Research for Marketing Decisions

Vaibhav Chawla

Department of Management Studies

Indian Institute of Technology Madras

Week - 08

Lecture - 36

Data Analysis using SPSS: Hypothesis Testing for Differences in Means P value, Sampling Distribution

I hope you have your internet data with the change with the changes one is there on your laptops. If it is there, please open this data. So we had to access we had to start with the second set of exercises. We finished the first set of exercises. If you remember where we had to do the data entry.

Data cleaning and descriptive statistics, cross-tabulation, chi-square, this we all finished in the last class and we understood what is p-value, we understood what is sampling distribution, so what I'm going to do today right now is, I'll run a video on p-value so that any confusion about that concept is quickly addressed. One of the most important concepts in statistics is the meaning of the p-value. Whenever you use Excel or other computer packages to analyze data, one of the key outputs is the p-value, or sig. In formal terms, the p-value is the probability that, if the null hypothesis were true, sampling variation would produce an estimate that is further away from the hypothesized value than our data estimate.

In less formal terms, The PWA tells us how likely it is to get a result like this if the law hypothesis is true. We will now go through this step by step with an example. Helen sells chocolate nutties. Recently she has received complaints that the chocolate nutties have fewer peanuts in them than they are supposed to.

The packet says that each 200 gram packet of chocolate nutties contains 70 grams of peanuts or more. Helen can't open up all the packets to check and then she wouldn't be able to sell any. So she decides to use a statistical test on a sample of the packets. The

null hypothesis, often called H0, is the thing we are trying to provide evidence against. For Helen, the null hypothesis is that the Choconutty's are as they should be.

The mean, or average, weight of peanuts in the packet is 70 grams. The alternative hypothesis, called H1 or HA, is what we're trying to prove. The customers have complained that the weight of peanuts is less than what it should be. So the alternative hypothesis is that the average rate of peanuts is less than 70 grams.

Helen decides to use a significance level of 0.05. If the p-value is lower than this, she will reject the null hypothesis. Having decided on her hypotheses and on the significance level, Helen takes a random sample of 20 packets of choconutties from her current stock of 400 packets. She melts down the choconutties and weighs the peanuts from each packet. If all the values were lower than 70 grams, with a mean of 30 grams for instance, it would be quite obvious that the bars did not have the required number of peanuts.

It is really unlikely that you would get 20 packets with a mean of 30 grams if the overall mean of all the packets in the population is 70 grams. Conversely, if all the values of the 20 packets are much higher than 70 grams, it would be obvious that there are enough peanuts and that there is nothing to complain about. However, in this case, the 20 packets contained the following rates of peanuts, and the mean is 68.7 grams. This caused Helen to ask herself, does this provide enough evidence that the bar was short of peanuts, or could this result just be from luck? She asked her brother to use Excel to find the p-value for the starter, comparing with the mean of 70 grams.

The p-value is 0.18. Judging from the data that we have, there is an 18% chance of getting a mean as low as this or lower if there is nothing wrong with the bars. That is, if the null hypothesis is true and the mean weight of nut is 70 grams or more. This key value of 0.18 does not provide enough evidence to reject the null hypothesis. In this case, someone does not have evidence to say that the bars are short of peanuts.

This is a relief. The smaller the p-value, the less likely it is that the result we got was simply a result of luck. If the p-value had turned out to be very small, we then would say that the result was significantly different from Cindy Grant's. In general, we start by saying that the null hypothesis is true. We take a sample and get a statistic.

We work out how likely it is to get a statistic like this if the null hypothesis is true. This is the p-value. If the p-value is really, really small, then our original idea must have been

wrong, so we reject the null hypothesis. p is low, null must go. A small p-value indicates a significant result.

The smaller the p-value is, the more evidence we have that the null hypothesis is probably wrong. If the p-value is large, then our original idea is probably correct. We do not reject the null hypothesis. This is called a non-significant result. The p-value tells us whether we get evidence from the sample that there is an effect in the population.

