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

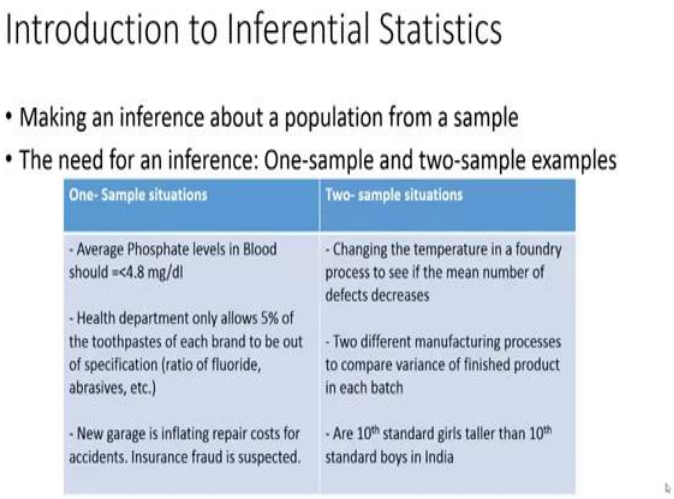
Lecturer - 09

Inferential Statistics - Motivation

Hello and welcome to our lecture Inferential Statistics. In this particular lecture, we are going to be talking more, giving you more about Motivation, for why we need inferential statistics and we will be talking a little bit about, what kinds of problems you can solve using these techniques and you know, where you would applied and when you would applied. In the subsequent lectures, we will talk a little bit about how you would applied and also, why you do some of the math that you do, why you do some of the computations.

So, understanding it a little bit more on the nuts and bolts level is something that will follow. But, in this lecture we are going to focus on why you need inferential statistics in the first place and we are going to do this through the lens or something called hypothesis testing. So, hypothesis testing is very widely used an excepted tool for a lot of data analysis and if you understand inferential statistics through hypothesis testing, many other concepts in inferential statistics just fall in place. So, things like confidence intervals and so on, which you might have heard become fairly easy to understand and process.

(Refer Slide Time: 01:29)



So, having said that let me jump in to the subject. The idea behind inferential statistics is to make some inference about the population from the sample. Just to jog your memories, I think we have spoken about population and sample a couple of times. But, this is different from, what you would have done with descriptive statistics. With descriptive statistics you do not care about population or sample, in the end of the day you have some data set.

And for simplicity sake, assume that was a sample that you got from population and this sample you will very content in descriptive statistics with describing that data set with describing that sample in case you have a sample. But, here the kinds of problems that you are more interested with inferential statistics, a ones where you have a population and you are getting as just a sample from that.

But, from this sample I do not want to just say something about the sample, I do not want to talk about the mean of the sample, I do not want to talk about the variance of the sample and I do not want to talk about the centrality dispersion. My goal is to say something about the population. I only have data, which is the sample. I do not have the data associated with the population, but with this sample can I say something about the population.

So, that is the core idea and just a kind of jogs some of your memories, the idea behind population and samples is fairly simple, we looked at it to the couple of examples. But,

you can think of the population in one of two ways. The first and the more obvious way is that, there exists this really large data set associated with the phenomena. So, let us say the phenomena was the height of all boys, who are in 10th standard in public school. So, that is, so that could be a very large data set and let us say that was for India.

So, it is very large data set, there are lot of children, who are in the 10th standard, who are in public schools across in India. And you can think of that as a population and you want to say something about that population and you might not have the data. So, you go to take a sample, so you select 5 schools or 10 schools or you select 200 students through a census process randomly and take that as a sample.

So, you take a sample of, you know some subset of students from the population, that is one way of thinking of population sample and another way could be that the population itself is moreover theoretical abstract concept. So, you could not have an actual data set, but it could be something like I have created this new machine and this new machine is going to start making certain products and let us say, you are very interested in a product dimensions. So, let us say the diameter of a product is machine makes.

So, you put a raw material in to this machine, the machine splits out some finished product and this finished product should have a diameter, I mean has some diameter. Here you do not, because it is a new machine you might not actually have a population; that is un really large data set that exists somewhere. The population here is the concept that, if this machine going to create infinite such products without any change in time space, the dimensions of these products would be the population and ultimately you might say, hey let us just for the first time run this machine and make 10 products and these 10 products ((Refer Time: 05:02)) have and I measured and so on, is the sample.

