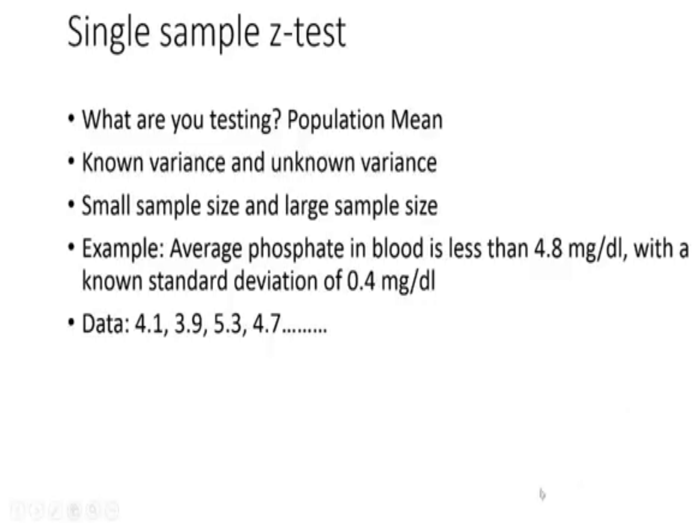
Introduction to Data Analytics

Prof. Nandan Sudarsanam and
Prof. B. Ravindran
Department of Management Studies and
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module – 02
Lecture - 10
Inferential Statistics – Single Sample Tests

In our previous lecture, we spoke about the need for… We tried to motivate the need for inferential statistics through the context of hypothesis testing. So, we spoke about why we needed it, where it would apply and so on. We concluded that lecture by coming up with a template, coming up with a rubric – essentially of what it is that one needs to do with hypothesis testing. So, we started off with instructions like you need a null hypothesis, an alternate hypothesis; and, the whole thing was fairly general. So, in today’s lecture, we continue our focus on hypothesis tests. And, we are going to talk about something called single sample tests. In the last class, you would have seen that we gave two sets of examples; we gave examples of single sample tests and two sample tests. And, today, we are to going focus on single sample tests. And, what we are going to do is we are going to illustrate that template that you saw with by illustrating one test. So, the test is going to be the single sample z-test and we are going to show you the mechanics behind it and kind of give you the reason of why we do some of the math that we do. And then, we will talk about some of the other tests that are there as well.

(Refer Slide Time: 01:41)



Single sample z-test

- What are you testing? Population Mean
- Known variance and unknown variance
- Small sample size and large sample size
- Example: Average phosphate in blood is less than 4.8 mg/dl, with a known standard deviation of 0.4 mg/dl
- Data: 4.1, 3.9, 5.3, 4.7………

So, the single sample z-test is a test that is used when you want to make some inferences about the population mean. Note that again there is the clarity here is that, you are not saying using the sample, but you are not interested in just reporting the sample mean, which is what you might have done with descriptive statistics; but, you are interested in ultimately making a statement about the population means. And, this is a test, where you need to know the variance of – you can think of it as variance or standard deviation; but, you need to know this variance of the population; and, that is a requirement. So, it is not the same thing as computing the variance from the sample. So, that is called the sample variance. And, there is a formula for that. There is actual data. But, this is more useful when there is pin historic data and you actually know the population variance. But, ultimately, you are doing a test about the population.

So, another case where this test finds an application is when the sample size is fairly large. So, this is the one exception to this rule that you need to know the variance. In some cases, when the sample size is fairly large and larger gets defined by approximately 30. So, that is the magic number that people use here. And, when the sample size is 30 or larger, you consider, you reason that you have a large enough sample; and, in that one instance, you do not need to know the variance; you can actually calculate the sample variance from the data itself and use that for this test. So, just to quickly summarize with the single sample z-test, you have a single sample and then you are testing… You are making some inferences about the population mean based up of the sample. And typically, in a single sample z-test, you need to know the variance with the small exception that you can also use the single sample z-test if you have a large enough sample size and you do not know the variance.

So, the example we are going to use to motivate this test is one that you have already seen in the previous class. So, we spoke about this problem on average phosphate levels in blood; and, I want to be very clear; we created this kind of a medical scenario; I am not… I am not saying this is medically accurate; we are interested in the statistics of it. So, we imagine this doctor or this public health system, which says that your average phosphate – the average phosphate level in your blood should be less than 4 point 8 milligrams per deciliter. And, the idea here is that, doctors or whoever understands that not every time. So, you take a blood reading; you take a blood and you take a reading of the phosphate level. Not every time is it going to be 4 point 8 or less; sometimes it might be more, sometimes it might be less. But, the whole idea is you are trying to see if the average is less.

