Practitioners Course in Descriptive, Predictive and Prescriptive Analytics

Prof. Deepu Philip
Dr. Amandeep Singh Oberoi
Prof. Sanjeev Newar
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur
National Institute of Technology, Jalandhar

Lecture - 17
Hypothesis Testing

Welcome you all to another lecture in the practitioner course on Descriptive Prescriptive and Predictive Analytics. I am Sanjeev Newar, a practitioner, a user of analytics all kinds of analytics and I would be covering a hypothesis testing today primarily from practitioners point of view relying more on the intuition part of it, the feel of it, so that you can play around with the tricks that this hypothesis testing offers and also beware of the traps and where you can trip down doing your analysis.

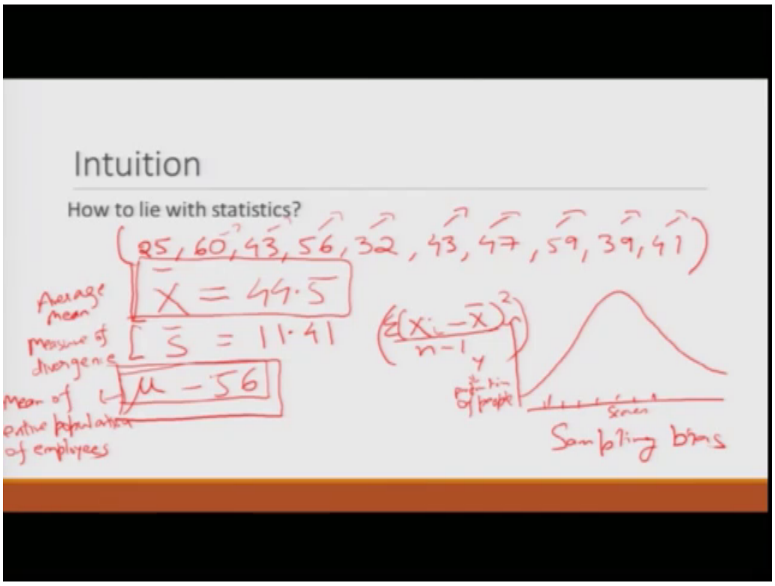
So, let us begin with creating an intuition about what this hypothesis testing and I am sure you would have the course on hypothesis testing may be some times in your under graduate or graduate studies, but let us have some kind of an intuition around it. So, if you ask me hypothesis testing is the art of lying with statistics.

So, there is a famous book by Derrel Hagf called How to Lie with Statistics. This was written, a master piece written in 1954 and as you would see and as you will discuss the traps, the trips, the pitfalls, you will see how it is easy to prove or disprove any kind of information using hypothesis testing or other statistics in general the goal of this lecture, however you will be to make sure that you are as unbiased as possible and you do not use statistics just for validating your hunch or proving your hunch, but rather doing an unbiased analysis to see whether your hunch or your guess, what you believe is true, is actually is true or not..

So, let us for example begin with a typical industry problem and to give you a background, most of this hypothesis testing is used a lot specially in the management domain, in the soft sciences, in the social sciences, of course medical pharmaceutical is a big taker of this hypothesis testing, but it has lot of use in other in other sciences also primarily where the data is not very clear.

So, let us for example, assume that you are HR manager of the firm. So, as Dilbert has said you know in a company there are people who work and there is HR, I mean just to give you a perspective now you as an HR has a job to make other people work. So, for example, you created some HR initiatives, you made some changes, you created some training programs as to improve the performance of the employees in the company. So, for example, you have a kind of a score let us further time being not get into what is the source of the score, but there is a score which measures the quality or the performance of the employees.

(Refer Slide Time: 03:55)



Intuition

How to lie with statistics?

So, for example, you know you have a score of 25 60 43 56 32 43 47 59 39 41. So, these numbers these are scores of ten randomly picked employees in your organization. So, you hr manager of a large organization and you had all this performance enhancing initiatives taken and then, you take a random sample of 10 people and you want to see whether their performance is good or bad or ugly or whatever.

So, if these be the 10 scores of 10 randomly chosen employees for sake of ease let me share the numbers of their mean. So, for example, mean of this score come out to be x bar comes out to be 44.5, the standard deviation of the mean or other. The sample we will talk about difference of standard deviations or the variance of the mean and variance of a sample which is another pitfall. So, for this sample, the sample standard deviation is 11.41 and just to give you a perspective, sample standard deviation is nothing, but each of these values minus the mean by n minus 1. So, you would have gone through in the previous lectures just to give you remind. So, again from purely from intuition perspective this x bar represents the average also called mean and this is a measure of divergence and data.

Now, given that you have this data suppose I tell you that the minimum acceptable score for a good employee as per your research is say 45. So, what it means is that if the let mu be the mean of entire populations of employees. So, if suppose this was the gold standard that you had made that the mean of entire population of my entire employees has to be 45 or above and the sample that you took you get a score of 44.5. So, does it mean the performance enhancement initiative way of taken are not to mark or because the diversions if you add for example, the mean and the standard deviation it kind of goes up to 55 or slightly above that. So, does it mean that this is just a sampling error and if you take the mean of the entire population, there is a probability that the actual mean of the population is actually above 45. So, how do you take these decisions?

