R for Categorical Data Analysis

Steele H. Valenzuela

March 11, 2015

Illustrations for Categorical Data Analysis

March 2015

Single 2X2 table

1. Introduction to Example

Example 1 Example 1 is used in Section 1.1 There is not an actual data set. Instead, you enter counts as part of the commands you issue.

Source: Fisher LD and Van Belle G. Biostatistics: A Methodology for the Health Sciences. New York: Wiley, 1993. Chapter 6, problem 5, page 232. Smith, Delgado, and Rutledge (1976) report data on ovarian carcinoma. Individuals had different numbers of courses of chemotherapy. The 5-year survival data for those with 1-4 and 10 or more courses of chemotherapy are shown below.

```
# Here's the best I could do with one command. It does the trick.
chemo <- matrix(c(21, 2, 2, 8), ncol = 2, byrow = TRUE,
                dimnames = list(Courses = c("1-4", ">= 10"),
                                "Five Year Status" = c("Dead", "Alive"))) # we have quotes around Five

chemo
```

```
##        Five Year Status
## Courses Dead Alive
##   1-4     21     2
##   >= 10    2     8
```

Do these data provide statistically significant evidence of an assocation of five-year survival with number of courses of chemotherapy?

Example 2 Example 2 is used in Section 1.2

The data set **single2x2.dta** contains the following 2x2 table of counts.

```
lung <- matrix(c(9, 31, 40, 2, 47, 49, 11, 78, 89), ncol = 3, byrow = TRUE,
               dimnames = list("Exposure (Smoking)" = c("Yes", "No", "Total"),
                               "Disease (Lung Cancer)" = c("Yes", "No", "Total")))
lung
```

```
##                     Disease (Lung Cancer)
## Exposure (Smoking) Yes No Total
##              Yes     9 31    40
##              No      2 47    49
##              Total  11 78    89
```

1a. Tests of Association Good to know. Sometimes, you want to be able to do a quick analysis of count data in a table and you want to, simply, type in the cell counts (instead of taking the time to create a R data set). For R, rather than type 5 lines of code, we simply insert the matrix into the function. It's nothing fancy, just the usual.

Semi-esoteric note: If you've been catching onto the syntax of R, we see that it is very object oriented. We create our object, the matrix, and put it inside of the function fisher.test(...). The immediate difference here between R and Stata is that we can name the matrix we created into an object. In the previous problems, I created the matrices and named them *chemo* and *lung*.

For Stata, we have a function, numbers, followed by an option. Stata is much more free in this sense but in R, we are allowed to carry around our object to the next line, the next phase, or however one imagines it.

```
fisher.test(matrix(c(21, 2, 2, 8), ncol = 2, byrow = TRUE))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  matrix(c(21, 2, 2, 8), ncol = 2, byrow = TRUE)
## p-value = 0.0001255
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    3.82 571.25
## sample estimates:
## odds ratio
##      34.05
```

Now, let's compute the likelihood ratio chi-square test. R does not have the function built into the program, but the source() function allows us to find a similar script online.

```
# took me over an hour to find but hey, it's here!
# the author of the function has named it g.test instead of LR chi-square test
library(RCurl)
```

```
## Loading required package: bitops
```

```
url2 <- getURL(url = "https://raw.githubusercontent.com/shv38339/practiSe/master/LRchisqtest.R",
               ssl.verifypeer = FALSE)
source(file = textConnection(url2))

g.test(matrix(c(21, 2, 2, 8), ncol = 2, byrow = TRUE), correct = "none")
```

```
##
##  Log likelihood ratio (G-test) test of independence without
##  correction
##
## data:  matrix(c(21, 2, 2, 8), ncol = 2, byrow = TRUE)
## Log likelihood ratio statistic (G) = 16.89, X-squared df = 1,
## p-value = 3.968e-05
```

1b. (More) Tests of Association To echo the above sentiment concerning objects and datasets and difference between R and Stata, there are no special commands, but rather different methods in that one may create an object. Let's continue and you'll be able to create a matrix in no time (the syntax is a bit tricky).