So, here is the case, where I still want to use the sample to say something about the machine in general. Not just the 10 products the machine is turned out, but the concept of the population here is not an actual finite large data set in my hands, it is more about concept. Now, having revisited population and sample here let us again see the statement, which is that inferential statistics you make, you want to say something about the population from the sample.

So, as I said the major aim of this lecture is to motivate you to see while inferential statistics is important. So, I felt that the best way to inference to that might be to give you

some examples. So, that is what we are going to do and I start with some simple examples. I have broken them down in to one sample and two sample examples and pretty soon, that will become clear what that distinction is.

So, let us start with the first example which is a one sample example on that upper left of your screen. So, the idea here and I am just, so you know we are in this part and the idea here is that, let us say we were interested in noting the average phosphate levels in our blood and I do not have a medical background or anything, so do not look at the medical aspect of this examples. But, let us say that your doctor or doctors in general or you know public health advocates, say that the average phosphate levels in bloods should be less than 4.8 milli grams per I do not know deciliter.

So, again irrespective the units, so the whole idea is that this number, which you can get if you go measure your blood should be on average less than 4.8. The key here is an understanding that they should be less than 4.8 on average. So; that means, the doctors or the public health advocates understand that sometimes it could be greater than 4.8 and that perhaps in this particular case is not a cause for along. Again do not focus on the medical aspect, I do not know if it is not, but this is the situation I am creating. So, but the important thing the doctors have told you, it is on average it should be less than 4.8.

So, let us say you say and you know this; obviously, variation. So, it really depends on what you ate that day, it depends on what time of the day you take the measurement, it depends on what instrument you used to take the measurement, it depends on how much water you had. So, let us say there are lot of factors that you do not seek to control and that is the whole idea behind this. But, you want to take a set of measurements and you want to take a set of measurements and answer the question us to whether the average phosphate levels in your blood.

In general, not just on the sample that you have taken, you do not I mean the sample could be anything, but you care about, in general is my average phosphate level and blood less than 4.8. So, why you… So, the question might arise you know, why you distinguishing between what the sample says and what reality is and that is going to become clear in second. So, let us first go about trying to answer this question. So, the first thing is, if you took a set of measurements and let us say, you got consistently very low values.

So, you let us say you got 2.4, 2.5, 2.1, 2.7, 2.3, 2.9 fill of four more numbers in the two points. Then, I guess you really do not need a statistician, you can look at the sample that you got and you can say, look I am fairly certain that even if I went on taking more and more samples or that if I woke up another day through all these data or took another sample or if I took infinite many samples, in either case it looks like my average is going to be less than 4.8, that is fine, you know intuitively that seems obvious.

Similarly, the flip sides, suppose you wanted to take this data and you consistently got 5.5, 6.1, 7.7, 6.9, so on, where every single data point is significantly greater than that 4.8 mark and approximately in that same region, meaning it is not widely moving around. So, that is also an intuition that you might have, that if one second it is 6.5 and the next second it is 2.1, you know it is widely moving around.

But, here I am giving you examples, where 6.1, 6.7, 7.1, 7.3, 7.4. So, it consistently significantly greater, again you do not need a statistician. Somewhere your intuition, you just say look, I mean based of the sample I am willing to bet that my average phosphate levels are less than 4.8 mg per dl. But, then it gets a little tricky. What happens if you had, you know numbers, you know some of them below some of them above.

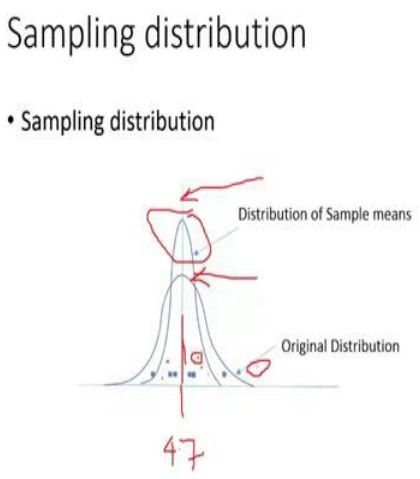
So, some of them are less than 4.8 some of them are greater than 4.8 and in some sense, what you do then and the instinct to the intuition there sometimes just to say well, let me take an average of the sample. If that average is greater than 4.8, then perhaps I should conclude that my average is greater than 4.8 and this is small problem with that kind of an approach. I mean, assume that you got the readings like as such as 5.1, 4.8, 4.9, 4.7, 4.7, 5.3 and you got some set of variance and you took the average and that average was 4.85.