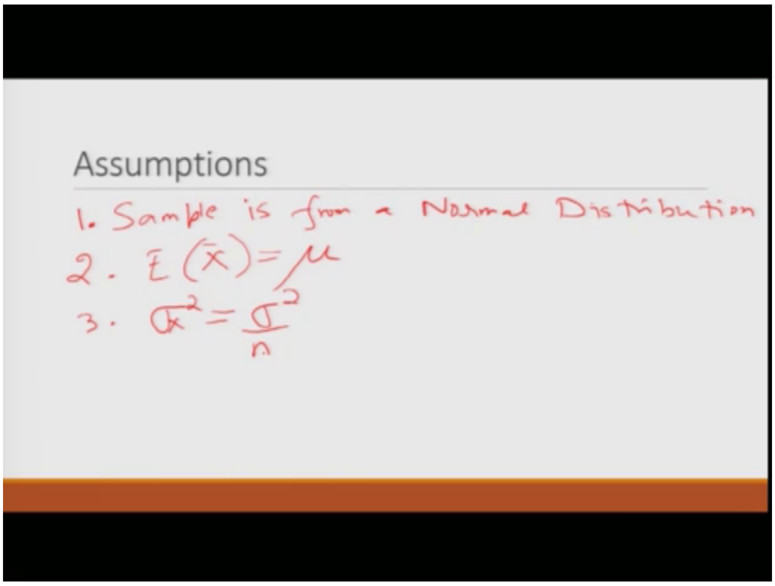
What if instead of 45 we had 56? Can you still make an assertion? So, these type of questions are answered by hypothesis testing. Now, there are two reasons why this value and this value are different. One is of course that this is just a set of ten values from an entire populations of may be hundreds or thousands or ten thousands of populations. So, it may not be a complete representative of the entire population. For example, if this be the distribution of your population where on the x axis, you have scores and on y axis, you have the proportion of people having a particular score.

So, I think we have talked about normal distribution. We will just get a feel of this later, but suppose this be the distribution, now what may have happened is that you might have because you have just chosen 10 people, you might have chosen more of the values which are tending to the lower side and hence, your score is slightly low that is one cause of error that may happen. So, this is you can call it as sampling bias. The second cause is that these numbers mind you these numbers that you are having these unlike some of these hard sciences where these numbers are obtained through a very sensitive instrument. These are numbers through some complex, some test etcetera.

So, these have a subjectivity and when there is subjectivity, there is a measurement error, there is also measurement error when you make calculations or you take out numbers or you derive values from any instruments, but these errors are actually much larger in action when we deal with subject like management or human resources or this softer sciences because these numbers for example would have come through some tests and those test itself may have some kind of a measurement errors. So, there is a measurement errors, there is a sampling bias and hence, these numbers may not be true representative of the entire stream of populations.

Now, we will talk about errors in measurement and how to deal with it later, but right now our focus is that given that we get these kind of numbers, what is the probability or what is the chance that we are still able to say that yes my sample as a mean which is above 56 or above 44 or whatever number you take. So, for this two happen again let me first up all also give you that whenever you give this kind of test, we assume few things number one, we assume that the sample is from a normal distribution.

(Refer Slide Time: 12:12)



Assumptions

1. Sample is from a Normal Distribution

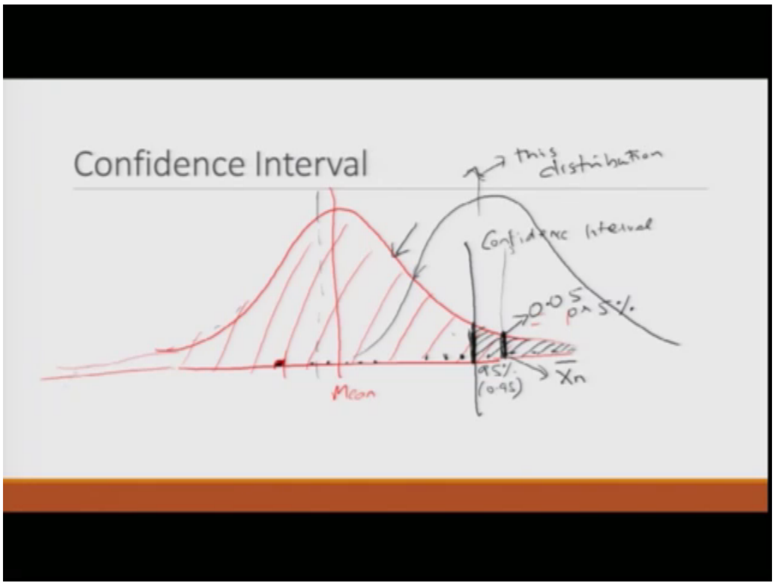
2. $E(\bar{x}) = \mu$

3. $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$

So, this is assumed for all kinds of hypothesis testing. Number two and this follows from this we assume that the samples have been taken as randomly as possible. There is no bias. Hence, the expected value, the mean of the sample is actually the mean of the population.

