Transcript

Live Video – Dataset Introduction

01. 00:00 / 00:07 - What we have here is a dataset on urine. Very exciting. So we have people's urine samples
02. 00:07 / 00:12 - and I don't really know the population to which this data applies, so it could very
03. 00:12 / 00:18 - well be people who have problems and are seeing an urologist. We have quite a few variables
04. 00:18 / 00:24 - to look at, we have their ID number, we have calcium oxalate crystal presence, so this
05. 00:24 / 00:29 - is whether you have these little calcium oxalate crystals in your urine. Then, we have some
06. 00:29 / 00:36 - measurements on the urine, so specific gravity, pH, osmolarity, conductivity, urea concentration,
07. 00:36 / 00:43 - and calcium concentration. All of these are quantitative variables, except for this yes/no
08. 00:44 / 00:51 - variable. That is the dataset, our outcome variable, even though we actually don't cover
09. 00:52 / 00:56 - this situation in the course very much. Right? There's one of them that we're not going to
10. 00:56 / 01:03 - cover --we're not going to cover Q predicts C. That's what this scenario is. So we're
11. 01:03 / 01:07 - going to talk a little bit about it but in terms of statistical methods. We aren't really
12. 01:07 / 01:13 - addressing how to officially predict a categorical variable from a quantitative variable in this
13. 01:13 / 01:19 - class. But in this problem this is our outcome the yes/no calcium oxalate crystal presence
14. 01:19 / 01:24 - variable. Some other terminology we could use for that we could call it our outcome
15. 01:24 / 01:29 - variable, we can call it our response variable, or we could call it our dependent variable.
16. 01:29 / 01:35 - And it is categorical so we have this categorical outcome, or response. And then we're going
17. 01:35 / 01:41 - to have some continuous predictors so then our other variables were again were specific
18. 01:41 / 01:47 - Gravity, pH, Osmolarity, Conductivity, urea, calcium concentration, and cholesterol. And
19. 01:47 / 01:54 - I think today we're focusing on calcium concentration as a predictor of the presence of these
20. 01:55 / 02:00 – crystals. So again these are quantitative variables we really don't need to care too much for
21. 02:00 / 02:05 - our application today whether they were discrete or continuous. But really I do think all of
22. 02:05 / 02:10 - them will be treated as continuous. In any analysis. We could do regression, scatter
23. 02:10 / 02:15 - plots between these two variables, all that good stuff. Determine which the covariates
24. 02:15 / 02:20 - are related to the outcome, again covariates can also be called independent variables,
25. 02:20 / 02:26 - predictors, or explanatory variables. In general, you've already learned that outcomes can be
26. 02:26 / 02:31 - categorical or quantitative and that predictors can be categorical or quantitative so we've
27. 02:31 / 02:38 - seen the four possible ways that we can sort of organize a statistical question. We can
28. 02:38 / 02:42 - have more than one outcome in some studies, we might have two outcomes that are of interest
29. 02:42 / 02:49 - to us we can have clearly many covariates in certain problems and can have any mixture
30. 02:49 / 02:56 - of variable types. And we've also seen that and so here's the box plot comparing those
31. 02:57 / 03:04 - with these calcium crystals to those without. If I have QC, we did mention that one approach
32. 03:05 / 03:11 - at this level is to just act like it was the other way around. Right? Even though it's
33. 03:11 / 03:16 - not exactly the question we're asking for, it will at least let us investigate the relationship
34. 03:16 / 03:23 - but not necessarily get us to prediction. So I can see clearly there's a difference
35. 03:23 / 03:29 - in calcium concentration. Here's my calcium concentration. There is a difference between
36. 03:29 / 03:36 - these two groups. Correct? Clearly those with the calcium crystals tend to have a higher
37. 03:36 / 03:43 - calcium concentration and it really is up to - not statistics - to prove which direction
38. 03:44 / 03:49 - that goes. Even though I might want to use a quantitative to predict a categorical or
39. 03:49 / 03:54 - categorical to predict a quantitative, just because we've predicted it doesn't mean its
40. 03:54 / 04:00 - causal. Right? So I don't know if having the crystals causes my calcium concentration to
41. 04:00 / 04:06 - increase, or if having the increase concentration causes the crystals to occur, or there could
42. 04:06 / 04:11 - be some third variable out there that's causing both. That's not what statistics can usually
43. 04:11 / 04:17 - do for us even though we really really want it to. And I we'll see that as we go through
44. 04:17 / 04:23 - the semester as a theme for how to be careful about what you say about what you've learned.
45. 04:23 / 04:27 - Right/? But clearly there is a distinction between these two groups those with these
46. 04:27 / 04:33 - crystals have a higher calcium concentration than those without. And then here we have
47. 04:33 / 04:38 - the summary statistics to kind of put some numbers to what we're seeing in those graphs.
48. 04:38 / 04:45 - So we see that the median in the no group is 2.16 millimoles per liter, and the median

49. 04:46 / 04:51 - in the yes group is 6.19 that's where the line would be through the boxplot. We also
50. 04:51 / 04:55 - have the mean which is where the diamond would be in those boxplots, as well as the whole
51. 04:55 / 05:01 - five number summary. The group that has these crystals present has much higher variability
52. 05:01 / 05:07 - than those without. Meaning really some people are normal. But yet they still have these
53. 05:07 / 05:12 - crystals. right? That's what it says some people are normal they're still down here
54. 05:12 / 05:18 - in a low-end, but yet they still have these crystals present but then most people are
55. 05:18 / 05:25 - unusual on that calcium concentration scale in this group but comparison you would be
56. 05:25 / 05:29 - able to say something like the distribution of individuals with calcium crystals has more
57. 05:29 / 05:35 - variability than those without and then you can compare their center by saying the average
58. 05:35 / 05:39 - in the median in this group is much higher than the average or median in the no group.
59. 05:39 / 05:46 - Right? So yes much larger calcium concentrations on average than no. And we will be quantifying
60. 05:46 / 05:52 - all this again in probability. That's the dataset that we're looking at later we'll
61. 05:52 / 05:56 - be able to conduct an actual hypothesis tests to compare those means. Or compare those
62. 05:56 / 06:01 - medians but now we're going to be using this data to illustrate the concepts of probability.
63. 06:01 / 06:05 - One of the things we are going to do to facilitate this discussion on probability is we're going
64. 06:05 / 06:12 - to categorize calcium concentration. So here we're going to create a variable where 1 represents
65. 06:12 / 06:19 - zero to 1.99 really 9999 but its rounded to two decimal places so I can just stop it there
66. 06:19 / 06:26 - and then 2 is 2 to 4.99, 3 is 5 to 7.99, 4 is eight or more. We're going to have these
67. 06:26 / 06:31 - four groups that we're going to look at to categorize this variable.