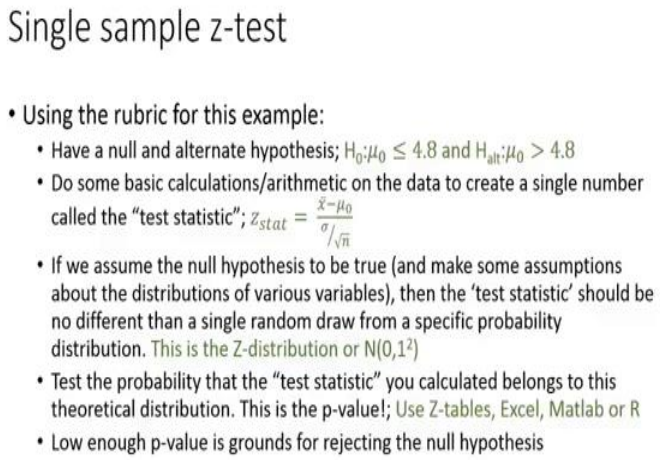
And, when you say I am interested in the average, you are not interested in the average of some sample that you have taken; and, here a sample would be if you took 5 blood tests; perhaps different machines give different results; perhaps different times of the day give different results; perhaps what you ate in the morning affects it; but, the whole point is you just have that sample at hand; but, you are interested in saying something about the population. So, let us just step back and go through each of the bullet points one more time in the context of this example. So, we said what are we testing? I am interested in the population mean. So, you might ask yourself in this particular example what the population is. And, you can think of the population in a couple of different ways; you can think of it as a concept of what your true average and true distribution is. So, let us say that, yes, you can take a blood test. And, when you take a blood test, it is like your taking a sample. But, there is the concept of what – of what the phosphate is in your blood at all times.

So, doing a test just gives you one peak into the reality; but, there exists this concept of reality, which is that, there is some distribution. And, we are assuming that distribution does not change over time; and, that distribution has a certain mean and we are very interested in this mean. And, this is the distribution of the phosphate levels in your blood. This is that reality, which we do not know and you can think of it as this oracle somewhere that knows what your true phosphate levels in blood is at all times and at some distribution. And, at any point of time, you go take a blood test, it is like you are getting a random, you are getting a number from this random variable, which is in the form of distribution. And, you are very interested in the mean of the distribution. So, that is the population; population is your true phosphate levels in your blood at any point of time. And so, it is a concept.

And, if you kind of like to think of population as actual data points; another way you can think of this is to say the population is what is this data set – this is very large data set that you would see if you were to continuously take blood tests – infinite number of times with infinite number of machines over multiple days or whatever. So, you can think of constructing this population in your head as a very large data set. But, the truth is ultimately, you never have the data set or you do not… otherwise, we would not be doing inferential statistics. But, what you do have is a sample. And, the sample can be of some size; we have not discussed that. It can be 5, 10, 20, 30, 40 data points. And, the idea here is that you know the variance. So, I have said that it is a known standard deviation of 0.4 milligrams per deciliter. So, just mind you that, this is the standard deviation, not of the data set, of the sample set you got; but, it is the standard deviation of the population. So, it is like this.

In this population distribution, we are trying to answer some questions about its mean. But, someone somehow you know what the standard deviation of this distribution is someone whispered it to you or you know it from fundamental principles or you might reason that historically it has been equal to this value and should not have changed. But, you know the standard deviation in this particular example. And, we have already talked about how – if you have a large enough sample size, you do not actually need to know the standard deviation, you can compute it. So, here is the data. So, I have told you that, we know the standard deviation; but, you also… This out here is the data and I have explicitly not given you the exact data points. So, just to give you an idea, each of these data points comes from going and doing a blood test. So, you did a blood test and you got 4.1; you did it again, you got 3.9 and so on. And, there is this list. And, this is what we call as a sample. So, what do you do with this data? How do you conduct the test?

(Refer Slide Time: 10:14)



Single sample z-test

- Using the rubric for this example:
  - Have a null and alternate hypothesis; H_0:$\mu_0 \leq 4.8$ and H_{alt}:$\mu_0 > 4.8$
  - Do some basic calculations/arithmetic on the data to create a single number called the "test statistic"; $z_{stat} = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$
  - If we assume the null hypothesis to be true (and make some assumptions about the distributions of various variables), then the 'test statistic' should be no different than a single random draw from a specific probability distribution. This is the Z-distribution or $N(0,1^2)$
  - Test the probability that the "test statistic" you calculated belongs to this theoretical distribution. This is the p-value!; Use Z-tables, Excel, Matlab or R
  - Low enough p-value is grounds for rejecting the null hypothesis

