Counting Data
Bios 662

Michael G. Hudgens, Ph.D.

mhudgens@bios.unc.edu

http://www.bios.unc.edu/∼mhudgens

2008-09-16 14:48

Outline

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

Binomial RV

- $X_1, \ldots, X_n \sim Bernoulli(\pi)$
- $Y = \sum X_i \sim Binomial(n, \pi)$
- Four key conditions
  1. Binary response (0/1)
  2. Observed a known number of times n
  3. Success probability (π) same
  4. Independence between trials
- Example 6.1 text: Smoke exposure

Binomial RV

- Hypothesis testing

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

- The statistic Y is the count of successes
- Under the null $H_0 : Y \sim \text{Binomial}(n, \pi_0)$
- 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

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

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

- 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)

Exact Test for Binomial: Example

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

y	$\Pr[Y \leq y]$
0	0.002
1	0.020
2	0.083
$\vdots$	
7	0.943
8	0.985
9	0.997
10	$>.999$
11	$>.999$
12	1

- Thus $y_{.025} = 1$, $y_{.975} = 9$, and

$$C_{.05} = \{Y : Y \leq 1 \text{ or } Y \geq 9\}$$

Exact Test for Binomial: Example

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

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

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

$$\sum_{y=0}^{y_\alpha} \binom{15}{i} .3^i .7^{15-i} \leq \alpha$$

- From table:

$$C_{.05} = \{Y : Y \leq 1\}$$

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: suppose we test a new therapy on 150 patients
- Then

$$C_{.05} = \{z : z < -1.645\}$$

where

$$Z = \frac{Y - 45}{\sqrt{150(.3)(.7)}}$$

Binomial: Small Sample CIs

- Invert exact test: find all π_0 such that $H_0 : \pi = \pi_0$ would not be rejected
- To get an exact $(1-\alpha)\%$ 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 *Clopper-Pearson* interval

Binomial: Small Sample CIs

- Can show that

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

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

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

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

- Can be "extremely conservative"; cf Wypij (*Encyc of Bios*, 1998)

Binomial: Small Sample CIs

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

$$\pi_L = \frac{4}{4 + 9 * F_{.975,18,8}}$$

- R

```
> 4/(4+9*qf(.975,18,8))
[1] 0.0992461
```

- SAS

```
data; x=4/(4+9*quantile('f',.975,18,8)); run;
```

Small Sample Binomial 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
```

Small Sample Binomial CIs

- SAS code

```
data; input event weight; cards;
  0 4
  1 8
 ;

proc freq;  tables event;  exact binomial;  weight weight; 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 $\pi_L = 0$ since

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

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 $n = 10$, $\alpha = 0.05$, $y = 0$
- $\pi_L = 0$, $\pi_U = 1 - .025^{1/10} = 0.3085$

- 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
- If n is sufficiently large,

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

- Thus an approximate $(1-\alpha)\%$ 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

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

- Approximate 95% CI on the proportion of binge drinkers is

$$0.36 \pm 1.96\sqrt{(.36)(.64)/886} = 0.36 \pm .03 = (.33, .39)$$

Comparing Two Proportions

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

Comparing Two Proportions

- $n_{11} \sim \text{Binomial}(n_1, \pi_1)$, $n_{21} \sim \text{Binomial}(n_2, \pi_2)$
- Put data in 2 x 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

- Hypotheses

$$H_0 : \pi_1 = \pi_2$$

versus

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

Fisher's Exact Test

- Assume margins m_1, m_2, n_1, n_2 fixed
- Then once we know n_{11}, the other values n_{12}, n_{21}, and n_{22} are immediately 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 smaller row total is the first row and 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 compute the probability
  4. Reiterate step 3 until one of the remaining 3 cells is 0
  5. This gives the CDF for n_{11}

FET: Example

- A study compared the surgical mortality for patients getting an emergency coronary bypass with those getting nonemergency bypass

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

- Null hypothesis

$$H_0 : \Pr[\text{dead}|\text{emer}] = \Pr[\text{dead}|\text{nonemer}]$$

$$H_0 : \pi_1 = \pi_2$$

FET: Example

- Set $n_{11} = 0$

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

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

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

FET: Example

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	$< .001$	$> .999$	$<.001$
5	$< .001$	$> .999$	$<.001$
6	$< .001$	$> .999$	$<.001$
7	$< .001$	$> .999$	$<.001$
8	$< .001$	1.000	$<.001$

