

Statistical Techniques for Data Science

Sampling Estimation

Dr. Subhabaha Pal
Manipal Global Academy of Data Science

Objective

After attending this session, you will be able to –

- **Explain Sampling Variation**
- **Describe Sampling Distribution**
- **Define Point Estimate, Interval Estimate and Confidence Interval**

Sampling Variation

- In the last section, we have learnt about sampling and different sampling techniques
- In this section, we will learn about sampling variation and sampling distribution
- Sampling variation is the variation in sample estimates, even if the samples are drawn from the same population
- Example – Below data states number of hours flight of 100 pilots of an airlines.

1861	2495	1000	2497	1865	791	2090	2637	1327	1678
1680	2858	795	2495	2496	2501	1160	1480	1860	2490
2090	2840	2490	2640	659	827	2646	2638	2643	868
1327	1866	1861	2486	2865	3011	2494	1489	1865	2855
2840	2499	2093	2660	1165	2600	2085	2640	2998	1861
2956	2495	2865	1865	3000	3019	1670	2858	2642	1680
3038	3000	1313	596	656	3240	590	2501	2485	3015
2092	1679	3024	2497	2825	2630	2070	2900	1861	2636
2495	2637	2497	1159	2640	3050	870	2896	2500	2638
926	2860	1481	875	2482	1860	2086	934	3200	2490

Sampling Variation Example

- 10 samples of different sample sizes are drawn randomly from the population of flight hours records of 100 pilots

Sample 1

3000 2486 820 1678 2070 2638 2490 1865 1000 2090 596 3200

Sample 2

2840 2858 3000 2490 2998 3050 2070 2896 3200 2490 3280

Sample 3

2858 3240 2497 2865 656 2093 934 1861 868 795

Sample 4

2086 1000 2497 596 656 875 2085 934 1313

Sample 5

820 1313 3000 2640 596 2640 2600 2495 934 2500

Sample 6

2840 2499 1327 1861 2495 3024 3038 2497

Sample 7

2858 2490 868 1670 1480 2643 1480 1680 2085 2490

Sample 8

2495 2858 1861 2092 2499 3000 2660 1000 1679 926 2660

Sample 9

795 791 3200 2085 2638 2497 2486 1159 2640

Sample 10

3019 3240 3200 3050 3000 3015 2900 2896 2998

Sampling Variation Example

- Examining the 10 samples, we get results in the right hand table
- The variation in the sample statistics even when the samples are drawn from the same population is termed as Sampling Variation

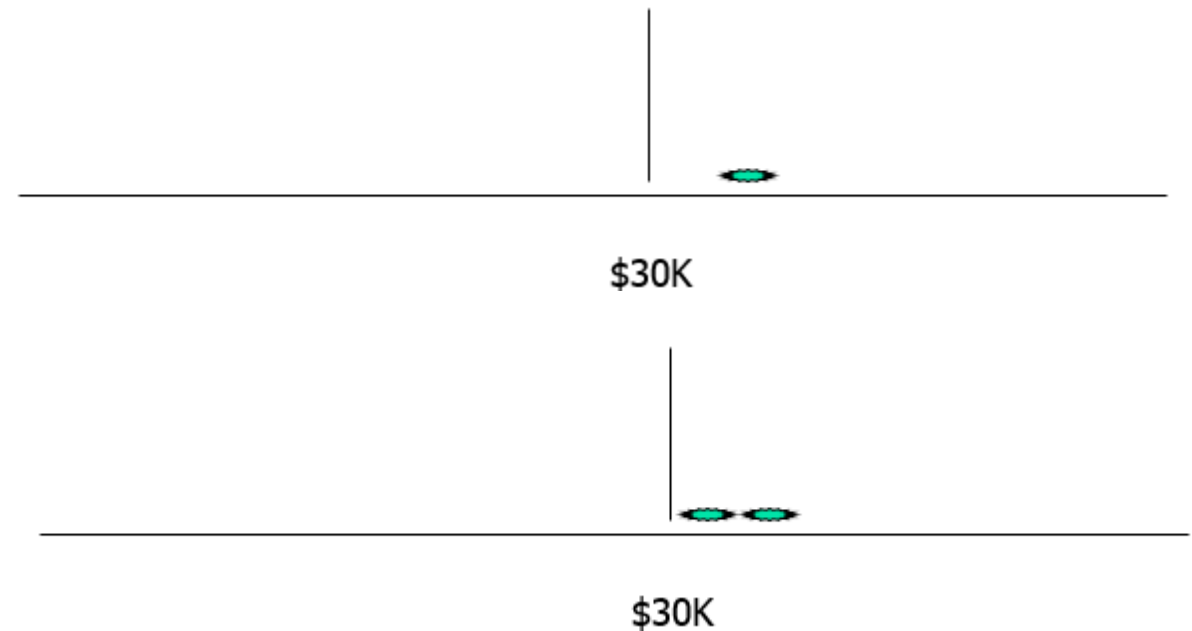
Sample No.	Sample size	Mean	SD
1	12	1994.42	843.23
2	11	2830.18	349.94
3	10	1866.70	988.57
4	9	1338.00	704.36
5	10	1953.80	920.44
6	8	2447.63	590.64
7	10	1974.40	638.05
8	11	2157.27	715.10
9	9	2032.33	891.53
10	9	3035.33	117.40
Overall	100	2162.24	732.26

