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.

1.1: 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
```

1.2: 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.

We have loaded our dataset into a variable called `wcgs`, which is saved as a dataframe. Let's take a look at one of the variables in the dataset: `Ncigs`, which is the number of cigarettes smoked in the study. We'll explore using a few important statistical functions to find the distribution of the variable, as well as finding any missing values in the data.

```
summary(wcgs)
```

```
##        id             age0          height0         weight0
##  Min.   : 2001   Min.   :39.00   Min.   :60.00   Min.   : 78
##  1st Qu.: 3741   1st Qu.:42.00   1st Qu.:68.00   1st Qu.:155
##  Median :11406   Median :45.00   Median :70.00   Median :170
##  Mean   :10478   Mean   :46.28   Mean   :69.78   Mean   :170
##  3rd Qu.:13115   3rd Qu.:50.00   3rd Qu.:72.00   3rd Qu.:182
##  Max.   :22101   Max.   :59.00   Max.   :78.00   Max.   :320
##
##       sbp0            dbp0            chol0          behpat0
##  Min.   : 98.0   Min.   : 58.00   Min.   :103.0   Min.   :1.000
##  1st Qu.:120.0   1st Qu.: 76.00   1st Qu.:197.2   1st Qu.:2.000
##  Median :126.0   Median : 80.00   Median :223.0   Median :2.000
##  Mean   :128.6   Mean   : 82.02   Mean   :226.4   Mean   :2.523
##  3rd Qu.:136.0   3rd Qu.: 86.00   3rd Qu.:253.0   3rd Qu.:3.000
##  Max.   :230.0   Max.   :150.00   Max.   :645.0   Max.   :4.000
##                                   NA's   :12
##      ncigs0        dibpat0          chd69           typechd
```

```
##  Min.   : 0.0   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.: 0.0   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median : 0.0   Median :1.0000   Median :0.00000   Median :0.0000
##  Mean   :11.6   Mean   :0.5038   Mean   :0.08148   Mean   :0.1363
##  3rd Qu.:20.0   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :99.0   Max.   :1.0000   Max.   :1.00000   Max.   :3.0000
##
##     time169         arcus0
##  Min.   :  18   Min.   :0.0000
##  1st Qu.:2842   1st Qu.:0.0000
##  Median :2942   Median :0.0000
##  Mean   :2684   Mean   :0.2985
##  3rd Qu.:3037   3rd Qu.:1.0000
##  Max.   :3430   Max.   :1.0000
##                 NA's   :2
```

We now have a non-parametric summary of the data we have acquired. While most things are unimportant to us now, we want to make sure that we catch any instances of `NAs`. If this shows up anywhere in the summary, we have found `NaNs` in our dataset. 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. This is not the behavior desired, and so it is a good idea in that situation to check for missing values in your variables.

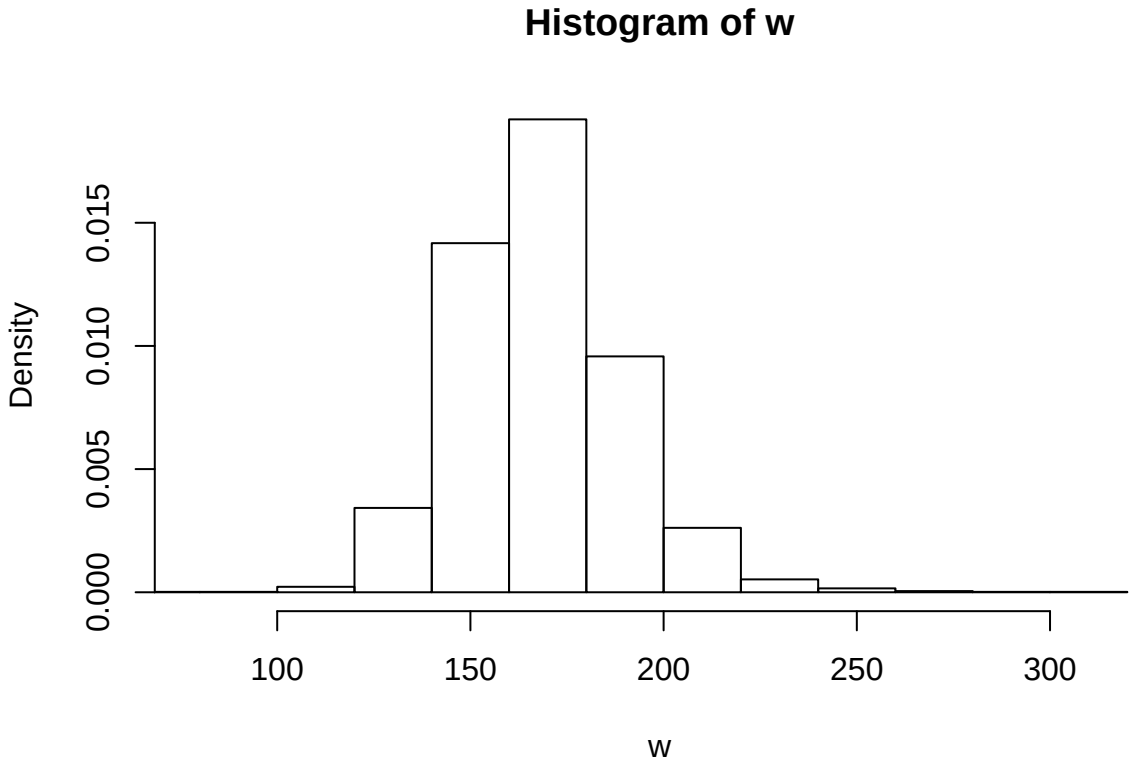
There are two variables with missing values in `wcgs`: `chol0` and `arcus0`. When doing kind of analysis or experiments, it is generally good practice to address these missing values. This can be dropping rows with missing values (`wcgs[-c(ROW IDX TO DROP)]`) or imputation (replacing with mean of the variable), which are the two most popular methods.

1.3: Look at the distribution of the `weight0` variable in particular.