We also this further implies that the square of the standard deviation of the sample is equal to the variance of the population divided by n. Now, these are basically derived from the assumption of normal distribution will not get into the mathematics of it, but the essence is that whenever we are doing a hypothesis testing, we will talk about assumptions later, but most of the case, most of the hypothesis testing we assume that the population is a normal distribution.

(Refer Slide Time: 13:50)



Confidence Interval

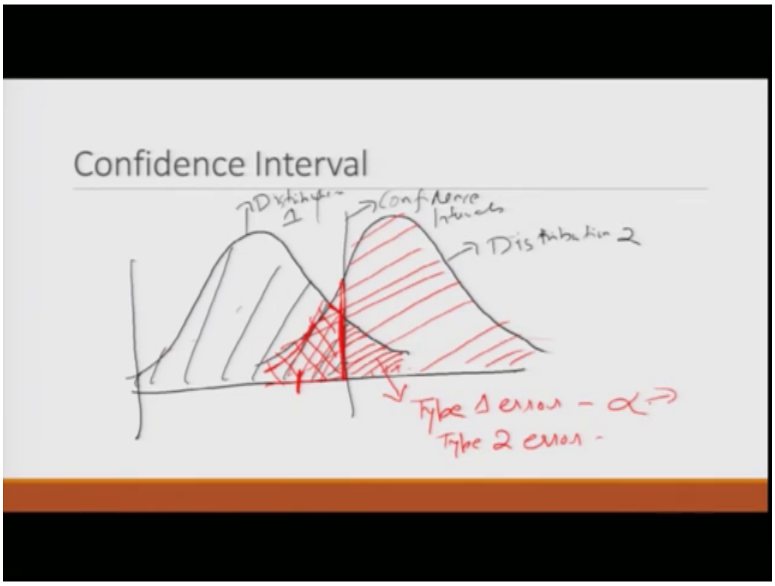
Now, let us talk about the confidence interval and then, understand how to make a reasonable hypothesis. So, as we said we assume that the population is a normal distribution. So, let this be the normal distribution curve, this be the mean. Now, if you recall when we talked about these probability distributions, we learned that the probability of any value happening in this entire area is equal to 1. Mind you do not ever touch the axis, but they just reach the axis somewhere in infinity same thing on the negative side.

So, when we are doing a hypothesis, we assume that the sample now all the values you have taken, you may have taken, chosen one value which is around this place. You may have chosen one value here, one value here, one value here, one value here, one value here. Hence, the mean may not be exactly this, but probably the mean that you are getting is slightly lower. So, what you do is, you define something called a confidence interval which means you are saying given that my sample is not the entire population. There is a chance of my sample parameters to be slightly diverged from the actual population parameter.

Hence, let me see whether that diversion is acceptable to me. So, for example, you define a confidence interval. Let us say 95 percent 0.95 which means that this extreme area of the normal distribution represents only 5 percent of your total sample. So, now what you are trying to say is that if I have taken a sample where the value ranges, if the value is less, then this 95 percent which means if I have to repeat the same experiment again and again and again and for the 95 percent of the time if my sample mean comes out to be less, then this particular 95 percent red line I would say that that sample has come from this particular, from this particular distribution as mentioned here.

However, suppose I have collected some samples and the values where say some where here, here, here, here, here, here, here and hence, the mean was coming somewhere here. Let this be the new mean. Then, I would say that it is only 5 percent likely that these values are coming from this particular distribution probably, then the values are coming from a different distribution which has. So, may be this is another distribution. Now, probably the values are coming from this distribution. So, if I have to simplify and tell you the choices to decide whether the values are coming from this distribution or this distribution.

(Refer Slide Time: 18:42)



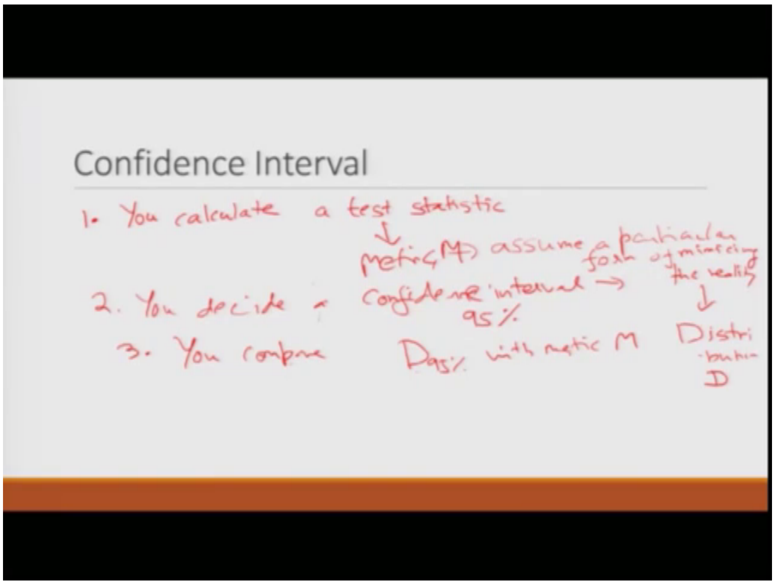
Confidence Interval

So, there are two distributions. So, what you do is you set up your cutoffs. These cutoffs are called confidence intervals.