A p-value less than .05 means that we have evidence of an effect. A p-value of more than .05 means that there is no evidence of an effect. Sometimes a significance level different from .05 is used, but .05 is the most common one. Disclaimer. This video uses plain language to get difficult ideas across.

So, I hope the p-value concept is to some extent clear right. So, in the last class I was explaining about the sampling distribution. Sampling distribution is very very important when we are doing hypothesis testing right?. We are today going to do the hypothesis testing for differences in means. Hypothesis testing as I said in the last class is of two types.

One is hypothesis testing for association between variables. Hypothesis testing for differences between means or medians. That is what we are going to look today. Hypothesis testing between means or medians. Now for that for me to begin explaining that as I said p-value concept is very important, and if you recall sampling distribution was very very important.

How many of you are able to recall sampling distribution? I'll quickly tell once again what is sampling distribution so that, you know, we can proceed. There is no doubt. Right? So let's say I want to measure the I want to get the average height of BTEC students at IIT campus now.

Let's say there are 5000 students. I want to know the average height. I cannot go to each and every one. And 5000 students. So what I do is I obtain a sample of size 30

or let's say 35. I obtain a sample of size 35 each time if I and let's say once. The first time I obtained a random sample of from 5000 students, I have, let's say, list. I obtained random sample of 35 first time. Their average height is 5 feet 6 inches.

Second time again, I Draw a random sample, 5 feet 7 inches. Third time, 5 feet 6 inches. Fourth time, 5 feet 8 inches. If I keep on drawing infinite number of samples, I will get infinite number of, I will get so many mean values.

And if I draw these mean values on the y-axis, frequency on the x-axis mean so I will get a curve like this especially when the sample size when the sample size is 30 or greater than 30, I will get get a sampling distribution like this. Now what is the sampling distribution? when I draw when I drew infinite number of samples from 5000 population of btech students 35 sample at a time. For each time I draw 35 respondents, 35 BTEC students, I take down their mean and I plot that mean on this plot. So, what is this?

This is the mean of sample and this is the on the y-axis is frequency. So, every time I draw a sample of 35, I note down their mean. I keep drawing 35 sample from 5000 students. I keep drawing infinite number of times.

I keep getting means. Once I have these infinite means, I plot them on y-axis frequency which means how many times a particular mean has occurred and on y-axis the mean what I got from the sample. So, this is a curve which I will get. It is a normal distribution where the peak would be the population mean.

Mu is the population mean. And most of these values will be within the 95%. There will be on the tails it is 2.5%. So when I draw this particular curve for all the means that I extract from the population of 5000 BTEC students, and my sample is 35 each time I put the mean for infinite samples and draw this curve I will get this curve which is called sampling distribution. In the sampling distribution y axis is a frequency with which the mean in the sample is occurring on the x axis is the sample mean, and wherever I get the maximum wherever I get the maximum frequency of means,

the maximum time and number of a particular mean occurring, that is called the population mean. Same number of sample each time. So, this is the sampling distribution. If a sample is here, so let's say sampling the population mean is 5 feet 7 inches, this is 6 feet 9 inches, if a person is if if for a particular sample i find a value here which means, it is very unlikely because it is falling in 2.5 percent, there so some of the samples are very very unlikely to happen because if the population mean is 5 feet 7 inches, it is very unlikely that I get a sample.

Still it is possible because I got a sample. So when I plot this sampling distribution, this is plotting of the means of samples of size 35 when I draw 35 samples infinite number of

times from a population. So this sampling distribution idea we are going to use in the hypothesis testing for differences. Now how we will use this for the hypothesis testing? In the hypothesis testing, we assume, we start with assuming that null hypothesis is correct.

And then we look for an evidence to reject the null hypothesis. So in hypothesis testing there is null, there is alternate. We start with the idea that assume that null hypothesis is correct. Then we look at evidence to reject it.

For example, as I gave the example last time also, I just brag about my house saying that in the backyard of my house, there are thousands of coconut trees with average height of coconut trees being 100 meter. Right? Now, some of you might accept it. There would be someone who is very smart, who says that how could this be possible? I don't I understand that he's our teacher, but he also cannot.

