Unit 4 – Categorical Data Analysis
Practice Problems

SOLUTIONS – *R Users*

#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.*

1A. 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]

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

4. Exponentiate to obtain 95% CI for OR

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

Solution using 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.*

It is of interest to study the association between coffee consumption and myocardial infarction and it is suspected that smoking confounds this relationship.

The following stratified 2x2 table data are available.

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: $OR_{MH} = 1.2746$

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

Solution using R

```
library(epiDisplay)

library(DescTools)
library(epiR)

# Questions #2A & 2B
# Enter K=7 2x2 tables using command array()
# NOTE!!!!  Each row is one 2x2, 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.

Answer: The test for homogeneity of proportions is significant (Chi square on df=6 is 48.76; p-value < .0001), suggesting that estimation of a coffee-MI association is different depending on level of cigarette smoking.

For completeness, the test of departure of the Mantel Haenszel OR from the null value of unity is marginally significant (Chi square on df = 1 is 3.42; p-value = .06)

The appropriate analysis would assess, first, the homogeneity of the stratum specific odds ratios. Looking at the stratum-specific estimated odds ratios in the STATA output above suggests that they are not equal. If true, the appropriate analysis concludes that the association between myocardial infarction and coffee consumption is modified by cigarette smoking.

Solution using R

```
# 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
```

#3. *Source: Fisher LD and VanBelle G. Biostatistics: A Methodology for the Health Sciences New York: John Wiley, 1993. Chapter 6 Problem #15, page 236.*

The paper of Remein and Wilkerson (1961) considers screening tests for diabetes. The Somogyi-Nelson (venous) blood test (data at one hour after a test meal and using 130 mg/100 ml as the blood sugar cutoff gives the following table:

Test	Diabetic	Non-diabetic	Total
Positive	59	48	107
Negative	11	462	473
Total	70	510	580

A. Compute the sensitivity, specificity, predictive value of a positive test and predictive value of a negative test.

Answer: **sensitivity = 0.843** **specificity = 0.906**
predictive value positive = 0.55 **predictive value negative = 0.98**

Solution by Hand:

$$\text{Sensitivity}=\frac{\#\text{ diabetics who test +}}{\text{Total \# diabetics}}=\frac{(59)}{(70)}=0.8429$$

$$\text{Specificity}=\frac{\#\text{ NON diabetics who test -}}{\text{Total \# NON diabetics}}=\frac{(462)}{(510)}=0.9059$$

$$\text{Predictive Value Positive}=\frac{\#\text{ Positive who have diabetes}}{\text{Total \# Positives}}=\frac{(59)}{(107)}=0.5514$$

$$\text{Predictive Value Negative}=\frac{\#\text{ Negatives who do NOT have diabetes}}{\text{Total \# Negatives}}=\frac{(462)}{(473)}=0.9767$$

Solution using R

```
# Question 3A


tableq3 <- as.table(rbind(c(59,48),c(11,462)))
dimnames(tableq3) <- list(
                  TEST=c("Positive","Negative"),
                  DISEASE=c("Diabetes","Non_diabetes"))
tableq3

##           DISEASE
## TEST       Diabetes Non_diabetes
##   Positive       59           48
##   Negative       11          462

# Question 3 - Sensitivity, Specificity, Predictive +, Predictive -
library(epiR)
epi.tests(tableq3)

##           Outcome +    Outcome -      Total
## Test +           59           48        107
## Test -           11          462        473
## Total            70          510        580
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------
## Apparent prevalence                   0.18 (0.15, 0.22)
## True prevalence                       0.12 (0.10, 0.15)
## Sensitivity                           0.84 (0.74, 0.92)
## Specificity                           0.91 (0.88, 0.93)
## Positive predictive value             0.55 (0.45, 0.65)
## Negative predictive value             0.98 (0.96, 0.99)
## Positive likelihood ratio             8.96 (6.72, 11.94)
## Negative likelihood ratio             0.17 (0.10, 0.30)
## -------------------------------------------------------------
```

B. ***OPTIONAL -***

Using the sensitivity and specificity values you got in part "A", consider the following various values of prevalence: .01, .05, .10, .20, .30, .40, .50, .60, .70, .80, .90, .95. Using these prevalence values, construct a plot of the predictive value of a positive test on the vertical axis versus prevalence on the horizontal axis. What you will have constructed is a plot of the probability of diabetes given a positive test result as a function of prevalence.

The solution is an application of Bayes Theorem.

Let

- D = Event of diabetes
- + = Event of positive test

What we want to calculate is Probability (D | +)

What we have as available information is

- Probability (+ | D) = **sensitivity** = .843
  Probability (+ | not D) = **1 – specificity** = .094
- Probability (D) = **prevalence** = .01, .05, …., .95
  Probability (not D) = **1 – prevalence** = .99, .95, …., .05

$$\Pr(D \mid +) = \frac{\Pr(D \text{ and } +)}{\Pr(+)}$$ by definition of conditional probability.

$$= \frac{\Pr(+ \mid D)\Pr(D)}{\Pr(+)}$$ by definition of conditional probability the other way.

$$= \frac{\Pr(+ \mid D)\Pr(D)}{\Pr(+ \mid D)\Pr(D) + \Pr(+ \mid \text{not } D)\Pr(\text{not } D)}$$ by theorem of total probabilities.

$$= \frac{[\text{Sensitivity}]\,(\text{Prevalence})}{[\text{Sensitivity}]\,(\text{Prevalence}) + [1 - \text{specificity}]\,(1 - \text{prevalence})}$$

$$= \frac{[.843]\,(\text{Prevalence})}{[.843]\,(\text{Prevalence}) + [.094]\,(1 - \text{prevalence})}$$

Solution using R

```
# FROM SCRATCH - Create data set named q3b with X=prevalence, Y=predictive value + values for plotting
# Input x
x <-c(.01,.05,.10,.20,.30,.40,.50,.60,.70,.80,.90,.95,.99)
q3b <- data.frame(x)
# Calculate y = predictive value + = (.843*prevalence)  /  [ (.843*prevalence) + (.094*(1-prevalence)) ]
q3b$y <- 0.843*q3b$x/((0.843*q3b$x)+(0.094*(1-q3b$x)))

# list
q3b

       x         y
 1  0.01 0.08306237
 2  0.05 0.32065424
 3  0.10 0.49911190
 4  0.20 0.69155045
 5  0.30 0.79353624
 6  0.40 0.85670732
 7  0.50 0.89967983
 8  0.60 0.93080604
 9  0.70 0.95439107
 10 0.80 0.97287940
 11 0.90 0.98776201
 12 0.95 0.99416548
 13 0.99 0.99887494

# plot(xvariable,yvariable,OPTION, OPTION, OPTION)
# The option type="b" will connect the points with segmented lines
plot(q3b$x,q3b$y,
     main="Pr[Diabetes] given Positive Test",
     xlab="Prevalence",
     ylab="Pr[diabetes | + test]",
     type="b")
```