Inferential Statistics

1. Probability
    A. Random Variable
        - Discrete Random variable — countable and have an exact number - filped coin
        - Continuous Random variable — not countable, predicted weather
    B. Probability Distributions
    C. Binomial Random variable — flied coin with 2 variables
    D. Geometric Random Variable — not countable veriable

2. Inference statistics
    A. Central Limit theorem
    B. Confidence Interval
    C. Hypohesis 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

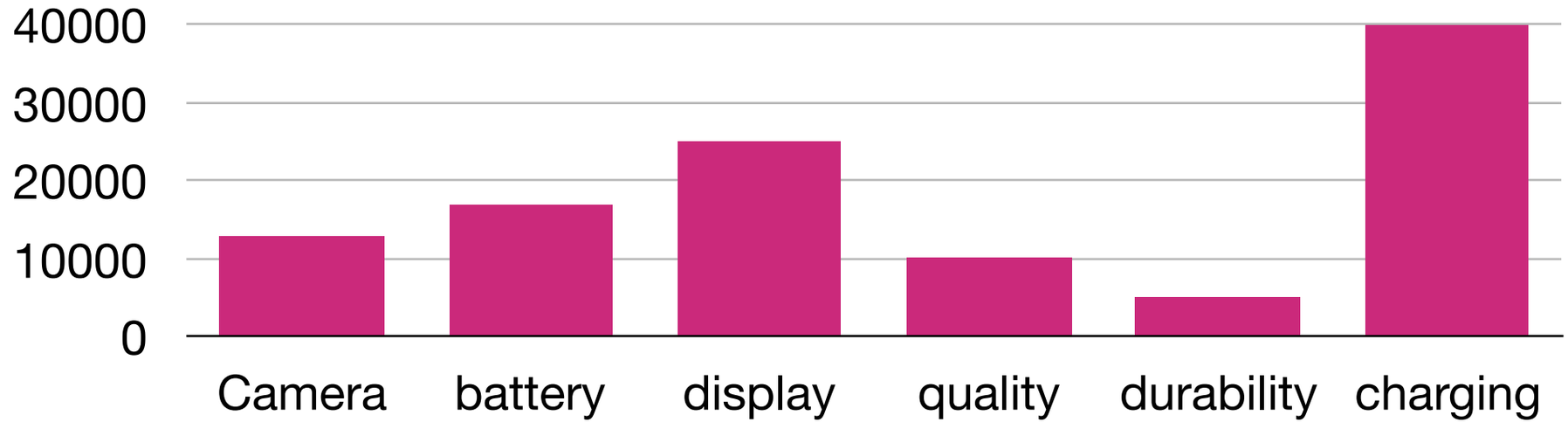
1. Using sample data to make an inference or draw a conclusion of the population.

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

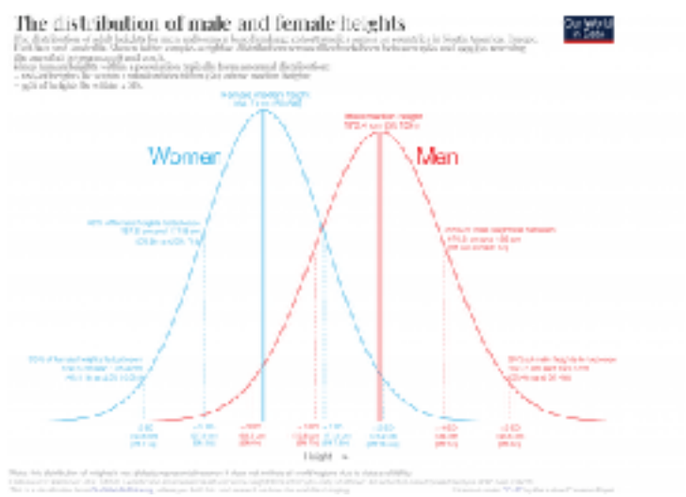
The mean engagement time of a particular website is 20 minutes. The owner of the website wants to redesign the website so that people spend even more time on his website. After redesigning the new website he observed that mean engagement time of that website is 25 minutes. How will you find that new 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 $\sigma/\sqrt{n}$ where σ is the standard deviation of population and n is the sample size.

