Lecture 6: Epi methods in Stata

BIOSTAT 212, Summer 2016

Kristen Aiemjoy
Department of Epidemiology and Biostatistics
University of California San Francisco

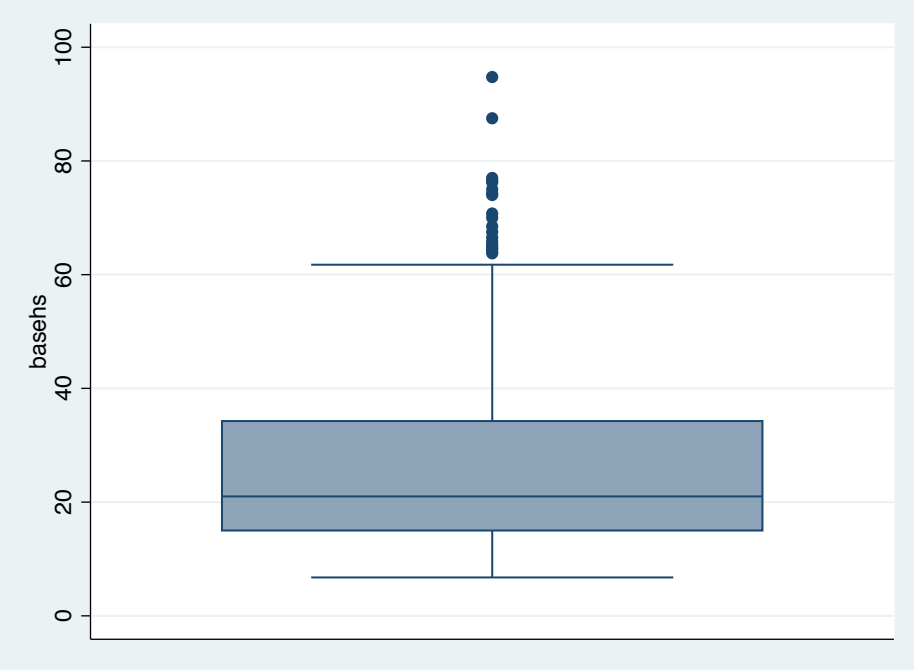
Review/ questions from last week?

P=0.000?

- P-value is a probability so it can't = 0
- Report as:
  - P<0.0005 [gives the most information]
    - P<0.001 [also ok]
    - P<0.01 [also ok]
- To get the exact p-value use the test statistic:
  - `display 1 - chi2(df, chi2)`
  - `display 2*normal(-abs(z))`
  - `display 2*ttail(df, abs(t))`

Outliers in `graph box`

- Whiskers are [lower quartile -1.5*IQR, upper quartile +1.5*IQ]
  - IQR: Inter-Quartile Range
  - 75% - 25%
- Any value greater than or less than whisker is an outlier

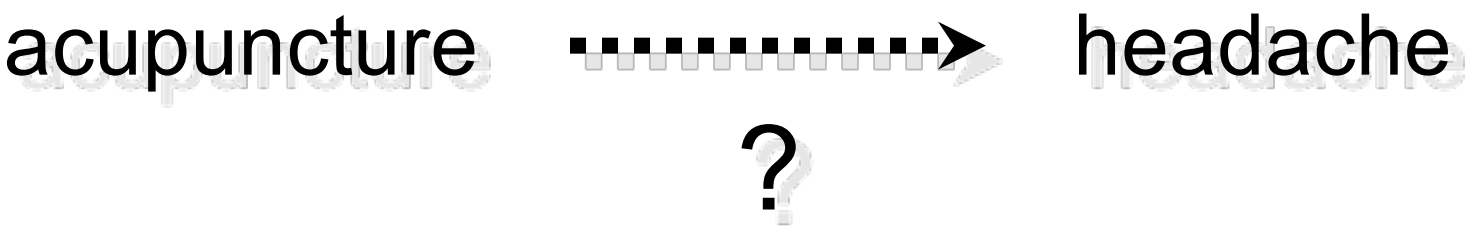


Outline for today

- Introduction to epi concepts
  - Exposure and Outcome
  - Confounding and Interaction
  - Measures of Association
- Estimating Measures of Association in Stata
  - `cs`
  - `cc`
  - `mhodds`
  - `tabodds`
  - Regression Models
    - `logistic`
    - `binreg`