FET: Example

- If $H_A : \pi_1 > \pi_2$, we would reject H_0 if n_{11} were large
- Eg

$$C_{.05} = \{n_{11} : n_{11} \geq 3\}$$

- P-value for this study

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

FET: P-values

- To compute p-values, consider all 2 x 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 or less likely as the observed table, given the fixed margins

FET

- Most statistical software packages compute the p-value for FET. Tables in text are difficult to use.
- SAS:

```
proc freq;
 exact fisher;
 tables arm*outcome;
 run;
```

- R:

```
fisher.test(TeaTasting,  alternative = "greater")
```

FET: SAS Output

```
        The FREQ Procedure

   Table of surgery by discharge

surgery     discharge

Frequency|dead    |alive   |  Total
---------+--------+--------+
emergenc |      1 |     19 |     20
---------+--------+--------+
other    |      7 |    369 |    376
---------+--------+--------+
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
```

FET: Example II

- Suppose another study yields

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

- Null hypothesis

$$H_0 : \Pr[\text{dead}|\text{emer}] = \Pr[\text{dead}|\text{nonemer}]$$

$$H_0 : \pi_1 = \pi_2$$

FET: Example II

- 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	+		+

FET: Example II

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

$$C_{.10} = \{n_{11} : n_{11} = 5, 6, \text{ or } 7\}$$

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

$$C_{.10} = \{n_{11} : n_{11} = 0\}$$

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

$$C_{.10} = \{n_{11} : n_{11} = 0, 6, \text{ or } 7\}$$

FET: 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 $p_i = n_{i1}/n_i$.
- The CLT shows that 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
- While among 1742 non-MI patients, 135 were OC users
- Let π_1 denote the probability of OC use given case (MI) and π_2 denote the probability of OC use given control (no MI)

Comparing Two Proportions: Example

- Hypotheses

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

- Rejection region

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

- Point estimates

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

- Test statistic

$$z = \frac{0.124 - 0.078 - 0}{\sqrt{\frac{(.124)(.876)}{234} + \frac{(.078)(.922)}{1742}}} = 2.42$$

Comparing Two Proportions: χ^2 Test

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

$$X^2 = \frac{N(n_{11}n_{22} - n_{12}n_{21})^2}{n_1 n_2 m_1 m_2} \sim \chi^2_1$$

- 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

- Aka "Pearson" chi-square statistic
- Equivalent form

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

  where $En_{ij} = n_i m_j / N$

- 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: $C_{0.05} = \{X^2 : X^2 > \chi^2_{1,.95} = 3.84\}$
- Test statistic

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

- R: chisq.test()

χ^2 Test: SAS Output

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

                        The FREQ Procedure

                      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
```

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 approx 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 FET
- For large samples and $H_A : \pi_1 \neq \pi_2$, use χ^2 or Z test, i.e.,

  $C_\alpha = \{X^2 : X^2 > \chi^2_{1,1-\alpha}\}$ or $C_\alpha = \{z : |z| > z_{1-\alpha/2}\}$

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

  $$C_\alpha = \{z : z < -z_{1-\alpha}\} \text{ or } C_\alpha = \{z : z > z_{1-\alpha}\}$$

Outline

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

Measures of Association

- Risk difference
- Relative risk (risk ratio)
- Odds ratio

Measures of Association

- In epidemiologic studies, we often obtain 2 x 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 cross-sectional, case-control, or prospective (cohort or clinical trial) study

Measures of Association: Estimands

- Let

  $\pi_1 = \Pr[\text{disease}|\text{exposure}]$ and $\pi_2 = \Pr[\text{disease}|\text{no exposure}]$

- Risk difference:

$$RD = \pi_1 - \pi_2$$

- Risk ratio (relative risk):

$$RR = \pi_1/\pi_2$$

- Odds ratio (cross product ratio):

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

Measures of Association: Estimands

- Independence or no association corresponds to

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

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

Measure of Association: Estimands

- Note

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

- 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); Rosner (1995, page 368) $\pi_1, \pi_2 \leq 0.10$; requires external knowledge

Measures of Association: Estimators

- Risk difference:

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

- Relative risk:

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

- Odds ratio:

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

Estimating RR in Case-Control Studies

- In case-control studies, $\widehat{RR}$ 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}
Unexposed	π_{21}	π_{22}

Estimating RR in Case-Control Studies

- Sample m_1 individuals with disease and m_2 without disease.
- Expected number of observations is

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

Estimating RR in Case-Control Studies

- Therefore

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

$$= \frac{\pi_{11}*(m_1\pi_{21}/\pi_{.1}+m_2\pi_{22}/\pi_{.2})}{\pi_{21}*(m_1\pi_{11}/\pi_{.1}+m_2\pi_{12}/\pi_{.2})}$$

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

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

Estimating RR in Case-Control Studies

- On the other hand, we would expect

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

- Under rare disease, π_{11} and π_{21} are both small,

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

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

Estimating OR in Case-Control Studies

- Intuitively why does $\widehat{OR}$ estimate OR in cc study?

$$OR = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}/\pi_{21}}{\pi_{12}/\pi_{22}} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

where

$$p_1 = \pi_{11}/(\pi_{11}+\pi_{21}) = \Pr[E+|D+]$$

$$p_2 = \pi_{12}/(\pi_{12}+\pi_{22}) = \Pr[E+|D-]$$

Measures of Association: RD

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

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

- Therefore a $(1-\alpha)\%$ CI for $\log(RR)$ is

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

- Thus

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

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

Measures of Association: RR

- 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 (p_2) is the sample proportion with disease given exposed (unexposed)

- See Rosner (1995) page 364

Measures of Association: Example

- In a recent study of the relationship between obesity and asthma, a cohort of 3792 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

- Null hypothesis

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

$$H_0 : \pi_1 = \pi_2 \iff H_0 : \pi_1 - \pi_2 = 0 \iff H_0 : \pi_1/\pi_2 = 1$$

- Rejection region

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

- Test statistic

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

Measures of Association: Example

- Estimate of RD

  $\widehat{RD} = p_1 - p_2 = 36/190 - 252/3602 = 0.189 - 0.070 = 0.12$

- 95% CI: (0.063, 0.176)

Measures of Association: Example

- Point estimate of RR

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

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

- RR 95% CI:

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

Measures of Association: SAS Code/Output

