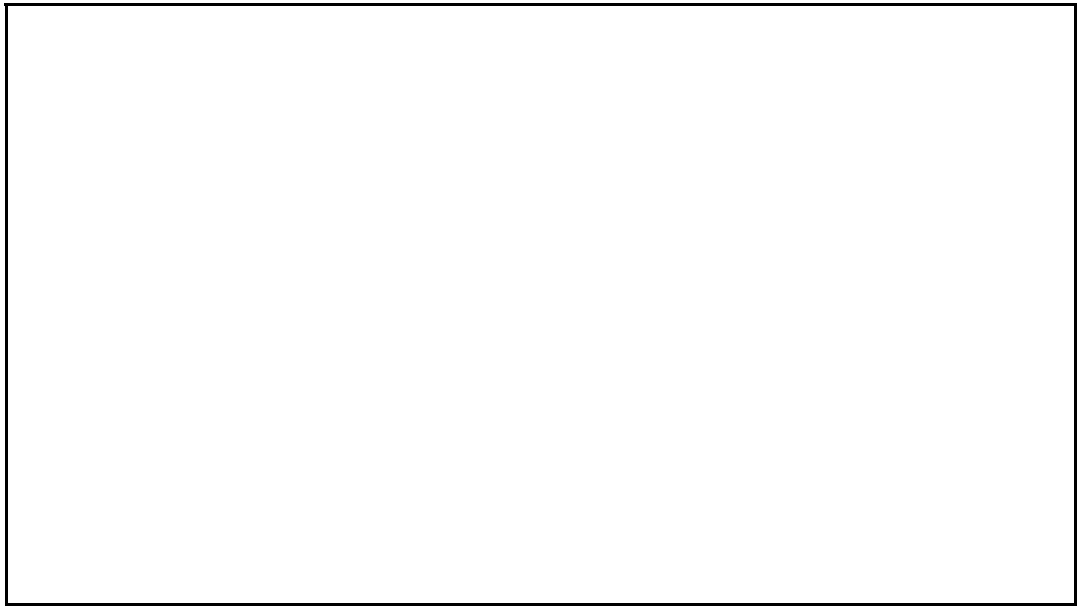
VARIABLES: WHAT THEY ARE AND WHAT THEY TELL US



4:43 / 4:43 1.0x

SPEAKER: MICHAEL J. MAHOMETA, Ph.D.

Once we have a hold of our data, we need to start making sense of it

- and the best way we can start to do that is by understanding

the KINDS of variables that make up our data.

Now basically, there are two major kinds of variables -

quantitative and qualitative

Now unfortunately that's a pretty broad statement ("two major kinds

of variables") - the fact is that there are actually nuances about

our variables that we actually need to understand before we move forward.

Help

So how will knowing the kinds of variables

that we have help us in our analysis?

Let's go ahead and take a look at variables a little bit closer.

Qualitative (or categorical) variables describe the "quality" of something

- a color, a size, a type, or an ethnicity.

Quantitative (or numerical) variables on the other hand

denote some sort of quantity - speed, height, or score.

Let's take a look at some actual data to get a feel

for this kind of variable classification.

Here's a dataset in R called mtcars.

It's data from 32 cars from Motor Trend magazine.

We can clearly see that some variables fit the idea of qualitative,

such as V or Straight, or Automatic or Manual.

They describe a quality of the car.

And some variables fit the idea of quantitative,

showing a quantity of a characteristic for the car,

such as miles per gallon or time it takes to run a quarter mile.

Help

Now, numerical variables can further be described

by assigning them as continuous or discrete.

Looking again at our dataset we see our time in the quarter mile.

Time is interesting because it can actually

be broken down into infinitely smaller and smaller and smaller units.

This infinite division in the numerical variable is what makes it continuous.

Now let's take a look at cylinders.

One the one hand, it's a number and it quantifies a characteristic of the car

- the number of cylinders in the engine.

But can we divide cylinders up into smaller and smaller units?