So, you say that if my value is anywhere in this range, then I assume that my value is coming from this distribution. Distribution 1, let this be distribution 2. However, if the value lies somewhere here even here I say it is coming from this particular distribution; so this particular dividing line that you have that shows that there is a scope of error. What may have happened is that actually the value came from this particular distribution which is distribution 1, but because it was in this extreme position, you ignored it. The other kind of error that may happen is that you got a value somewhere here. You assumed it to be coming from distribution 1, but actually it was coming from this distribution 2 and hence, you accepted it.

So, this error that happens where you reject something despite it coming from the original distribution, this is called type 1 error or called alpha and the error where despite this coming from this part, despite it coming from this part, you still reject it. That is called type 2 error. So, both the errors are there. So, we will talk about the errors later. For your intuition, we must focus primarily on the type 1 error for now and let us understand that type 1 error is primarily the error of rejecting a distribution or rejecting a sample, rejecting a hypothesis despite it coming from the distribution, because it was beyond the confidence interval level. So, this is the primary intuition behind the entire philosophy of hypothesis testing you primarily do three steps to do any kind of hypothesis testing in step 1.

(Refer Slide Time: 22:17)



Confidence Interval

1. You calculate a test statistic

↓

metric (M) → assume a particular form of mimicking the reality

↓

Distribution D

2. You decide a confidence interval →

95%

3. You compare D95% with metric M

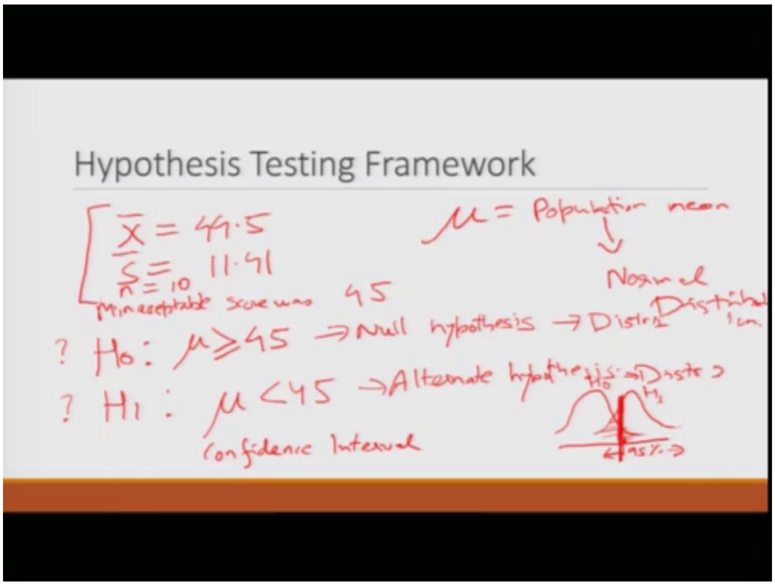
You calculate a test statistics which is nothing, but based upon your kind of hypothesis that you want to perform, the kind of question you want to answer and based upon the kind of values that you got, you calculate a metric. This is nothing, but a metric, then you decide a confidence interval.

So, for example, you may say that I want to be 95 percent sure that the values I have got from this particular distribution. So, you decide a confidence interval. So, in case it is 95 percent, then what you do is you now here when you are calculating a test statistic or a metric, you assume a particular form of mimicking the reality what you says that because my sample size is lower because this is a particular kind of a question that I am trying to answer, you assume that a different kind of a distribution, not exactly a normal distribution is what is relevant here.

So, this is suppose to mimic the reality. Now, this is supposed to be a distribution. This reality is supposed to be a distribution d. So, what you do is you compare the value of d at 95 percent with this metric m and if you believe that whatever be the 95 percent confidence interval, this metric m still lying within that interval. You say that well whatever values I have got that belongs to this particular distribution. If they do not, then you reject it.

So, this is the broad intuition behind it.

(Refer Slide Time: 25:37)



Hypothesis Testing Framework

$\bar{X} = 44.5$

$\bar{S} = 11.41$

$n = 10$

Min acceptable score was 45

μ = Population mean → Normal Distribution

? $H_0 : \mu \geq 45$ → Null hypothesis → Distri[illegible]

? $H_1 : \mu < 45$ → Alternate hypothesis → Distr

Confidence Interval

So, if we go back to our example of you getting score from 10 employees where if you recall, we found that the mean was 44.5, the standard deviation was 11.41 and minimum acceptable score was 45. So, what you do is, you create a question. So, hypothesis is basically it all start with your hunch. For example, your hunch was that well after all the efforts that I have taken, the score is still above 45.

