

Descriptive Analytics

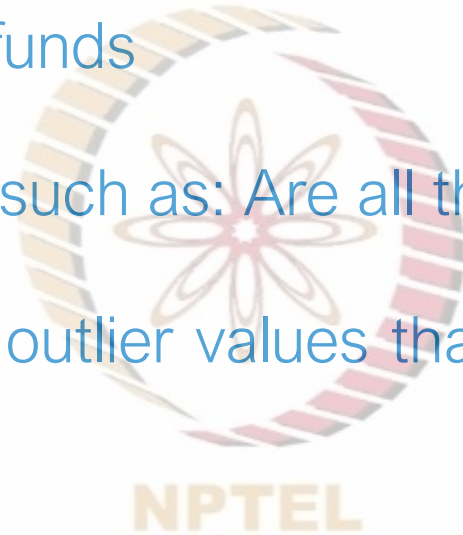
Prof. Abhinava Tripathi

NPTEL



Introduction

- As a fund manager, several prospective clients are requesting to compare the performance of different funds
- They have several questions such as: Are all the values relatively similar?
- And does any variable have outlier values that are either extremely small or extremely large?
- While doing a complete search of the retirement funds data could lead to answers to the preceding questions, you wonder if there are better ways than extensive searching to uncover those answers



Introduction

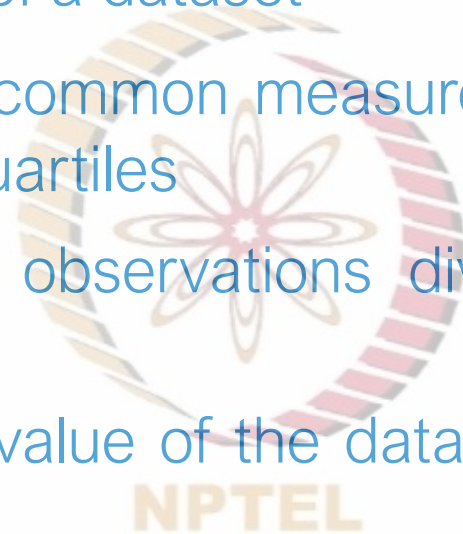
- Descriptive analytics is a commonly used form of data analysis whereby historical data is collected, organized, and then presented in a way that is easily understood
- In Descriptive analysis, we describe our data with the help of various representative methods like charts, graphs, tables, excel files, etc
- The descriptive statistic can be categorized into three parts:
 - Measures of central tendency
 - Measures of variation
 - Measures of shape

Measures of central tendency



Measures of central tendency

- A measure of central tendency is a summary statistic that represents the center point or typical value of a dataset
- In statistics, the three most common measures of central tendency are the mean, median, mode, and quartiles
 - **Mean:** It is the sum of observations divided by the total number of observations
 - **Median:** It is the middle value of the data set. It splits the data into two halves
 - **Mode:** It is the value that has the highest frequency in the given data set
 - **Quartiles:** Quartiles are measures of central tendency that divide a group of data into four subgroups or parts (Q1, Q2, Q3, Q4)



Measures of central tendency: Mean

- The arithmetic mean (in everyday usage, the mean) is the most common measure of central tendency
- To calculate a mean, sum the values in a set of data and then divide that sum by the number of values in the set

$$\bar{X} = \frac{\text{sum of } n \text{ values}}{n} \text{ or } \bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} \text{ or } \bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

- Consider the following data on typical time-to-get-ready for the office in the morning

Day:	1	2	3	4	5	6	7	8	9	10
Time (minutes)	39	29	43	52	39	44	40	31	44	35

Measures of central tendency: Mean

- Consider the following data on typical times to get ready for the office in the morning

$$\bullet \bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{39 + 29 + 43 + 52 + 39 + 44 + 40 + 31 + 44 + 35}{10} = \frac{396}{10} = 39.6$$

- On Day 3, a set of unusual circumstances delayed the person getting ready by an extra hour, so that the time for that day was 103 minutes

$$\bullet \bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{39 + 29 + 103 + 52 + 39 + 44 + 40 + 31 + 44 + 35}{10} = \frac{456}{10} = 45.6$$

Measures of central tendency: Median

- It is the middle value of the data set as It splits the data into two halves
- Extreme values do not affect the median, making the median a good alternative to the mean
- $Median = \frac{n+1}{2}th \text{ ranked value}$
- Calculate the median by following one of two rules
 - **Rule 1:** If the data set contains an odd number of values, the median is the measurement associated with the middle-ranked value
 - **Rule 2:** If the data set contains an even number of values, the median is the measurement associated with the average of the two middle-ranked values

Measures of central tendency: Median

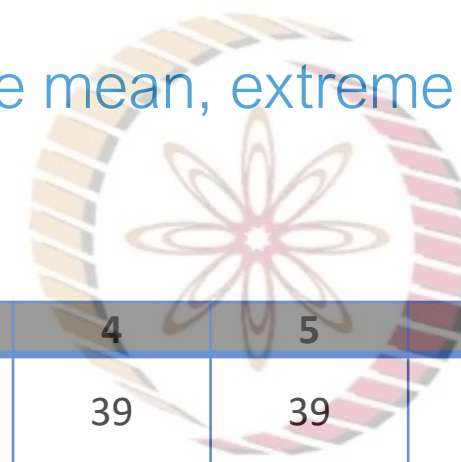
- We will again use the example of 10 time-to-get-ready values, first we will rank them from low to high

Day:	1	2	3	4	5	6	7	8	9	10
Ranked values	29	31	35	39	39	40	43	44	44	52

- The result of dividing $n + 1$ by 2 for this sample of 10 is $(10 + 1)/2 = 5.5$
- As per rule two: Median = $(39 + 40)/2 = 39.5$
- Substituting 103 minutes on Day 3 (As earlier) does not affect the value of median, which would remain 39.5
- This example illustrates that the median is not affected by extreme values

Measures of central tendency: Mode

- The mode is the value that appears most frequently
- Like the median and unlike the mean, extreme values do not affect the mode

A decorative graphic in the background of the table, consisting of a stylized flower or star shape with multiple petals, surrounded by concentric arcs in shades of orange and red.

Day:	1	2	3	4	5	6	7	8	9	10
Ranked values	29	31	35	39	39	40	43	44	44	52

- There are two modes, 39 minutes and 44 minutes, because each of these values occurs twice

Measures of central tendency: Mode

- The mode is the value that appears most frequently
- Like the median and unlike the mean, extreme values do not affect the mode

Observed Data	1	3	0	3	26	2	7	4	0	2	3	3	6	3
Ranked values	0	0	1	2	2	3	3	3	3	3	4	6	7	26

- Because 3 occurs five times, more times than any other value, the mode is 3

Measures of central tendency: Quartiles

- Quartiles are measures of central tendency that divide a group of data into four subgroups or parts
- The three quartiles (Q1, Q2, Q3, Q4) split a set of data into four equal parts.
- First quartile, Q1, $Q1 = (n + 1)/4$ th ranked value
- Third quartile, Q3, $Q3 = 3(n + 1)/4$ th ranked value
- The second quartile (Q2), the median, divides the set such that 50% of the values are smaller than or equal to the median, and 50% are larger than or equal to the median

Measures of central tendency: Quartiles

- Rules for Calculating the Quartiles from a Set of Ranked Values
 - **Rule 1:** If the ranked value is a whole number, the quartile is equal to the measurement that corresponds to that ranked value
 - **Rule 2:** If the ranked value is a fractional half (2.5, 4.5, etc.), the quartile is equal to the measurement that corresponds to the average of the measurements corresponding to the two ranked values involved
 - **Rule 3:** If the ranked value is neither a whole number nor a fractional half, round the result to the nearest integer and select the measurement corresponding to that ranked value

Measures of central tendency: Quartiles

- Consider our example of time-to-get-ready values

Day:	1	2	3	4	5	6	7	8	9	10
Ranked values	29	31	35	39	39	40	43	44	44	52

- Q1: $(n + 1)/4 = (10 + 1)/4 = 2.75$, thus Q1= 35
- Q3: $3(n + 1)/4 = 3(10 + 1)/4 = 8.25$, thus Q3= 44
- Q2 is same as median= 39.5 (corresponding to 5.5)
- Percentiles: Related to quartiles are percentiles that split a variable into 100 equal parts

Measures of central tendency: The Interquartile Range

- The interquartile range (also called the midspread) measures the difference in the center of a distribution between the third and first quartiles
- Interquartile range (IQR) = $Q_3 - Q_1$

Day:	1	2	3	4	5	6	7	8	9	10
Ranked values	29	31	35	39	39	40	43	44	44	52

- IQR= 44-35= 9

Measures of variation



Measures of variability

- Measures of variability describe the spread or the dispersion of a data set
- Measures of variability are
 - **Range:** The Range describes the difference between the largest and smallest data point in our data set
 - **Variance:** The variance is the average of the squared deviations about the arithmetic mean for a set of numbers
 - **Standard Deviation (SD):** Standard deviation measures the dispersion of a dataset relative to its mean. It is defined as the square root of the variance
 - **Mean Absolute deviation:** The mean absolute deviation (MAD) is the average of the absolute values of the deviations around the mean for a set of numbers.

Measures of variability: Range

- A simple measure of variation, the range is the difference between the largest and smallest value and is the simplest descriptive measure of variation for a numerical variable
- $Range = X_{largest} - X_{smallest}$

Day:	1	2	3	4	5	6	7	8	9	10
Ranked values	29	31	35	39	39	40	43	44	44	52

- As per the formula, the range is $52 - 29 = 23$ minutes
- The range measures the total spread in the set of data
- However, the range does not take into account how the values are distributed between the smallest and largest values

Measures of variability: Variance or Standard Deviation

- Two commonly used measures of variation that account for how all the values are distributed are the variance and the standard deviation
- Two commonly used measures of variation that account for how all the values are distributed are the variance and the standard deviation
- The calculation of variance squares the difference between each value and the mean and then sums those squared differences
- For sample variance these sum of squares are divided by sample size-1
- For population variance these sum of squares are divided by population size (N)

Measures of variability: Variance or Standard Deviation

- For a sample containing n values X_1, X_2, \dots, X_n , the sample variance (S^2) is defined as
- Sample variance
$$S^2 = \frac{[(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2]}{n-1}$$
- For a Population containing N values X_1, X_2, \dots, X_n , the Population variance (σ^2) is defined as
- Population variance
$$\sigma^2 = \frac{[(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_N - \bar{X})^2]}{N}$$
- Observe that the difference between dividing by n and by $n - 1$ becomes smaller as the sample size increases and converges to large population size N

Measures of variability: Variance or Standard Deviation

- This can be put in a more compact manner as shown here.
- $S^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}$ or in standard deviation form
- $S = \sqrt{\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}}$
- For population SD: $\sigma = \sqrt{\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n}}$
- Observe that the difference between dividing by n and by n - 1 becomes smaller as the sample size increases and converges to large population size N

Measures of variability: Variance or Standard Deviation

- Consider the example of 10 observations from time-to-get-ready

Time (X)	Step 1: $(X_i - \bar{X})$	Step 2: $(X_i - \bar{X})^2$
39	-0.60	0.36
29	-10.60	112.36
43	3.40	11.56
52	12.40	153.76
39	-0.60	0.36
44	4.40	19.36
40	0.40	0.16
31	-8.60	73.96
44	4.40	19.36
35	-4.60	21.16
Mean=40		Sum =412.40
		Sum Divide by (n-1)=45.82

- $$S^2 = \sqrt{\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}} = \frac{[(39-39.6)^2 + (29-39.6)^2 + \dots + (35-39.6)^2]}{10-1} = \frac{412.4}{9} = 45.82$$
- $$S = 6.77$$