Let's load the Stata data set into R.

```
library(readstata13) # remember to install this package first
url2 <- "http://people.umass.edu/biep640w/datasets/single2x2.dta"
dat <- read.dta13(file = url)
```

1. For small to moderate sample sizes, use the function `fisher.test()` to obtain a Fisher Exact Test.
2. If the cell sizes are too small, R will allow the function `chisq.test()` to obtain a Pearson Chi-Square Test, but not without the following warning:

   Warning message: In chisq.test(object) : Chi-squared approximation may be incorrect

3. R's base functions do not include a likelihood ratio chi-square test, but the function is sourced from a website in the first example (one of the disadvantages of having an open source software is that sometimes you don't have everything and you have to trust your sources).

Let us continue and compute the Fisher's Exact Test on the data set we just uploaded.

```
table(dat$smoking, dat$lungca) # this is what we're going to work with
```

```
##
##               0  1
##   Non-smoker 47  2
##   Smoker     31  9
```

```
# alternative will specify a one-sided or two-sided p-value
fisher.test(table(dat$smoking, dat$lungca), alternative = c("two.sided"))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(dat$smoking, dat$lungca)
## p-value = 0.0108
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   1.264 67.751
## sample estimates:
## odds ratio
##      6.683
```

```
fisher.test(table(dat$smoking, dat$lungca), alternative = c("greater"))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(dat$smoking, dat$lungca)
## p-value = 0.01001
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  1.548   Inf
## sample estimates:
## odds ratio
##      6.683
```

Now let's compute the Likelihood Ratio Chi-Square Test on the same data set.

```
tbl1b <- table(dat$smoking, dat$lungca)
g.test(tbl1b, correct = "none")
```

```
##
##  Log likelihood ratio (G-test) test of independence without
##  correction
##
## data:  tbl1b
## Log likelihood ratio statistic (G) = 7.212, X-squared df = 1,
## p-value = 0.007242
```

1c. Cohort Design Luckily, `R` has a package, `epicalc`, to mirror `Stata`'s **cs** and **cc** commands. Here, `cs` represent a cohort study design and provides the Risk Ratio plus other tidbits of information.

```
# ?cs for more information...I'll admit, the helpfile is daunting so beware.
library(epicalc)
```

```
## Loading required package: foreign
## Loading required package: survival
##
## Attaching package: 'survival'
##
## The following object is masked _by_ '.GlobalEnv':
##
##     lung
##
## Loading required package: MASS
## Loading required package: nnet
```

```
cs(outcome = dat$lungca, exposure = dat$smoking)
```

```
##
##            Exposure
## Outcome    Non-exposed Exposed Total
##   Negative 47          31      78
##   Positive 2           9       11
##   Total    49          40      89
##
##            Rne         Re      Rt
##   Risk     0.04        0.22    0.12
##
##                                           Estimate Lower95ci Upper95ci
##  Risk difference (attributable risk)     0.18     0.04      0.32
##  Risk ratio                              5.51     1.29      23.55
##  Attr. frac. exp. -- (Re-Rne)/Re         0.82
##  Attr. frac. pop. -- (Rt-Rne)/Rt*100 %   66.98
##  Number needed to harm (NNH)             5.43     3.11      26.13
##    or 1/(risk difference)
```

You'll notice that although this package mirrors `Stata`'s commands, the table is also mirred or backwards. BUT, if you know your outcome and exposure, everything works out. Although I don't know exactly what's going on under the hood of the **Risk Ratio** and the confidence interval, it differs from Stata. Additionally, the p-value is not provided in this function.

1d. Case-control Design Similar to the `cs` function, `cc` function is a case-control study design and provides the odds ratio. If we are using Fisher's Exact Test method, it is specified with the option *fisher.or = TRUE.*

```
cc(outcome = dat$lungca, exposure = dat$smoking, fisher.or = TRUE)
```

```
##
##            dat$smoking
## dat$lungca Non-smoker Smoker Total
##      0             47     31    78
##      1              2      9    11
##      Total         49     40    89
##
## OR =  6.68
## Exact 95% CI =  1.26, 67.75
## Chi-squared = 6.9, 1 d.f., P value = 0.009
## Fisher's exact test (2-sided) P value = 0.011
```

You'll notice that the information not provided is *Proportion Exposed*, a one-sided p-value, and a few other things that were in the previous function.

Stratified Analysis of k-tables

1. Introduction to Example Note: this is a subset of the data used in the Unit 4 (Categorical Data Analysis) practice problems.

