THE CONTINGENCY TABLE

8:01 / 8:01 1.0x

SPEAKER: MICHAEL J. MAHOMETA, Ph.D.

We know how to look at a the relationship between two

numerical or quantitative variables.

But what about categorical or qualitative variables?

Can we examine the relationship between two categorical variables?

Is that possible?

Well, absolutely.

Help

But to get there, we first need to examine

the two categorical variables AT THE SAME TIME.

Sure, examining one categorical variable is easy - say the breakdown of men

and women in a typical Introductory Stats class.

That would be simple.

It's just two percentages that add up to 100.

But now, how do we examine, say the sex in a class and its relationship

to the final letter grade?

Can we compare those two categorical variables at the same time?

Or - if you are a fan of cruises and blockbuster movies

- can we examine an even more interesting question?

We know that both variables are categorical.

We know that we could describe both variables individually

with a table of counts - remember we can count the number of rows in our data

frame that meet a certain criteria of our table.

If we do this for the passenger type, we'll

end up with 325 data rows that correspond to the first class

passengers, 285 data rows that correspond to the second class

passengers, and so on.

We can do the same thing for the Survivor status.

Finding that 711 passengers survived, while 1490 did not.

What's important to note here, is that these single variable values -

these two table of counts - both add up to 2201 -

the full contingent of persons on the Titanic.

We'll talk about "full" data a bit more in a little bit.

Now, looking at the two tables of counts individually,

we really don't actually answer our question of interest.

We know that we need to find the relationship between two

categorical variables and looking at the individual tables

just won't get us there.

Much like when we combined numerical data into a scatterplot,

and saw a single point represented as the intersection between two

Help

continuous variables, so we can now "combine" categorical data

and examine the "intersection" of characteristics.

How do we do this?

Help

It's called a contingency table.

The contingency table literally shows us the intersection

of characteristics between two variables.

The characteristic count of one variable contingent on the characteristic

of a second variable.

To create the contingency table, we do much the same procedure

as we do with the table of counts.

We count the number of times we have an intersection of the two

variable characteristics.

Here's the blank contingency table for our Titanic data.

The number of rows in the data frame (or counts)

that meet the criteria of first class and not survived

is 122, while the count of rows that meet

the criteria of first class, survived is 203.

Now, in reality, we would never do this by hand -

going though and counting all the pieces of information in our data frame.

Instead, we simply ask software to do the counting for us.

Here's the full contingency table for the Titanic passengers.

Help

Each cell contains the count of rows that

have the combined characteristic of the two variables.

Now, by itself, the contingency table is OK.

It's helpful, but it doesn't really help answer our question of interest.

To do that, we need to ask for what are called

the marginal sums or the marginal distributions for each variable

- represented here on the table.

This total column represents the marginal distribution

for the Survivorship variable, and this total row

represents the marginal distribution for the Class of passenger.

Notice that the values in each of the total groups

match the individual tables of counts for each of our two variables.

This will ONLY occur when there is no missing data.

Help

For example, if a third class passenger was missing their Survivor status,

that passenger would NOT be included in the contingency table of Class

and Survivorship - we wouldn't know where to count them - did they survive

or did they not?

We just wouldn't know, so we can't include them in the contingency table.

The take home here?

Always use the marginal distributions or marginal totals

from the contingency table.

They will always be the correct ones, representing

the valid data used in the contingency table.

We'll end up using these marginal totals to help

answer our question of interest.

Now examining the just the counts in the contingency table

doesn't really help answer our question.

We can't readily see the relationship between two categorical variables.

But, if we use proportions (or percentages),

we can start seeing a pattern.

We technically have three choices of percentages to use:

Help

the table percentages, the row percentages,

and the column percentages.

The table percentages are usually ignored.

They really don't help us in answering the question.

Instead, we turn to our column or row percentages.

And we have a choice to make - which one do we use?

Well it depends.

What story is the most impactful from our table?

First, the row percentages.

These are the conditional distributions of Class, by Survivorship.

So for each level of Survivorship, we have the break down of Class level.

If we sum across the Class percentages for each level of Survivorship,

we'll get 100%.

The column percentages work the same way.

They represent a conditional distribution of Survivorship

for each level of Class.

Again, summing the percentages in each Class will give us 100%.

It's important that we can really understand

Help

the contingency table.

And that we can readily use it to our advantage.

For example, what's the percentage of survivors

that were second class passengers?

Can you find the percentage?

It's about 17%.

What about the percent of Crew that did not survive?

Well that's about 76%.

All we've just done is determine our conditional probabilities -

one for Class, and one for Survivorship.

We did this by using our row and column percentages and our row and column

totals.

So which do we use?

Row or column percentages?

Well, which conditional distribution gives us an interesting or impactful

story to tell about our contingency table?

That's the better question.

I'll tell you what - you work on that question

for a little bit: which conditional distribution should we use?

We'll answer it in the next video, along with continuing

to answer our question of interest.

Help

Comprehension Check

Here is data from a research study on the relationship between sex and fear of heights.

	Men	Women
Expressed a Fear of Heights	68	109
Did Not Express a Fear of Heights	94	89

(6/6 points)

1) What fraction represents the proportion of people in the study that were afraid of heights?

- 162/360
- 177/360 ✔
- 183/360
- 198/360

Help

2) How many simple events (outcomes) were possible for participants in this study?

- 0
- 2
- 3
- 4 ✔

3) **Of those participants that were afraid of heights**, what percentage were men? (*Round to 1 decimal place with no % sign.*)

38.4

38.4

4) What percentage of **women** were afraid of heights? (*Round to 1 decimal place with no % sign.*)

55.1

55.1

5) What percentage of **all participants** were not afraid of heights? (*Round to 1 decimal place with no % sign.*)

Help

50.8

50.8

6) A contingency table is used to simultaneously display counts of

- (•) two categorical variables. ✔
- ( ) two continuous variables.
- ( ) two quantitative variables.
- ( ) two variables that cannot be graphed.

Check | Show Answer