```
w <- wcgs$weight0
summary(w)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      78     155     170     170     182     320
```

```
hist(w, prob=TRUE, xlim=c(min(w), max(w)))
```



The distribution for this variable looks roughly normal. We decide that weight is a possible confounding variable in determining the probability of a person having coronary heart disease (`chd69 == 1`) or not (`chd69 == 0`). To determine a range that would be statistically significant, we will turn the `weight0` column into a categorical variable.

1.4 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

The best solution to do this is to use base R's function `cut()` and label these into a new column in our dataframe. An example is shown below.

```
wcgs$weight.cats<-cut(wcgs$weight0, c(0, 150, 160, 170, 180, max(wcgs$weight0) + 1) , right=FALSE, label
summary(wcgs)
```

```
##        id             age0          height0         weight0
##  Min.   : 2001   Min.   :39.00   Min.   :60.00   Min.   : 78
##  1st Qu.: 3741   1st Qu.:42.00   1st Qu.:68.00   1st Qu.:155
##  Median :11406   Median :45.00   Median :70.00   Median :170
##  Mean   :10478   Mean   :46.28   Mean   :69.78   Mean   :170
##  3rd Qu.:13115   3rd Qu.:50.00   3rd Qu.:72.00   3rd Qu.:182
##  Max.   :22101   Max.   :59.00   Max.   :78.00   Max.   :320
##
##       sbp0            dbp0            chol0          behpat0
```

```
##  Min.   : 98.0   Min.   : 58.00   Min.   :103.0   Min.   :1.000
##  1st Qu.:120.0   1st Qu.: 76.00   1st Qu.:197.2   1st Qu.:2.000
##  Median :126.0   Median : 80.00   Median :223.0   Median :2.000
##  Mean   :128.6   Mean   : 82.02   Mean   :226.4   Mean   :2.523
##  3rd Qu.:136.0   3rd Qu.: 86.00   3rd Qu.:253.0   3rd Qu.:3.000
##  Max.   :230.0   Max.   :150.00   Max.   :645.0   Max.   :4.000
##                                   NA's   :12
##      ncigs0        dibpat0          chd69           typechd
##  Min.   : 0.0   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.: 0.0   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median : 0.0   Median :1.0000   Median :0.00000   Median :0.0000
##  Mean   :11.6   Mean   :0.5038   Mean   :0.08148   Mean   :0.1363
##  3rd Qu.:20.0   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :99.0   Max.   :1.0000   Max.   :1.00000   Max.   :3.0000
##
##     time169         arcus0       weight.cats
##  Min.   :  18   Min.   :0.0000   1:442
##  1st Qu.:2842   1st Qu.:0.0000   2:465
##  Median :2942   Median :0.0000   3:649
##  Mean   :2684   Mean   :0.2985   4:617
##  3rd Qu.:3037   3rd Qu.:1.0000   5:981
##  Max.   :3430   Max.   :1.0000
##                 NA's   :2
```

- We passed in the relevant column `wcgs$weight0`
- Specified a vector of values for the limit points `c(0, 150, 160, 170, 180, max(wcgs$weight0) + 1)`. The +1 is there because we need to include the max value, and because we are looking at the left inclusive ranges (`right=FALSE`).
- Finally, labels specify the value associated to that range `labels=c(1:5)`

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

