Lab 6

Public Health 241: Statistical Analysis of Categorical Data

YOUR NAME / YOUR STUDENT ID HERE

TODAY'S DATE

In this lab, we'll cover tools you'll need to complete Homework 7 (concepts covered up to and including Chapter 9), and recap/add to the general-use functions we've learned in previous labs. Functions introduced in this lab will allow us to consider the concepts of interaction and confounding as they affect our analysis of the main effects of a principal exposure variable.

1. Stratified Analyses

Today, we'll again be using the Western Collaborative Group Study data.

1. Load the WCGS dataset into R.

```
wcgs <- read.dta("data/wcgs.dta")
```

```
## Warning in read.dta("data/wcgs.dta"): cannot read factor labels from Stata
## 5 files
```

2. Familiarize yourself with the variables in your dataset. A Word document containing a description of the study and all the variables in the dataset is available on bCourses.

Note: The inspect command is useful for data exploration in a different way than we've seen with other commands. On the left side of the Results window, inspect draws a crude histogram of the distribution of the variable. Also, it can give you clues about problems in your data; for instance, if there were negative or missing values for a variable that should only be positive, inspect could identify this problem. Missing values are particularly important when generating new variables from pre-existing variables, since (as pointed out in past labs) a missing value will be treated as $+\infty$ in an inequality (for instance, in a replace command). This is often not the behavior desired, and so it is a good idea in that situation to check for missing values in your variable. Another way to check for missing values:

The above command only reports on variables for which there are more than 0 missing values. (? does this have an R equivalent?)

Label your data and variables using information from the description Word document linked to above.

3. Look at the distribution of the `weight0` variable in particular.

4. Weight is a suspected confounder in the relationship between two variables of interest, Behavior Pattern and Coronary Heart Disease. Generate a categorical variable for weight and divide the continuous weight variable into the following categories:

- $<$ 150 lbs
- $\geq$ 150 lbs and $<$ 160 lbs
- $\geq$ 160 lbs and $<$ 170 lbs
- $\geq$ 170 lbs and $<$ 180 lbs
- $\geq$ 180 lbs

5. Examine the odds ratio for coronary heart disease associated with behavior pattern. Take a look at the relative risk.

6. We can now examine the odds ratio and relative risk for coronary heart disease and behavior pattern for each of the weight categories defined above using the by option, for Relative Risk:

and for Odds Ratio:

By adding one more option to the " function, we can tell R to use the Woolf method for calculating weights, rather than the default Mantel-Haenszel method:

7. What do these estimates tell you qualitatively about interaction and confounding? How can you use the Mantel-Haenszel to sum up your opinions about confounding? Make sure that you can interpret the results of this test. Compare the CMH test statistic with the overall χ^2 test statistic from the unstratified analysis.

Some notes:

- You cannot get the Cochran-Mantel-Haenszel test results without using the _____.??????
- Since this lab focuses on odds ratios, which can be calculated using the same equations for cohort, population-based, and case-control studies, *you may use either the cs or the cc command* _____ regardless of your study design. The output should be identical.

2. Optional

If you'd like to explore stratification further, here are some additional questions you could explore in the Titanic dataset, `titanicdata.dta`.

Using the Titanic data from bCourses, generate a new variable died, that will take on a value of 1 if the individual didn't survive the trip, and 0 otherwise. Examine the possible confounding effects of age (a simple adult/child dummy variable) on the association between `sex` and `died` (for passengers only). What is the relative risk of death for adults? How about children? Use the Cochran-Mantel-Haenszel test for independence, to determine the evidence for death being independent of sex, controlling for the simple age variable. What kind of causal graph do you imagine in this case? Now look at the age as an exposure, and sex as a possible confounder. Is sex a confounder? What form of causal graph underlies your reasoning in this case?