So, you saying, you know what the sample showed that it is greater and you conclude, for instance that the average phosphate level in my blood in general is greater than 4.8, based off of the sample that I just saw. The problem with that could be that may be if you just took two more data points. Let us say you took two more data points, you increased your sample by two more and you got a 4.6 and a 4.7 and all of a sudden, because of these new data points, your average you know slights just below 4.8.

Something about doing this process, we just take the average of the sample and make a conclusion, does not seem correct for this reason. And you know, another way of looking

at it is also that, if we looked at this notion of sampling distributions. So, we know that, let us take a look at the same graph that we looked at the end of the last class.

(Refer Slide Time: 12:14)



So, you had this thing, which we described as the original distribution. And let us say for now, let us just say for now, the truly your… This distribution by the way represents the amount of phosphate in mg per dl that you will see in your blood if you do a test at any given point of time. So, you get numbers from this distributions, so sometimes you get a 5.1; that is what, that is the dot there, sometimes you get a 4.4 that is the other dot there.

But, in the end of the day it looks like on average, given how I have drawn it. On average it is only 4.7 mg per dl, which is less than the 4.8 mark that we were interested. Now, you go and take a sample, this is the same example as a last time, so you took the sample and you some data points. So, you took 6 data points and this is what you got. Now, if I want to just take the sample average and I am just eye balling it here, if I just took the average of these numbers I would say that average falls somewhere here, would it to see.

So, this might be the average of these numbers or maybe it will fall right on this data actually. So, let us say this is somewhere out here is the average and the true is this sample average is greater than the 4.7 and if you want to just go by the sample average, we might have conclude it that the amount of mg per dl is greater than 4.8 or whatever. But, here is a good part, if you now had a class in statistics that told you that the average that you get from the sample is not always going to fall on this 4.7. As your number of

samples tends to infinity? Yes, it will converge to this point.

But, if you got a finite number of samples and that is not, so it is not going to always be exactly on 4.7. So, what is it going to be? What it is going to be is another distribution. So, if you had N samples and this distribution will change with more samples, if there are many, many, many, many samples, then literately if as the number N tends to infinity, this distribution will pretty much like flat on this line.

But, if not, you still getting the sample mean, the mean that you calculate from the samples literally a random variable that you getting from this distribution. So, it is literally like you just pick random points from this distribution. So, you get a point any where here, you probably will not get a point here, because it does not… The probability of getting this point from this distribution is very low, it is almost 0.

So, you will be getting points from this distribution and as a result, just because this number is greater, that 4.7 should not make you conclude that this mean, which is what you are trying to conclude. You are trying to say something about this line; you are trying to say something about this line, which is the mean of the population and, because you get a sample mean which is nothing but, the number from a distribution should not make you conclude that therefore, it is greater than 4.8.

So, that is why you need to do something more, you need to do something more complex than just blindly taking the sample average. So, again we are going to talk later about how you do it, but and what and when you do, but right now I am just trying to motivate for you why you need something else. So, take another example and in this example, we take a problem of proportions.

So, let us say the health department or some dentist related body says; only 5 percent of the toothpastes of any given brand can be out of specification. So, out of specification might mean that, you have some ratings on the amount of fluorides tooth paste can have. So, let us say you are allowed any 1000 parts per million of fluoride and you, there are other chemical limits. But, the health department understands, that not every toothpaste can match exactly the ideal requirements.

So, let us say the set out limits on the chemicals and they say, look if you are a toothpaste manufacturer, only I am going to only allow 5 percent of your toothpastes to, you know

be out of specification, 95 percent of your toothpastes that I see in the market need to fall within my guidelines. Same problem comes up again. You can take a sample of 10, 20 toothpastes and it could very well be that truly this toothpaste brand is involved in a chemical process or manufacturing process, that creates on average only 4 percent. Only 4 percent of the toothpaste that this company makes are actually out of specification.

