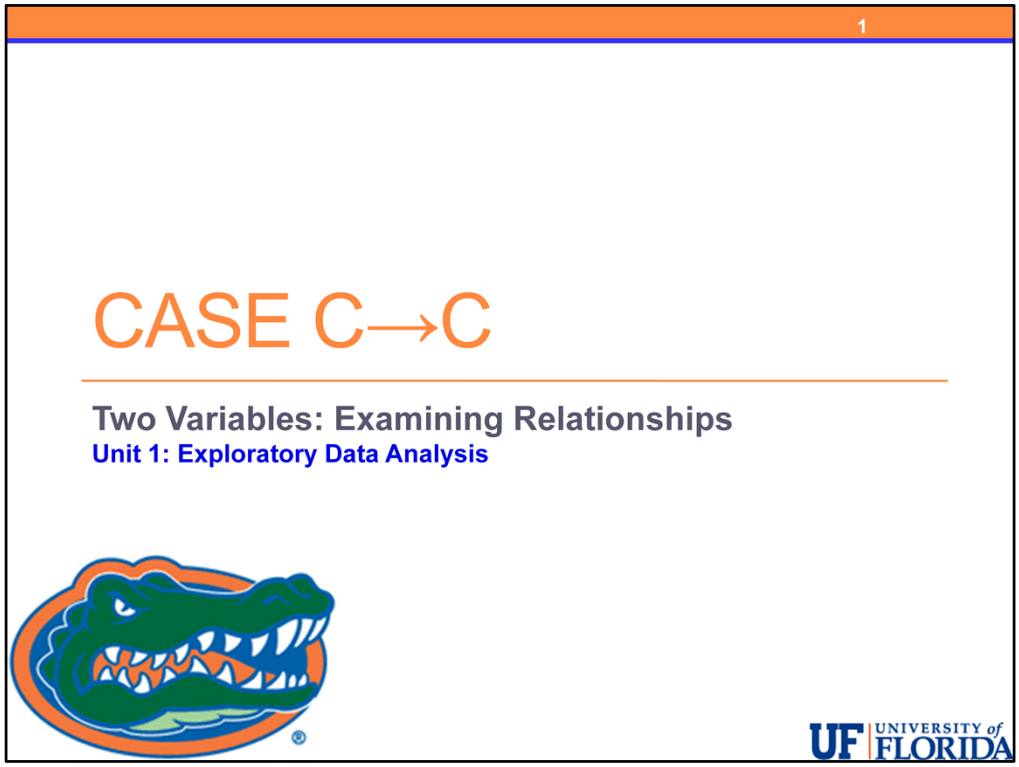
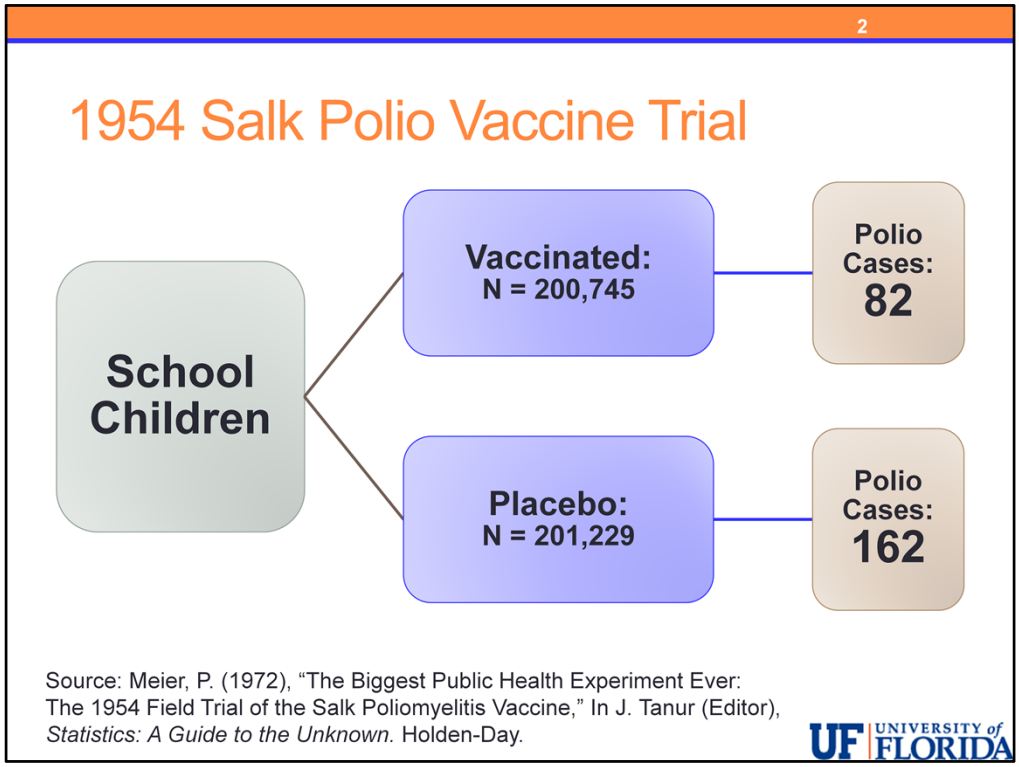



