Unit 8

Stata for Categorical Data Analysis

version 16

"Vive la difference!"

Categorical data are count data and are discrete.

Methods for their description include frequencies, relative frequencies, risks, odds, risk ratios (relative risk), odds ratios (relative odds) and more.

Methods for their analysis include Fisher Exact Tests of association, chi square tests of association, chi square tests of trend and eventually modeling (logistic regression, poisson regression, and more

This unit describes the use of Stata for summarizing and graphing count data and the analysis of contingency tables.

Important! As recommended for continuous data, always precede a regression analysis for discrete data with contingency table analyses, such as those described here!

Table of Contents

Learning Objectives

When you have finished this unit, you should be able to use Stata to:

- Perform a Fisher Exact Test and a Chi Square test of association for data in a single 2x2 table.
- Perform an analysis of risk and relative risk for cohort data in a single 2x2 table.
- Perform an analysis of odds and relative odds (OR) for case-control data in a single 2x2 table.
- Perform an analysis of association for paired data in a single 2x2 table using *McNemar's Test*.
- Produce a graphical summary of data from stratified 2x2 tables.
- Perform and interpret an analysis of stratified 2x2 tables, using *Mantel-Haenszel methods*.
- Produce a graphical summary of data from a 2xC table, ordered
- Perform an analysis of trend in data from a 2xC table, ordered.
- Perform and interpret the *test of trend* for RxC tables of counts of ordinal data that are suitable for explorations of dose-response.
- Perform and interpret a *chi square goodness-of-fit (GOF)* test.

Preliminary #1: Introduction to Entering and Working with Tabular Data

Tip! Good news. It is possible to input data in the form of a table and analyze it in Stata. In brief, this entails the following:

__1. Entering into Stata the cell identifiers using row and column variables;
__2. Entering for each identified cell, the frequency of that combination; and
__3. Finally, using the command **expand** to create the full (individualized) data set.

Data in tabular form are a ***collapsing*** of n=157 individual records. Consider the following.

	Heart Attack)	
Cups coffee/day	MI (mi=1)	CONTROL (mi=2)
≥ 5 cups (coffee=1)	**7**	18
< 5 (coffee=2)	20	112

The table entry 7 represents a **frequency of 7** individual observations of **coffee=1, mi=1.**

In a spreadsheet of *individual data* with one row for each subject, we'd have 7 rows of **coffee=1, mi=1:**

coffee	mi
1	1
1	1
1	1
1	1
1	1
1	1
1	1
etc	etc

In contrast, in a spreadsheet of *tabular data*, we'd have just 1 row for **coffee=1, mi=1.** But we would also have an additional variable that keeps track of the frequency. Here I've chosen to name my frequency count variable **tally**:

coffee	mi	tally
1	1	7

Happily for us, Stata can convert data in tabular summary form to a new dataset of individual observations through the use of the command expand.

Illustration –

Input tabular data and create a Stata data set of individual observations:

Consider again the 2x2 table of frequencies on the previous page. Below is our *tabular data* in a spreadsheet with just 4 rows.

coffee	mi	tally = # repeats (cell frequency)
1	1	7
1	2	18
2	1	20
2	2	112

Step 1 – Create a coding manual *Suggestion (not required)* – Use "lower case" variable names.

Example -

Variable	Variable Label	Format	Format code definitions
coffee	Cups Coffee Per Day	coffeef	1 = 5+ cups/day 2 = Less
mi	Myocardial Infarction	mif	1 = MI 2 = Non-MI
tally	Frequency (weight)	-	0, 1, 2, ….

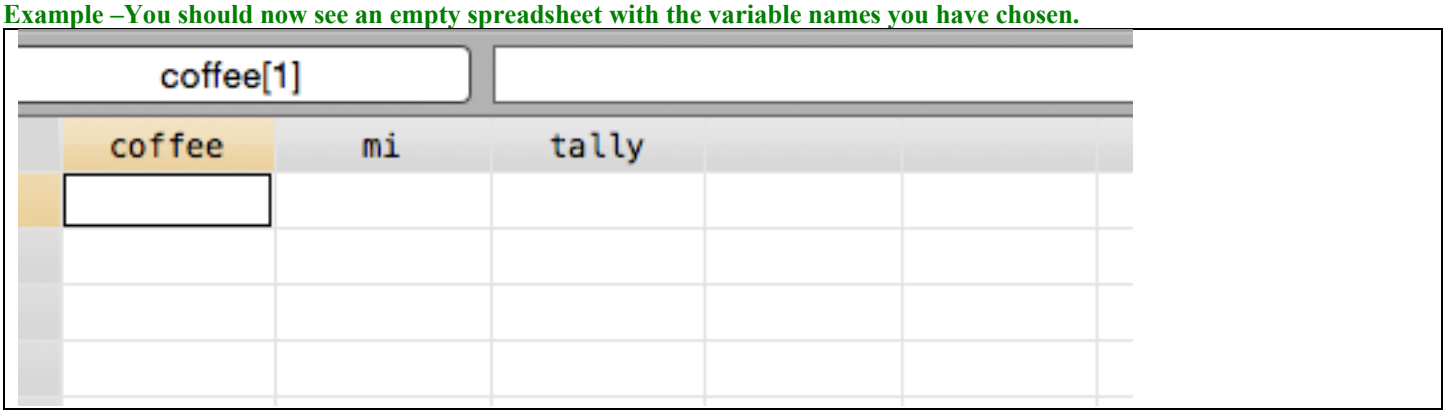
Step 2 – In Stata, define variables and initialize to missing.

Example -

```
. generate coffee=.
. generate mi=.
. generate tally=.
```

Step 3 – From the top menu, click on the data editor icon.

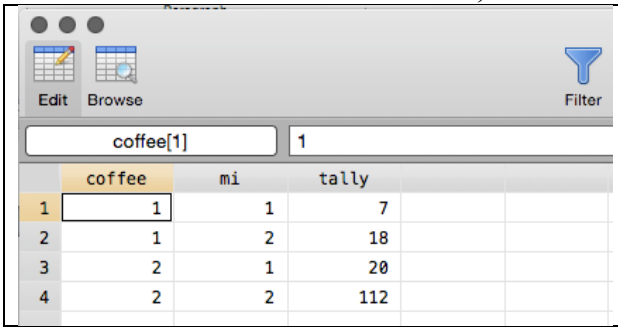
Example –You should now see an empty spreadsheet with the variable names you have chosen.



Step 4 – In EXCEL, enter your data, taking care to match the column headings with your empty Stata spreadsheet and remembering to format your excel data as numeric.

A	B	C
coffee	mi	tally
1	1	7
1	2	18
2	1	20
2	2	112

Step 5 – In EXCEL highlight and select the rows of data only (not the row of variable names). Do an EDIT>COPY. Then in STATA, in the data editor window, do an EDIT>PASTE. You should have:



	coffee	mi	tally
1	1	1	7
2	1	2	18
3	2	1	20
4	2	2	112

Step 6 – Close the data editor window to return to the command line (Don't worry – your data is not lost.)
In the command window, assign variable names, create value labels, and assign value labels.
Example –

```
. label variable coffee "Cups of Coffee/Day"
. label variable mi "MI - Myocardial Infarction"
. label define mif 1 "MI" 2 "Non-MI"
. label define coffeef 1 "5+cups" 2 "Less"
. label values coffee coffeef
. label values mi mif
```

Step 7 (Optional) -
Save a Stata version of your tabular data (To be honest, I'm not sure you need to do this)
Example –

```
. save "/Users/cbigelow/Desktop/illustration_tabular.dta"
```

Step 8– Key step – Here is where you create individual observations.
Use the command expand *variablename* where *variablename* is the variable that contains the cell frequencies.
Example – Stata tells me that it has created 153 *additional* observations.

```
. expand tally
(153 observations created)
```

Design ..... Data Collection ..... Data Management ..... Data Summarization ..... Statistical Analysis ..... Reporting

But wait!! ***"(153 observations created)"*** **seems to be wrong**
No worries. It all checks out just fine:
The total sample size is (7+18+20+112) = 157
But in creating the table you started off with 4 observations. You already had these. So now,
4 observations (tabular) + 153 additional observations = sum of table totals = 157, match! Phew.

Step 9 – SAVE. Before you save, now that you have a fully expanded data set, you can drop the variable that keeps track of the cell frequencies.

Example –

```
. save "/Users/cbigelow/Desktop/illustration_full.dta"
file /Users/cbigelow/Desktop/illustration_full.dta saved
```

Preliminary #2: 0's and 1's, 1's and 2's, and flipped rows and columns

Glitch. Sometimes a 2-way cross tabulation in Stata is not laid out as you'd like. It might be a slew of 1's and 2's when you really rather have 0's and 1's. And vice versa. There's an explanation and a fix.

Example – Is exposure to HIGH coffee consumption (exposure=yes) associated with MI (event=yes)?
Suppose you choose to use 1/2 variables (as for example, in entering tabular data).

