BIOS 662 Fall 2018

Count Data

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

https://sakai.unc.edu/portal

Outline

- One sample binary outcome
- Two sample binary outcome
- Measures of association
- Confounding - Mantel-Haenszel
- Matching - McNemar

Binomial Random Variable

- $X_1, \ldots, X_n \sim \text{Bernoulli}(\pi)$
- $Y = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, \pi)$
- Four key conditions

  1. Binary response (0/1)
  2. Observed a known number of times n
  3. Success probability (π) the same each time
  4. Independence between trials

- Example 6.1 in the text: Smoke exposure

Binomial Random Variable

- Hypothesis testing

$$H_0 : \pi = \pi_0 \text{ vs. } H_A : \pi \neq \pi_0$$

- The statistic Y is the count of the successes
- Under the null, $Y \sim \text{Binomial}(n, \pi_0)$
- Need to find $y_{\alpha/2}$ and $y_{1-\alpha/2}$ such that

$$\Pr[Y \leq y_{\alpha/2} | H_0] \leq \alpha/2$$

and

$$\Pr[Y \geq y_{1-\alpha/2} | H_0] \leq \alpha/2$$

Exact Test for Binomial Proportion

- For small samples, compute exact CR using

$$\Pr[Y \le y_{\alpha/2}] = \sum_{i=0}^{y_{\alpha/2}} \binom{n}{i} \pi_0^i (1-\pi_0)^{n-i}$$

$$\Pr[Y \ge y_{1-\alpha/2}] = \sum_{i=y_{1-\alpha/2}}^{n} \binom{n}{i} \pi_0^i (1-\pi_0)^{n-i}$$

- Binomial probabilities are computed or read from a table; e.g., in R using `pbinom` or `dbinom`;

  in SAS using `CDF('BINOMIAL',m,p,n)`

  where `m` is the number of successes

Exact Test for Binomial: Example

- Suppose $n = 12,\ \ \pi_0 = 0.4,\ \ \alpha = 0.05$

y	$\Pr[Y \leq y]$	$\Pr[Y \geq y]$
0	0.00218	1.00000
1	0.01959	0.99782
2	0.08344	0.98041
$\vdots$	$\vdots$	
7	0.94269	0.15821
8	0.98473	0.05731
9	0.99719	0.01527
10	0.99968	0.00281
11	0.99998	0.00032
12	1.00000	0.00002

- Thus $y_{0.025} = 1,\ \ y_{0.975} = 9$, and

$$C_{0.05} = \{Y : Y \leq 1 \text{ or } Y \geq 9\}$$

Exact Test for Binomial: Example II

- Suppose it is known that the 1-year death rate for a particular form of cancer is 30%.
- A new therapy designed to decrease the death rate is to be tried on 15 patients

$$H_0 : \pi = 0.3 \quad \text{vs. } H_A : \pi < 0.3$$

- Then want $C_\alpha = \{Y : Y \leq y_\alpha\}$ where

$$\sum_{i=0}^{y_\alpha} \binom{15}{i} 0.3^i 0.7^{15-i} \leq \alpha$$

- From table or R:

$$C_{0.05} = \{Y : Y \leq 1\} = \{Y : Y \in \{0, 1\}\}$$

Binomial: Large Sample

- Test of hypothesis for binomial data when n is large
- Normal approximation to binomial
- If $Y \sim \text{Binomial}(n, \pi)$, then for large n the distribution of
$$Z = \frac{Y - n\pi}{\sqrt{n\pi(1-\pi)}}$$
is approximately $N(0, 1)$
- Approximation improves as $n \to \infty$
- Rule of thumb: $n\pi(1-\pi) \geq 10$

Binomial: Example

- Revisit cancer example: Now suppose we test the new therapy on 150 patients
- Then

  $$C_{0.05} = \{z : z < -1.645\}$$

  where

  $$Z = \frac{Y - 45}{\sqrt{150(0.3)(0.7)}}$$

Binomial: Small Sample CIs

- Invert the exact test: Find all π_0 such that $H_0 : \pi = \pi_0$ would not be rejected
- To get an exact $100(1-\alpha)\%$ CI for π, solve these equations for π_L and π_U:

$$\Pr[Y \geq y | \pi = \pi_L] = \sum_{k=y}^{n} \binom{n}{k} \pi_L^k (1-\pi_L)^{n-k} = \alpha/2$$

$$\Pr[Y \leq y | \pi = \pi_U] = \sum_{k=0}^{y} \binom{n}{k} \pi_U^k (1-\pi_U)^{n-k} = \alpha/2$$

- Known as the *Clopper-Pearson* interval

Binomial: Small Sample CIs

- Can show that

$$\pi_L = \frac{y}{y + (n - y + 1) \times F_{2(n-y+1),2y,1-\alpha/2}}$$

for $1 \leq y \leq n$ ($\pi_L = 0$ for $y = 0$); and

$$\pi_U = \frac{y + 1}{y + 1 + (n - y)/F_{2(y+1),2(n-y),1-\alpha/2}}$$

for $0 \leq y \leq n - 1$ ($\pi_U = 1$ for $y = n$)

- This CI can be "extremely conservative"; cf. Wypij (*Encyclopedia of Biostatistics*, 1998)

Binomial: Small Sample CIs

- For example, suppose $n = 12$ and $y = 4$
- Then

$$\pi_L = \frac{4}{4 + 9 \times F_{18,8,0.975}}$$

- R

```
> 4/(4+9*qf(0.975,18,8))
[1] 0.0992461
```

- SAS

```
data; x=4/(4+9*quantile('f',0.975,18,8));
```

Binomial: Small Sample CIs

- R code

```
> binom.test(4,12)

        Exact binomial test

data:  4 and 12
number of successes = 4, number of trials = 12, p-value = 0.3877
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.0992461 0.6511245
```

Binomial: Small Sample CIs

- SAS code

```
data; input event count; datalines;
  0 4
  1 8
 ;

proc freq;  tables event;  exact binomial;  weight count; run;
```

```
               Binomial Proportion for event = 0
              -----------------------------------
              Proportion (P)               0.3333
              ASE                          0.1361
              95% Lower Conf Limit         0.0666
              95% Upper Conf Limit         0.6001

              Exact Conf Limits
              95% Lower Conf Limit         0.0992
              95% Upper Conf Limit         0.6511
```

