Inferential Statistics for Data Science

Outline

- Motivation
- Population and Sample
- Sampling and its types
- Inferential Statistics
- Sampling Distribution and Central Limit Theorem
- Estimating Population Mean and Population Proportion

Motivation

- Recall Descriptive Statistics:
  - Describes the characteristics of the dataset
  - Distribution, central tendencies (mean, median, mode) and variability (standard deviation, variance, etc.) are used to describe the given data
- Question: What if we want to make some inferences or predictions from the data which is not fully available or is too large?
- Examples:
  - What is the battery life of a particular mobile model ?
  - What is the average salary of a data scientist in India?
  - What is the most preferred OTT platform for watching movies in India?
- Question: How to make inferences about data which is partially known or is too large to analyse?

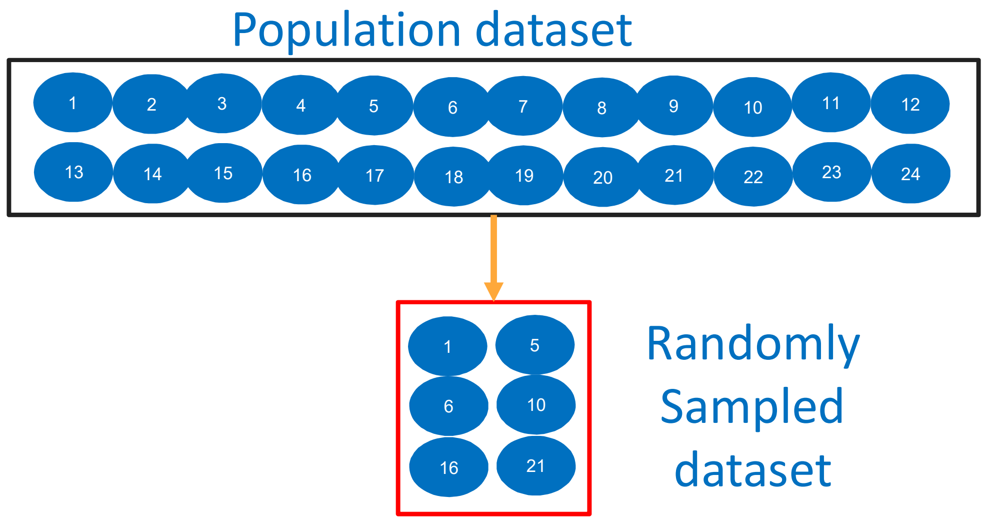
Population and Sample

Population and Sample

- Population:
  - Refers to the whole group or set of data points on which inferences or predictions are to be made
  - Size of the population could be very large depending on the inference to be made
  - Battery life example: Population is the set of all mobiles of that particular model
  - OTT example: Population is the set of all people in India who watch movies on OTT platforms
- Sample:
  - A set of data points which are representative of the population
  - Size of sample set is generally much smaller than the size of population

Obtaining Sample from Population

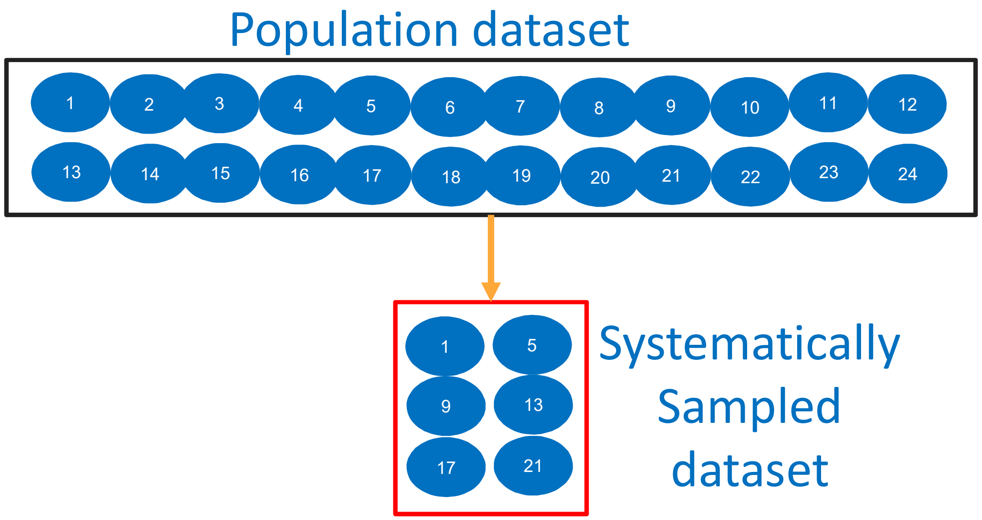
- Sampling is performed to get a sample set from the actual population
- 3 types of Sampling:
    - Random sampling: Each data point of population is picked with equal probability



OTT Example: How would random sampling be done in this case?

