Sample size Calculations & Contingency Tables

- Sample size calculations using Stata
- The Chi-squared test for independence

Sample size calculations

- Set significance level, α, and the desired power (80-90%)
- Specify your hypothesized mean value for each group
- Estimate the sample standard deviation for each group

- Example: We conduct an intervention that we think increases the CD4 count among HIV-infected persons. We randomize a control versus intervention, and think it will increase CD4 count from 350 to 450. The SD for CD4 count for our previous studies was 250.

Sample size Calculations

```
. power twomeans 350 450, sd(250)

Performing iteration ...

Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =  100.0000
           m1 =  350.0000
           m2 =  450.0000
           sd =  250.0000

Estimated sample sizes:

            N =       200
  N per group =       100
```

Sample size Calculations

We write this up as:

With a sample of size 200 (100 in each group), we will have 80% power to detect a difference of 100 or greater in CD4 cell count if the CD4 cell count in the control group is 350 and the standard deviation in both groups is 250, at significance level=0.05.

- Writing sample size statements: https://www-users.york.ac.uk/~mb55/guide/size.htm

Other sample size calculators:
- UCSF online http://www.sample-size.net/
- Other online calculators (try http://powerandsamplesize.com/Calculators/)

Statistical hypothesis tests

Data and comparison type	**Alternative hypotheses**	**Parametric test Stata command**	**Non-parametric test Stata command**
Numerical; One mean	H_a: $\mu \neq \mu_a$ (two-sided) H_a: $\mu > \mu_a$ or $\mu < \mu_a$ (one-sided)	Z or t-test •ttest var1=hypoth val.*	
Numerical; Two means, <u>paired data</u>	H_a: $\mu_1 \neq \mu_2$ (two-sided) H_a: $\mu_1 > \mu_2$ or $\mu < \mu_a$ (one-sided)	Paired t-test •ttest var1=var2*	Sign test • signtest var1=var2 •Wilcoxon Signed-Rank signrank var1=var2)
Numerical; Two means, independent data	H_a: $\mu_1 \neq \mu_2$ (two-sided) H_a: $\mu_1 > \mu_2$ or $\mu < \mu_a$ (one-sided)	T-test (equal or unequal variance) •ttest var1, by(byvar) unequal	Wilcoxon rank-sum test •ranksum var1, by(byvar)
Numerical, Two or more means, independent data	H_a: $\mu_1 \neq \mu_2$ or $\mu_1 \neq \mu_3$ or $\mu_2 \neq \mu_3$ etc.	ANOVA •oneway var1 byvar	Kruskal Wallis test •kwallis var1, by(byvar)
Dichotomous; One proportion	H_a: $p \neq p_a$ (two-sided) H_a: $p > p_a$ or $p < p_a$ (one-sided)	Proportion test • prtest var1=hypoth value* • bitest var1=hypoth value	
Dichotomous; two proportions	H_a: $p_1 \neq p_2$ (two-sided) H_a: $p_1 > p_2$ (one-sided)	Proportion test (z-test) • prtest var1, by(byvar)	Chi-square test •tab var1 var2, chi exact McNemar's for paired data: •mcc var1 var2
Categorical by categorical (nxk)	H_a : The rows not independent of the columns		Chi-square test •tab var1 var2, chi exact

Categorical outcomes

- With the exception of the proportion test, all the previous tests were for comparing numerical outcomes and categorical variables (e.g. predictors)
  - e.g., CD4 count by alcohol consumption
- We often have dichotomous outcomes and variates
  - e.g. Having any children by sex

```
. tab children_any gender

children_a |
 ny - Do   |
 you have  | gender - What is your
      any  |       gender?
children?  |      Male     Female |     Total
-----------+----------------------+----------
        0  |        24         23 |        47
        1  |         2          6 |         8
-----------+----------------------+----------
    Total  |        26         29 |        55
```

Contingency tables

- Often summaries of counts of disease versus no disease and exposed versus not exposed
- Frequently 2x2 but can generalize to n x k
  - n rows, k columns
- Note that Stata sorts on the numeric value, so for 0-1 variables the disease state will be the 2nd row