He's assumed he's expected not to lie. But here is somebody who is lying. Who might be lying. But since he knows that I have been teaching with honesty, so he might think that. It might be there is some

possibility it might be correct as well. So, what he does is or he or she does is, they create a hypothesis, say null hypothesis is that population mean is 100 meter, and alternate is is not equal to 100 meter. So what that person does is he goes he assumes that null hypothesis is correct which means assuming that the population mean is assuming that the population mean is 100 meter then he conducts the hypothesis testing, he or she goes and draws and goes to my house yes they find that there are coconut trees. And he draw the sample of 35. And so he uses some machine to go up and down and measures the height.

I don't know how. And 35 sample and he got that the sample size is the mean height in that particular sample of 35 is 7 meters. Now we will come to sampling distribution to see whether to reject the null hypothesis or not. Now assuming that null hypothesis being correct which means assuming that the population mean is 100 meter, which means assuming that the population mean is 100 meter mu. How likely am I to get a sample with the

mean of 7 meters. So let's say, this is 100 and this is 125 meters and this side is 75 meters on the two tails so I am getting a sample which is 7 meters which is far on the left this is 2.5 percent 2.5 percent, which means the probability that I will draw a sample with such a

less mean is very very less less than 2.5 percent and if you actually it will be very very less even 0.001 percent that is the probability. So if the probability is such that it is less than 5 percent assuming that the population mean is this what is the probability of my obtaining a sample with a particular mean? if that probability is here or here which means towards the tails, which means it is very unlikely. When it is unlikely that you obtain a such sample, you reject your initial assumption of null hypothesis. I am not saying it is not possible, it is possible. So we take this risk that with 5% chance that we may be wrong, the alpha error, 5% chance but mostly 95% confidence we can say that null hypothesis is rejected.

Still there is some error probability, right? Because if you draw in the sampling distribution you are likely to get some samples with on the tails this could be one of the samples on the tails but you accept this risk to simplify if you obtain a sample which is towards the tails and your assumption is that null hypothesis is true as in this case, null hypothesis is 100 meters. Null hypothesis is that the population mean is 100 meter and you get a sample which is very unlikely then you reject the null hypothesis itself.

This is what is the use of sampling distribution. So whenever the sample size is 30 or above the sampling distribution will be normal distribution you need not conduct any normality of sample type of test but when the sample size is less than 30 you need to be sure that this the you need to conduct the test for the normal distribution of sample, but if the sample size is 30 or greater than 30 the sampling distribution will always be close to normal distribution. If the sample size is less than 30 you cannot say that the sampling distribution will be normal then you will have to first conduct the certain test to look at the normality of the population because the sampling distribution will not will is likely not to be normal, so always the idea is always have the sample size 30 or greater than 30. Now let's move to certain tests for today that we have to conduct.

First one is do the graphical and non-graphical test to examine the normality of internet usage data. I assume you have your laptops and the sheets are open. You are not preparing for the quiz. Anyways, the quiz is whatever is going to happen will happen. You will miss this also and that also if you are preparing.

Be present as you know, all the meditation, yoga, mental health, all they say that focus on the present. Don't go into the future and or look into the past. It will only worry. So the first one is do graphical and non-graphical test to examine the normality of internet usage data. Now internet usage data sample size is 30.

So we need not actually require doing it. But because there is a possibility in case you have the sample size less than 30, then you need to conduct this normality test whether the population is normally distributed if the sample size is 30 or greater, than 30 then you need not worry whether population is normally distributed or not because sampling distribution will be normally distributed, but if the sample size is less than 30 you need to worry whether the population is normally distributed, so you need to conduct this graphical and non graphical test for the normality of internet usage data, so we will quickly do it although they are not very very important, but we will do it still analyze descriptive statistics one is graphical test you can choose either pp or qq plot if you click on them you will get a window like this and internet usage in hours per week if you move to the variable box just click ok you will get a graph like this, if in this graph the points are close to the middle line if they are close to the central line, very, very close, almost sticking to the central line, then you say that the normality assumption is met.

