**MATH 243 - Project Two**

[00:00:00.41] SPEAKER: Hi, class. In this video, we will go over the second project in MAT-243. This project is in Module Five. So let's go over the template for the project. This is the template that you're going to use to write your summary report, and then we will also look at the Python script.

[00:00:19.28] So to download the template, you'll go to Module Five folder and then scroll down until you see Project Two Submission. Click on requirements and rubric. Then you'll scroll down to summary report Word document. This is not a zip file. It's a Word document file.

[00:00:37.81] So when you click this, this will open up the template. And here is the Project Two Summary Report template. Let's open the Python script in Codio. So I've opened that up here, and this is going to be Project Two Jupyter Script. So this is the Notebook for this project.

[00:01:01.33] So once this opens, here is a little bit of an introduction to the second project. Here are the variables that we're going to be working with. And again, you have your team that you have picked already in the first m and then the assigned team that's going to be given in step 1.

[00:01:23.98] So here is step 1. The assigned team is Chicago Bulls from years '96 to '98. This is the same team from the first project. So once you have run the first step, you should be able to answer the question in section 2 here. So your team and the assigned team.

[00:01:50.14] Now, the problem statement, you're going to discuss what is a problem that you're trying to solve, what data set are you using, and what methods are you going to use. So in this project, you're going to do a bunch of hypothesis tests.

[00:02:05.92] Are you going to do a hypothesis test for the mean, for abortion, and then for the difference in two means? So be sure to give a little bit of an introduction about what hypothesis testing is and how it's carried out and what are the various steps in a hypothesis test.

[00:02:30.37] So let's go to section 3 here. Section 3 is hypothesis test for the population mean. So suppose a relative skill level of 1340 represents a critically low skill level in the league. And the management of your team has hypothesized that the average relative skill level of your team is greater than 1340, and you tested this claim. And this is step 3 in the Python script.

[00:03:00.64] So the first thing you should do here is decide on what the null and the alternative hypotheses will be for this question. So I'll give you a hint for this one, and then I'll let you do the rest of the three hypothesis tests on your own. So in this scenario here, the management of your team is hypothesizing that the average relative skill level of your team is greater than 1340.

[00:03:37.69] Now, the null hypothesis always involves the equal sign. So the null hypothesis here will be that the average relative skill level of your team is equal to 1340, and then the alternative hypothesis is what your management is hypothesizing. The alternative hypothesis will be that the average relative skill level of your team is greater than 1340.

[00:04:02.95] The level of significance here is 5% or 0.05. So be sure to specify that as well. And here are basically the steps involved. So you should first write out the null hypothesis. Write out both the statistical notation and description in words. Alternative hypothesis. Level of significance.

[00:04:26.98] And now, you have to report the test statistic and the p-value in this table. So be sure to include this table and clearly specify the test statistic and the p-value and then the conclusion of the hypothesis test. So let's run this step here.

[00:04:44.73] So first, I will run the first step. And that step was successful here. And then we have to pick a team. So let's-- I will pick the same team I picked in the first project. It's the Suns. So I will run this step. And this is fine as well. And now, here is the hypothesis test.

[00:05:14.50] So step 3. So first you should make sure to read the instructions here. So it says replace this with the name of your team's dataframe. So your team's dataframe is your underscore team underscore df because that's where we picked the team.

[00:05:37.24] So I'm going to copy this. Right-click. Copy. And I have to replace this. So this is given right here. So I'm going to delete this. And then I will paste. Well, it looks like it didn't copy.

[00:05:56.99] So what I'm going to do is I'm just going to highlight this Control C and then get back here and then Control V. Now, it will copy it. You can also type it out. So if you read, it says your underscore team underscore df. So you can type this out here as well.

[00:06:18.26] The second edit is to replace relative skill with the name of the variable for relative skill. So to get the name of the variable for relative skill, there are two ways we can do this. First, you can clearly see in this printed dataframe. These are the first five rows.

[00:06:38.60] You have this elo underscore n, and that represents the relative skill level. So the variable name here is elo underscore n. But if you scroll up, elo underscore n is also given here in this table, and this is a measure of relative skill level of the team in the league. So this is the name of the variable-- elo underscore n.

[00:07:01.88] So I'm going to delete this. Now, when you're typing this, this is a variable name. So we have to-- and it says here enclose this variable in single quotes. So we have to put two single quotes and then type out the name.

[00:07:23.68] And now, the last edit is to replace null hypothesis value with the mean value of relative scale under the null hypothesis. So under the null hypothesis, the mean value of relative scale is 1340. That is given in the question, also in the template here, 1340.

[00:07:42.07] So I'm going to delete this and type 1340. And this sets up the hypothesis test. So notice we're using a Python library called scipy. it has a function scipy dot stats And. Then scipy dot stats. I'm using an alias here st, st dot ttest 1sample.

