LLO 8200: Variable Classification

Youtube Generated subtitles

Before we get too deep into our dive of data science/data analytics and how they can ultimately be used to make better decisions within your organizations, we need to spend some time revisiting some of the key statistical concepts that you learned in your stats class. These asynchronous sessions will discuss the different types of variable classification, descriptive and inferential statistics, and when we will use which procedures to explain or predict the thing that we are ultimately interested in.

So let's start with a very basic but essential topic of classifying the nature of our variables first thing we need to do is ensure that we even have a variable you would be surprised at the number of times that I have seen someone try to perform an analysis myself included where they believed they had a variable but it turned out to not be a variable we need to ensure that what we are noting what we are measuring what we are taking care to unpack that we are representing that it has it is in fact a variable that has some sort of variation that is to say that there's more than one possible value for it if we are doing a study looking at the spending behavior of lottery winners and we are only looking at lottery winners then being a lottery winner is not a variable it's a characteristic of our sample and the population that we are trying to generalize to so it may seem obvious but check your observations is this a constant that's present in all cases or is this something that can differ from case to case okay so let's say that you have insured yep I got variables now how are we going to go about classifying them I just want to say that for something that one would think is very straightforward you will actually find quite a bit of disagreement as to the best framework for thinking about how to classify variables I think that what I'm articulating here is going to be a practical way for you to quickly identify variable types I don't think it's the best or even the most theoretically appropriate but I think that it's a convenient heuristic to use to sort of move forward okay so you may have heard someone ask whether a variable is categorical or continuous usually what the person wants to know

is whether the variable is categorical or numeric or put another way

whether it's qualitative or quantitative okay we are terribly unoriginal in statistics when it comes to naming things odds are it's named after a person or it's just some sort of modification of its definition so qualitative variables are variables that express some quality they have a quality to them okay they can be expressed non-numerically and will typically be used to categorize

on the other hand quantitative variables will Express a quantity they will be expressed exclusively numerically and we can perform mathematical operations on them so when someone asks is it categorical or continuous what they're really asking is can I manipulate the values with arithmetic can I add subtract multiply divide if you can then the variable is quantitative if you can't then it's qualitative okay key to note here some qualitative variables will be expressed numerically but when you can easily replace those numbers with other labels like names of fruits or first the actual words first second third and it conveys the same kind of information the same meaning to the variables

then it's not it's qualitative and it's not quantitative right remember it's expressing the quality it's not expressing sort of the quantity

so once we have established whether it's qualitative or quantitative we then need to ask ourselves what's the scale of measurement for this qualitative or quantitative variable if it's qualitative then it's going to be either nominal or ordinal when it's nominal it's just a name right Nom name something like a student ID number is qualitative you could easily replace it with a student's name or a musical genre like the blues that is a level of a category that is just a name we contrast that with an ordinal scalar measurement where the category level has some sort of implied rank to it right think of things like placing in a race or letter grades first place is sooner than or better than second place and a is better than a b which is better than a c okay there's an order implied

if it's quantitative okay then the scale of measurement will be either interval or ratio if it's interval that means that we have equal intervals between values but not a meaningful zero okay so something like degrees Celsius zero degrees Celsius is not a meaningful zero it's just the point at which water freezes okay degrees Celsius does have equal intervals a 10 degree change is a 10 degree change okay SAT scores interval

a 100 Point difference doesn't matter where you are along the Continuum a 100 Point difference is a 100 Point difference okay

a ratio scale is an interval scale that also has a meaningful zero so something like the Kelvin scale right zero Kelvin means no heat that is a meaningful to theoretically meaningful zero points earned in a class a zero is a meaningful number in that case zero points earned in a class means zero points earned nothing zero

so let's work through a couple of examples here okay let's start with the number of votes in an election okay it's going to be qualitative or is it gonna be quantitative a number of votes in an election well it's quantitative right you are counting up the votes there's a number there's a numeric quantity attached what's a scale of measurement is it interval or is it ratio is there a meaningful zero well yeah right zero votes means no one voted for them zero support presumably so in that case we know that the number of votes in election is going to be a quantitative ratio scale okay so what about the highest degree that you achieve okay is it going to be qualitative or quantitative well it's going to be qualitative right you have an associate's degree you have a bachelor's degree you have a master's degree you have a doctorate degree but you could just as easily change those names to

banana apple orange kiwi okay but

what's a scale of measurement is it nominal are they just names or is there an implied order well there's an implied order right doctorate is at the top followed by Masters followed by bachelors followed by Associates it's not to say that any degree type is better than another it's just that there is an order a doctorate is a much is a more advanced degree than a bachelor's degree

okay so we know that it's going to be qualitative and will be ordinal now what about men's pants sizes men's jeans okay qualitative or

quantitative well if we think about men's genes for a second men's genes are expressed uh two numbers okay