Sampling Distribution

- **Sampling distributions constitute the theoretical basis of statistical inference and are of considerable importance in business decision-making**
- **If numerous different samples of equal size from the same population are taken, the probability distribution of all possible values of a given statistic from all the distinct possible samples of equal size is called a sampling distribution**
- **The sampling distribution depends on the underlying distribution of the population, the statistic being considered, the sampling procedure employed and the sample size used**
- **A Sampling distribution is the distribution of statistic that would be produced in repeated random sampling (with replacement) from the same population**
- **Sampling distributions are used to calculate the probability that sample statistics could have occurred by chance and thus to decide whether something that is true of a sample statistic is also likely to be true of a population parameter**

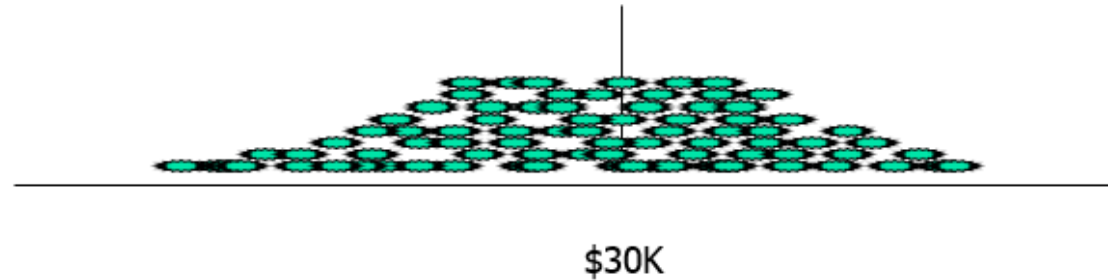
Sampling Distribution - Example

- The following example will help to understand the concept of Sampling distribution
- From the census data, it is revealed that the mean income of US population is 30K USD
- In the 1st attempt, we select 1500 US citizens in a sample and determine their mean income
- We find the mean income to be 42K USD
- We map that in a plot and get the graph on the right hand side
- We draw a second sample of same size and after plotting the mean income 33K USD, we get the graph of the right hand side

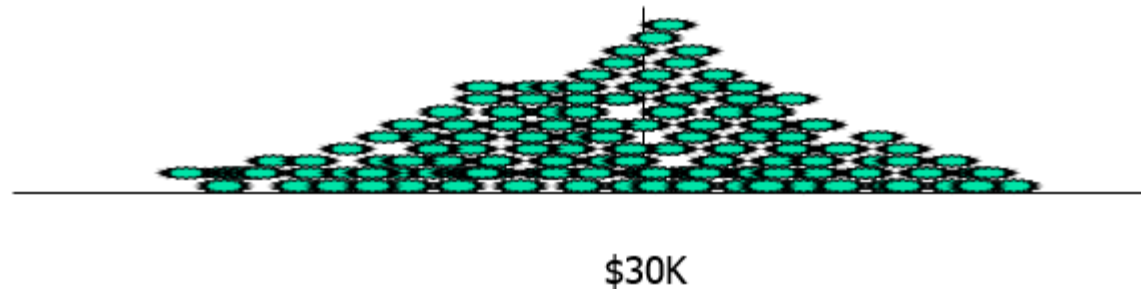


Sampling Distribution - Example

- After carrying out drawing a number of samples, we shall be getting a graph something of the form below