```
chd.dibpat <- sum(wcgs$chd69 & wcgs$dibpat0)
chd.no.dibpat <- sum(wcgs$chd69 & !wcgs$dibpat0)
no.chd.dibpat <- sum(!wcgs$chd69 & wcgs$dibpat0)
no.chd.no.dibpat <- sum(!wcgs$chd69 & !wcgs$dibpat0)

matr <- matrix(c(chd.dibpat, no.chd.dibpat, chd.no.dibpat, no.chd.no.dibpat), ncol=2)
tabl <- as.table(matr)
epi.2by2(tabl)
```

```
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +          178           79        257              69.3
## Exposed -         1411         1486       2897              48.7
## Total             1589         1565       3154              50.4
##                  Odds
## Exposed +        2.25
## Exposed -        0.95
## Total            1.02
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------------
```

```
## Inc risk ratio                                  1.42 (1.30, 1.56)
## Odds ratio                                      2.37 (1.80, 3.12)
## Attrib risk *                                   20.56 (14.63, 26.48)
## Attrib risk in population *                     1.67 (-0.85, 4.20)
## Attrib fraction in exposed (%)                  29.68 (23.09, 35.71)
## Attrib fraction in population (%)               3.32 (2.28, 4.36)
## ---------------------------------------------------------------------
##  X2 test statistic: 39.898 p-value: < 0.001
##  Wald confidence limits
##  * Outcomes per 100 population units
```

This is an exercise we've now done many times! What we're more interested in seeing is whether this number for relative risk changes for different weight categories.

1.6: 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:

```
for (i in 1:5) {
  temp <- wcgs[wcgs$weight.cats == i,]
  chd.dibpat <- sum(temp$chd69 & temp$dibpat0)
  chd.no.dibpat <- sum(temp$chd69 & !temp$dibpat0)
  no.chd.dibpat <- sum(!temp$chd69 & temp$dibpat0)
  no.chd.no.dibpat <- sum(!temp$chd69 & !temp$dibpat0)

  matr <- matrix(c(chd.dibpat, no.chd.dibpat, chd.no.dibpat, no.chd.no.dibpat), ncol=2)
  tabl <- as.table(matr)
  print(epi.2by2(tabl))
}
```

```
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +           16            9         25              64.0
## Exposed -          206          211        417              49.4
## Total              222          220        442              50.2
##                   Odds
## Exposed +        1.778
## Exposed -        0.976
## Total            1.009
##
## Point estimates and 95 % CIs:
## ---------------------------------------------------------------------
## Inc risk ratio                                  1.30 (0.95, 1.77)
## Odds ratio                                      1.82 (0.79, 4.21)
## Attrib risk *                                   14.60 (-4.82, 34.02)
## Attrib risk in population *                     0.83 (-5.86, 7.52)
## Attrib fraction in exposed (%)                  22.81 (-5.20, 43.37)
## Attrib fraction in population (%)               1.64 (-0.66, 3.90)
## ---------------------------------------------------------------------
##  X2 test statistic: 2.011 p-value: 0.156
##  Wald confidence limits
##  * Outcomes per 100 population units
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +           17            8         25              68.0
## Exposed -          185          255        440              42.0
```