Measures of variability: Variance or Standard Deviation

- Consider the example of 10 observations from time-to-get-ready

Time (X)	Step 1: $(X_i - \bar{X})$	Step 2: $(X_i - \bar{X})^2$
39	-0.60	0.36
29	-10.60	112.36
43	3.40	11.56
52	12.40	153.76
39	-0.60	0.36
44	4.40	19.36
40	0.40	0.16
31	-8.60	73.96
44	4.40	19.36
35	-4.60	21.16
Mean=40		Sum =412.40
		Sum Divide by (n-1)=45.82

- $$\sigma^2 = \sqrt{\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n}} = \frac{[(39-39.6)^2 + (29-39.6)^2 + \dots + (35-39.6)^2]}{10} = \frac{412.4}{10} = 41.24$$
- $$\sigma = 6.42$$

Measures of variability: MAD

- The steps to calculate the mean absolute deviation are shown provided here
 - Step 1: Calculate the mean
 - Step 2: Calculate how far away each data point is from the mean using positive distances. These are called absolute deviations
 - Step 3: Add those deviations together
 - Step 4: Divide the sum by the number of data points

- $$MAD = \frac{[\sum_{i=1}^n |(x_i - \bar{x})|]}{n}$$

Measures of variability: MAD

- Consider the example of 10 time-to-get-ready values and MAD computation for the data

Time (X)	S2: absolute($X_i - \bar{X}$)
39	0.60
29	10.60
43	3.40
52	12.40
39	0.60
44	4.40
40	0.40
31	8.60
44	4.40
35	4.60
S1: Mean=40	S3: Sum=50.00
	S4: Sum/10=5

Measures of shape

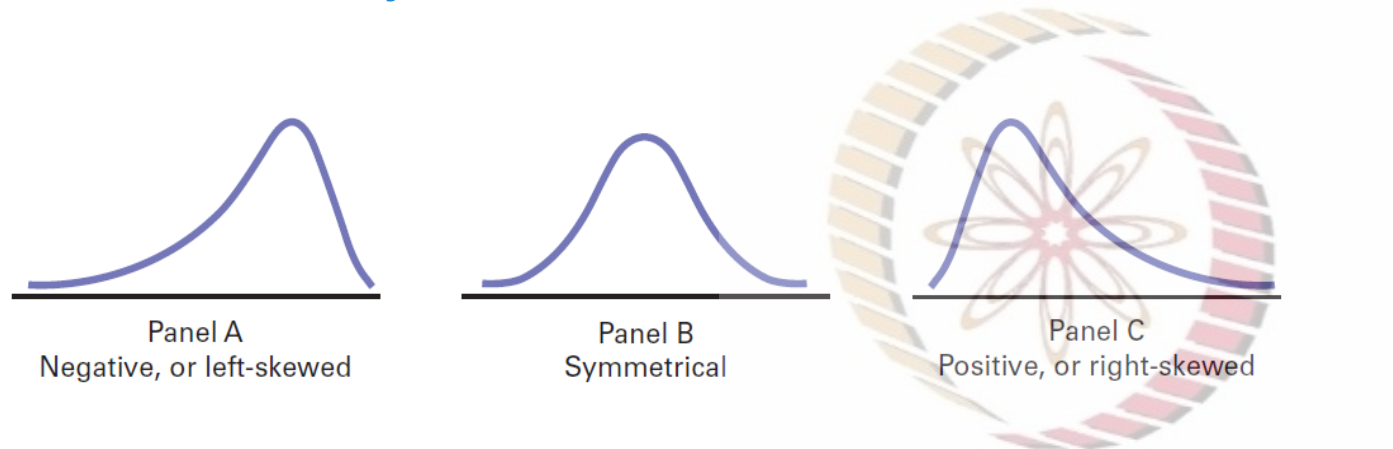
- A measure of shape is the tool that can be used to describe the shape of a distribution of data



- **Skewness:** Skewness refers to a distortion or asymmetry that deviates from the symmetrical nature of data around its mean
- **Kurtosis:** Kurtosis measures the peakedness of the curve of the distribution

Measures of shape: Skewness

- The distribution of data in which the right half is a mirror image of the left half is said to be symmetrical

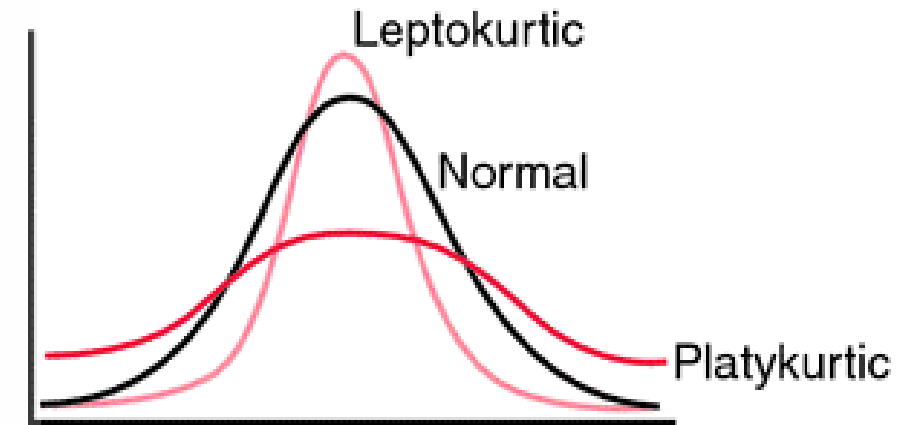


- Panel A: $\text{Mean} < \text{median}$: negative, or left-skewed distribution
- Panel B: $\text{Mean} = \text{median}$: symmetrical distribution (zero skewness)
- Panel C: $\text{Mean} > \text{median}$: positive, or right-skewed distribution

Measures of shape: Kurtosis

- Kurtosis measures the peakedness of the curve of the distribution

- That is, how sharply the curve rises approaching the center of the distribution



- Leptokurtic:** A distribution that has a sharper-rising center peak than the peak of a normal distribution has positive kurtosis
- Platykurtic:** A distribution that has a slower-rising (flatter) center peak than the peak of a normal distribution has negative kurtosis



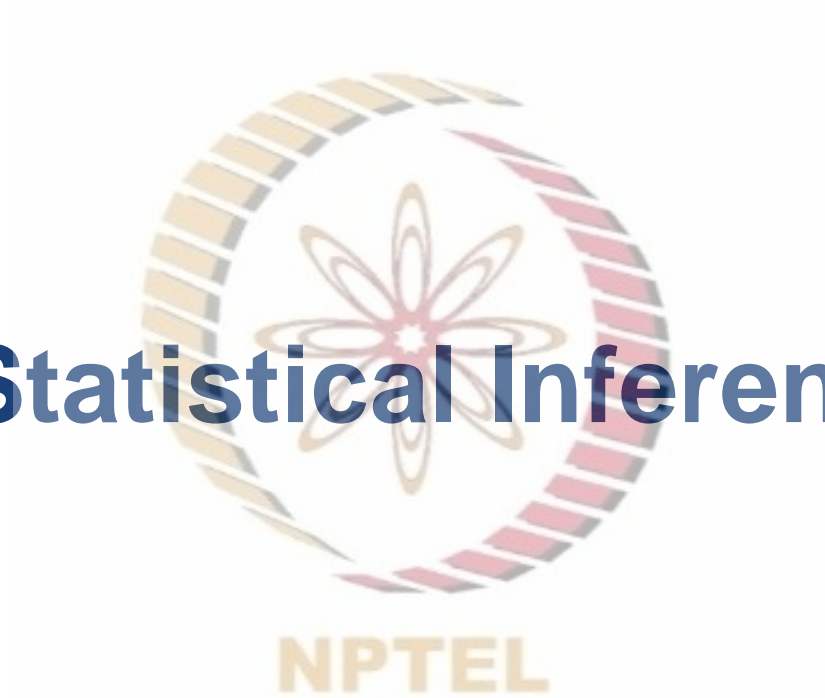
Statistical Inference: Sampling and Confidence Interval Estimation

Prof. Abhinava Tripathi

NPTEL



Introduction: Statistical Inference



Introduction

- Since one does not have the luxury of working with populations, one has to make only inferences and exact solutions or estimates are not available
- Let us start with a simple example from manufacturing industry
- You are working as a part of food regulator to examine the quality of food
- You can not go to all factory outlets and check each packet
- A feasible way is to take small sample that is representative of the population to make appropriate inferences

Introduction

- Assume that the company has 30000 packets out of which you select 100 samples
- You find that the lead content in these packets is 2.2 ppm with a standard deviation of 0.7 ppm
- Can we say that the population mean and standard deviation parameters would be same as the sample
- Is it possible that these sample parameters would be very different from population parameters

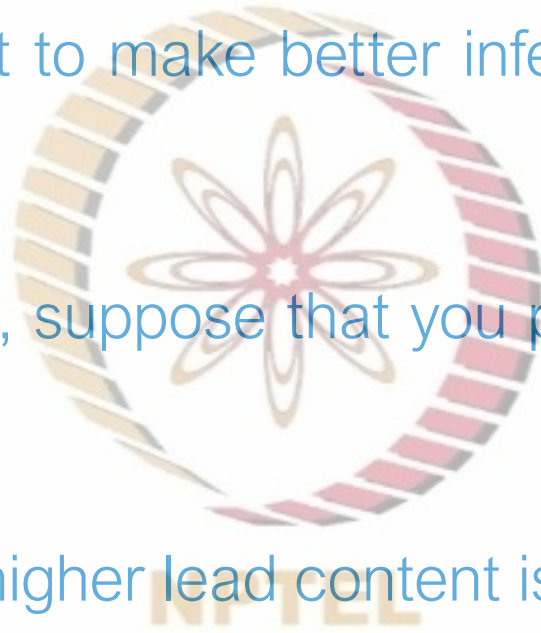


Types of Sampling: Probability Sampling



Introduction to Sampling

- Good sampling is important to make better inferences about the population parameters
- In the food sample problem, suppose that you pick all the 100 samples from a single factory
- It may be possible that the higher lead content is specific to that this factory
- This requires that sampling procedure is fair and unbiased so that inferences are accurate



Simple Random Sampling

- Let us discuss some of the ways in which we can select a sample of 100 noodle packets
- One, though very less efficient way, is to collect all the 30000 packets randomly and select 100 out of these
- This is called simple random sampling
- This is like a blindfolded person picking sample units from population: the process is completely random
- Let us discuss this in more detail



Stratified Sampling

- In the previous example, suppose that 70% of the noodles are from factory A and 30% from factory B
- You sample 70 packages from factory A and 30 packages from factory B randomly
- This kind of sample is expected to be more representative of the population
- It carries proportions of packets from factory A and B, similar to that in the population
- The units are divided into homogeneous strata (sub-groups) and then samples are taken randomly
- This approach is known as Stratified random sampling