Obtaining Sample from Population

- Sampling is performed to get a sample set from the actual population
- 3 types of Sampling:
  - Random sampling: Each data point of population is picked with equal probability
  - Systematic sampling: Every k^{th} data point is picked from the population set

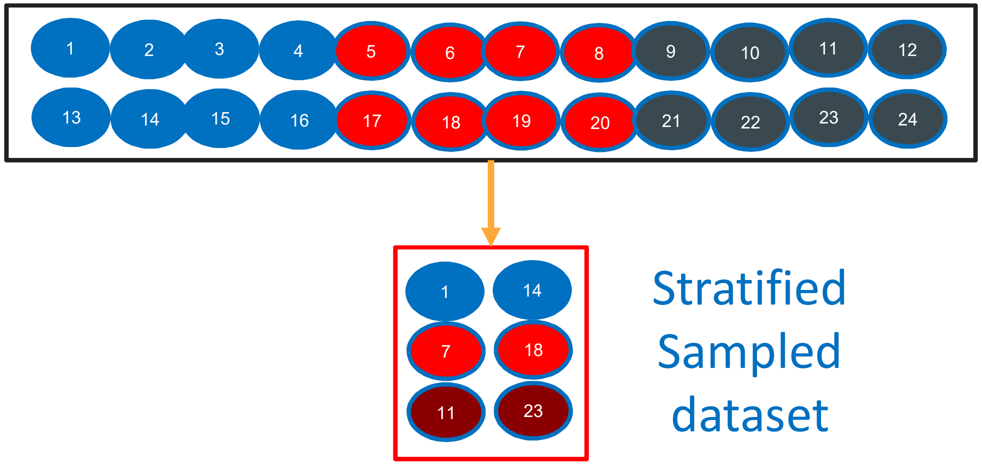


OTT Example: How would systematic sampling be done in this case?

Obtaining Sample from Population

- Sampling is performed to get a sample set from the actual population
- 3 types of Sampling:
  - Random sampling: Each data point of population is picked with equal probability
  - Systematic sampling: Every k^{th} data point is picked from the population set
  - Stratified sampling: Population is divided into subsets (stratums) based on some criteria. Random sampling is performed on each stratum

Stratified Population dataset



Stratified Sampled dataset

OTT Example: How would stratified sampling be done in this case?

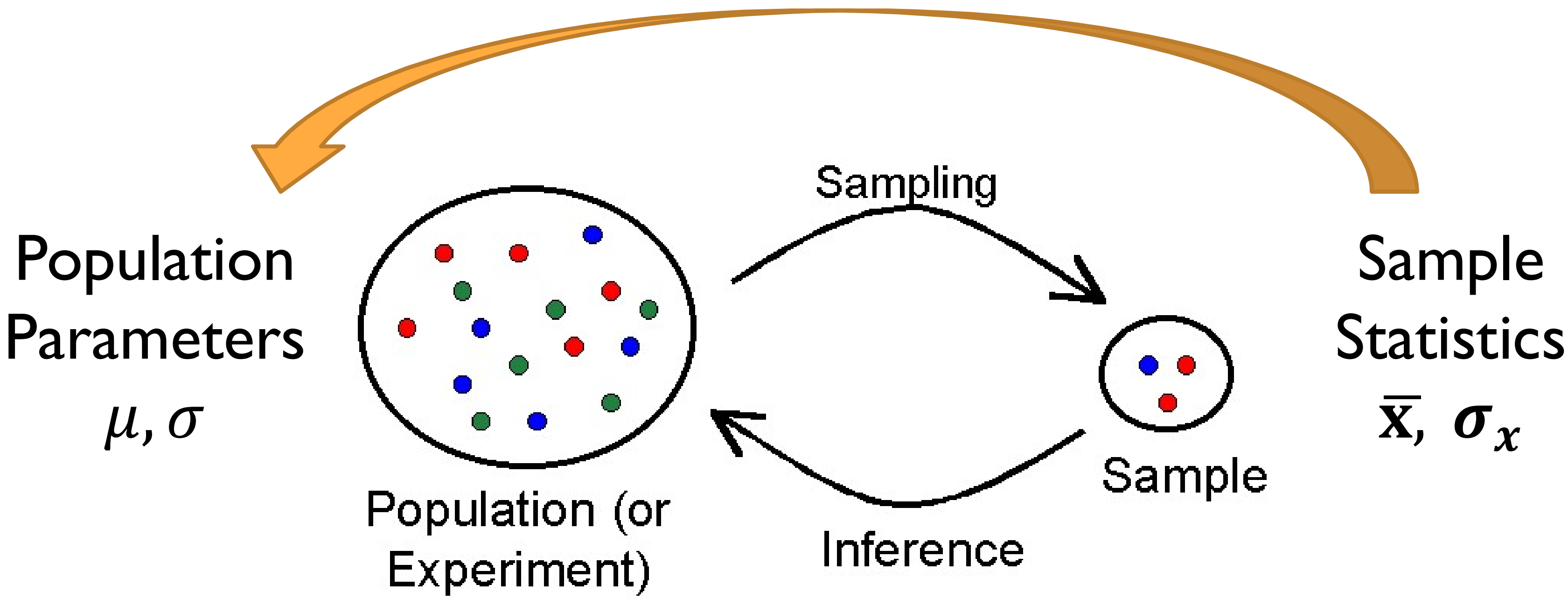
Inferential Statistics

Inferential Statistics

- Makes predictions about the population based on a sample set collected from the population
- Generalises over the population by analysing the sample set
- Comprises of:
  - Estimating parameters: Estimating the parameters (such as mean, standard deviation, etc.) of the population using sample data
    - Example: What is the battery life of a particular mobile model ? – Mean battery of all mobiles of that particular model
  - Hypothesis testing: Testing a claim on a parameter or distribution of the population (hypothesis) using the sample data
    - Example: 'Hotstar' is preferred by more than 50% OTT users in India. How to test this claim?
- Note: Parameter of a sample (such as mean, standard deviation, etc.) is referred to as 'statistic'