		Exposure		
		+	-	Total
Disease	+	a	b	a+b
	-	c	d	c+d
Total		a+c	b+d	n=a+b+c+d

Example of data set		
Obs.	Exposure (1=yes; 0=no)	Disease (1=yes; 0=no)
1	1	1
2	1	0
3	1	1
4	0	0
5	1	1
6	1	0
7	0	0
…	…	…
n	0	0

Categorical outcomes

- Say we want to know whether the proportion with children varies by gender in biostats classes
- We could test the null hypothesis that the proportion of males with children equals that of females. This is the test of proportions from Mod 5
- H_0: $p_{w/children\ males} = p_{w/children\ females}$
- H_A: $p_{w/children\ males} \neq p_{w/children\ females}$

- There are other methods to do this (chi-square test)
- Why?
  - These methods are more general – can be used when you have more than 2 levels in either variable
- We will start with the 2x2 example however

Categorical outcomes

- Overall, the probability of having any children overall in this sample is 8/55=0.1455 This is the marginal probability of having children (in this class)
- There were 26 males and 29 females
- Under the null hypothesis, the expected proportion in each group is the overall proportion
- So the expected number of students who have children:

  Males 26*.1455=3.8

  Females: 29*.1455=4.2

Categorical outcomes

- We can also calculate the expected number without children under the null hypothesis of no difference
  - Males: 26*(1-.1455) = 22.2
  - Females: 29*(1-.1455) = 24.8
- Observed & Expected counts

Observed data

```
. tab children_any gender

children_a |
  ny - Do |
 you have | gender - What is your
      any |        gender?
children? |      Male     Female |     Total
----------+----------------------+----------
        0 |        24        23 |        47
        1 |         2         6 |         8
----------+----------------------+----------
    Total |        26        29 |        55
```

EXPECTED COUNTS UNDER THE NULL HYPOTHESIS

```
.tab children_any gender, expected nof

children_a |
   ny - Do |
  you have | gender - What is your
       any |        gender?
 children? |      Male     Female |     Total
-----------+----------------------+----------
         0 |      22.2       24.8 |      47.0
         1 |       3.8        4.2 |       8.0
-----------+----------------------+----------
     Total |      26.0       29.0 |      55.0
```

Categorical outcomes

- Generically

Expected counts		Exposure		
		+	-	Total
Disease	+	(a+b)(a+c)/n	(a+b)(b+d)/n	a+b
	-	(c+d)(a+c)/n	(c+d)(b+d)/n	c+d
Total		a+c	b+d	n=a+b+c+d

Categorical outcomes

- The Chi-squared test compares the observed frequency (O) in each cell with the expected frequency (E) under the null hypothesis of no difference
- The differences O-E are squared, divided by E, and added up over all the cells
- The sum of this is the test statistic and follows a chi-square distribution
- Interestingly, the square root of the Chi-squared distribution is the Normal distribution

Chi-squared test of independence

- The chi-squared test statistic (for the test of independence in contingency tables) for a 2x2 table (dichotomous outcome, dichotomous exposure)

$$X_1^2 = \sum_{i=1}^{4}\left(\frac{(O_i - E_i)^2}{E_i}\right)$$

- i is the index for the cells in the table – there are 4 cells
- This test statistic is compared to the chi-squared distribution with 1 degree of freedom

