Introduction to Data Analytics

Prof. Nandan Sudarsanam and Prof. B. Ravindran

Department of Management Studies and

Department of Computer Science and Engineering

Indian Institute of Technology , Madras

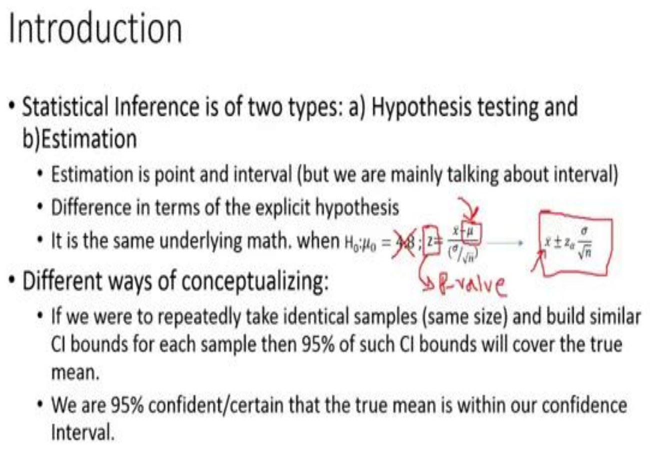
Module – 03

Lecture – 13

Inferential Statistics – Confidence Intervals

Hello and welcome to our next lecture in inferential statistics. Today, we are going to be talking about confidence intervals.

(Refer Slide Time: 00:21)



Introduction

- Statistical Inference is of two types: a) Hypothesis testing and b)Estimation
  - Estimation is point and interval (but we are mainly talking about interval)
  - Difference in terms of the explicit hypothesis
  - It is the same underlying math. when $H_0{:}\mu_0 = \cancel{48}$; $\boxed{z} = \frac{\bar{x} - \boxed{\mu}}{(\sigma/\sqrt{n})} \longrightarrow \boxed{\bar{x} \pm z_\alpha \frac{\sigma}{\sqrt{n}}}$
- Different ways of conceptualizing:
  - If we were to repeatedly take identical samples (same size) and build similar CI bounds for each sample then 95% of such CI bounds will cover the true mean.
  - We are 95% confident/certain that the true mean is within our confidence Interval.

So, statistical inference even in terms of classification mainly is discussed in terms of hypothesis testing and estimation. So, these are the two broad categories in statistical inference. And, hypothesis testing is something that we have discussed in good detail in this class; we have talked about single and two-sample tests – various tests. Today, we are going be primarily talking about estimation. And, you might notice that, almost any text that talks about statistical inference, talks about these two topics. And, some of them might be introducing estimation or confidence intervals before hypothesis testing. But, that does not really matter. Essentially, these are two sides of the same coin if you ((Refer Slide Time: 01:07)) So, today, we are going to be talking about estimation.

Now, the idea behind estimation is that estimation can be in terms of point or interval. What we mean by that is it is a same core concept as what we introduced with inferential statistics during hypothesis testing; which is that, we are interested in some population parameter. What we mean by that is that, there is this concept. So, the examples that we have used in this class are things like amount of phosphate in our blood, the height of tenth standard students in public schools in India. In all these cases, you define some population and you are interested in some parameter. More often than not, we would discuss the parameter being the mean. So, what is the average amount of phosphate in blood? What is the average height? But, it does not… That is just one of the parameters. That is the most common one. But, it can be other parameters. So, stepping back, we are interested in some population parameter. But, you do not know this parameter. It is like it is some truth that you do not already know. If you did know that, there would be no need for any of this or any of the statistics. But, what you do have is a sample. So, you have 5 data points, 10 data points, 20 data points, 30 data points. Some sample from this population.

And, what you are most interested is you are most interested about this population parameter. So, in hypothesis testing, you would hypothesize that, this population parameter is equal to 4.8 or 2.3 or it is less than 4.2. And then, you would go about and look at this sample and see if that is true or not. With estimation, you are not having any hypothesis; you are not having any hypotheses in mind in that sense. What you are trying to do is you are trying to take the sample. And, with point estimation, you are trying to come up with a single point estimate of the population parameter. And, that might seem fairly straightforward. So, for instance, let us say you are interested in the population parameter, which is the average amount of phosphate in blood. And, you took a data, you took some sample. And, that sample was about 20 data points. A simple point estimate of the population mean could be the sample mean. So, you take the sample of 20 data points; take their average. And, that is your best; that could be one of your best point estimates of the population mean. So, point estimate just means you are making as good a guess on the population parameter based on the data that you have.