Otherwise, it is not. In this case, it does not look to be satisfied. Likewise, a QQ plot is also there. we will look at now statistical test to do that because this is just a graphical plot go to you have to go to analyze descriptives explore internet usage move to dependent list and statistics here go to plots and click on normal plots with tests continue and let me do it again analyze descriptive statistics explore move internet usage into dependent list click on plots normality plots with test continue and okay you are getting this test of normality Kolmogorov Smirnov test and Shapiro will Kolmogorov

Smirnoff you have to look at if your sample size is greater than 2000 and if it is less than 2000 you have to look at Shapiro-Wilk anyways when the sample size is greater than 2000 you need not worry so Shapiro-Wilk is when the sample size is less than 2000 in this case it is 30 we need not do these tests but we are just looking at how and where to conduct this so Shapiro-Wilk value is saying 0.001 what does it mean It is, we are inferring about the population that the population, in the population the internet usage data is not normally distributed because 0.001. So, what is the null hypothesis here? Normally distributed. Yes, there is null hypothesis and there is alternate hypothesis.

What would have been happening in the background? of the test null hypothesis is that the data in the population is normally distributed and alternative it is not normally distributed right now in the background there is some some calculations are going on regarding when you assume that null assuming that null hypothesis is correct then SPSS is trying to fit the data on to that particular line and seeing the difference and thereby the p value is being calculated. So, the MBA student need not actually you know go further

to think about what is happening at the background unless they want to specialize into data science with the PhD student should actually because this is very very important. So this is what we did.

We need not do it because ampersand is 30. Now the next one is do a t-test to see the hypothesis that the mean familiarity rating exceeds 4, which is the neutral point. Let me simplify it. Do a t-test to see the hypothesis that the mean familiarity rating is equal to 4. When I say exceeds, it becomes a directional one-tailed.

I am looking at only positive tail. So, do a t-test to see the hypothesis that the mean familiarity rating exceeds 4, which is the neutral point in 7-point scale. So, if it is not a normal data, then how are we using the t-test and all of those things? That is what I said. 30 or greater than 30, your sampling distribution will be almost normal.

And because we are utilizing sampling distribution for Hypothesis testing population, sample and difference. Since that is close to normal, we could do these T tests. Here, isn't it failing the Shapiro test? It is failing because the population is not normally distributed.

It is showing, right? But we are worried about sampling distribution. If the sample size is less than 30, then we need to worry about because the sampling distribution will also be not normal. And then we need to worry about the population distribution of the data. If the sample size is greater than 30, even when the population is not normally distributed, we do not worry because sampling distribution will be close to normal and we could do all these test.

Most of the statistical test we will be able to do when the sample size is greater than 30. Now, the second one is do a t-test to see the hypothesis whether the mean familiarity rating exceeds 4. Let me simplify it. The mean familiarity rating is equal to 4. What will be your null hypothesis?

Let me make it even simpler. Mean familiarity rating For example, let's say somebody, let's say somebody works in Uber and they say, please conduct the hypothesis test that our mean satisfaction of our customers on five point scale is equal to four. Suppose this task is given to you. You will have to do it.

How you will do it? That is what it is. If you are not able to see how it will be used in your job role, this is how it will be. What will be the null hypothesis? New is equal to 4, right?

Alternate is new is not equal to 4. It is a very equal to, this is the same example that I gave. Mu is equal to 100 meters. Right? So now there is a sampling distribution and you have drawn this sample of size 30.

And you will see what is the p-value. How to do this? And I change equal to 4. Then we will come to exceeds 4. Let's do it first equal to 4.

Which means it can be less than, greater than. Let us do the bidirectional. We will look at p value on both the tails. That it should not be either on the left side or the right side. So, we will go to analyze.

Compare means. And this is one sample t-test where we are comparing. There is one sample we are comparing whether it is the population mean is different from 4. So, always when we are doing hypothesis testing, we are using the sample to infer something about the population. Sample is only a, it is only one of the ways by which we infer about the population.