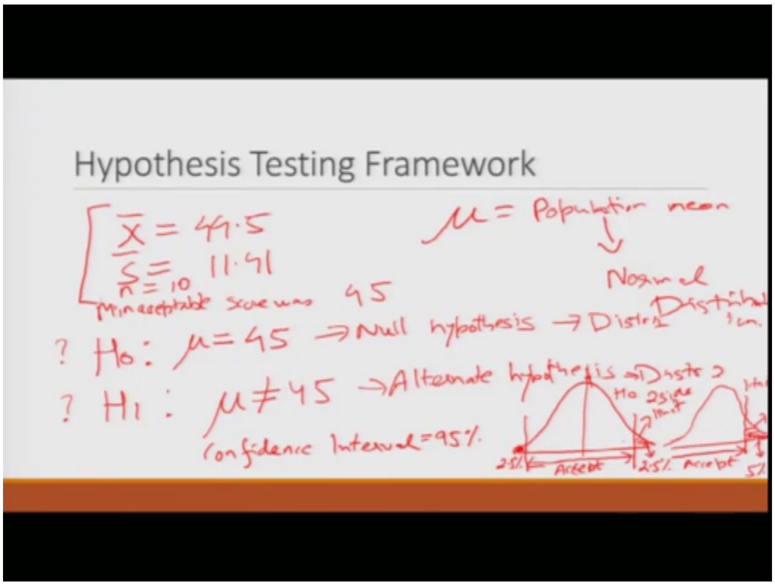
So, your hunch is that mu which is the average score, mu is population mean. So, the mean of the population or the average score of this mean, so your hunch is that your mu is greater than 45 given that these were the ten observations, given that this was the sample that you received, sample you know n was 10. So, you made ten observations, 44.5 was the mean, 11.4 was the standard deviation given that you get this information, you make a hunch or you make a hypothesis that the mean of the population is greater than 45. So, this is just a hypothesis. Whatever hypothesis you make, you call it the null hypothesis.

Now, what is the alternate hypotheses, but may be wrong. The alternate hypothesis is that h 1 denoted by h 1 is that the mean is less than 45. So, this is called the alternate hypothesis. Now, your goal is to decide whether you should choose this one or this one. How do you choose number 1? You assume that all this population is a normal distribution; you assume that you do not know it is a normal distribution, but you do not know what that mu is or what are the distribution looks like.

So, you have two choices. This represents one distribution, distribution 1. This represents another distribution, distribution 2. You have to choose this, one of these distributions. The complexity comes because these two distributions are not completely separate. There is a common area as let me just briefly make the diagram again. So, this is one distribution, this is another distribution, this is the distribution of the null hypothesis, this is the distribution of the alternate hypothesis. The issue is that there is a common area out here. So, how do you decide if a point comes here that whether it belongs to this distribution or this distribution?

Hence, you take recourse of the confidence interval. You say I make a cut off here. Let this cutoff be at 95 percent. It can be 95 percent, it can be 99 percent or it may be whatever. 95 percent here means simply that of my entire population, 95 percent of the values are in this region. So, 95 percent simply means that of my entire population, 95 percent of the population is going to be in the left side of it. So, now you want to choose whether you go with this or you go with this. There can be a different kind of hypothesis testing assignments where instead of an inequality, you are testing for equality.

(Refer Slide Time: 30:52)



Hypothesis Testing Framework

$\bar{x} = 44.5$

$s = 11.41$

$n = 10$

Min acceptable score was 45

μ = Population mean → Normal Distribution

? $H_0 : \mu = 45$ → Null hypothesis → Distr 1

? $H_1 : \mu \neq 45$ → Alternate hypothesis → Distr 2

Confidence Interval = 95%

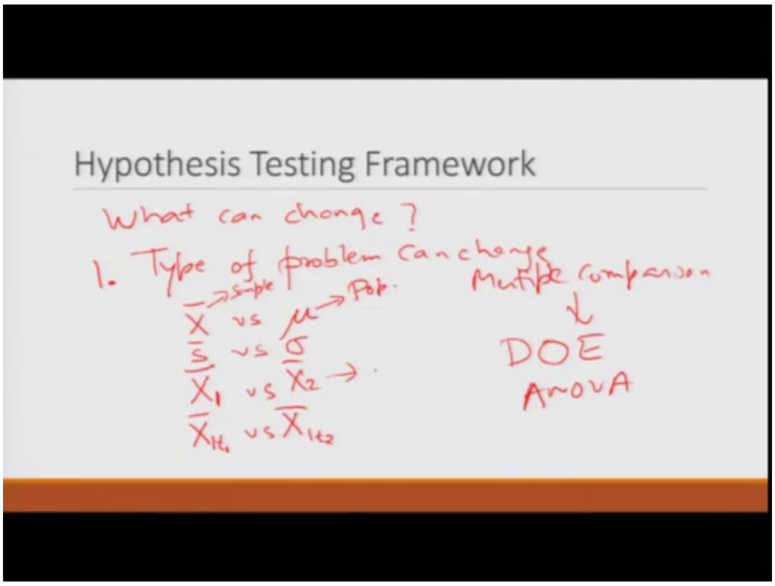
You do not know whether the score is good or bad. You just want to see whether for example, you could did some experiment on a set of people or set of patients in pharmaceutical contexts or some other kind of experiments and when you got some new observations.