Case C-C, where we want to determine the relationship between two categorical variables, is relatively easy but it does have a few issues that can trip you up when you are asked to answer questions based upon data.




Let's look at the data from a very famous study. It was ground breaking in public health here in the united states. The 1954 Salk polio vaccine trial.

This is a study where a little over four hundred thousand school-age children were randomly assigned to one of two groups, a vaccine group or a placebo group. Meaning effectively someone flipped a coin for each child and if it came up heads they were assigned to one group and if it came up tails then they were assigned to the other. The assignment was purely due to chance. There was nothing about each child that predetermined which group they would be in.

Roughly two hundred thousand children were randomized to each of the two groups.

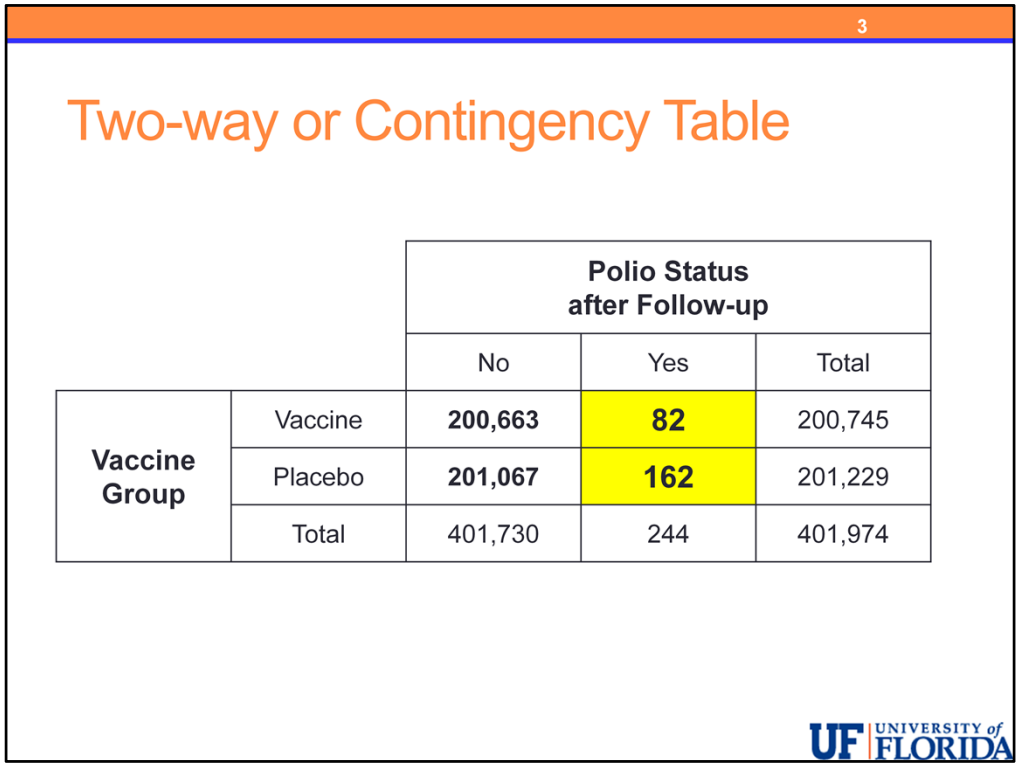
After a follow-up period there were about half the number of polio cases in the group that was vaccinated compared to those that were not.

