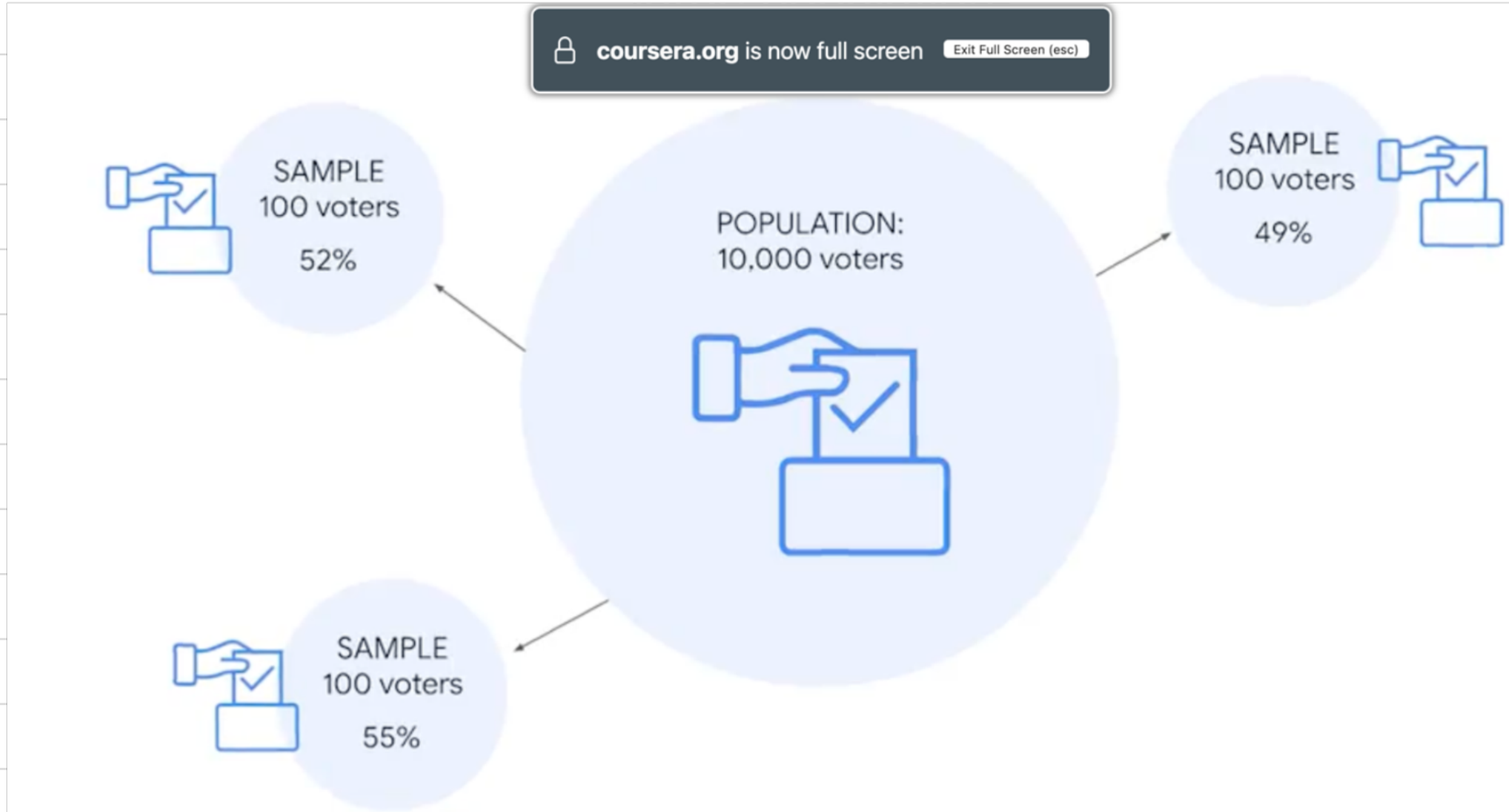
↳ 45% favor to B

Example: This is the 1 Random Sample



1st 100 RS
2nd 100 RS
3rd 100 RS
↓
gives different
of favor so

Instead of point



Estimate we can use sample data to construct a confidence interval.