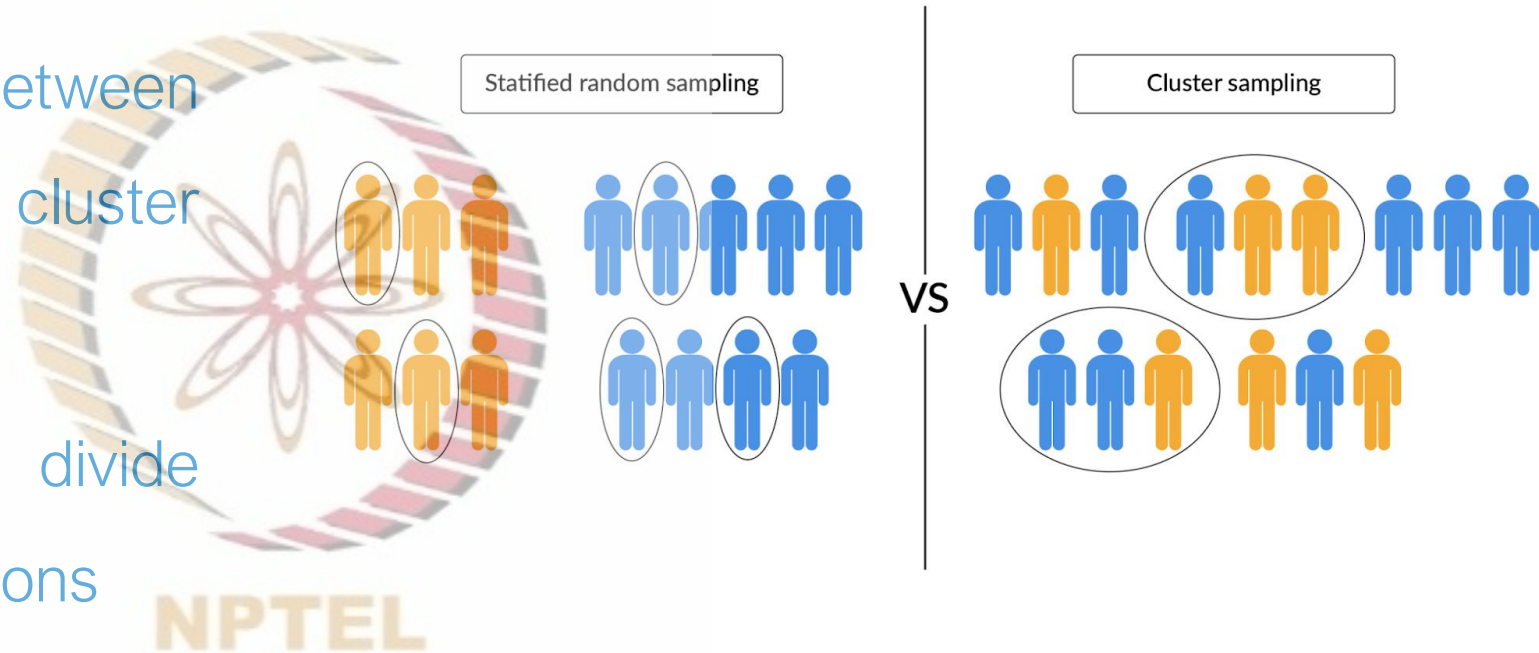


Cluster Sampling

- In the previous example, suppose that there are 20 warehouses in the country
- You would not like to collect your sample from each warehouse (i.e., consider each warehouse as cluster)
- You can select 3-4 warehouses as clusters (may be through random sampling) and then consider desired number of samples from these clusters
- Cluster sampling is usually used when you see that the population can be divided into different groups or clusters that have different characteristics
- Then you do sampling from dissimilar clusters

Cluster Sampling vs. Stratified Sampling

- One may get confused between stratified sampling and cluster sampling
- Since, in both cases we divide populations in sub-populations



- In stratified sampling, we divide the population into sub-populations and then select the sample units in the same proportion as the sub-populations so that the sample is as representative as the parent population

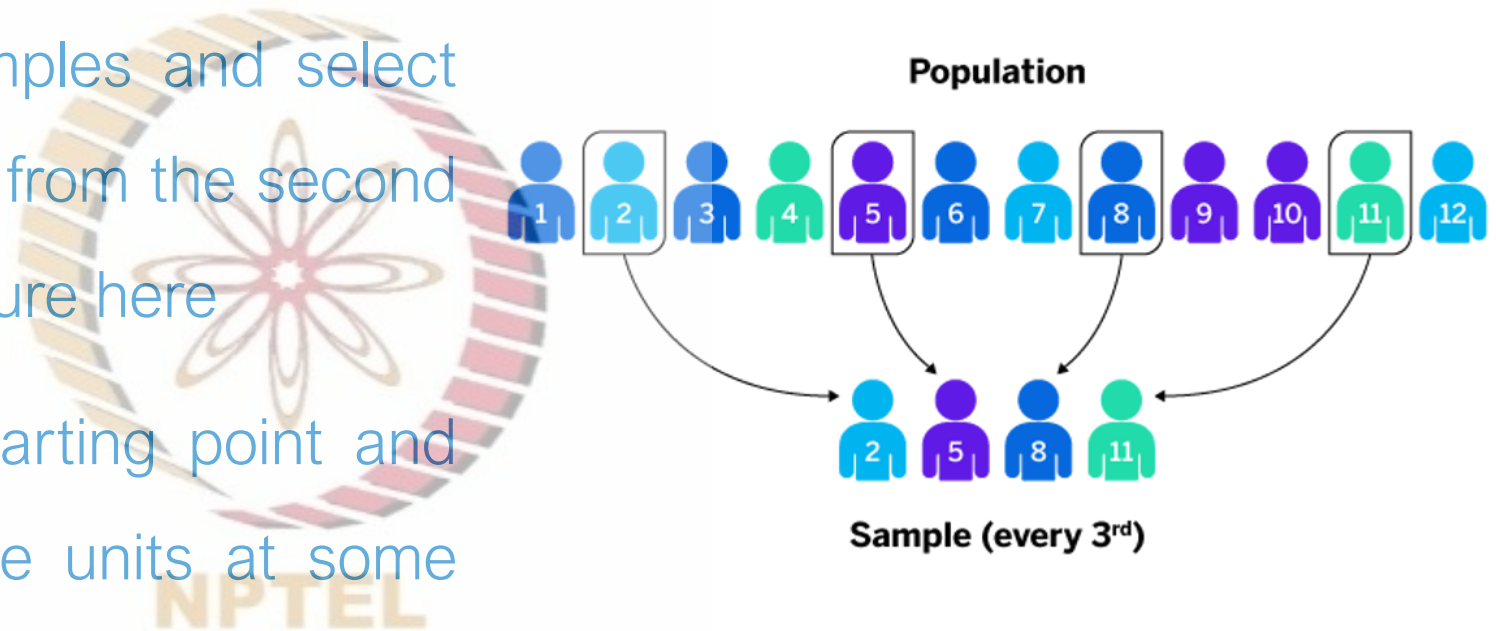
Cluster Sampling vs. Stratified Sampling

- In cluster sampling also, we divide the population into sub-population
- But here we only study the selected clusters, not all the clusters



Systematic Sampling

- Let say you label the samples and select every third packet starting from the second packet, as shown in the figure here
- We selected a random starting point and started picking out sample units at some fixed and periodic interval
- This is called systematic sampling



Sampling methods

- We studied four kinds sampling methods
 - Simple random sampling
 - Stratified random sampling
 - Cluster sampling
 - Systematic sampling
- Which kind of sampling is more suitable for our example
- For many of the food regulators, stratified sampling and simple random sampling is considered as more suitable



Heterogeneous population

- What could be those cases where population is not heterogeneous in nature
- If all the noodle packets are of the same nature and manufactured in a single factory, then simple random sampling would be the most straightforward
- If the packets came in different flavors and were manufactured in different factories, then ideally stratified sampling would be the recommended method
- All these sampling techniques fall under the category of probability sampling
- In these sampling techniques, every unit of the population has a certain known chance of being included in the sample

Types of Sampling: Non-Random Sampling

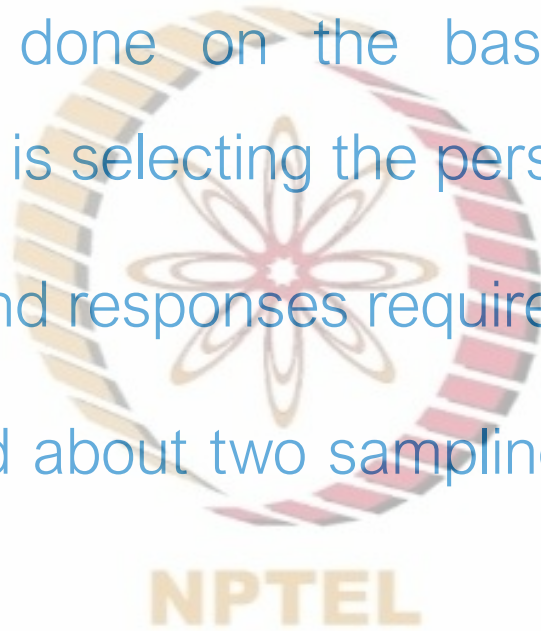


Non-Random Sampling: Convenient sampling

- There is another sampling method called as non-random sampling
- Here, the odds of a sample unit getting selected can not be calculated
- **Convenient sampling:** You choose 100 packets that were closer to you and most easily available
- This sampling method is based on the convenience of the person selecting the sample
- This method has a high probability of being biased

Non-Random Sampling: Judgement sampling

- **Judgement sampling:** It is done on the basis of the knowledge and judgement of the person who is selecting the person
- Often the survey questions and responses require highly specialized skillset
- In this discussion, we learned about two sampling techniques that fall under non-random sampling
- In these methods, it is often important to understand the implication of sampling techniques on the nature of sample being acquired



Statistical Inference I: Central Limit Theorem



Introduction

- In the previous example, 30000 packets is called the population, and the small collection to be examined is called sample

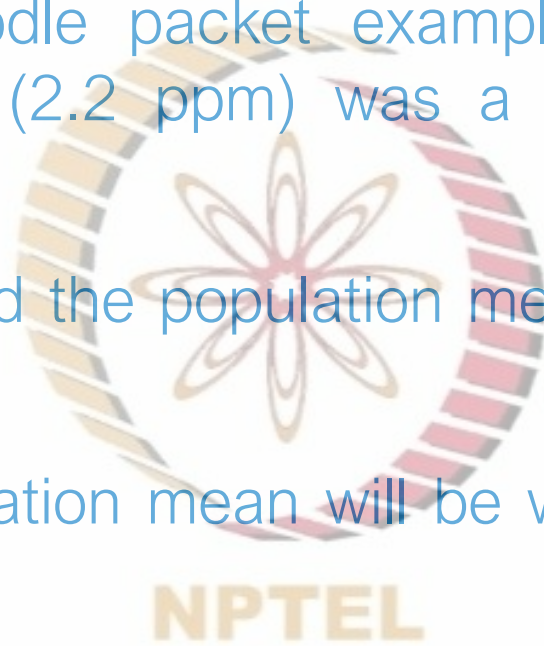
Parameter	Population	Sample
Size	N	n
Mean	μ (or \bar{X})	\bar{x}
Sigma	σ	s
Variance	$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{(N)}$	$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n-1)}$

NPTEL

- The population size is denoted by capital N, its mean by μ and its standard deviation by σ
- The sample size is denoted by a lowercase n, and the mean by \bar{x}

Introduction

- Remember that in our noodle packet example, we wanted to validate whether our sample mean (2.2 ppm) was a true representation of the population
- It is impossible to exactly find the population mean from sample mean with zero error
- All we can say is that population mean will be within 2.2 plus minus some error
- If we are able to estimate that error, let us say 0.2 ppm; then still we are able to add some value to analysis
- We know that the levels of led will be less than 2.5 ppm and more than 2 ppm

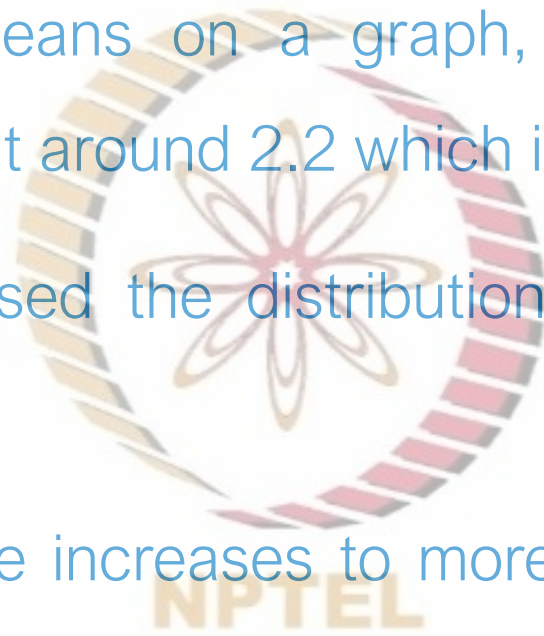


Introduction

- How to establish if the sample is indeed a true representation of the population
- Consider that you have all the $N=30000$ packets, that is the population data
- If the mean of this data is $\mu=2.199$ and the standard deviation is $\sigma=0.132$; these are essentially population parameters
- Let us now consider a sample of size 5 with a mean $\bar{x}=2.145$
- In another sample, the mean comes out to be 2.27 ppm
- So instead of 2, we will choose 100 samples with a sample size of 5