Binomial: Small Sample CIs

- Suppose $y = 0$
- Then $\pi_L = 0$ because

$$\Pr[Y \geq 0 | \pi = \pi_L] = \sum_{k=0}^{n} \binom{n}{k} \pi_L^k (1 - \pi_L)^{n-k} = 1$$

for any $\pi_L \neq 0$

- For the upper bound

$$\Pr[Y \leq 0 | \pi = \pi_U] = \sum_{k=0}^{0} \binom{n}{k} \pi_U^k (1 - \pi_U)^{n-k} = \alpha/2$$

implies $\pi_U = 1 - (\alpha/2)^{1/n}$

Binomial: Small Sample CIs

- Suppose $n = 10,\ \ \alpha = 0.05,\ \ y = 0$
- $\pi_L = 0,\ \ \pi_U = 1 - 0.025^{1/10} = 0.3085$
- R

```
> binom.test(0,10)

        Exact binomial test

data:  0 and 10
number of successes = 0, number of trials = 10, p-value = 0.001953
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.0000000 0.3084971
```

Binomial: Large Sample CIs

- Let $p = Y/n$ where Y is the number of successes in n trials
- Can think of this as a random sample $X_1, X_2, \ldots, X_n$ in which $X_i = 1$ for a success and 0 otherwise, with $Y = \sum_1^n X_i$, and so $p = \bar{X}$
- If n is sufficiently large,

$$p \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

- Thus an approximate $100(1-\alpha)\%$ CI for π is

$$p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

- Rule of thumb: $np(1-p) \geq 10$

Binomial: Example

- Suppose a random sample of 886 undergrads at a college finds that 321 report binge drinking at least once in the past year

- Then point estimate for π is

$$p = \frac{321}{886} = 0.36$$

- An approximate 95% CI for the proportion of binge drinkers is:

$$0.36 \pm 1.96\sqrt{\frac{(0.36)(0.64)}{886}} = 0.36 \pm 0.03 = (0.33, 0.39)$$

Comparing Two Proportions

- Small sample sizes
  - Fisher's exact test
- Large sample sizes
  - normal approximation to the binomial
  - χ^2 test

Comparing Two Proportions

- Put the data in a 2×2 table

	Success	Failure	
Sample 1	n_{11}	n_{12}	n_1
Sample 2	n_{21}	n_{22}	n_2
	m_1	m_2	N

- Suppose $n_{11} \sim \text{Binomial}(n_1, \pi_1)$ and $n_{21} \sim \text{Binomial}(n_2, \pi_2)$

- Hypotheses

$$H_0 : \pi_1 = \pi_2$$

versus

$$H_A : \pi_1 \neq \pi_2 \quad \text{or} \quad H_A : \pi_1 < \pi_2$$

Fisher's Exact Test

- Assume the margins m_1, m_2, n_1, n_2 are fixed
- Then once we know n_{11}, the other values n_{12}, n_{21}, and n_{22} are uniquely determined
- Under H_0, can show

$$\Pr[n_{11} = k | m_1, n_1, n_2] = \frac{\binom{n_1}{k}\binom{n_2}{m_1-k}}{\binom{N}{m_1}}$$

$$= \frac{n_1!\, n_2!\, m_1!\, m_2!}{N!\, n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!}$$

- This is the *hypergeometric* distribution

Fisher's Exact Test

- For Fisher's exact test, we use the hypergeometric distribution

1. Rearrange the table so that the row with the smaller row total is the first row and the column with the smaller column total is the first column
2. Set $n_{11} = 0$ and compute $\Pr[n_{11} = 0]$ using the hypergeometric distribution
3. Construct the next table by increasing n_{11} by 1 and re-compute the probability
4. Repeat step 3 until one of the remaining 3 cells is 0
5. This gives the CDF for n_{11}

Fisher's Exact Test: Example

- A study compared the surgical mortality for patients receiving an emergency coronary bypass with those receiving a non-emergency bypass

	Dead	Alive	
Emergency	1	19	20
Non-emergency	7	369	376
Total	8	388	396

- Null hypothesis

$$H_0 : \Pr[\text{dead}|\text{emergency}] = \Pr[\text{dead}|\text{non-emergency}]$$

$$H_0 : \pi_1 = \pi_2$$

Fisher's Exact Test: Example cont.

- Set $n_{11} = 0$

	Dead	Alive	
Emergency	0	20	20
Non-emergency	8	368	376
Total	8	388	396

$$\Pr[n_{11} = 0 \mid \text{observed margins}] = \frac{20!\ 376!\ 388!\ 8!}{396!\ 0!\ 20!\ 8!\ 368!} = 0.658$$

- Similarly for $\Pr[n_{11} = 1]$, $\Pr[n_{11} = 2]$, ...

Fisher's Exact Test: Example cont.

a	$\Pr[n_{11} = a]$	$\Pr[n_{11} \leq a]$	$\Pr[n_{11} \geq a]$
0	0.658	0.658	1.000
1	0.285	0.943	0.342
2	0.051	0.994	0.057
3	0.005	0.999	0.006
4	<0.001	>0.999	<0.001
5	<0.001	>0.999	<0.001
6	<0.001	>0.999	<0.001
7	<0.001	>0.999	<0.001
8	<0.001	1.000	<0.001

Fisher's Exact Test: Example cont.

- If $H_A : \pi_1 > \pi_2$, we would reject H_0 for large n_{11}
- For example

$$C_{0.05} = \{n_{11} : n_{11} \geq 3\}$$

- P-value for this study

$$\Pr[n_{11} \geq 1] = 1 - 0.658 = 0.342$$

Fisher's Exact Test: P-values

- To compute p-values, consider all 2×2 tables possible given the observed margins
- One-sided p-value: sum the probabilities of the observed table and all tables more extreme than the observed table in the direction of H_A
- Two-sided p-value: sum the probabilities of tables that are as likely as or less likely than the observed table, given the fixed margins

Fisher’s Exact Test: P-values

- Most statistical software packages compute the p-value for Fisher’s exact test. The tables in the text are difficult to use.
- SAS:

```
data;
  input surgery $ discharge $ count;
  datalines;
  emergency dead 1
  emergency alive 19
  other dead 7
  other alive 369
 ;

proc freq order=data;
  tables surgery*discharge / nopct nocol;
  exact fisher;
  weight count;
```