Inferential Statistics

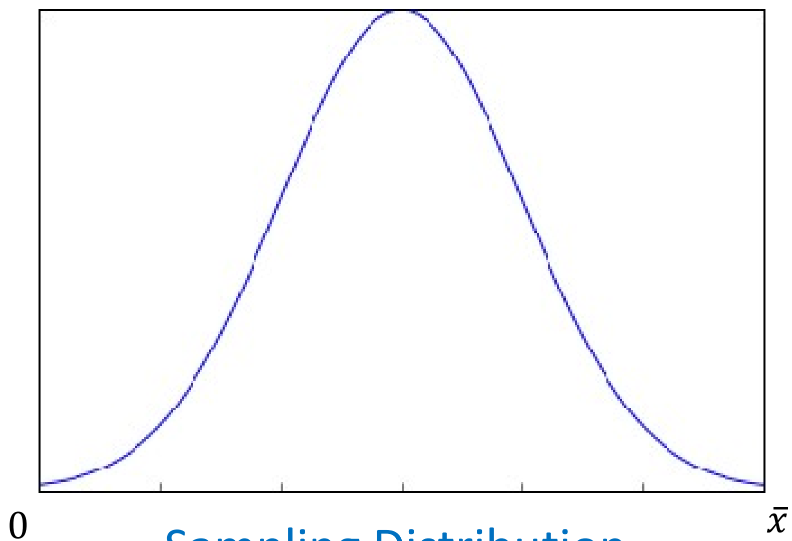
- Makes predictions about the population based on a sample set collected from the population



Sampling Distribution
and
Central Limit Theorem

Sampling Distribution

- Suppose multiple samples (sets) are sampled from a population of size N
- Sampling distribution is the probability distribution of a particular statistic computed using each of the sample sets
- Example: Suppose population size is N and many sample sets $(x_1, x_2, \ldots)$ of size n are drawn from the population
- Mean of each sample set is computed $(\text{say } \overline{x_1}, \bar{x}_2 \ldots)$ and plotted as a distribution
- Note: Here, variable takes numerical values



Sampling Distribution of Sample Means

Parameters of the Sampling Distribution

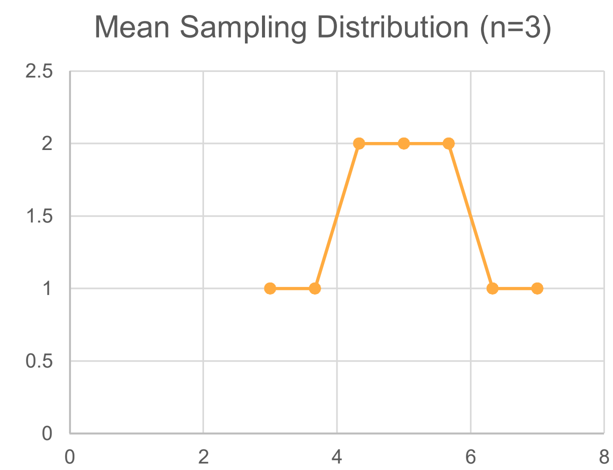
- Mean of the sampling distribution of sample means: $\mu_{\bar{x}}$
- Standard deviation of the sampling distribution (referred to as Standard error): $\sigma_{\bar{x}}$
- Sampling distribution mean and standard error have special properties in relation to population mean (μ) and standard deviation (σ)
- Central Limit Theorem describes the relation between μ & $\mu_{\bar{x}}$ and σ & $\sigma_{\bar{x}}$

Example of Mean Sampling Distribution

- Population set: $\{1,3,5,7,9\}; N = 5; \mu = 5; \sigma = 2.83$
- Consider multiple samples of size 3 i.e., $n = 3$