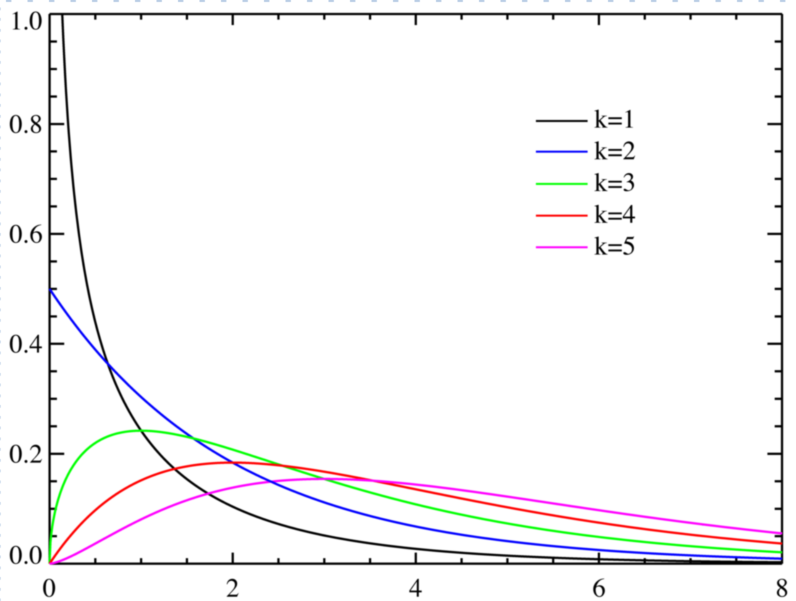
Chi-squared test of independence

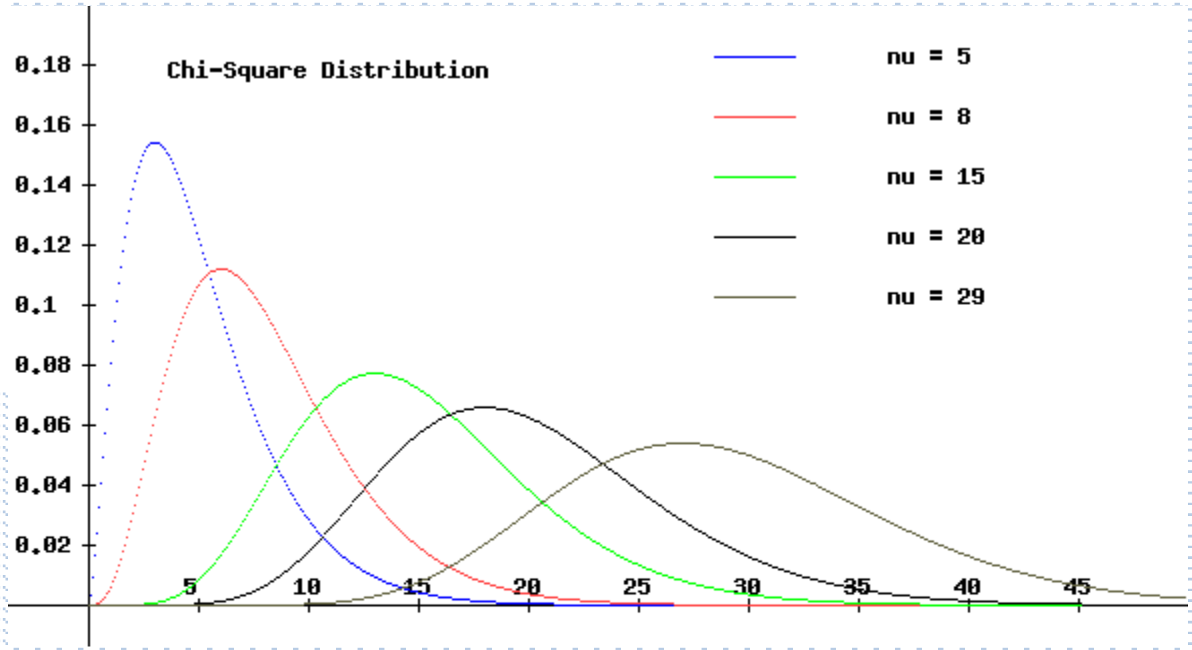
- The chi-square test statistic for the test of independence in an nxk contingency table is

$$X^2_{(n-1)*(k-1)} = \sum_{i=1}^{nk}\left(\frac{(O_i - E_i)^2}{E_i}\right)$$

- This test statistic is compared to the chi-squared distribution
- The degrees of freedom for the this test are (n-1)*(k-1), so for a 2x2 there is 1 degree of freedom
  - n=the number of rows; k=the number of columns in the nxk table
- Expected cell sizes should all be >1 and fewer than 20% should be <5
- The Chi-squared test is for two sided hypotheses