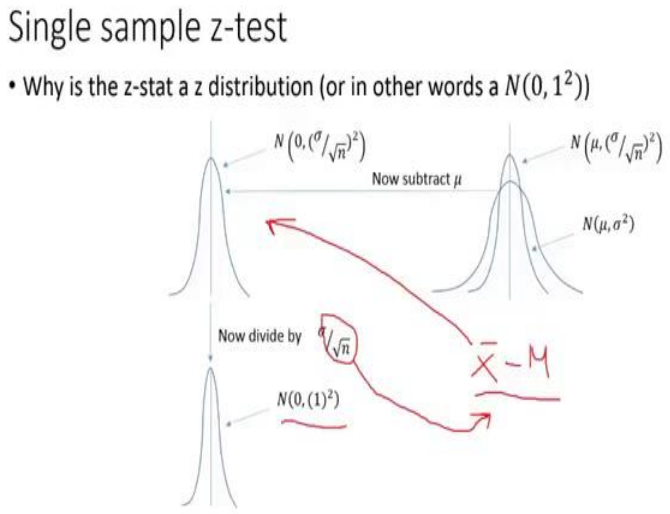
Let us go back to this rubric that we created that is just it is like this template that we discussed in the end of last class and to conduct any hypothesis test. So, the first bullet point has a null and alternate hypothesis. And, that is what we are going to do. The null hypothesis here is to say that, μ_0, which means – which refers to the population mean. And, you can use the… I have used mu naught here, you can use mu as well; that also you might see text books do that. But, the idea is that, the null hypothesis says that, the true mean of the population is less than 4 point 8 I do not know the answer to this question; that is what I am hypothesizing. We are going to do some mechanics; we are going to go through some process. At the end of it, we are going to see if the null hypothesis is true or not colloquially speaking. So, a null hypothesis like we said, the null and alternate together need to be mutually exclusive, collectively exhaustive. μ_0 says that, null is less than 4 point 8; that is the null hypothesis. So, the alternate hypothesis should be that μ_0 is greater than 4 point 8.

The next step we said is do some basic calculations – arithmetic on the data to create a single number called the test statistic. And, the math that we are going to be doing here is fairly straightforward. We are going to take $\bar{x}$, which means a sample mean. The sample mean here would just mean that, you take these data points out here and you take their average. So, you take all the data points that you have collected and you take their average. So, we do that. And, we then calculate $\bar{x}$ - μ_0. And, here μ_0 is 4.8; it is the number that you are hypothesizing. And, you divide that by $\frac{\sigma}{\sqrt{n}}$. So, here σ refers to that standard deviation that you already know. So, that is the 0 point 4 that we spoke about out here. So, we are already given that, the standard deviation is 0 point 4.

So, given this, we compute $\overline{x}$ - μ_0. So, $\overline{x}$ is the average of the sample; μ_0 is the 4 point 8 – the number that you are hypothesizing; σ is the standard deviation that you are given; and, n is the number of data points in that sample. So, if your sample size is 10, 15, you would substitute 10 of 15. So, that is how you calculate something called the z-statistic. So, why you are calculating this value? We said in the next bullet point that, if we assume the null hypothesis to be true; and, make some assumptions about. Distributions – I would not say that every time; but, if we assume the null hypothesis to be true, then technically, the test statistics should be no different than pulling a random number from a specific probability distribution. So, if that… The whole idea is if the test statistic – if the null hypothesis is true, then this test statistics z-stat should be equivalent calculating z-stat. So, you can plug in numbers for $\overline{x}$, μ, μ_0, σ and n. And, you will calculate a z-stat. Once you calculate the z-stat, if the null hypothesis is true, then the z-stat that you are getting should be equivalent, should be the same thing as pulling a random number out of a specific probability distribution.

What is this specific probability distribution? In the case of the single sample z-test, this distribution is called the z-distribution. And, the z-distribution is nothing but a normal distribution with mean of 0. So, I have used this nomenclature; we have discussed how this is standard nomenclature; but, the n here means it is a normal distribution with a mean of 0, which is the first number that you see. So, you have a normal distribution with a mean of 0 and a standard deviation of 1. And, 1 square is a convenient way to represent it because you know very clearly that, you are talking about… 1 refers to the standard deviation and 1^2 is also equal to 1. So, the standard deviation or the variance is equal to 1. And so, this is what is known as z-distribution. So, we are saying that, if the null hypothesis is true, the z-statistic that you compute with this data should be the same or should be equivalent to pulling a random number from a z-distribution. This is a very useful thing to say. And, what we are going to do is we are going to come back to what we do next, because of the statement. But, before we do that, I want to make sure that, you understand why. If the null hypothesis is true, that the z-statistic – computing a z-statistic from the data is equivalent to pulling a random number from a distribution that is normally distributed with mean 0 and standard deviation – 1.

(Refer Slide Time: 15:58)



Single sample z-test

- Why is the z-stat a z distribution (or in other words a $N(0,1^2)$)

$N\left(0,\left(\sigma/\sqrt{n}\right)^2\right)$

$N\left(\mu,\left(\sigma/\sqrt{n}\right)^2\right)$

Now subtract μ

$N(\mu,\sigma^2)$

Now divide by $\sigma/\sqrt{n}$

$N(0,(1)^2)$

$\bar{X}-\mu$

So, let us go to the next slide to kind of do that. So, at the start, you had this distribution. So, this distribution is the population distribution. So, this is the population. The population – if the null hypothesis is true, would have a mean equaling μ ; correct? I am using μ and μ_0 little interchangeably out here. But, I do not want you to get confused by that; but, the idea is whatever your hypothesis is, if your null hypothesis is true, then the mean of this distribution is equal to 4.8. Technically, it is less than or equal to 4.8; but, we are going to take the extreme case. So, we are going to come to one end of it and say it is equal to 4.8. So, this is the extreme situation, where the mean is actually equal to 4.8. And, we are already given the standard deviation. So, you have already been told that, sigma is 0.4 and you are given that value.