But, today, we are going to be talking… And more interestingly, this is what gets… This is what people are more interested in, which is interval – estimation in the form of an interval. So, this goes back to again the core concept with hypothesis testing, which is fine. You do not know the population mean; you have a sample mean; and, you acknowledge that, if you go to take another sample of another 20 points, you might not get the exact same value. And, both these values the first time around and the second time around might not be exactly equal to population mean. If it is not exactly equal to the population mean, then can I come up with some range around my point estimate. So, I have a point estimate, which is actually my sample mean; my sample mean is my best bet; let us say at my population mean. But, I acknowledge that, I might not have exactly hit target.

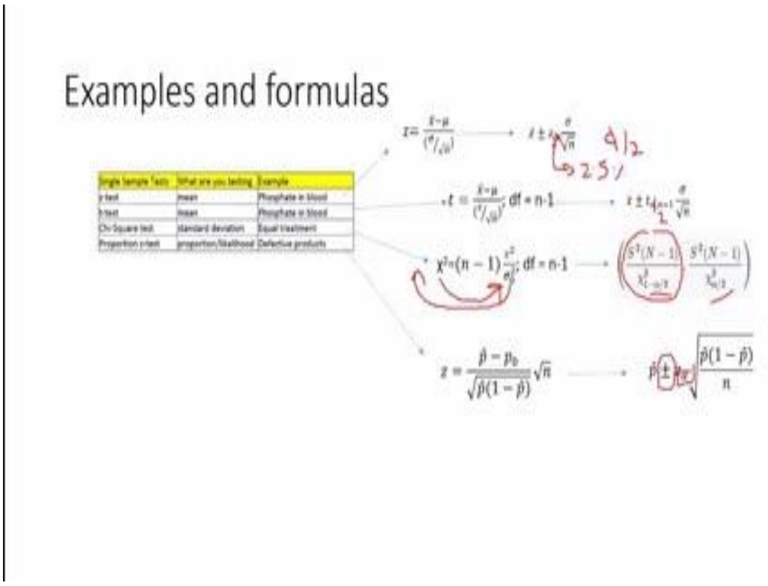
The population mean might be a little higher or a little lower than my sample mean; in which case, I ask the question – can I come up with the range around this sample mean? By which means it is essentially like I am giving myself a margin of errors by which I am fairly certain that I have covered the population mean. So, that is the goal that we are going to embark upon. And, in many ways, it is the same map, because we have introduced hypothesis testing; it becomes a little easier, so that we can reason by the same logic in map that we have already discussed. So, let us do that. So, the core idea is that, let us take an example that we have looked at many times. So, which is that we might be hypothesizing the amount of phosphate in blood is equal to exactly 4.8.

Now, we discussed that, in confidence intervals, you do not have this number; you do not come up with the hypothesis. You are only interested in coming up with some bounds around your point estimate. So, what you essentially do? One way of thinking about confidence intervals given that we have already introduced hypothesis tests is well, for different values of mu naught. So, let us say you have some dataset and you calculate some x bar. Essentially, 4.8 gets plugged in out here to calculate your z-statistic. If you are doing a z-test, you are given a sigma; if you are doing a t-test, you take an s; but, in neither case, that gets plugged in; n is again the number of data points. So, you get some z-value – some z value; that is out here. And, based out of that z-value, you calculate some probability; you calculate a p value. Now, the core idea with hypothesis test was that, if this p value is really small, you reject the null hypothesis.