Stata orders 1/2 possibilities with "1" first and "2" second.
Suppose your variables **coffee** and **mi** are defined as follows:

coffee = 1 if HIGH (≥ 5 cups/day)
2 otherwise (< 5 cups/day)

mi = 1 if YES, an MI occurred
2 otherwise (no MI)

All is well. A cross-tabulation in Stata using **tab2** produces the layout you want (exposure is in the first row, event is in the first column)

		mi
coffee	**MI** (mi=1)	**Control** (mi=2)
≥ 5 (coffee=1)	7	18
< 5 (coffee=2)	20	112

Now suppose you would rather work with these variables as 0/1 variables (we will want to do this later, when we learn the epidemiology tables commands such as cs and cc).

Stata orders 0/1 possibilities with "0" first and "1" second.

Suppose you want your variables **coffee01** and **mi01** to be defined as follows:

coffee01 = 1 if HIGH (≥ 5 cups/day)
0 otherwise (< 5 cups/day)

mi01 = 1 if YES, an MI occurred
0 otherwise (no MI)

Oh dear. A cross-tabulation in Stata using **tab2** will *not* produce the layout you want (exposure is in the first row, event is in the first column). Instead, you get the following flipped arrangement:

		mi01
coffee01	**MI** (mi=0)	**Control** (mi=1)
< 5 (coffee=0)	112	20
≥ 5 (coffee=1)	18	7

```
. * HOW TO:  Create 0/1=event variables from 1=event/2 variables
. generate coffee01=coffee
. recode coffee01 (2=0)
. generate mi01=mi
. recode mi01 (2=0)
```

1. Single Proportion

Note - The following is duplicated from page 18 of notes, *7. Stata for Analysis of 1, 2, 3+ Samples* for completeness sake.

Command	Example
Exact Confidence Interval for Binomial π **ci *variable*, binomial level*(#)*** **This produces Clopper-Pearson "exact" confidence interval**	**.ci foreign, binomial level(90)** **This produces an exact 90% confidence interval estimate of the binomial parameter π for the variable foreign**
Confidence Interval for Binomial π, *"immediate"* **cii *n observedproportion*, binomial level*(#)***	**.cii 74 .2973, level(90)** **This produces an exact 90% confidence interval estimate of the binomial parameter π for an UNNAMED variable**
Exact test for Binomial π **bitest *variable=nullpi*** **The option level( ) is NOT allowed**	**.bitest foreign=.28** **This produces an exact test of significance of the null hypothesis that the binomial parameter π = .28 for the variable foreign**
Exact test for Binomial π, *"immediate"* **bitesti *n #successes nullpi*** **The option level( ) is NOT allowed**	**.bitesti 74 22 .28** **This produces an exact test of significance of the null hypothesis that the binomial parameter π = .28 for an UNNAMED variable in the setting where N=74, # successes = 22 and the null hypothesis that π = .28**
Normal Approximation test for Binomial π **prtest *variable=nullpi*, level*(#)***	**.prtest foreign=.28, level(95)** **This produces a normal aproximation test of significance of the null hypothesis that the binomial parameter π = .28 for the variable foreign. The output includes a 95% confidence interval estimate of π.**
Normal Approximation test for Binomial π, *"immediate"* **prtesti *n #successes nullpi*, count level*(#)***	**.prtesti 74 22 .28, count level(95)** **This produces a normal approximation test of significance of the null hypothesis that the binomial parameter π = .28 for an UNNAMED variable in the setting where N=74, # successes = 22 and the null hypothesis that π = .28. The output includes a 95% confidence interval estimate of π.**

2. Single 2x2 Table

2.1. Tests of Association Using tabi with Direct Entry of Counts

Illustrative Data for this Section

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.

	Five Year Status	
Courses	**Dead**	**Alive**
1-4	21	2
≥ 10	2	8

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

Sometimes the data you have is in tabular form and you will enter them in the command line
You have a 2x2 table with cell counts. You do not have individual observations. Stata has an immediate command that lets you analyze the data in this table. It is the command **tabi**

How to Use the Command tabi

Command and Examples	Notes
tabi *row1col1 row1col2\row2col1 row2col2* **Example using illustrative data above** **tabi 21 2\2 8**	**Use this to obtain Fisher Exact Test.** **This is recommended for small to moderate sample sizes.**
tabi *row1col1 row1col2\row2col1 row2col2*, lrchi **Example using illustrative data above** **tabi 21 2\2 8, lrchi**	**Use option lrchi to obtain Chi Square Test.** **This is recommended for small to moderate sample sizes.**

2.2 Tests of Association Using tabulate

Illustrative Data for this Section
Download from the course website: **single2x2.dta.**
It contains the following 2x2 table of counts.

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

Introduction to the Command tabulate

Note – The command tabulate is NOT an immediate command. It requires a stata data set in working memory

How to Use the Command tabulate

Command and Examples	Notes
tabulate *rowvariable columnvariable*, exact **Example using illustrative data above** **tabulate smoking lungca, exact**	**Use this to obtain Fisher Exact Test.** **This is recommended for small to moderate sample sizes.**
tabulate *rowvariable columnvariable*, lrchi **Example using illustrative data above** **tabulate smoking lungca, lrchi**	**Use option lrchi to obtain Chi Square Test.** **This is recommended for small to moderate sample sizes.**

2.3 Analysis of a *Cohort Study* using the command csi or cs

IMPORTANT!!! - The commands csi and cs require 0/1 coding **(see again page 8 for how to create)**

You must use **0 or 1 coding** for your case variable, with **1=case**
You must use **0 or 1 coding** with your exposure variable with **1=exposed**

Illustrative Data for this Section
Download from the course website: **single2x2.dta.**
It contains the following 2x2 table of counts.

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

How to Use the Commands csi and cs

Command and Examples	Notes
csi *row1col1 row1col2 row2col1 row2col2* **csi *#a #b #c #d*** **Example using illustrative data above** **csi 9 31 2 47, exact**	**Use option exact to obtain Fisher Exact Test.** **This is recommended for small to moderate sample sizes.**
cs *casevariable01 exposurevariable01*, lrchi **Example using illustrative data above** **cs lungca smoking, or**	**Use option or to obtain odds ratios.**

Tip -
Type **help cs** to familiarize yourself with other options that are possible.

2.4 Analysis of a *Case-Control* Study using the command cci or cc

IMPORTANT!!! - The commands csi and cs require 0/1 coding **(see again page 8 for how to create)**

You must use **0 or 1 coding** for your case variable, with **1=case**
You must use **0 or 1 coding** with your exposure variable with **1=exposed**

Illustrative Data for this Section
Download from the course website: **single2x2.dta.**
It contains the following 2x2 table of counts.

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

How to Use the Commands cci and cc

Command and Examples	Notes
cci *row1col1 row1col2 row2col1 row2col2* **cci *#a #b #c #d*** **Example using illustrative data above** **csi 9 31 2 47, exact**	**Use option exact to obtain Fisher Exact Test.** **This is recommended for small to moderate sample sizes.**
cc *casevariable01 exposurevariable01*, lrchi **Example using illustrative data above** **cc lungca smoking, or**	**Use option or to obtain odds ratios.**

Tip -
Type **help cc** to familiarize yourself with other options that are possible.

2.5 *Matched Case-Control Data*: McNemar's Test using command mcci

Illustrative Data for this Section

Source:
Forster et al (1995) Risk factors in clinically diagnosed presenile dementia of the Alzheimr type: a case-control study in northern England. *J. Epidemiol. Comm. Health* 49: 253-258.

Forster et al conducted a ***matched*** case-control study of 109 clinically diagnosed patient aged < 65 years. Each case was 1:1 matched with a community control person of the same sex and age. The exposure variable of interest in this study was whether or not a relative had dementia. The following are the data:

	Control has a relative with dementia?	
Case has a relative with dementia?	Yes	No
Yes	6	25
No	12	66

→ **Thus, the data are comprised of 109 *pairs*, representing a total of 218 individuals.**

Do these data provide statistically significant evidence of an association of family history of dementia with case status of dementia?

Quick Review of Matched Study Designs and Their Analysis

Matched study designs are considered when it is of interest to control for confounding. In a matched case-control study, the selection of controls defined as "matches" to the cases is done to achieve similar distributions on key variables. In the illustration presented here, matching yields case and control distributions of age and sex that are similar. *Whenever a matched study design is used, the analysis must take into account the matching. This is because, in matched data, the cases and controls are not independent*. As such, they are more similar to each than they would have been had independent sampling been employed.