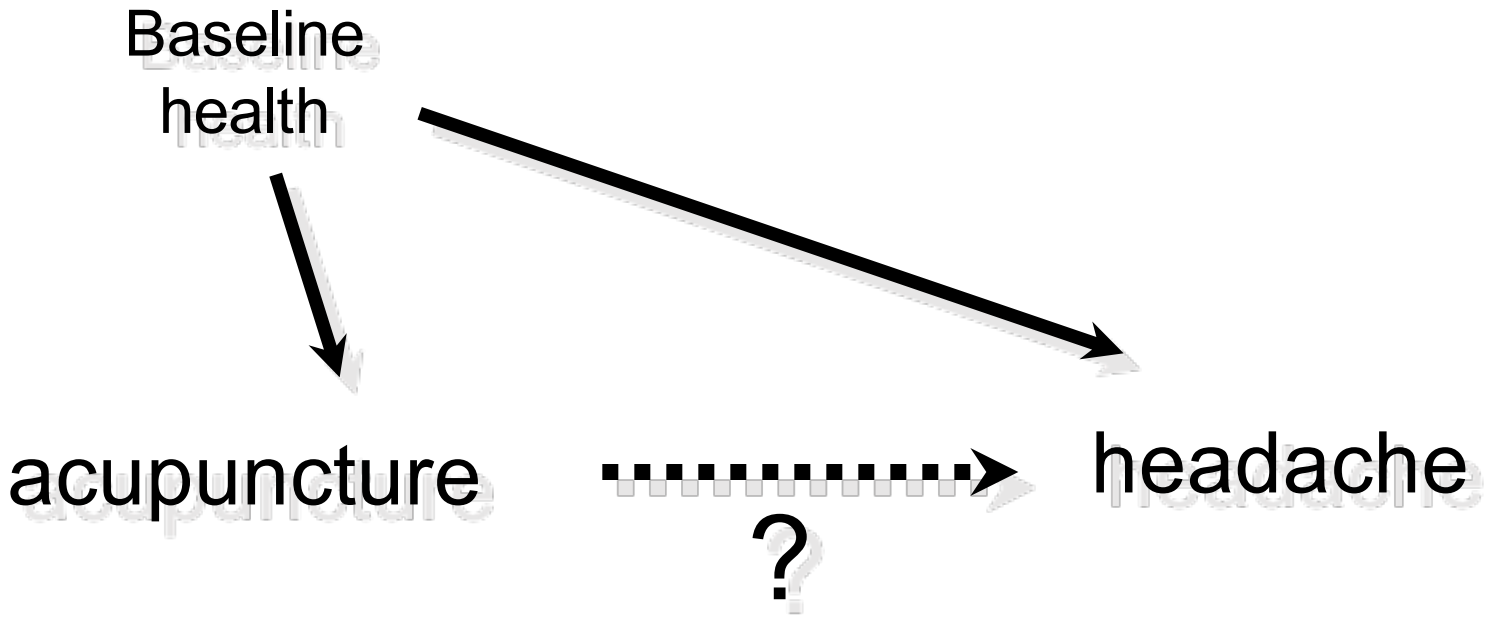
Exposures and Outcomes

- Does “exposure” cause “outcome”?