Fisher’s Exact Test: SAS Output

```
surgery     discharge

Frequency|
Row Pct  |dead    |alive   |  Total
---------+--------+--------+
emergenc |      1 |     19 |      20
         |   5.00 |  95.00 |
---------+--------+--------+
other    |      7 |    369 |     376
         |   1.86 |  98.14 |
---------+--------+--------+
Total          8      388      396

       Fisher’s Exact Test
----------------------------------
Cell (1,1) Frequency (F)         1
Left-sided Pr <= F          0.9434
Right-sided Pr >= F         0.3419

Table Probability (P)       0.2854
Two-sided Pr <= P           0.3419
```

Fisher's Exact Test: P-values

- R

```
> fisher.test(matrix(c(1,19,7,369),nrow=2),alternative="greater")

        Fisher's Exact Test for Count Data

data:  matrix(c(1, 19, 7, 369), nrow = 2)
p-value = 0.3419
alternative hypothesis: true odds ratio is greater than 1
```

```
> fisher.test(matrix(c(1,19,7,369),nrow=2))

        Fisher's Exact Test for Count Data

data:  matrix(c(1, 19, 7, 369), nrow = 2)
p-value = 0.3419
alternative hypothesis: true odds ratio is not equal to 1
```

Fisher's Exact Test: Example II

- Suppose another study yields

	Dead	Alive	
Emergency	2	23	25
Non-emergency	5	30	35
Total	7	53	60

- Null hypothesis

$$H_0 : \Pr[\text{dead}|\text{emergency}] = \Pr[\text{dead}|\text{non-emergency}]$$
$$H_0 : \pi_1 = \pi_2$$

Fisher's Exact Test: Example II cont.

- p-value computation

a	$\Pr[n_{11} = a]$	$H_A : \pi_1 > \pi_2$	$H_A : \pi_1 < \pi_2$	$H_A : \pi_1 \neq \pi_2$
0	0.017		$+$	$+$
1	0.105		$+$	$+$
2	0.252	$+$	$+$	$+$
3	0.312	$+$		
4	0.214	$+$		$+$
5	0.082	$+$		$+$
6	0.016	$+$		$+$
7	0.001	$+$		$+$

Fisher's Exact Test: Example II cont.

- Critical region for $H_A : \pi_1 > \pi_2$

$$C_{0.10} = \{n_{11} : n_{11} = 5, 6, \text{ or } 7\}$$

- Critical region for $H_A : \pi_1 < \pi_2$

$$C_{0.10} = \{n_{11} : n_{11} = 0\}$$

- Critical region for $H_A : \pi_1 \neq \pi_2$

$$C_{0.10} = \{n_{11} : n_{11} = 0, 6, \text{ or } 7\}$$

Fisher's Exact Test: Comments

- Justification/ramification of conditioning on margins
- Alternative: Barnard's test, more powerful for small sample sizes. Available in StatXact. R?

Comparing Two Proportions: Large Samples

- If n_1 and n_2 are large, we can use the normal distribution
- Let n_{i1} be the number of successes in the i^{th} sample; $i = 1, 2$
- Estimator of π_i is $p_i = n_{i1}/n_i$
- From the CLT, if n_i is large

$$p_i \sim N\left(\pi_i, \frac{\pi_i(1-\pi_i)}{n_i}\right)$$

Comparing Two Proportions: Large Samples

- If samples are independent and π_i known for $i = 1, 2$, it follows

$$\frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \sim N(0, 1)$$

- This approximation is good if $n_i\pi_i(1 - \pi_i) \geq 10$ for $i = 1, 2$

Comparing Two Proportions: Large Samples

- If samples are independent and π_i unknown for $i = 1, 2,$ Slutsky/CLT imply

$$\frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0, 1)$$

  for sufficiently large n_1 and n_2

  (rule of thumb: $n_i p_i(1 - p_i) \geq 10$ for $i = 1, 2$)

Comparing Two Proportions: Example

- A case-control study was conducted to investigate the association between oral contraceptive use and myocardial infarction
- Among 234 MI patients, 29 were OC users
- Among 1,742 non-MI patients, 135 were OC users
- Let π_1 denote the probability of OC use given a case (MI) and π_2 denote the probability of OC use given a control (no MI)

Comparing Two Proportions: Example cont.

- Hypotheses

$$H_0 : \pi_1 = \pi_2 \quad \text{vs.} \quad H_A : \pi_1 \neq \pi_2$$

- Rejection region

$$C_{0.05} = \{|z| > 1.96\}$$

- Point estimates

$$p_1 = 29/234 = 0.124; \;\; p_2 = 135/1742 = 0.078$$

- Test statistic

$$z = \frac{0.124 - 0.078 - 0}{\sqrt{\frac{(0.124)(0.876)}{234} + \frac{(0.078)(0.922)}{1742}}} = 2.42$$

Comparing Two Proportions: χ^2 Test

- Alternative test of $H_0 : \pi_1 = \pi_2$ is the χ^2 test
- Recall 2 × 2 table

	Success	Failure	
Sample 1	n_{11}	n_{12}	n_1
Sample 2	n_{21}	n_{22}	n_2
	m_1	m_2	N

- It can be shown that under H_0, the statistic

$$X^2 = \frac{N(n_{11}n_{22} - n_{12}n_{21})^2}{n_1 n_2 m_1 m_2} \sim \chi_1^2$$

- Critical region for $H_A : \pi_1 \neq \pi_2$

$$C_\alpha = \{X^2 : X^2 \geq \chi^2_{1,1-\alpha}\}$$

Comparing Two Proportions: χ^2 Test

- Also known as the “Pearson” chi-square statistic
- Equivalent form

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(n_{ij} - E(n_{ij}))^2}{E(n_{ij})}$$

where $E(n_{ij}) = n_i m_j / N$

- We will see this again for $r \times c$ tables

Comparing Two Proportions: χ^2 Test

- OC-MI example:

	OC Users	Non-users	
MI Cases	29	205	234
Controls	135	1607	1742
	164	1812	1976

- Rejection region: $C_{0.05} = \{X^2 : X^2 > \chi^2_{1,0.95} = 3.84\}$
- Test statistic

$$X^2 = \frac{1976(29 \times 1607 - 135 \times 205)^2}{234 \times 1742 \times 1812 \times 164} = 5.84$$