Steps for Constructing a Confidence Interval

① Identify a Sample Statistic : proportion

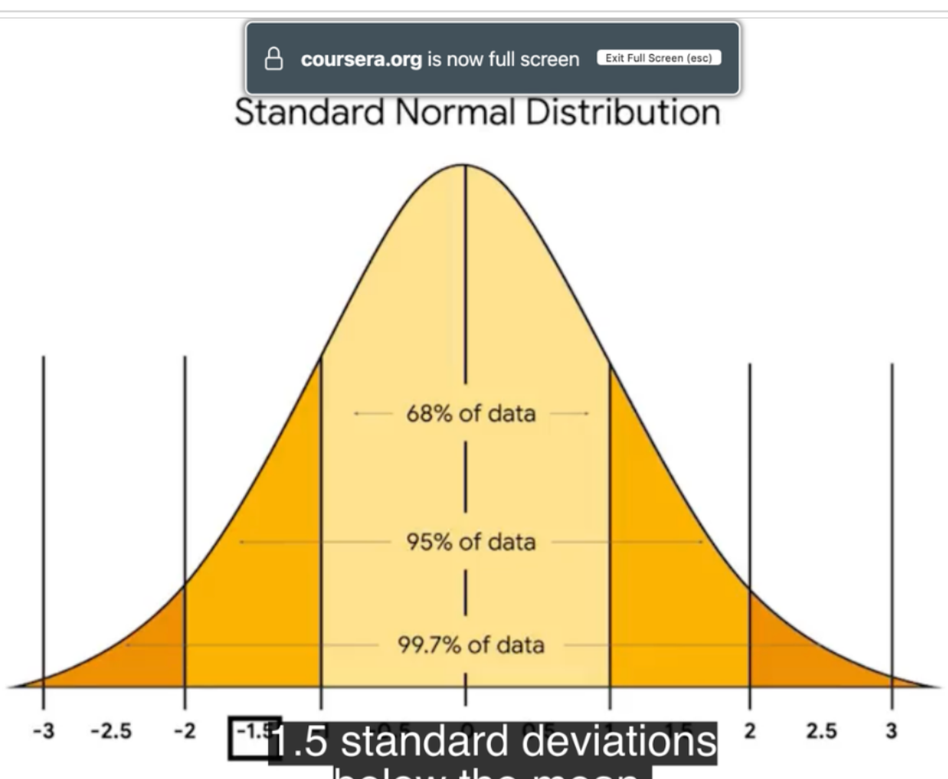
② Choose a Confidence Interval : 95%.

③ find the Margin of Error : Range of values above & below the Sample Statistic

④ Calculate the Interval :

* Margin of error = 2-Score * SE

2-Score measures the distance between the data point from the population mean and a standard normal distribution. 2-Score 1 means 1 Std deviations above the mean (μ).



↳ SE measures the Sample variability.

SE of the proportion

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$n = 100$

So, 2-Score * SE ⇒ 1.96 * 0.05

⇒ 0.098

Confidence level	Z-score
90%	1.645
95%	1.96
99%	2.58

1.645 for 90 percent

Calculate the Interval : [45.2 %, 64.8 %]

upper limit = Sample proportion + Margin of Error

0.55 + 0.098 = 0.648 ⇒ 64.8 %.

Lower limit = Sample proportion − MI of Error

0.55 − 0.098 ⇒ 0.452

45.2%.

[45.2% < 50%]

Previous elections Here, the new poll reports that 54% of voters prefered candidate Davis. Using point Estimation it is 50.9%. But using CI of 95, integral will now above 50% and is 57.1%. this gives more Confidence and 5% of Margin Error need to be considered too.

Sample size = 100 [45.2, 64.8] ⇒ 19.6

Sample size = 1,000 [50.9, 57.1] ⇒ 6.2

Sample size increase: ↑ SE decreases: ↓

So, the first step of constructing a CI is identifying a sample statistic → Next, Confidence Level is chosen. Then the margin of error is found. Finally, the interval is calculated.