So, first, let… This is just the distribution of the population. So, you have built the distribution of the population. Now, when you go to take a sample from this population of some size n, what do you get? What you get is as we have discussed, if I take sample of 5 data points, 6 data points; and then, compute a mean from that sample, we know that, the arithmetic mean that you compute from a sample need not always be equal to the μ exactly. Sometimes it is μ , sometimes it is little higher than μ ; sometimes it is a little lower than μ . But, we have already discussed as to how. That is also a distribution and that distribution called the sampling distribution is also normally distributed with the same mean μ , but with a standard deviation of $\frac{\sigma}{\sqrt{n}}$; where, n is the number of samples you took to compute that mean. So, it you took 5 data…

So, if you technically just take one sample and compute the mean from it, you will get the same distribution again – get the same original distribution. So, if you substitute n is equal to 1, nothing changes. And, that should be intuitive. If your sample size is just one data point; when you are computing the average of one data point, it is literally like you are recreating the distribution. If that n goes to infinity, then your variance goes to 0. So, as your sample size keeps on increasing, you are literally going to be sitting on top of this line and you really would not have this distribution. But, for all finite sizes of samples, what you have is a distribution for samples, which is normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Now, what we are going to do? Now, mind you, this is the distribution of sample means. So, in some sense, this is the distribution of $\bar{x}$. Now, what we are going to do is we are going to subtract μ from this distribution. From this distribution, we are going to subtract μ . What effect does that have? From a distribution – from a normal distribution, essentially, if you just subtract a number, it is literally like just shifting the distribution and centering it. So… because you are just subtracting a number, you are not affecting the standard deviation. So, this gets unaffected. But, when you just subtract a number from the distribution, you can think of subtracting from a distribution as you take each data point and then subtract the same number from it. Or, you can think of it as computing that $\bar{x}$ and then subtracting it. But, in either case, it has the effect of just shifting the distribution; it has the effect of centering the distribution at another location. And, that is what happens. When you subtract μ , when you take μ out of the $\bar{x}$; so, it is… This is what I have done; I have taken mu out of x bar. It has an effect of moving this distribution to another location, which is now centered around the mean equal to 0.

Now, what happens if you then take this distribution and divide it by the standard deviation? So, what happens if we divide it by this number – $\frac{\sigma}{\sqrt{n}}$? The effect that has is in re-scaling the standard deviation such that you now have a distribution, which is normally distributed. When you divide, something that is already centered around 0. So, the distribution is already centered around 0; which means that it has some positive values, some negative values. And now, you go divide by a number. The effect of the division is that, because it is already centered around 0; it is not actually going to change the central location of the distribution; it is instead only going to widen it or narrow it depending on what you divide it by. So, I have kind of shown the distribution getting narrower; but, that need not be the case; the distribution could have just got wider. It just depends on whether $\frac{\sigma}{\sqrt{n}}$was greater than or less than 1. So, if it is greater than 1, then you would by dividing by $\frac{\sigma}{\sqrt{n}}$, you will be making the distribution narrower.

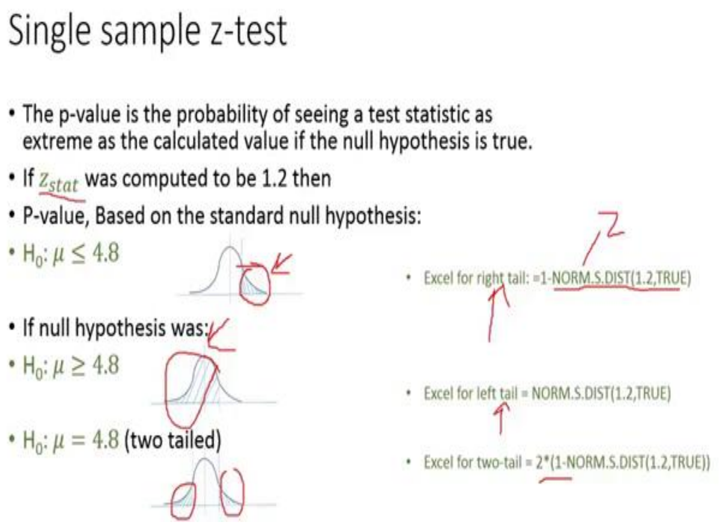
But, if $\frac{\sigma}{\sqrt{n}}$ was smaller than 1, then it would have the effect of widening the distribution. But, ultimately, when you go divide the normal distribution, which is already centered around 0, all you are doing is you are either stretching the distribution or kind of crunching it. You can think it as scaling it. And, now, you have a standard deviation of 1. So, that is how you get the whole idea of x. So, on the first step, we took $\bar{x}$ and then we subtracted the μ from it. That is where we got this value. And, now, after dividing by this number sigma by square root of n, you went and divided this by this number to get your normal distribution with mean 0 and standard deviation 1; which is what we said was the z-distribution; got it; perfect.