χ^2 Test Example: SAS

```
proc freq order=data; tables patient*oc / norow nocol nopercent chisq;

                             Table of patient by oc

                      patient     oc

                      Frequency|yes     |no      |  Total
                      ---------+--------+--------+
                      mi       |     29 |    205 |    234
                      ---------+--------+--------+
                      non-mi   |    135 |   1607 |   1742
                      ---------+--------+--------+
                      Total         164     1812     1976

                     Statistics for Table of patient by oc

             Statistic                     DF       Value      Prob
             ------------------------------------------------------
             Chi-Square                     1      5.8443    0.0156
```

χ^2 Test Example: R

```
> chisq.test(matrix(c(29,205,135,1607),nrow=2),correct=FALSE)

        Pearson's Chi-squared test

data:  matrix(c(29, 205, 135, 1607), nrow = 2)
X-squared = 5.8443, df = 1, p-value = 0.01563


> chisq.test(matrix(c(29,205,135,1607),nrow=2))

        Pearson's Chi-squared test with Yates' continuity correction

data:  matrix(c(29, 205, 135, 1607), nrow = 2)
X-squared = 5.2501, df = 1, p-value = 0.02195
```

Comparing Two Proportions: χ^2 Test

- Note: $\sqrt{5.84} = 2.42$ and $\sqrt{3.84} = 1.96$
- Intuition: If $Z \sim N(0,1)$, then $Z^2 \sim \chi^2_1$
- Indeed, for 2-sided tests, the χ^2 and Z test are approximately equivalent
- In fact, if we use

$$Z = \frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

where $p = (n_{11} + n_{21})/N$,

then exactly equivalent for two-sided H_A

Comparing Two Proportions: Summary

- For small samples, use Fisher's exact test
- For large samples and $H_A : \pi_1 \neq \pi_2$, use χ^2 or Z test, i.e.,

$$\begin{aligned} C_\alpha &= \{X^2 : X^2 > \chi^2_{1,1-\alpha}\} \\ \text{or} \quad C_\alpha &= \{z : |z| > z_{1-\alpha/2}\} \end{aligned}$$

- For large samples and $H_A : \pi_1 < \pi_2$ or $H_A : \pi_1 > \pi_2$, use Z test, i.e.,

$$\begin{aligned} C_\alpha &= \{z : z < -z_{1-\alpha}\} \\ \text{or} \quad C_\alpha &= \{z : z > z_{1-\alpha}\} \end{aligned}$$

Outline

- One sample binary outcome
- Two sample binary outcome
- Measures of association
  - Risk difference
  - Relative risk (risk ratio)
  - Odds ratio
- Confounding - Mantel-Haenszel
- Matching - McNemar

Measures of Association

- In epidemiologic studies, we often obtain 2 × 2 tables

	Disease	No disease	
Exposed	n_{11}	n_{12}	n_1
Unexposed	n_{21}	n_{22}	n_2
	m_1	m_2	N

- Source could be a cross-sectional, case-control, or prospective (cohort or clinical trial) study

Measures of Association: Estimands

- Let

$$\begin{aligned} \pi_1 &= \Pr[\text{ disease} \mid \text{exposed }] \\ \text{and} \quad \pi_2 &= \Pr[\text{ disease} \mid \text{not exposed }] \end{aligned}$$

- Risk difference:

$$\mathrm{RD} = \pi_1 - \pi_2$$

- Risk ratio (relative risk):

$$\mathrm{RR} = \pi_1/\pi_2$$

- Odds ratio (cross product ratio):

$$\mathrm{OR} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

Measures of Association: Estimands

- Independence or no association corresponds to

$$\mathrm{RR} = 1 \quad \text{and} \quad \mathrm{OR} = 1$$

- $\mathrm{OR}, \mathrm{RR} \in [0, \infty)$
- $\mathrm{RR} = 4$ implies an exposed person is 4 times as likely to have the disease as an unexposed person
- $\mathrm{OR} = 4$ implies the odds of disease in the exposed is 4 times that in the unexposed

Measure of Association: Estimands

- Note

$$\text{OR/RR} = \left[\frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}\right] \Big/ \left[\frac{\pi_1}{\pi_2}\right] = \frac{1-\pi_2}{1-\pi_1}$$

- If disease rare,

$$1-\pi_1 \approx 1-\pi_2 \approx 1$$

- In this case, OR $\approx$ RR; this is important in case-control studies

- Rule of thumb:
  $\pi_1, \pi_2 \leq 0.05$ (text page 165);
  $\pi_1, \pi_2 \leq 0.10$ (Rosner, 1995, page 368);
  requires external knowledge

Measures of Association: Estimators

- Risk difference:

$$\widehat{\mathrm{RD}} = p_1 - p_2 = (n_{11}/n_1) - (n_{21}/n_2)$$

- Relative risk:

$$\widehat{\mathrm{RR}} = p_1/p_2 = (n_{11}/n_1)/(n_{21}/n_2)$$

- Odds ratio:

$$\widehat{\mathrm{OR}} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$

Estimating RR in Case-Control Studies

- In case-control studies, $\widehat{\text{RR}}$ should not be used to estimate RR. Why?
- Intuitively, RR describes $\Pr[D^+|E^+]$ and $\Pr[D^+|E^-]$, while case-control studies provide information about $\Pr[E^+|D^+]$ and $\Pr[E^+|D^-]$

Estimating RR in Case-Control Studies

- Formally: Suppose the joint distribution of exposure and disease in the population is denoted by

	Disease	No disease	
Exposed	π_{11}	π_{12}	$\pi_{1\cdot}$
Unexposed	π_{21}	π_{22}	$\pi_{2\cdot}$
	$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	

Estimating RR in Case-Control Studies

- Sample m_1 individuals with disease and m_2 without disease.
- The expected numbers of observations are

	Disease	No disease
Exposed	$\frac{m_1\pi_{11}}{\pi_{\cdot 1}}$	$\frac{m_2\pi_{12}}{\pi_{\cdot 2}}$
Unexposed	$\frac{m_1\pi_{21}}{\pi_{\cdot 1}}$	$\frac{m_2\pi_{22}}{\pi_{\cdot 2}}$
	m_1	m_2