So, the question with confidence intervals – one way of thinking about is given that, I do not have some hypothesis, I ask myself the question within what range can my mu's be such that I will calculate a zee such that I will get a p value, which I will not reject. I am just going to repeat that; given that you do not have a mu naught, you can think of a confidence interval as what is the range of values that mu could potentially be such that given that, for a given dataset, you will get some x bar, some sigma, some root n such that you will calculate a value z such that you will get a p value, which you will not reject. So, if you would not reject, that means you need to have some bounds. Suppose you start off by saying well, I am going to reject any p value less than 0.05. So, that is something you started off with. Now, given that you started off with that; then, is there some range… For a given data set, is there some range of mu's such that you would not be rejecting this hypothesis test. And essentially, to compute that, all you do is just rearrange the terms out here. So, you keep the mu naught on… – the mu or the mu naught – I am using those two terms here interchangeably. But, you keep that on one side and you essentially move the terms to the other side to get this formula for confidence interval.

So, typically, if you know the formula for the hypothesis test or the test statistic, you can just essentially rearrange it. But, the core idea is that, this is your point estimate, which is your x bar. So, for your best estimate of mu is your x bar; but, you create a margin of error or you create a range around it as plus or minus the z associated with this alpha. And, alpha here is that 0.05 that you said. Essentially, you said within a certain range. So, within that range and sigma and square root of n are the same. Another way a more formal definition associated with confidence interval is that, if we were to repeatedly take identical samples of the same size and build similar confidence interval bounds for each sample, then you are building a bound such that 95 percent of such confidence interval bounds will cover the true mean. Or, in other words, we are 95 percent confidence slash certain that the true mean mu is within our confidence.

(Refer Slide Time: 09:36)



Examples and formulas

Single Sample Tests	What are you testing	Example
z-test	mean	Phosphate in blood
t-test	mean	Phosphate in blood
Chi Square test	standard deviation	Equal treatment
Proportion z-test	proportion/likelihood	Defective products

$$z = \frac{\bar{x}-\mu}{(\sigma/\sqrt{n})} \longrightarrow \bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

α/2

2.5%

$$t = \frac{\bar{x}-\mu}{(s/\sqrt{n})};\ \text{df} = \text{n-1} \longrightarrow \bar{x} \pm t_{\frac{\alpha}{2},n-1}\frac{\sigma}{\sqrt{n}}$$

$$\chi^2 = (n-1)\frac{s^2}{\sigma_0^2};\ \text{df} = \text{n-1} \longrightarrow \left(\frac{S^2(N-1)}{\chi^2_{1-\alpha/2}}\ \frac{S^2(N-1)}{\chi^2_{\alpha/2}}\right)$$

$$z = \frac{\hat{p}-p_0}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{n} \longrightarrow \hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

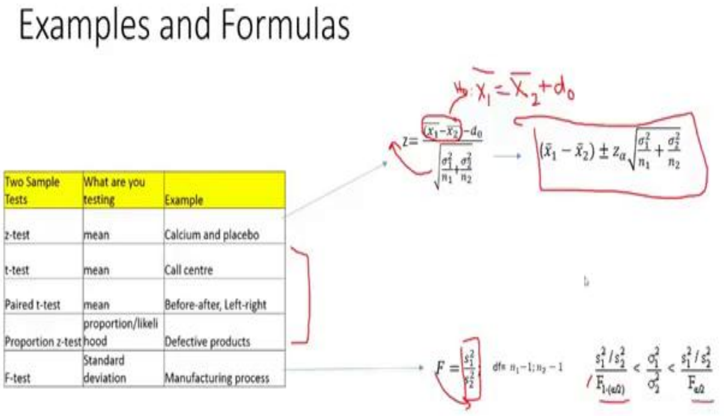
So, for single sample tests, the process of creating confidence interval is fairly straightforward. I explain to you how in the z-test, it essentially just becomes a rearrangement of terms and this sigma by… This concept of sigma by square root of n essentially goes towards the z and then the x bar goes here such that it is a plus or minus. And, that is how you get the formula for that interval. The same core idea for the t-distribution; the formula is no different, except that, the t-distribution gets also defined by the number of degrees of freedom; so, not just the alpha, but the number of degrees of freedom. And, in all these cases, mind you, because I am putting a plus or minus, this is the equivalent of the two-tailed z-test or the two-tailed t-test. You could also create a one -sided bound if you were interested. And, that would be again the formulae equivalent of having one sample test. Again it would be the same sigma divided by square root of n. The formula itself would not be different. But, the way this alpha gets used up will be different. Again it goes back to the core concept of how you would shade that region under the probability distribution.