Now, you want to see whether has something changed from the population. So, here you are not concerned with whether it is greater or less than, you want to be concerned whether it is different or not. So, in this case if this be my hypothesis, I define cut offs on both sides and say that if the value I get is within this range, then I accept it. So, in both cases a normal distribution, in other case let me just draw for comparison, you just make one side and you say that if this is if it is anywhere here till minus infinity, you accept it.

So, these are two different kinds of problems almost the same variety in ones case you are making two side limits or it is called two tailed and here we are having just one single tailed, one tail. So, typically what happens is because a normal distribution is symmetric, if this area is 5 percent for example you are doing a confidence interval of 95 percent for the difference that comes is if it is a single tailed, you take 5 percent on this side and it can be on the left side also.

In this case because it is evenly distributed, you divide it between 2.5 percent and 2.5 percent. So, this is the problem assignment. Now, what can change? In this a lot of things can change.

(Refer Slide Time: 34:00)



Hypothesis Testing Framework

What can change?

1. Type of problem can change

$\bar{X} \xrightarrow{} \text{Sample}$ vs $\mu \rightarrow \text{Pop.}$

$\bar{s}$ vs $\bar{\sigma}$

$\bar{X}_1$ vs $\bar{X}_2 \rightarrow$

$\bar{X}_{1t_1}$ vs $\bar{X}_{1t_2}$

Multiple comparison

$\downarrow$

DOE

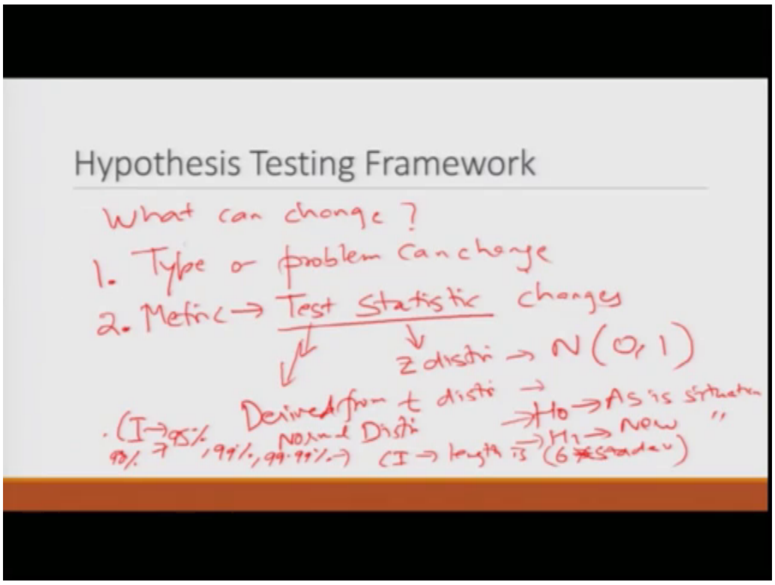
Anova

So, this being a frame work what can change number one, your type of problem can change. For example, here I give you example of comparing a sample mean verses population mean. Now, whenever you read the literature, I mean just keep a note of this that whenever there is a mu, this refers to populations and this x bar refers to sample. So, this is the general notation. So, one is you compare these verses. This you can compare your standard deviation with the populations standard deviations. Here you have compared verses population, you can compare one sample mean with the another sample mean, you can compare the same sample when same sample one say at time T 1 versus time T 2.

Now, these are just one verses, one comparison, then you can make multiple comparisons. As we move into multiple comparison, we move into a field called design of experiments. The basic one start with something called anova, where what you are doing is you are comparing multiple samples like this and not just two, but multiple samples at one go.

So, the type of problem can change and when type of problem changes, this also means that the metric or the test statistic.

(Refer Slide Time: 36:12)



Hypothesis Testing Framework

What can change ?

1. Type or problem can change

2. Metric → Test Statistic changes

z distri → N(0, 1)

Derived from t distri → Normal Distri

→ H0 → As is situation

→ H1 → Now "

. CI → 95% 90% → , 99%, 99.99% → CI → length is (6 * standev)

This is the word which is used in the domain, the test statistics changes. So, in one case we assume because when it was simple suppose you take a large sample from a normal distribution and you just want to compare the means that you assume to be a normal distribution. So, this test statistics actually comes from z distribution which is nothing, but a normal distribution with mean of 0 and standard deviation of 1. If you are taking a smaller sample, you have T distribution which is a different distribution and hence, the formula of this metric changes.