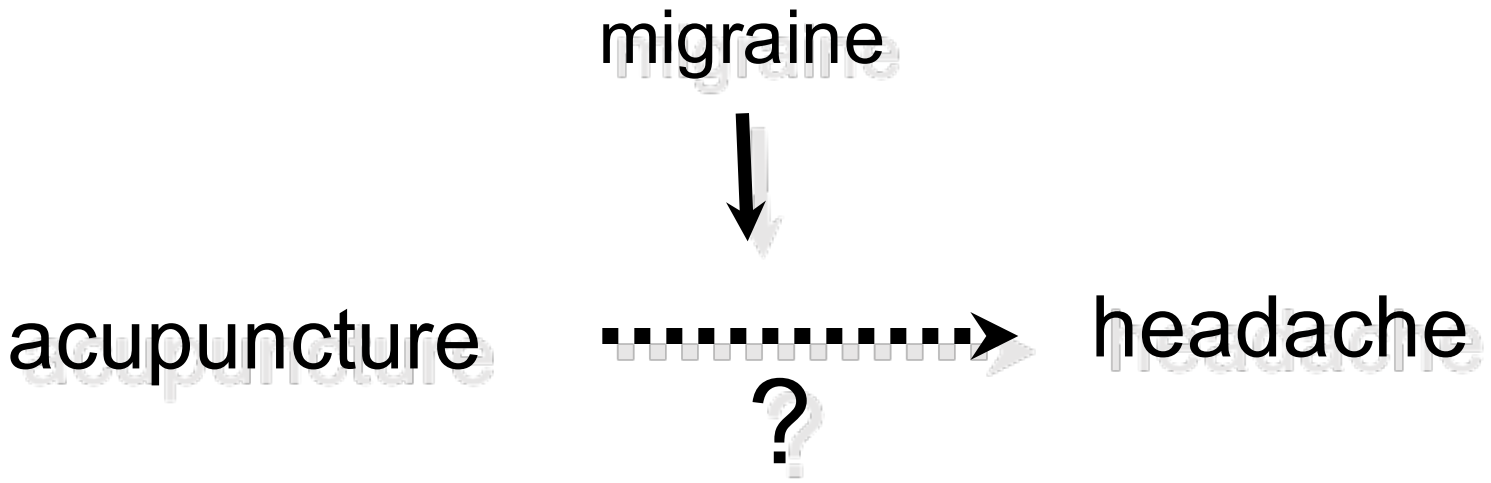
Confounding

- Do other variables mask or exaggerate the association between exposure and outcome?



Interaction

- Is the effect of exposure on outcome different across strata (categories) of a third variable?



Confounding vs Interaction

- **Confounding:** bad, you want to minimize it

- **Interaction:** not bad: you want to report it

Measures of Association

- An association between an exposure and outcome can be assessed by comparing the probability of outcome in the exposed group to the probability of outcome in unexposed group.

- **Ratio (division)**

  [Pr(outcome) in exposed]

  [Pr(outcome) in unexposed]

- **Difference (subtraction)**

  [Pr(outcome) in exposed] – [Pr(outcome) in unexposed]

Ratio Measures of Association

Study Design	Ratio Measure	Stata command
Cross-sectional study	Prevalence Ratio (PR)	`cs O E, by(C)` `logisitic O E C*`

Cohort study	Risk Ratio (RR)	`cs O E, by(C)`
		`binreg O E C, rr`
		`regress O E C*`
		`logistic O E C*`

	Rate Ratio (RR)	`ir O E T` `poisson O E C, irr`
	Hazard Ratio (HR)	`stcox O E C`

Case-control study	Odds Ratio (OR)	`cc O E, by(C)`
		`mhodds O E, by(C)`
		`logistic O E C, or`
		`tabodds O E, adjust(C)`

*needs additional lincom statement

`O = outcomevar E = exposurevar C = covariates T = time`

Difference Measures of Association

Study Design	Ratio Measure	Stata command

Cross-sectional study	Prevalence Difference	`cs O E`
		`logistic O E C*`
		`regress O E C`

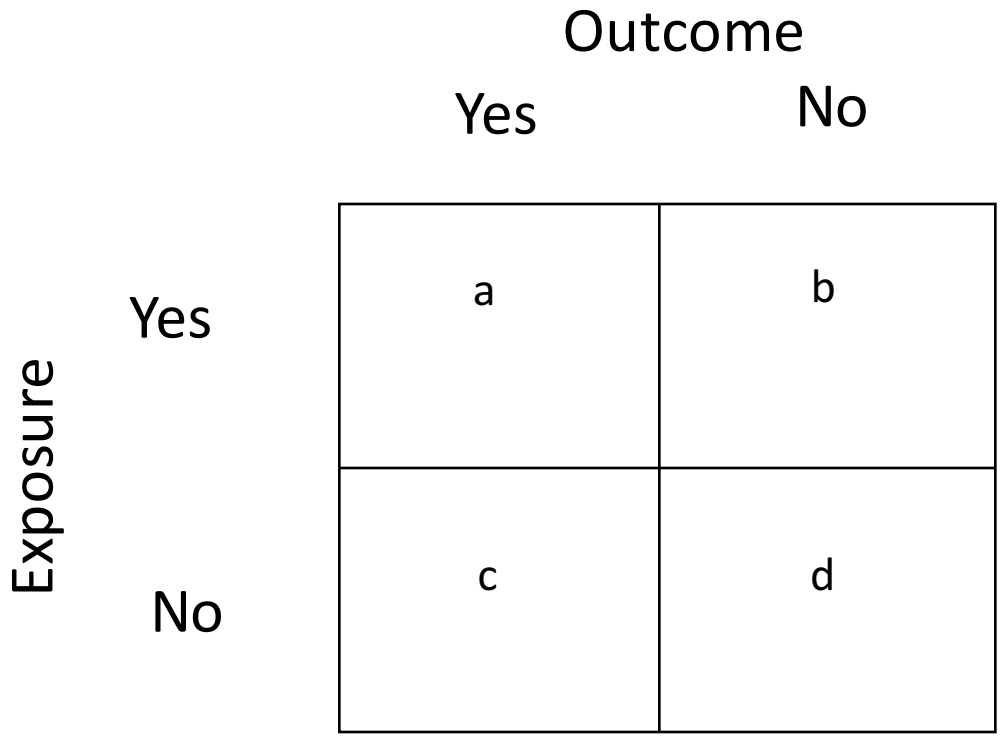
Cohort study	Risk Difference	`cs O E`
		`regress O E C`
		`binreg O E C, rd`