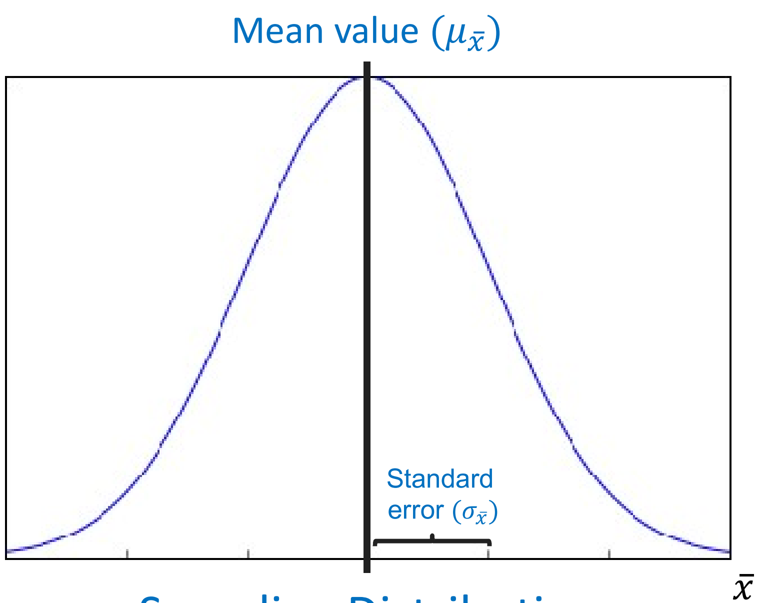
Samples			Sample Mean $\bar{x}$
1	3	5	3.00
1	3	7	3.67
1	3	9	4.33
1	5	7	4.33
1	5	9	5.00
1	7	9	5.67
3	5	7	5.00
3	5	9	5.67
3	7	9	6.33
5	7	9	7.00

Mean ($\mu_{\bar{x}}$)	5
Standard distribution ($\sigma_{\bar{x}}$)	1.21



Central Limit Theorem (CLT)

- Proves that:
  1. Sampling distribution of means approaches a normal distribution as the sample size (n) increases, irrespective of the distribution of the population
  2. Mean of the sampling distribution is equal to the mean of the population distribution i.e., $\mu = \mu_{\bar{x}}$
  3. Standard error is related to the standard deviation of population distribution as follows: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