[00:08:12.61] So we're doing a one sample ttest. So this is also a function in scipy dot stats, and that function takes these inputs here. And then it will run the hypothesis test. We will get our test statistic and the p-value. And here, we're printing those. So let's print this. Let's run the script. And so we get 1495.59 as the mean relative scale of your team.

[00:08:43.36] So now, notice this is higher than 1340. But this being higher, is it statistically significant? So let's look at that. Here is the test statistic, and here is the p-value. p-value is 0. So because the p-value is 0 and the level of significance is 5%-- 0.05-- since the p-value is less than the level of significance, the test is statistically significant.

[00:09:08.62] And what this means is that the management of your team is correct that the average relative skill level of your team is greater than 1340. And that difference is statistically significant. So it is significantly different or significantly greater than 1340. That is the conclusion and the interpretation based on the p-value.

[00:09:34.64] And then be sure to discuss what are the implications of this finding. So what does it mean when we say that your team's relative skill level is greater than 1340? And the difference is significant. What are the practical-- what is the practical significance of this result? So be sure to discuss this as well.

[00:09:56.71] The second hypothesis test is for the population mean, again. But this time, the variable that we're working with is average number of points scored by your team. Now, the team's coach has hypothesized that the average number of points scored by your team in the team's years is less than 106 points.

[00:10:20.87] So the average number of points scored is less than 106. So just like before, you're going to use this statement to set up your null and alternative hypothesis. This time, the level of significance is 1%. And so now, we're going to go to step 4 in the Python script to run this test.

[00:10:41.18] So if we go to step 4. So here's step 4. And you can see you have to write your code on your own in this code block section. And the instructions are provided here. Now, the easiest way to do this test is, well, first, you'll see that this is pretty much the same test as step 3 except the variable is different.

[00:11:07.97] So to make it simple, I can just copy all of this. Well, I will highlight it and then Control C on the keyboard. And here, I will do Control V to type it out or to paste it. For max, there is probably a different way of copying and pasting. So I'm sure that will work here as well.

[00:11:35.64] Now, the dataframe for your team is called your team df. So make sure you use your underscore team underscore df. And we did that added previously. So we have the correct dataframe here as well as here. The variable pts represents points scored by your team.

[00:11:59.84] So the variable name is provided to you. So instead of elo underscore n, which was for relative skill, we will edit this and make it pts so that we are working with points. And the same thing here. Calculate and print the mean points scored by your team.

[00:12:20.71] So we're printing that in this line and in this line. In this line, we're calculating it. And then this line, we're printing it. Now, this is not the relative skill. It is mean points scored by your team. So I will make that edit so that it prints the correct line.

[00:12:44.41] Identify the mean score under the null hypothesis. You only have to identify this value and do not have to print it. So the mean value that is given is 106 points so that your null hypothesis will be based on the average that is provided to you in the question. So that's 106.

[00:13:07.26] So instead of 1340, which is what we used in the previous step for average relative scale, we will use 106 now. And then assuming that the population standard deviation is unknown, use Python method to carry out the hypothesis test. And again, the hypothesis test will still be one sample ttest just like before. And we will print the test statistic and the p-value.

[00:13:35.18] So I think everything is set up correctly. So let's print this out. So I will highlight this step and hit run. And so I get mean points scored by your team is 100. So for the Suns, the average score in years 2013 to 2015 is 100.94.

[00:13:58.36] Now, notice it is less than 106. The question is if this being less than 106 is statistically significant or not. So is it? Well, let's see. Test statistic is minus 6.41, and the p-value is 0. So the p value is 0.

[00:14:17.35] The level of significance in the question was 1%. So this is statistically significant. So again, you're going to answer these questions and all of these questions here as well as what are the implications of your findings from this hypothesis test, what is its practical significance.

[00:14:39.04] The next step involves hypothesis test for population proportion. So obviously, this step is different than the previous two steps because in those steps, we were working with averages. Now, we're working with proportion, and the proportion has a numerator and denominator. That's how you calculate a proportion.

[00:14:59.56] Suppose that the management claims that the proportion of games that your team wins when scoring 102 or more points is 90% or 0.90. So what is the claim here? The claim is that when your team scores more than 102 points, then the management is claiming that the team wins 90% of the games. And this is in step 5 of the Python script.

[00:15:28.13] So let's go to step 5. So here in step 5, you will see there are three edits that you have to make. Now, because we're not calculating the average, we're calculating a proportion. So we need to somehow calculate that proportion. And a proportion involves a numerator and denominator.

[00:15:47.65] So what is the numerator? The numerator is the number of times your team wins when it scores more than 102 points. That's the numerator. And the denominator is basically the total number of games your team has played when it scores more than 102 points. So that's the denominator.

