

Inferential Statistics



Discrete vs Continuous

- | | |
|---|--|
| ■ Countable | ■ Uncountable |
| ■ Discrete Points | ■ Continuous Intervals |
| ■ $p(x)$ is probability distribution function | ■ $f(x)$ is probability density function |
| ■ $p(x) \geq 0$ | ■ $f(x) \geq 0$ |
| ■ $\sum p(x) = 1$ | ■ Total Area under curve = 1 |

Number of Cars	Probability
0	0.03
1	0.13
2	0.70
3	0.10
4+	0.04

1. Probability

A. Random Variable

- Discrete Random variable
- Continuous Random variable

B. Probability Distributions



C. Binomial Random variable

D. Geometric Random Variable

2. Inference statistics

A. Central Limit theorem

B. Confidence Interval

C. Hypothesis testing

Statistics

Descriptive

1. Organizing and summarizing data using numbers & graphs
2. Data Summary:
Bar Graphs, Histograms, Pie Charts, etc.
Shape of graph & skewness
3. Measures of Central Tendency:
Mean, Median, & Mode
4. Measures of Variability:
Range, variance, & Standard deviation

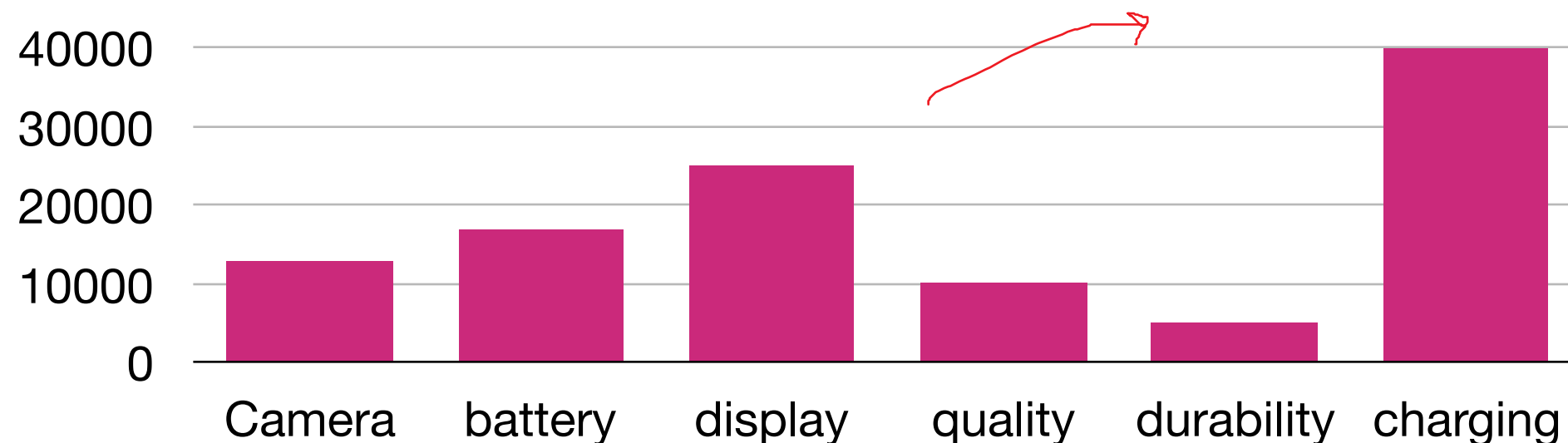
Inferential

1. Using sample data to make an inference or draw a conclusion of the population.
2. Uses probability to determine how confident we can be that the conclusions we make are correct.
(Confidence Intervals & Margins of Error)

Case studies

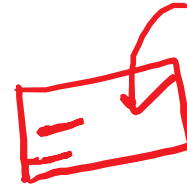
Case 1

Consider you are working as a data scientist for a company ABC and Samsung is the client of ABC company. When Samsung launches a Galaxy s9, you conducted a survey of that product and you collected a million feed-backs. And following diagram represents the distribution of the product attribute.



When you presented this data to your client, he is very happy and he wants to know for each attribute how many of them are feeling good and how many of them are feeling bad. Let's say while taking the feedback you haven't collected the data about whether they are feeling good or bad. How will you come up with the new dataset that is good representation of the current dataset and it has attribute sentiments?

