Transcript

Video – 0424 Unit4B Case CQ Two Independent Samples C

01. 00:01 / 00:05 - Before we move on to our examples, we are going to talk about the nonparametric test
02. 00:05 / 00:12 - that were going to see in this scenario. So this is the Wilcoxon Rank-Sum test or sometimes
03. 00:12 / 00:18 - it's called the Mann-Whitney U-test. In fact the Wilcoxon Rank-Sum test is technically the correct
04. 00:18 / 00:23 - test when the sample sizes are equal. It was developed by Wilcoxon and then Mann-Whitney
05. 00:23 / 00:29 - extended it to the case of unequal sample sizes, but the two names are used relatively
06. 00:29 / 00:36 - interchangeably in textbooks and literature. The test is used to compare two distributions
07. 00:36 / 00:41 - in independent samples. The idea behind the test is somewhat similar to the signed rank
08. 00:41 / 00:47 - test. In that it is going to be based upon the ranks of the values instead of the magnitude.
09. 00:47 / 00:53 - The same reason for this, ranking procedures help moderate the effect if any outliers or
10. 00:53 / 00:59 - any extreme skewness. For our purposes, we're going to assume that the quantitative variable
11. 00:59 / 01:05 - is a continuous random variable or can be treated as a continuous random variable and
12. 01:05 / 01:09 - that we're going to be interested in testing whether there's a shift in the distribution.
13. 01:09 / 01:13 - In other words, we are going to assume the distribution is the same except that one group
14. 01:13 / 01:20 - is sort of shifted higher or lower than the other. So we're going to assume that the distributions
15. 01:21 / 01:26 - of the two populations are the same except for horizontal shift and if that is the case
16. 01:26 / 01:32 - then the hypotheses can be formulated that the medians of the two populations are equal
17. 01:32 / 01:37 - versus the medians of the two populations are not equal. And as always, one-sided test
18. 01:37 / 01:42 - are possible, but for this test we're just going to talk about the two-sided test. Then
19. 01:42 / 01:48 - we obtain our data, check conditions, and summarize the data. Here we have to have two
20. 01:48 / 01:52 - independent random samples; all observations in each sample must be independent of all
21. 01:52 / 01:58 - other observations. This version that we're going to use, we're going to assume a continuous
22. 01:58 / 02:04 - random variable. So we should have a relatively continuous random variable. And we assume
23. 02:04 / 02:09 - that there's only a location shift. So we should check to make sure the two distributions
24. 02:09 / 02:16 - are similar except possibly for their center. So we're just shifting one group left or right.
25. 02:16 / 02:22 - The data are summarized by a test statistic which counts the sum of the sample one ranks.
26. 02:22 / 02:26 - And the way that we rank the observations is we combine all observations in both samples.
27. 02:26 / 02:31 - We rank them from smallest to largest. Then we determine which ranks came from sample
28. 02:31 / 02:38 - 1 and find the sum of these ranks. And the idea is very similar to the Signed-Rank test,
29. 02:38 / 02:42 - if the distributions were the same we would expect an equal distribution of large and
30. 02:42 / 02:49 - small ranks in population 1's sample and population 2's sample. And if we see a drastic difference
31. 02:51 / 02:55 - there, all of the small ranks come from group 1 and all the large ranks come from group
32. 02:55 / 03:01 - 2, then that's going to give us evidence that there's a difference in the location of these
33. 03:01 / 03:06 - two distributions. You won't be conducting this test by hand at all we're just interested
34. 03:06 / 03:10 - in explaining again a little bit of the logic behind where these tests come from and how
35. 03:10 / 03:17 - they work so you have some idea of what they're doing when you use the p-value. As for finding
36. 03:18 / 03:23 - the p-value and writing our conclusions, we will use the software to obtain the p-value
37. 03:23 / 03:27 - for us and then the conclusion will either be worded in terms of the medians, that the
38. 03:27 / 03:32 - medians of the two populations are either different or there's not enough evidence that
39. 03:32 / 03:38 - they're different. Or we can word it back into is the fact that this shows a relationship
40. 03:38 / 03:43 - between our categorical explanatory variable X and the response variable Y.