		Controls	
		Exposed	Not Exposed
Cases	Exposed	aa	bb
	Not exposed	cc	dd

***Tip* – The entries denoted "aa", "bb", "cc" and "dd" remind us that we are working with pairs!**

Design ..... **Data Collection** ..... **Data Management** ..... **Data Summarization** ..... **Statistical Analysis** ..... **Reporting**

In a matched case-control 2x2, only the discordant pairs ("bb" and "cc") contribute to the analysis.
In this illustration, **"bb" = 25** and **"cc"= 12".**

Matched Odds Ratio

$$\text{Matched OR} = \frac{\#\text{pairs: cases exposed \& controls NOT exposed}}{\#\text{pairs: controls exposed \& cases NOT exposed}} = \frac{bb}{cc} = \frac{25}{12} = 2.08$$

95% Confidence Interval for Matched Odds Ratio

Lower limit = $\exp[\ln(\text{matched OR}) - 1.96*SE_{\ln(\textit{matched OR})}] = \exp[\ln(2.08) - 1.96*.3512] = 1.0450$

Upper limit = $\exp[\ln(\text{matched OR}) + 1.96*SE_{\ln(\textit{matched OR})}] = \exp[\ln(2.08) + 1.96*.3512] = 4.1401$

where you have previously calculated

$$SE_{\ln(\textit{matched OR})} = \sqrt{\frac{1}{bb}+\frac{1}{cc}} = \sqrt{\frac{1}{25}+\frac{1}{12}} = 0.3512$$

Test of Null: Matched Odds Ratio = 1

In words, null: "No association of exposure and disease"

This is the formula that Stata Uses	
With ***No*** Continuity Correction	With Continuity Correction
$\text{McNemar } \chi^2_{DF=1} = \frac{(bb - cc)^2}{(bb + cc)}$ $= \frac{(25 - 12)^2}{(25 + 12)} = \frac{169}{37} = 4.57$	$\text{McNemar } \chi^2_{DF=1} = \frac{(\lvert bb - cc \rvert - 1)^2}{(bb + cc)}$ $= \frac{(\lvert 25 - 12 \rvert - 1)^2}{(25 + 12)} = \frac{144}{37} = 3.89$
(stata command not shown): p-value = .0325	***(stata command not shown): p-value = .0486***

Interpretation:
The McNemar test of association of exposure and disease in these matched data is statistically significant (Chi Square with df=1 = 4.57 or 3.89, depending on continuity correction; both p-values are < .05). The assumption of the null hypothesis of no association, when applied to the observed data, has led to a modestly unlikely event. The null hypothesis is rejected. Conclude that there is statistically significant evidence of an association of family history of dementia with case status of dementia.

Command and Example	Notes
Immediate command: **mcci *aa bb cc dd*** **Example** **mcci 6 25 12 66**	**Use option level(#) to change confidence level.**

```
. mcci 6 25 12 66

                 | Controls                |
Cases            |   Exposed   Unexposed   |      Total
-----------------+-------------------------+-----------
         Exposed |         6          25   |         31
       Unexposed |        12          66   |         78
-----------------+-------------------------+-----------
           Total |        18          91   |        109

McNemar's chi2(1) =      4.57    Prob > chi2 = 0.0326
Exact McNemar significance probability       = 0.0470

Proportion with factor
        Cases       .2844037
        Controls    .1651376     [95% Conf. Interval]
                    ---------    --------------------
        difference  .1192661      .0030318   .2355003
        ratio       1.722222      1.039684   2.852836
        rel. diff.  .1428571      .0215646   .2641497

        odds ratio  2.083333      1.008624    4.55128  (exact - slightly different than my
                                                        approximation)
```

Interpretation is the same:
The McNemar test of association of exposure and disease in these matched data is statistically significant (Chi Square with df=1 = 4.57 or 3.89, depending on continuity correction; both p-values are < .05). The assumption of the null hypothesis of no association, when applied to the observed data, has led to a modestly unlikely event. The null hypothesis is rejected. Conclude that there is statistically significant evidence of an association of family history of dementia with case status of dementia..

3. Stratified Analysis of K 2x2 Tables

Illustrative Data for this Section

Dear class – in the next few pages, I show you how to input these tables into Stata (as tables!) and from there create a full data set of individual records. If you have trouble, no worries. On page 21, we use a previously created such stata dataset called coffeemi_full.dta. You can download the latter to your computer.

Stratum 1: FORMER SMOKER (smoking=1), n=157

Cups Coffee per day	MI (mi=1)	Control (mi=2)
≥ 5 (coffee=1)	7 (tally=7)	18
< 5 (coffee=2)	20	112

Stratum 2: 1-14 CIGARETTES/DAY (smoking=2), n=75

Cups Coffee per day	MI (mi=1)	Control (mi=2)
≥ 5 (coffee=1)	7	24
< 5 (coffee=2)	33	11

Stratum 3: 35-44 CIGARETTES/DAY (smoking=3), n=164

Cups Coffee per day	MI (mi=1)	Control (mi=2)
≥ 5 (coffee=1)	27	24
< 5 (coffee=2)	55	58

Stratum 4: 45+ CIGARETTES/DAY (smoking=4), n=98

Cups Coffee per day	MI (mi=1)	Control (mi=2)
≥ 5 (coffee=1)	30	17
< 5 (coffee=2)	34	17

Preliminary - Do these preliminary steps to learn how to create a Stata data set for the 4 2x2 tables shown on the previous page.

Step 1 – Create a coding manual

Tip **– Use "lower case" only variable names.**

Example -

Variable	Variable Label	Format	Format code definitions
smoking	Smoking Status	smokingf	1 = Former smoker 2 = 1-14 cigs/day 3 = 35-44 cigs/day 4 = 45+ cigs day
coffee	Cups Coffee Per Day	coffeef	1 = 5+ cups/day 2 = Less
mi	Myocardial Infarction	mif	1 = MI 2 = Non-MI
tally	Frequency (weight)	-	0, 1, 2, ….

Step 2 – In Stata, define variables and initialize to missing.

Example -

```
. generate smoking=.
. generate coffee=.
. generate mi=.
. generate tally=.
```

Step 3 – From the top menu, click on the data editor icon.

Example –You should now see an empty spreadsheet with the variable names you have chosen.

Step 4 – Enter your tabular data. Then close the data editor window (again, no worries - nothing lost).
Example – When you're done, you should have the following.

tally[17]

	smoking	coffee	mi	tally
1	1	1	1	7
2	1	1	2	18
3	1	2	1	20
4	1	2	2	112
5	2	1	1	7
6	2	1	2	24
7	2	2	1	33
8	2	2	2	11
9	3	1	1	27
10	3	1	2	24
11	3	2	1	55
12	3	2	2	58
13	4	1	1	30
14	4	1	2	17
15	4	2	1	34
16	4	2	2	17

Step 5 – In the command window, assign variable names, create value labels, and assign value labels.
Example –

```
. label variable smoking "Stratum of Smoking"
. label variable coffee "Cups of Coffee/Day"
. label variable mi "MI - Myocardial Infarction"
. label define smokingf 1 "Former Smoker" 2 "1-4 cigs/day" 3 "35-44 cigs/day" 4 "45+ cigs/day"
. label define mif 1 "MI" 2 "Non-MI"
. label define coffeef 1 "5+cups" 2 "Less"
. label values smoking smokingf
. label values coffee coffeef
. label values mi mif
```

Step 6 (optional) – Save a Stata version of your tabular data
Example –

```
. save "/Users/cbigelow/Desktop/coffeemi_tabular.dta"
```

Step 7– Create full data set of individual records using command expand *variablename* where *variablename* is the variable that contains the cell frequencies.

Example – Here, Stata tells me that it has created 478 *additional* observations.

```
. expand tally
(478 observations created)
```

CHECK that all is well:
The sum of the table totals on page 17 is (157+75+164+98) = 494
16 observations (tabular) + 478 additional observations = sum of table totals (157+75+164+98) = 494, match!.

Step 8 (optional) – Drop the variable that keeps track of the cell frequencies

Example –

```
. drop tally
```

Step 9 – Save your data.

Example –

```
.
. save "/Users/cbigelow/Desktop/coffeemi_full.dta"
file /Users/cbigelow/Desktop/coffeemi_full.dta saved
```

3.1 Descriptives - Numerical

Illustrative Data for this Section is coffeemi_full.dta and you may have just created it.
You can also download it from the course website.

