Homework #4
Unit 4 – Categorical Data Analysis (1 of 2)
Practice Problems
Solutions

#1. *Source: Fisher LD and VanBelle G. Biostatistics: A Methodology for the Health Sciences New York: John Wiley, 1993. Chapter 6 Problem #12, page 234.*

This exercise reviews your understanding of odds ratios and the comparison of crude versus adjusted odds ratios. Peterson et al (1979) studied the patterns of infant deaths, in particular SIDS, in King County Washington during the years 1969-1977. They compared the SIDS deaths with a 1% sample of all births during the specified time period. Tables relating the occurrence of SIDS with maternal age less than or equal to 19 years of age, and to birth order greater than one, follow. The following are the data for singleton births.

A. Compute odds ratios and 95% confidence intervals for the four tables.

1st table

	Child	
Birth Order	**SIDS**	**Control**
> 1	201	689
=1	92	626

Answer: OR = 1.985 95% CI = (1.52, 2.599)

Solution by Hand (*1st table only*):

1. Obtain OR and ln(OR)

$$\hat{OR} = \frac{ad}{bc} = \frac{(201)(626)}{(92)(689)} = 1.9850 \quad \rightarrow \quad \ln(\hat{OR}) = \ln(1.9850) = 0.6856$$

2. Obtain var(ln[OR]) and se(ln[OR])

$$\hat{var}(\ln[\hat{OR}]) \approx \frac{1}{201} + \frac{1}{92} + \frac{1}{689} + \frac{1}{626} = .0189 \quad \rightarrow \quad \hat{se}(\ln[\hat{OR}]) = \sqrt{\hat{var}(\ln[\hat{OR}])} \approx \sqrt{.0189} = 0.1375$$

3. Obtain 95% CI for ln[OR]

$$.6856 - 1.96\sqrt{.0189} \le \ln(OR) \le .6856 + 1.96\sqrt{.0189} \quad = \quad (0.4161, 0.9551)$$

4. Exponentiate to obtain 95% CI for OR

$$(\exp[.4161], \exp[.9551]) = (1.5161, 2.5988)$$

R

```
library(DescTools)

# Table 1
tableq11 <- as.table(rbind(c(201,689),c(92,626)))        # I like entering data doing row by row using rbind()
dimnames(tableq11) <- list(
                   BIRTH_ORDER=c(">1","=1"),
                   CHILD=c("SIDS","Control"))
tableq11

##             CHILD
## BIRTH_ORDER SIDS Control
##          >1  201     689
##          =1   92     626

OddsRatio(tableq11,conf.level=0.95)

## odds ratio     lwr.ci     upr.ci
##   1.985013   1.516221   2.598748

# Table 2
tableq12 <- as.table(rbind(c(76,164),c(217,1151)))
dimnames(tableq12) <- list(
                   MATERNAL_AGE=c("<=19","> 19"),
                   CHILD=c("SIDS","Control"))


tableq12

##              CHILD
## MATERNAL_AGE SIDS Control
##         <=19   76     164
##         > 19  217    1151

OddsRatio(tableq12,conf.level=0.95)

## odds ratio     lwr.ci     upr.ci
##   2.458020   1.806011   3.345417

# Table 3
tableq13 <- as.table(rbind(c(26,17),c(267,1298)))
dimnames(tableq13) <- list(
                   EXPOSURE=c("Yes","No"),
                   CHILD=c("SIDS","Control"))
tableq13

##         CHILD
## EXPOSURE SIDS Control
##      Yes   26      17
##      No   267    1298

OddsRatio(tableq13,conf.level=0.95)

## odds ratio     lwr.ci     upr.ci
##   7.435118   3.978336  13.895502
```

```
# Table 4
tableq14 <- as.table(rbind(c(26,17),c(42,479)))
dimnames(tableq14) <- list(
                  EXPOSURE=c("Yes","No"),
                  CHILD=c("SIDS","Control"))
tableq14

##          CHILD
## EXPOSURE SIDS Control
##      Yes   26      17
##      No    42     479

OddsRatio(tableq14,conf.level=0.95)

## odds ratio     lwr.ci     upr.ci
##  17.442577   8.767214  34.702414
```

B. Which table of the last two do you think reflects best the risk of both risk factors at once? Comment. There is no single right answer here.