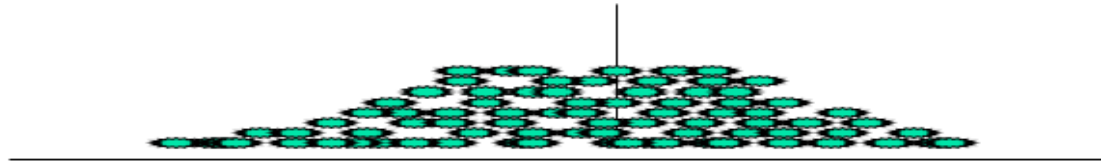


- If we go on continuing the similar actions, we shall be getting graphs like this –

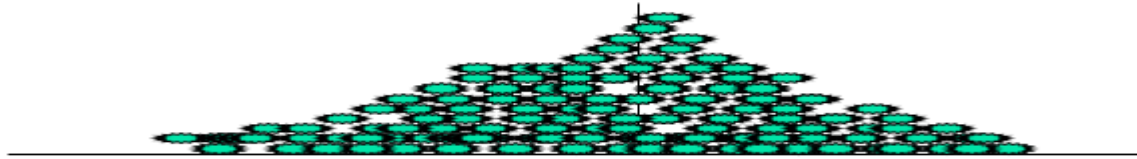


Sampling Distribution - Example

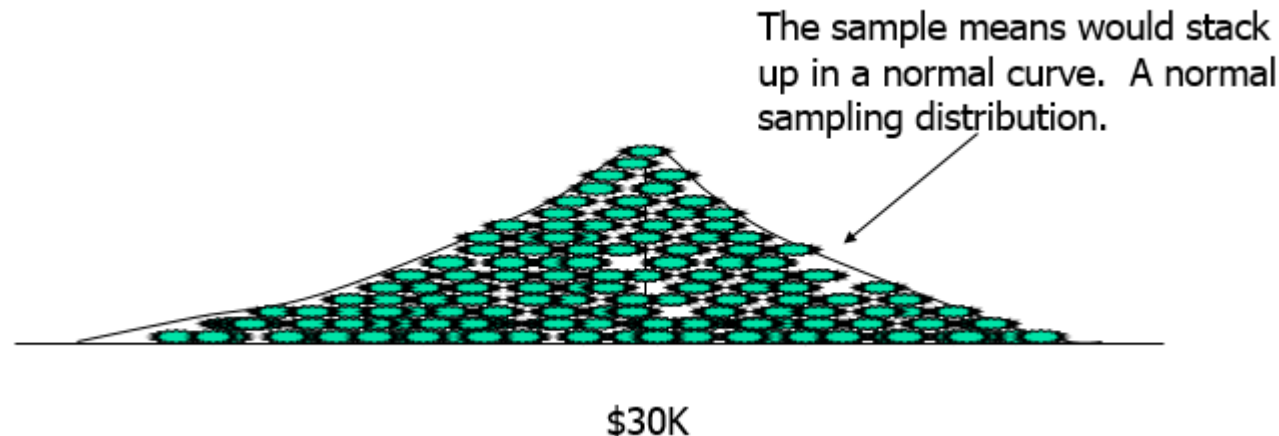
- After carrying out drawing a number of samples, we shall be getting a graph something of the form below



- If we go on continuing the similar actions, we shall be getting graph like this –



- After continuing the process huge number of times, the graph will look like this -



Relationship between Population, Sample & Sampling Distribution

- Mean of sampling distribution is same as the mean of the population
- Standard deviation of a sample is close to standard deviation of the population values
- Standard deviation of samples is taken as good approximation and estimate of the corresponding population standard deviation
- In order to use s of the sample to estimate σ of the population, the following adjustment is done which contributes to the greater accuracy of the estimate –
- We use the expression of s as $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$ instead of $\sqrt{\frac{\sum(x-\bar{x})^2}{n}}$
- The adjustment decreases the denominator and therefore gives a larger result
- The estimated standard deviation of the population is slightly larger than the observed standard deviation of the sample