```
proc freq order=data;
  tables wt*lungs/relrisk riskdiff;
  weight weight;
run;
```

Statistics for Table of wt by lungs

Column 1 Risk Estimates

	Risk	ASE	(Asymptotic) 95% 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

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

  Woolf (1955)

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

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

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 (p_2) is the sample proportion with disease given exposed (unexposed)

- 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^* (p_2^*) is the sample proportion exposed among cases (controls)

- See Rosner (1995) page 369

Measures of Association: OR

- Recall OC-MI example:

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

- Point estimate

$$\widehat{OR} = \frac{(29)(1607)}{(205)(135)} = 1.68$$

- 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:

The FREQ Procedure

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

```
> library(epitools)

>example <-
        array(c(29,205,135,1607),
        dim = c(2, 2),
        dimnames = list(
          OC = c("User", "Non-user"),
          MI = c("Case", "Control")))
```

Measures of Association: OR

- R

```
> oddsratio.wald(example)

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

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

$p.value
          two-sided
OC          midp.exact fisher.exact chi.square
  User              NA           NA         NA
  Non-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 can alter the measured association between exposure and disease
- A confounding variable can mask a true disease-exposure relationship or can cause the observed relationship 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 hyp

$$H_0 : \pi_1 = \pi_2 \iff H_0 : OR = 1$$

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

Confounding: Example

- However, men work outdoors more than women
- Stratified analysis
- Outdoor occupation $\widehat{OR} = 1.06$

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

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

	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 x 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: MH

- Thus

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

and

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

- Let

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

and

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

Confounding: MH

- The Mantel-Haenszel statistic is given by

$$X_{MH} = \frac{(|O - E| - .5)^2}{V}$$

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

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

- X_{MH} has power against the alternative hyp 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: MH

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

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

Confounding: MH

- 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 $Var(\log(\widehat{OR}_{MH}))$ 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 $(1-\alpha)$ x 100% CI is

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

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

Confounding: Malaria Example Revisited

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

  $$O_1 = 53; E_1 = \frac{68 * 63}{81} = 52.889; V_1 = \frac{68 * 13 * 63 * 18}{81^2 * 80} = 1.9099$$

- Indoor:

  $$O_2 = 35; E_2 = 34.9589; V_2 = 12.662$$

- MH test statistic

  $$X^2 = \frac{(|(53 + 35) - (52.889 + 34.9589)| - .5)^2}{1.8633 + 12.662} = 0.008$$

  without continuity correction $X^2 = 0.0016$

Confounding: Malaria Example using SAS

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

  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(53,15,10,3,
        35,53,52,79),
      dim = c(2, 2, 2),
      dimnames = list(
          Gender = c("Male", "Female"),
          Malaria = c("Yes", "No"),
          Job = c("Out", "In")))
```