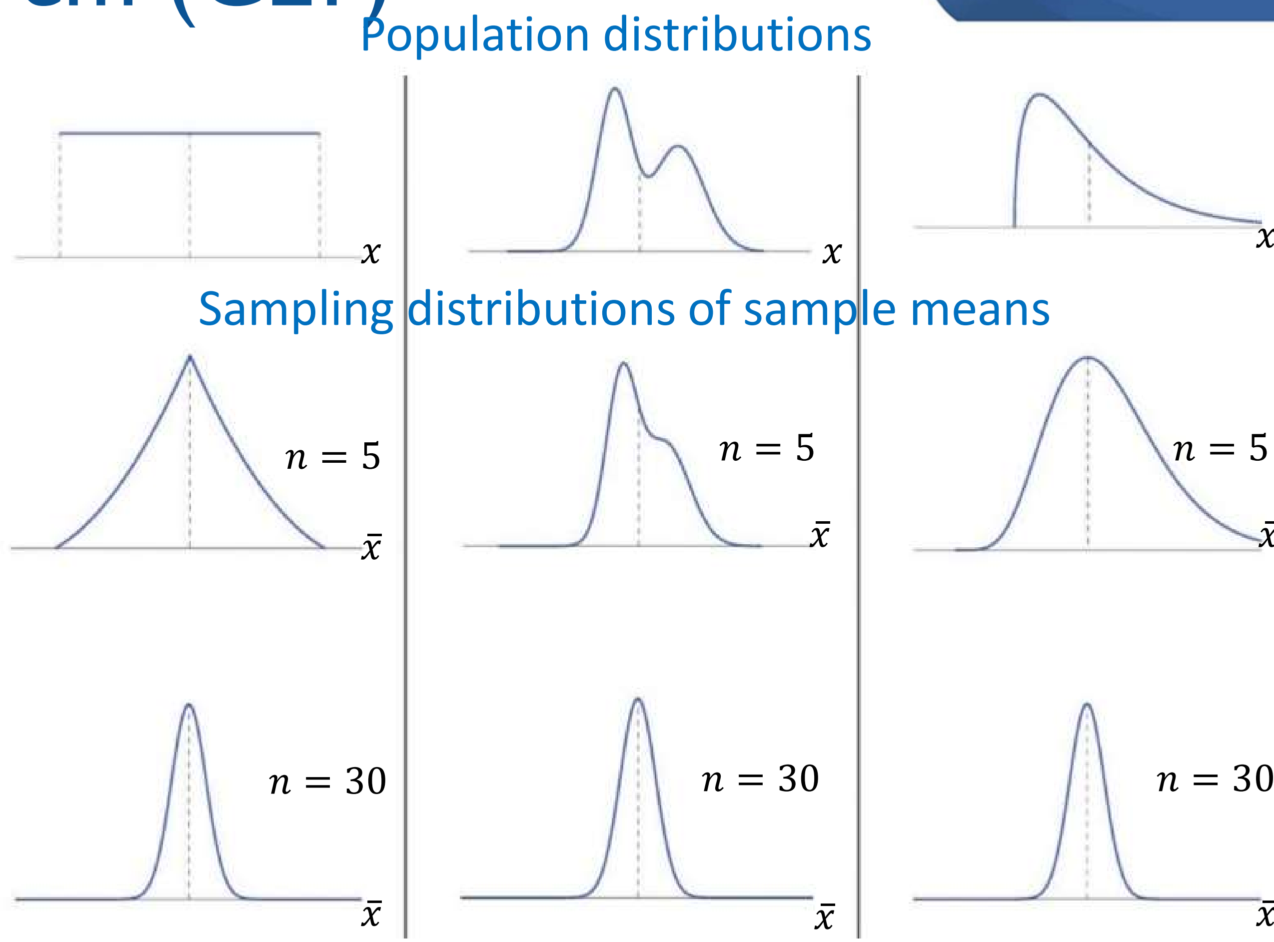


Sampling Distribution of Sample Means

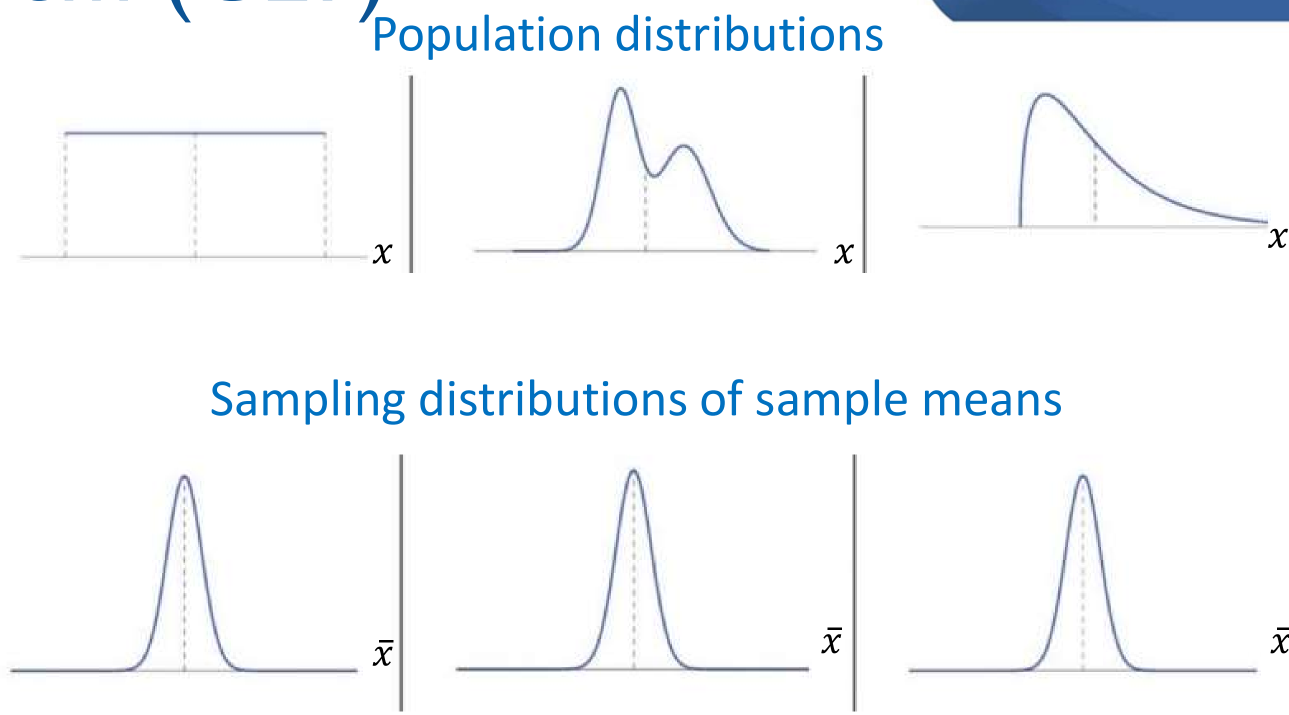
Central Limit Theorem (CLT)

Point 1:

- Sampling distribution of means approaches a normal distribution as the sample size (n) increases
- Shape of the population distribution does not matter
- Generally normal distribution is observed when $n \geq 30$
- Note: If population distribution is normal, then value of n does not matter



Central Limit Theorem (CLT)



Point 2:

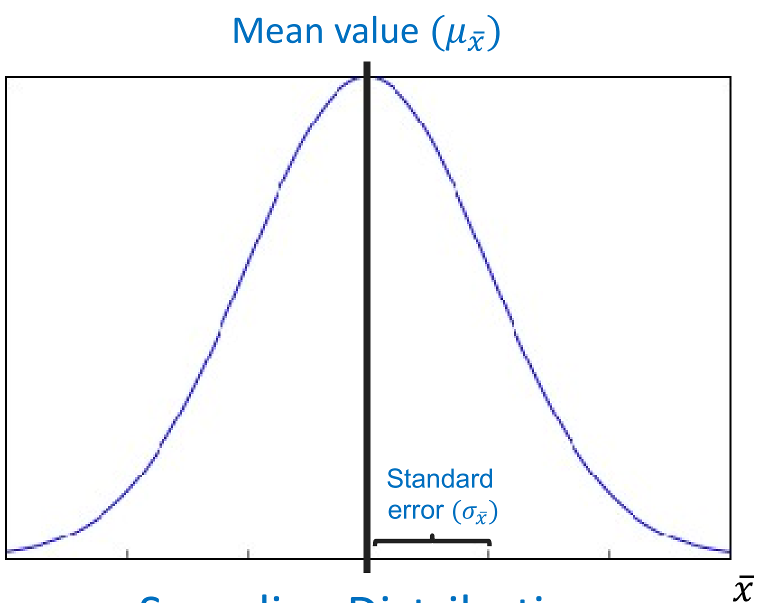
- Mean of the sampling distribution is equal to the mean of the population distribution i.e., $\mu = \mu_{\bar{x}}$
- Shape of the population distribution does not matter

Central Limit Theorem (CLT)

Point 3:

- Standard error is related to the standard deviation of population distribution as follows: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- Standard error decreases as the sample size n increases
- Note: Standard error becomes zero when $n = N$

What is the use of CLT?