82 cases out of the two hundred thousand seven hundred forty-five that were vaccinated verses 162 in the two hundred and one thousand two hundred and twenty nine given the placebo.

A few comments about the design of this study. It was actually double-blind which means neither the patient, the student or their family, nor the physician who was administering the treatment or placebo knew which group they were in. And the objective of the randomization was that the groups should be equivalent except for the factor of the

vaccine being investigated.

What are the standard visual displays and numerical measures which are appropriate in this case?




In the case where both of our categorical variables are binary, we can summarize the situation fairly easily in a variety of simple ways – often verbally.

For the general case of two categorical variables, the typical visual display which accompanies an analysis in Case C-C is a two-way table or contingency table. Occasionally, we may also present an actual graphical display such as a grouped bar chart, but often, like pie charts and bar chars, this uses more valuable space than simply explaining the result in words, providing numerical support as required.

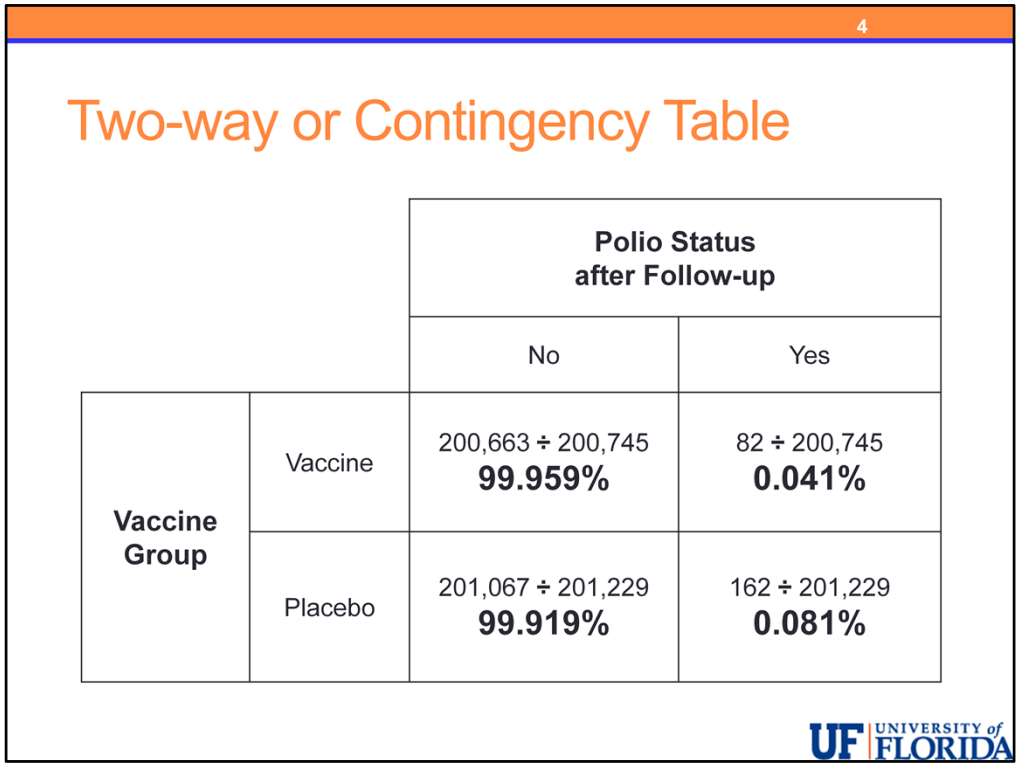
Two-way or contingency tables present one of the variables in the columns and one of the variables in the rows.