So, here compare means one sample t-test, you will get this window. Now, we want to compare it with 4, whether the mean familiarity rating in the population is equal to 4. Mu is equal to 4. so familiarity we will move in the test variable test value we will keep it is for click options don't do anything with it confidence level is 95% and you click on ok you will get this result one sample t-test what is the significance to tell you are getting are you able to do that or shall I do it again okay you have to go to analyze compare means one sample t-test in one sample t-test the test value that you are going to test is 4 whether the mean familiarity is equal to 4 so your mu is so as the ola as the uber manager wanted you to do whether the mean satisfaction out of 5 is 4 so you put 4 here

and draw a random sample and their satisfaction values let's say whatever is it move it there and click on ok and you are getting this test this value significance two tailed point zero zero two what does it mean assuming that population mean would have been four what is the likelihood of obtaining a sample with the mean Here mean difference is given. What is the likelihood of obtaining a sample with the mean of 6.6? Here 6.6. 0.002 it is written.

What does it mean? It is very unlikely. Less than 0.05. Which means you reject the null hypothesis. What was your null hypothesis?

That a mu is equal to 4. Oh sorry I did with internet usage. I am sorry I have to do with familiarity. Sorry. Previously but I was sorry familiarity 4 and ok.

Yeah. So, we are getting 0.011. Again it is less than 0.05. So, our null hypothesis was mu is equal to 4. And alternate wise it is not equal to 4.

Then we ran this test and we are looking at whether this particular sample falls in the last two tails. And actually it is falling because the p value it is 0.011. 1.1% on the positive side. 0.011. Which means less than 0.05.

Which means we have to reject the null hypothesis. It is not in population mean is not equal to in population mean familiarity rating is not equal to 4. Now, if this would have been whether mean satisfaction of customers over customers is on a scale of 5 is greater than 4. Now, we are looking at positive one direction only not negative. Then one way is for 1 tailed, whatever 2 tailed value did you get, you have to divide it by 2.

If you divide 0.011 by 2, what it will be? Divide 0.011 by 2. 0.005 and then you have to compare it with 0.05 because now you are looking at only one tailed 5% at one tailed whatever probability value you got from two tailed multiplied or divided by two that is the p value for one tailed test now the mean difference is positive 0.86667 And you see familiarity with the rating mean is 4.8667 and mean difference is 0.8667. So you know what is there.

You know that when the test would have been conducted, the first one is the sample mean. The second one is the population mean. The difference would have been taken and the mean difference is positive, which means it exceeds 4. Because it is positive towards the positive side. See, when is equal to 4, when we did with whether the mean familiarity is equal to 4, we rejected the null hypothesis, it is not equal to 4.

Now, if it is not equal to 4, either it exceeds 4 or less than 4. And if you look at the mean difference, it is positive. And when this test is conducted, first is the sample minus sample mean minus the x bar minus mu. If you remember x bar sample mean minus mu. divided by S divided by under root N. So, first comes the sample mean, then the mu.

So, if mean difference is positive, what is the meaning? And P value is also 0.011 divided by 2, 0.005, less than 0.05. What does it mean? The mean familiarity rating exceeds 4. So, if you will have to do with for this Uber, for your employees, whether their satisfaction, job satisfaction is this,

But if you read your one-sided P here? One-sided P is, so this is two-sided. When directional hypothesis is there, we only look at one side. So, earlier this is 2.5%, 2.5%

here. Now, in directional hypothesis, we look at 5% on either side, whichever side is the hypothesis made.

So which means earlier it was only 2.5% area here, now it is 5% area. And this is mu is four, this is familiarity rating four, and this area is let's say 4.6, and your sample is somewhere here, not somewhere here, 4.8886. Which is in the 5 percent positive side. Mean familiarity rating exceeds 4. Because that was two tailed.

In two tailed you have the 2.5 percent here, 2.5 percent here. Now in one tailed you look at all the 5 percent on one side.