Source: Fisher LD and Van Belle G. Biostatistics: A Methodology for the Health Sciences. New York: Wiley, 1993. Chapter 6, problem 14, page 235. Rosenburg et al (1980) performed a retrospective study of the association of coffee drinking (exposure) and the occurrence of myocardial infarction (MI) (outcome) among n = 494. Information on smoking was also available. The analysis investigated possible modification of the coffee-MI relationship with smoking status (stratification). The sample size is n = 494.

A **stratified analysis of K 2x2** tables is used to assess:

1. evidence of modification of an exposure-disease relationship by changes in the value of a third (stratifying) variable; or
2. in the absence of modification, a Mantel-Haenszel analysis of an exposure disease relationship controlling for confounding.

Here is the data in tabular form.

Stratum 1: Former Smoker (smoking = 1)

	MI (mi = 1)	Control (mi = 0)
>= 5	7	18
< 5	20	112

Stratum 2: 1-14 Cigarettes/Day (smoking = 2)

	MI (mi = 1)	Control (mi = 0)
>= 5	7	24
< 5	33	11

Stratum 3: 35-44 Cigarettes/Day (smoking = 3)

	MI (mi = 1)	Control (mi = 0)
>= 5	27	24
< 5	55	58

Stratum 4: 45+ Cigarettes/Day (smoking = 4)

	MI (mi = 1)	Control (mi = 0)
>= 5	30	17
< 5	34	17

```r
library(Hmisc)
```

2a. How to enter tabular data

```
## Loading required package: grid
## Loading required package: lattice
##
## Attaching package: 'lattice'
##
## The following object is masked from 'package:epicalc':
##
##     dotplot
##
## Loading required package: Formula
##
## Attaching package: 'Hmisc'
##
## The following object is masked from 'package:epicalc':
##
##     fillin
##
## The following objects are masked from 'package:base':
##
```

```
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
names(dat)[1] <- "smoking"
label(dat$smoking) <- "smoking status"
label(dat$lungca) <- "Cups coffee per day"
describe(dat)
```

```
## dat
##
##  3  Variables      89  Observations
## ---------------------------------------------------------------------------
## smoking : smoking status
##       n missing  unique
##      89       0       2
##
## Non-smoker (49, 55%), Smoker (40, 45%)
## ---------------------------------------------------------------------------
## lungca : Cups coffee per day
##       n missing  unique    Info     Sum    Mean
##      89       0       2    0.32      11  0.1236
## ---------------------------------------------------------------------------
## tally
##       n missing  unique    Info    Mean
##      89       0       4    0.81   36.57
##
## 2 (2, 2%), 9 (9, 10%), 31 (31, 35%), 47 (47, 53%)
## ---------------------------------------------------------------------------
```

```
contents(dat)
```

```
##
## Data frame:dat   89 observations and 3 variables    Maximum # NAs:0
##
##                        Labels Levels   Class Storage
## smoking      smoking status      2         integer
## lungca  Cups coffee per day         numeric  double
## tally                                        double
##
## +--------+-----------------+
## |Variable|Levels           |
## +--------+-----------------+
## | smoking|Non-smoker,Smoker|
## +--------+-----------------+
```

Let's load the data.

```
url3 <- "http://people.umass.edu/biep640w/datasets/coffeemi_full.dta"
dat2 <- read.dta13(file = url3)
```

```
## Warning: duplicated levels in factors are deprecated
```

```
contents(dat2)
```

```
##
## Data frame:dat2  494 observations and 4 variables    Maximum # NAs:0
##
##         Levels Storage
## smoking      4 integer
## coffee       2 integer
## mi           3 integer
## tally          double
##
## +--------+------------------------------------------------------+
## |Variable|Levels                                                |
## +--------+------------------------------------------------------+
## | smoking|Former Smoker,1-4 cigs/day,35-44 cigs/day,45+ cigs/day|
## +--------+------------------------------------------------------+
## | coffee |5+cups,Less                                           |
## +--------+------------------------------------------------------+
## | mi     |Non-MI,MI,Non-MI                                      |
## +--------+------------------------------------------------------+
```