mean, max, SD

Population



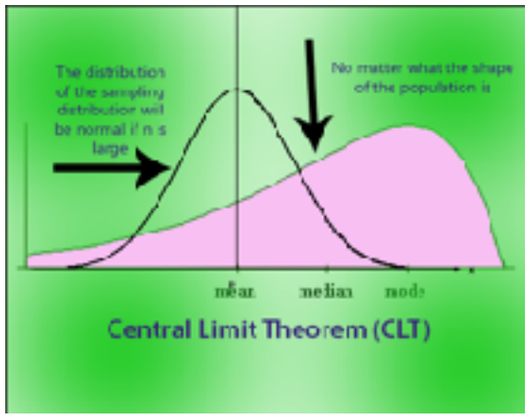
Samples

sample 1 SD1

sample 2 SD2

SD3 sample 3

sample 4

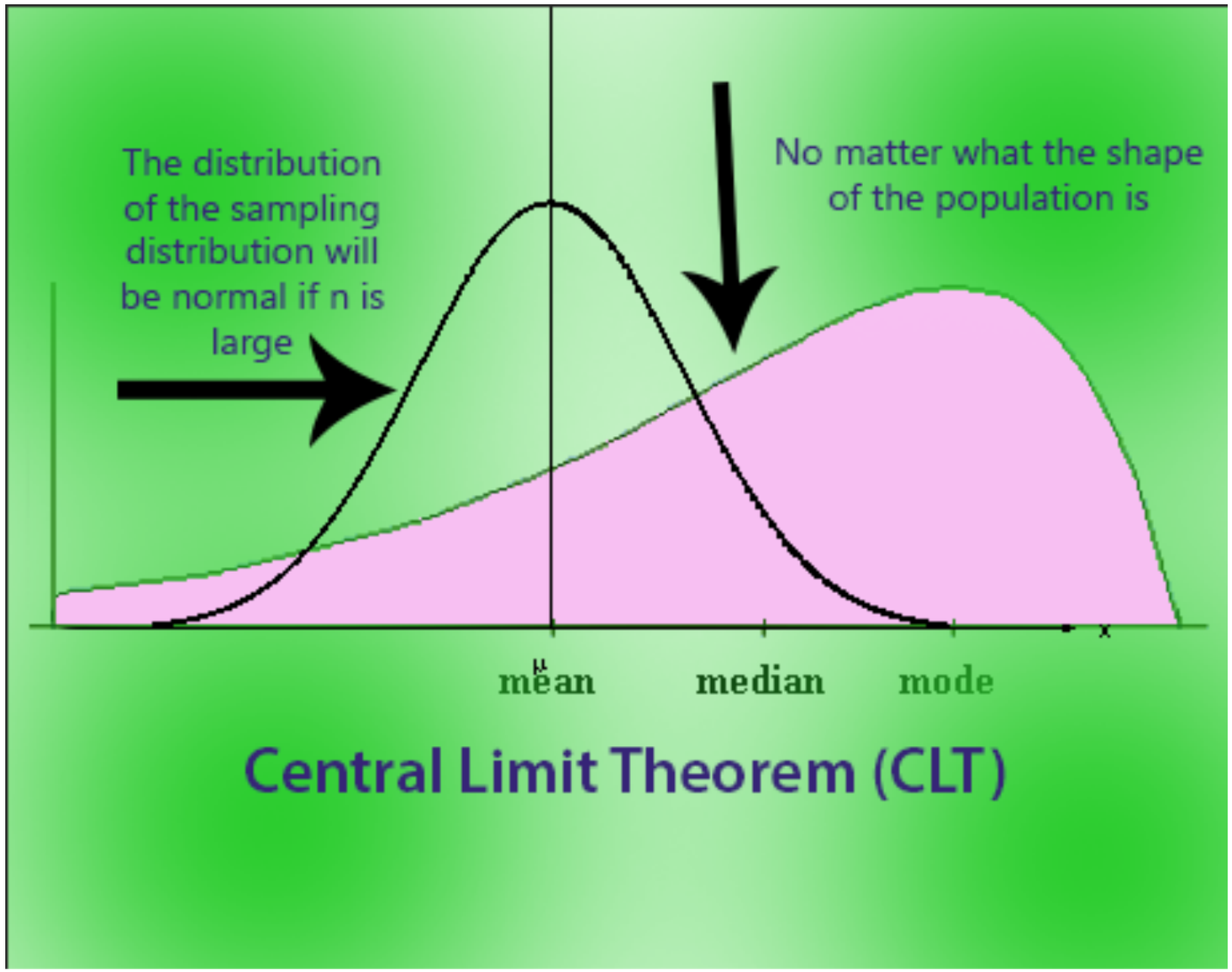


we should average all samples.
will be entire population.
AVG(SD1 + SD2+SD3 + SD4)