	Rate Difference	`ir O E T, ird` `poisson O E C*`
Case-control study	N/A	

`O = outcomevar E = exposurevar C = covariates T = time`

*needs additional lincom statement

Odds Ratio



Odds Ratio =

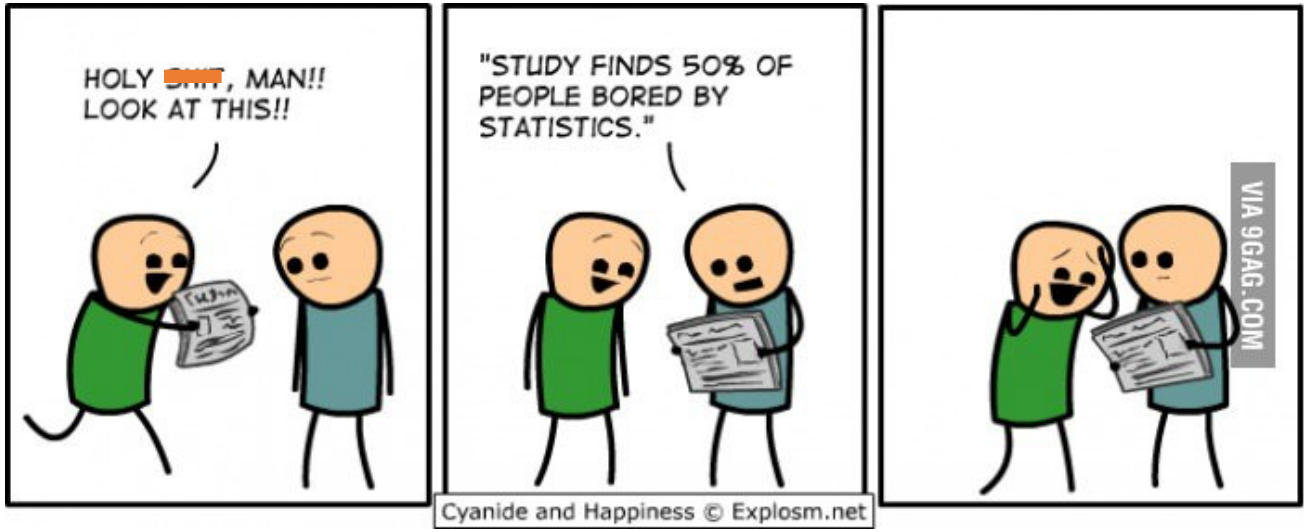
$$\frac{a/b}{c/d}$$

For cross-control studies

The odd thing about Odds Ratios….

