Biostat 200
Lecture 8

Where are we

- Types of variables
- Descriptive statistics and graphs
- Probability
- Confidence intervals for means and proportions
- Hypothesis testing for means and medians and proportions (alone or grouped by a categorical variable)
- Today: Testing for categorical variables grouped by another categorical variable
- Lectures 9+10: Correlations and linear regression
- Lecture 11: Logistic regression

Assumptions of hypothesis tests

- For t-test and ANOVA, the underlying distribution of the random variable being measured (X) should be approximately normal
  - In reality the t-test is rather robust, so with large enough sample size and without very large outliers, it is ok to use the t-test
- For the ANOVA, the variance of the subgroups should be approximately equal
- For the Wilcoxon Rank Sum Test and the Kruskal-Wallis Test the underlying distributions must have the same basic shape

Categorical outcomes

- With the exception of the proportion test, all the previous tests were for comparing numerical outcomes and categorical predictors
  - E.g., CD4 count by alcohol consumption
- We often have dichotomous outcomes and predictors
  - E.g. Owning an iPhone 6 by sex

- We can make tables of the number of observations falling into each category
- These are called contingency tables
- E.g. Owning an iPhone by sex

```
 tab iphone6_01 sex

           |          sex
iphone6_01 |      Male     Female |     Total
-----------+----------------------+----------
        no |         9         21 |        30
       yes |         8         17 |        25
-----------+----------------------+----------
     Total |        17         38 |        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**

Contingency tables

- Contingency tables are usually summaries of data that originally looked like this.

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**

```
list iphone6_01 sex in 1/10

     +-------------------+
     | iphon~01      sex |
     |-------------------|
  1. |      yes   Female |
  2. |      yes     Male |
  3. |       no     Male |
  4. |       no     Male |
  5. |      yes     Male |
     |-------------------|
  6. |      yes     Male |
  7. |       no   Female |
  8. |      yes   Female |
  9. |       no     Male |
 10. |       no   Female |
     +-------------------+
```

```
list iphone6_01 sex in 1/10, nolabel

     +----------------+
     | iphon~01   sex |
     |----------------|
  1. |        1     2 |
  2. |        1     1 |
  3. |        0     1 |
  4. |        0     1 |
  5. |        1     1 |
     |----------------|
  6. |        1     1 |
  7. |        0     2 |
  8. |        1     2 |
  9. |        0     1 |
 10. |        0     2 |
     +----------------+

.
.
```

- Say we want to know whether the proportion with iPhones varies by gender.
- We could test the null hypothesis that the proportion of males owning iPhones equals that of females.
- H_0: $p_{\text{iPhone 6 males}} = p_{\text{iPhone 6 females}}$
- H_A: $p_{\text{iPhone 6 males}} \neq p_{\text{iPhone 6 females}}$

- This is the test of proportions from Lecture 6

The Proportion test

```
. prtest iphone6_01, by(sex)

Two-sample test of proportions                  Male: Number of obs =       17
                                              Female: Number of obs =       38
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Male |   .4705882   .1210578                      .2333193    .7078572
      Female |   .4473684   .0806601                      .2892775    .6054593
-------------+----------------------------------------------------------------
        diff |   .0232198   .1454684                      -.261893    .3083326
             |  under Ho:   .1452891     0.16   0.873
------------------------------------------------------------------------------
        diff = prop(Male) - prop(Female)                          z =   0.1598
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.5635         Pr(|Z| < |z|) = 0.8730          Pr(Z > z) = 0.4365
```

- 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

- Overall, the probability of owning an iPhone 6 overall in this sample is 25/55=0.455 This is the marginal probability of owning an iPhone 6
- There were 17 males and 38 females
- Under the null hypothesis, the expected proportion in each group is the overall proportion
- So the expected number of iPhone 6s:

  Males 17*.455=7.7

  Females: 38*.455=17.3

- We can also calculate the expected number without iPhone 6s under the null hypothesis of no difference
  - Males: 17*(1-.455) = 9.3
  - Females: 38*(1-.455) = 20.7
- We can make a table of the expected counts

EXPECTED COUNTS UNDER THE NULL HYPOTHESIS