But, it is perfectly possible that you went and took 10 toothpastes from the market and your luck, 7 of those 10 are out of the specification; that is perfectly possible. It might not be the most probabilistic thing, but it is perfectly possible. It is possible that, this toothpaste company is involved in a manufacturing process and chemical process that creates toothpastes and on average, 4 percent of all the toothpastes they make.

So, when I say toothpaste, think of it as a toothpaste tube; on average 4 percent of all the tubes they make. Have a chemical composition that is not acceptable, which is fine, because the health department says you cannot go more than 5 percent and this toothpaste company has rising it is hand and saying hey you know, which only 4 percent. But, I now go and randomly sample 5 toothpastes, 10 toothpastes and I find that out of the 5 toothpastes that I randomly sampled, 3 of them are defective. All of a sudden, I am saying 3 out of 5 that 60 percent, you say only 5 percent is allowed and I find 60 percent.

And so, is that does not mean the company is not creating toothpastes less than 5 percent rate, which are in conformance less than 5 percent rate, probably no. Again you need a little bit more new on thinking and you need little more statistics to actually answer this question, based off of the sample you cannot just take the sample average. Another example, the third example that we have on the one sample cases, imagine that you are in an insurance company and you find, that there is this particular mechanic shop that is new garage, which does repairs and because most people are required to have insurance.

Once the garage kind of writes out an invoice and people who file the insurance claim attach this mechanics invoice and tell the company to reimburse them for this rectification that is claim to the car. And let us say, this is a new garage and you know the insurance company is suspecting that these guys are cheating, that their set up as a place to not do any real work, but just write really high invoices. So, that the insurance company, so they are involved in some kind of a fraud.

So, one thing that the insurance company can do is saying, I am going to look at that next

10, 20 or whatever repairs. So, let us say up to this point this garage is not made a single you know claim, but it is just being set up, but the inside word is that they are trying to cheat this system. So, once a garage gets set up, this insurance company is ready. The first 10 claims or 20 claims at this garage files or 30 claims, they take those claims and they see how that compares to the national average in terms of average claims.

Again the problem is just, because this sample is greater than the national average, can we conclude yes these guys are cheaters and just, because his sample average is less than the national average, can we conclude these guys are not cheaters and the answer to both those questions is no. In some cases, like when I was talking to you about the case, it might be brutally obvious, where every single data point is so high or so low, that you are like I do not need a statistician to tell me that the answer to this question.

But, when it is not that case you need little more you need inferential statistics to answer that question. So, let us we look at the single sample cases by that we essentially mean that there was one data set and you are essentially comparing that data set to some bench mark number that you had in your head. Lets now, move to two sample situations the example here is let us say that I am running up foundry and some guy comes in consultant comes in says you know if you change the temperature a little bit of the molten metal that you pouring in whatever.

Let us say change the temperature down by degree over two and I assure you that average number of defects that you see in your costs in metal costs will decrease. So, like a mechanical engineering application you like may be the consultant knows what is talking, but how do I test it, how do I test it and the answer to that question is it is you can do the following, which is before you do that before you go change things you can measure the average number of defects in your costs and you do that for 10, 20 data points.

And then, you do with the consultant, which is change the temperature and then, you collect the another 10, 20 data points. Now, let us go back to the question if the average of the sample the first sample, so now, we have sample a sample b sample a corresponds to before the temperature was change sample b corresponds after the temperature was change. Now; obviously, I mean in all likely hood there is going to be some average to sample a and there will be some average to sample b.

In all likely hood these two are not going to be the same one is going to be higher or lower than the other just like in the single sample case if they. So, dramatically different these samples and when I say dramatically different I mean dramatically different with respect to some amount of variability there is in the two samples as well. Then, you do not need a statistician to tell you it is like, so obvious that changing the temperature dramatically reduce the number of defects.

But, in many cases you do not know and in those cases it is not obvious to say the average of sample A was different from the average of sample B, I mean go back to I am going to erase this, but erase the red ink, but go back to this example. Let us say that, lets say the this is the original distribution of the number of defects, so this is the original distribution and let us say this is 3 defects and this is like 9 defects, now that is too high let us say this is 5 defects and this is 1 defect. So, this original distribution is what I care about that goes from here to hear these numbers of defects you will see in a costs.