- Odds are not-intuitive, difficult to communicate and interpret

- They are farther from the null than the risk ratio
  - exaggerate the association if interpreted as a probability



Odds Ratios: `cc`

`cc outcomevar predictorvar`

– calculates odds ratios

Options

`,by (varname)`	-stratify by a third variable [to look for interaction]
`[, missing]`	-include missing data as a category of the stratified variable

Limitations

Cannot adjust for more than 1 covariate

Practice

- Fill out cells of table

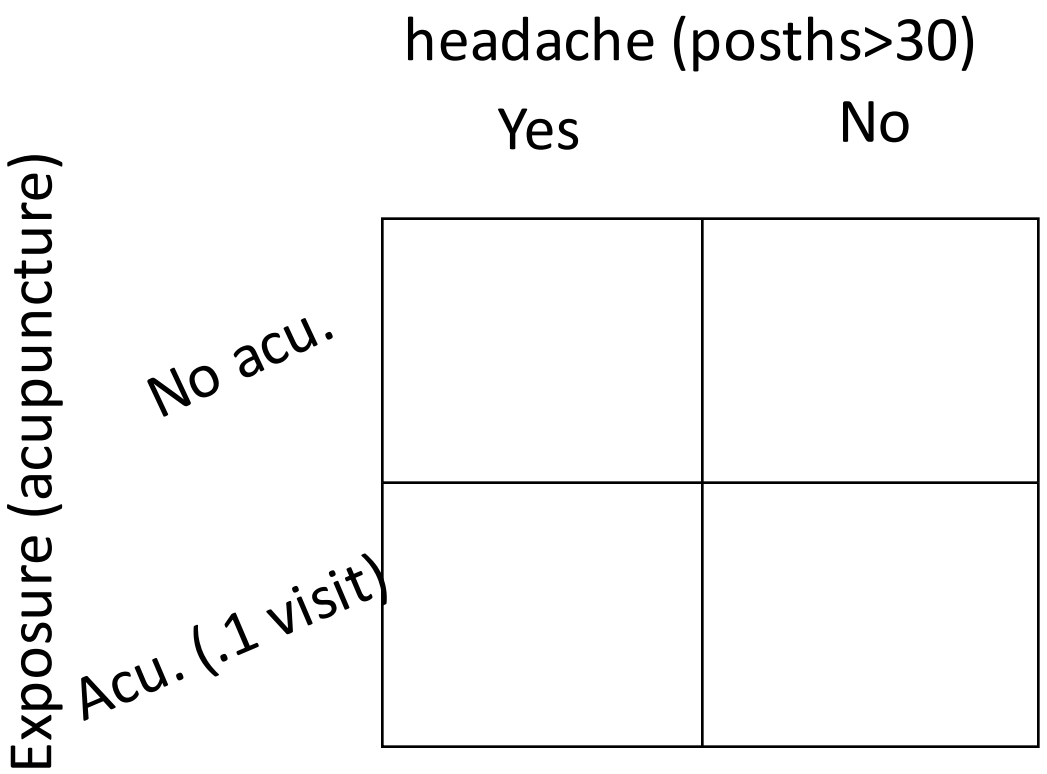
```
tab no_acu headache
```

- Calculate the OR manually using the di function

```
di (98/117)/(45/141)
```

- Compare OR with `cc` command

```
cc headache no_acu
```



Odds Ratios: `mhodds`

`mhodds outcomevar predictorvar [covariates]`

– calculates odds ratios and can adjust for covariates

Options

`,by (varname)`	-stratify by a third variable [to look for interaction]
`[, missing]`	-include missing data as a category of the stratified variable

Limitations

- Uses stratification to adjust for covariates
  - with several covariates, strata start to get “thin”

Practice

- What is the OR for the effect of acupuncture on headache?

```
mhodds headache no_acu
```

- What is the OR adjusted for baseline health score?

```
mhodds headache no_acu health_bl
```

Odds Ratios: `tabodds`

`tabodds outcomevar predictorvar, adjust[covariates]`

– calculates odds ratios & allows for more than 2 exposure categories (test for trend)

Options

`,adjust (varname)` adjust for covariates

Limitations

- Cannot look for interaction
- Uses stratification → strata start to get thin with many convariates/categories

Practice

- Does there appear to be a trend in the number of acupuncture treatments and headache?

```
tabodds headache acu_cat
```

What is a **regression**?