Estimating RR in Case-Control Studies

- Therefore

$$\widehat{\mathrm{RR}} \approx \frac{\left(\frac{m_1\pi_{11}}{\pi_{\cdot 1}}\right) / \left(\frac{m_1\pi_{11}}{\pi_{\cdot 1}} + \frac{m_2\pi_{12}}{\pi_{\cdot 2}}\right)}{\left(\frac{m_1\pi_{21}}{\pi_{\cdot 1}}\right) / \left(\frac{m_1\pi_{21}}{\pi_{\cdot 1}} + \frac{m_2\pi_{22}}{\pi_{\cdot 2}}\right)}$$

$$= \frac{\pi_{11} \times \left(\frac{m_1\pi_{21}}{\pi_{\cdot 1}} + \frac{m_2\pi_{22}}{\pi_{\cdot 2}}\right)}{\pi_{21} \times \left(\frac{m_1\pi_{11}}{\pi_{\cdot 1}} + \frac{m_2\pi_{12}}{\pi_{\cdot 2}}\right)}$$

$$= \frac{\frac{m_1\pi_{11}\pi_{21}}{\pi_{\cdot 1}} + \frac{m_2\pi_{11}\pi_{22}}{\pi_{\cdot 2}}}{\frac{m_1\pi_{11}\pi_{21}}{\pi_{\cdot 1}} + \frac{m_2\pi_{12}\pi_{21}}{\pi_{\cdot 2}}}$$

- This depends on the choice of m_1 and m_2;
  for instance, $\widehat{\mathrm{RR}} \to 1$ as $m_1 \to \infty$ for fixed m_2

Estimating RR in Case-Control Studies

- On the other hand, we would expect

$$\widehat{\mathrm{OR}} \approx \frac{\left(\frac{m_1\pi_{11}}{\pi_{\cdot 1}}\right) / \left(\frac{m_2\pi_{12}}{\pi_{\cdot 2}}\right)}{\left(\frac{m_1\pi_{21}}{\pi_{\cdot 1}}\right) / \left(\frac{m_2\pi_{22}}{\pi_{\cdot 2}}\right)}$$

$$= \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}}$$

- For a rare disease, π_{11} and π_{21} are both small, so

$$\frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} \approx \frac{\pi_{11}/(\pi_{11}+\pi_{12})}{\pi_{21}/(\pi_{21}+\pi_{22})}$$

Thus $\widehat{\mathrm{OR}} \approx \mathrm{RR}$ in this case.

Estimating OR in Case-Control Studies

- Intuitively, why does $\widehat{\text{OR}}$ estimate OR in a case-control study?

$$\text{OR} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}}$$

$$= \frac{\pi_{11}/\pi_{21}}{\pi_{12}/\pi_{22}} = \frac{\frac{\pi_{11}}{\pi_{11}+\pi_{21}}\Big/\frac{\pi_{21}}{\pi_{11}+\pi_{21}}}{\frac{\pi_{12}}{\pi_{12}+\pi_{22}}\Big/\frac{\pi_{22}}{\pi_{12}+\pi_{22}}}$$

$$= \frac{\omega_1/(1-\omega_1)}{\omega_2/(1-\omega_2)}$$

where

$$\omega_1 = \pi_{11}/(\pi_{11}+\pi_{21}) = \Pr[E+|D+]$$

$$\omega_2 = \pi_{12}/(\pi_{12}+\pi_{22}) = \Pr[E+|D-]$$

Measures of Association: RD

- Similarly, $\widehat{\text{RD}}$ should not be used to estimate RD in case-control studies
- For prospective or cross-sectional studies, a $100(1-\alpha)\%$ CI for RD is given by

$$p_1 - p_2 \pm z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

when n_1 and n_2 are sufficiently large

Measures of Association: RR

- It can be shown that

$$\widehat{\text{Var}}(\log(\widehat{\text{RR}})) = \frac{n_{12}}{n_{11}n_1} + \frac{n_{22}}{n_{21}n_2}$$

  and

$$\log(\widehat{\text{RR}}) \sim N(\log(\text{RR}), \text{Var}(\log(\text{RR})))$$

- Therefore a $100(1-\alpha)\%$ CI for $\log(\text{RR})$ is

$$\log(p_1/p_2) \pm z_{1-\alpha/2}\sqrt{\frac{n_{12}}{n_{11}n_1} + \frac{n_{22}}{n_{21}n_2}}$$

Measures of Association: RR

- Thus

$$\mathrm{CI}_{\mathrm{lower}} = \frac{p_1}{p_2} \exp\left\{ -z_{1-\alpha/2} \sqrt{\frac{n_{12}}{n_{11}n_1} + \frac{n_{22}}{n_{21}n_2}} \right\}$$

$$\mathrm{CI}_{\mathrm{upper}} = \frac{p_1}{p_2} \exp\left\{ z_{1-\alpha/2} \sqrt{\frac{n_{12}}{n_{11}n_1} + \frac{n_{22}}{n_{21}n_2}} \right\}$$

- In a prospective or cross-sectional study, these CIs are recommended when $n_i p_i (1 - p_i) \geq 5$ for $i = 1, 2$ where p_1 and p_2 are the sample proportions with the disease given exposed and unexposed, respectively

- See Rosner (1995) page 364

Measures of Association: Example

- In a study of the relationship between obesity and asthma, a cohort of 3,792 children free of asthma were followed for 5 years

	Asthma	No asthma	
Obese	36	154	190
Not obese	252	3350	3602
	288	3504	3792

Measures of Association: Example cont.

- Null hypothesis

$$H_0 : \Pr[\text{asthma} \mid \text{obese}] = \Pr[\text{asthma} \mid \text{not obese}]$$

$$H_0 : \pi_1 = \pi_2$$

Equivalently:

$$H_0 : \pi_1 - \pi_2 = 0 \quad \text{or} \quad H_0 : \pi_1/\pi_2 = 1$$

- Rejection region

$$C_{0.05} = \{X^2 > 3.84\}$$

- Test statistic

$$X^2 = \frac{(3792)(36 \times 3350 - 252 \times 154)^2}{3602 \times 190 \times 288 \times 3504} = 36.73$$

Measures of Association: Example cont.