Case studies

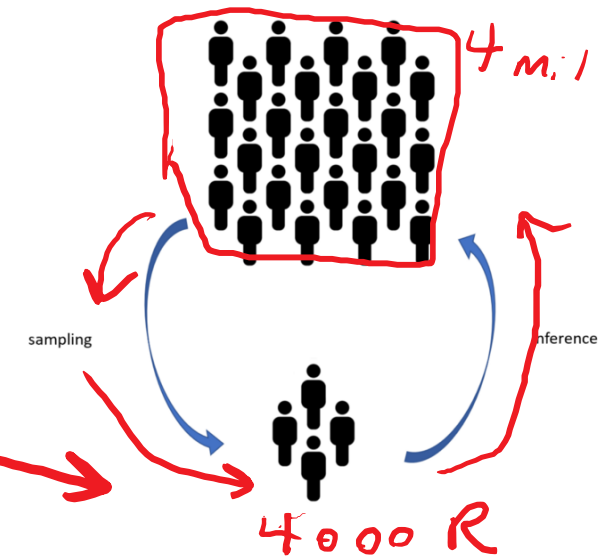


Case 2

The mean engagement time of a particular website is 20 minutes. The owner of the website wants to redesign the website so that people spend even more time on his website. After redesigning the new website he observed that mean engagement time of that website is 25 minutes. How will you find that new design actually helped to improve the user engagement time?

Central Limit theorem

It states that when plotting a sampling distribution of means, the mean of sample means will be equal to the population mean. And the sampling distribution will approach a normal distribution with variance equal to σ^2/n where σ is the standard deviation of population and n is the sample size.