- Fit a mathematical equation to your data
- Where the probability of your outcome is a function of the exposure and other covariates
  - Pr(outcome) = B(exposure) + B(covariate1) +B(covariate2) +…
- Advantages
  - Allows for interaction and adjustment of many covariates
- Disadvantages
  - Requires more assumptions
  - Can be more difficult to interpret and communicate

Odds Ratios: `logistic`

`logistic outcomevar predictorvar covariates`

– multivariate logistic regression

Options

`,vce(robust)` -options for adjusting SEs

`##` -Specify an interaction between 2 variables

If using any categorical predictors with >2 categories, use `i.catvar`

Logistic regression annotated output

```
logistic headache no_acu health_bl

Logistic regression                               Number of obs[b]   =        400
                                                  LR chi2(2)[c]      =      31.89
                                                  Prob > chi2[d]     =     0.0000
Log likelihood =  -244.2483[a]                    Pseudo R2[e]       =     0.0613

------------------------------------------------------------------------------
  headache[f] | Odds Ratio[j]   Std. Err.[k]      z[l]    P>|z|[l]     [95% Conf. Interval][m]
--------------+---------------------------------------------------------------
      no_acu |   2.784887[g]   .6255929     4.56   0.000     1.79306     4.32534
  health_bl |   .9834639[h]    .004774    -3.43   0.001    .9741515    .9928654
       _cons |   .8250564[i]   .2625809    -0.60   0.546    .4421624     1.53952
------------------------------------------------------------------------------
```

Logistic regression annotated output, cont.

a Log likelihood - This is the log likelihood of the model. The value -244.2483 has no meaning in and of itself; rather, this number can be used to help compare nested models.

b Number of obs - This is the number of observations that were used in the analysis. This number may be smaller than the total number of observations in your data set if you have missing values for any of the variables used in the logistic regression.

c LR chi2(2) - This is the likelihood ratio (LR) chi-square test. The number in the parenthesis indicates the number of degrees of freedom. In this model, there are 2 predictors, so there are 2 degrees of freedom.

d Prob > chi2 - This is the probability of obtaining the chi-square statistic given that the null hypothesis is true. In other words, this is the probability of obtaining this chi-square statistic (31.89) if there is in fact no effect of the independent variables, taken together, on the dependent variable. In this case, the model is statistically significant because the p-value is <0.001

e Pseudo R2 - This is the pseudo R-squared. Logistic regression does not have an equivalent to the R-squared that is found in linear regression; however, many people have tried to come up with one. There are a wide variety of pseudo-R-square statistics. Because this statistic does not mean what R-square means in linear regression (the proportion of variance explained by the predictors), interpret this statistic with great caution.

Logistic regression annotated output, cont.

f headache - This is the outcome in our logistic regression. The variables listed below it are the predictors (exposure and covariates)

g Odds Ratio – These are the odds ratios for the logistic regression equation for predicting the outcome variable (headache) from the exposure (acupuncture) and covariate (baseline health score). They are converted from the log-odds units used for the logistic regression equation.

log(p/1-p) = b0 + b1*no_acu + b2*health_bl

g no_acu – Patients who received no acupuncture had 2.87 times higher odds of severe headache (post headache score >= 30) compared to patients who had at least 1 acupuncture treatment, holding baseline health constant.

h health_bl – For every one-unit increase in baseline health score the odds of severe headache (post headache score >= 30) decrease by .98, holding acupuncture treatment constant.

i constant - This is the expected value of the log-odds of **headache** when all of the predictor variables equal zero.

k Std. Err. - These are the standard errors associated with the ORs. The standard error is used for testing whether the parameter is significantly different from 0; by dividing the parameter estimate by the standard error you obtain a z-value (see the column with z-values and p-values). The standard errors can also be used to form a confidence interval for the parameter, as shown in the last two columns of this table.

l z and **P>|z|** - These columns provide the z-value and 2-tailed p-value used in testing the null hypothesis that the OR is 0.

m [95% Conf. Interval] - This shows a 95% confidence interval for the OR.

Practice

- What is the OR for the effect of acupuncture on headache?