Construct a Confidence Interval for a Mean (μ):

↳ Claim of 20 hrs of battery life is accurate for battery life of the new phone?

- Sample mean = 20.5 hrs
- Sample Std. deviation = 1.7 hrs.
- Population Std. deviation = 1.5 hrs.

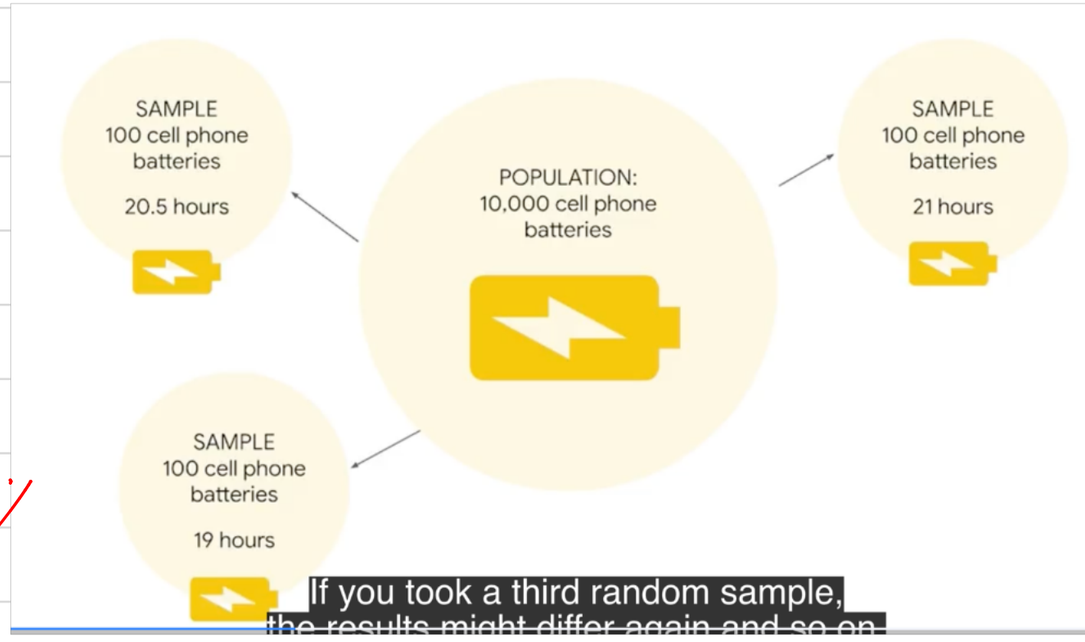
Using Random 100 Samples

1st 100 Random Sample

2nd 100 R. ..

3rd 100 ...

yields different μ or results.



Steps for Constructing a Confidence Interval

1. Identify a sample statistic: mean
2. Confidence Level: 95%
3. Margin Error = 2-Score * SE

Z-Score => distance between data point.

Standard Error of the Mean:

$$SE \Rightarrow \frac{\sigma^*}{\sqrt{n}}$$

*'σ' when std of population is known, otherwise 's';

SE => 0.15

So, Margin Error = 2-Score * SE => 1.96 * 0.15

=> 0.294

④ Calculate the Interval: [20:12, 20:48] → means 20 hrs 12 mintes to 20 hrs 48 mintes.

$20.5 + 0.294 \Rightarrow 20.794$ → upper limit.

$20.5 - 0.294 \Rightarrow 20.206$ → lower limit

⑤ Lower limit of CI = 20:12 (Hypothesized) ? 20:00 (Claim).

What if HEAD of Marketing? doesn't Satisfy → for that we do Some Sample data using 99%. [20:07, 20:53] Still above 20 hours in my the Confidence Interval

Confidence Intervals

- Confidence level = 95%.

  [20:12, 20:48] = 36 min.

- Confidence Level = 99%

  [20:07, 20:53] = 46 min.

#What if σ is unknown? If we don't know the σ we have to use different method of calculation?

Source: the power of Statistics offered by Google Coursera.