Unit 4 – Categorical Data Analysis
Practice Problems

SOLUTIONS – *Stata 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) Stata 95% CI = (1.51, 2.60)

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]

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

4. Exponentiate to obtain 95% CI for OR

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

2nd table

	Child	
Maternal Age	**SIDS**	**Control**
≤ 19	76	164
> 19	217	1151
Answer: OR = 2.458	**95% CI = (1.80, 3.36)**	

3rd table

	Child	
	SIDS	**Control**
Birth order > 1 and Maternal Age ≤ 19	26	17
Birth order = 1 **OR** Maternal Age > 19	267	1298
Answer: OR = 7.435	**95% CI = (3.94, 14.05)**	

4th table

	Child	
	SIDS	**Control**
Birth order > 1 and Maternal Age ≤ 19	26	17
Birth order = 1 **AND** Maternal Age > 19	42	479
Answer: OR = 17.443	**95% CI = (8.14, 37.35)**	

Solution using STATA

The following assumes you have launched Stata and started a log of your session!

```
. set more off

. *
. ******  1a)  table #1
. generate birth=.
. generate sids=.
. generate tally=.

. *---- Click on data editor icon to enter data. Then close data editor window.  ---*
. label define birthf 0 ">1" 1 "=1"
. label define sidsf 0 "SIDS" 1 "Control"
. label values birth birthf
. label values sids sidsf
. tab2 birth sids [freq=tally]

-> tabulation of birth by sids

           |         sids
     birth |      SIDS    Control |     Total
-----------+----------------------+----------
        >1 |       201        689 |       890
        =1 |        92        626 |       718
-----------+----------------------+----------
     Total |       293      1,315 |     1,608


. tabodds birth sids [freq=tally], or

------------------------------------------------------------------------------
        sids |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SIDS |    1.000000          .           .               .           .
     Control |    1.985013      25.45       0.0000       1.512824    2.604583
------------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(1)  =    25.45
                                  Pr>chi2  =   0.0000

Score test for trend of odds:     chi2(1)  =    25.45
                                  Pr>chi2  =   0.0000
```

```
. *
. ******  1a)  table #2
. clear
. generate age=.
. generate sids=.
. generate tally=.

. *---- click on data editor icon to enter data.  Then close data editor window.  ---*
. label define agef 0 "<=19" 1 ">19"
. label define sidsf 0 "SIDS" 1 "Control"
. label values age agef
. label values sids sidsf
. tab2 age sids [freq=tally]

-> tabulation of age by sids

           |         sids
       age |      SIDS    Control |     Total
-----------+----------------------+----------
      <=19 |        76        164 |       240
       >19 |       217      1,151 |     1,368
-----------+----------------------+----------
     Total |       293      1,315 |     1,608


. tabodds age sids [freq=tally], or

------------------------------------------------------------------------------
        sids |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SIDS |    1.000000          .          .               .           .
     Control |    2.458020      34.20       0.0000      1.799826    3.356913
------------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(1)   =    34.20
                                  Pr>chi2   =   0.0000

Score test for trend of odds:     chi2(1)   =    34.20
                                  Pr>chi2   =   0.0000
```

```
. *
. ******  1a)  table #3
. clear
. generate birthage=.
. generate sids=.
. generate tally=.

. *---- click on data editor icon to enter data ---*
. label define birthagef 0 "order>1 AND age <=19" 1 "order=1 OR age>19"
. label define sidsf 0 "SIDS" 1 "Control"
. label values birthage birthagef
. label values sids sidsf
. tab2 birthage sids [freq=tally]

-> tabulation of birthage by sids

                    |         sids
          birthage |      SIDS    Control |     Total
--------------------+----------------------+----------
order>1 AND age <=19 |        26         17 |        43
   order=1 OR age>19 |       267      1,298 |     1,565
--------------------+----------------------+----------
              Total |       293      1,315 |     1,608


. tabodds birthage sids [freq=tally], or

------------------------------------------------------------------------------
        sids |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SIDS |    1.000000          .          .            .           .
     Control |    7.435118      52.88       0.0000      3.935808   14.045648
------------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(1)   =    52.88
                                  Pr>chi2   =   0.0000

Score test for trend of odds:     chi2(1)   =    52.88
                                  Pr>chi2   =   0.0000
```

```
. *
. ******  1a)  table #4
. clear
. generate birthage=.
. generate sids=.
. generate tally=.


. *---- click on data editor icon to enter data.  Then close data editor window.  ---*
. label define birthf 0 "order>1 AND age <=19" 1 "order=1 AND age> 19"
. label define sidsf 0 "SIDS" 1 "Control"
. label values birthage birthf
. label values sids sidsf
. tab2 birthage sids [freq=tally]

-> tabulation of birthage by sids

                    |         sids
           birthage |      SIDS    Control |     Total
--------------------+----------------------+----------
order>1 AND age <=19 |        26         17 |        43
 order=1 AND age> 19 |        42        479 |       521
--------------------+----------------------+----------
              Total |        68        496 |       564


. tabodds birthage sids [freq=tally], or

------------------------------------------------------------------------------
        sids |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SIDS |    1.000000          .          .               .           .
     Control |   17.442577     102.69       0.0000      8.145526   37.350994
------------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(1)   =   102.69
                                  Pr>chi2   =   0.0000

Score test for trend of odds:     chi2(1)   =   102.69
                                  Pr>chi2   =   0.0000
```