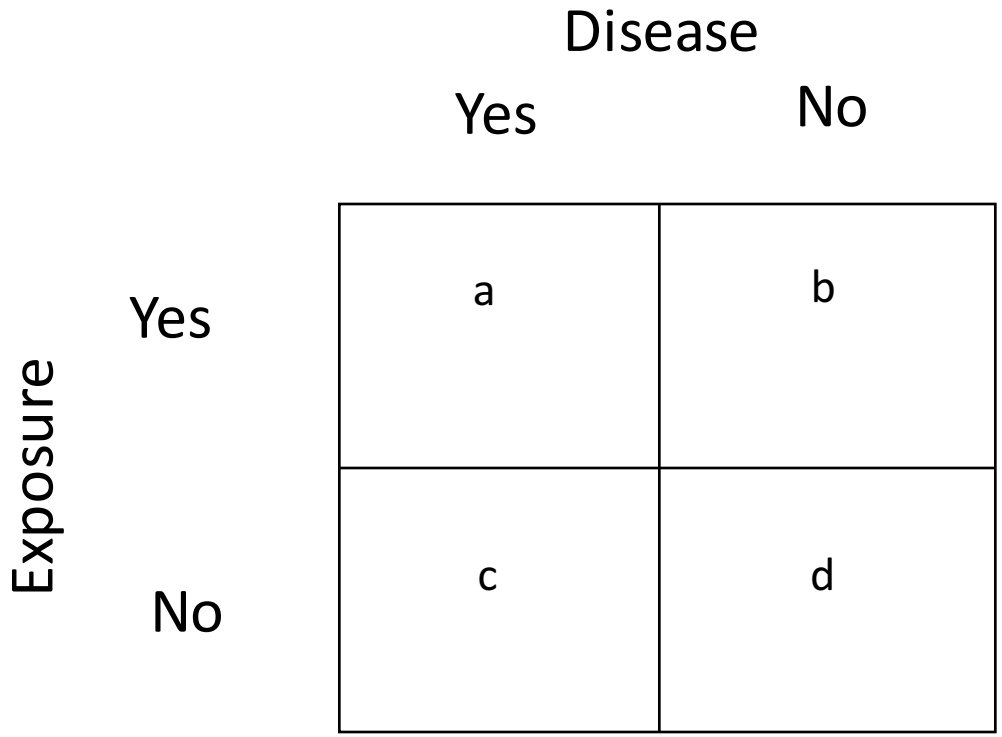
```
logistic headache no_acu
```

- What is the OR adjusted for baseline health?

```
logistic headache no_acu health_bl
```

- Bonus (Prize!) Interpret the adjusted odds ratio in a sentence.

Risk Ratio and Risk Difference



Risk Ratio =

$$\frac{a/(a+b)}{c/(c+d)}$$

Risk Difference =

$$a/(a+b) - c/(c+d)$$

Risk = Probability

For cohort studies & RCTs

Risk Ratios/Differences: `cs`

`cs outcomevar predictorvar`

– calculates risk ratios and risk differences

Options

`,by (varname)`	-stratify by a third variable [to look for interaction]
`[, missing]`	-include missing data as a category of the stratified variable

Limitations

- Cannot adjust for more than 1 covariate

Practice

```
cs headache no_acu
```

- What is the risk of headache in the exposed group (no acupuncture)?

```
di (98/(98+117))
```

- What is the risk of headache in the unexposed group (acupuncture)?