- Point estimate of RD

$$\begin{aligned}\widehat{\mathrm{RD}} &= p_1 - p_2 \\ &= 36/190 - 252/3602 \\ &= 0.189 - 0.070 \\ &= 0.12\end{aligned}$$

  Interpretation: we estimate that obese children have a 12 percentage point greater chance of developing asthma within 5 years than non-obese children

- 95% CI: (0.063, 0.176)

Measures of Association: Example cont.

- Point estimate of RR

  $$\widehat{\mathrm{RR}} = 0.189/0.070 = 2.7$$

  Interpretation: we estimate that obese children are 2.7 times more likely to develop asthma within 5 years than non-obese children

- 95% CI for RR:

  $$2.7 \exp\left\{\pm 1.96\sqrt{\frac{154}{36(190)} + \frac{3350}{252(3602)}}\right\} = (1.97, 3.72)$$

Measures of Association: SAS Code/Output

```
data;
   input asthma $ obese $ count;
   datalines;
   yes yes 36
   yes no 252
   no yes 154
   no no 3350
   ;
proc freq order=data;
   tables obese*asthma / norow nocol nopercent relrisk riskdiff;
   weight count;
```

```
                           Table of obese by asthma

                     obese     asthma

                     Frequency|yes     |no      |  Total
                     ---------+--------+--------+
                     yes      |     36 |    154 |    190
                     ---------+--------+--------+
                     no       |    252 |   3350 |   3602
                     ---------+--------+--------+
                     Total         288     3504     3792
```

Measures of Association: SAS Code/Output

```
             Statistics for Table of obese by asthma

                    Column 1 Risk Estimates

                                     (Asymptotic) 95%
                    Risk        ASE     Confidence Limits
    -------------------------------------------------------
    Row 1          0.1895     0.0284     0.1338     0.2452
    Row 2          0.0700     0.0043     0.0616     0.0783
    Total          0.0759     0.0043     0.0675     0.0844

    Difference     0.1195     0.0287     0.0632     0.1759


           Estimates of the Relative Risk (Row1/Row2)

Type of Study                    Value      95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)       3.1076        2.1151        4.5659
Cohort (Col1 Risk)              2.7083        1.9720        3.7195
Cohort (Col2 Risk)              0.8715        0.8131        0.9341
```

Measures of Association: OR

- Can show

$$\widehat{\mathrm{Var}}(\log(\widehat{\mathrm{OR}})) = \frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{12}} + \frac{1}{n_{22}}$$

  and

$$\log(\widehat{\mathrm{OR}}) \sim N(\log(\mathrm{OR}), \mathrm{Var}(\log(\mathrm{OR})))$$

  (Woolf, 1955)

- Thus for large n, a $100(1-\alpha)\%$ CI is

$$\widehat{\mathrm{OR}} \exp\left\{ \pm z_{1-\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{12}} + \frac{1}{n_{22}}} \right\}$$

Measures of Association: OR

- In a prospective or cross-sectional study, Woolf CIs are recommended when
$$n_i p_i (1 - p_i) \geq 5$$
for $i = 1, 2$ where p_1 and p_2 are the sample proportions with disease given exposed and unexposed, respectively
- In a case-control study, Woolf CIs are recommended when
$$m_i p_i^* (1 - p_i^*) \geq 5$$
for $i = 1, 2$ where p_1^* and p_2^* are the sample proportions exposed among cases and controls, respectively
- See Rosner (1995) page 369

Measures of Association: OR

- Recall the oral contraceptive use and MI example:

	OC Users	Non-users	
MI Cases	29	205	234
Controls	135	1607	1742
	164	1812	1976

- Point estimate

$$\widehat{\mathrm{OR}} = \frac{29 \times 1607}{205 \times 135} = 1.68$$

- 95% CI

$$1.684 \exp\left\{ \pm 1.96 \sqrt{\frac{1}{29} + \frac{1}{205} + \frac{1}{135} + \frac{1}{1607}} \right\} = (1.10, 2.58)$$

Measures of Association: OR

- SAS ouput:

```
                 Table of patient by oc

           patient     oc

           Frequency|yes     |no      |  Total
           ---------+--------+--------+
           mi       |     29 |    205 |    234
           ---------+--------+--------+
           non-mi   |    135 |   1607 |   1742
           ---------+--------+--------+
           Total         164     1812     1976

        Estimates of the Relative Risk (Row1/Row2)

Type of Study                    Value      95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)       1.6839       1.0991        2.5800
Cohort (Col1 Risk)              1.5992       1.0967        2.3320
Cohort (Col2 Risk)              0.9497       0.9033        0.9984
```

Measures of Association: OR

- R

```
> # First need to install the "epitools" package
> library(epitools)

> # Rows should be the exposures, columns the case status
> # Unexposed controls should be in top left cell
> example <-
        array(c(1607,135,205,29),
        dim = c(2, 2),
        dimnames = list(OC = c("Non-user", "User"),
                        MI = c("Control", "Case")))
```

Measures of Association: OR

- R

```
> oddsratio.wald(example)

$data
          MI
OC         Control Case Total
  Non-user    1607  205  1812
  User         135   29   164
  Total       1742  234  1976

$measure
          odds ratio with 95% C.I.
OC         estimate    lower    upper
  Non-user 1.000000       NA       NA
  User     1.683939 1.099069 2.580045

$p.value
          two-sided
OC         midp.exact fisher.exact chi.square
  Non-user         NA           NA         NA
  User     0.02158681   0.02228029 0.01562785
```

Confounding

- *Confounding*: A confounding variable is a variable that is associated with both the disease and the exposure.
- Such a variable may bias the measured association between exposure and disease
- A confounding variable may mask a true disease-exposure association or may cause the observed association to be too large

Confounding: Example

- Malaria and gender (case-control study)

	Malaria	No malaria	
Males	88	68	156
Females	62	82	144
	150	150	300

- Null hypothesis

$$H_0 : \pi_1 = \pi_2 \;\Leftrightarrow\; H_0 : \text{OR} = 1$$

- $\widehat{\text{OR}} = 1.71$; $X^2 = 5.34$ $(p = 0.02)$

- However, men work outdoors more than women

Confounding: Example cont.

- Stratified analysis
- Outdoor occupation $\widehat{\text{OR}} = 1.06$

	Malaria	No malaria	
Males	53	15	68
Females	10	3	13
	63	18	81

- Indoor occupation $\widehat{\text{OR}} = 1.00$

	Malaria	No malaria	
Males	35	53	88
Females	52	79	131
	87	132	219

Confounding: Mantel-Haenszel

- Adjust for possible confounding by stratification and combining 2×2 tables.