Chi-squared distribution





Mean = degrees of freedom
Variance = 2*degrees of freedom

Categorical outcomes

```
. . tab children_any gender, chi

children_a |
   ny - Do |
  you have | gender - What is your
       any |        gender?
 children? |      Male     Female |     Total
-----------+----------------------+----------
         0 |        24         23 |        47
         1 |         2          6 |         8
-----------+----------------------+----------
     Total |        26         29 |        55

          Pearson chi2(1) =   1.8632   Pr = 0.172
```

Test statistic (df)

p-value

```
 tab children_any gender, row col chi expected

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
|   row percentage   |
| column percentage  |
+--------------------+

children_a |
   ny - Do |
  you have | gender - What is your
       any |        gender?
 children? |      Male     Female |     Total
-----------+----------------------+----------
         0 |        24         23 |        47
           |      22.2       24.8 |      47.0
           |     51.06      48.94 |    100.00
           |     92.31      79.31 |     85.45
-----------+----------------------+----------
         1 |         2          6 |         8
           |       3.8        4.2 |       8.0
           |     25.00      75.00 |    100.00
           |      7.69      20.69 |     14.55
-----------+----------------------+----------
     Total |        26         29 |        55
           |      26.0       29.0 |      55.0
           |     47.27      52.73 |    100.00
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =   1.8632   Pr = 0.172
```

Test of independence

- For small cell sizes in 2x2 tables, use the Fisher exact test
- It is based on a discrete distribution called the hypergeometric distribution
- For 2x2 tables, you can choose a one-sided (results as or more extreme only in one direction) or a two-sided test

```
. . tab children_any gender, chi exact

children_a |
 ny - Do |
 you have | gender - What is your
      any |        gender?
children? |      Male     Female |     Total
-----------+----------------------+----------
         0 |        24         23 |        47
         1 |         2          6 |         8
-----------+----------------------+----------
     Total |        26         29 |        55

          Pearson chi2(1) =   1.8632   Pr = 0.172
           Fisher's exact =                 0.257
   1-sided Fisher's exact =                 0.164
```

Comparison to test of two proportions

```
prtest children_any, by(gender)

Two-sample test of proportions                     Male: Number of obs =       26
                                                 Female: Number of obs =       29
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Male |   .0769231   .0522589                     -.0255026    .1793487
      Female |   .2068966   .0752216                      .0594649    .3543282
-------------+----------------------------------------------------------------
        diff |  -.1299735    .091593                     -.3094925    .0495456
             |  under Ho:   .0952197    -1.36   0.172
------------------------------------------------------------------------------
        diff = prop(Male) - prop(Female)                          z =  -1.3650
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.0861         Pr(|Z| > |z|) = 0.1723          Pr(Z > z) = 0.9139
```

Note that for 2x2 tables the chi-squared statistic is equal to the z statistic squared

. di (-1.365)^2

1.863

Chi-squared test of independence

- The chi-squared test can be used for more than 2 levels of exposure (with a dichotomous outcome)
  - The null hypothesis is $p_1 = p_2 = \ldots = p_k$
  - The alternative hypothesis is that not all the proportions are the same
- Note that, like ANOVA, a statistically significant result does not tell you which level differed from the others
- Also when you have more than 2 groups, all tests are 2-sided
- The degrees of freedom for the test are k-1

Chi-squared test of independence

```
. tab sex lastalc_3  if hiv==1, row col chi exact

+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
| column percentage |
+-------------------+

           | RECODE of lastalc (E1. Last time
           |          took alcohol)
   A1. Sex |     Never  >1 year a  Within th |     Total
-----------+---------------------------------+----------
      male |       110         64        203 |       377
           |     29.18      16.98      53.85 |    100.00
           |     29.49      35.36      45.72 |     37.78
-----------+---------------------------------+----------
    female |       263        117        241 |       621
           |     42.35      18.84      38.81 |    100.00
           |     70.51      64.64      54.28 |     62.22
-----------+---------------------------------+----------
     Total |       373        181        444 |       998
           |     37.37      18.14      44.49 |    100.00
           |    100.00     100.00     100.00 |    100.00

          Pearson chi2(2) =  23.2657   Pr = 0.000
           Fisher's exact =                 0.000

.
```