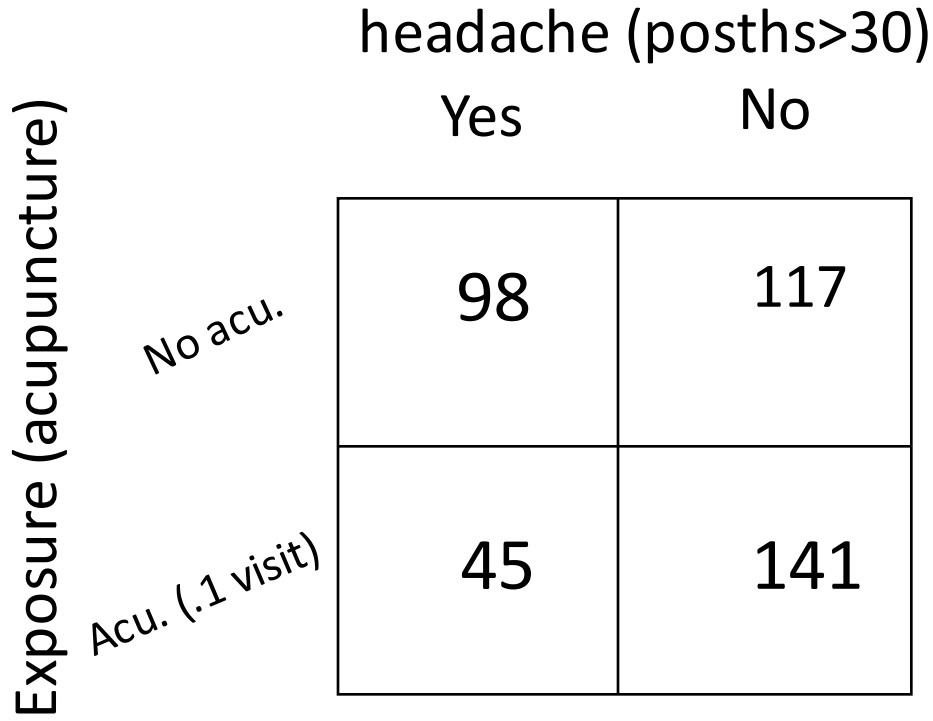
```
di (45/(45+141))
```

- What is the RR?

```
di .45/.24
```

- What is the risk difference?

```
di .45-.24
```



Multiplicative interaction with `cs`

```
cs headache no_acu, by(migraine)

        migraine |       RR       [95% Conf. Interval]   M-H Weight
-----------------+------------------------------------------------
               0 |    1.071429     .4059152   2.828077     2.333333
               1 |     1.96481     1.444798   2.671986     21.85942
-----------------+------------------------------------------------
           Crude |    1.884031     1.405061   2.526276
    M-H combined |    1.878646     1.401877    2.51756
-----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =    1.368  Pr>chi2 = 0.2422
```

MH Test of Homogeneity

Test for multiplicative interaction:

- **Is the RR for acupuncture and headache different in people with migraine compared to people without migraine?**

Null hypothesis: The stratum-specific RRs are the same (homogenous)

Alternative: The stratum-specific RRs are different (heterogeneous)

If the test of homogeneity is “significant” (small p value), we reject the null in favor of the alternative hypothesis (interaction)

Quiz

```
cs headache no_acu, by(migraine)

        migraine |       RR       [95% Conf. Interval]   M-H Weight
-----------------+------------------------------------------------
               0 |    1.071429     .4059152   2.828077     2.333333
               1 |     1.96481     1.444798   2.671986     21.85942
-----------------+------------------------------------------------
           Crude |    1.884031     1.405061   2.526276
    M-H combined |    1.878646     1.401877    2.51756
-----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =    1.368  Pr>chi2 = 0.2422
```

- (prize!) Interpret the p-value

When to report/ignore interaction

Stratum-Specific Risk Ratios for a Given Exposure and Disease			
Potential Effect Modifier		P value for heterogeneity	Report or Ignore Interaction
Present	Absent		
2.3	2.6	0.45	Ignore
2.3	2.6	0.001	Ignore
2.0	20.0	0.001	Report
2.0	20.0	0.10	Report
2.0	20.0	0.40	Ignore
3.0	4.5	0.30	Ignore
3.0	4.5	0.001	+/-
0.5	3.0	0.001	Report
0.5	3.0	0.15	+/-

More to come in epi 203!

Risk Ratio/difference: `binreg`

`binreg outcomevar predictorvar covariates`
– generalized linear model, binomial family (regression)

Options

`rr`	- report risk ratios
`rd`	- report risk differences
`,vce(robust)`	-options for adjusting SEs
`##`	-Specify an interaction between 2 variables

If using any categorical predictors with >2 categories, use `i.catvar`

Practice

- What is the Risk Ratio (RR) for the effect of acupuncture on headache?

```
binreg headache no_acu, rr
```

- What is the Risk Ratio (RR) adjusted for baseline health score?

```
binreg headache no_acu health_bl, rr
```

Practice

- What is the Risk Difference (RD) for the effect of acupuncture on headache?

```
binreg headache no_acu, rd
```

- What is the Risk Difference (RD) adjusted for baseline health score?

```
binreg headache no_acu health_bl, rd
```

Rate Ratio and Rate Difference

Rate Ratio =

[Cases in exposed/exposed person-time]
[Cases in unexposed/unexposed person-time]

Rate Difference =

[Cases in exposed/exposed person-time] -[Cases in unexposed/unexposed person-time]

Use:

```
ir O E T
poisson O E C, irr
cox O E C
```

```
O = outcomevar
E = exposurevar
C = covariates
T = time
```

For cohort studies