With the chi square distribution, that rearrangement is not obvious. To some extent it is. But, the plus or minus is not, because you do not have a plus or minus term; it is essentially like this sigma naught goes out here, the chi square distribution comes down here. But, the way we differentiate between the lower bound and the upper bound is by changing the alpha in the bottom of the chi square distribution. So, it is the same rearrangement; it is a same core concept of taking the hypothesis test and rearranging;

that is, in this case, it would be to put the sigma naught out here and bring the chi square down here and you would have the same formula that you see here. But, you get an upper bound and a lower bound by looking at the 1 minus alpha by 2 and alpha by 2.

And, by the way, this notion of alpha by 2 depends on how you define it. Now, if you say I want a 95 percent confidence bound; that means, you are left with 5 percent. And, if it is two-tailed test, that 5 percent gets divided into 2.5 percent times 2. That is how you get the alpha by 2. So, it really depends if someone starts by stating alpha and you know it is a two-tailed test; then, technically, the correct way to do this would not be to just have an alpha out here, but it would be to have an alpha by 2, so that you are being technically correct. And, the same thing goes for here as well; you will have an alpha by 2. Again the same core idea with respect to the z-test. If this alpha is more generic term that I have used out here. So, if somebody comes and says I need a two-tailed test; so, there is a plus or minus and the alpha gets divided by 2. But, if it is a one-sided test, that alpha can stay as alpha. So, depends on how it gets firmly defined typically. If it is a two-tailed test, you will represent it as alpha by 2. But, it is… Again if you look at it, it is a same rearrangement from your test statistic to create the confidence interval.

(Refer Slide Time: 13:06)



Examples and Formulas

Two Sample Tests	What are you testing	Example
z-test	mean	Calcium and placebo
t-test	mean	Call centre
Paired t-test	mean	Before-after, Left-right
Proportion z-test	proportion/likelihood	Defective products
F-test	Standard deviation	Manufacturing process

$H_0: \overline{X}_1 = \overline{X}_2 + d_0$

$$z = \frac{\boxed{\overline{x}_1 - \overline{x}_2} - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightarrow \boxed{(\bar{x}_1 - \bar{x}_2) \pm z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$F = \boxed{\frac{s_1^2}{s_2^2}} \quad df = n_1 - 1; n_2 - 1 \qquad \frac{s_1^2 / s_2^2}{\underline{F_{1-(\alpha/2)}}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2 / s_2^2}{\underline{F_{\alpha/2}}}$$

So, the same idea goes towards two-tailed tests. Just to give you an idea of how it works for a two-tailed test, I have given a single sample for the z-test, for the t-test and the proportional z-test. We have consciously left it out; that might be a part of your assignments that you could work on. But, the idea is the same. We are interested in this term. We are interested in the term x 1 bar minus x 2 bar. And so, we want to create a confidence bound around x 1 bar minus x 2 bar. If you remember this simple way of thinking about x 1 bar minus x 2 bar was to say to test the null hypothesis that x 1 bar is equal to x 2 bar. But, that is logically the same as asking the question – what is the confidence bounds around x 1 bar minus x 2 bar? And, seeing if that essentially covers a 0 or not. Of course, this d naught is an extra term.

Suppose you were interested in a hypothesis that looked more like this; if this was your null hypothesis, then you could – you would have this d naught being something nonzero; otherwise, d naught is just equal to 0. But, this is the idea. So, you are interested in some bounds around the term x 1 bar minus x 2 bar. And, again you would do the same logical rearrangement. This goes here and the bounds go around x 1 bar minus x 2 bar. And, you are essentially creating bounds around this value. So, there is x 1 bar minus x 2 bar and what is your range around that; great. Again similar to the chi square test, the F-test is not so straightforward. So, I am including the formula associated with that out of here. You do the same thing that you did with the chi square test, which is you have the same s 1 square by s 2 square. That is your core upper bound and lower bound. But, you divide by this F. This F kind of comes down and you divide by that. But, the lower bound is 1 minus alpha by 2 and the upper bound is alpha by 2. Of course, also remember that, when this is a two-tailed test, the alpha that you see here becomes alpha divided by 2. If it is a one-tailed test, meaning you just had a plus or a minus; you can keep that as alpha. I hope that clarified this concept of confidence intervals.

Thank you.