If you examine the variable **mi**, there are some funky things going on as it has 3 levels when in actuality, it should only have 2. Let's fix that along with reordering the levels of **mi** and **coffee**.

```
dat2$mi <- droplevels(dat2$mi)
```

```
## Warning: duplicated levels in factors are deprecated
```

```
dat2$mi <- relevel(dat2$mi, "Non-MI")
dat2$coffee <- relevel(dat2$coffee, "Less")
contents(dat2)
```

```
##
## Data frame:dat2  494 observations and 4 variables    Maximum # NAs:0
##
##         Levels Storage
## smoking      4 integer
## coffee       2 integer
## mi           2 integer
## tally          double
##
## +--------+------------------------------------------------------+
## |Variable|Levels                                                |
## +--------+------------------------------------------------------+
## | smoking|Former Smoker,1-4 cigs/day,35-44 cigs/day,45+ cigs/day|
## +--------+------------------------------------------------------+
## | coffee |Less,5+cups                                           |
## +--------+------------------------------------------------------+
## | mi     |Non-MI,MI                                             |
## +--------+------------------------------------------------------+
```

Let's create a bit more of a sophisticated table than we have in the past with a new library, `descr`.

```r
library(descr) # remember to install the package
# first, specify the row and column
# next, prop.r and prop.c specify row and column proportions
# by default, a mosaic plot is displayed, which I have turned off by setting plot to FALSE
crosstab(dat2$coffee, dat2$mi, prop.r = TRUE, prop.c = TRUE, plot = FALSE)
```

```
##    Cell Contents
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |          Column Percent |
## |-------------------------|
##
## =======================================
##                 dat2$mi
## dat2$coffee    Non-MI       MI    Total
## ---------------------------------------
## Less              198      142      340
##                58.235   41.765   68.826
##                70.463   66.667
## ---------------------------------------
## 5+cups             83       71      154
##                53.896   46.104   31.174
##                29.537   33.333
## ---------------------------------------
## Total             281      213      494
##                56.883   43.117
## =======================================
```

And if we want to stratify over the variable **smoking**, just do the following:

```r
# unfortunately, the crosstab function is giving us a hard time...
table6 <- with(dat2, table(coffee, mi, smoking))
ftable(table6) # so, we don't have any row or column percentages :(
```

```
##               smoking Former Smoker 1-4 cigs/day 35-44 cigs/day 45+ cigs/day
## coffee mi
## Less   Non-MI                   112           11             58           17
##        MI                        20           33             55           34
## 5+cups Non-MI                    18           24             24           17
##        MI                         7            7             27           30
```

```r
ftable(prop.table(table6))
```

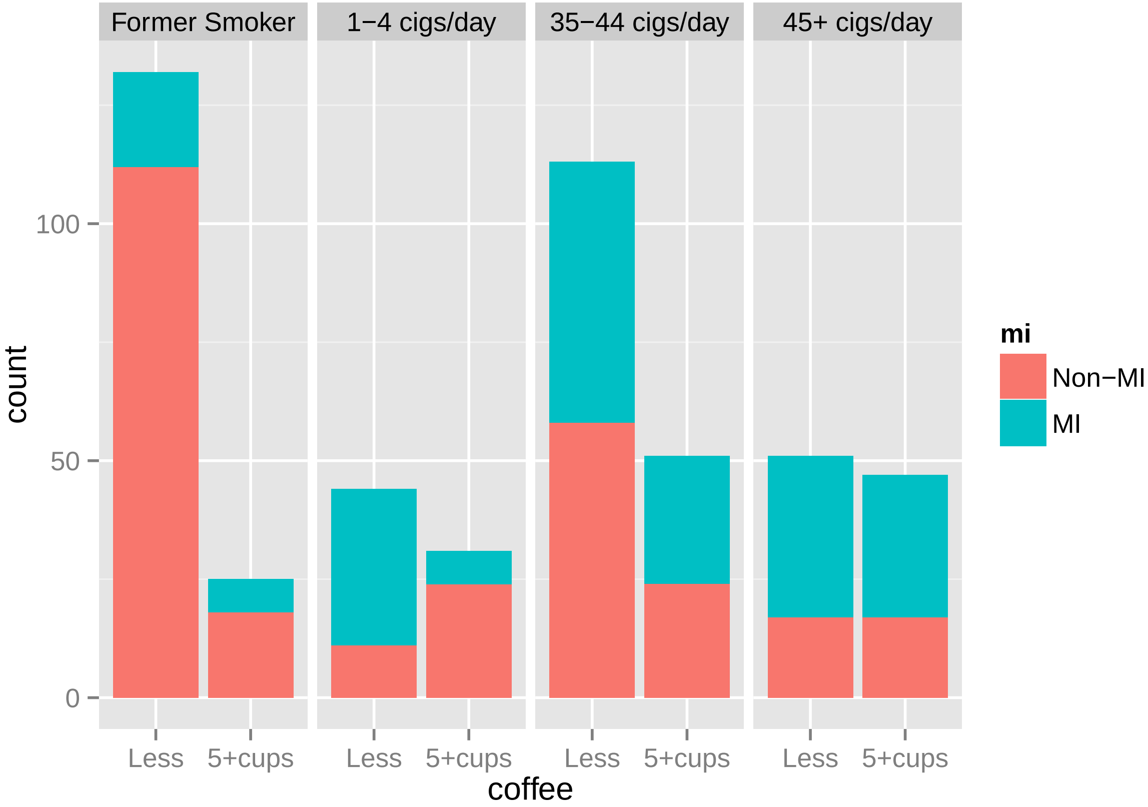
```
##               smoking Former Smoker 1-4 cigs/day 35-44 cigs/day 45+ cigs/day
## coffee mi
## Less   Non-MI               0.22672      0.02227        0.11741      0.03441
##        MI                   0.04049      0.06680        0.11134      0.06883
## 5+cups Non-MI               0.03644      0.04858        0.04858      0.03441
##        MI                   0.01417      0.01417        0.05466      0.06073
```