Stratum 1: FORMER SMOKER (smoking=1), n=157

Cups Coffee per day	MI (mi=1)	Control (mi=2)
≥ 5 (coffee=1)	7 (tally=7)	18
< 5 (coffee=2)	20	112

Stratum 2: 1-14 CIGARETTES/DAY (smoking=2), n=75

Cups Coffee per day	MI (mi=1)	Control (mi=2)
≥ 5 (coffee=1)	7	24
< 5 (coffee=2)	33	11

Stratum 3: 35-44 CIGARETTES/DAY (smoking=3), n=164

Cups Coffee per day	MI (mi=1)	Control (mi=2)
≥ 5 (coffee=1)	27	24
< 5 (coffee=2)	55	58

Stratum 4: 45+ CIGARETTES/DAY (smoking=4), n=98

Cups Coffee per day	MI (mi=1)	Control (mi=2)
≥ 5 (coffee=1)	30	17
< 5 (coffee=2)	34	17

Command and Examples	Notes
Overall cross tab of exposure x disease, collapsing over strata: **tab2 *exposurevar diseasevar*** **Example** **tab2 coffee mi, row column**	**Use options row and column to get row and column perecentages.** **There are other options too..**
Cross tab of exposure x disease, separately for each stratum: **sort *stratumvar*** **by *stratumvar*: tab2 *exposurevar diseasevar*** **Example** **sort smoking** **by smoking: tab2 coffee mi, row column**	**This will produce lots and lots of output, depending. Consider using the tabulate command described below.**

NIFTY! – Here is a way to obtain a much more compact display of descriptives, separately for each stratum. Required – In order for this to work, the exposure and disease variables must be 0/1
 	
tabulate *stratumvar exposurevar01*, summarize(*diseasevar01*) means
 	
Example
`. generate mi01=mi`
`. replace mi01=0 if mi==2`
`. generate coffee01=coffee`
`. replace coffee01=0 if coffee==2`
`. label variable coffee01 "Cups of Coffee/Day"`
`. label define coffeef2 0 "Less" 1 "5+cups"`
`. label values coffee01 coffeef2`
`. tabulate smoking coffee01, summarize(mi01) means`	

```
. tabulate smoking coffee01, summarize(mi01) means
                           Means of mi01

Stratum of | Cups of Coffee/Day
   Smoking |      Less     5+cups |     Total
-----------+----------------------+----------
 Former Sm | .15151515        .28 | .17197452
 1-4 cigs/ |       .75  .22580645 | .53333333
 35-44 cig | .48672566  .52941176 |        .5
 45+ cigs/ | .66666667  .63829787 | .65306122
-----------+----------------------+----------
     Total | .41764706  .46103896 | .43117409
```

Key

It works!! Because the variable mi01 that we created is coded as 0=NON MI and 1=MI, the value of the sample mean of mi01 will be equal to the % who experience MI. Thus, we see the following:
(1) Overall, 43% experienced an MI
(2) Among former smokers whose coffee consumption is "LESS", 15% experienced an MI
Etc.

3.2 Descriptives - Graphical

Stata also offers several graphical options. Two are shown here. One is a **bar graph**, which is often used but not always a great choice. The second is a plot of the **odds ratios**, together with their 95% confidence limits.

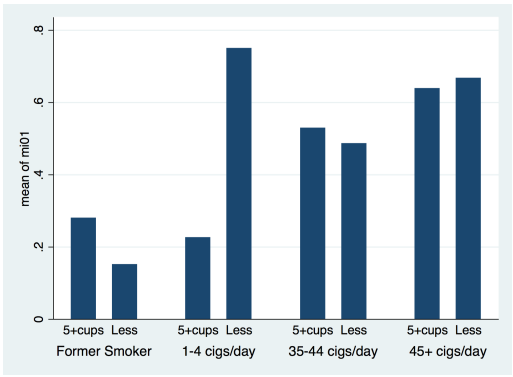
3.2.a Bar Graph

Required – The following graph requires that your column variable (outcome) be coded 0/1.

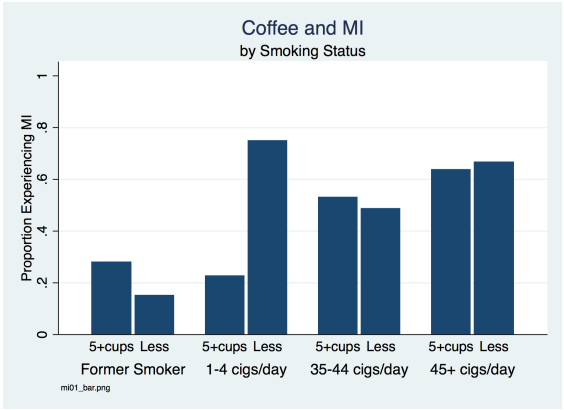
Goal -
Display the % experiencing outcome, over exposure, separately for each stratum

. * Command is **graph bar** with options **over** and **over**
. * graph bar ***diseasevar***, over(***exposurevar***) over(***stratavar***)

```
. ** Example – No frills
. graph bar mi01, over(coffee) over(smoking)
```




```
. **  Example – this time with lots of aesthetics
. graph bar mi01, over(coffee, gap(10)) over(smoking, gap(80)) outergap(50) ytitle("Proportion
Experiencing MI") title("Coffee and MI") subtitle("by Smoking Status") ylabel(0(.2)1)
caption("mi01_bar.png", size(vsmall))
```



3.2.b Odds Ratios and 95% CI

Be forewarned - This requires a number of steps. It involves creating a little data set that contains the values that you want plotted (the OR estimates and the lower and upper CI limits)

Goal -
Display the odds ratio OR, with 95% CI – separately for each stratum

. * Command is **graph twoway (scatter *ORvar exposurevar*) (rcap *lowerCI upperCI exposurevar*)**.
. * graph twoway (scatter ***dORvar exposurevar)*** *(*rcap ***lowerCI upperCi exposurevar)***

Step 1 – Obtain stratum specific OR and 95% CI limits. Command is mh with option by( )

Example -

```
. mhodds mi01 coffee01, by(smoking)

Maximum likelihood estimate of the odds ratio
Comparing coffee01==1 vs. coffee01==0
by smoking

------------------------------------------------------------------------------
  smoking | Odds Ratio        chi2(1)         P>chi2       [95% Conf. Interval]
----------+-------------------------------------------------------------------
 Former S |   2.177778           2.42         0.1197         0.79694    5.95115
 1-4 cigs |   0.097222          19.81         0.0000         0.02717    0.34791
 35-44 ci |   1.186364           0.25         0.6139         0.61031    2.30612
 45+ cigs |   0.882353           0.09         0.7693         0.38203    2.03793
------------------------------------------------------------------------------


--- some output omitted --
```

Step 2 – Create a new "little" data set containing the information to be plotted, taking care to have saved your full data set first.

```
. clear
. generate or=.
. generate high=.
. generate low=.
. generate smoking=.
```

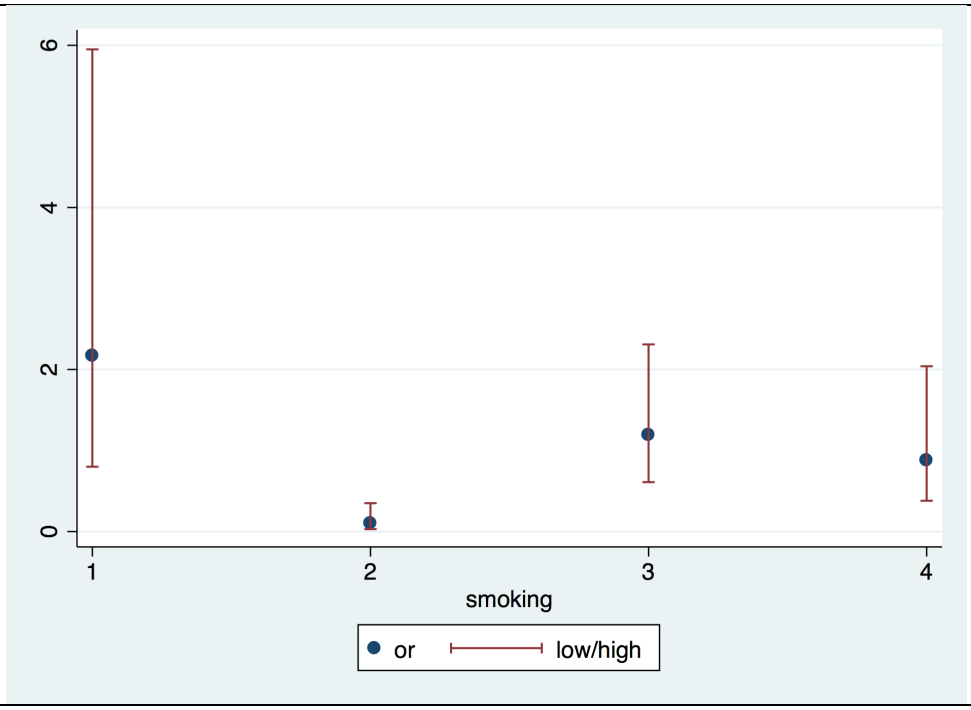
```
. *  In the DATA EDITOR window, enter the values needed for plotting.  These are in the
output from the command mhodds mi01 coffee01, by(smoking).  See previous page.
```

