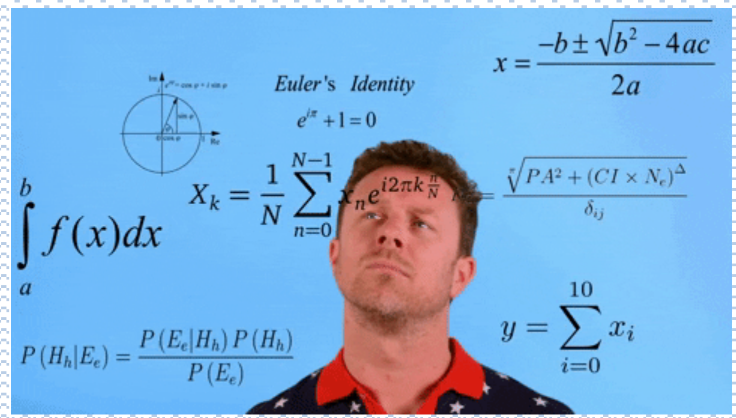
Sample size Calculations & Contingency Tables

- Sample size calculations using Stata
- More Chi-squared tests for categorical data
- Odds Ratios and Risk Ratios!!



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

Example: 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:

  Treatment failure: HIV viral load after 6 months of ART >400 copies
- Controls:

  No treatment failure: 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”

Paired dichotomous data

- 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-squared distribution with 1 degree of freedom
- The test is called McNemar's test
  - The -1 is a continuity correction, not all versions of the test use this, some use .5

Paired dichotomous data

- 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)
```

Paired dichotomous data

- Note that the McNemar test is only for MATCHED data
- It is quite possible to collect unmatched case control data. Then you analyze using the chi-square methods presented earlier.
- For before and after 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

Paired dichotomous data

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)
```

Comparison of disease frequencies across groups

- The chi-squared 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 and Odds Ratio

- 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 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

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: $\bar{X} - 1.96\sigma / \sqrt{n}$ Upper Confidence Limit: $\bar{X} + 1.96\sigma / \sqrt{n}$

- 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

```
. 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
```

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

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)
```