So, now, what we are going to do is go back to this rubric. So, great; so, we have reasoned that you have to calculate the z-stat. And, we have already said that, the z-stat if the null hypothesis is true, should be like pulling a random number from a normal distribution. What we are going to do now is we are going to test the probability that, this statistics that you got; we are going to say – if your null hypothesis was true, I should have gotten a random number from here. But, let us actually look at the actual number that I got. Does my z-stat actually look like it could be something that I could have pulled out of a normal 0 comma 1 square, because if it does not, then something that I assumed was wrong. I said that, this z-stat should look like something that I pulled out of a $N(0,1^2)$ if the null hypothesis is true.

So, let me go to take an actual look at the z-stat number. And, if it looks like it could not have come from this distribution; if it looks like it was unlikely to have come from this distribution, then I can reason that, maybe the null hypothesis was not true and I can reject the null hypothesis. So, that is what we are going to do the next step. The next step is to take the z-stat and plug it in to this $N(0,1^2)$ and see how extreme is this actual number given that we know the $N(0,1^2)$; and so, we know the potential values it can take. For instance, I have already told you it is a normal distribution – mean 0 comma 1 standard deviation. So, from this distribution, if I were to pull a random number, how likely is it that I would see a number like 55. That is too high. The standard deviation is

1, the mean is 0; it is almost impossible that you would pull a number like 55 or -20; so, for that matter. So, if it turns out that the z-stat is too extreme a value to be coming from this normal distribution, then we can maybe make a statement about the null hypothesis. So, that is what we are going to do as a next step. Let me just clean this up.

(Refer Slide Time: 25:33)



Single sample z-test

- The p-value is the probability of seeing a test statistic as extreme as the calculated value if the null hypothesis is true.
- If z_{stat} was computed to be 1.2 then
- P-value, Based on the standard null hypothesis:
- H_0: $\mu \leq 4.8$
- If null hypothesis was:
- H_0: $\mu \geq 4.8$
- H_0: $\mu = 4.8$ (two tailed)

- Excel for right tail: =1-NORM.S.DIST(1.2,TRUE)
- Excel for left tail = NORM.S.DIST(1.2,TRUE)
- Excel for two-tail = 2*(1-NORM.S.DIST(1.2,TRUE))

So, I have just restated the official definition. What we are going to be calculating is something called a p value. And, this is the probability of seeing a test statistic as extreme as the calculated value if the null hypothesis is true. With the core idea being that, if it looks too extreme that, if the p value of the probability of seeing this test statistic is really low; then, perhaps the null hypothesis was not true to start. So, for instance out here, if the z-statistic you computed was 1.2; so, I just… I am just giving you a number to go with. Then, the core idea is that, you would calculate a p value based on the standard null hypothesis, which is that, μ_0<1 4.8. And, you will say… Let me take this distribution, which is the z-distribution or the $N(0,1^2)$. So, this is the $N(0,1^2)$. And, I am going to go place 1.2 here. And, I am going to compute a probability; and, this – the probability is a probability to the right side of 1.2. It is the area under the curve out here. And the idea is because you are then quantifying the probability of seeing something as extreme as 1.2 or greater is equal to the area under this curve.

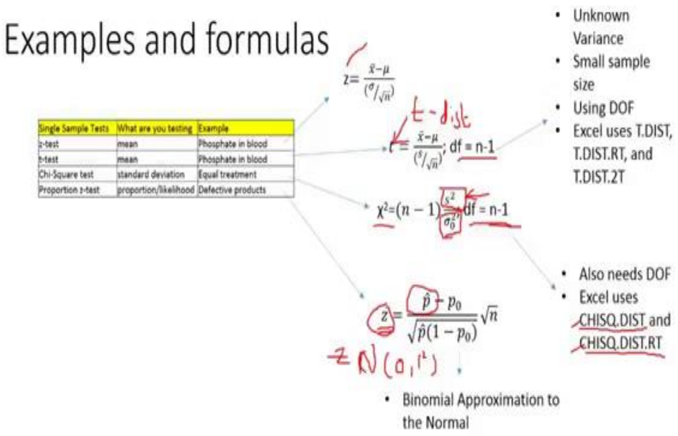
And, that might happen to be any value. So, I mean I think in this case, it happens to be something like 13 percent or whatever. But, if this number was really low; if this number was 0.001, you might say look – this number is so low that I am going to reject the null hypothesis. And, if you cannot find something that extreme, the standard thing that you do is you fail to reject the null hypothesis; you technically never accept the null hypothesis. So, that is the core idea. And, you can also think of this in a couple of other ways. For instance, if your null hypothesis were that mu is greater than 4.8; then, you would be looking at the area to the left of your curve. But, typically, in a situation like that, you actually would not do the statistical test. So, it is not common to see p values greater than 0.5 because at that point, you start by saying look I have computed a z-statistics that is already positive, that is, 1.2. And so, I know even before I go put this line out here that, I am going to get a probability greater than 0.5.