Samples

Population mean

$$1, 1, 2, 3 = \frac{3+2+1+1}{4} = \frac{7}{4}$$

M



sample 1

m_1



sample 2

m_2



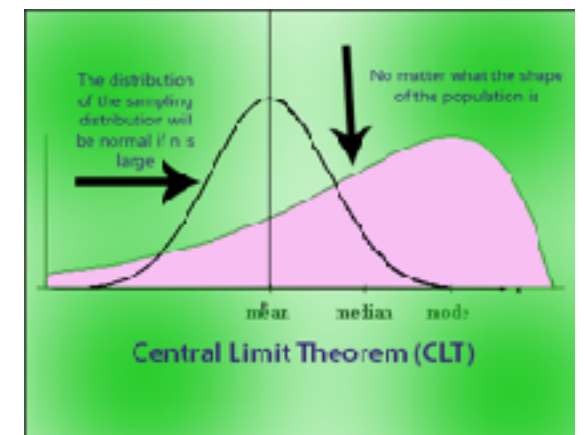
sample 3

m_3



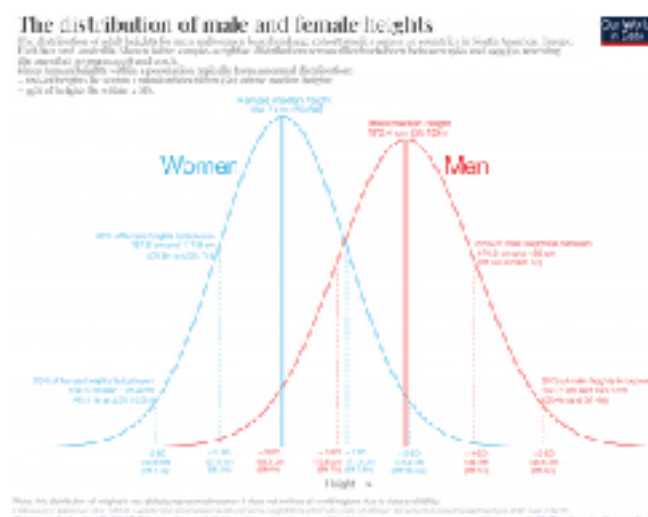
sample 4

m_4



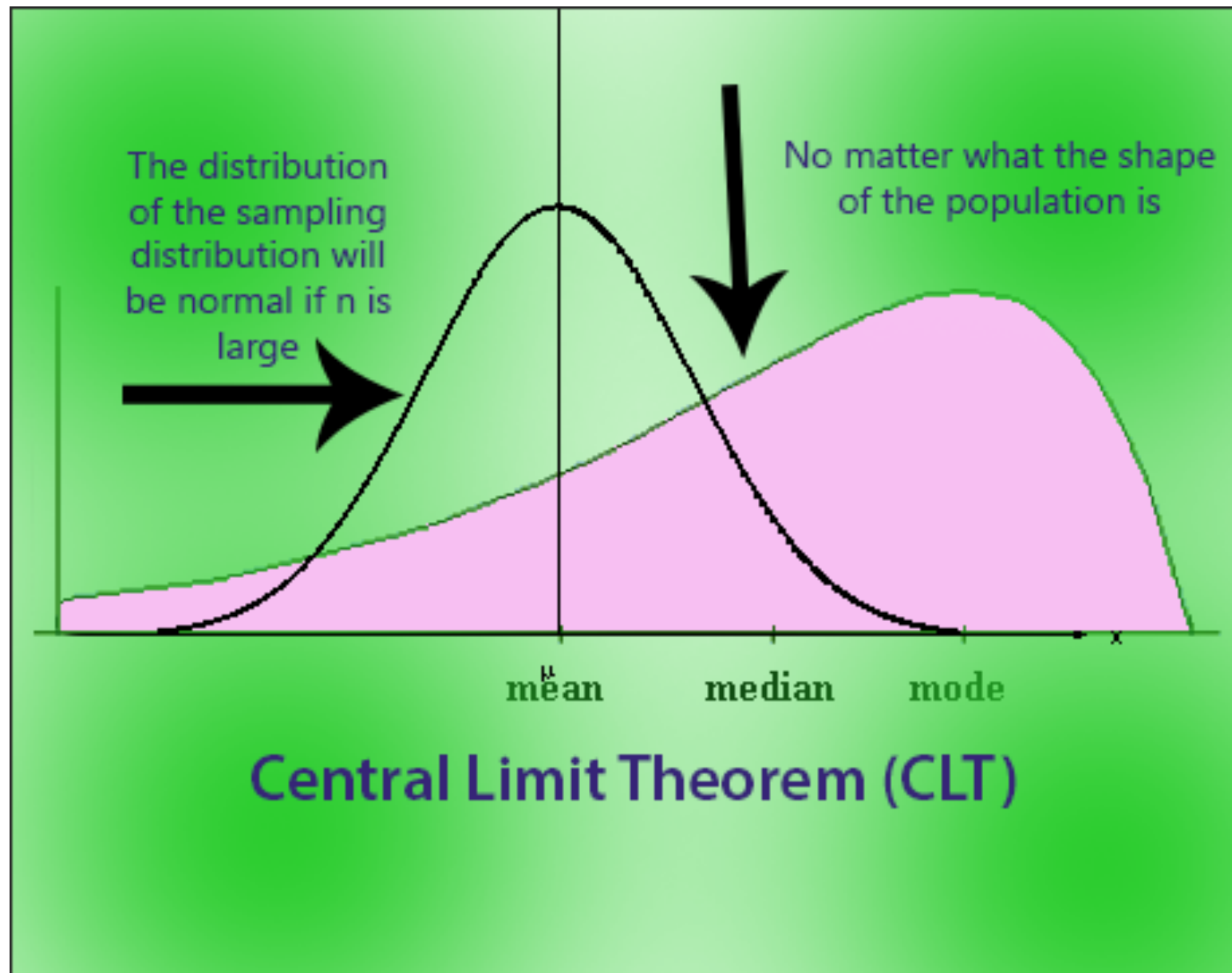
Mean of mean

$$\frac{\sum (m_1 + m_2 + m_3 + m_4)}{4} = M$$



Proving Central Limit theorem

Click the image to visit the notebook



Confidence interval

The confidence interval is a type of interval estimate from the sampling distribution which gives a range of values in which the population statistic may lie.