```
. . tab iphone6_01 sex, expected nof

           |          sex
iphone6_01 |      Male     Female |     Total
-----------+----------------------+----------
        no |       9.3       20.7 |      30.0
       yes |       7.7       17.3 |      25.0
-----------+----------------------+----------
     Total |      17.0       38.0 |      55.0
```

Observed data

```
. tab iphone6_01 sex

           |          sex
iphone6_01 |      Male     Female |     Total
-----------+----------------------+----------
        no |         9         21 |        30
       yes |         8         17 |        25
-----------+----------------------+----------
     Total |        17         38 |        55
```

- 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

- The Chi-square 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

Chi-square test of independence

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

$$\mathrm{X}_1^2 = \sum_{i=1}^{4} \left(\frac{(O_i - E_i)^2}{E_i} \right)$$

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

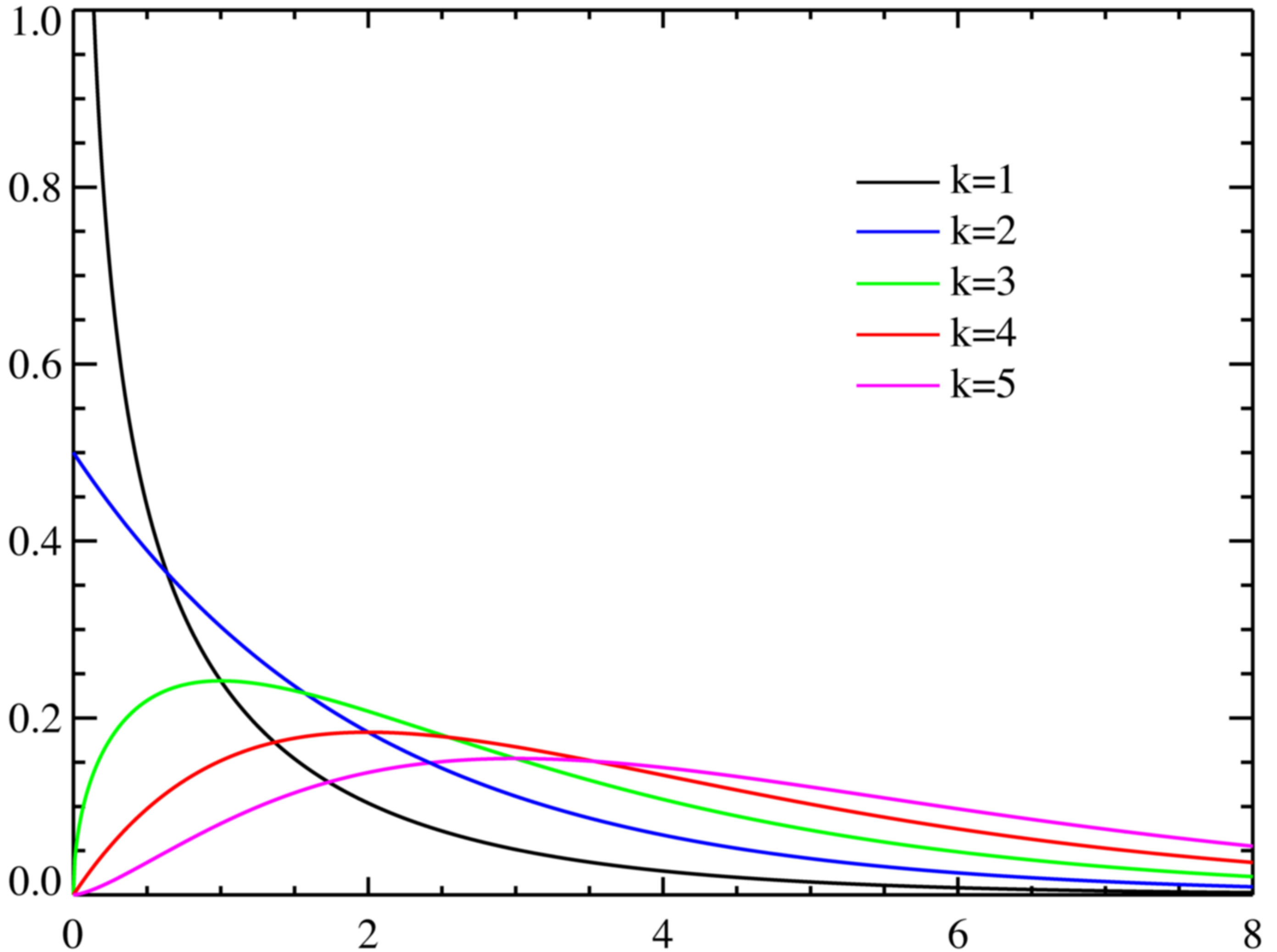
Chi-square test of independence

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

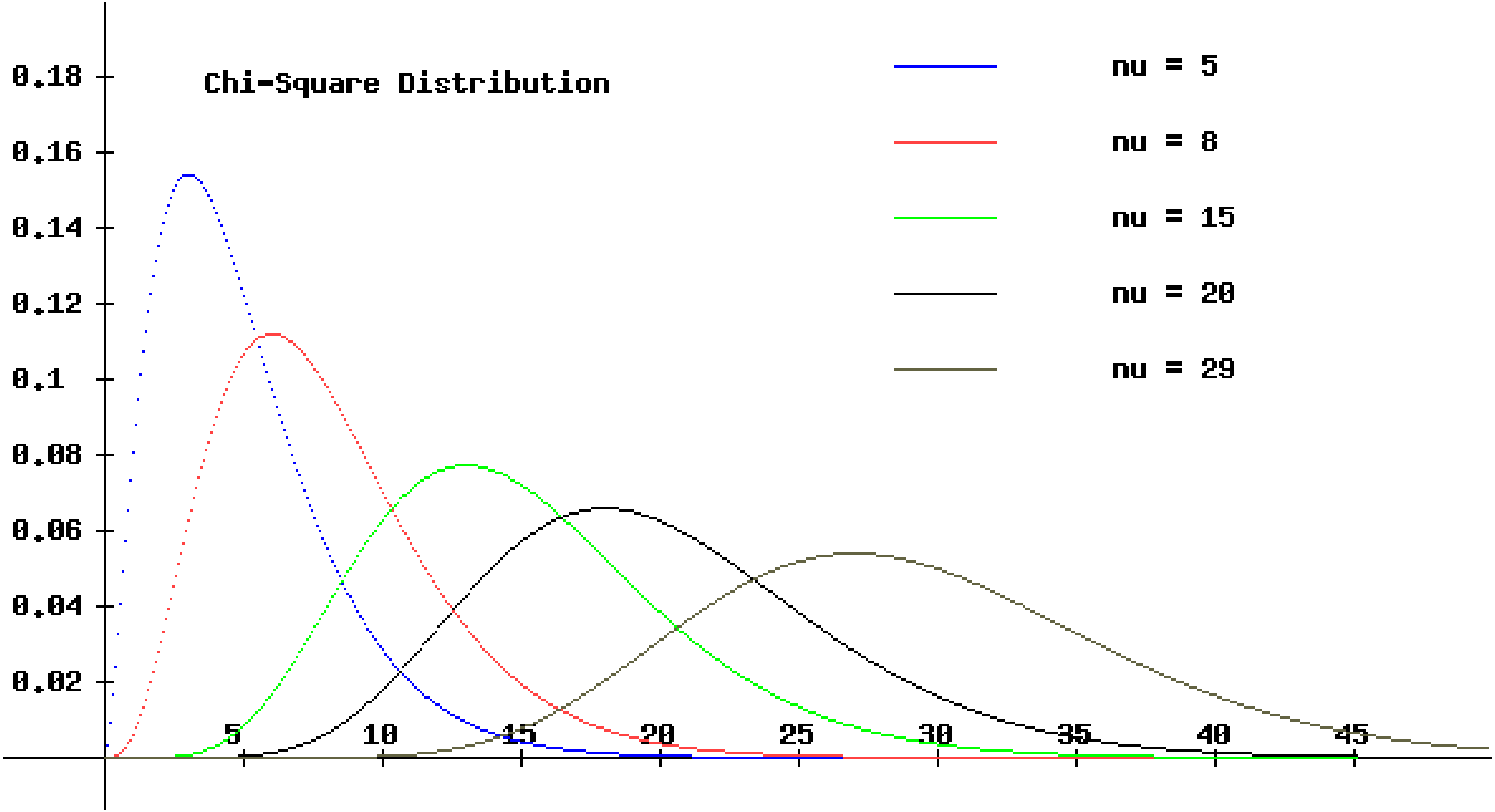
$$X^2_{(n-1)*(k-1)} = \sum_{i=1}^{nk} \left(\frac{(O_i - E_i)^2}{E_i} \right)$$

- This test statistic is compared to the chi-square 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
  - The chi-square distribution with 1 degree of freedom is actually the square of a standard normal distribution
- Expected cell sizes should all be >1 and fewer than 20% should be <5
- The Chi-square test is for two sided hypotheses

Chi-square distribution



Chi-square distribution



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

Chi-square test of independence

- For the example, the chi-square statistic for our 2x2 is