Now; obviously, if I take a sample of 6 or, so I get a random point I let us say this is the random point I get and the erase the other one done erased. So, I basically I took the original distribution I took the sample of 6 costs and when I take the sample of 6 costs like we discuss we the average is not going to always we exactly 3 is going to be some number that falls in to this distribution the distribution or sample means and I drawn a random number that I got out here from that distribution.

So, good now let us say this consultant who is telling you to go increase or decrease the temperature was completely around. Let us say he had no clue, what he will say he was just lying, but fortunately, unfortunately whatever is said does not make a different I mean he lied in that you know he said is going to improve the process it did not improve the process luckily it dint make the process was.

Now, you go take and because a processes not change the original distribution is not change after that temperature is change the number of defects you going to be receiver also exactly in conformance of this distribution it is in conformance to this outside distribution the original distribution. So, that has not changed essentially this mean has not changed this mean has not changed. Now, you go take a sample of sets you get another sample average and this sample average again is going to belong to this distribution and let us say this sample average was this value.

So, this is the new sample average, so this is new it is that is new now you cannot say hey this new sample average is higher than the old sample average. Therefore, the population mean is different it is not I mean I just gave you an example, where the population mean truly did not changed, but how you could have seen two different sample averages in concluded that one is greater than the other.

So, this is why again you need more than looking at the sample average of a and sample average b and saying hey one is higher than the other. So, we should believe that what the consultancy said was correct or the other way around you know if the if you conclude that what the consultancy said is definitely made things was that is also marked correct perfectly possible that there was truly a change, but because of luck you know you saw things other way.

Now, another example you can think of and this the next example is one where I wanted to emphasis variation the rather than just the mean, let us say you have two different manufacturing processes and you want compare their variance of the finished product in each batch. So, you have manufacturing process A manufacturing process B and they make batches of you know finished material finished product and within each batch this some amount of variance of each products that some amount of variance right variances the inherent variation between one part to the next.

And, let us say I care about that the variance let us say I do not in a particular batch I do not want one product to look very different from the next product I do not care about the mean. But, I want them to all be consistent again you would use the same concept you would take the first batch, which is made from machine A calculate the variance of it calculate the variance of that sample take the second batch made from machine B calculate the variance of that sample. And again the overall concept exists you cannot just say sample of sample variance of machine A is lower than sample of machine B.

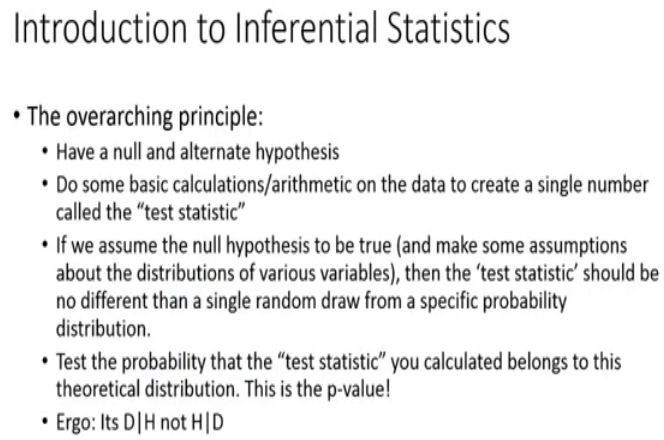
Therefore, I conclude that machine A better than machine B it is perfectly possible than machine A actually verse in machine B. But, because you ultimately only have a sample that machine A got lucky ultimately it goes back to this idea that you see in this distribution. But, sometimes you get number on this side of the distribution sometimes you get a number on this side of the distribution. The more data point should take the overall variance of this distribution reduces, which is why if you had infinite number of

points right none of these problem is exists, but you do not.

And, if you dealing with that then there I think to do this to use inferential statistics to look closely at the data. Another example that is often coated is things like are tenth standard girls taller than tenth standard boys in India for instance we all know that in terms of average heights men have larger average heights than women, but we also heard that girls start growing taller earlier. So, I do not know is 10th standard breakeven point at least some statistics text books sink to think, so.

So, you could have a simple question like the population here is 10th standard girls in India tenth standard boys in India you are ultimately taking the sample and based of the sample, what can you say about the population. And you know sometimes the story is obvious from the sample itself sometimes you need inferential statistics to come in and tell you whether you can say something concrete or perhaps you cannot say anything concrete and that is also something inferential statistics will tell you. But, ultimately it is not as simple as just saying the average of sample A is greater than the average of sample B, therefore I am going to conclude one way or the other.