- For each stratum, $j = 1, 2, \ldots, S$, we have

	Disease	No disease	
Exposed	n_{11j}	n_{12j}	n_{1j}
Unexposed	n_{21j}	n_{22j}	n_{2j}
	m_{1j}	m_{2j}	N_j

- Recall that if the margins $(m_{1j}, m_{2j}, n_{1j}, n_{2j})$ are fixed, n_{11j} follows the hypergeometric distribution

Confounding: Mantel-Haenszel

- Thus

$$E(n_{11j}) = \frac{n_{1j}m_{1j}}{N_j}$$

and

$$\mathrm{Var}(n_{11j}) = \frac{n_{1j}n_{2j}m_{1j}m_{2j}}{N_j^2(N_j - 1)}$$

- Let

$$O_j = n_{11j};\;\; E_j = E(n_{11j});\;\; V_j = \mathrm{Var}(n_{11j})$$

and

$$O = \sum_{j=1}^{S} O_j;\;\; E = \sum_{j=1}^{S} E_j;\;\; V = \sum_{j=1}^{S} V_j;$$

Confounding: Mantel-Haenszel

- The Mantel-Haenszel statistic is given by

$$X^2_{\text{MH}} = \frac{(|O - E| - 0.5)^2}{V}$$

- Under $H_0 : \text{OR} = 1$ within strata, $X^2_{\text{MH}} \sim \chi^2_1$

$$C_\alpha = \{X^2_{\text{MH}} : X^2_{\text{MH}} > \chi^2_{1,1-\alpha}\}$$

- X_{MH} has power against the alternative hypothesis of consistent patterns of association; it has low power for detecting association in opposite directions. However, it always preserves type I error (Stokes, Davis, Koch 1995)

Confounding: Mantel-Haenszel

- Assuming homogeneous OR across strata, we can also use the MH approach to estimate the overall or common OR
- MH estimator of OR

$$\widehat{\mathrm{OR}}_{\mathrm{MH}} = \frac{\displaystyle\sum_{j=1}^{S} n_{11j}\, n_{22j}/N_j}{\displaystyle\sum_{j=1}^{S} n_{12j}\, n_{21j}/N_j}$$

Confounding: Mantel-Haenszel

- Let

$$P_j = (n_{11j} + n_{22j})/N_j; \;\; Q_j = (n_{12j} + n_{21j})/N_j$$

$$R_j = (n_{11j}\, n_{22j})/N_j; \;\; W_j = (n_{12j}\, n_{21j})/N_j$$

- Then $\mathrm{Var}(\log(\widehat{\mathrm{OR}}_{\mathrm{MH}}))$ is

$$\frac{\sum_j P_j R_j}{2(\sum_j R_j)^2} + \frac{\sum_j (P_j W_j + Q_j R_j)}{2(\sum_j R_j)(\sum_j W_j)} + \frac{\sum_j Q_j W_j}{2(\sum_j W_j)^2}$$

- A $100(1-\alpha)\%$ CI is

$$\widehat{\mathrm{OR}}_{\mathrm{MH}} \exp\left\{ \pm z_{1-\alpha/2} \sqrt{\mathrm{Var}(\log(\widehat{\mathrm{OR}}_{\mathrm{MH}}))} \right\}$$

- Robins, Breslow, Greenland (Biometrics, 1986); See Rosner 1995 p 410

Confounding: Malaria Example Revisited

- Unstratified: $X^2 = 5.34$
- Outdoor $\widehat{\mathrm{OR}} = 1.06$; indoor $\widehat{\mathrm{OR}} = 1.00$
- Outdoor:

$$O_1 = 53;\ \ E_1 = \frac{68 \times 63}{81} = 52.889;$$

$$V_1 = \frac{68 \times 13 \times 63 \times 18}{81^2 \times 80} = 1.9099$$

- Indoor:

$$O_2 = 35;\ \ E_2 = 34.9589;\ \ V_2 = 12.6620$$

Confounding: Malaria Example cont.

- MH test statistic

$$X^2_{\mathrm{MH}} = \frac{(|(53+35)-(52.889+34.9589)|-0.5)^2}{1.9099+12.6620}$$

$$= 0.008$$

without continuity correction $X^2_{\mathrm{MH}} = 0.0016$

Confounding: Malaria Example Using SAS

```
** Note that the confounder is the first variable;
** listed in the tables statement;

proc freq order=data;
  tables job*gender*malaria / cmh;
  weight count;
```

```
            Summary Statistics for gender by malaria
                       Controlling for job

  Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
---------------------------------------------------------------
    1        Nonzero Correlation        1      0.0016    0.9682
    2        Row Mean Scores Differ     1      0.0016    0.9682
    3        General Association        1      0.0016    0.9682
```

Confounding: Malaria Example using R

```
example <- array(c(53,10,15,3,35,52,53,79),
      dim = c(2, 2, 2),
      dimnames = list(Gender = c("Male", "Female"),
                      Malaria = c("Yes", "No"),
                      Job = c("Outdoors", "Indoors")))

> mantelhaen.test(example)

   Mantel-Haenszel chi-squared test without continuity correction

data:  example
Mantel-Haenszel X-squared = 0.0016, df = 1, p-value = 0.9682
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.6041733 1.6902399
sample estimates:
common odds ratio
         1.010543
```

Matched or Paired Observations

- In some studies, subjects occur naturally in pairs or matches; e.g., twins or a matched case-control design
- If we want to compare binary responses in matched pairs, the assumption of independence is violated
- The data are of the form (Y_{i1}, Y_{i2}), where $Y_{ij} = 1$ if exposed and $= 0$ if unexposed; $i = 1, 2, \ldots, n$; $j = 1, 2$

		D^+		
		$Y_{i1} = 1$	$Y_{i1} = 0$	
D^-	$Y_{i2} = 1$	n_{11}	n_{12}	
	$Y_{i2} = 0$	n_{21}	n_{22}	
				n