```
di ((9-9.3)^2)/9.3 + ((21-20.7)^2)/20.7 + ((8-7.7)^2)/7.7
 + ((17-17.3)^2)/17.3
. 03091587
```

There is 1 degree of freedom

- Probability of observing a chi-square value of 0.0309 with 1 degree of freedom

```
. di chi2tail(1,0.0309)
. 86046367
```

Fail to reject the null hypothesis of independence

```
. . tab iphone6_01 sex, chi

           |          sex
iphone6_01 |      Male     Female |     Total
-----------+----------------------+----------
        no |         9         21 |        30
       yes |         8         17 |        25
-----------+----------------------+----------
     Total |        17         38 |        55

          Pearson chi2(1) =   0.0255   Pr = 0.873
```

Test statistic (df)

p-value

If you want to see the row or column percentages, use row or col options

```
 tab iphone6_01 sex, row col chi expected

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

           |          sex
iphone6_01 |      Male     Female |     Total
-----------+----------------------+----------
        no |         9         21 |        30
           |       9.3       20.7 |      30.0
           |     30.00      70.00 |    100.00
           |     52.94      55.26 |     54.55
-----------+----------------------+----------
       yes |         8         17 |        25
           |       7.7       17.3 |      25.0
           |     32.00      68.00 |    100.00
           |     47.06      44.74 |     45.45
-----------+----------------------+----------
     Total |        17         38 |        55
           |      17.0       38.0 |      55.0
           |     30.91      69.09 |    100.00
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =   0.0255   Pr = 0.873
```

- Because we are using discrete cell counts to approximate a chi-squared distribution, for 2x2 tables some use the *Yates correction*

$$\mathrm{X}_1^2 = \sum_{i=1}^{4} \left(\frac{(|O_i - E_i| - 0.5)^2}{E_i} \right)$$

- Not computed in Stata

Lexicon

- When we talk about the chi-square test, we are saying it is a test of independence of two variables, usually exposure and disease.
- We also say we are testing the "association" between the two variables.
- If the test is statistically significant ($p<0.05$ if $\alpha=0.05$), we often say that the two variables are "not independent" or they are "associated".

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 iphone6_01 sex, chi exact

           |          sex
iphone6_01 |      Male     Female |     Total
-----------+----------------------+----------
        no |         9         21 |        30
       yes |         8         17 |        25
-----------+----------------------+----------
     Total |        17         38 |        55

          Pearson chi2(1) =   0.0255   Pr = 0.873
           Fisher's exact =                 1.000
   1-sided Fisher's exact =                 0.551
```

Comparison to test of two proportions

```
. prtest iphone6_01, by(sex)

 prtest iphone6_01, by(sex)

Two-sample test of proportions                     Male: Number of obs =       17
                                                 Female: Number of obs =       38
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Male |   .4705882   .1210578                      .2333193    .7078572
      Female |   .4473684   .0806601                      .2892775    .6054593
-------------+----------------------------------------------------------------
        diff |   .0232198   .1454684                      -.261893    .3083326
             |  under Ho:   .1452891     0.16   0.873
------------------------------------------------------------------------------
        diff = prop(Male) - prop(Female)                          z =   0.1598
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.5635         Pr(|Z| < |z|) = 0.8730          Pr(Z > z) = 0.4365
```

For 2x2 tables the chi-square statistic is equal to the z statistic squared

```
. di (0.1598)^2
.02553604
```

Chi-square test of independence

- The chi-square 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-square 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-square test has 1x2=2 degrees of freedom

- Another way to state the null hypothesis for the chi-square 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)

- If you run a chi-square test and fail to reject the null, can you say Factor A is not associated with Factor B?