[00:16:18.00] And so if you look at the code here, you'll see these two comments. First comment is the number of games won when your team scores over 102 points and the total number of games when your team scores over 102 points. So those two can be used to calculate the proportion. And you'll see that that's what is being done here. p equals counts, which is the numerator divided by this n observations or nobs, which is the total number of games when your team has scored over 102 points. So here is how we calculate this proportion, and then we're printing it out here.

[00:17:11.44] So the first edit is to replace count underscore var with the variable name that represents number of games won when your team scores over 102 points. So variable name. So this is a variable name that represents the number of games won when your team scores over 102 points.

[00:17:36.24] So clearly, that is counts because here, we're calculating-- first, we subset the dataframe to only the games where when your team has scored more than 102 points. So this is that subset data. And then we subset again on when the team has actually won, and then we sum it. So that gives us the count. And the count is the count bar here.

[00:18:09.54] So I will take this count underscore var here, and I will edit this and make it count. So this will be the variable counts here. The next one is replace nobs underscore var with the variable name that represents the total number of games when your team scores over 102 points. So this here will then be simply nobs, which is this variable name here.

[00:18:45.28] Null hypothesis value. Replace null hypothesis value with the proportion under the null hypothesis. So under the null hypothesis, remember, the proportion is not 102 points. The proportion is 90% or 0.90. So I'm going to edit this and make it 0.90.

[00:19:04.99] And then this prop underscore var, I also have to make this 0.90. And that is it. Now, I will print out. I will run the step and this will give me my test statistic and the p-value. And so the test statistic is negative 6.86, and the p-value is 0.0. And the proportion of games won by your team when scoring more than 102 points is 0.7009 or 70%.

[00:19:35.06] So we can see that's actually lower than 90%. But is that difference statistically significant? And you can see p-value is 0. The level of significance here is 5%, given in the template. So clearly, this difference is statistically significant.

[00:19:54.56] And then again, you will answer these questions. And be sure to discuss this, the practical significance or the practical implications of your finding. This is very important. What can you say after reaching the conclusion of your hypothesis test?

[00:20:11.81] And now, the last question, section 6, hypothesis test for the difference between two population means. So now, again, we're working with means, but now, we're working with two population means. We're trying to see whether they are statistically different or not.

[00:20:27.77] So you were asked to compare your team's skill level with the assigned team skill level. So we're comparing your team's average relative skill with the assigned teams average relative skill. And the level of significance is 1%. And this is in step 6 of the Python script.

[00:20:46.85] So if we go to step 6 here, again, we have three edits to do. So replace dataframe assigned team with the name of assigned teams dataframe. See step 1 for the name of assigned teams dataframe. So if we go to step 1, the assigned team's dataframe is here-- assigned team underscore df.

[00:21:11.12] So this is the-- I'm going to copy this. This is where we're subsetting for the Bulls, so this is the assigned team df. So when we go back to step 6 and let's look for dataframe assign team. And so the dataframe assigned team is right here. So I will delete this and paste on the dataframes name here.

[00:21:39.84] Now, replace dataframe your team with the name of your team's dataframe. Now, your team's dataframe was your underscore team underscore df. So I will edit this and just so I can just type it-- your underscore team underscore df. This is the name of the dataframe of your team.

[00:22:04.01] And then replace the relative skill with the name of the variable for relative scale. So remember, the relative scale, the variable name, again, this will be in single quote. Enclose this variable in single quote. So I will delete this here, single quote, and then the name was elo underscore n.

[00:22:25.54] And what are we doing here? We're running a ttest again for the difference between these two-- between the average relative skill of your team and the assigned team. And by the way, I have to edit this here as well. So I'm going to single quote elo underscore n, and that should do it. So we will run this step. We will get our test statistic and the p-value again. And by the way, this function here is also in scipy dot stat module.

[00:23:06.45] So if we run this. So the mean relative skill of the assigned team in years '96 to '98 is 1739, and the mean relative skill of your team in 2013 to 2015 is 1495. So clearly, it's lower than the assigned team. But is that difference-- is this different statistically significant? Well, our p-value is 0 again. And the level of significance was 1%.

[00:23:41.34] So clearly 0 is less than 1%, so the test is statistically significant. And so the conclusion here is that the mean relative scale of your team is different than the mean relative scale of the assigned team. And again, you have to answer these questions and be sure to, again, discuss the practical significance of this result.

[00:24:11.56] So this here concludes the second project. So just be sure for each of these hypothesis tests, write out the null hypothesis, both in statistical notation as well as description in words. Same for the alternative hypothesis.

[00:24:32.50] Clearly specify the level of significance. Report your test statistic in the p-value in this kind of a table. Then provide a conclusion for the test, and that will be based on comparing the p-value to the level of significance. And then interpret what it means.

[00:24:52.75] And then discuss the practical significance or the implications of these findings. And that is it. And be sure to submit the summary report as well as your-- the Python script here, which, again, will be in HTML format. So you can go to file, download as, HTML.