Central Limit Theorem (CLT)

- If we plot these sample means on a graph, it will look like a normal distribution with a center point around 2.2 which is close to population mean
- If the sample size is increased the distribution keeps getting closer and closer to normal distribution
- Moreover, as the sample size increases to more than 30, the mean of the sample distribution approaches the population mean
- This experiment is the basis for central limit theorem



Central Limit Theorem (CLT)

- The central limit theorem states that when you take a large number of samples, the mean of the sampling distribution thus formed, will be approximately equal to the population mean
- The second part of the theorem states that the standard deviation of this sampling distribution will be equal to σ , which is our population standard deviation, divided by the square root of n where n is the sample size
- Finally, the central limit theorem states that if the sample size that you take is greater than 30, the sampling distribution will become normally distributed

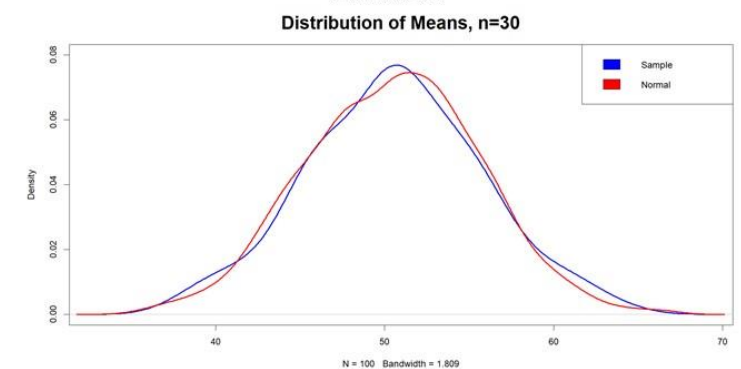
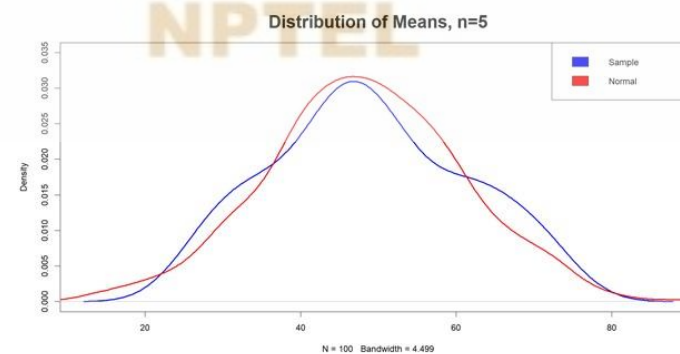
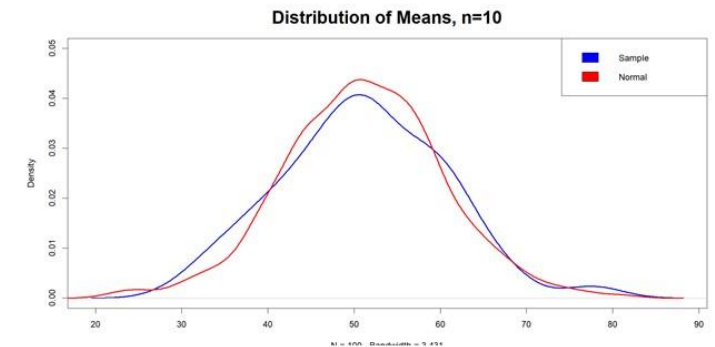
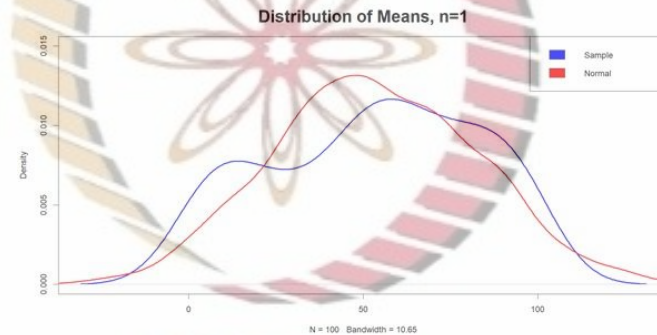
Central Limit Theorem (CLT)

- The central limit theorem states that when you take a large number of samples, the mean of the sampling distribution thus formed, will be approximately equal to the population mean
- The second part of the theorem states that the standard deviation of this sampling distribution will be equal to σ , which is our population standard deviation, divided by the square root of n where n is the sample size
- Finally, the central limit theorem states that if the sample size that you take is greater than 30, the sampling distribution will become normally distributed

Central Limit Theorem (CLT)

- To concretize your understanding of the central limit theorem, let's try and visualize the central limit theorem.

- If you plot the distribution of sample means, or what is also known as the sampling distribution, then this distribution approaches a normal distribution



Central Limit Theorem (CLT)

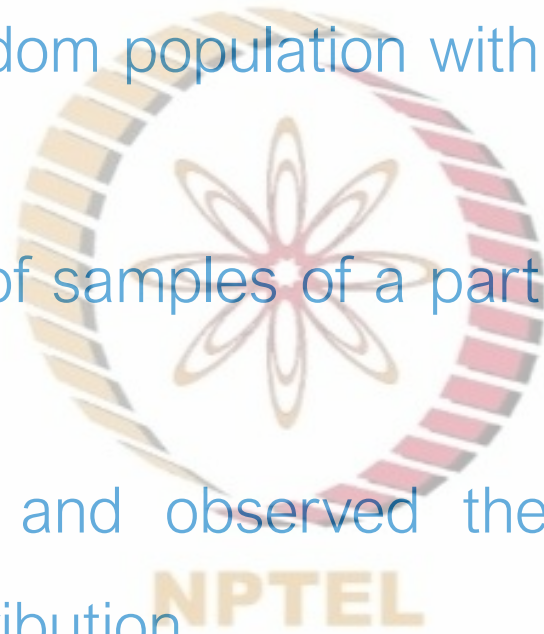
- Notice as the sample size increases how the sample distribution follows the postulations of CLT
- As you increase the size of n , so basically you are bringing the n value closer to the population size value
- The sample mean approaches the population mean as the sample size is increased
- What will happen if you increase the sample size to $n=50$ or even higher

NPTEL

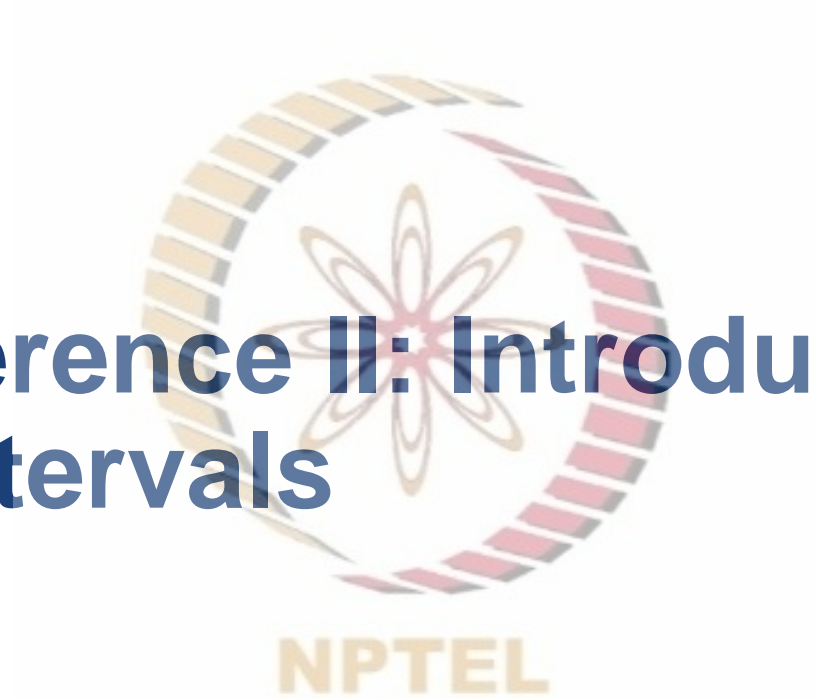
Parameter	Population	Sample ($n=1$)	Sample ($n=5$)	Sample ($n=10$)	Sample ($n=30$)
Mean	50.78	52.97	51.80	50.45	50.60
SD	28.80	29.72	12.91	9.32	5.30
$28.80/\sqrt{n}$		28.80	12.88	9.11	5.26

Central Limit Theorem (CLT)

- We started with absolutely random population with a mean μ and standard deviation of σ
- We then took a large number of samples of a particular sample size and plotted the means of these samples
- We varied the sample sizes and observed the behavior of resulting sampling distribution vis-à-vis normal distribution
- As the sample size increased, the probability distribution became close and closer to normal distribution, and the mean of sample converged to the mean of population



Statistical Inference II: Introduction to Confidence Intervals

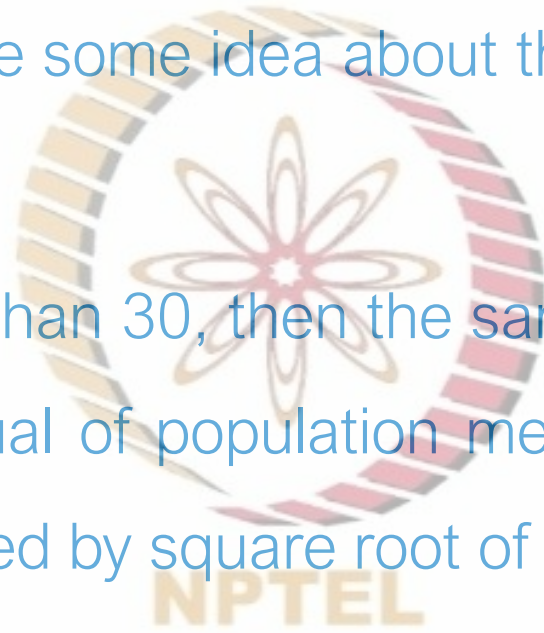


Introduction to Confidence Intervals

- Let us go back to the noodle example, and derive conclusions about the population using the sample
- We took a sample of 100 packets and found out that its sample mean was 2.2 ppm and standard deviation was 0.7 ppm
- We will make use of sampling distribution properties
- Sampling distribution is nothing but the distribution of all the possible sample means that can be generated from this population

Introduction to Confidence Intervals

- With the help of CLT, we have some idea about the properties of this sampling distribution
- If the sample size is greater than 30, then the sampling distribution is normally distributed with a mean equal of population mean and a standard deviation equal to population SD divided by square root of sample size
- We do not know the exact population mean and standard deviation
- There are cases where you have some idea about the population standard deviation, and in some cases you don't, and you employ sample SD for that



Introduction to Confidence Intervals

- Sample standard deviation ($s=0.7$) and $n=100$, we get the SD of sampling distribution as $0.7/10 = 0.07$
- Here, we are using the sample standard deviation as the substitute for population standard deviation
- Now we will make use of normal distribution properties as elaborated earlier (1-2-3 rule)
- For example, using this rule, we can say that the probability that sample mean lies from $\mu - 2 \cdot 0.07$ to $\mu + 2 \cdot 0.07$ is 95%