As for displaying descriptive statistics, `R` does not win in having nice, compact tables such as `Stata`.

2c. Descriptives - Graphical One of R's strengths comes from its ability to plot/graph beautiful data visualizations. Let us create a bar graph as well as an odds ratio with 95% confidence limits.

2c.a. Bar Graph This bar graph will display the percentage experiencing the outcome, over exposure, separately for each value of the stratification variable. That's a lot to take in, but the graph will make sense. As for the syntax of the code, we will use the `ggplot2` package to make sense of it.

```r
library(ggplot2)
ggplot(data = dat2, aes(x = coffee, fill = mi)) + geom_bar(position = "stack") + facet_grid(. ~ smoking
```



Unfortunately, the proportions are a bit tricky to display, but the best I can show is the COUNT proportion of non-MI and MI against the binvary variable **coffee**, stratified by smoking.

2c.b. Odds Ratio and 95% CI Finding a suitable package in R has proven challenging, but they are out there. Follow the instructions below to obtain the odds ratio as well as plotting the 95% confidence intervals. It should be noted that the lower and upper bound are similar to the `Stata` handout, which I am unsure of what is going on under the hood.

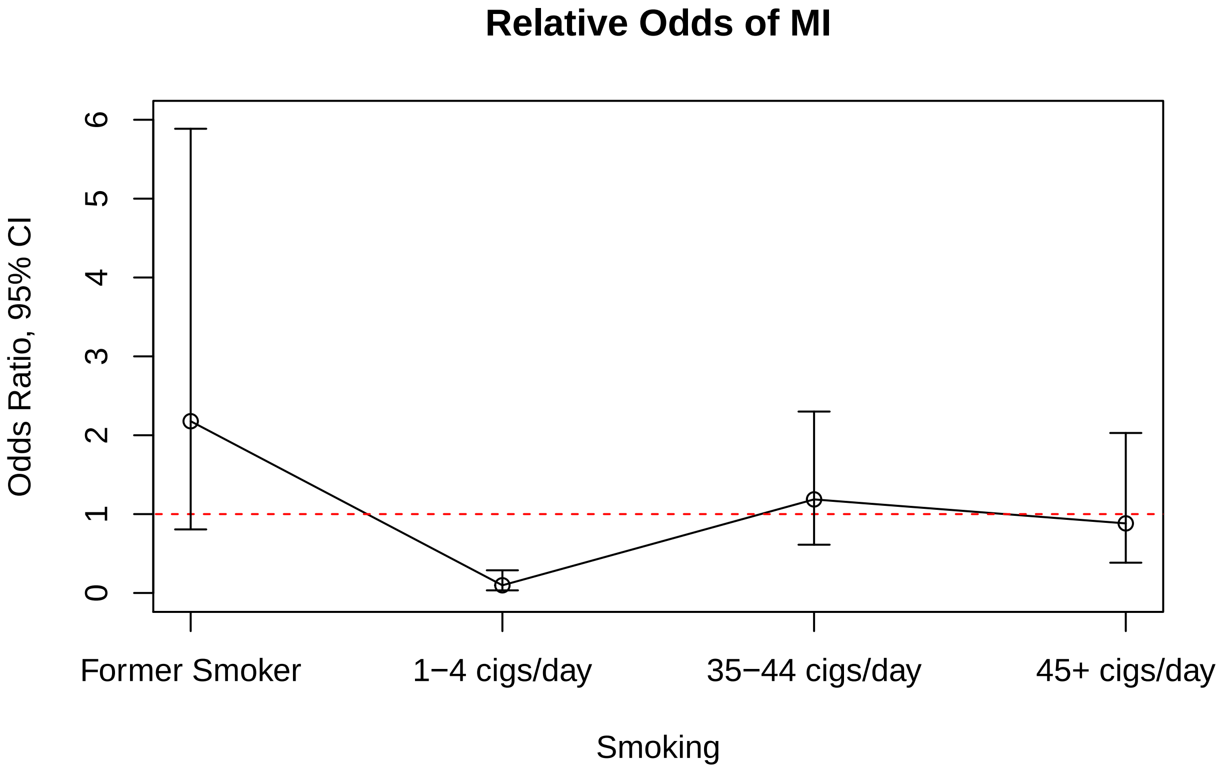
```r
library(vcd) # install the package first
odds.2cb <- oddsratio(table6, log = FALSE) # computes the odds ratio
summary(odds.2cb) # summary displays the odds ratio
```

```
##                 Odds Ratio
## Former Smoker         2.18
## 1-4 cigs/day          0.10
## 35-44 cigs/day        1.19
## 45+ cigs/day          0.88
```

```
confint(odds.2cb) # displays the confidence intervals
```

```
##                      lwr    upr
## Former Smoker  0.80577 5.8860
## 1-4 cigs/day   0.03289 0.2874
## 35-44 cigs/day 0.61187 2.3003
## 45+ cigs/day   0.38381 2.0285
```

```