Note that this is a 2x3 table, so the chi-squared test has 1x2=2 degrees of freedom

Chi-squared test of independence

- Another way to state the null hypothesis for the chi-squared test:
  - Factor A is not associated with Factor B
- The alternative is
  - Factor A is associated with Factor B
- For more than 2 levels of the outcome variable this would make the most sense
- The degrees of freedom are (r-1)*(c-1) (r=rows, c=columns)

Odds ratio in a cohort study

	Exposure		
Disease	+	-	**Total**
+	**a**	**b**	**a+b**
-	**c**	**d**	**c+d**
Total	**a+c**	**b+d**	**n=a+b+c+d**

- Odds of disease among the exposed persons

  = P(disease | exposed) / (1-P(disease | exposed))

  = [a / (a + c)] / [c / (a + c)] = a/c
- Odds of disease among the unexposed persons

  = P(disease | unexposed) / (1-P(disease | unexposed))

  = [b / (b + d)] / [d / (b + d)] = b/d
- Odds ratio = a/c / b/d = ad/bc

Odds ratio

- Note that the odds ratio is also equal to

  [P(exposed | disease)/(1-P(exposed |disease)] /

  [P(exposed | no disease)/(1-P(exposed | no disease)]

- This what we calculate for case-control studies because the proportion with disease is fixed (so you can't calculate the probability of disease or the odds of disease)

Interpretation of ORs and RRs

- If the OR or RR equal 1, then there is no effect of exposure on disease.
- If the OR or RR >1 then disease is increased in the presence of exposure. (Risk factor)
- If the OR or RR <1 then disease is decreased in the presence of exposure. (Protective factor)

Comparing measures of association

- When a disease is rare, i.e. the risk is <10%, the odds ratio approximates the risk ratio
- The odds ratio overestimates the risk ratio
- Why use it? – statistical properties, usefulness in case-control studies

The association of having any children with gender

```
tab children_any gender

children_a |
  ny - Do |
 you have | gender - What is your
      any |        gender?
children? |      Male     Female |     Total
-----------+----------------------+----------
        0 |        24         23 |        47
        1 |         2          6 |         8
-----------+----------------------+----------
    Total |        26         29 |        55
```

What is the (estimated) odds ratio (where sex=female is the exposure)?

```
. di (6*24)/(2*23)
3.1304348
```

95% Confidence interval for an odds ratio

- Remember the 95% confidence interval for a mean μ

Lower Confidence Limit: Upper Confidence Limit:

$$\bar{X} - 1.96\sigma / \sqrt{n} \qquad \bar{X} + 1.96\sigma / \sqrt{n}$$

- The odds ratio is <u>not</u> normally distributed (it ranges from 0 to infinity)
  - But the natural log (ln) of the odds ratio is approximately normal
  - The estimate of the standard error of the estimated ln OR is

$$SE(\ln(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

95% Confidence interval for an odds ratio

- We calculate the 95% confidence interval for the log odds

$$\left(\ln OR - 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}, \ln OR + 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)$$

- Then exponentiate back to obtain the 95% confidence interval for the OR

$$\left(e^{\left(\ln OR - 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)}, e^{\left(\ln OR + 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)}\right)$$

Calculating an odds ratio and 95% confidence interval in Stata using tabodds command

Tabodds outcomevar exposurevar , or

```
. tabodds children_any gender, or

---------------------------------------------------------------------------
      gender |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+-------------------------------------------------------------
        Male |    1.000000          .            .             .          .
      Female |    3.130435       1.83       0.1762      0.546769  17.922776
---------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(1)  =     1.83
                                  Pr>chi2  =   0.1762

Score test for trend of odds:     chi2(1)  =     1.83
                                  Pr>chi2  =   0.1762
```