Introduction to Confidence Intervals

- While we do not know the population mean ' μ ', but we know the population standard deviation 0.07
- Rearranging this a little bit, we can say that $P(2.2 - 2 \cdot 0.07 \text{ to } 2.2 + 2 \cdot 0.07) = 95\%$
- Or the probability that population mean ' μ ' lies in the interval $P(2.2 - 2 \cdot 0.07 \text{ to } 2.2 + 2 \cdot 0.07)$ is 95%
- Or you can say with 95% probability that the mean will lie between 2.06 ppm to 2.34 ppm

Introduction to Confidence Intervals

- The probability associated with this claim is called the confidence level
- Since we are concluding about the population mean with 95% probability, we can say that the confidence level is 95% or alternatively level of significance or alpha value =5% (i.e., 1- confidence level)
- Next, you have the margin of error, which is the maximum error= $2 \times 0.07 = 0.14$
- Final the interval of values or the confidence interval= 2.06 to 2.34
- Since the upper bound of the confidence is less than 2.5, we can conclude with 95% confidence that noodles do not contain lead content that is more than the prescribed limit of 2.5 ppm

Statistical Inference II: Confidence Interval Construction



Confidence Interval Construction

- Till now we have understood estimation of population mean with the construction of unbiased confidence interval
- Often getting population data is not feasible and you need to rely on inferential statistics
- The objective here is to estimate population mean; the population need not be normal
- To solve the problem we start with the sample, using appropriate sampling technique

Confidence Interval Construction

- You select a sample of small size n and calculate the mean of the sample \bar{x} and sample standard deviation 's'
- To solve this problem, let us recall central limit theorem (CLT)
- CLT suggests that sampling distribution will behave like a normal distribution as you increase the sample size (>30), with a mean of μ (population mean) and SD of σ/\sqrt{n}
- Using these values, we will estimate the population mean μ

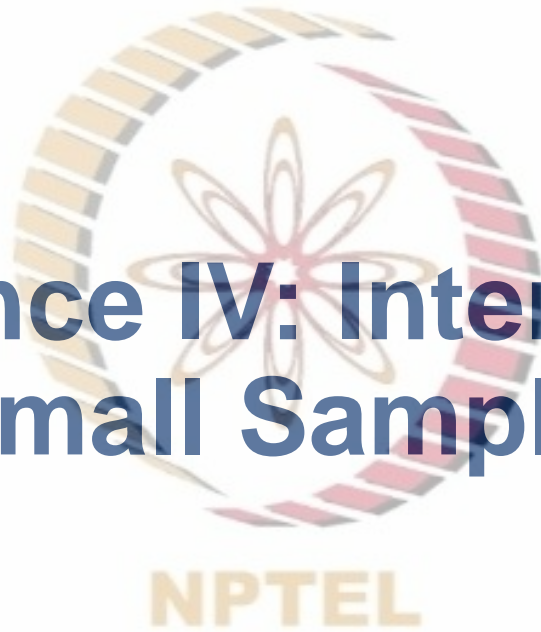
Confidence Interval Construction

- If we consider a confidence level of $y\%$ and apply the CLT, we can estimate that population mean lies in the range: $\bar{x} - z * \frac{s}{\sqrt{n}}$ to $\bar{x} + z * \frac{s}{\sqrt{n}}$; where z is the critical value associated with $y\%$ confidence level
- For confidence levels 90%, 95%, 99%, the values are 1.65, 1.96, 2.58
- You want to be highly confident in the noodle packet example, and 99% confidence makes more sense, or may be you have higher tolerance levels and ok to go ahead with 90% confidence levels

Confidence Interval Construction

- Collect a sample of $n \geq 30$ from the population
- Compute the mean and standard deviation of the sample
- Based on the CLT, assume that the sampling distribution is normal with a mean same as the population mean and SD which is same as population SD divided by square root of n ; population SD is proxied by sample SD
- Select the appropriate confidence level and based on that and decide the appropriate confidence interval : $\bar{x} - z * \frac{s}{\sqrt{n}}$ to $\bar{x} + z * \frac{s}{\sqrt{n}}$

Statistical Inference IV: Interval Estimation for Small Samples



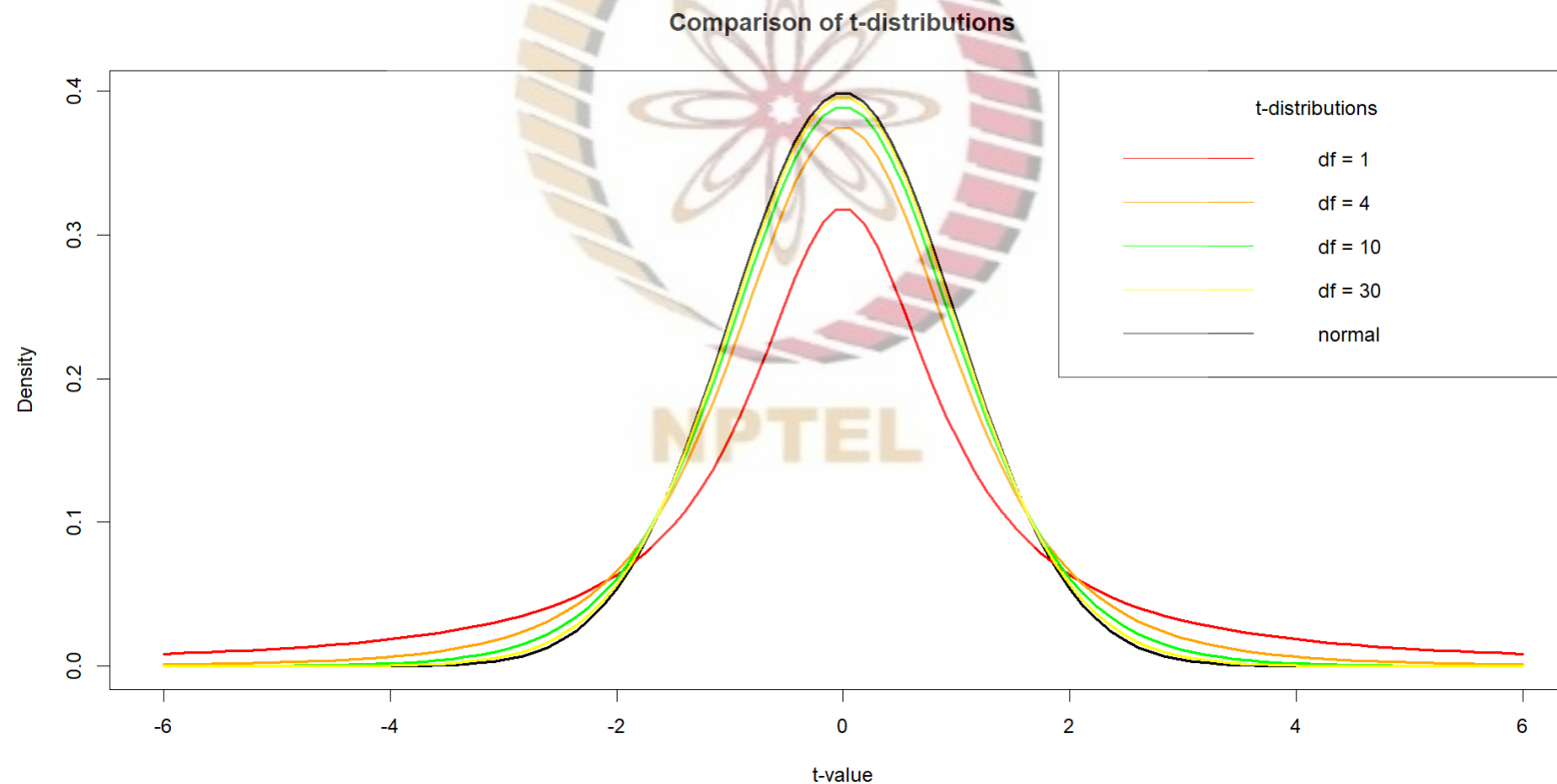
Interval Estimation for Small Samples

- Often times, large samples are not available and one has to work with small samples
- For example, you are in pharma company and in a medicine trial you only have 15 volunteers
- In such cases with less than 30 sample size, you work with t-distribution, where population SD is not known and the same is proxied using the sample standard deviation
- A t-distribution is similar to z-distribution only that it has shorter peak and wider tails



Interval Estimation for Small Samples

- A t-distribution is similar to z-distribution only that it has shorter peak and wider tails



Interval Estimation for Small Samples

- You work for a pharma company and are testing the effects of a medicine on 15 volunteers
- The medicine increases the presence of a particular hormone XYZ in a patient's blood, by 10.038 micro units
- We estimate the population SD using sample SD= 0.072
- We will use the procedure similar to interval estimation using Z-distribution
- But due to sample size restrictions, we will use t-distribution

Interval Estimation for Small Samples

- Population SD is proxied using sample SD; sample mean $\bar{x} = 10.038$ and sample SD = 0.072
- In the sampling distribution we are assuming t-distribution; each t-distribution is distinguished by its degrees of freedom
- For a sample size of 'n', the corresponding degrees-of-freedom (DOF) would be 'n-1'
- For smaller sample sizes, t-distributions are flatter than for larger sample sizes
- For large DOF, t-distribution is similar to the standard normal distribution (at sample size n=30)

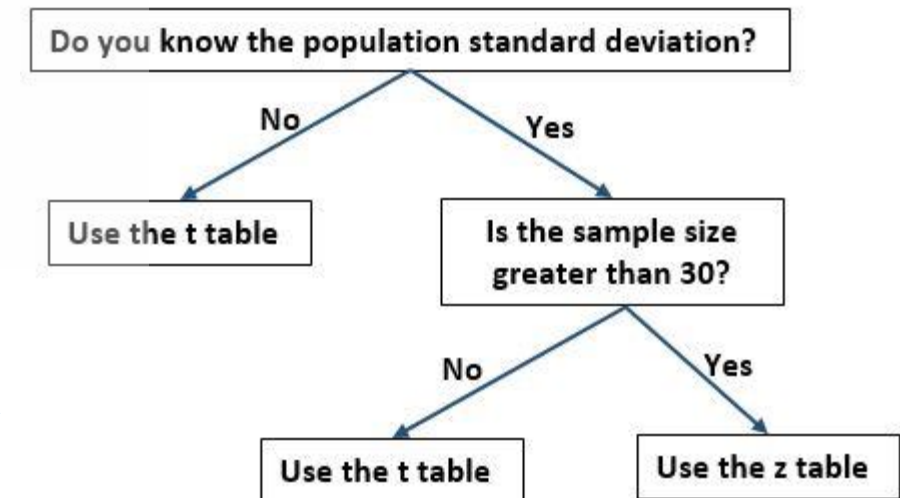


Interval Estimation for Small Samples

- For this case ($n=15$) $DOF=14$, and we will use t-distribution
- We select a confidence level of 95%; the relevant confidence interval will be:
 $\bar{x} - t * \frac{s}{\sqrt{n}}$ to $\bar{x} + t * \frac{s}{\sqrt{n}}$; the corresponding t-value is 2.145
- The lower bound is $10.038 - 2.145 * 0.072 / \sqrt{15} = 9.998$
- The upper bound is $10.038 + 2.145 * 0.072 / \sqrt{15} = 10.077$
- So the 95% confidence interval lies in the range of 9.998 to 10.077
- t-distribution is preferred when sample size is small and population SD is unknown