SD4

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

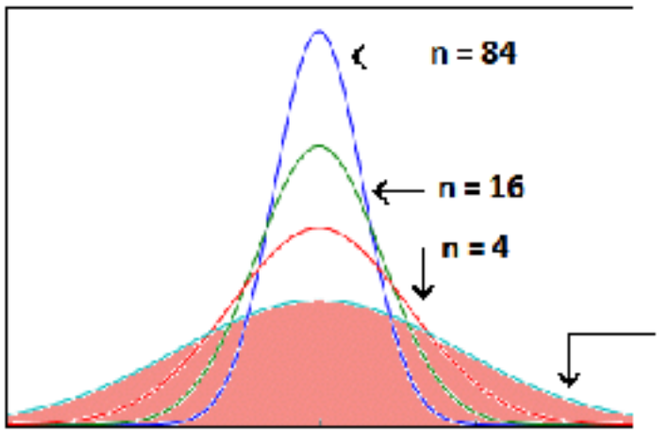
n=4

sample 2

n=16

sample 3

n=84



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

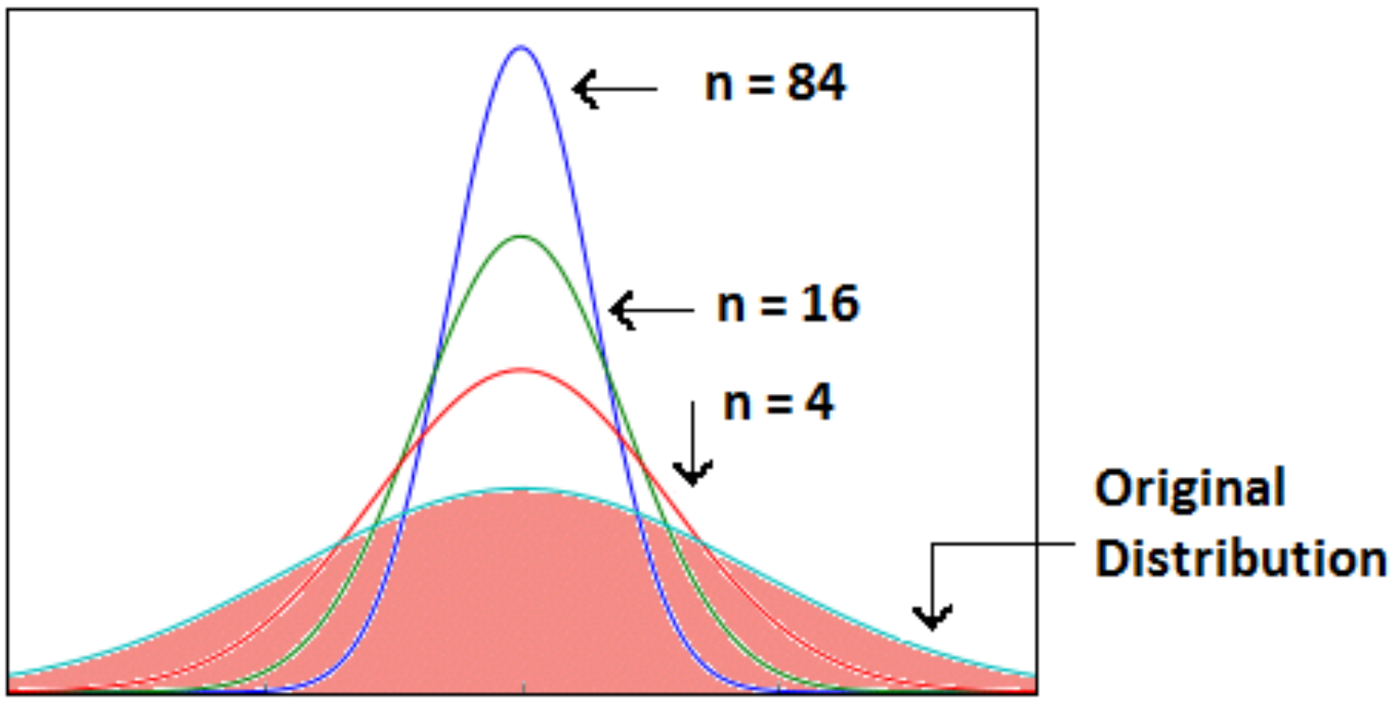
$\bar{X}$ = mean of the sample distribution

$Z_{a/2}$ = 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 there 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:

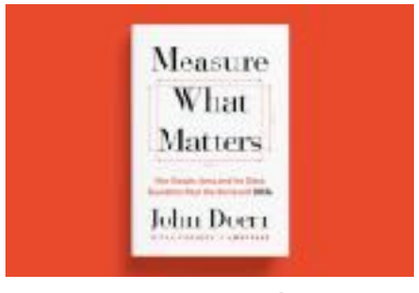
X = number of times not getting picked
L = All possible outcomes or Sample space
L = { A, B, C, D}

Find:

P(X=4) = ?

Hypothesis testing: null hyp.

User Engagement on a website before and after changing the design



Old design

Mean user engagement time = 20min

New design

Mean user engagement time = 25min

1. Let's say,
    1. H_o : new mean user engagement time = old mean user engagement time
    2. H_a : new mean user engagement time != 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>=25 minutes | H_o :True)

5. P-value < Significance Level : Reject H_o
   P-value < Significance Level : Do not Reject H_o

Thank you