Unit 4 – Categorical Data Analysis
Practice Problems (1 of 2)
Due Thursday October 3, 2024
Last Date to submit Late for credit (-20 points): Thursday October 10, 2024

#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.

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

	Child	
Maternal Age	**SIDS**	**Control**
≤ 19	76	164
> 19	217	1151

	Child	
	SIDS	**Control**
Birth order > 1 and Maternal Age ≤ 19	26	17
Birth order = 1 **OR** Maternal Age > 19	267	1298

	Child	
	SIDS	**Control**
Birth order > 1 and Maternal Age ≤ 19	26	17
Birth order = 1 **AND** Maternal Age > 19	42	479

a. Compute the estimated odds ratio measures of association and 95% confidence intervals for the four tables.

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.

#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). The following data were obtained.

Stratum = NEVER SMOKED

Cups Coffee per day (0/1)	**MI** (micase=1)	**Control** (micase=0)
1 = ≥ 5	7	31
0 = < 5	55	2691

Stratum = FORMER SMOKER

Cups Coffee per day	**MI** (micase=1)	**Control** (micase=0)
≥ 5	7	18
< 5	20	112

Stratum = 1-14 CIGARETTES/DAY

Cups Coffee per day	**MI** (micase=1)	**Control** (micase=0)
≥ 5	7	24
< 5	33	11

Stratum = 15-24 CIGARETTES/DAY

Cups Coffee per day	**MI** (micase=1)	**Control** (micase=0)
≥ 5	40	45
< 5	88	172

Stratum = 25-34 CIGARETTES/DAY

Cups Coffee per day	**MI** (micase=1)	**Control** (micase=0)
≥ 5	34	24
< 5	50	55

Stratum = 35-44 CIGARETTES/DAY

Cups Coffee per day	**MI** (micase=1)	**Control** (micase=0)
≥ 5	27	24
< 5	55	58

	Stratum = 45+ CIGARETTES/DAY (micase=0)	
Cups Coffee per day	**MI** (micase=1)	**Control**
≥ 5	30	17
< 5	34	17

A. Assume there is no effect modification. Under this assumption, the true stratum-specific associations are all equal to a single common association of coffee consumption with MI. Given this assumption, compute the Mantel Haenszel estimate of the common odds ratio.

B. Compute the appropriate chi square test for association.

C. In 1-2 sentences, interpret your findings to a client who is not an expert in biostatistics.

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
```