or	high	low	smoking
2.17	5.95	.8	1
.1	.35	.03	2
1.19	2.31	.61	3
.88	2.04	.38	4

Step 2 – Graph the odds ratios ± 95% confident limits

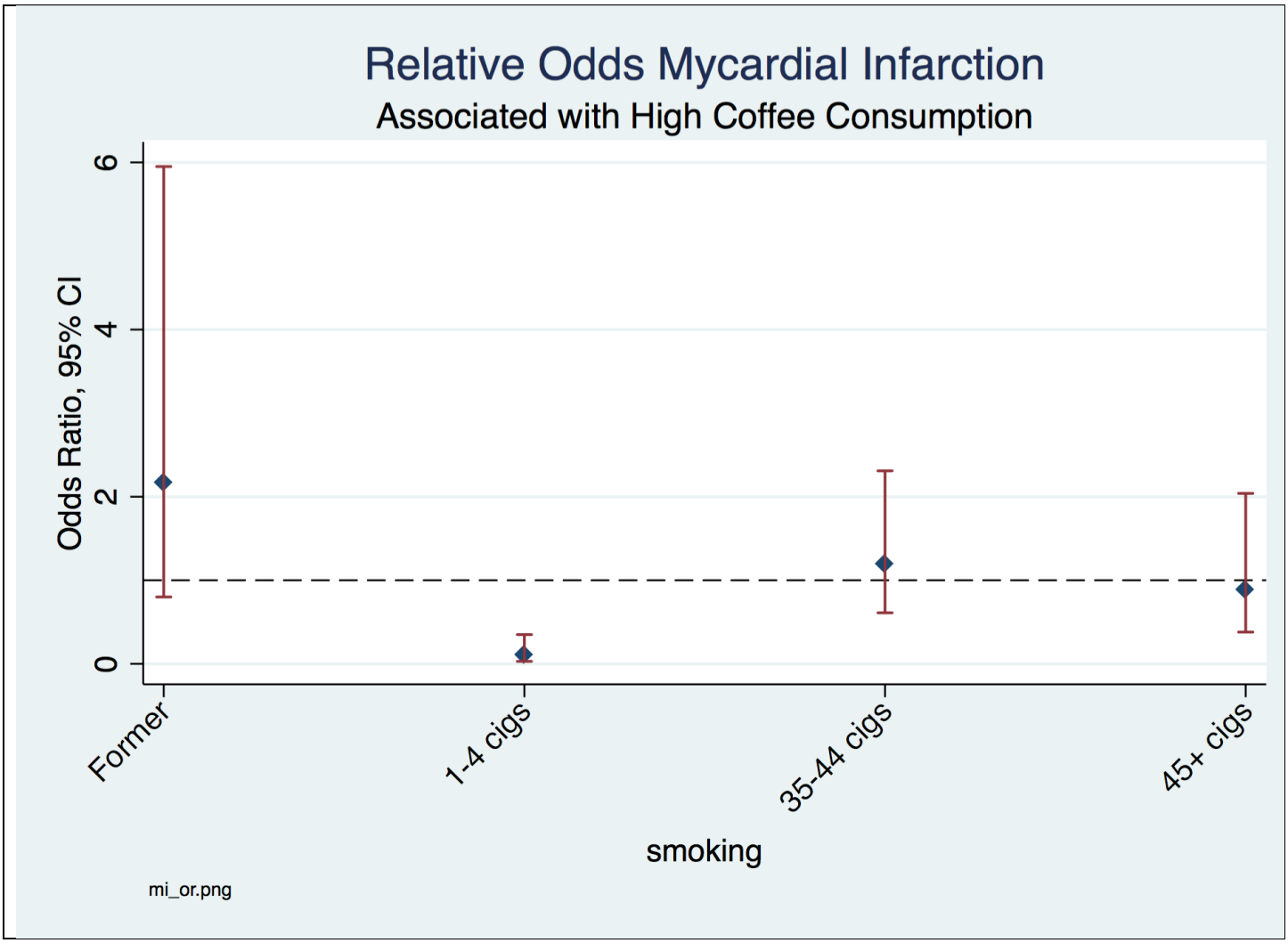
Example – No frills. As you'll see, it needs some aesthetics!

```
. graph twoway (scatter or smoking) (rcap low high smoking)
```



Example – With Aesthetics!

```
. graph twoway (scatter or smoking, msymbol(d)) (rcap low high smoking), yline(1,lwidth(thin)
lpattern(dash) lcolor(black)) xlabel(1 "Former" 2 "1-4 cigs" 3 "35-44 cigs" 4 "45+ cigs",
angle(45)) title("Relative Odds Mycardial Infarction") subtitle("Associated with High Coffee
Consumption") ytitle("Odds Ratio, 95% CI") legend(off) caption("mi_or.png", size(vsmall))
```



3.3. Mantel-Haenszel Test of Null: Homogeneity of the Odds Ratio

The command **cc** , will produce the results of both the Mantel-Haenszel **test of homogeneity of odds ratios** and the Mantel-Haenszel **test of the common odds ratio =1.** Recall - "cc" stands for "case-control"

Command and Examples	Notes
sort *stratavar* **cc *diseasevar01 exposurevar01,* by*(stratavar)*** **Example** **cc mi01 coffee01, by(smoking)**	**Required -** **The exposure and disease variables must be 0/1**

Example – continued.

```
. cc mi01 coffee01, by(smoking)
Stratum of Smoki |       OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
   Former Smoker |   2.177778      .6752163   6.360992     2.292994 (exact)
    1-4 cigs/day |   .0972222      .0281342   .3212659        10.56 (exact)
  35-44 cigs/day |   1.186364      .5805495   2.429507      8.04878 (exact)
    45+ cigs/day |   .8823529       .353352   2.204447     5.897959 (exact)
-----------------+-------------------------------------------------
           Crude |   1.192771      .7976463   1.781013              (exact)
    M-H combined |   .7751256      .5172801   1.161498
-----------------+-------------------------------------------------
Test of homogeneity (M-H)       chi2(3) =    19.92  Pr>chi2 = 0.0002

                  Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =      1.65
                                               Pr>chi2 =    0.1992
```

Interpretation:

The Mantel Haenszel test of homogeneity of odds ratio is statistically significant (Chi Square with df=3 = 19.92, p-value = .0002). The assumption of the null hypothesis of no association, 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.

3.4. Mantel-Haenszel Test of Null: Common Odds Ratio = 1

The command **cc** , will produce the results of both the Mantel-Haenszel **test of homogeneity of odds ratios** and the Mantel-Haenszel **test of the common odds ratio =1.** **Recall - "cc" stands for "case-control"**

Command and Examples	Notes
sort *stratavar* **cc *diseasevar01 exposurevar01*, by*(stratavar)*** **Example** **cc mi01 coffee01, by(smoking)**	**Required -** **The exposure and disease variables must be 0/1**

Example – continued.

```
. cc mi01 coffee01, by(smoking)

Stratum of Smoki |       OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
   Former Smoker |   2.177778      .6752163   6.360992     2.292994 (exact)
    1-4 cigs/day |   .0972222      .0281342   .3212659        10.56 (exact)
  35-44 cigs/day |   1.186364      .5805495   2.429507      8.04878 (exact)
    45+ cigs/day |   .8823529       .353352   2.204447     5.897959 (exact)
-----------------+-------------------------------------------------
           Crude |   1.192771      .7976463   1.781013              (exact)
    M-H combined |   .7751256      .5172801   1.161498
-----------------+-------------------------------------------------
Test of homogeneity (M-H)       chi2(3) =    19.92  Pr>chi2 = 0.0002

                    Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      1.65
                                                Pr>chi2 =    0.1992
```

Interpretation:

Beware!!! 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 association of high coffee consumption with event of myocardial infarction Chi Square on df=1 = 1.65, p-value = .1992).

4. 2xC Table Analysis of Trend

Illustrative Data for this Section is esophageal_cancer_full.dta
You can download it from the course website. Following is a tabular summary

	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

Description: Data are excerpted 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.

Source: Tuyns AJ, Pequignot G and Jenson OM (1977) Le cancer de l'oesophage en Ille-et-Villaine en function des niveaux de consummation d'alcool et de tabac. *Bull Cancer* 64: 45-60.

4.1 Descriptives - Numerical

Frequencies and Percentages

Commands and Examples	Notes
tab2 *diseasevar dosevar* **Example** **tab2 case alcohol, row column**	**Tip – Recommendation.** **Because this is an analysis of trend, in your tab2 command, specify the disease variable first.** **Use options row and column to get row and column perecentages.**
tabulate *diseasevar dosevar* **Example** **tabulate case alcohol, row column**	
table *dosevar*, contents(mean *diseasevar01* sd *diseasevar01*) **Example** **table alcohol, contents(mean case sd case)**	**Required – The disease variable must be 0/1**

Odds

Commands and Examples	Notes
tabodds *diseasevar dosevar* **Example** **tabodds case alcohol**	

Relative Odds (OR) with Increasing Dose and Associated Hypothesis Tests

Commands and Examples	Notes
tabodds *diseasevar dosevar*, or **Example** **tabodds case alcohol, or**	**Two hypothesis tests** **1. Global test of homogeneity of odds** **2. Test of trend**

4.2 Descriptives - Graphical

Again, Stata also offers several graphical options. Two are shown here. The first requires a few steps.

4.2.a. Mean Percent Event ± 95% CI, Over Dose

Goal -
Display the % with disease in relationship to increasing dose of exposure.