Sampling Distribution of the Mean

- Suppose that a random sample of size n has been taken from a normal population with mean μ and variance σ^2 . Now each observation in the sample X_1, X_2, \dots, X_n is a normally and independently distributed random variable with mean μ and variance σ^2 . Then by the reproductive property of normal distribution,

The sample mean $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$

has a normal distribution with mean $\mu_{\bar{X}} = \frac{\mu + \mu + \dots + \mu}{n} = \mu$

and variance $\sigma_{\bar{X}}^2 = \frac{\sigma^2 + \sigma^2 + \dots + \sigma^2}{n} = \frac{\sigma^2}{n}$

and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \rightarrow$ It is also called standard error of mean

Sampling Distribution of Mean

➤ If a population distribution is normal, the sampling distribution of the mean (\bar{x}) is also normal for samples of all sizes

➤ The following are important properties of sampling distribution of mean –

(1) It has a mean equal to the population mean, i.e., $\mu_{\bar{x}} = \frac{\mu + \mu + \dots + \mu}{n} = \mu$

(2) It has a standard deviation equal to the population standard deviation divided by the square root of the sample size, that is, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Where $\sigma_{\bar{x}}$ is the measure of the spread of \bar{X} values around μ or a measure of average sampling error or simply stated standard error of mean

(3) The sampling distribution of mean is normally distributed

➤ In case the population distribution is non-normal and provided σ is finite, the distribution of sample means for large samples is always normal

➤ In practice, standard deviation of population is rarely known, and therefore, the standard deviation of the samples which closely approximates the standard deviation of the population is used in place of σ . Hence, the formula for standard error takes the following form -

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where s refers to standard deviation of sample

Distribution of Sample Median

- If a population is large and follows normal distribution with mean μ and standard deviation σ , the medians of the random samples drawn of size n drawn from the population are also normally distributed with a mean μ and standard deviation $1.2533 \frac{\sigma}{\sqrt{n}}$ (if n is large)
- The standard deviation of the distribution of sample median is called standard error of the sample median and is denoted by $\sigma_{Median} = 1.2533 \frac{\sigma}{\sqrt{n}}$

Distribution of Sample Standard Deviation

- If a population is large and normally distributed with standard deviation σ , the standard deviations of the random samples of size n (where n is large), are closely approximated by a normal distribution with a standard deviation $\sigma/\sqrt{2n}$
- The standard deviation of the distribution of standard deviations of samples drawn from a normal population is called the standard error of the standard deviation and is denoted by $S = \sigma/\sqrt{2n}$ where S = standard error of the standard deviations

- Suppose 2 normal populations (first with size N_1 , mean μ_1 and standard deviation σ_1 and the second with size N_2 , mean μ_2 and standard deviation σ_2) are there and samples of size n_1 is drawn from the first population and size n_2 is drawn from second population, then the properties of sampling distribution of the difference of means of 2 samples, i.e., $\bar{x}_1 - \bar{x}_2$ will be as follows –
1. With simple random sample from 2 independent normal populations, the mean of the sampling distribution of $\bar{x}_1 - \bar{x}_2$, denoted by $\mu_{\bar{x}_1 - \bar{x}_2}$ is equal to the difference between the population means, i.e., $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$
 2. The standard deviation of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ (also known as standard error of $\bar{x}_1 - \bar{x}_2$) is given by $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$ (since \bar{x}_1 and \bar{x}_2 are independent random variables, the variance of their difference is the sum of their variances)
 3. If \bar{x}_1 and \bar{x}_2 are the means of two independent samples drawn from two large or infinite populations, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ will be normal if the samples are of sufficiently large size

Sampling Distribution of Proportions

- Population proportion $P=X/N$, where X is the number of elements which possess a certain trait and N is the total number of items in the population
- Sample proportion $p=x/n$, where x is the number of items in the sample which possess a certain trait and n is the sample size
- Suppose a population is infinite and the probability of occurrence of the event (called success or presence of the trait) is P while the probability of non-occurrence of the event is $(1-P)$, then considering all possible samples of size n drawn from this population and for each sample, determining the proportion p of successes, a sampling distribution of sample proportion p will be obtained with mean μ_p and standard deviation σ_p such that –