Answer: The 3rd table, where OR = 7.435 95% CI = (3.94, 14.05)

Remarks:
In the first table, the risk factor explored is birth order > 1.
In the second table, the risk factor explored is maternal age ≤ 19
An exploration of both risk factors at once occurs in both the 3rd and 4th tables
The advantage of the 3rd table is that the referent is the presence of either risk factor.
Thus, the estimated OR =7.45 may be interpreted as a measure of the relative odds of SIDS beyond that accompanying either birth order > 1 alone OR maternal age ≤ 19 alone.

#2. *Source: Fisher LD and VanBelle G. Biostatistics: A Methodology for the Health Sciences New York: John Wiley, 1993. Chapter 6 Problem #14, page 235.*

This exercise gives you practice performing a stratifed analysis of rates. Consider again the example given in the introduction to the notes for this unit. Researchers are investigating whether or not there is a relationship between coffee consumption and cardiovascular risk and want to take into account the potential role of a third variable, smoking. Smoking is the stratification variable. Separately in each stratum of smoking, low coffee drinkers are compared with high coffee drinkers with respect to proportion suffering a myocardial infarction (MI).

NEVER SMOKED

Cups Coffee per day	MI	Control
≥ 5	7	31
< 5	55	2691

FORMER SMOKER

Cups Coffee per day	MI	Control
≥ 5	7	18
< 5	20	112

1-14 CIGARETTES/DAY

Cups Coffee per day	MI	Control
≥ 5	7	24
< 5	33	11

15-24 CIGARETTES/DAY

Cups Coffee per day	MI	Control
≥ 5	40	45
< 5	88	172

25-34 CIGARETTES/DAY

Cups Coffee per day	MI	Control
≥ 5	34	24
< 5	50	55

35-44 CIGARETTES/DAY

Cups Coffee per day	MI	Control
≥ 5	27	24
< 5	55	58

45+ CIGARETTES/DAY

Cups Coffee per day	MI	Control
≥ 5	30	17
< 5	34	17

A. Compute the Mantel Haenszel estimate of the odds ratio.

Answer: $OR_{MH} = 1.275$

$$OR_{MH} = \frac{\sum_{\text{stratum1}}^{\text{stratum7}} a_i d_i / T_i}{\sum_{\text{stratum1}}^{\text{stratum7}} b_i c_i / T_i}$$

R

```
library(epiDisplay)

library(DescTools)
library(epiR)

# Questions #2A & 2B
# Enter K=7 2x2 tables using command array()
# NOTE!!!!  Each row is one 2x2 table that is entered column by column: a,c,b,d
# dim=c(#rows, # columns, #strata)
tableq2 <- array(c(7,55,31,2691,
                   7,20,18,112,
                   7,33,24,11,
                   40,88,45,172,
                   34,50,24,55,
                   27,55,24,58,
                   30,34,17,17),
                 dim=c(2,2,7),
                 dimnames=list(
                   COFFEE=c(">= 5 cups"," < 5 cups"),
                   MI=c("MI-case", "Control"),
                   STRATUM=c("Never Smoked", "Former", "1-14 cigs/day",
                             "15-24 cigs/day", "25-34 cigs/day", "35-44 cigs/day",
                             "45+ cigs/day")))

# List Data, table by table
tableq2
```

```
STRATUM = Never Smoked

           MI
 COFFEE     MI-case Control
   >= 5 cups       7      31
    < 5 cups      55    2691



STRATUM = Former

           MI
 COFFEE     MI-case Control
   >= 5 cups       7      18
    < 5 cups      20     112


STRATUM = 1-14 cigs/day

           MI
 COFFEE     MI-case Control
   >= 5 cups       7      24
    < 5 cups      33      11


STRATUM = 15-24 cigs/day

           MI
 COFFEE     MI-case Control
   >= 5 cups      40      45
    < 5 cups      88     172
```

```
STRATUM = 25-34 cigs/day

            MI
 COFFEE      MI-case Control
   >= 5 cups      34      24
    < 5 cups      50      55



STRATUM = 35-44 cigs/day

            MI
 COFFEE      MI-case Control
   >= 5 cups      27      24
    < 5 cups      55      58



STRATUM = 45+ cigs/day

            MI
 COFFEE      MI-case Control
   >= 5 cups      30      17
    < 5 cups      34      17
```

```
# Mantel-Haenszel Odds Ratio (95% CI limits)
mhor(mhtable=tableq2,decimal=2,graph=FALSE,design="case control")
```

```
 Stratified analysis by STRATUM
                            OR lower lim. upper lim.  P value
 STRATUM Never Smoked    11.018     3.9232     26.990 1.41e-05
 STRATUM Former           2.165     0.6752      6.363 1.47e-01
 STRATUM 1-14 cigs/day    0.101     0.0281      0.321 1.47e-05
 STRATUM 15-24 cigs/day   1.735     1.0222      2.941 3.81e-02
 STRATUM 25-34 cigs/day   1.554     0.7767      3.144 1.94e-01
 STRATUM 35-44 cigs/day   1.185     0.5805      2.430 7.36e-01
 STRATUM 45+ cigs/day     0.883     0.3534      2.205 8.33e-01
 M-H combined             1.275     0.9727      1.670 6.46e-02

 M-H Chi2(1) = 3.42 , P value = 0.065
 Homogeneity test, chi-squared 6 d.f. = 48.76 , P value = 0
```