. * Command is **graph twoway (scatter *diseasevar dosevar*) (rcap *lowerCI upperCI sdosevar*)**.
. * graph twoway (scatter ***diseasevar dosevar)*** **(**rcap ***lowerCI upperCi dosevar)***

Step 1 – Use the command collapse to obtain the mean and Sem of the disease variable, separately for each stratum defined by dose.

Example -

```
. collapse (mean) case (sem) semcase=case, by(alcohol)
. list

     +-------------------------------+
     | alcohol       case    semcase |
     |-------------------------------|
  1. |   0-39g   .0408163   .0285593 |
  2. |   40-79       .225   .0668667 |
  3. | 80-119g         .5   .1212678 |
  4. |    120+         .5   .1666667 |
  5. |       .          .          . |
     +-------------------------------+

. drop if alcohol==.
(1 observation deleted)
```

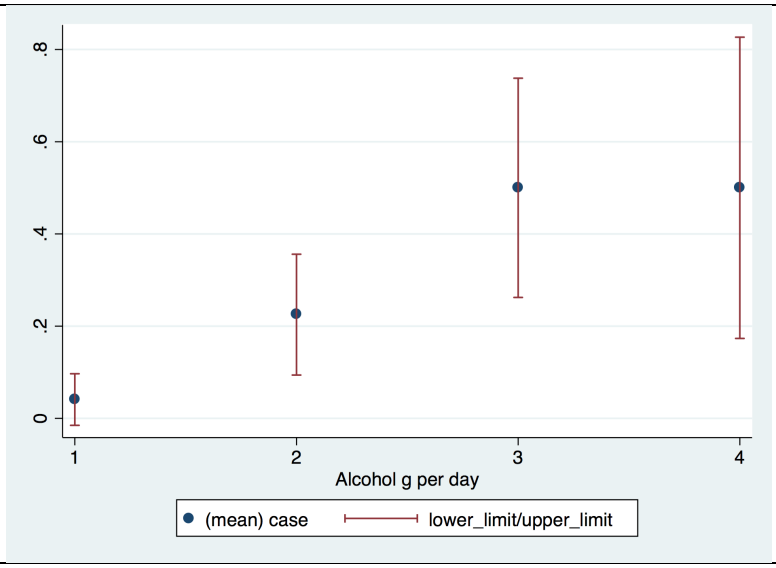
Step 2 – Obtain values of upper and lower 95% CI using the new variables Stata created in collapse
Example –

```
. generate lower_limit=case-1.96*semcase
. generate upper_limit=case+1.96*semcase
```

Step 3 – Graph % Disease ± 95% confident limits

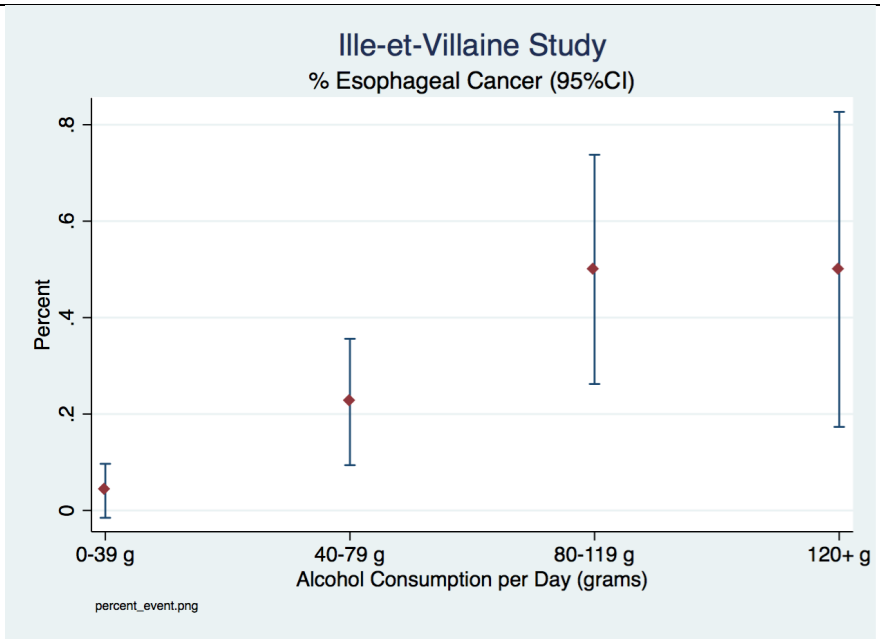
Example – Basic. Again. Needs some aesthetics!

```
. graph twoway (scatter case alcohol) (rcap lower_limit upper_limit smoking)
```



Example – This time with aesthetics

```
. graph twoway (rcap lower_limit upper_limit alcohol) (scatter case alcohol, msymbol(d)), xlabel(1
"0-39 g" 2 "40-79 g" 3 "80-119 g" 4 "120+ g") xtitle("Alcohol Consumption per Day (grams)")
ytitle("Percent") title("Ille-et-Villaine Study") subtitle("% Esophageal Cancer (95%CI)")
legend(off) caption("percent_event.png", size(vsmall))
```



4.2.b. Odds of Event ± 95% CI, Over Dose

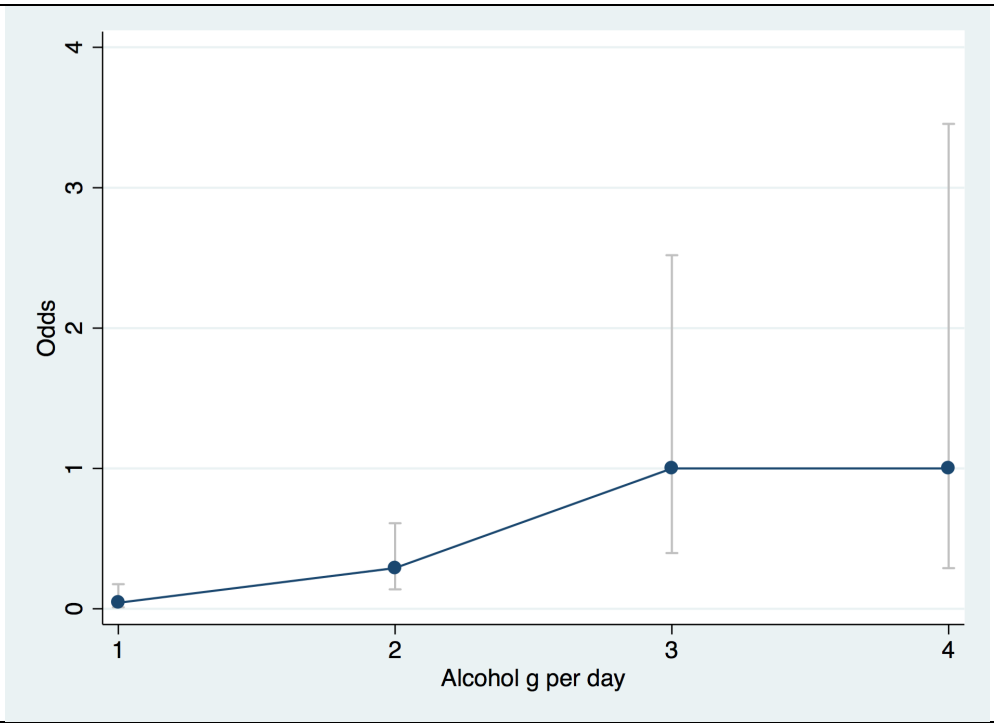
Dear class – to be honest, in this situation, I'm more likely to want to plot the relative odds (OR) over dose. So I would then follow the steps described on pp 24-26.

Goal -
Display the estimated odds of event of disease in relationship to increasing dose of exposure.

. * Command is **tabodds** **with option** **ciplot**.
. * tabodds ***diseasevar dosevar***, ciplot