# Lastly, let's plot the odds ratio and their respective confidence intervals
plot(odds.2cb, main = "Relative Odds of MI", xlab = "Smoking", ylab = "Odds Ratio, 95% CI")
```



2d. Mantel-Haenszel Test of Null: Homogeneity of Odds Ratio In order to test for the homogeneity of the odds ratio, we compute another function from the `vcd` package, called the Woolf Test.

```
woolf_test(table6)
```

```
##
##  Woolf-test on Homogeneity of Odds Ratios (no 3-Way assoc.)
##
## data:  table6
## X-squared = 19.79, df = 3, p-value = 0.0001873
```

From Professor Bigelow's on how to interpret the above statistic: > The Mantel Haenszel test of homogeneity of odds ratio is statistically significant (Chi-Square with df = 3, X-squared = 19.92, p-value = 0.0002). The assumption of the null hypothesis of no assocation, when applied to the observed data, has led to an extremely unlikely event. The null hypothesis is rejected. Conclude that there is statistically significant evidence that the association of high coffee consumption with event of myocardial infarction is different, depending on smoking status.

You'll notice the numbers are a bit off, again. Also, the crude odds ratio is not available. . . ???

2e. Mantel-Haenszel Test of Null: Odds Ratio Well, we've been touting the Mantel-Haenszel test for quite some time and now we're finally going to use it.

```
# for now, don't worry about the explanation of the options, just input the table
mantelhaen.test(table6, correct = FALSE, exact = FALSE)$statistic
```

```
## Mantel-Haenszel X-squared
##                     1.648
```

```
mantelhaen.test(table6, correct = FALSE, exact = FALSE)$p.value
```

```
## [1] 0.1992
```

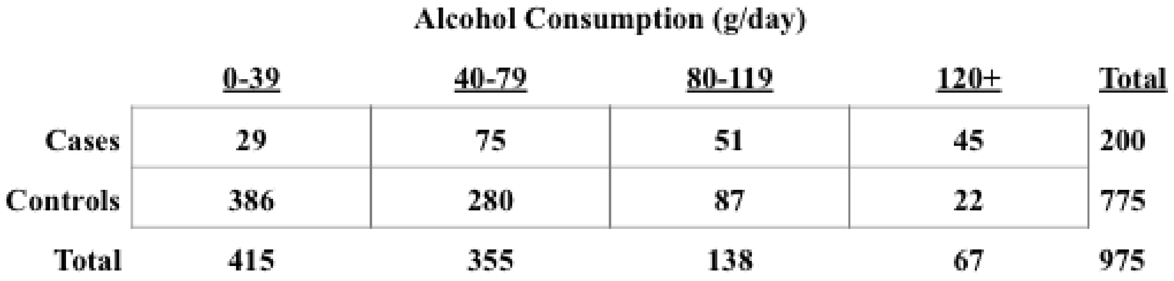
From the output, we receive the Mantel-Haenszel Chi-Square statistic of 1.65 (rounded) and a p-value of .1992 (also rounded). Professor Bigelow interprets it as such: > In real world practice, because we have evidence of effect modification of the coffee-MI relationship, depending on smoking status, we would not actually perform this test.