Well, no.

Can we have a car that has 4.3 cylinders?

I hope not.

Not unless it was in some major accident that we don't know about.

The point is, for discrete data, we can't make that division up

into smaller and smaller units.

There's yet another classification to variables called levels of measurement.

And there are four of them.

We have Nominal, Ordinal, Interval, and Ratio.

Our car names are the Nominal level of measurement

- it's simply the name of the car.

Imagine you classified each car in terms of size - small, medium, and large -

just like t-shirts.

This classification would be an example of an Ordinal level of measurement.

There's a specific order to the values within the measure,

but the spacing doesn't actually have to be equidistant all the way through.

The space from small to medium doesn't have to be the same as the spacing

from medium to large.

Now the Interval measurement level fixes that problem

by having distances between values be equally spaced (or at fixed intervals)

- but zero for the interval level of

measurement

doesn't actually mean anything.

It's just a place holder.

The most common example is temperature.

In our mtcars data, we actually don't have an Interval level of measurement.

Instead we have what are called Ratio data:

Data that has fixed intervals between values, just like interval,

but now zero actually means something - a lack of a characteristic.

Displacement and miles per gallon are in fact Ratio variables.

A zero means a car with no displacement, or a car

that gets zero miles per gallon.

So to wrap up, the variables in our datasets (in our data frames)

have specific characteristics that allow us to classify them.

This classification will ultimately help inform our decisions

as to which statistical tool we can use to answer our question about the data.

Just like the kind of paint I use when painting a room

helps dictate how I actually clean my brush,

so the kind of variables we have dictate the

kind of statistical tool

we can use when it comes to data analysis.

Categorical data lends itself to certain techniques,

while numerical can be used with others.

Comprehension Check

Help

Below are the first four rows of data from a survey of incoming university freshmen:

ID	Ethnicity	Major	Freshman Orientation	HS Sports	Math GPA
001	White	Biology	Yes	1	3.1
002	Asian	Art History	No	0	3.4
003	Hispanic	Computer Science	Yes	0	3.9
004	Hispanic	Undecided	Yes	1	2.3

(4/4 points)

The ethnicity of each student is what kind of variable?

- identifier variable
- categorical (qualitative) ✔
- numerical (quantitative)
- cannot be determined without additional information

The Math GPA for each student is what kind of variable?

- identifier variable
- categorical (qualitative)
- numerical (quantitative) ✔
- cannot be determined without additional information

Help

Each student was asked, "Did you play sports in high school?" Their responses were recorded as "1" for YES and "0" for NO. What kind of data do you have now?

- It is still categorical data, even though numbers are used to represent categories. ✔
- It is numerical data, because 0 and 1 are numbers.
- It is no longer data. The data has been modified from its original format.
- It is numerical data because we can count how many 0's and 1's we have, and you can't do that with categorical data.

Which of the following would NOT be considered raw data?

- Height of each skyscraper in Manhattan.
- Highest grade level completed by each of 16 taxicab drivers in Queens.
- Average age of ballet dancers attending The Juilliard School. ✔
- Highest temperature recorded each day of June 2014 in the Bronx.

Check Show Answer

Help

Below are examples of data:

- Jersey numbers of five players on an NBA basketball team
- Length (in seconds) of each song on a Michael Jackson album.
- Top five best-selling books on the New York Times bestsellers list.
- Height of children in a kindergarten class

(2/2 points)

Only one of the following scales of measurement is not represented above. Which is it?

- ordinal
- interval ✔
- nominal
- ratio

Which of the following is an example of interval data being converted to ordinal data?

- Finish times for a marathon reported as first, second and third place, etc.
- Age recorded as infant, youth, adult, and senior.
- Winter temperatures graphed in categories of 30-39, 40-49, 50-59, and 60-69 degree Fahrenheit.
- Happiness scores (measured on a scale of 1-10) recorded as not happy, happy, and very happy. ✔

Help

Check Show Answer