Interval Estimation for Small Samples

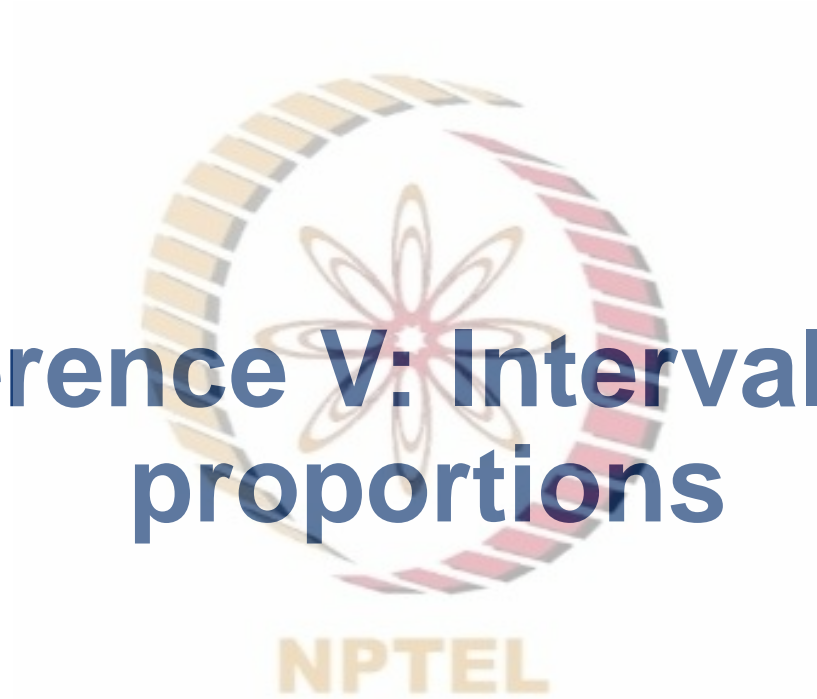
- t-distribution depends on degrees-of-freedom (df) or sample size -1
- For a large sample size, both the t-distribution and normal distribution
- The decision rule flowchart to use the t-distribution and z-distribution is provided below
- If the population standard deviation is unknown and the sample size is greater than or equal to 30, then the z distribution is preferred over the t distribution



Interval Estimation for Small Samples

- If the sample size is less than 30, then even if the population standard deviation is known, it is best to use the t-test as it is ideally suited to dealing with small samples
- The lower and upper bound is given by $\bar{x} - t * \frac{s}{\sqrt{n}}$ to $\bar{x} + t * \frac{s}{\sqrt{n}}$; using this we can estimate the confidence interval

Statistical Inference V: Interval Estimation for proportions



Interval Estimation for proportions

- Many times the values are categorical in nature: for example, in an exit poll survey, a sample of people voted one of the two parties
- How to extrapolate this value to the entire population, given that sample mean and standard deviation driven approaches are not valid
- Consider for example, you are working as part of a political science company that specializes in voter polls and designs surveys to keep political office seekers informed of their position in a race

Interval Estimation for proportions

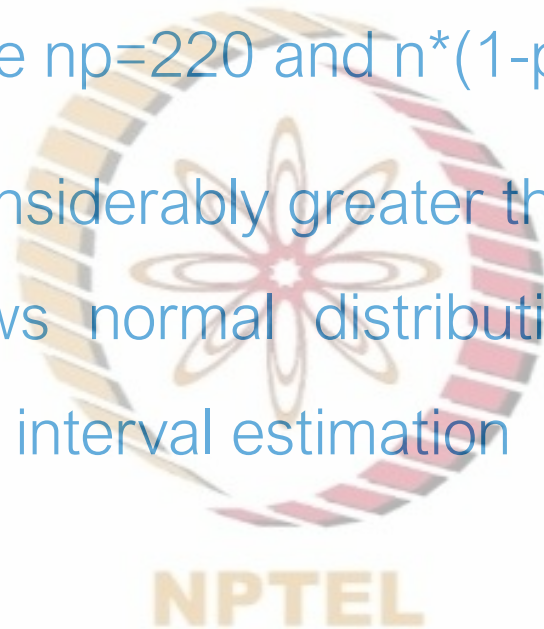
- Through these surveys you found that 220 registered voters, out of 500 contacted, favor a particular candidate. You want to develop 95% confidence interval estimate for the population of registered voters
- The data is categorical in nature: Voted for a party or not voted
- The proportion of voters who voted is $\bar{p} = 220/500 = 0.44$; we want the confidence interval around this proportion (e.g., 0.43 to 0.45)
- The approach to estimate the confidence interval remains the same

Interval Estimation for proportions

- Step 1: is to collect a sample of size $n=500$
- Step 2: Since data is categorical, we computed the proportions ($\bar{p}=0.44$)
- Step 3: Here we generate the sampling distribution of sample proportions and then find the interval estimate
- For being able to apply the sampling distribution of sampling proportion: $n \cdot p > 5$ and $n \cdot (1-p) > 5$;
- The best estimate of population proportion p here is the sample proportion $\bar{p}=0.44$

Interval Estimation for proportions

- Since, $n=500$ here, therefore $np=220$ and $n*(1-p)= 280$
- Both of these values are considerably greater than 5, so we can assume that sampling distribution follows normal distribution and go ahead with the formula for 95% confidence interval estimation

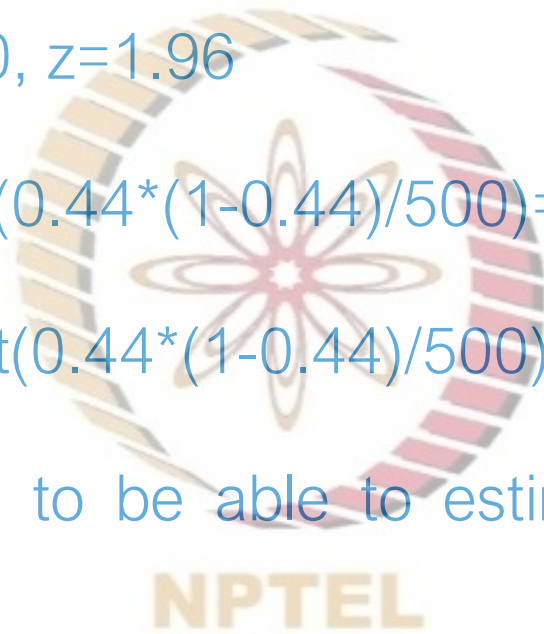


- The appropriate confidence interval here is $\bar{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{(1-\bar{p})*\bar{p}}{n}}$; here SD is

taken as $\sqrt{\frac{(1-\bar{p})*\bar{p}}{n}}$

Interval Estimation for proportions

- In this case, $\bar{p}=0.44$, $n=500$, $z=1.96$
- Lower limit= $0.44-1.96*\sqrt{0.44*(1-0.44)/500}=0.44-0.0435=0.3965$
- Upper limit= $0.44+1.96*\sqrt{0.44*(1-0.44)/500}=0.44+0.0435= 0.4835$
- The aim of the problem is to be able to estimate an interval around the sample proportion \bar{p}



INDIAN INSTITUTE OF TECHNOLOGY KANPUR



Statistical Inference: Hypothesis Testing

Prof. Abhinava Tripathi

NPTEL



Introduction: Hypothesis Testing



Introduction

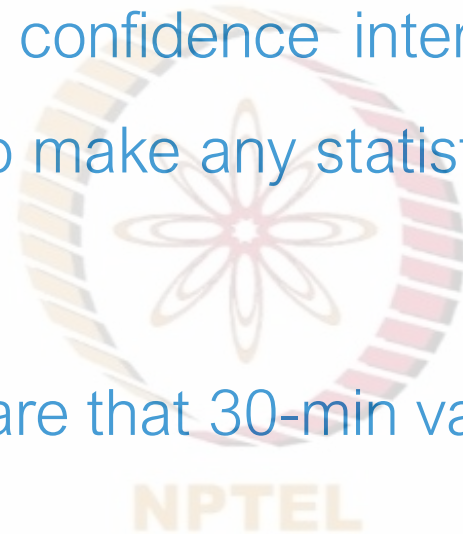
- Inferring insights from sample data is called inferential statistics
- Consider an example, where you have joined a mobile phone manufacturer as business analyst
- Company has come up with a new brand that is expected to charge in 30-mins for a full day operation : A day's power within 30-mins
- You want to check whether this claim is valid or not: some of the units do not conform to this number
- Marketing team wants you to conduct statistical test to be 95% sure of this claim

Introduction

- One approach to solve this problem is to collect a sample of 100 phones
- Chances are that some of these phones may take less than 30-mins and some may take more than 30-mins
- With this information you can come-up with a confidence interval that the charging time is in the range of 24-29 minutes with 95% confidence: I can say that charging time is less than 30 mins
- But what if the confidence interval is from 26-32 mins: I can not say that charging time is less than 30-mins

Introduction

- This confidence interval approach is perfectly valid
- However, as I increase my confidence interval from 95% to 99%, it will become increasingly difficult to make any statistical claim that charging time is less than 30-mins
- This is so because chances are that 30-min value may fall in this interval
- A more robust and efficient method to test this claim is provided by hypothesis testing

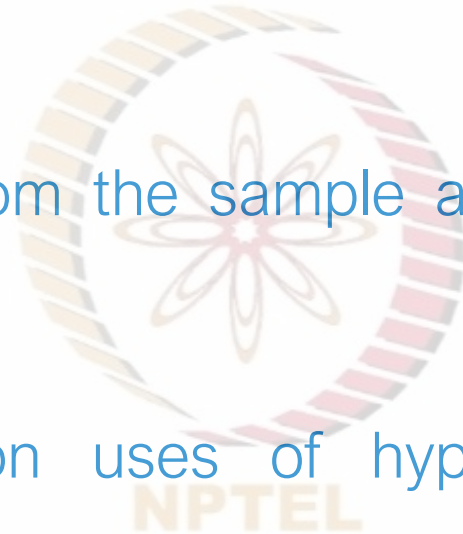


Applications of Hypothesis Testing



Applications of Hypothesis Testing

- Hypothesis testing is an efficient way to test the statistical credibility of a claim
- You gather the evidence from the sample and check if the claim can be rejected or not
- One of the most common uses of hypothesis testing is campaign effectiveness
- A pizza delivery firm plans to test the effectiveness of their campaign in the test population and control population



Applications of Hypothesis Testing

- These kind of problems are referred to as AB testing: for example, you want to compare the response rate of two different webpages
- You divide the population in two groups (Group 1 and 2) that are exposed to different versions of the product: version A and version B
- Another example is the application of quality claim checks: a lightbulb manufacturer claims that his product will last more than 5000 hours
- Here, hypothesis testing can add a lot of value to analysis

Hypothesis Testing: Part I

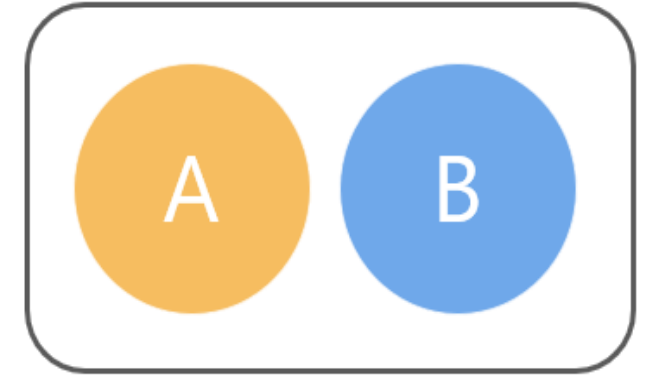


Hypothesis Testing

- We will continue with our charging time problem
- Let us start by assuming that the claim is true: It takes 30-minutes to charge the phone
- As a first step, we define the null and alternate hypothesis
- Null H_0 : The phone fully charges in exactly 30-mins
- Alternate H_1 : The phone does not charge in 30-mins
- If the null is rejected then either it takes more or less than 30-minutes to charge the phone