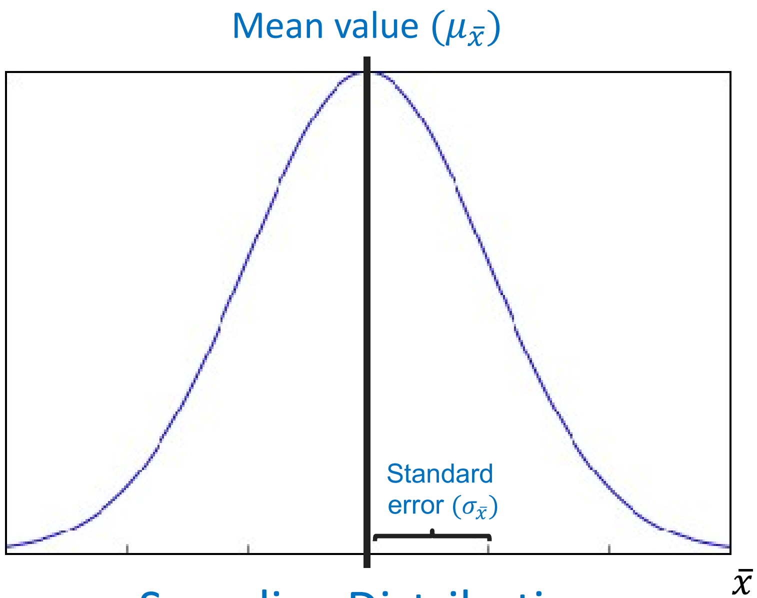


Sampling Distribution of Sample Means

Estimating Population Parameters: Mean

Estimating Mean of Population

- Question: Given a sample x, how to estimate the mean of the population?
- Mean of the sample $\bar{x}$ lies somewhere on the sampling distribution
- $\bar{x}$ is most likely close to the mean of sampling distribution ($\mu_{\bar{x}}$) because it is a normal distribution as per CLT
- Therefore, $\bar{x}$ can be considered to be an estimate of μ since $\mu = \mu_{\bar{x}}$ as per CLT
- Question: How good is the estimate? What is the margin of error for the estimate?



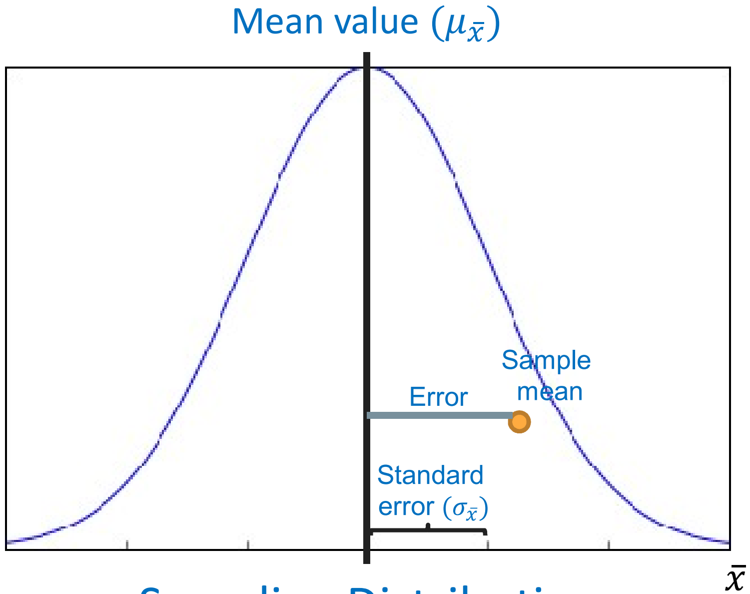
Sampling Distribution of Sample Means

Margin of Error

- Margin of error: Range of possible error between the sample mean and population mean (mean of sampling distribution)

$$\mu = \bar{x} \pm \epsilon$$

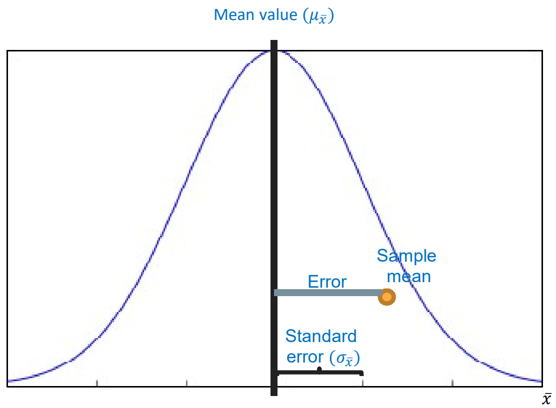
- Question: How to find epsilon?
- Idea: Use the standard error $\sigma_{\bar{x}}$ to quantify the margin of error
- For a normal distribution,
  - 68.3% of data falls within 1 standard deviation of the mean
  - 95.4% of data falls within 2 standard deviations of the mean
  - 99.7% of data falls within 3 standard deviations of the mean