How well the sample is representing the population

Samples



sample 1

n=4



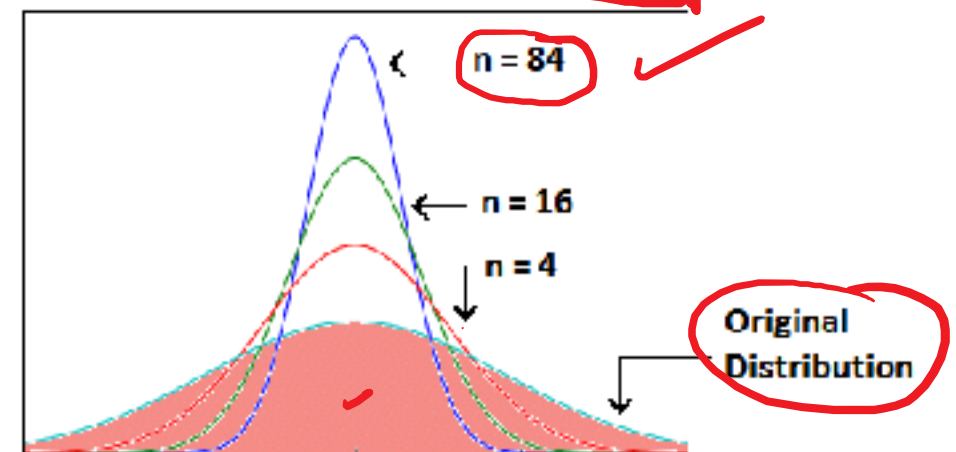
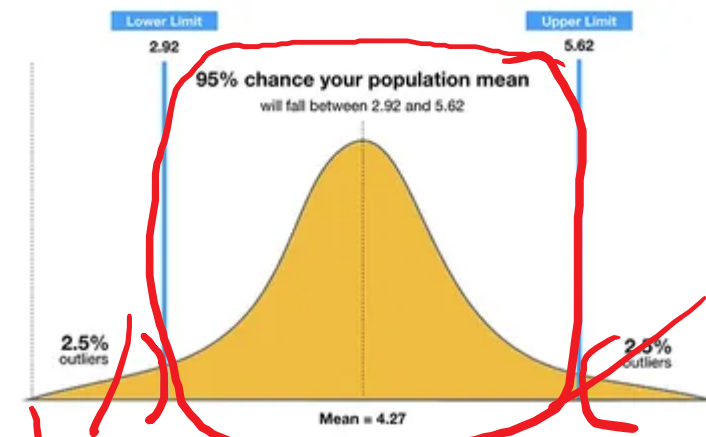
sample 2

n=16



sample 3

n=84



$$\text{C.I} = \bar{X} \pm Z_{\alpha/2} \sigma / \sqrt{n}$$

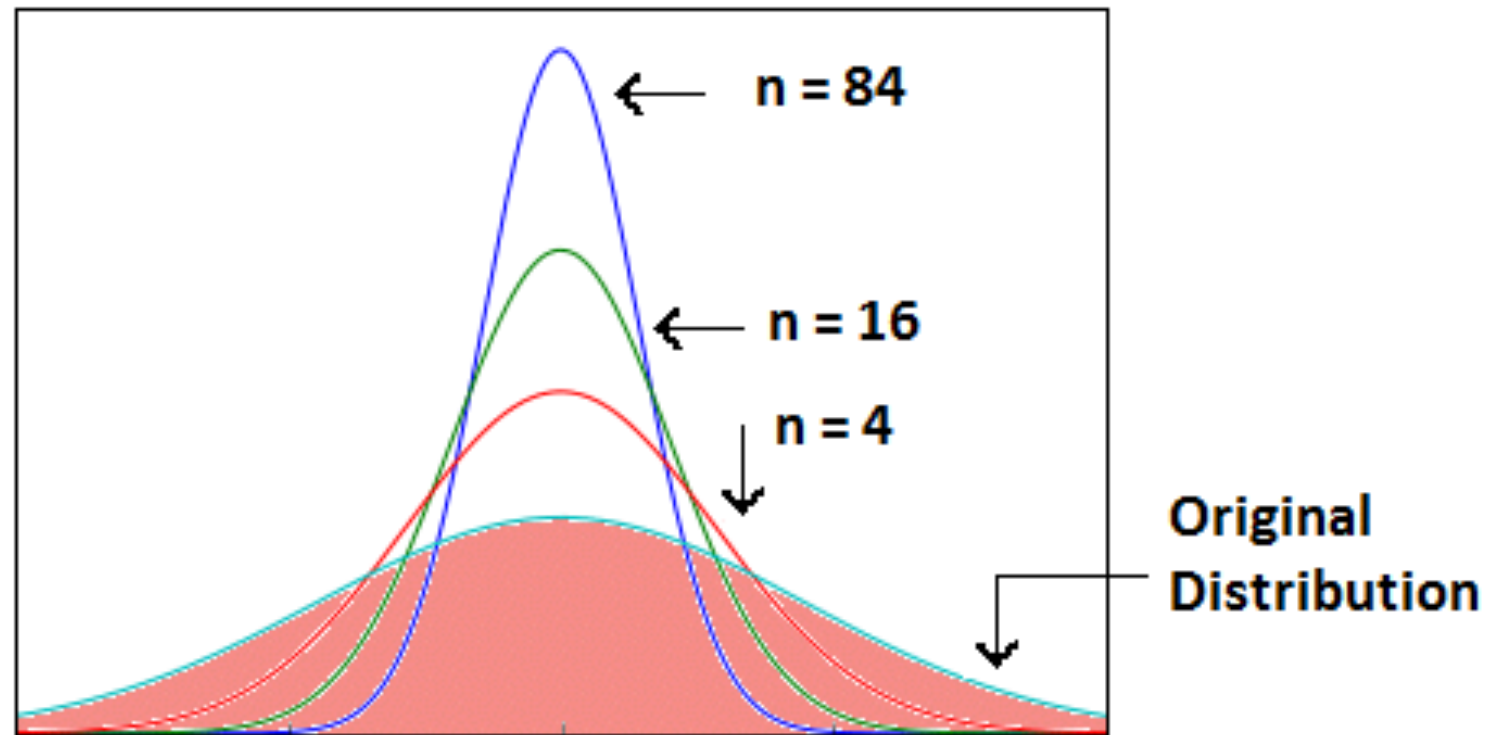
\bar{X} = mean of the sample distribution