```
## Total              202          263        465              43.4
##                  Odds
## Exposed +       2.125
## Exposed -       0.725
## Total           0.768
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                  1.62 (1.21, 2.16)
## Odds ratio                                      2.93 (1.24, 6.93)
## Attrib risk *                                   25.95 (7.10, 44.81)
## Attrib risk in population *                     1.40 (-5.05, 7.84)
## Attrib fraction in exposed (%)                  38.17 (17.33, 53.75)
## Attrib fraction in population (%)               3.21 (0.53, 5.82)
## -------------------------------------------------------------------
##  X2 test statistic: 6.486 p-value: 0.011
##  Wald confidence limits
##  * Outcomes per 100 population units
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +           33           17         50              66.0
## Exposed -          297          302        599              49.6
## Total              330          319        649              50.8
##                  Odds
## Exposed +       1.941
## Exposed -       0.983
## Total           1.034
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                  1.33 (1.07, 1.65)
## Odds ratio                                      1.97 (1.08, 3.62)
## Attrib risk *                                   16.42 (2.69, 30.14)
## Attrib risk in population *                     1.26 (-4.29, 6.82)
## Attrib fraction in exposed (%)                  24.87 (6.88, 39.39)
## Attrib fraction in population (%)               2.49 (0.27, 4.65)
## -------------------------------------------------------------------
##  X2 test statistic: 4.977 p-value: 0.026
##  Wald confidence limits
##  * Outcomes per 100 population units
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +           42           18         60              70.0
## Exposed -          278          279        557              49.9
## Total              320          297        617              51.9
##                  Odds
## Exposed +       2.333
## Exposed -       0.996
## Total           1.077
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                  1.40 (1.17, 1.69)
## Odds ratio                                      2.34 (1.32, 4.17)
## Attrib risk *                                   20.09 (7.77, 32.41)
## Attrib risk in population *                     1.95 (-3.77, 7.68)
```

```
## Attrib fraction in exposed (%)               28.70 (14.18, 40.76)
## Attrib fraction in population (%)            3.77 (1.24, 6.23)
## ------------------------------------------------------------------
##  X2 test statistic: 8.757 p-value: 0.003
##  Wald confidence limits
##  * Outcomes per 100 population units
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +           70           27         97              72.2
## Exposed -          445          439        884              50.3
## Total              515          466        981              52.5
##                  Odds
## Exposed +        2.59
## Exposed -        1.01
## Total            1.11
##
## Point estimates and 95 % CIs:
## ------------------------------------------------------------------
## Inc risk ratio                               1.43 (1.25, 1.65)
## Odds ratio                                   2.56 (1.61, 4.06)
## Attrib risk *                                21.83 (12.32, 31.33)
## Attrib risk in population *                  2.16 (-2.38, 6.70)
## Attrib fraction in exposed (%)               30.24 (19.77, 39.35)
## Attrib fraction in population (%)            4.11 (2.12, 6.06)
## ------------------------------------------------------------------
##  X2 test statistic: 16.697 p-value: < 0.001
##  Wald confidence limits
##  * Outcomes per 100 population units
```

By adding one more option to the `epi.2by2()` function, we can tell R to use the Woolf method for calculating weights, rather than the default Mantel-Haenszel method:

```
chd.dibpat <- sum(wcgs$chd69 & wcgs$dibpat0)
chd.no.dibpat <- sum(wcgs$chd69 & !wcgs$dibpat0)
no.chd.dibpat <- sum(!wcgs$chd69 & wcgs$dibpat0)
no.chd.no.dibpat <- sum(!wcgs$chd69 & !wcgs$dibpat0)

matr <- matrix(c(chd.dibpat, no.chd.dibpat, chd.no.dibpat, no.chd.no.dibpat), ncol=2)
tabl <- as.table(matr)
epi.2by2(tabl, homogeneity ="woolf") # Woolf!
```

```
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +          178           79        257              69.3
## Exposed -         1411         1486       2897              48.7
## Total             1589         1565       3154              50.4
##                  Odds
## Exposed +        2.25
## Exposed -        0.95
## Total            1.02
##
## Point estimates and 95 % CIs:
## ------------------------------------------------------------------
## Inc risk ratio                               1.42 (1.30, 1.56)
## Odds ratio                                   2.37 (1.80, 3.12)
## Attrib risk *                                20.56 (14.63, 26.48)
## Attrib risk in population *                  1.67 (-0.85, 4.20)
```

```
## Attrib fraction in exposed (%)               29.68 (23.09, 35.71)
## Attrib fraction in population (%)            3.32 (2.28, 4.36)
## ---------------------------------------------------------------------
##  X2 test statistic: 39.898 p-value: < 0.001
##  Wald confidence limits
##  * Outcomes per 100 population units
```

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.

One Last Note:

- 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 `cohort.count` or `case.control`, 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?