# 

**Data Analysis with Categorical Data - Stata**

Today we’re talking about how to determine if you have interaction or confounding in categorical data analysis, and all need to know about stratified analyses.

# 

# Categorical Data

## Hypothesis Testing

To begin analyzing categorical data, you might start with hypothesis testing. For two independent binomial proportions, you can use the Chi-square test if you have a large amount of data, or the Fisher’s exact Test for smaller samples. For two dependent binomial proportions in which we have paired data, you can use McNemar’s test. But what if you have multiple levels of a categorical variable such as different age groups, or different levels of exposure?

When we are interested in summarizing a series of tables across the levels of a categorical variable, then we can use categorical data analysis. There are three main steps:

1. Determine whether the association between the risk factor and the outcome differs for each level of the third variable - that is interaction
2. If there is interaction, report the results separately for each level of the variable
3. If there is no interaction, then you need to
   1. Determine whether the association between the risk factor and the outcome is not due to random variation
   2. Determine the magnitude of the overall association between the risk factor and the outcome

Let’s go through each step one by one.

## Inputting Data

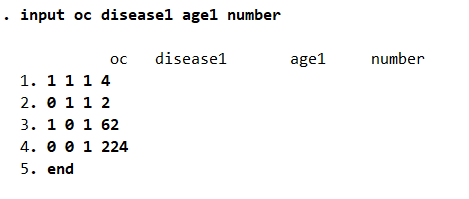
In most biostatistics courses they will ask you to input data manually. In reality, you will rarely do this because a dataset will be provided to you, but it’s a good skill to know how to do just in case.

In Stata, begin inputting data with the following command:

input *riskfactor outcome category* number

Where riskfactor is your exposure, outcome if your outcome variable, category is your categorical variable and number is the cell value for that combination of variables.

For example, you might be interested in the effect of oral contraceptive use on the risk of a heart attack in different age groups. Here is a simple 2x2 table and coordinating code.The code would look like:



When you are ready to stop entering data, type end.

## Testing for Interaction

Once your data has been entered, you are ready to start testing for interaction. You can then use the following command to test for interaction:

cc *outcome riskfactor* [freq=number], by(*category)*

In the output, look for the “Test of homogeneity (M-H) chi2” line. This line tests the hypothesis that the odds ratios are the same. If the “Pr>chi2” value is <0.05, you can reject the null hypothesis and conclude that the odds ratios (crude vs. adjusted) are significantly different and you have interaction.

In this case, you should report the odds ratio by each group. This is considered a stratified analysis. You do not need to assess for confounding if you have interaction.

*Note:* Epidemiologists typically say “effect modification” while biostatisticians use “interaction”. They mean different things, but are unfortunately used interchangeably. You can [read more about the difference here](https://journals.lww.com/epidem/Fulltext/2009/11000/On_the_Distinction_Between_Interaction_and_Effect.16.aspx).

If you have data inputted as variables, you can easily do the same thing as above by using the logit command in Stata. Learn more about the logit and assessing for interaction [in this guide from UCLA](https://stats.oarc.ucla.edu/stata/webbooks/logistic/chapter2/).

## Testing for Confounding

If you do not have interaction, you’ll want to test for confounding of the association between the risk factor and the outcome. In Stata, look for “Test that combined OR = 1” which tests if the Mantel-Haenszel combined odds ratio = 1. If the “Pr>chi2 =” is <0.05, then we can reject the null hypothesis and conclude that there is a significant effect of the risk factor on the outcome, adjusted for the categorical variable.

## Magnitude of Confounding

Using the output from Stata, you can assess the magnitude of the confounding by determining if the difference between the adjusted and the unadjusted odds ratios are different by more than 10%. This is a general rule, and some professors use a different percentage. However, the 10% rule is usually good enough to assess for confounding. Here is the equation in Stata:

display ((ORadjusted - ORcrude)/ORadjusted) \* 100

Where ORadjusted is the M-H combined OR, and ORcrude is the Crude OR in the Stata output. If this number is >10, then you can conclude there is confounding and you should report the M-H odds ratio. If it is <10, you do not have confounding and you should present the crude odds ratio.

## **Quick Review**

To assess for confounding and interaction in categorical data analysis, first start by calculating the odds ratio in each group of the categorical variable. If they are significantly different, you have interaction and you don’t need to assess for confounding. If they are not, you should assess for confounding. If your odds ratios are different by more than 10% then you have confounding and you should present the M-H odds ratio. If they are not different by more than 10%, then you don’t have interaction or confounding, and you should present the crude odds ratio.

| Crude = Strata1 = Strata2 | No interaction or confounding |
| --- | --- |
| Crude ≠ Strata1 = Strata2 | Confounding, report M-H odds ratio |
| Crude ≠ Strata1 ≠ Strata2 | Interaction, report the stratified odds ratios |