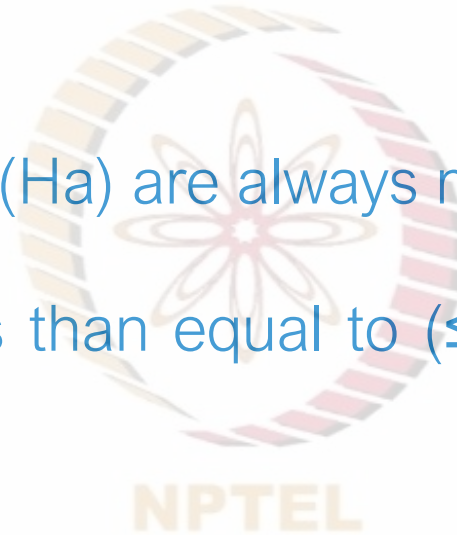
Hypothesis Testing

- This was a simple case of framing null and alternate hypothesis
- Null (A) and alternate (B) hypothesis are always mutually exclusive events
- This was a simple case of framing null and alternate hypothesis
- Null (A) and alternate (B) hypothesis are always mutually exclusive events
- Any other non-mutually exclusive configuration will be rejected



Hypothesis Testing

- Also the null and alternate must be collectively exhaustive, i.e., at least one of them must be true
- So the Null (H_0) and alternate (H_a) are always mutually exclusive and exhaustive
- Null H_0 : Charging time is less than equal to (\leq) 30-mins; Alternate H_1 : Charging time is more than ($>$) 30-mins
- As per the convention, Null hypothesis must contain equality sign



Hypothesis Testing

- In a converse manner, if someone claimed that the phone takes at least 30-mins then the following null and alternate would be framed
- Null H_0 : Charging time is more than (\geq) 30-mins
- Alternate H_1 : Charging time is less than ($<$) 30-mins
- To summarize, (a) Null hypothesis captures the status quo, (b) alternate, which we are trying to prove, is the complement to the null, (c) as a convention, null contains equality ($=, \leq, \geq,$)

Hypothesis Testing

- In our earlier example, H_0 is that 'Charging time is = 30 min'
- Alternative hypothesis, is that 'Charging time is NOT equal to 30 min'
- If the computations favor alternate hypothesis, then you reject the null or accept the alternate hypothesis
- That is, the claim that charging time is not equal to 30-mins is true
- And the required action is based on whether the charging time is more or less than 30-mins

Hypothesis Testing

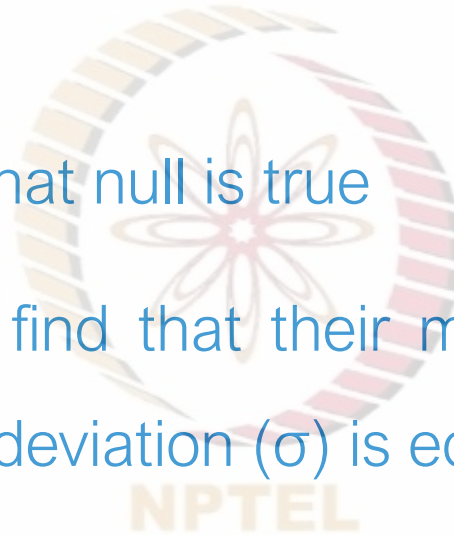
- If the calculations favor the null hypothesis that charging time is equal to 30-mins that means you fail to reject the null (not that null is correct or accepted)
- Not that null hypothesis has been provided to be true
- Sample properties may different from population, and as more and more sample arrive, null may be disapproved
- For example, all swans are white, till the one single black swan was found to reject the null

Hypothesis Testing Part II: Critical Value Method



Hypothesis Testing: Critical Value Method (CVM)

- We will discuss the CVM method for testing the claim that charging time is equal to 30-mins
- We start with the assumption that null is true
- You sample 100 phones and find that their mean charging time is 30.37-mins and population standard deviation (σ) is equal to 2.477
- Please note that these 100 data points make-up for one sample
- We will employ CLT here to implement the hypothesis testing



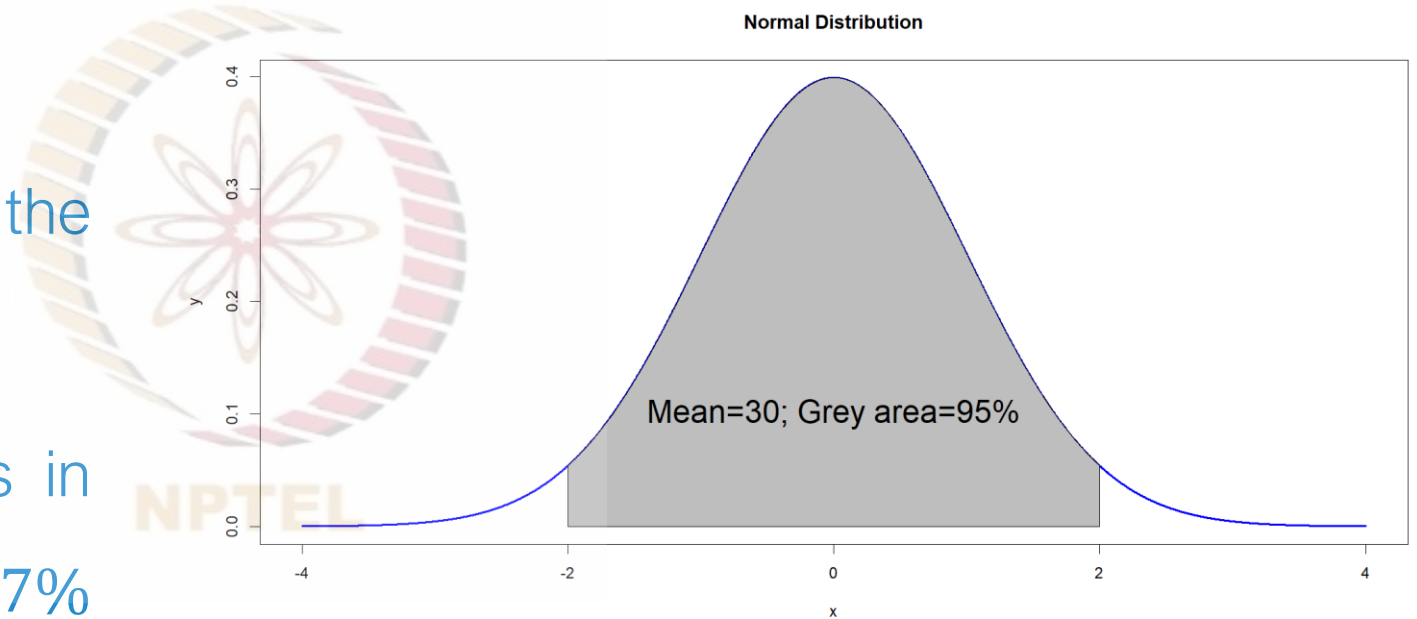
Hypothesis Testing: Critical Value Method (CVM)

- Choose a sufficiently large sample (>30) so that sampling distribution is same as normal distribution
- CLT suggests that the mean of sampling distribution is the same as population mean: 30
- Also, if the population SD is 2.477, the SD of sampling distribution is 0.2477
- You are happy as long as the sample mean lies in the 95% confidence level interval of 30-mins



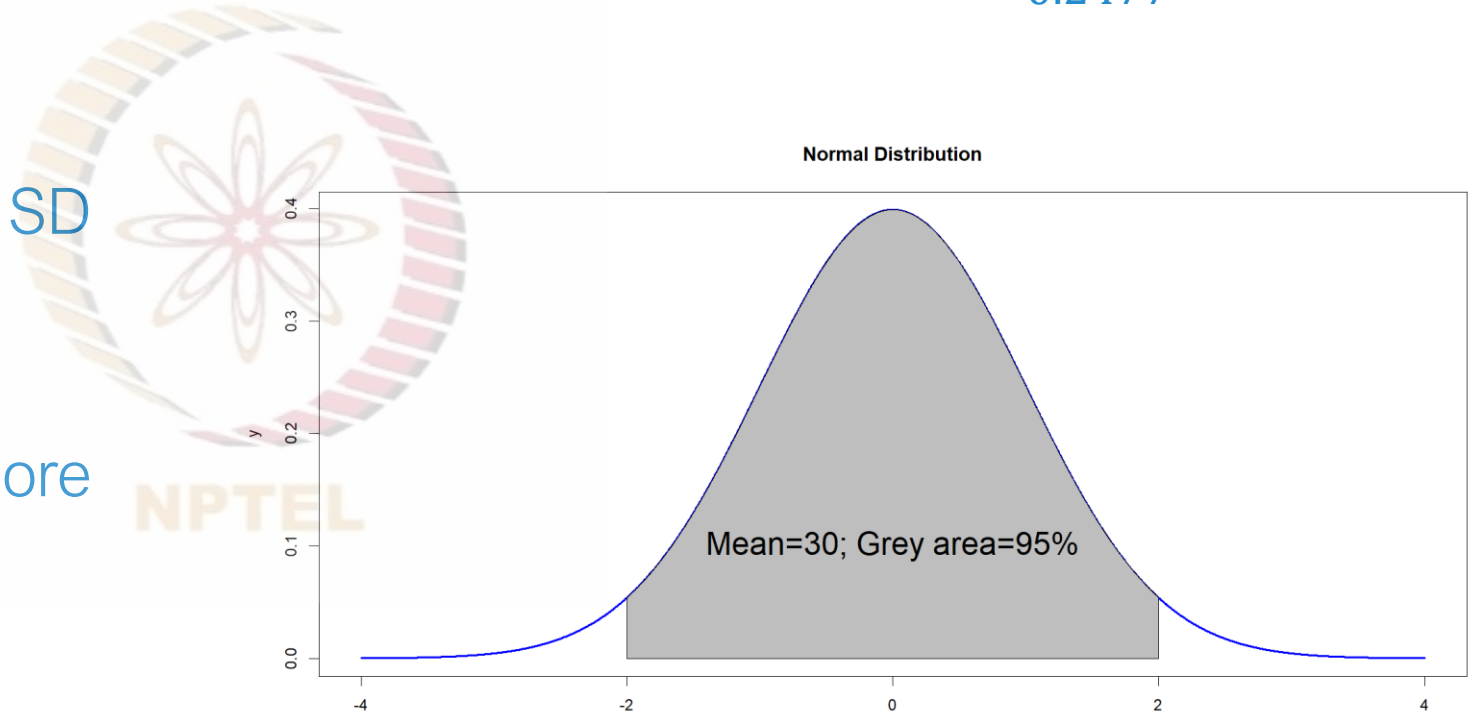
Hypothesis Testing: Critical Value Method (CVM)

- You are happy as long as the sample mean lies in the 95% confidence level interval of 30-mins
- Recall the properties of the normal curve
- Around 68% of the area lies in ± 1 SD, 95% in ± 2 SD, and 99.7% in ± 3 SD



Hypothesis Testing: Critical Value Method (CVM)

- The z-score corresponding to 30.37 is calculated as shown here: $\frac{30.37-30}{0.2477} = 1.4937$
- Sample mean lies 1.4937 SD away from the center
- This is less than the critical score of $z=2$
- Since this is less than the critical score, you fail to reject the null



Hypothesis Testing: Critical Value Method (CVM)

- For example, if sample mean was 30.62, then $z = \frac{30.62 - 30}{0.2477} = 2.5$
- This would fall outside the 95% region and you would be able to reject the null
- Let us recap the problem: (1) Frame the null and alternate hypothesis; (2) Decide the appropriate confidence interval; (3) Calculate the critical z value; (4) Compute the sample z-score; (5) Compare the sample z-score with the critical z value

Hypothesis Testing: Critical Value Method (CVM)

- For example, if sample mean was 30.62, then $z = \frac{30.62 - 30}{0.2477} = 2.5$
- This would fall outside the 95% region and you would be able to reject the null
- Let us recap the problem: (1) Frame the null and alternate hypothesis; (2) Decide the appropriate confidence interval; (3) Calculate the critical z value; (4) Compute the sample z-score; (5) Compare the sample z-score with the critical z value

Hypothesis Testing Part III: One Tailed Test (CVM)