now these numbers correspond to something what do they correspond to well they correspond to inches okay so a 36 is 36 inch waist and the 29 is a 29 inch inseam okay so if we take just the waist size for a second okay so the 36. what type of variable is it is it qualitative or quantitative

well since it measures since it is rooted in inches and a 36 is a 36 inch waist it tells you it's quantitative okay now what's the scale of measurement well we're measuring length so it's got to be a ratio scale because 0 inches is no inches or no length okay so we have Gene men's Gene sizes are expressed as two quantitative numbers that are on a ratio scale both of which happen to be length one measuring the waist one measuring the insane

what about women's Gene sizes is it qualitative or quantitative well as some of you know Gene sizes are as particularly in women's Gene sizes are notoriously inconsistent from manufacturer to manufacturer or from Cut to cut so if we ignore all of that and we just take a single manufacturer's list of sizes okay and we look at just this is going to be qualitative or quantitative well if we look at it just right off the bat it looks like it's quantitative right not so much not so much the sizes the double zero to zero to two to four those sizes but the corresponding waist size that looks fairly quantitative it's just going up incrementally by one okay but what's the waist size represent well it's representing again that length around the waist but notice something okay here we have a difference of one inch

one inch one inch one inch one inch one inch an inch and a half two inches an inch and a half again and then another inch and a half and then we go back to one

even though the waist size is incrementally increasing by one we see that the corresponding circumference of the waist is increasing by amounts that are not consistent okay so even though it seems to look like it might be a quantitative interval scale it looks more like it's going to be qualitative and it'll be ordinal okay maybe it's not the waste maybe it comes down to hips well we see the same problem and we actually see it sooner right here we have a difference of an inch and a half and here we have a difference of an inch okay again an inch an inch an inch an inch and then I go down to an inch again but now we're not starting on that half unit like we did in the previous ones okay so what we're seeing here is we're seeing here that when it comes to women's jean sizes the type of variable is actually qualitative we would actually be just get as much information by saying smallest size next size next size we could just name them size one through whatever right the amount of space between the units is not consistent so we know it's not going to be interval the next one down is going to be ordinal so the scale of measurement we have here is ordinal because a 2 is smaller than a 4. again working within the same manufacturers particular cut of jeans a size 2 is going to be smaller than a size 4. okay and a size 6 will be smaller than a size eight there is a rank implied but that order is not an equal distance when it comes to length which is what we're interested in here so women's jeans in contrast demands are going to actually be qualitative and they will be ordinal

here's a tricky one if that wasn't tricky enough here's a tricky one language proficiency okay what type of variable is it

that's a little less clear right it's a little less clear right how did we measure proficiency is this some score on a test if it is a score on a test and it's numeric then it's probably quantitative and they probably interval

but what if we didn't measure it as scores on a test but we measured it on a letter grade it's just a b c d e or we used a pass fail system is it still quantitative

not really it's qualitative okay so for the purposes are of our statistics and models that we'll use we would have to treat it as though it were qualitative in that case so we can't actually establish the scale of measurement uh unless we know specifically how we measured it the other examples were fairly intuitive okay but when it comes to something like language proficiency we need to think very clearly about how proficiency is operationalized okay because operationalized one way it will be quantitative and possibly interval or ratio but if it's not and it's not continuous it'll be categorical and it'll be either nominal or ordinal okay

it's essential that we be able to identify the types of variables that we are working with and the scale of measurement that the variable is on the flavor of the variable the type and scale is largely going to guide the kinds of models and statistics that we're ultimately going to be able to utilize okay

I have here at the bottom of this slide it's a modified form of the cheat sheet that you got at the end of Applied stats right if you work your way through it after identifying what the dependent and independent variables are you can look at the type of variable quantitative or qualitative and then how many of each of them that you have and you can sort of decide What statistic you end up needing okay now this cheat sheet is a guideline this is not the end all be-all of how What statistic what statistical test is most appropriate okay I know some of you want a flowchart like this so that you can just say my data is shaped like this and I'm going to be running this analysis because that's the analysis you run when your data is shaped like this okay the reality is is not that clear okay I can't provide you with the flowchart that's going to cover all instances in this class or in your day-to-day throughout this course and your work with variables outside of this course you're going to deviate from this chart and that's okay as long as those deviations are intentional and principled okay our hope is to prepare you in such a way that you can ask appropriate questions to determine which analysis or analyzes is better going is Will better serve my decision making for my organization

all right so over the next few videos you are going to review these analyzes emphasizing specific areas in this that this class will later elaborate on none of this should be completely new to you if you don't quite remember the minutia of the formula or the calculations that's fine that's totally okay what's going to be more important here is that you understand what these statistics are conceptually doing and to be able to read and interpret the results okay so from this point going forward we will be reviewing hypothesis testing and parameter estimation going through specific examples again nothing should be new but you may see

things framed a little bit differently so that we can sort of build on um we can build on these Concepts later down the road