So, for instance, your z-stat was computed to be exactly equal to 0. You know that, if your null hypothesis was mu is greater than 4.8; that, it will be 0.5. So, any z-stat even greater than that is bound to be greater than 0.5 when you do not need statistics for that. But, another interesting situation, which a lot of people work with, is called the two-tailed case. These two are called 1 tailed – 1 tail. So, you also have the 2 tailed case, where your null hypothesis is really that, mu is equal to 4.8. And, you are interested in rejecting this null hypothesis whether that mu is too large; meaning it is large enough that you can say that it cannot be equal to 4.8; or, if it is small enough. So, you are happy to reject if you see evidence that shows that mu cannot be 4.8 on either account; maybe because it is the data suggests that it is too large, maybe because the data suggests that it is too small. And, that is called the 2-tailed case.

Now, there is lots of different software, where many of the steps that we have done we are actually taking care of and it does not take much. Even a simple excel sheet if you just go down the data and say do a z-test; it will do it. But, somewhere understanding the mechanics of this and getting it to the stage of the z-statistic, at least brings about some sense of control and transparency in your understanding. But, once you get off the stage, computing this area – whether it is on the left-hand side, right-hand side or either side can be done fairly; it is not something that can easily be done by hand. So, what text books do is – if you have taken, most statistics text book will have these pages towards the end. And, they are given in the form of tables. So, you can take z-statistic number that you computed and go plug that in to this table. And, it will tell you the probabilities.

And, usually there will be a diagram on top to hint whether they are giving you the probabilities to the left hand side or the right hand side or both sides. And, you know – as long as you know what they are giving you, it is fairly easy to figure out whatever it is that you need. If you want the right-hand side; but, they are giving you only the left-hand side; then, you can just subtract the number they are giving you from 1, because you know the total area under the curve is equal to 1. But, what I am going do is I am just going to give you some simple Excel functions that do this for you. For instance, in Excel, you can just… The convention is to give you the area to the left-hand side. So, instance, what I do here is I subtract that from 1. And, the norm s dist is what refers to the z-distribution. And, the true refers to the fact that I am not interested in just height, I need the area under the curve. So, that is what that… So, for the right-hand side tail, you can use this; for the left-hand side tail, you can use this. A simple multiplication by 2 with the left-hand side case gives you the 2-tail situation. So, with this, we have discussed greater detail the single sample z-test.

(Refer Slide Time: 32:11)



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

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

t-dist

$t = \frac{\bar{x}-\mu}{(s/\sqrt{n})}$; df = n-1

- Unknown Variance
- Small sample size
- Using DOF
- Excel uses T.DIST, T.DIST.RT, and T.DIST.2T

$\chi^2 = (n-1)\frac{s^2}{\sigma_0^2}$; df = n-1

- Also needs DOF
- Excel uses CHISQ.DIST and CHISQ.DIST.RT

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

$z \sim N(0,1^2)$

- Binomial Approximation to the Normal

So, now, let us look at a couple of other single sample test. What I provide you with here is the list of them and the formulas. And, I will give you some idea of the context in which they are used. But, we would not derive it or go through it in the same detail as the z-test. So, we have already discussed the z-test. Let us now look at the next test, which is the t-test. So, we have finished with the z-test. So, now, we are going to look at the t-test. So, with the t-test, it is a very useful test and it also tries to test this… It essentially tries to do the same job the z-test is doing; which is to make some statement about the population mean; so, same problem statement in some sense. The one big difference is you are not given the variance.

And, in most situations in life, in statistics, you would not know the variance; I mean just think about how fairly unrealistic it is that you already know the variance of the population in a situation, where you are trying to make a statement about the mean. I mean the only reason you are doing this test is because you do not know what the population mean is. So, you are trying to… You are making a hypothesis, you are taking a sample, and then you are testing that sample, you are working with that sample to make a statement about the population mean. So, there is… I mean think of it as there is some uncertainty about the population mean in the first place and that is why you are doing this test. To assume that in such a situation, you already know the population variance is not very – need not a very realistic. So, this test works the same way.

So, if you look at it, it has got the same $\overline{x}$; it has got the same μ; it has got the same $\sqrt{n}$. But, this s is different from this σ. And, the difference is here sigma was given in this z-test. So, in the z-test, sigma is given. But, in this test, the s is computed; it is computed from the data. So, you actually go back to the sample data and you calculate the variance or standard deviation from the data using the same formula for dispersion that we would have discussed when we spoke about standard deviation and descriptive statistics using the n - 1 idea. And, if you do not remember, you can go back and see that lecture. But, the idea is that, you compute the standard deviation and you plug that value in to get the t-distribution.