One Tailed Test

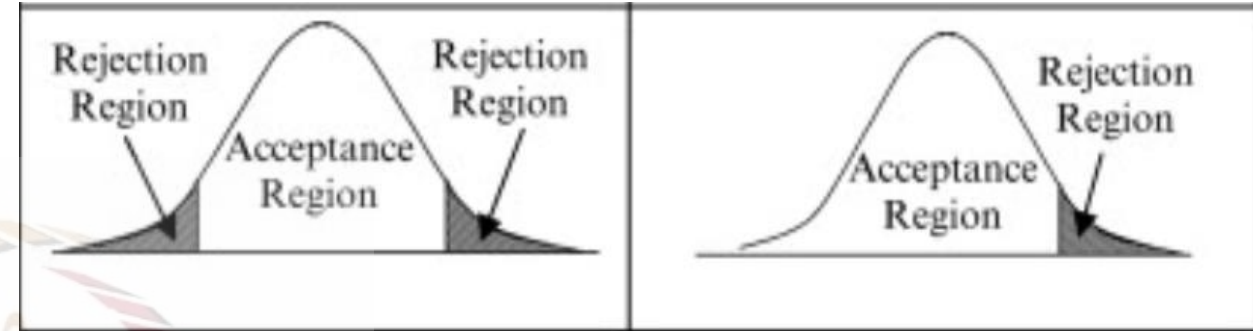
- It is sometimes sufficient to test only one side of the sample mean distribution; this is called a one-tailed test
- In the previous example, we performed the test on both sides of the normal distribution, that is, we would have rejected the null hypothesis if the sample was significantly different in either direction
- As a customer, you want to test whether a day's charge is less than 30-mins

One Tailed Test

- Step 1 is the same, i.e., frame the null and alternate hypothesis
- H_0 : Charging time is less than or equal (\leq) 30-mins
- H_1 : Charging time is less than or equal (\geq) 30-mins
- Step 2: Decide the confidence level (95%);
- Step 3: Find the corresponding critical z-value; here, we can reject the null if we can prove that the sample mean is significantly greater than 30-minutes
- Hence the rejection region is on the right side of the curve: right tailed test

One Tailed Test

- Let us compare the rejection regions in one-tailed vs two-tailed test



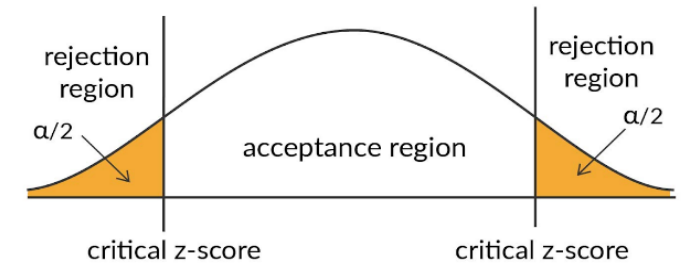
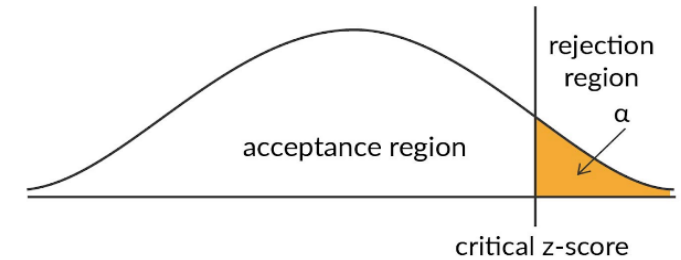
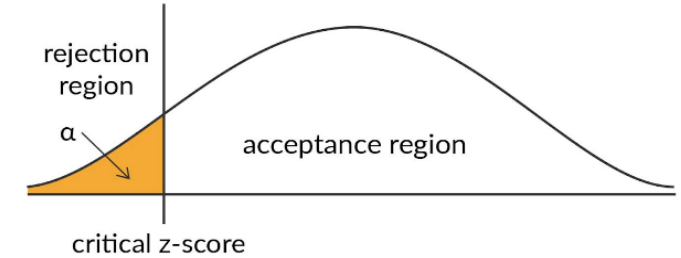
- In the two tailed test, the unshaded region is the 2.5% on the right and left
- In the single tailed test, the entire 5% is on the right side, thus the critical value will not be the same
- In the previous case, critical z value was 1.96. Now the complete area on the left of the curve is 95% and the corresponding z -value is 1.645
- The rejection region is on the right of this z -value

One Tailed Test

- Now that we have our z-value, let us compute the sample z-score
- For our sample of 100 phones, the sample mean was 30.79, and we found the population SD to be 2.477
- Sample z-score = $\frac{30.79 - 30}{\frac{2.477}{\sqrt{100}}} = 3.19$; which is much larger than 1.645
- So we reject the null for a right tailed test

One Tailed Test

- So there are three types of tests: (1) two tailed tests; (2) Right tailed tests; (3) Left tailed tests
- If the alternate hypothesis has $<$ sign, then it is a left tailed test
- If the alternate hypothesis has $>$ sign, then it is a right tailed test
- If the alternate hypothesis includes a ' \neq ' sign, then it is a two tailed test



One Tailed Test

- Let us consider one hypothetical case of left-tailed test
- Cadbury states that the average weight of a particular brand of its chocolate is 60g; as an analyst, you want to test if the weight is less than 60g or not at 2% significance
- Here H_0 : Weight is less than (\geq) 60g; H_1 : Weight is less than ($<$) 60g
- Again the hypothesis can be solved in 5 standard steps
- We need to compare the critical z value (corresponding to 2% level) on the left tail with the corresponding z-statistic from the sample

Hypothesis Testing Part IV: P-value method



P-value method

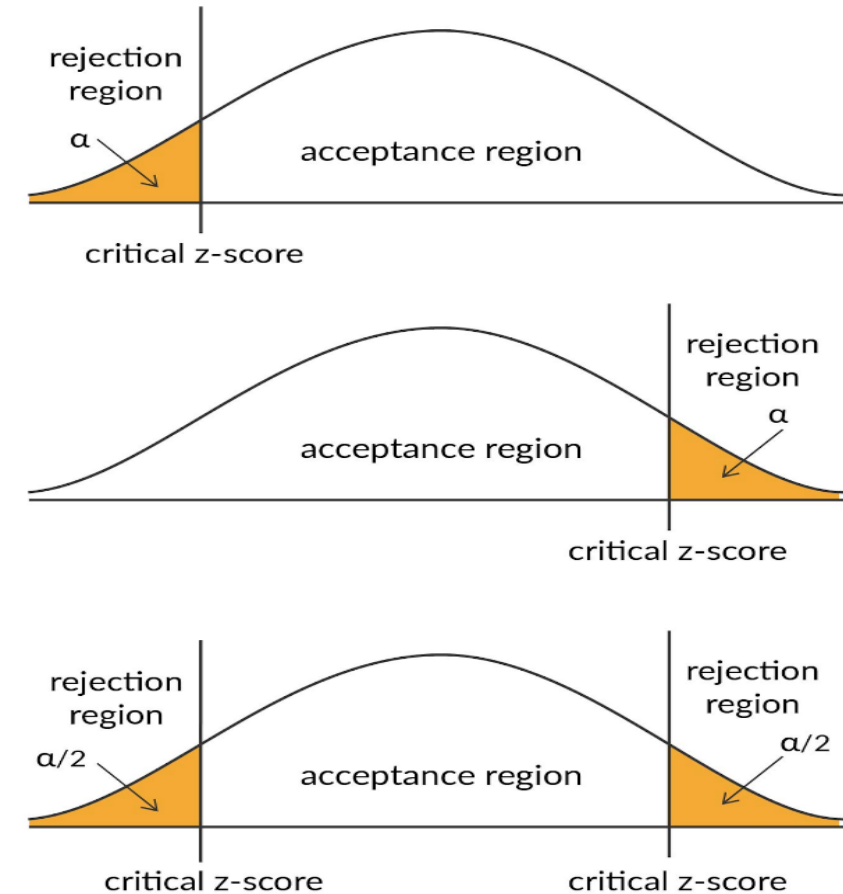
- Let us get introduced to the P-value approach to hypothesis testing
- In the same mobile phone example, first we formulate null and alternate
- H_0 : Mean charging time=30-mins; H_1 : Mean charging time \neq 30-mins
- Decide the level of significance ($\alpha=5\%$) or level of confidence $1-0.05=0.95$
- To compute the p-value, we need the corresponding z-score: we already calculated this value as 1.4937
- P-value or significance level or the area in the tail of the normal probability distribution

P-value method

- In our mobile phone example, p-value represents area from $-\infty$ to -1.4937 and 1.4937 to $+\infty$
- Since the curve is symmetric, we can calculate one value and multiply it with 2
- The value can be computed with R software. The value corresponding to one tail-works out to 0.068 and thus, the total p-value becomes $2*0.068=0.136$
- If p-value is less alpha, then we reject the null

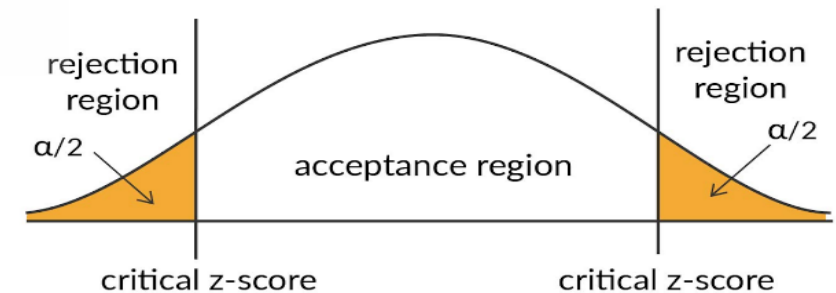
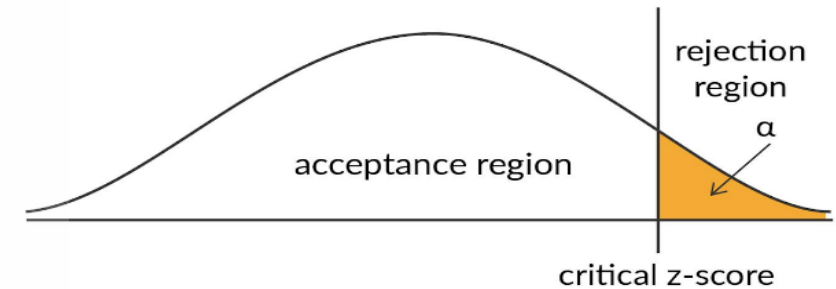
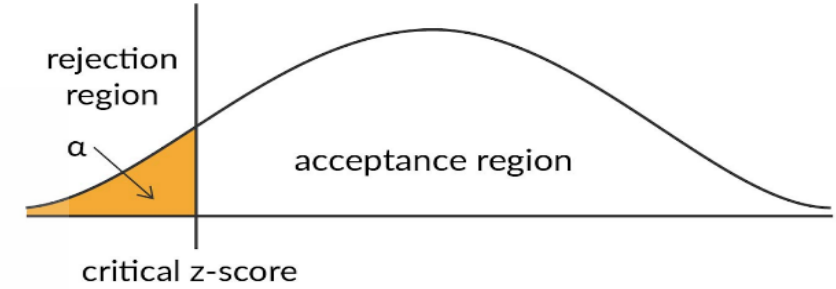
P-value method

- Here our p-value is 0.136, which is more than 0.05, we fail to reject our null at 95% confidence (or 5% significance level)
- For a two tailed test, p-value falls on both sides
- For the left tailed test, p-value will be on the left side of the curve



P-value method

- For the right tailed test, p-value will be on the right side of the curve
- Once the p-value is available, the condition to reject the null remain the same



Summary and Concluding Remarks



Summary

- We discussed hypothesis testing to check the validity of claims with statistical rigor
- You defined the null and alternate hypothesis
- You computed the z-score from sample data points and compared it to the critical z-value
- You also saw the p-value method, which was the probability at tails (significance level)
- A lower p-value meant the higher chance (confidence) of rejecting the null