```
> 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
```

Matched or Paired Observations

- In some studies, subjects occur naturally in pairs or matches; e.g., twins
- 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 $Y_{ij} = 1$ for a success and 0 for a failure; $i = 1, 2, \ldots, n$; $j = 1, 2$

	$Y_{i1} = 1$	$Y_{i1} = 0$	
$Y_{i2} = 1$	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

  $$\pi_1 - \pi_2 = \Pr[Y_{i1} = 1] - \Pr[Y_{i2} = 1]$$

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

Matched or Paired Observations

- Hypotheses

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

- McNemar's test statistic

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

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

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

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

Matched or Paired Observations: Example

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

Matched or Paired Observations: Example

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

- McNemar's test statistic

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

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

$$p = \Pr[\chi^2_1 \geq 4.94] = 0.026$$

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

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

Matched or Paired Observations: Example

- Suppose we want to compare 2 lotions in 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

Matched or Paired Observations: Example

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

Matched or Paired Observations: Example

- 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

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

- Do not reject H_0

- R: mcnemar.test()

Matched or Paired Observations: Example

```
proc freq order=data;
  tables lot2*lot1/nopct norow nocol;
  exact agree; weight wt;
```

The FREQ Procedure

Table of lot2 by lot1

lot2	lot1		
Frequency	relief	norelief	Total
relief	11	6	17
norelief	10	24	34
Total	21	30	51

Statistics for Table of lot2 by lot1

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 not a test for agreement; consider

		Rater 1	
		+	-
Rater 2	+	0	65
	-	65	0

for these data: $M = 0$; $p = 1$

Matched or Paired Observations

- Odds ratio for matched data

$$\widehat{OR}_M = n_{12}/n_{21}$$

  this is just $\widehat{OR}_{MH}$ w/ stratum for each matched pair

- Estimated variance (page 180 text)

$$\widehat{Var}(\widehat{OR}_M) \approx (1 + \widehat{OR}_M)^2(\widehat{OR}_M)/(n_{12} + n_{21})$$

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

$$\widehat{OR}_M \pm z_{1-\alpha/2}\sqrt{Var(\widehat{OR}_M)}$$

Matched or Paired Observations: CMV Example

- Odds ratio estimate

$$\widehat{OR}_M = 65/42 = 1.55$$

- Corresponding estimate of variance

$$\widehat{Var}(\widehat{OR}_M) = 2.55^2 * 1.55/(65 + 42) = 0.0942$$

- Approximate 95% CI

$$1.55 \pm 1.96 * .31 = (0.95, 2.15)$$

Estimated variance justification

Let $p = n_{12}/(n_{12} + n_{21})$. Denote the corresponding population parameter by π. Let $q = 1 - p$. Now

$$\widehat{OR}_M = \frac{p}{q} = \frac{1-q}{q} = \frac{1}{q} - 1.$$

Thus

$$Var(\widehat{OR}_M) = Var(\frac{1}{q}) = Var(\frac{n_{12}+n_{21}}{n_{21}}) = (n_{12}+n_{21})^2 Var(\frac{1}{n_{21}}) \qquad (1)$$

Using the usual approximation

$$Var(f(n_{12})) \approx (f'(E(n_{12}))^2 Var(n_{12})$$

with $f(x) = 1/x$ and the fact that $n_{21} \sim Binomial(n_{12} + n_{21}, 1 - \pi)$, we have

$$Var(\frac{1}{n_{21}}) \approx (-\frac{1}{E(n_{21})^2})^2 (n_{12}+n_{21})\pi(1-\pi)$$

or equivalently

$$Var(\frac{1}{n_{21}}) \approx \frac{1}{(n_{12}+n_{21})^4(1-\pi)^4}(n_{12}+n_{21})\pi(1-\pi)$$

which equals

$$Var(\frac{1}{n_{21}}) \approx \frac{\pi}{(n_{12}+n_{21})^3(1-\pi)^3}.$$

Substituting into (1), we have

$$Var(\widehat{OR}_M) \approx \frac{\pi}{(n_{12}+n_{21})(1-\pi)^3}.$$

Since we don't know π we subsitute in p, yielding

$$\widehat{Var}(\widehat{OR}_M) = \frac{p}{(n_{12}+n_{21})q^3} = \widehat{OR}_M \frac{1}{(n_{12}+n_{21})q^2}.$$

Finally, $\widehat{OR}_M + 1 = 1/q$, so we have

$$\widehat{Var}(\widehat{OR}_M) = \widehat{OR}_M(1+\widehat{OR}_M)^2 \frac{1}{(n_{12}+n_{21})}.$$