Now, a couple of things that are worth noting is that, we spoke about how if you know the variance, you can use the z-distribution; if you do not know the variance, you can use t-distribution. But, there is this exception. We said if your sample size is large enough; then, you can technically use the z-distribution and just compute the variance and consider the variance to be the truth; consider the variance to be the sigma and go ahead. I personally find that a little confusing; I think that is fine; if… That is what is there in tax; that is what people do and that is the reasonable approximation. But, the point is you cannot go wrong with using the t test when you do not have the variance. So, even if you have a large enough sample size, the idea is that the t-distribution becomes… It approximates z-distribution quite well when your sample size is greater than 30. There for all practical purposes, the t-distribution looks exactly like the z-distribution. But, keep the things really simple; you can just follow this simple rule that, you do not know. If you know the variance, just use the z; if you do not know the variance, just use the t. And, that should keep you clear.

The other thing to mention out here is that, this DOF or degrees of freedom – we have mentioned that, out here we briefly spoke about that concept when we were talking again about the standard deviation. Without going into too much detail into degrees of freedom again, the simple thing to keep in mind is that, the t-distribution is not one distribution. I mean it is one distribution, but in the sense that, the t-distribution – you can think of the degrees of freedom as a parameter that goes with the t-distribution. So, just like the normal distribution, if you say the normal distribution, you need to mention the mean and the variance for you to have a… to actually draw the exact distribution or to do some computation on it. It is no point coming to someone and saying how likely is it to see a 1.2 in a normal distribution? That question does not make sense. Normal distribution with what mean and what variance? And then, I can answer your question.

You can think of degrees of freedom in a similar light; which is that, the t-distribution itself is not completely defined until I mention to you what the degrees of freedom are. So, t-distribution with three degrees of… – with degrees of freedom equal to 3 looks different from a t-distribution with degrees of freedom 4. And, the core idea that you need to know is that, the t-distribution has a mean of 0. And, it looks very similar to the normal distribution of mean 0 and standard deviation 1. But, the exception that as the degrees of freedom keep increasing; so, when you go to degrees of freedom of 30 and greater; and, at some point, it is exactly the normal distribution. So, the t-distribution with a large enough degrees of freedom is exactly like the normal distribution with mean 0 and standard deviation 1. But, as the degrees of freedom keep decreasing and come all the way down to, the lowest degrees of freedom you can have is 1. When it comes all the way down to degrees freedom equal to 1, you will find that it still looks a little bit like the normal distribution with mean 0 and standard deviation 1; but, it is a little shorter – shorter in the center and has fatter tails on the sides. And so, that is how it deviates from the normal distribution.

But, all that you need to know is that, the degrees of freedom get defined by the concept n - 1. So, number of data points minus 1 tells you the degrees of freedom. And, once you know the degrees of freedom, you know which t-distribution to look up in the tables. So, you know how to draw the curve and then calculate probabilities from it. Again Excel uses – Excel has some slightly nicer functions for it. So, if you are just interested in looking at the left-hand side of the distribution, you just use T-DIST – T dot DIST. If you are interested in T dot… On the right-hand side, you do T dot DIST dot RT; or, on both sides, you do T dot DIST dot 2T. So, you do not need to actually do the 1 minus and so on that we were talking with the z-distribution. Excel already has some inbuilt functions to just point to which side of the distribution you are interested.

So, we go now to the next. We are finished with the t–distribution; we go now next to the next test, which is the chi square test. And, the chi square test has a couple of different types of tests. But, the one that we are interested now is the chi square test for variance. I am using the words variance, standard deviation a little interchangeably; one is just the square of the other. The test is ultimately one for variance. And, if you are testing variance, you are essentially testing standard deviation. So, if it is easy for you to think standard deviation, you can keep that in mind. And, an example for instance of the chi square test is you are really interested in looking at a sample, but you are not interested in making a statement about the population mean. You are instead interested in making a statement about the population variance. So, you are interested in saying is the population… Just like in this z-test and the t-test, you are interested in saying something like – is the population mean equal to 4.8? Or, is the population mean less than 4.8?

Similarly, here you would be interested in saying things like – is the population variance equal to 0.5, 0.3 – whatever number you have in mind. The important thing is you have a number in mind and you are trying to see if the sample that you are taking… With the sample that you are taking, can you say something about the population variance being equal to this magical number that you have in your head. And, the mechanics of the test is fairly straightforward. And, it is here the σ_0 is the hypothesized variance. So, this is the number that you want to compare it to. This is the equivalent of the 4.8 that was there for means. The s^2 is the sample, is the variance that you compute from the data, from the sample. You take that data set of the samples and you compute standard deviation, you compute a variance from that. And, that is s^2. And, the way you do that again to remind you is this that, $\frac{1}{n-1}$ in the formula for the calculation of standard deviation; you would be using that. And, that is how you calculate the test statistics, which then gets compared to a chi square distribution with n - 1 degrees of freedom; just like in the first case, it got compared to z-distribution and this got compared to… The t-statistics got compared to t-distribution. This is the same way the chi square distribution gets compared to a chi square distribution; great.