B. Compute the appropriate chi square test for association.

R

Answer: This test is only appropriate if we can assume homogeneity of OR in all 7 strata
But note!!! The test of homogeneity actually says NO!! (Chi Square statistic$_{df=6}$ = 48.76, p-value <<< .0001).
For illustration purposes, we'll forge on ….
From part "A", we have OR$_{MANTEL\text{-}HAENSZEL}$ = 1.275 (95% CI: 0.97, 1.67)
This is marginally statistically significantly different from the null value of 1 (p-value = .065).

```
# Mantel-Haenszel Odds Ratio (95% CI limits)
mhor(mhtable=tableq2,decimal=2,graph=FALSE,design="case control")


 Stratified analysis by STRATUM
                          OR lower lim. upper lim.  P value
 STRATUM Never Smoked    11.018     3.9232     26.990 1.41e-05
 STRATUM Former           2.165     0.6752      6.363 1.47e-01
 STRATUM 1-14 cigs/day    0.101     0.0281      0.321 1.47e-05
 STRATUM 15-24 cigs/day   1.735     1.0222      2.941 3.81e-02
 STRATUM 25-34 cigs/day   1.554     0.7767      3.144 1.94e-01
 STRATUM 35-44 cigs/day   1.185     0.5805      2.430 7.36e-01
 STRATUM 45+ cigs/day     0.883     0.3534      2.205 8.33e-01
 M-H combined             1.275     0.9727      1.670 6.46e-02

 M-H Chi2(1) = 3.42 , P value = 0.065
 Homogeneity test, chi-squared 6 d.f. = 48.76 , P value = 0
```

Supplement (NOT part of your homework assignment; NOTHING to turn in)

Learn R

Practice with variable types, factors, and creating a contingency table

Practice #1 - Create objects of various types using functions c( ) and factor( ). From these, create a dataframe.

1a. Create a character variable object called **`name.`** Show.

```
# character variable values must be in single or double quotes
name <- c("Piper", "Chyke","Kelsey","Anand","Shirin","Nina","Serena","Isaac","Paige",
          "Daria")
name
##  [1] "Piper"  "Chyke"  "Kelsey" "Anand"  "Shirin" "Nina"   "Serena" "Isaac"
##  [9] "Paige"  "Daria"
```

1b. Create a character variable object called **`cilantro`** that contains missing values. Show.

```
# Missing is NA with NO QUOTES
cilantro <- c("love", NA, "hate", "love", "love", "love", "hate","hate","love", "love")
cilantro                                                    # character data prints in quotes
##  [1] "love" NA     "hate" "love" "love" "love" "hate" "hate" "love" "love"
```

1c. Create a factor variable object called **`coffee`**. Show.

```
# factor variable requires using factor(c())
coffee <- factor(c("small", "small", "medium", "small", "small", "large", "large", "medium", "large", "large"
))
coffee                                                       # factor data prints WITHOUT quotes
##  [1] small  small  medium small  small  large  large  medium large  large
## Levels: large medium small
```

1d. Create an integer variable object called **`dentist`**. Show.

```
# integer variable values require L
dentist <- c(1L,2L,2L,2L,1L,1L,NA,1L,4L,1L)
dentist
##  [1]  1  2  2  2  1  1 NA  1  4  1
```

1e. Create a numeric variable object called **`mvpa`** that includes missing values. Show.

```
# missing numeric variable value is NA
mvpa <- c(30.2,89.3,57.4,45.8,NA,126.9,190.5,64.2,NA,120.0)
mvpa
##  [1]  30.2  89.3  57.4  45.8    NA 126.9 190.5  64.2    NA 120.0
```