> The results shown here indicate that, on average, there is no assocation of high coffee consumption with the event of myocardial infarction Chi-Square on df = 1, statistic = 1.65, and p-value = .1992.

2xC Table Analysis of Trend

Introduction to Example *Source:* Tuyns AJ, Pequignot G and Jenson OM (1997) Le cancer de l'oesphage en Ille-et-Villaine en function des niveaux de consummation d'alcool et de tabac. *Bull Cancer* 64: 45-60.

The following are excerpted data from a case-control study of the relationship between alcohol consumption at 4 increasing levels ("doses") and case-control status for the disease of esophageal cancer.



Alcohol Consumption (g/day)

	0-39	40-79	80-119	120+	Total
Cases	29	75	51	45	200
Controls	386	280	87	22	775
Total	415	355	138	67	975

Because the study design is **case-control**, an appropriate measure of association is the **odds ratio** measure of association. We are specifically interested in how the relative odds of esophageal cancer changes with increasing alcohol consumption. Thus, there are at least two research questions:

1. Does the odds of esophageal cancer differ by level of alcohol consumption? (Test of general assocation)

   H_0: No association between exposure and disease H_A: Any association between exposure and disease (unspecified)

2. If the odds of esophageal cancer differs by level of alcohol consumption, then does the odds of esophageal cancer increase with increasing levels of alcohol consumption? (Test of trend)

   H_0: No association between exposure and disease H_A: Monotone increasing (or decreasing) association between exposure and disease (trend)

3a. Descriptives - Numerical Let's load the data.

```
library(foreign)
url4 <- "http://people.umass.edu/biep640w/datasets/esophageal_cancer.dta"
dat3 <- read.dta(file = url4)
des(dat3)
```

```
##
##  No. of observations =  10
##   Variable      Class           Description
## 1 alcohol       factor          Alcohol g per day
## 2 case          factor
## 3 tally         numeric
```

We see that we have three variables, two of which are factors, one numeric, and 10 observations.

Moving on, let's put it in a table.

```
# I don't use this command much, but hey, it's a part of R's base package
xtabs(tally ~ case + alcohol, data = dat3)
```

```
##          alcohol
## case       0-39g 40-79 80-119g 120+
##   control     47    31       9    5
##   case         2     9       9    5
```

Next, let's expand however many counts for each tabulation into a data frame. It's a bit of doozy so if you don't get it the first time around, try again or shoot me an email.

```
dat3.edit <- dat3[-(9:10), ] # let us rid the frame of the totals. There is some odd behavior if includ
dat3.expand <- lapply(dat3.edit, function(x) rep(x, dat3.edit$tally)) # apply the number of counts
dat3.expand.df <- as.data.frame(dat3.expand) # let's ensure it's a data frame
dat3.new <- dat3.expand.df[, -3] # let's remove the counts/tallies
with(dat3.new, table(case, alcohol)) # and we should end up with the original table
```

```
##          alcohol
## case       0-39g 40-79 80-119g 120+
##   control     47    31       9    5
##   case         2     9       9    5
```

```
# and if we want the margin counts, we do the following
addmargins(with(dat3.new, table(case, alcohol)))
```

```
##          alcohol
## case       0-39g 40-79 80-119g 120+ Sum
##   control     47    31       9    5  92
##   case         2     9       9    5  25
##   Sum         49    40      18   10 117
```