The orientation of variables is up to you, however, often we place the explanatory variable, in this case, vaccine group (vaccine or no vaccine) in the rows and the response variable – Polio status at follow-up - in the columns.

You can see that in this case, a graph might not add much information given the rarity of the disease. Pie charts or bar charts overall or for each group would provide little benefit.

What numerical measures would be appropriate?

Although in this case, the groups are very similar in size, we still would like a more precise comparison.




At this stage, we usually desire a comparison of the distribution of the response variable – Polio status between the levels of the explanatory variable – Vaccine group.

If we only had the variable Polio status, we would use a frequency table - which has the count (given above for this data) and the percentages within each Polio status group.

Here we desire the same percentages – those with and without Polio – taken within each of the two Vaccine groups. These are generally called conditional percentages – the word conditional indicating that we are given some additional information or conditions which we must consider – "among those vaccinated" (this is the condition) what percent (or probability) developed Polio during the follow-up period? (This is the percentage or later … probability).

When we discuss conditional probability, we will point out that it is equivalent, for similar data structures, to what we are doing in Case C-C for exploratory data analysis with conditional percentages.

These percentages are also often called "row" and "column" percentages, especially by software packages which often give these conditional percentages for both variables.

Although the difference might not seem like much – 0.08% in the placebo group and 0.04% in the Vaccine group, for such a rare event and a large sample size, the question becomes, could those results – the results of cutting this percentage in half in our study – be due to

chance alone? If we had conducted the study again would we have found the opposite results or is that very unlikely.

What do you think? We'll come back and answer this question of statistical significance later!

Statistical methods will tell us how to make these probability calculations. How to figure out whether what we saw was likely if the vaccine was no better than the placebo or was unusual.

For this data,  it was presented, policies were made, and polio vaccines became the norm for younger children.

# Another Example – Using SAS

| Table of gender by cvd | | | |
|---|---|---|---|
| gender(Gender) | cvd(History of Cardiovascular) | | |
| Frequency<br>Percent<br>Row Pct<br>Col Pct | No | Yes | Total |
| Male | 89<br>17.80<br>29.67<br>71.20 | 211<br>42.20<br>70.33<br>56.27 | 300<br>60.00 |
| Female | 36<br>7.20<br>18.00<br>28.80 | 164<br>32.80<br>82.00<br>43.73 | 200<br>40.00 |
| Total | 125<br>25.00 | 375<br>75.00 | 500<br>100.00 |

- In this case we want to compare the prevalence of CVD between Males and Females
  - Males: 70.33% have history of CVD
  - Females: 82% have history of CVD

UF UNIVERSITY of FLORIDA

Here is another example, using SAS software. We have gender in the rows and history of cardiovascular disease in the columns.

This table provides a complete summary. In each cell it first gives the frequency, then the overall percent (which is rarely of interest), followed by the row percentages and finally the column percentages.

Here we can see that among males (in the male row), 70.33% have a history of CVD and among females, 82%. You might think this is a very high number, however, this is the WHAS data and each of these individuals is in the dataset due to their admittance to the hospital with a heart attack.

# Example: SPSS (Column Percent)

Smoke_Cigarettes * Gender Crosstabulation

| | | | Gender | | |
|---|---|---|---|---|---|
| | | | Female | Male | Total |
| Smoke_Cigarettes | No | Count | 120 | 89 | 209 |
| | | % within Gender | 94.5% | 89.9% | 92.5% |
| | Yes | Count | 7 | 10 | 17 |
| | | % within Gender | 5.5% | 10.1% | 7.5% |
| | Total | Count | 127 | 99 | 226 |
| | | % within Gender | 100.0% | 100.0% | 100.0% |

