Transcript

Video – 0423 Unit4B Case CQ Paired Samples A

01. 00:01 / 00:08 - So we'll start again with the case of matched pairs. Here we have an X variable as our two-valued
02. 00:08 / 00:15 - categorical explanatory variable, which breaks our population into two subpopulations. Which
03. 00:15 / 00:21 - we could call subpopulation 1 and subpopulation 2. Or we could just rethink the problem and
04. 00:21 / 00:28 - think about it as population 1 and population 2 either terminology is fine. Our samples
05. 00:28 / 00:35 - here are dependent. We have paired data. The way we analyze this data is to reduce what
06. 00:35 / 00:42 - is now two samples into one sample. For each observation in sample one there is a paired
07. 00:42 / 00:48 - observation in sample 2. We can collect those pairs. Each of those values is going to be
08. 00:48 / 00:55 - a numeric quantity. One for sample 1 and one for sample 2. And then we simply take the
09. 00:55 / 01:00 - differences. We can calculate the difference between sample 1 and sample 2 and we can do
10. 01:00 / 01:06 - that in either order. Here we've written sample 1 minus sample 2 but it could just as easily
11. 01:06 / 01:12 - be sample 2 minus sample 1. And you can do that in whichever way makes sense to you in
12. 01:12 / 01:19 - a given problem. We are denoting our differences by d_1 to d_n. Again n is our number of pairs.
13. 01:21 / 01:28 - Notice that we have twice n numbers. We have n for sample 1 and n for sample 2 and that
14. 01:28 / 01:34 - gives us two times n total numbers. But because they are paired and we're reducing them to
15. 01:34 / 01:40 - differences, were left with just how many pairs. We get one difference for each pair.
16. 01:40 / 01:44 - Once we have the differences calculated we actually don't need the sample 1 in sample
17. 01:44 / 01:49 - 2 values anymore. If we were going to do this by hand we would be basing it solely on the
18. 01:49 / 01:54 - differences. In software packages you can usually do it either way. You can have the
19. 01:54 / 02:00 - software package take sample 1 and 2, it knows how they are paired, and it takes care of
20. 02:00 / 02:05 - the differences and solves the problem. Or you could calculate the differences yourself
21. 02:05 / 02:10 - and then have the software work with those differences. We will actually look at both
22. 02:10 / 02:17 - methods in our software. We are going to run through the steps and then we'll do an example.
23. 02:17 / 02:23 - The steps here are the same as they were for a one sample t-test. The hypotheses basically
24. 02:23 / 02:28 - are still the same. That mu is either equal to some number mu_0 or less than, or greater
25. 02:28 / 02:35 - than, or not equal to, that same number mu_0. But since we're interested in is there a difference
26. 02:35 / 02:41 - versus is there not a difference, the mean null value that we're going to be testing
27. 02:41 / 02:46 - against is going to be 0 (zero). We are also going to change the notation a little to clarify
28. 02:46 / 02:52 - that this is a paired difference problem. So we're going to use mu-sub-d to represent
29. 02:52 / 02:57 - the mean of the differences. It is important to note that this really is the same as mu_1
30. 02:57 / 03:03 - minus mu_2, or if you have it the other way around, mu_2 minus mu_1. It might be easier
31. 03:03 / 03:10 - for you to start out by thinking about your hypotheses in terms of population 1 and population
32. 03:10 / 03:17 - 2, but then convert them into hypotheses about difference. So mu-sub-d represents the mean
33. 03:18 / 03:24 - difference in the population. We are testing the null hypothesis that mean difference is
34. 03:24 / 03:30 - 0 versus the alternative hypothesis that it is either less than zero, greater than 0,
35. 03:30 / 03:35 - or not equal to 0. It could be a different number there. It doesn't have to be 0. In
36. 03:35 / 03:39 - this class we're going to focus solely on is there a difference versus is there not
37. 03:39 / 03:45 - a difference. So either there is a difference between these two populations or there is
38. 03:45 / 03:51 - not. We set up our hypotheses and then in step 2 we obtain our data, check our conditions,
39. 03:51 / 03:58 - and summarize the data. Since this is really just a special case of the one sample t-test,
40. 03:58 / 04:02 - we can use this under the same conditions that we talked about. The sample is random.
41. 04:02 / 04:06 - In this case the sample that we're talking about is the sample of the differences. So
42. 04:06 / 04:12 - the sample of the differences is random or at least can be considered random in context.
43. 04:12 / 04:18 - We are in one of the three situations marked with a green check mark in the table below.
44. 04:18 / 04:24 - So we have that the differences are either varying normally or the differences do not
45. 04:24 / 04:30 - vary normally. If the difference is vary normally, we are fine. If we have a large sample we
46. 04:30 / 04:35 - are fine. But if we have a small sample and the differences are not normally distributed,

47. 04:35 / 04:41 - then we will have a problem. And we will be talking about some ways to solve those problems.
48. 04:41 / 04:48 - In the case of small samples you want to check to see if normality is a reasonable assumption.
49. 04:48 / 04:53 - We can do that with the histogram or a normal probability plot. Once we are sure that we
50. 04:53 / 04:58 - can safely use the test, the data are summarized by a test statistic which is very similar
51. 04:58 / 05:02 - to what we saw. We are using a little bit of a different notation to emphasize that
52. 05:02 / 05:08 - we're doing a paired difference problem. We use y-bar-sub-d to represent the sample mean
53. 05:08 / 05:15 - of the differences and s-sub-d to represent the sample standard deviation of the differences.
54. 05:15 / 05:22 - The test statistic measures, in standard errors, how far our data are from the null hypothesis
55. 05:22 / 05:29 - value, which in our case is 0. The p value is calculated from t-distribution. So once
56. 05:30 / 05:37 - we replace sigma by s, we have a t-distribution with n -1 degrees of freedom and under that
57. 05:38 / 05:42 - distribution we calculate the p-values. We are going to allow the software to calculate
58. 05:42 / 05:49 - these p-values for us and focus on interpreting the results. Our conclusions are drawn in
59. 05:50 / 05:55 - the same way. If the p-value is small, there is a significant difference between what we
60. 05:55 / 05:59 - observed in the sample and what was claimed in the null hypothesis so we can reject the
61. 05:59 / 06:05 - null hypothesis and conclude that the categorical explanatory variable does affect (is related
62. 06:05 / 06:10 - to) the quantitative response variable as we specified in Ha, it is greater than, it
63. 06:10 / 06:16 - is less than, it is not equal to. If the p-value is not small then we don't have enough statistical
64. 06:16 / 06:22 - evidence to reject the null hypothesis. We go back in and compare our p-value to the
65. 06:22 / 06:29 - alpha level cutoff which is almost always 0.05. We can also support our result of our
66. 06:30 / 06:34 - test with a confidence interval. In this case our confidence interval is going to be for
67. 06:34 / 06:39 - the mean difference and we would then interpret that interval in the context of the problem.
68. 06:39 / 06:46 - If the null value of 0 falls outside the confidence interval, then we can reject the null hypothesis.
69. 06:46 / 06:52 - If the null value of 0 falls inside the confidence interval then we fail to reject the null hypothesis.