$\mu_p = P$ and $\sigma_p = \sqrt{\frac{P(1-P)}{n}}$ where σ_p = standard error of the sample proportion (standard error of sample proportion measures the chance variations of sample proportions from sample to sample)

- For large values of n ($n \geq 30$), the sampling distribution is very closely approximated as normally distributed

Sampling Distribution of Difference of 2 Proportions

- Suppose 2 binomial populations (first with parameter P_1 and size N_1 and the second with size N_2 and parameter P_2) are there and samples of size n_1 is drawn from the first population and size n_2 is drawn from second population, then the properties of sampling distribution of the difference of proportions of 2 samples, i.e., $p_1 - p_2$ will be as follows –
1. With simple random sample from 2 independent binomial populations, the mean of the sampling distribution of $p_1 - p_2$, denoted by $\mu_{p_1-p_2}$ is equal to the difference between the population means, i.e., $\mu_{p_1-p_2} = \mu_{p_1} - \mu_{p_2} = P_1 - P_2$
 2. The standard deviation of the sampling distribution of $p_1 - p_2$ (also known as standard error of $p_1 - p_2$) is given by $\sigma_{p_1-p_2} = \sqrt{\sigma_{p_1}^2 + \sigma_{p_2}^2} = \sqrt{P_1(1 - P_1)/n_1 + P_2(1 - P_2)/n_2}$ (since p_1 and p_2 are independent random variables, the variance of their difference is the sum of their variances)
 3. If p_1 and p_2 are the proportions of two independent samples drawn from two large or infinite populations, the sampling distribution of $p_1 - p_2$ will be normal if the samples are of sufficiently large size

Statistical Estimation

- **Statistical Estimation is the procedure of using a sample statistic to estimate a population parameter**
- **A statistic used to estimate a parameter is called an estimator and the value taken by the estimator is called an estimate**
- **Statistical estimation can be broadly classified into 2 categories –**
 - **Point Estimation**
 - **Interval Estimation**

- **An estimate of a population parameter given by a single number is called a point estimate of the parameter**
- **Example – If a firm takes a sample of 50 salesmen and finds out that the average amount of time each salesman spends with his customers is 80 minutes and then it uses this estimate as the average amount of time spent by all salesmen of the firm, it is termed as Point Estimation**
- **Properties of a Good Estimator are –**
 - ❖ **Unbiasedness**
 - ❖ **Consistency**
 - ❖ **Efficiency**
 - ❖ **Sufficiency**

Point Estimation - Unbiasedness

- If θ is the parameter being estimated and $\hat{\theta}$ is an unbiased estimator of θ , then $E(\hat{\theta}) = \theta$
- Sample mean is an unbiased estimator of the population mean, since –

$$E(\bar{x}) = E\left[\frac{x_1 + x_2 + x_3 + \dots + x_n}{n}\right] = \frac{1}{n} [E(x_1) + E(x_2) + E(x_3) + \dots + E(x_n)] = \frac{1}{n} n\mu = \mu$$

- p (i.e, x/n) is an unbiased estimator P , since –

$$E(p) = E(x/n) = \frac{1}{n} E(x) = \frac{1}{n} \cdot nP = P$$

i.e., sample proportion is an unbiased estimator of P (population proportion)

- If the sampling distribution of $\hat{\theta}$ is such that $E(\hat{\theta}) \neq \theta$, then the estimator is said to be biased

Point Estimation - Consistency

- As the sample size increases, the difference between the sample statistic and the population parameter should become smaller and smaller
- If the difference continues to become smaller and smaller as the sample size becomes larger, then sample statistic converging in probability to a parameter is termed as consistent estimator of that parameter
- Symbolically, if $\hat{\theta}$ is a sample statistic computed from a sample of size n and θ is the parameter to be estimated, then
$$Pr[|\hat{\theta} - \theta| \leq d] \rightarrow 1 \quad \text{as } n \rightarrow \infty$$
for any positive arbitrary d , then $\hat{\theta}$ is said to be a consistent estimator of θ
- \bar{x} and s^2 are consistent estimators of μ and σ^2
- The sample median is a consistent estimator of the population mean only if the population distribution is symmetrical