1f. Create a dataframe called **`mydataframe`** that binds together the objects created in #1a - #1e. Show.

```
# create dataframe using data.frame()
mydataframe <- data.frame(name,cilantro,coffee,dentist,mvpa)
mydataframe
##      name cilantro coffee dentist  mvpa
## 1   Piper     love  small       1  30.2
## 2   Chyke     <NA>  small       2  89.3
## 3  Kelsey     hate medium       2  57.4
## 4   Anand     love  small       2  45.8
## 5  Shirin     love  small       1    NA
## 6    Nina     love  large       1 126.9
## 7  Serena     hate  large      NA 190.5
## 8   Isaac     hate medium       1  64.2
## 9   Paige     love  large       4    NA
## 10  Daria     love  large       1 120.0
```

1g. Examine the structure of the dataframe you just created

```
str(mydataframe)                                                    # Entire dataframe
## 'data.frame':    10 obs. of  5 variables:
##  $ name    : chr  "Piper" "Chyke" "Kelsey" "Anand" ...
##  $ cilantro: chr  "love" NA "hate" "love" ...
##  $ coffee  : Factor w/ 3 levels "large","medium",..: 3 3 2 3 3 1 1 2 1 1
##  $ dentist : int  1 2 2 2 1 1 NA 1 4 1
##  $ mvpa    : num  30.2 89.3 57.4 45.8 NA ...

str(mydataframe$name)                                               # Single variable in dataframe
##  chr [1:10] "Piper" "Chyke" "Kelsey" "Anand" "Shirin" "Nina" "Serena" ...
```

Practice #2 - Quick and easy: Produce descriptvies on every variable in your dataframe.

```
# Using summary( ) in package {base}, no installation required.
# Note - NO descriptives are produced for character variable objects
summary(mydataframe)
##      name            cilantro           coffee     dentist
##  Length:10          Length:10          large :4   Min.   :1.000
##  Class :character   Class :character   medium:2   1st Qu.:1.000
##  Mode  :character   Mode  :character   small :4   Median :1.000
##                                                   Mean   :1.667
##                                                   3rd Qu.:2.000
##                                                   Max.   :4.000
##                                                   NA's   :1
##       mvpa
##  Min.   : 30.20
##  1st Qu.: 54.50
##  Median : 76.75
##  Mean   : 90.54
##  3rd Qu.:121.72
##  Max.   :190.50
##  NA's   :2
```

Practice #3 - Working with Factors in R

NOTE 1: For analyzing categorical data, R requires that your categorical variables be of type FACTOR
NOTE 2: By default, R stores factor value levels alphabetically.

3a. Convert from character to factor. Show.

```
cilantrof <- factor(cilantro)
cilantrof
##  [1] love <NA> hate love love love hate hate love love
## Levels: hate love
```

3b. Don't like alphabetic storage? **TIP - Always set factor levels explicitly.** Check.

```
coffee <- factor(coffee,
                 levels=c("small", "medium", "large"))
attributes(coffee)
## $levels
## [1] "small"  "medium" "large"
##
## $class
## [1] "factor"
```

3c. Set factor levels explicitly and declare as ORDERED. Check.

```
coffee <- factor(coffee,
                 levels=c("small", "medium", "large"),
                 ordered=TRUE)

attributes(coffee)
## $levels
## [1] "small"  "medium" "large"
##
## $class
## [1] "ordered" "factor"
```

Practice #4 - Direct entry of a 2x2 Table

4a. Create by direct entry a 2x2 table called **`mytable`**, ROW by ROW**.** Show.

```
# Row by row (a,b,c,d)
mytable <- as.table(rbind(c(59,48),c(11,462)))
mytable
##     A   B
## A  59  48
## B  11 462
```

4b. Make it readable! Label row variable, column variable, row variable values and column variable values. Show.

```
# dimnames( ) labels row variable first and column variable second.
dimnames(mytable) <- list(
                     TEST=c("Positive","Negative"),            # ROW_VAR = c("value1", "value2")
                     DISEASE=c("Diabetes","Non_diabetes"))     # COLUMN_VAR = c("value1", "value2")
mytable
##           DISEASE
## TEST       Diabetes Non_diabetes
##   Positive       59           48
##   Negative       11          462
```