(Refer Slide Time: 30:55)



So, let us go to what it is that we will be trying to do with inferential statistics I am just going to go through the overeating principle and keep this in mind will review this slide a couple of times. But, I think the ultimate test in some sense of you understanding, what it is in, what it is and how it is and why it is will really become clear once we do the

actual math with each test. And ultimately these will come in the form of test I mean some of you might have heard of these test like t test, z test, chi square test, f test, ANOVA and so on and we will go through each of these tests.

But, here is the overarching principle with respect to hypothesis testing is to have this have a null hypothesis and an alternate hypothesis. So, for instance in the fluoride case the null hypothesis could very well be let me erase this, the null hypothesis could very well be that you have less than 4.8 mg per dl. So, the null hypothesis is that the actual phosphate levels in the average phosphate levels in blood for person x is less than or equal to 4.8.

And, so the alternate hypothesis would be that its greater than 4.8 the important thing is the null hypothesis and the alternate hypothesis in some sense together should be mutually exclusive, which means that if it is greater than 4.8, then it cannot be less than or equal to 4.8 and vise versa and you know collectively exhaustive there should essentially cover the entire space that you are interested in talking about.

So, I mean meaning that the average phosphate level is either less than or equal to 4.8 or greater than 4.8 it cannot be neither. So, its collectively exhaustive, so then you what we will be doing and you not been talk this yet, but you will be doing some basic calculations or arithmetic on the data to create a single number call the test statistic. So, you do not know what that is yet, but what you will do is you do the reason I am explaining this to you is to give you an idea that you are going to be working with the sample.

So, it is not like a magic in that you are not going to say something about the population without dealing with the sample. So, you will be doing some math you know and it some of that might involved taking things like the sample mean sample standard deviation, but you will be doing some math on that and when you finish with that math you will getting something called the test statistic some of you might have heard of the these test statistics it may be called the z statistics or the t statistic and so on.

The crucks of null hypothesis the crucks of hypothesis test is that if we assumed the null hypothesis is to be true. And make some assumptions about distributions of various variables and those we won't go into that much, but if we assume the null hypothesis is to be true. Then technically the test statistics should be no different than drawing a

random number from a specific probability distribution.

So, in some sense what we saying is if the null hypothesis is to if that if the true it is a true mean is equal to 4.8, 4.8 this time, because the null hypothesis is true the null hypothesis was is it less than or equal to 4.8, so here the null hypothesis is true. So, let us take the extreme case the true mean is 4.8 and let us say there are some assumptions like that this distribution is normal may be that these vary the samples that you are taking are independent of each other some set of assumptions that you have to take.

If all that of its true, then we want to do certain things such that you will get another new distribution you will be doing some math with these data points. See these data points that you got from the sample you will be doing some math with those data points you will do that math's such that the test statistic would be no different than a single random number that you draw from a very specific probability distribution.

If that is the case, then you test the probability that the test statistic you calculated belongs to this theoretical distribution you basically say hey it looks to me like if the null hypothesis is true, then whatever I have calculated of here with the sample should be like drawing a random number from this distribution. So, let me calculate the actual test statistic and see how likely it is that this number came from that probability distribution and that is what we call is a P value.

Now, once you have done this process you might say look if the null hypothesis is true, then in the test statistic that I should have calculated should have come from this distribution. But, look at the number that I got in my hand it is, so unlikely that I could have gotten this test statistic from this distribution. Therefore, the null hypothesis perhaps was not true and I am going to reject the null hypothesis or in some cases the test statistics that you get looks like it does belong to this distribution the specific distribution.

And therefore, you can only you cannot really say anything it is like you just have no grounds for rejecting the null hypothesis you can just say I feel to reject the null hypothesis the important thing for you might want to look at the this procedure a couple of times, but the important thing that you might want to digest from this is that the P value itself is associated with the probability of seeing this data if the null hypothesis were true and not really the of probability of saying of this hypothesis being true given the data.

So, it is probability of data given hypothesis not probability of hypothesis given data. So, I hope that kind of clarifies inferential statistics for you and in the next class we will look at some specific tests and go over how you actually do these tests and even go deeper and talk about why we do some of the mathematical operations that we do.

Thank you.