Paired dichotomous data

- Matched pairs
  - Matched case-control study
  - Before and after data
- You cannot just put each individual into an exposure and disease box, because then you would lose the benefits of pairing (and the observations would not be independent!)
- Instead you have a table that tabulates each of the 4 possible states for each pair

Paired dichotomous data

- For a 1:1 matched case/control study, in all pairs, one has the disease (case) and the other does not (control). The table then counts the number of pairs in which
  - 1. Both were exposed
  - 2. Neither were exposed
  - 3. The case was exposed, the control was not
  - 4. The case was not exposed, the control was exposed

Case-control study HIV positives on ART in Uganda

- The study question was: Is alcohol consumption in the past 3 months associated with treatment failure?
  - The null hypothesis is that alcohol consumption is not associated with treatment failure

- Cases:
  HIV viral load after 6 months of ART >400 copies
- Controls:
  HIV viral load <400
- Matched on sex, duration on treatment, and treatment regimen class

```
. list matched_pair lastalc_case lastalc_control

     +--------------------------------+
     | matche~r   lastal~e   lastal~l |
     |--------------------------------|
  1. |        1        Yes         No |
  2. |        2        Yes        Yes |
  3. |        3         No         No |
  4. |        4        Yes         No |
  5. |        5         No         No |
     |--------------------------------|
  6. |        6         No        Yes |
  7. |        7        Yes        Yes |
  8. |        8        Yes         No |
  9. |        9         No         No |
 10. |       10         No         No |
     |--------------------------------|
 11. |       12        Yes        Yes |
 12. |       13         No        Yes |
 13. |       14        Yes         No |
 14. |       15        Yes         No |
 15. |       16        Yes         No |
     |--------------------------------|
 16. |       17         No         No |
 17. |       18         No         No |
 18. |       19         No         No |
 19. |       20        Yes         No |
 20. |       21         No         No |
     |--------------------------------|
 21. |       22        Yes         No |
 22. |       23         No         No |
 23. |       24        Yes        Yes |
 24. |       25         No         No |
 25. |       26        Yes         No |
     |--------------------------------|
 26. |       27         No        Yes |
 27. |       28         No         No |
     +--------------------------------+

.
```

```
. tab lastalc_case lastalc_control

lastalc_ca |    lastalc_control
        se |        No        Yes |     Total
-----------+----------------------+----------
        No |        11          3 |        14
       Yes |         9          4 |        13
-----------+----------------------+----------
     Total |        20          7 |        27
```

Data are in “treatment outcomes case control.dta”

- The test statistic is $$X^2 = \frac{[|r-s|-1]^2}{(r+s)}$$
- *r* and *s* are the number of discordant pairs
  - Concordant pairs provide no information
- Under the null hypothesis, r and s would be equal
- This statistic has an approximate chi-square distribution with 1 degree of freedom
- The test is called <u>McNemar's test</u>
  - The -1 is a continuity correction, not all versions of the test use this, some use .5

- r=9, s=3
- Test statistic = (9-3-1)^2/12 = 2.083

```
. di chi2tail(1,2.083)
.14894719
```

- Test statistic = (6)^2/12 = 3 (Not using the continuity correction)

```
 di chi2tail(1,3)
.08326452
```

In Stata, use mcc for Matched Case Control

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)
```

Use mcci if you only have the table, not the raw data

mcci #both_exposed #case_exposed_only #control_exposed_only #neither_exposed

```
. mcci 4 9 3 11

                 | 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)
```

- Note that the McNemar test is only for <u>MATCHED</u> data
- It is quite possible to collect unmatched case control data. Then you analyze using the chi-square methods presented earlier.

Paired dichotomous data

- For <u>before and after</u> data, the pairs are the individual participant, and the four outcomes might be:
    1. "Yes" before + "Yes" after (no change)
    2. "No" before + "No" after (no change)
    3. "Yes" before + "No" after
    4. "No" before + "Yes" after
- E.g. Positive biomarker test at baseline and 3 months follow up

Self-reported alcohol consumption in Uganda McNemar's test for paired data

- Null hypothesis: No change in the proportion PEth+ (≥50 ng/ml) from baseline to 3 months in persons entering HIV care in Uganda