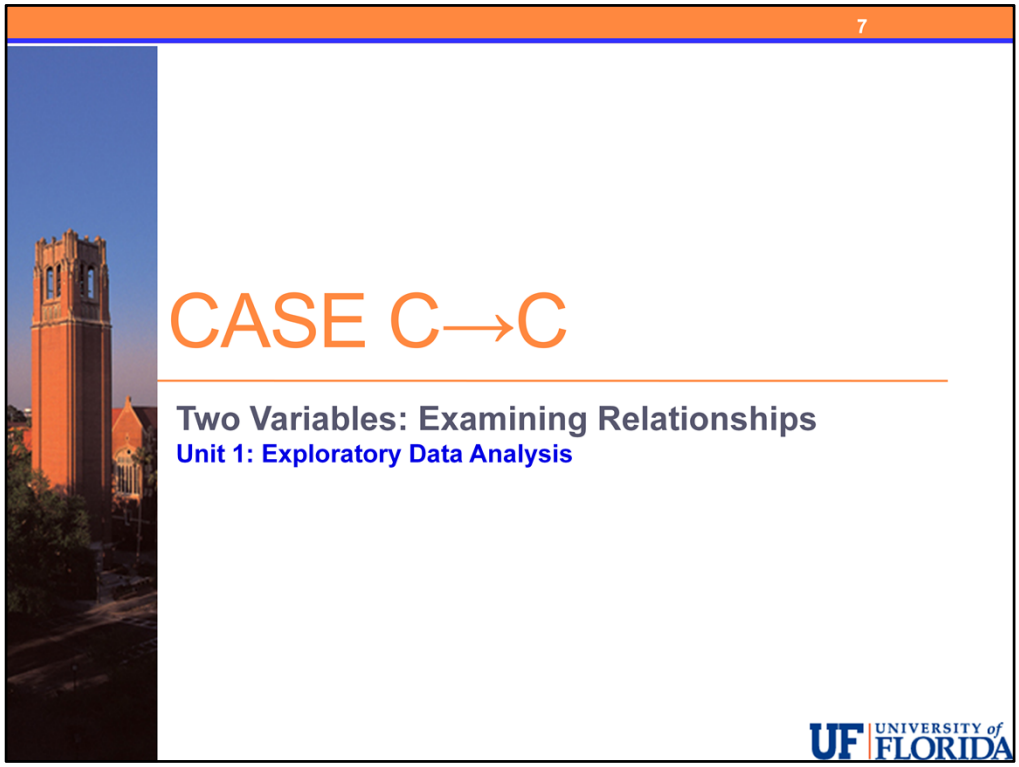
UF UNIVERSITY of FLORIDA

We can obtain similar output from SPSS, although it is labeled differently.

Here we have the frequency or count and what is labeled % within gender. This indicates the percentages presented in the row, are taken by dividing the counts by the gender totals. These are column percentages in this case. SPSS chooses to label them by their variables.

What we can see here is that in this sample, 5.5% of females and 10.1% of males smoke cigarettes.

In SPSS, we could also choose to obtain the row percentages, overall percentages, and some other quantities we will learn about later in the course.




In summary, for two categorical variables, we present a two-way table or contingency table which contains the frequency broken down by both variables simultaneously. This is the visual display we most often use to represent our data.

We want to determine how does the pattern of the distributions containing conditional percentages for the response variable compare for the different levels of our explanatory variable.

These conditional percentages are row percentages if the explanatory variables is in the rows – we ask: what percent WITHIN each row fall into each of the categories of our response variable (in each column). We divide by the row total.

These conditional percentages are column percentages if the explanatory variable is in the columns – we ask: what percent WITHIN each column fall into each of the categories of our response variable (in each row). We divide by the column total.

If we are simply looking for an association between two categorical variables and the role of the variables is not specified, we might find both sets of conditional percentages useful (those based upon row percentages and those based upon column percentages).

Overall, the material in Unit 1 is the foundation of statistical analysis for one and two variables, so far we have covered exploratory data analysis for One Categorical Variable, One Quantitative Variable, Two Variables with One of Each Type – including Case C-Q and

to a lesser degree Case Q-C, and Two Categorical Variables – Case C-C.