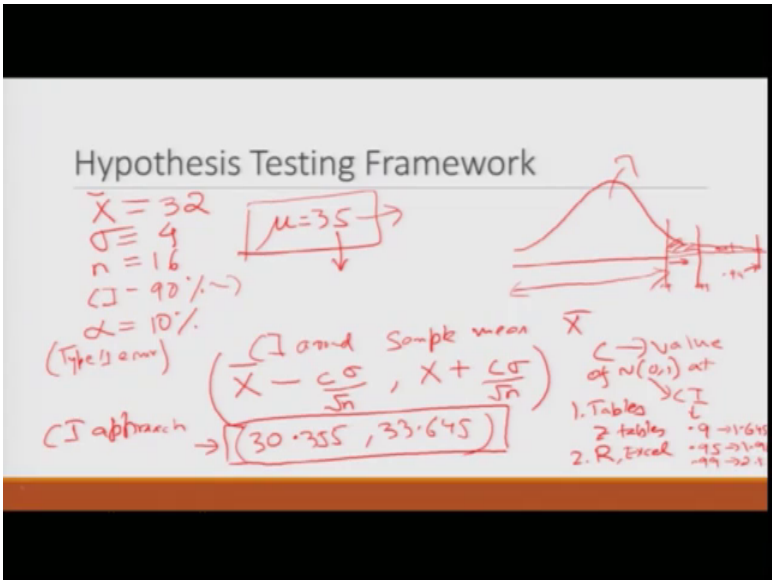
When you make comparison of variance, the square term involved and hence, you no more use these distributions. You make distributions which are more complex combinations of linear of normal distribution. So, in most of the cases, you will see that these test statistics are derived from normal distribution, but because they are derived from normal distribution, they may not be an exactly normal distribution. So, this test statistics changes and what you finally do is that and the other thing that change is the confidence interval. You can decide that you want to take 95 percent confidence interval which means that if 95 percent of the values are within that range, you accept it. You can extend it to 99 percent, you can extend it to 99.99 percent.

So, typically what you do is that the null hypothesis is normally as is situation and alternate hypothesis is the new situation. So, for example, if you are testing on a particular impact of a drug, so here null hypothesis that no change happens and H 1 alternate hypothesis is that there is a change. So, depending upon how much conservative you want to be between accepting this one, this one, you change your confidence interval. So, when you talk about things like six sigma, so six sigma is nothing, but a confidence interval whose length is six times the standard deviation; so 6 into standard deviation which is some 99.99 whatever.

So, it is a very extreme. So, when you are into quality control, you typically tend to keep this very high confidence interval. Then, you are into more exploratory or more softer sciences, the typical norm. For example, in most of the management field the typical norm is 95 percent. Sometimes you may also go to 90 percent depending upon the quality of data that you get and how conservative you want to be about any kind of a measurement (refer time: 40:00).

So, this is the broad hypothesis testing frame work. Now, let us just take an example. So, all that we have just discussed it actually starts making some sense. You get a feel of what we are trying to do. So, for example, we have a case where you took some samples.

(Refer Slide Time: 40:30)



Hypothesis Testing Framework

You took 16 samples. The mean was 32, the standard deviation was 4 and as we said there were 16 samples which means that you took 16, you did something and then, you took 16 values out of the populations and then, you want to test the hypothesis that this mean of the new population or whatever the distribution that do not know is say 35. So, this is what you want to test whether a value of 35 is possible or not.

So, what you do is, suppose you want to be taking a confidence interval of 90 percent which means the alpha which is type 1 error is 10 percent. Type 1 error which means that you are open to the fact that you may be making a mistake in 10 percent of the cases. So, what you want to actually assess is that a value of 35 if I draw the distribution curve again, you want to be sure that a value of 35 lies somewhere here. So, this is what you are trying to do. So, now there are two approaches of doing it. In one you actually create given that 32 is the mean, you create a confidence interval around sample mean x bar. So, in this case the confidence interval around this is given by x bar minus c sigma by root n x plus c sigma by root n. This c is nothing, but the value of at confidence interval.

So, for example if you take confidence interval of 0.9, then this value comes to 1.645. If it is 0.95, this comes to 1.97 and so on. It keeps increasing. If it is 0.99, it comes to 2.58 because these curve never reaches, touches the x axis and it goes till infinity. It gets thinner and thinner. So, any incremental movement that you make, if you move your confidence inter from here to here to here to here, you know just for a 5 percent for example 0.9. This was 0.9 and this was 0.95 and this was 1.99. You will see that the distance you have to cover is actually, so this graph is not the right way. So, if this be the 99, so probably 99 will lie somewhere much farther 1.642, 0.963, 0.58. So, it keeps because the curve is getting thinner and thinner. There are less number of observations that you can make in this region.

So, probability is very low here you know. So, this is like high probability. So, this increases. So, what you do is you create 90 percent confidence interval. In this case, this comes out to be if you do the calculations, 30.355, 33.645. Now, how do you get these values of confidence interval? For different normal distributions, you have tables, the tables of these are called z tables. If you are using r or you are using xl, there is a formula that you can use which is nothing, but the formula of the normal distribution and hence, you can get these values.