Calculating an odds ratio and 95% confidence interval in Stata using cc command

```
. gen gender_01=gender-1

.
. cc children_any gender_01
                                                         Proportion
                 |   Exposed   Unexposed  |      Total      Exposed
-----------------+------------------------+------------------------
           Cases |         6           2  |          8       0.7500
        Controls |        23          24  |         47       0.4894
-----------------+------------------------+------------------------
           Total |        29          26  |         55       0.5273
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         3.130435       |    .4837995    34.13049 (exact)
 Attr. frac. ex. |         .6805556       |   -1.066972    .9707007 (exact)
 Attr. frac. pop |         .5104167       |
                 +-------------------------------------------------
                               chi2(1) =     1.86  Pr>chi2 = 0.1723
```

Exact confidence intervals use the hypergeometric distribution

Chi-square test when the exposure has several levels (outcome dichotomous)

- e.g. Is low CD4 count (<250) at HIV testing associated with alcohol consumption category?
-

```
tab cd4_low lastalc_3,  col chi

           | RECODE of lastalc (E1. Last time
           |          took alcohol)
   cd4_low |     Never  >1 year a  Within th |     Total
-----------+---------------------------------+----------
         0 |       203         89        246 |       538
           |     54.42      49.44      55.78 |     54.12
-----------+---------------------------------+----------
         1 |       170         91        195 |       456
           |     45.58      50.56      44.22 |     45.88
-----------+---------------------------------+----------
     Total |       373        180        441 |       994
           |    100.00     100.00     100.00 |    100.00

          Pearson chi2(2) =   2.0894   Pr = 0.352
```

Odds ratio when the exposure has several levels

- One level is the "unexposed" or reference level

This is as if you partitioned your data off and did a chi-square test comparing cd4<250 in the past drinker versus lifetime abstainer categories

```
. tabodds cd4_low lastalc_3, or

---------------------------------------------------------------------------
   lastalc_3 |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       Never |    1.000000          .            .             .          .
   >1 year~o |    1.220952       1.21       0.2722      0.854441   1.744676
   Within ~r |    0.946557       0.15       0.6979      0.717263   1.249151
---------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(2)   =     2.09
                                  Pr>chi2   =   0.3522

Score test for trend of odds:     chi2(1)   =     0.19
                                  Pr>chi2   =   0.6623
```

Stata lets you choose the reference level

```
. tabodds cd4_low lastalc_3, or base(2)

---------------------------------------------------------------------------
   lastalc_3 |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+-------------------------------------------------------------
       Never |    0.819033       1.21       0.2722       0.573172   1.170355
   >1 year~o |    1.000000          .            .              .          .
   Within ~r |    0.775261       2.06       0.1509       0.547265   1.098244
---------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(2)   =     2.09
                                  Pr>chi2   =   0.3522

Score test for trend of odds:     chi2(1)   =     0.19
                                  Pr>chi2   =   0.6623
```

Odds ratio for matched pairs

- The odds ratio is r/s
- The standard error of ln(OR) is

$$SE(\ln[OR]) = \sqrt{\frac{r+s}{rs}}$$

- So the 95% confidence interval for the estimated OR is

$$\left(e^{\left(\ln OR - 1.96\sqrt{\frac{r+s}{rs}} \right)}, e^{\left(\ln OR + 1.96\sqrt{\frac{r+s}{rs}} \right)} \right)$$

Odds ratio for matched pairs

- For alcohol vs. case/control status:
- OR= 9/3=3
- SE (lnOR) = sqrt((9+3)/27) = .667
- 95% CI: exp(ln(3)-1.96*.667), exp(ln(3)+1.96*.667)

  (0.81 – 11.1)

Data are in: treatment outcomes case control.dta

In Stata

mcc case_exposed control_exposed

```
. . mcc lastalc_case lastalc_control

                 | Controls               |
Cases            |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
         Exposed |         4           9  |         13
       Unexposed |         3          11  |         14
-----------------+------------------------+------------
           Total |         7          20  |         27

McNemar's chi2(1) =      3.00    Prob > chi2 = 0.0833
Exact McNemar significance probability       = 0.1460

Proportion with factor
        Cases       .4814815
        Controls    .2592593     [95% Conf. Interval]
                   ---------     --------------------
        difference  .2222222     -.0518969   .4963413
        ratio       1.857143      .9114712    3.78397
        rel. diff         .3      .0159742   .5840258

        odds ratio         3      .7486845     17.228   (exact)
```