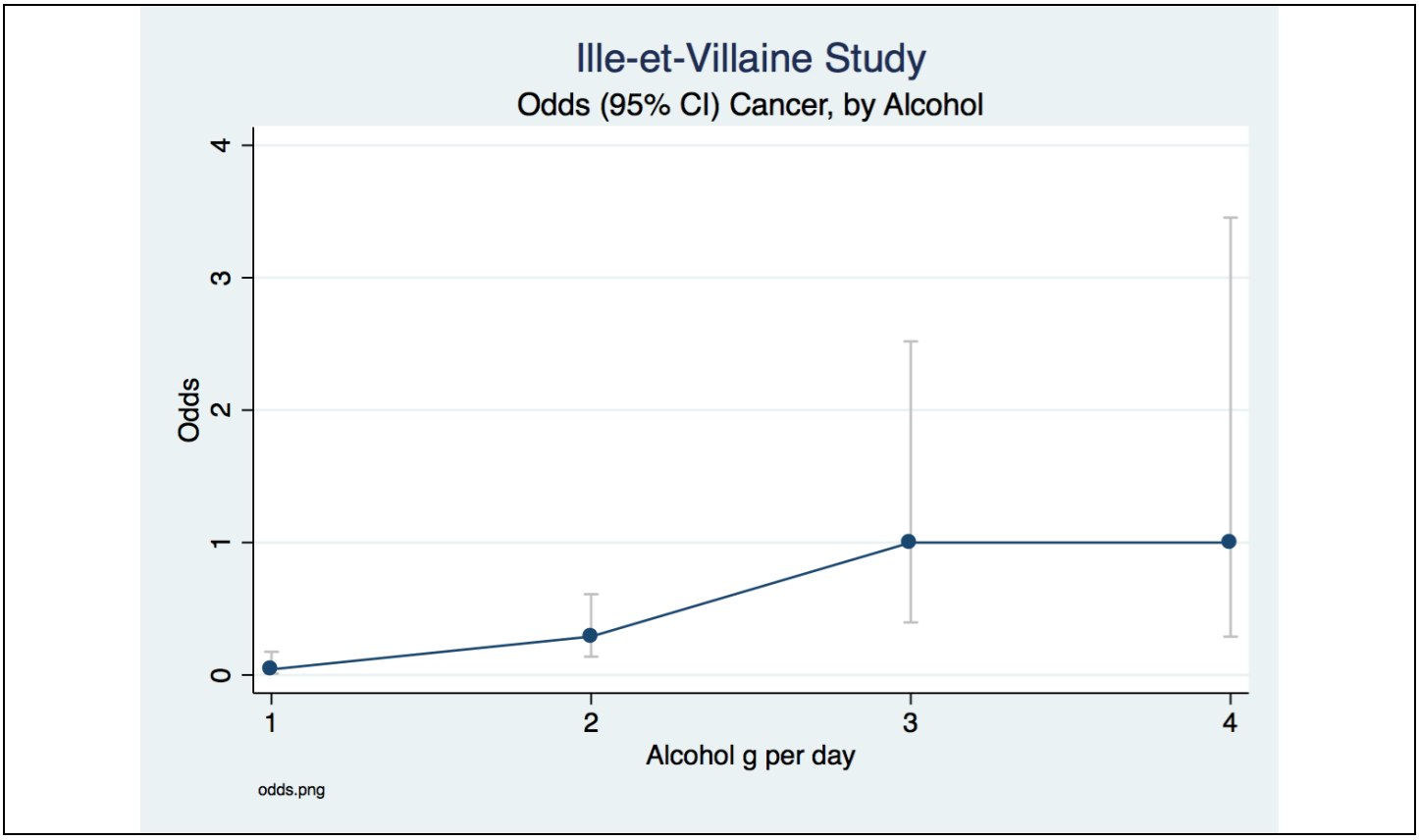
Example – No frills. Again. Needs some aesthetics!

```
. * ALERT!  If you've been following along, then you will need to reload the full data.
. clear
. use "/Users/cbigelow/Desktop/esophageal_cancer_full.dta"
. tabodds case alcohol, ciplot
```



Example – With aesthetics

```
. tabodds case alcohol, ciplot title("Ille-et-Villaine Study") subtitle("Odds (95% CI) Cancer, by
Alcohol")caption("odds.png", size(vsmall))
```



4.2.c Odds Ratios and 95% CI

Note - The stata commands are very similar to what is described on pp 24-26.

Goal -
Display the odds ratio OR, with 95% CI – separately for each increasing level of dose of exposure (alcohol consumption). Use as the referent group alcohol=1 (0-33 g)

. * Command is **graph twoway (scatter *ORvar exposurevar*) (rcap *lowerCI upperCI exposurevar*)**.
. * graph twoway (scatter ***dORvar exposurevar)*** *(*rcap ***lowerCI upperCi exposurevar)***

Step 1 – Obtain stratum specific OR and 95% CI limits. Command is tabodds with option or
Example -

```
. tabodds case alcohol, or

-------------------------------------------------------------------------------
     alcohol |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+-----------------------------------------------------------------
       0-39g |    1.000000          .          .            .           .
       40-79 |    6.822581       6.82     0.0090     1.280402   36.353890
     80-119g |   23.500000      19.93     0.0000     3.061934  180.359887
        120+ |   23.500000      16.46     0.0000     2.496605  221.200386
-------------------------------------------------------------------------------

--- some output omitted --
```

Step 2 – Create a new "little" data set containing the information to be plotted, taking care to have saved your full data set first.

```
. clear
. generate or=.
. generate high=.
. generate low=.
. generate alcohol=.
```

```
. *  In the DATA EDITOR window, enter the values needed for plotting.  These are in the
output from the command tabodds case alcohol, or.  See previous page.
```

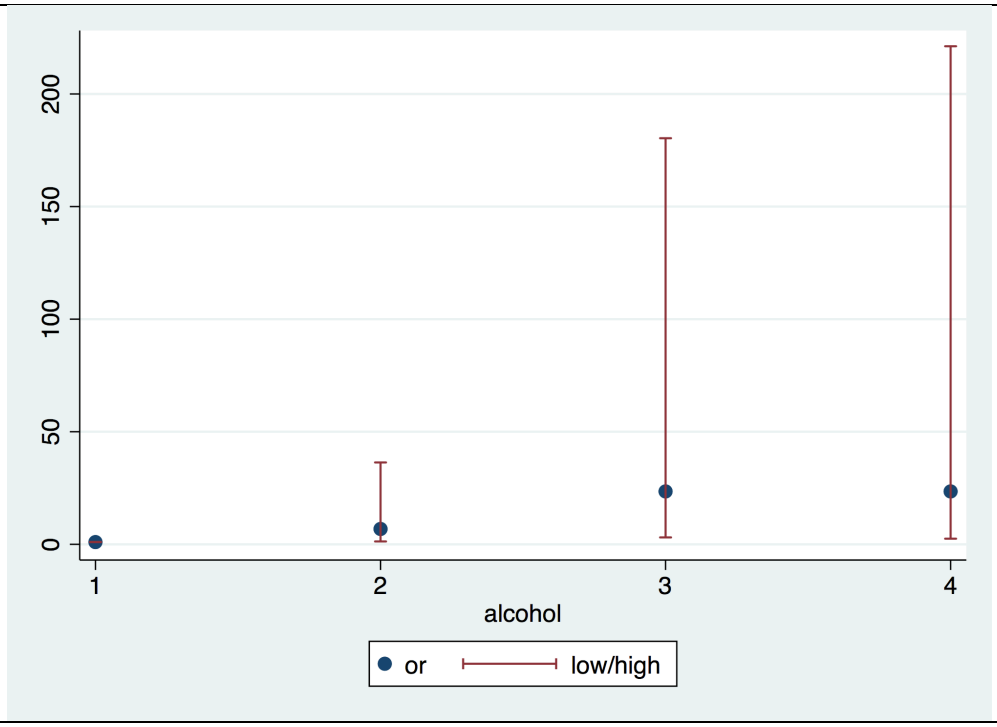
var11[15]

	or	high	low	alcohol
1	1	1	1	1
2	6.8226	36.3539	1.2804	2
3	23.5	180.3599	3.0619	3
4	23.5	221.2004	2.4966	4

Step 2 – Graph the odds ratios ± 95% confident limits

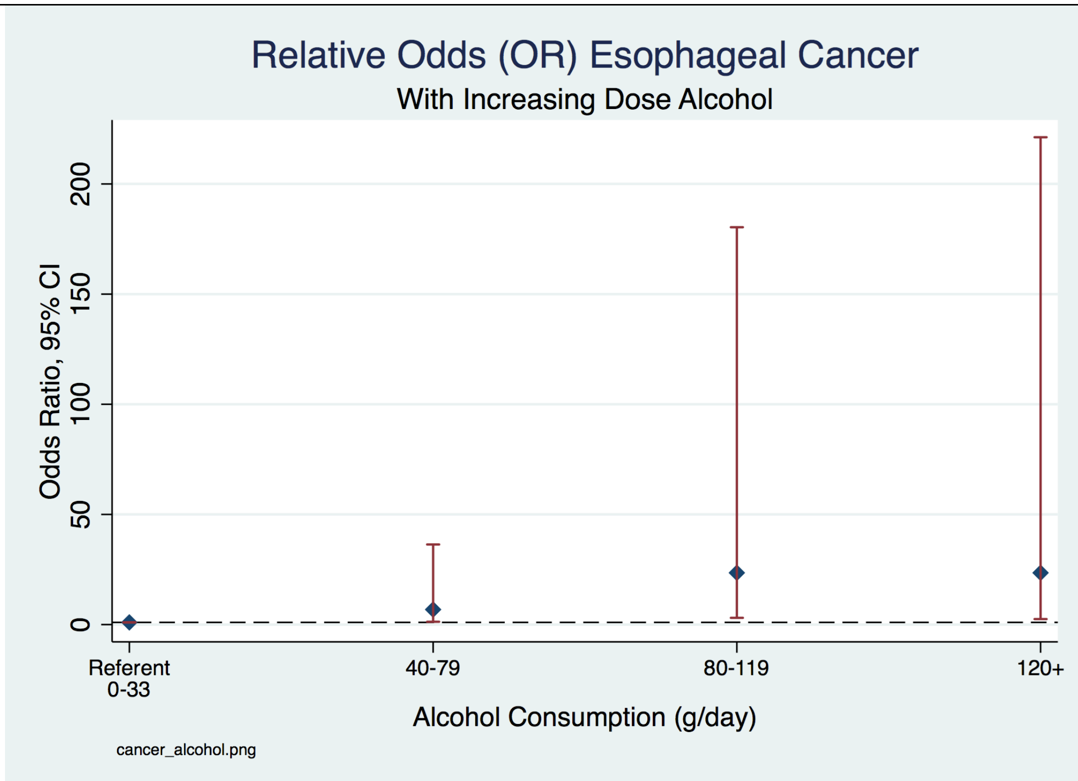
Example – No frills.

```
. graph twoway (scatter or alcohol) (rcap low high alcohol)
```