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: $OR_{MH} = 1.2746$

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

By Hand Using Excel

Worksheet for problem 2A

a	b	c	d	T	ad/T	bc/T
7	31	55	2691	2784	6.766164	0.612428
7	18	20	112	157	4.993631	2.292994
7	24	33	11	75	1.026667	10.56
40	45	88	172	345	19.94203	11.47826
34	24	50	55	163	11.47239	7.361963
27	24	55	58	164	9.54878	8.04878
30	17	34	17	98	5.204082	5.897959
				Totals:	58.95374	46.25239

OR(MH) = 1.2746

Solution using STATA

Answer: $OR_{MH} = 1.2746$

The following assumes you have downloaded hw_categorical.dta, launched Stata and used FILE > OPEN to read in the data.

```
. set more off

. *
. *****  2a & b)
. sort cigs
. mhodds micase coffee [freq=tally], by(cigs)

Maximum likelihood estimate of the odds ratio
Comparing coffee==1 vs. coffee==0
by cigs

------------------------------------------------------------------------------
     cigs | Odds Ratio        chi2(1)         P>chi2       [95% Conf. Interval]
----------+-------------------------------------------------------------------
  1=never |  11.048094          46.38         0.0000         4.62874   26.37008
 2=former |   2.177778           2.42         0.1197         0.79694    5.95115
 3=1-14 c |   0.097222          19.81         0.0000         0.02717    0.34791
  4=15-24 |   1.737374           4.78         0.0288         1.05209    2.86903
  5=25-34 |   1.558333           1.80         0.1798         0.81068    2.99550
  6=35-44 |   1.186364           0.25         0.6139         0.61031    2.30612
 7=45+ ci |   0.882353           0.09         0.7693         0.38203    2.03793
------------------------------------------------------------------------------

  Mantel-Haenszel estimate controlling for cigs
    ------------------------------------------------------------------
     Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
    ------------------------------------------------------------------
       1.274610      3.42         0.0646         0.984792   1.649719
    ------------------------------------------------------------------

Test of homogeneity of ORs (approx): chi2(6)    =   64.62
                                     Pr>chi2    =  0.0000
```

B. Compute the appropriate chi square test for association.

Answer: The test for homogeneity of proportions is significant (Chi square on df=6 is 64.62; 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 STATA

```
. *
. *****  2a & b) - It's the same set of commands that you executed for #2a.. sort cigs
. mhodds micase coffee [freq=tally], by(cigs)

   Mantel-Haenszel estimate controlling for cigs
    ------------------------------------------------------------------
     Odds Ratio    chi2(1)        P>chi2        [95% Conf. Interval]
    ------------------------------------------------------------------
       1.274610       3.42        0.0646         0.984792   1.649719
    ------------------------------------------------------------------

Test of homogeneity of ORs (approx): chi2(6)    =    64.62
                                     Pr>chi2    =   0.0000
```

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

Solution:
Examination of the data stratum by stratum suggests that, among former or current smokers, there is little evidence of a coffee consumption – heart attack relationship. Among the other strata, only one 95% CI estimate of the odds ratio excludes unity and the lower 95% CI limit here is 1.05; this is the 15-24 cigarettes per day stratum.

Among the never smokers, the data suggest a positive association, with 5+ cups/day coffee drinkers having an estimated relative odds of MI that is approximately 11 times greater than the lesser coffee drinkers. However, this is an imprecise estimate as the associated 95% CI is very wide (4.5 to 25.7); this is because 96.7% (2691) drink less than 5 cups of coffee per day and did not experience an MI.

#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:

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

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

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

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

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 STATA

```
. *
. ***** 3b) Plot of Prob[diabetes | positive test] as a function of prevalence.
. generate sensitivity=.
. generate specificity=.
. generate prevalenc=.

. *  Again, click on the data editor to enter data.  Then close the window.

. generate x=prevalenc
. generate numerator=sensitivity*prevalenc
. generate denominator=(sensitivity*prevalenc) + ( (1-specificity)*(1-prevalenc) )
. generate y=numerator/denominator

. label variable x "Prevalence"
. label variable y "Pr[diabetes | + test]"


. list x y, clean

        x          y
  1.   .01   .0830624
  2.   .05   .3206543
  3.    .1    .499112
  4.    .2   .6915506
  5.    .3   .7935363
  6.    .4   .8567074
  7.    .5   .8996798
  8.    .6    .930806
  9.    .7   .9543911
 10.    .8   .9728795
 11.    .9    .987762
 12.   .95   .9941655
 13.   .99   .9988749
```

```
. graph twoway (connected y x, symbol(d)), title("Probability Diabetes Given + Test")
xlabel(0(.2)1) ylabel(0(.2)1) caption("predictedpos.png", size(vsmall))

. save "/Users/carolbigelow/Desktop/hw4_q3b.dta"
file /Users/cbigelow/Desktop/hw4_q3b.dta saved
```