$Z_{\alpha/2}$ = Z value for desired confidence level α

σ = standard deviation of the population

Finding confidence interval

Click the image to visit the notebook



Hypothesis testing ✓

Hypothesis testing is used to infer the result of a *hypothesis* performed on sample data from a larger population

✓ **Problem:** There are 4 siblings (A,B,C,D) who have written their names on the sheet and put them into a bowl. Hence there are 4 sheets in the bowl. What is the probability of A not getting picked 4 times in a row?

Hypothesis: Random selection of names from a bowl



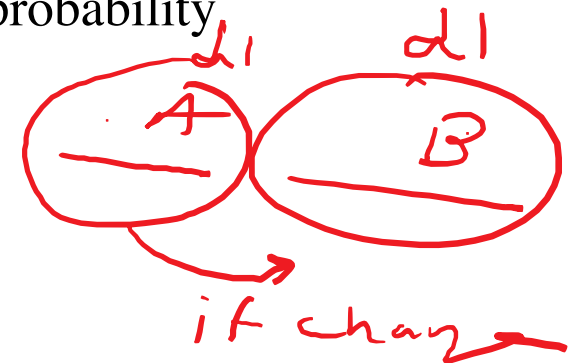
Given:

- ✓ X = number of times not getting picked
- ✓ L = All possible outcomes or Sample space
- ✓ $L = \{ A, B, C, D \}$

Find:

$$P(X=4) = ?$$

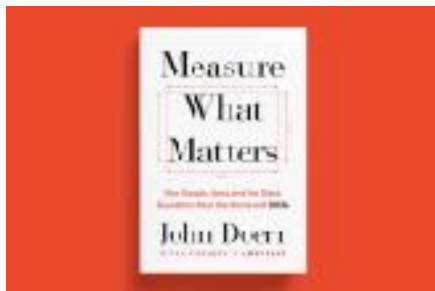
$$P(X = 4)$$



Hypothesis testing: null hyp.

User Engagement on a website before and after changing the design

web → data 1



Old design

Mean user engagement time = 20min

web → data 2



New design

Mean user engagement time = 25min

1. Let's say.

1. H_0 : new mean user engagement time = old mean user engagement time
2. H_a : new mean user engagement time \neq old mean user engagement time

2. Significance Level = 5%

3. Take sample engagement time for new design

1. $N = 100$
2. Mean = 25 minutes

4. p-value = $p(X \geq 25 \text{ minutes} \mid H_0: \text{True})$

5. P-value < Significance Level : Reject H_0

P-value > Significance Level : Do not Reject H_0

not

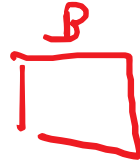
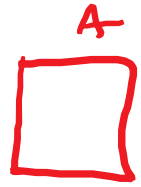
20/25

20

\neq $>$ $<$

reject
change

less than 0.05



H_0 reject
 H_1 change accepted

P-value less
~~0.15~~
0.05
0.04

Thank you