```
. tab peth0_50 peth3_50

           |       peth3_50
  peth0_50 |         0          1 |     Total
-----------+----------------------+----------
         0 |        78          7 |        85 
         1 |        23         69 |        92 
-----------+----------------------+----------
     Total |       101         76 |       177 
```

using BREATH wide data.dta

Matched case-control study command

```
mcc peth0_50 peth3_50

                 | Controls               |
Cases            |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
         Exposed |        69          23  |         92
       Unexposed |         7          78  |         85
-----------------+------------------------+------------
           Total |        76         101  |        177

McNemar's chi2(1) =      8.53    Prob > chi2 = 0.0035
Exact McNemar significance probability       = 0.0052

Proportion with factor
        Cases        .519774
        Controls     .4293785     [95% Conf. Interval]
                     ---------     --------------------
        difference   .0903955      .0255752    .1552158
        ratio        1.210526      1.064678    1.376355
        rel. diff.   .1584158      .0609088    .2559229

        odds ratio   3.285714       1.36498    9.066655   (exact)
```

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

Comparison of disease frequencies across groups

- The chi-square test and McNemar's test are tests of independence
- They do not give us an estimate of how much the two groups differ, i.e. how much the disease outcome varies by the exposure variable
- We use odds ratios (OR) and relative risks (RR) as measures of ratios of disease outcome (given exposure or lack of exposure)
- The odds ratio and the relative risk are just two examples of "measures of association"

Comparison of disease frequencies – relative risk

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

- Risk ratio (or relative risk or relative rate)

  = P (disease | exposed) / P(disease | unexposed)

  = R_e / R_u = a/(a+c) / b/(b+d)

Comparison of disease frequencies – relative risk

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

- Note that you cannot calculate this entity when you have chosen your sample based on disease status
  - I.e. Case-control study – you have fixed the probability of disease a priori (i.e. usually 50%)! Relative risk is a NO GO!
  - You can calculate it but it won't have any meaning…

Odds

- If an event occurs with probability *p*, the odds of the event are *p/(1-p) to 1*
- If an event has probability .5, the odds are 1:1
- Conversely, if the odds of an event are a:b, the probability of a occurring is *a/(a+b)*
  - The odds of horse A winning over horse B winning are 2:1 → the probability of horse A winning is .667.

Odds ratio

	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

- 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 is needed for case-control studies in which the proportion with disease is fixed (so you can't calculate 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)

Comparison of 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 owning an iPhone with gender

```
tab iphone6_01 sex

           |          sex
iphone6_01 |      Male     Female |     Total
-----------+----------------------+----------
        no |         9         21 |        30
       yes |         8         17 |        25
-----------+----------------------+----------
     Total |        17         38 |        55
```

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

```
. di (9*17)/(21*8)
.91071429
```

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 <u>log odds</u>

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

- Then exponentiate back to obtain the 95% confidence interval for the <u>OR</u>

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

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

Tabodds outcomevar exposurevar , or

```
.  tabodds iphone6_01 sex, or

---------------------------------------------------------------------------
         sex |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
-------------+-------------------------------------------------------------
        Male |    1.000000          .            .             .          .
      Female |    0.910714       0.03       0.8742      0.286076   2.899230
---------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(1)  =     0.03
                                  Pr>chi2  =   0.8742

Score test for trend of odds:     chi2(1)  =     0.03
                                  Pr>chi2  =   0.8742
```

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

```
. gen sex_01=sex-1

.
. cc iphone6_01 sex_01
                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |        17           8  |         25       0.6800
        Controls |        21           9  |         30       0.7000
-----------------+------------------------+------------------------
           Total |        38          17  |         55       0.6909
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         .9107143       |    .2491868     3.37443 (exact)
 Prev. frac. ex. |         .0892857       |    -2.37443    .7508132 (exact)
 Prev. frac. pop |            .0625       |
                 +-------------------------------------------------
                               chi2(1) =     0.03  Pr>chi2 = 0.8730
```

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

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

- 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)
```

For next time

- Read Pagano and Gauvreau
  - Pagano and Gauvreau Chapter 15 (review)
  - Pagano and Gauvreau Chapters 17 and 18