Matched or Paired Observations

- Note

$$\Pr[Y_{i1} = 1] = \Pr[Y_{i1} = 1, Y_{i2} = 1] + \Pr[Y_{i1} = 1, Y_{i2} = 0]$$

and

$$\Pr[Y_{i2} = 1] = \Pr[Y_{i1} = 1, Y_{i2} = 1] + \Pr[Y_{i1} = 0, Y_{i2} = 1]$$

- Therefore

$$\begin{aligned} \pi_1 - \pi_2 &= \Pr[Y_{i1} = 1] - \Pr[Y_{i2} = 1] \\ &= \Pr[Y_{i1} = 1, Y_{i2} = 0] - \Pr[Y_{i1} = 0, Y_{i2} = 1] \end{aligned}$$

Matched or Paired Observations

- Hypotheses

$$H_0 : \pi_1 = \pi_2 \quad \text{vs.} \quad H_A : \pi_1 \neq \pi_2$$

- McNemar's test statistic

$$\mathrm{M} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

- Under H_0, $\mathrm{M} \sim \chi^2_1$ if $n_{12} + n_{21}$ is sufficiently large (i.e. ≥ 30)

$$C_\alpha = \{\mathrm{M} : \mathrm{M} > \chi^2_{1,1-\alpha}\}$$

$$p = \Pr[\chi^2_1 \geq m]$$

Matched/Paired Observations: Example

- A case-control study was conducted to investigate the association between cytomegalovirus (CMV) and atherosclerosis
- Study participants with atherosclerosis, as measured by ultrasound of the carotid artery, were matched with persons without atherosclerosis, matching on age, sex, ethnicity, geographic site, and date of ultrasound
- Cytomegalovirus antibodies were measured in each person

Matched/Paired Observations: Example cont.

		Cases	
		CMV+	CMV−
Controls	CMV+	214	42
	CMV−	65	19

- McNemar's test statistic

$$\mathrm{M} = \frac{(42-65)^2}{42+65} = 4.94$$

- Reject $H_0 : \pi_1 = \pi_2$ for $\alpha = 0.05$;

$$p = \Pr[\chi_1^2 \geq 4.94] = 0.026$$

Matched or Paired Observations

- The χ^2 approximation for McNemar's test is adequate if $n_{12} + n_{21} \geq 30$
- For smaller samples, can compute the exact p-value
- Key: recognize this as a one sample binomial test
- Let $c = n_{12} + n_{21}$. If $n_{12} < c/2$, then

$$p = 2 \sum_{k=0}^{n_{12}} \binom{c}{k} 2^{-c}$$

otherwise

$$p = 2 \sum_{k=n_{12}}^{c} \binom{c}{k} 2^{-c} = 2 \sum_{k=0}^{n_{21}} \binom{c}{k} 2^{-c}$$

Matched/Paired Observations: Example II

- Suppose we want to compare 2 lotions for the treatment of poison ivy
- Persons with poison ivy on both arms are selected for the study
- One arm is randomly assigned to receive lotion 1, while the other is treated with lotion 2

		Lotion 1	
		Relief	No relief
Lotion 2	Relief	11	6
	No relief	10	24

Matched/Paired Observations: Example II cont.

- Let $\pi_i = \Pr(\text{itching relief using lotion } i)$

$$H_0 : \pi_1 = \pi_2 \text{ vs. } H_A : \pi_1 \neq \pi_2$$

- Exact p-value

$$p = 2 \sum_{k=0}^{6} \binom{16}{k} 2^{-16} = 2 \times 0.2272 = 0.4544$$

- Do not reject H_0

$$\mathrm{M} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} = \frac{(6-10)^2}{6+10} = 1$$

- R: mcnemar.test()

Matched/Paired Observations: Example II cont.

```
proc freq order=data;
   tables lotion2*lotion1 / norow nocol nopercent;
   exact agree; weight count;

                      The FREQ Procedure

                  Table of lotion2 by lotion1

              lotion2     lotion1

              Frequency|relief  |norelief|  Total
              ---------+--------+--------+
              relief   |     11 |      6 |     17
              ---------+--------+--------+
              norelief |     10 |     24 |     34
              ---------+--------+--------+
              Total          21       30       51


           Statistics for Table of lotion2 by lotion1

                         McNemar's Test
                  ----------------------------
                  Statistic (S)         1.0000
                  DF                         1
                  Asymptotic Pr >  S    0.3173
                  Exact      Pr >= S    0.4545
```

McNemar's Test

- *Marginal homogeneity*

$$H_0 : \Pr[Y_{i1} = 1] = \Pr[Y_{i2} = 1]$$

- This is a test of association with a risk factor, not a test for agreement between the members of a pair; consider

		Rater 1	
		+	−
Rater 2	+	0	65
	−	65	0

for these data: $\mathrm{M} = 0$; $p = 1$

- We'll look at a measure of agreement (kappa statistic) later in the semester

Matched or Paired Observations

- Odds ratio for matched data

$$\widehat{\mathrm{OR}}_{\mathrm{M}} = n_{21}/n_{12}$$

this is just $\widehat{\mathrm{OR}}_{\mathrm{MH}}$ with a stratum for each matched pair

- Confidence interval obtained by starting on the log scale

$$\widehat{\mathrm{Var}}(\ln(\widehat{\mathrm{OR}}_{\mathrm{M}})) \approx \frac{1}{n_{12}} + \frac{1}{n_{21}}$$

- For $n_{12} + n_{21} \geq 30$, an approximate $100(1-\alpha)\%$ CI

$$\exp\left(\ln(\widehat{\mathrm{OR}}_{\mathrm{M}}) \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\ln(\widehat{\mathrm{OR}}_{\mathrm{M}})}\right)$$

Matched or Paired Observations: CMV Example

- Odds ratio estimate

$$\widehat{\mathrm{OR}}_{\mathrm{M}} = 65/42 = 1.55$$

- Corresponding estimate of variance of $\ln(\widehat{\mathrm{OR}}_{\mathrm{M}})$

$$\widehat{\mathrm{Var}}(\ln(\widehat{\mathrm{OR}}_{\mathrm{M}})) = \frac{1}{65} + \frac{1}{42} = 0.0392$$

- Approximate 95% CI on the log scale

$$\ln(1.55) \pm 1.96 \times \sqrt{0.0392} = (0.0502, 0.8263)$$

- So an approximate 95% CI on the original scale is

$$(e^{0.0502}, e^{0.8263}) = (1.05, 2.28)$$