Example – With Aesthetics!

```
. graph twoway (scatter or alcohol, msymbol(d)) (rcap low high alcohol), yline(1,lwidth(thin)
lpattern(dash) lcolor(black)) xlabel(1 `" "Referent" "0-33" "' 2 "40-79" 3 "80-119" 4
"120+",labsize(small)) title("Relative Odds (OR) Esophageal Cancer") subtitle("With Increasing Dose
Alcohol") ytitle("Odds Ratio, 95% CI") xtitle("Alcohol Consumption (g/day)") legend(off)
caption("cancer_alcohol.png", size(vsmall))
```



HACK!! **HOW TO SPLIT X-Axis labels Over Multiple Lines – In the graph above, I split the first label on the x-axis so that it would be on two lines. Take a look at my xlabel( ) option; here it is:**

```
xlabel(1 `" "Referent" "0-33" "' 2 "40-79" 3 "80-119" 4 "120+",labsize(small))
```

The label for the value "1" is split over two lines. To do this in Stata, each line needs to be in its own set of quotes and then the entire label for the value of "1" must be enclosed in yet another set of boundaries. These are shown in red below. Take care with the single quotes. The opening one is a forward single quote and closing one is a closing single quote.

```
        value   `"   "firstline"   "secondline"      "'

e.g      1     `"    "Referent"     "0-33"           "'
```

4.3 Chi Square tests of General Association

Null Hypothesis: No association of disease with dose

Preliminary – If you don't already have it, download and install the module `tab_chi`

```
. ssc install tab_chi
```

Recommended: Using command tabodds

Commands and Examples	Notes
tabodds *diseasevar dosevar*, or **Example** **tabodds case alcohol, or**	**Tabodds actually provides 2 hypothesis tests** **1. Test of General Association (df = C-1)** **2. Test of Trend (df = 1)**

Using command tabchi

Commands and Examples	Notes
tabchi *diseasevar dosevar*, pearson cont **Example** **tabchi case alcohol, pearson cont**	**tabulate gives both a Pearson chi square test and a likelihood ratio test. They're close** **Key to options** **pearson = pearson residuals** **cont = contribution to chi square test** **= $(O-E)^2/E$**

Using command tabulate

Commands and Examples	Notes
tabulate *diseasevar dosevar*, col expected chi2 **Example** **tabulate case alcohol, col expected chi2**	**tabulate gives a Pearson chi square test with Degrees of freedom = (#rows-1)(#col-1)** **Example** **Df = (2-1)*(4-1) = 3**

4.4 Chi Square tests of Trend Using tabodds

Null Hypothesis: No association of disease with dose
Alternative Hypothesis: Linear trend in occurrence of disease in relationship to dose

Recommended: Using command tabodds

Commands and Examples	Notes
tabodds *diseasevar dosevar*, **or** **Example** **tabodds case alcohol, or**	**Tabodds actually provides 2 hypothesis tests** **1. Test of General Association (df = C-1)** **2. Test of Trend (df = 1)**

Example -

```
. tabodds case alcohol, or
```

```
------------------------------------------------------------------------------
     alcohol |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1. 0-39g |    1.000000          .          .               .           .
    2. 40-79 |    6.822581       6.82     0.0090        1.280402   36.353890
   3. 8~119g |   23.500000      19.93     0.0000        3.061934  180.359887
     4. 120+ |   23.500000      16.46     0.0000        2.496605  221.200386
------------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(3)  =    22.22          Test of General Association
                                  Pr>chi2  =   0.0001
                                                  <--

Score test for trend of odds:     chi2(1)  =    20.85
                                  Pr>chi2  =   0.0000          Test of Trend
                                                  <--
```

Interpretation:

Test of General Association The null hypothesis of no association is rejected.
These data suggest statistically significant evidence of a departure from independence.
But we don't learn much beyond that (we could look at the table of course)

Test of Trend The null hypothesis of no trend is rejcted.
These data suggest a trend in the odds of disease with increasing dose of alcohol
Examination of the table suggests that the trend is positive.

5. RxC Table Analysis of Trend Using nptrend

Null Hypothesis: No association of row and column variables
Alternative Hypothesis: Monotone association of row and column variables

Illustrative Data for this Section is esophageal_cancer_full.dta

	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

Which to Use? tabodds or nptrend? -

For data in a 2x2 table
Use either tabodds or nptrend. Recommendation: tabodds.

For data in a RxC table with MORE than 2 rows and MORE than 2 columns
Must use command nptrend

Commands and Examples	Notes
nptrend *diseasevar*, by*(dosevar)* score*(dosevar)*	**Required – Both row and column variables must be ordered**
Example **Nptrend case, by(alcohol) score(alcohol)**	**Note - nptrend does NOT allow tabular data; you must have expanded your data first (see again pp 10-13)**

```
. nptrend case, by(alcohol) score(alcohol)

   alcohol      score        obs      sum of ranks
         1          1         49            2395.5
         2          2         40            2386.5
         3          3         18            1363.5
         4          4         10             757.5

         z  =  4.57
  Prob > |z| = 0.000
```

Interpretation. Reject the null (p-value << .0001). Conclude there is statistically significant evidence of trend in the odds of esophageal cancer with increasing alcohol consumption.

6. Chi Square Goodness of Fit Test Using chitesti

Preliminary – If you don't already have it, download and install the module `tab_chi`

```
. ssc install tab_chi
```

Chitesti is an "immediate" command. User must supply the observed and expected frequencies, O and E.

Command	Notes
chitesti ***O# O# … O#\E# E# … E#,*** **nfit(*#parametersestimated*)**	

Example (Taken from BIOSTATS 640 Notes)
Source: Zar, JH. Biostatistical Analysis, third edition. Upper Saddle River: Prentice Hall, 1996 p. 461

A plant geneticist wishes to know if a sample of n=250 seedlings comes from a population with an **hypothesized 9:3:3:1 ratio** of yellow smooth: yellow wrinkled: green smooth: green wrinkled seeds. Thus, expected counts are computed using the hypothesized 9:3:3:1 phenotype ratios as shown below.

i	Phenotype	O_i	Expected Count, E_i	Component $\frac{(O_i - E_i)^2}{E_i}$
1	Yellow smooth	152	$(n)[Pr(phenotype)]=(n)\left[\frac{9}{9+3+3+1}\right]=(250)[.5625]=140.625$	0.9201
2	Yellow wrinkled	39	$(n)[Pr(phenotype)]=(n)\left[\frac{3}{9+3+3+1}\right]=(250)[.1875]=46.875$	1.3230
3	Green smooth	53	$(n)[Pr(phenotype)]=(n)\left[\frac{3}{9+3+3+1}\right]=(250)[.1875]=46.875$	0.8003
4	Green wrinkled	6	$(n)[Pr(phenotype)]=(n)\left[\frac{1}{9+3+3+1}\right]=(250)[.0625]=15.625$	5.9290
	TOTAL	**250**	**250**	**8.972**

Hand Calculation

Degrees of freedom = K(total # bins) - [1] - [# parameters estimated]

= [4] – [1] - [0, because we didn't have to estimate any!]

= 3

$\chi^2_{\text{goodness of fit; df=3}} = 8.972$

p-value = Prob [Chi square w df = 3 $\geq$ 8.972] = 0.02967

```
. chitesti 152 39 53 6\140.625 46.875 46.875 15.625,nfit(0)

observed frequencies from keyboard; expected frequencies from keyboard

         Pearson chi2(3) =   8.9724   Pr =  0.030
likelihood-ratio chi2(3) =  10.8325   Pr =  0.013

  +---------------------------------------+
  | observed   expected   obs - exp   Pearson |
  |---------------------------------------|
  |      152    140.625      11.375     0.959 |
  |       39     46.875      -7.875    -1.150 |
  |       53     46.875       6.125     0.895 |
  |        6     15.625      -9.625    -2.435 |
  +---------------------------------------+
```

Interpretation: The null hypothesis is rejected by both the Pearson and Likelihood Ratio tests. The distribution of phenotypes in this sample is statistically significantly different than what was hypothesized (9:3:3:1 of yellow smooth: yellow wrinkled: green smooth: green wrinkled.