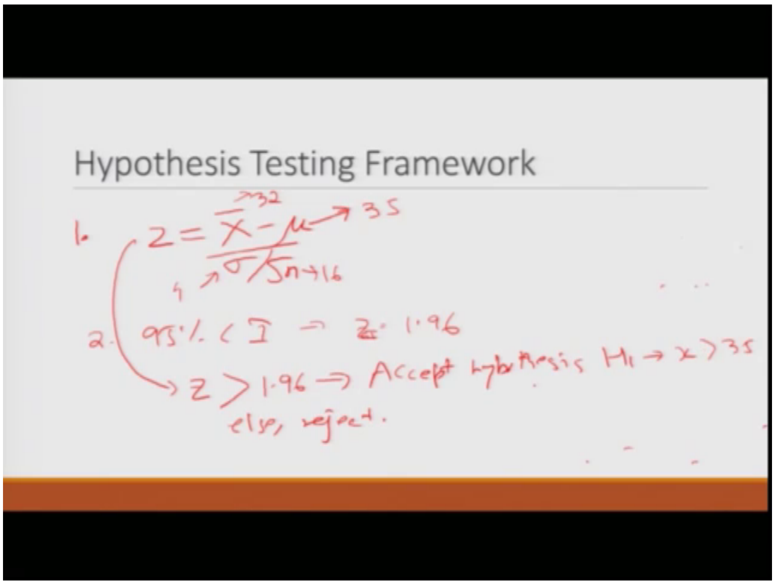
So, as an analytics practitioner, how you get these values c will be list of concern, but to understand this c and see whether this c actually applies here or not is what you should be focused upon. So, here you for example you see that with 90 percent confidence, you can say that if I observe a value of 32, the confidence interval is 30 to 33.645. Here what we have done is you assume that your population mean is equal to your sample mean and you want to see what is the variation.

You assume that your sample standard deviation is also equal to a population standard deviation because in most cases because you do not know the population, you never know the populations standard deviation. So, standard deviation of your sample is considered a good proxy for or the only option, the only proxy even if it is not that good and if you go to the literature, you will see why it actually creates a lot of problem, but in absence of any other information that is the best the only proxy we have. So, we have no choice. So, then this be the range of values that will be there if this was the distribution.

So, here if you say mu is equal to 35 or you want to see whether a value of 35 is part of this range, you say no because 35 is out of this range. Hence, you reject that 35 is the value acceptable in this range. So, if your alternate hypothesis was or null hypothesis whatever was that 35 is part of this particular distribution. You reject it. You say the value of x cannot be 35 assuming that 90 percent confidence interval is there. So, one way of doing is to create this confidence interval. So, this is the ci approach of doing hypothesis testing which is very popular among people who come from engineering or a laboratory science field because they have always made this confidence interval on anything that they create.

So, people like systems engineers and people like production engineers who are about quality control, they tend to approach this problem from a confidence interval approach. If you look at the management disciplines, the social sciences, the pharmaceutical industries, so in these disciplines the alternate approach actually emerged. So, both are actually coming from the same family or same kind of approach. The formulation is exactly same, but you know this is the different way of doing and this is what is more popular. So, this is the table based or the test statistic approach.

(Refer Slide Time: 49:18)



Hypothesis Testing Framework

1. $$z = \frac{\overset{\to 32}{\bar{x}} - \mu \to 35}{\overset{4 \nearrow}{\sigma} / \sqrt{n} \to 16}$$

2. 95% CI → z_{ci} 1.96

→ $z > 1.96$ → Accept hypothesis $H_1 \to x > 35$
else, reject.

So, here what you do is as we said in this case you calculate your z equal to x bar minus mu by sigma by root n. So, mu in this case was 35 which we wanted to test. A x bar was 32 which we got from the sample, this was 4 n was 16.

So, you do this calculation. For example, if you wanted to test at 95 percent confidence interval which is very common in this domain, then the z for 95 percent confidence interval is equal to 1.96. So, now, let me call it z ci. So, now if your z that you get here this is greater than 1.96, then you say that there is a difference which means 35 is an acceptable value. If it is less, then accept hypothesis else reject.

Now, note that this accept and reject also come with a kind of rider. Basically if it is greater than this, you accept the hypothesis h 1 which says that the value is 35 or z is greater than 35, x is greater than 35 or whatever, but that does not mean that you accept this. It means just that you have sufficient widens to reject the null hypothesis. If you get a lower value, it just means that you do not have sufficient evidence to reject the null hypothesis. That is it. It does not say that it is true or false because there is that 5 percent error limit that we have.

So, we should always be very clear on this that you are accepting the alternate hypothesis does not mean that the null hypothesis is wrong. In all situations, there is always a margin of error in which we are playing.

So, in the next lecture, we will take some examples and see how this hypothesis test works for different scenarios. We will talk about hypothesis testing for independent samples, for dependent samples. We will also talk about some non-parametric tests where you do not assume any distributions and still go about doing your statistical inference.

Thank you very much.