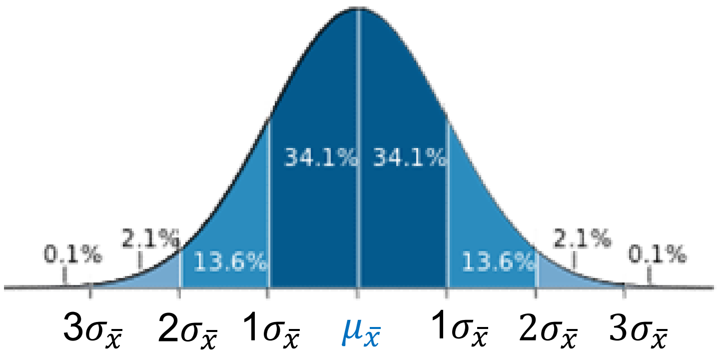
Sampling Distribution of Sample Means

Margin of Error and Probability

- Margin of error: $\mu = \bar{x} \pm \epsilon$
- In the mean sampling distribution,
  - 68.3% of data falls within $1\sigma_{\bar{x}}$
  - 95.4% of data falls within $2\sigma_{\bar{x}}$
  - 99.7% of data falls within $3\sigma_{\bar{x}}$
- Example: Suppose a sample mean is calculated to be 10 and standard error $\sigma_{\bar{x}} = 0.5$
- Implies the following:
  - $\mu = 10 \pm 0.5$ with a probability of 68.3%
  - $\mu = 10 \pm 1$ with a probability of 95.4%
  - $\mu = 10 \pm 1.5$ with a probability of 99.7%

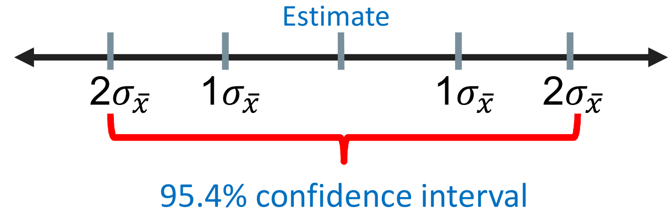


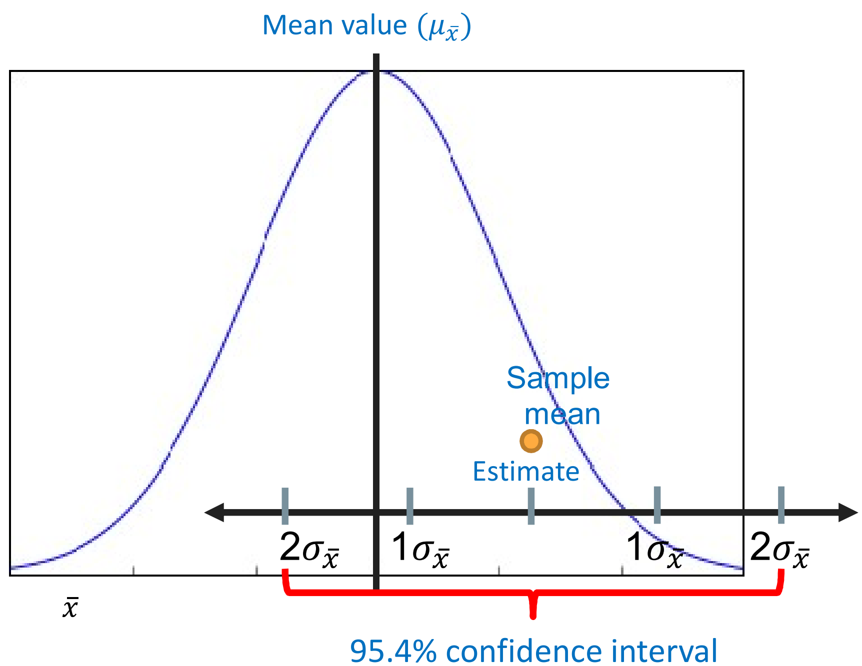
Sampling Distribution of Sample Means



Confidence Levels and Intervals

- **Confidence level:** Probability that the population parameter lies within an error margin of the sample statistic
- **Confidence interval:** Range in which the population parameter could lie with a given confidence level
- **Example:** $\bar{x} = 10$ and $\sigma_{\bar{x}} = 0.5$
- Implies the following:
  - Confidence interval of $\mu = (9.5, 10.5)$ with confidence level of 68.3%
  - Confidence interval of $\mu = (9, 11)$ with confidence level of 95.4%
  - Confidence interval of $\mu = (8.5, 11.5)$ with confidence level of 99.7%