Next, let's calculate tabulate the row and column percentages. I believe I have done this before with many other functions but here is the definitive one I will use from now on.

```
with(dat3.new, tabpct(case, alcohol, graph = FALSE))
```

```
##
## Original table
##          alcohol
## case       0-39g  40-79  80-119g  120+  Total
##   control     47     31        9     5     92
##   case         2      9        9     5     25
##   Total       49     40       18    10    117
##
## Row percent
##          alcohol
## case         0-39g   40-79  80-119g   120+  Total
##   control       47      31        9      5     92
##             (51.1)  (33.7)    (9.8)  (5.4)  (100)
##   case           2       9        9      5     25
##                (8)    (36)     (36)   (20)  (100)
##
## Column percent
##          alcohol
## case       0-39g       %  40-79       %  80-119g      %  120+      %
##   control     47  (95.9)     31  (77.5)        9   (50)     5   (50)
##   case         2   (4.1)      9  (22.5)        9   (50)     5   (50)
##   Total       49   (100)     40   (100)       18  (100)    10  (100)
```

Now, let's calculate the mean and standard deviation of the variable **case** for each level of the variable **alcohol**. Before we do, let's alter the variable **case** into a numeric variable, actually a binary variable of values 0 and 1.

```
dat3.new$newcase <- as.numeric(dat3.new$case) - 1
aggregate(dat3.new$newcase, by = list(dat3.new$alcohol), FUN = mean)
```

```
##   Group.1 mean.dat3.new$newcase
## 1   0-39g               0.04082
## 2   40-79               0.22500
## 3 80-119g               0.50000
## 4    120+               0.50000
```

```
aggregate(dat3.new$newcase, by = list(dat3.new$alcohol), FUN = sd)
```

```
##   Group.1 sd.dat3.new$newcase
## 1   0-39g              0.1999
## 2   40-79              0.4229
## 3 80-119g              0.5145
## 4    120+              0.5270
```

========================== **missing tabodds function in R** ===============

3b. Descriptives - Graphical ALSO INCOMPLETE

3c. Chi-Squared Tests of General Association Let's calculate the expected frequency, column percentage, and chi-squared tests.

```
# original table
with(dat3.new, table(case, alcohol)) # table
```

```
##          alcohol
## case      0-39g 40-79 80-119g 120+
##   control    47    31       9    5
##   case        2     9       9    5
```

```
with(dat3.new, chisq.test(table(case, alcohol))) # chi-squared test
```

```
## Warning: Chi-squared approximation may be incorrect
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  table(case, alcohol)
## X-squared = 22.41, df = 3, p-value = 5.368e-05
```

```
with(dat3.new, chisq.test(table(case, alcohol))$expected) # expected cell frequencies
```

```
## Warning: Chi-squared approximation may be incorrect
```

```
##          alcohol
## case      0-39g  40-79 80-119g  120+
##   control 38.53 31.453  14.154 7.863
##   case    10.47  8.547   3.846 2.137
```

```
with(dat3.new, tabpct(case, alcohol, graph = FALSE, percent = "col")) # column frequencies
```

```
## 
## Column percent
##          alcohol
## case       0-39g       %  40-79       %  80-119g      %  120+      %
##   control     47  (95.9)     31  (77.5)        9   (50)     5   (50)
##   case         2   (4.1)      9  (22.5)        9   (50)     5   (50)
##   Total       49   (100)     40   (100)       18  (100)    10  (100)
```

Next, let's examine the Pearson residuals.

```
with(dat3.new, chisq.test(table(case, alcohol))$residuals)
```

```
## Warning: Chi-squared approximation may be incorrect
##          alcohol
## case          0-39g    40-79  80-119g     120+
##   control  1.36455 -0.08077 -1.36992 -1.02108
##   case    -2.61766  0.15495  2.62796  1.95876
```

As for the code for the contribution to the chi-squared statistic, that may be calculated manually from the previous input, implementing the given and expected frequencies. Please see Prof. Bigelow's Stata handout for more information.

3d. Test of Trend INCOMPLETE Once again, I have not found an equivalent of tabodds from Stata to R.

4. RxC Table Test of Trend INCOMPLETE

Unfortunately, this still needs to be further looked at as R differs from Stata for this function.

```
prop.trend.test(x = c(47, 31, 9, 5), n = c(49, 40, 18, 10))
```

```
##
##  Chi-squared Test for Trend in Proportions
##
## data:  c(47, 31, 9, 5) out of c(49, 40, 18, 10) ,
##  using scores: 1 2 3 4
## X-squared = 21.03, df = 1, p-value = 4.519e-06
```