Transcript

Video – 0422 Unit4B Case CQ Introduction

01. 00:01 / 00:08 - So here we are in case CQ. We have a categorical explanatory variable and a quantitative response
02. 00:08 / 00:14 - variable. Back in the exploratory data analysis unit, we would have examined the relationship
03. 00:14 / 00:21 - between X and Y by creating side by side boxplots and by supplementing that display with the
04. 00:22 / 00:28 - corresponding descriptive statistics. So in general in this entire module we're making
05. 00:28 / 00:34 - inferences about the relationship between X and Y when X is a categorical variable and
06. 00:34 / 00:41 - Y is a quantitative variable. And this boils down to comparing the means in subpopulations
07. 00:42 / 00:49 - of our overall quantitative response. So for example if we measured heights as our response
08. 00:49 / 00:56 - variable and we looked at the categorical variable gender, we would be comparing the
09. 00:56 / 01:01 - heights of males to the heights of females and we'd be able to say that there is a relationship
10. 01:01 / 01:07 - between gender and height because once I know you're male or female it tells me something
11. 01:07 / 01:12 - about your height. So this figure summarizes the idea. We are going to have an overall
12. 01:12 / 01:18 - population of really measurements of Y, is one way to think about it, and then we're
13. 01:18 / 01:25 - going to split that population into subpopulations based upon which group you fall in. And to
14. 01:26 / 01:30 - infer about the relationship here, basically we're going to need to compare the means of
15. 01:30 / 01:36 - these subpopulations. We're going to split this into two different cases. The case for
16. 01:36 / 01:42 - k is equal to 2, we have exactly two groups, we will get through that and then we'll talk
17. 01:42 / 01:47 - about the case where k is greater than 2. So in the case where k is equal to 2 we are basically
18. 01:47 / 01:54 - going to be comparing two population means. Here we have some mu-sub-1 and mu-sub-2
19. 01:55 / 02:02 - being represented. I provided a table to summarize the tests that we're going to learn. Again
20. 02:03 / 02:08 - in the case where k is greater than 2, we have multiple groups, more than two groups
21. 02:08 / 02:14 - and we will still be comparing means but now I have to compare more than two means and
22. 02:14 / 02:17 - that becomes a little bit more difficult. It will be very easy for us to say there's
23. 02:17 / 02:22 - some difference here. It will be more difficult for us to find out where those differences
24. 02:22 / 02:29 - are. We have a table here summarizing what tests will apply in the case when k is greater
25. 02:30 / 02:37 - than 2. Before we can begin talking about any specific tests we have to define what we mean by
26. 02:37 / 02:44 - dependent verses independent samples. The difference between dependent and independent
27. 02:45 / 02:50 - samples is somewhat subtle depending on the scenario and it has to do with how the samples
28. 02:50 / 02:57 - were chosen. In some cases we have one group that has one of the categorical variables
29. 02:57 / 03:03 - and another independent group which has another value. For example if we took a random sample
30. 03:03 / 03:09 - of US adults, recorded their height and we also recorded their gender and then we split
31. 03:09 / 03:15 - the data into males and females as our two groups, those two groups would be completely
32. 03:15 / 03:20 - independent of each other. Notice that in the case of independent samples we allow the
33. 03:20 / 03:26 - sample sizes to be different. We have a simple random sample from subpopulation 1 of size n_1.
34. 03:26 / 03:33 - It has population mean mu_1. We have a simple random sample of size n_2 from subpopulation
35. 03:34 / 03:41 - 2 and it has a population mean mu_2. The idea here again is in order to have independent
36. 03:43 / 03:50 - samples, the individuals in the first group can't have any relationship at all to individuals
37. 03:50 / 03:56 - in the second. In other cases a matched pairs design may be used where each observation
38. 03:56 / 04:02 - in one sample is matched, paired, or linked with an observation in another sample. These
39. 04:02 / 04:08 - are often called dependent samples. One example here would be what if he went back to the
40. 04:08 / 04:14 - idea of heights. If we had heights again of males and females but now instead of having
41. 04:14 / 04:20 - a random sample of people, we have husbands and wives. We basically sample pairs. We sample
42. 04:20 / 04:26 - married couples and then we split them off. We have the male part of the pair and the
43. 04:26 / 04:33 - female part of the pair. Each observation in our sample from subpopulation 1 is matched,
44. 04:33 / 04:40 - paired, or linked with an observation in the sample from subpopulation 2. The matching
45. 04:40 / 04:45 - could be by person. We could have the same person measured twice. It could be individualswho
46. 04:45 / 04:51 - are paired in some other way, husband-and-wife, siblings, twins. Or we might even match people

47. 04:51 / 04:58 - based upon common demographics like age, race, and gender. Notice, by design in this case,
48. 04:58 / 05:05 - the sample size is n. And n represents the number of pairs. Now we are ready to begin
49. 05:05 / 05:11 - talking about comparing two means and we will begin with the matched pair case. Later we'll
50. 05:11 / 05:16 - come back and talk about the two independent sample case and then at the very end we'll
51. 05:16 / 05:20 - talk about more than two independent samples.