Point Estimation - Efficiency

- If the variance of the estimator is small, the estimator value will be closer to the parameter value
- Some estimators are more efficient than other estimators
- If $\hat{\theta}_1$ is an unbiased estimator of θ and $\hat{\theta}_2$ is another unbiased estimator of θ , then the relative efficiency of $\hat{\theta}_1$ and $\hat{\theta}_2$ is given by –
Relative efficiency = $\text{Var}(\hat{\theta}_2)/\text{Var}(\hat{\theta}_1)$
- For symmetrical distribution, both the sample mean and sample median are unbiased and consistent estimators of the population mean
- In such case, we choose them on the basis of relative efficiency, i.e., we select the one which has smaller variance
- \bar{x} is more efficient estimator of μ than median

Point Estimation - Sufficiency

- A sufficient estimator is the one that uses all information about the population parameter contained in the sample
- The sample mean is sufficient estimator of the population mean since all the information in the sample is used in the computation
- Sample median is not sufficient estimator

Interval Estimation

- An estimate of a population parameter given by two numbers between which the parameter may be considered to lie is called as interval estimate of the parameter
- Interval estimates indicate the precision or accuracy of an estimate and are, therefore, preferable to point estimate
- The interval estimate or a 'confidence interval' consists of an upper confidence limit and lower confidence limit and a probability is assigned that this interval contains the true population value
- The first step in constructing a confidence interval is to decide how much confidence we want this interval will contain the population value (Example – 95% confidence interval)

Interval Estimation Procedure

- The procedure of determining interval estimate comprises of 3 steps –
1. The particular statistics, say, the mean of the sample or standard deviation of the sample is determined
 2. The confidence level is decided, i.e., 95%, 99% etc.
 3. The standard error of the particular statistic is calculated
 4. Confidence limits are calculated as (sample statistic $\pm z_c$ (S.E.) where z_c is the critical value of z
 5. 95% confidence limit for estimation of the population mean μ are given by $\bar{x} \pm 1.96\sigma_{\bar{x}}$
 6. 99% confidence limit for estimation of the population mean μ are given by $\bar{x} \pm 2.58\sigma_{\bar{x}}$

Estimate Mean, Variance Known

Interval estimate of age of students on campus?

Population Variance = 6.25

95% confidence level is 97.5th percentile of the normal distribution at the upper tail




Console ~/

```
> n = length (survey$Age)
> sd = 2.5
> stderr = sd / sqrt (n)
> #error
> error = qnorm (0.975) * stderr
> lowerint = mean (survey$Age) - error
> upperint = mean (survey$Age) + error
> lowerint
[1] 20.05623
> upperint
[1] 20.6928
> error
[1] 0.3182834
> |
```

Estimate Mean, Variance Unknown

Interval estimate of age of students on campus?

95% confidence level is 97.5th percentile of the normal distribution at the upper tail




```
Console ~/     
> n = length (survey$Age)  
> sdev = sd (survey$Age)  
> sderror = sd / sqrt (n)  
> sderror  
[1] 0.1623925  
> error.t = qt (0.975, df = n - 1) * sderror  
> lower = mean (survey$Age) - error.t; lower  
[1] 20.05459  
> upper = mean (survey$Age) + error.t; upper  
[1] 20.69444  
>
```

Estimate Sample Size

Assume campus population SD age = 6.4 years

What is the sample size needed to have a margin of error = 1.2 years

Confidence = 95%

```
Console ~/     
> zconf = qnorm (0.975)  
> sd = 6.4  
> err = 1.2  
> samsize = ((zconf ^ 2) * (sd ^ 2)) / err ^ 2  
> samsize  
[1] 109.2682  
> |
```



Copyright Manipal Global Education Services Pvt. Ltd. All Rights Reserved.

All product and company names used or referred to in this work are trademarks or registered trademarks of their respective holders.

Use of them in this work does not imply any affiliation with or endorsement by them.

This work contains a variety of copyrighted material. Some of this is the intellectual property of Manipal Global Education, some material is owned by others which is clearly indicated, and other material may be in the public domain. Except for material which is unambiguously and unarguably in the public domain, permission is not given for any commercial use or sale of this work or any portion or component hereof. No part of this work (except as legally allowed for private use and study) may be reproduced, adapted, or further disseminated without the express and written permission of Manipal Global Education or the legal holder of copyright, as the case may be.



**THANK
YOU!**