So, a couple of things to note is that, if again chi square also uses the concept of degrees of freedom; so, think of the degree of freedom as something that defines the exact distribution you are interested in, because a chi square distribution with 3 degrees of freedom is a different distribution than a chi square. It is a different density function. It looks different. It has different mathematical properties than a chi square distribution with 4 degrees of freedom, 5 degrees of freedom. So, the degrees of freedom help you define the exact distribution and its parameters and its mean variance and so on. But, essentially, that is what you would use. You would need to use the degrees of the freedom and that is also the same as before; it is n minus 1. So, number of data points minus 1. And, chi square… With Excel out here just uses chi square dist; this is the left side and chi square dist dot rt is the right side. I do not see them having something for 2-tailed, but I might be wrong. But, as long as you have these two, you can quite easily just draw that graph in your head and figure out which side; if you are interested in a 2-tail distribution, how you would compute that; great.

So, we finally, come to our last single sample test, which is called the proportion z-test. And, the idea here is you are testing something that is a proportion. So, you are testing a hypothesis like less than 30 percent of the shoppers, who come to my online store, are women. So, you can say again; we can go the 2-tailed way or you can go the 1-tailed way; you can say less than 30 percent; you can say 30 percent of my shoppers in my online store are women or you can say greater than 30 percent of the shoppers are women. The key out here is that, whatever sample you collect to actually test this hypothesis, the hypothesis… So, let us fix on the hypothesis. Let us say the hypothesis is less than or equal to 30 percent of the shoppers, who come to my online store are women.

You have this hypothesis; you will now go and collect a sample. The sample in this particular example could be something like you actually give a survey at the end of the purchase or something and people actually say them – male or female. So, you collect a sample. How does this sample look? The answer is that the sample unlike in the previous examples, where you would have seen an actual number. So, in the previous example, in the phosphate examples, you saw numbers like 4.1, 3.5 – these were actually readings from blood tests. Here you are going to get something that is binary. The person is either going to say that either they are female or not. So, it is a series of 1s and 0s – very similar to the idea behind Bernoulli trials. And, what you are doing is you are now looking at the sample data of 1s and 0s, which… and then answering the question of whether… and then saying something about the hypothesis, which is less than 30 percent of the people, who come to my shop are women. And, this has the same intuition as all the other forms of inferential statistics, which is if for instance, you take 100 samples and you know all 100 of them point to the shoppers being women; then, you are likely to reject the idea that only less than 30 percent of the shoppers have come to my online store are women.

But, the idea is if you notice something, this looks in every way shape and form like the Bernoulli distribution. And, the fact that you are counting how many. So, the Bernoulli distribution had to do with probability of heads or tails. So, that would be probability of it being male or female. But, if you are interested in out of size of n people, who arrive; how many of them? Sort of hundred people who arrive are… Is it less than 30 who are women? That ties us to the binomial distribution. And, yet you do not see the binomial distribution being used in the test, you instead see the same idea of z. So, you are again calculating a z-statistic with this test and you are comparing to the z-distribution, which is normal – which is $N(0,1^2)$.. And, the idea here is fairly simple.

If you remember, we spoke about the binomial approx… – we are approximating this binomial distribution to a normal distribution. That is a right way to put it. And, that is what you are doing out here. And, it should also be intuitive in that the $\hat{p}$ out here is a calculated proportion. So, you will take a data set. Let us say you took 30 people or 40 people or 50 people who came to your store and you actually found that, exactly 23 of those 50 were females. So, that proportion is what? The proportion that you get from the sample is what you have is $\hat{p}$. p_0 is the hypothesized proportion. So, p_0 would be the 30 percent. So, this would be sample. And, this is the equivalent of μ_0. So, this is the population – proportion – the number that you are hypothesizing; and of course, is the sample size.

And, if you look at it, this also looks very much like that $\bar{x}$ - μ concept. And, in some way, this formula out here in the denominator should remind you of the formula that you saw for standard deviation in the binomial distribution class. So, you are doing something very similar to the $\frac{\bar{x} - \mu}{\sigma\sqrt{n}}$; it is in construct identical to that. But, you are using the binomial distribution for calculating things like the variance. But, you are also saying – hey, this I believe should be… you can approximate to the normal distribution. So, I am just going to use the z-distribution to calculate my p values or I am going to use the z-tables.

Just in quick conclusion, just going back to the rubric that we created, I will just clean this up. The final idea is that, you use these z-tables; you can use z–tables; by z-tables, I mean like you can use the back of the statistics text book, you can use Excel or Matlab or R. Just important that you know how to do with at least one of these softwares. And, the core idea is that, if you get a low enough p value; all of these help you calculate p value or a probability. And, if you get a low enough p value, you can use that as grounds for rejecting the null hypothesis. And, I am saying I reject the hypothesis that mu naught is less than or equal to 4.8. And also, just keep in mind – on the flip side, you never say I accept the null hypothesis and you can only say that, I fail to reject the null hypothesis. I hope that give you one illustrative example of the single sample test and the idea behind the mechanics of it; and, introduced you to the other tests. And, in the next class, what we are going to do is we are going to talk about 2-sample tests and we are going to go also beyond that. We are going to talk about the idea of having multiple samples.

Thank you.