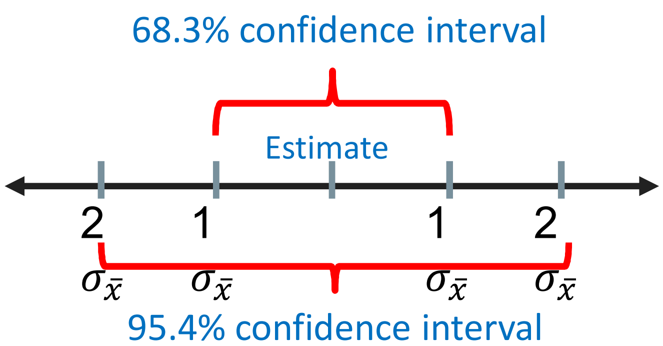




Confidence Levels and Intervals

- Question: How to determine the confidence interval for a particular confidence level?
- Idea: Use z-score corresponding to the confidence level
- Confidence interval of $\mu = (\bar{x} - z\sigma_{\bar{x}}, \bar{x} + z\sigma_{\bar{x}})$
- Z scores for some confidence intervals:

Confidence Level	Z-score
90%	1.65
95%	1.96
98%	2.33



Estimating Mean of Population: Summary

- Steps to estimate the mean of the population given a representative sample of size n:
    1. Calculate the mean of the sample $\bar{x}$ and the standard error $\sigma_{\bar{x}}$
    2. Decide the confidence level with which the population mean is to be estimated
    3. Take the z-score corresponding to the chosen confidence level $(\%C)$
    4. Compute the confidence interval of the population mean
    $$\mu = (\bar{x} - z\sigma_{\bar{x}}, \bar{x} + z\sigma_{\bar{x}})$$
    5. Conclude that population mean lies in the range $(\bar{x} - z\sigma_{\bar{x}}, \bar{x} + z\sigma_{\bar{x}})$ with a confidence level (probability) of $\%C$

Estimating Mean of Population: Points to Note

- Question: What to do if the range of the confidence interval is to be reduced?
  - Value of $\sigma_{\bar{x}}$ is to be reduced
  - Increase the sample size n because $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- Question: What if the population standard deviation (σ) is not known?
  - Standard deviation of the sample σ_x is taken as an estimate of σ
  - For a reasonable estimate of σ, $n \geq 30$ is recommended
  - Standard error is estimated to be $\frac{\sigma_x}{\sqrt{n}}$
  - Z-score for interval calculation is replaced by t-score
  - Note: While Z-score is fixed for a given confidence level, t-score is dependent on the sample size n

Estimating Population Parameters: Proportion

Estimating Proportion of Population

- Proportion is considered when the variable is a categorical variable
- Example: For how many people in India, 'Hotstar' is the preferred OTT platform?
- Suppose 10 crore people in India subscribe to OTT platforms
- Question: How to estimate the proportion of 'Hotstar' preferred users?
- Proven result: Sampling distribution of proportions approaches a normal distribution as sample size increases
- Note: Proportions have binomial distribution

Population of OTT users in India

Preferred OTT	# People	Proportion of population
Hotstar	3.5 Crore	0.35
Prime	2.5 Crore	0.25
Netflix	0.5 Crore	0.05
Zee5	2 Crore	0.2
Sonyliv	1.5 Crore	0.15
Total	**10 Crore**	**1**

$$\text{Proportion, } p = \frac{\#Items\ in\ Category}{Population\ size} = \frac{c}{N}$$

Estimating Proportion of Population: Summary

- Steps to estimate the proportion of a category of the population given a representative sample of size n:
  1. Calculate the proportion of the category in the given sample:
  $$p_x = \frac{c_x}{n}$$
  2. Estimate the standard error of sampling distribution of proportions:
  $$\sigma_p = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{p_x(1-p_x)}{n}}$$
  3. Decide the confidence level with which the proportion is to be estimated
  4. Take the z-score or t-score corresponding to the chosen confidence level $(\%C)$
  5. Compute the confidence interval of the population mean at $\%C$
  $$p = \left(p_x - z\sigma_p, p_x + z\sigma_p\right) \text{ or } p = \left(p_x - t\sigma_p, p_x + t\sigma_p\right)$$

Summary

- Inferential statistics is the study of techniques which are used to makes inferences about the population using a representative sample
- Sample can be obtained by using different sampling techniques
- Central limit theorem relates the population parameters with sample statistics through the sampling distribution
- CLT can be used to estimate population mean from a sample mean with certain level of confidence
- Confidence interval gives the range in which population mean lies with certain probability
- For categorical variables, the confidence interval of proportion of a category in the population can be estimated with certain probability

peration == "MIRROR_X":
irror_mod.use_x = True
irror_mod.use_y = False
irror_mod.use_z = False
_operation == "MIRROR_Y":
irror_mod.use_x = False
irror_mod.use_y = True
irror_mod.use_z = False
_operation == "MIRROR_Z":
irror_mod.use_x = False
irror_mod.use_y = False
irror_mod.use_z = True

THANK YOU