Design and Analysis Techniques for Epidemiologic Studies (I)

outline

1. Introduction
2. Study design
3. Measures of effect for categorical data
   - Risk Difference
   - Risk Ratio
   - Odds Ratio
4. Confounding and standardization
5. Mantel-Haenszel Test
6. Power and Sample-Size Estimation for Stratified Categorical Data

Introduction

- In epidemiologic applications, the rows of the table refer to disease categories and the columns to exposure categories (or vice versa).

Hypothetical exposure–disease relationship

		disease		
		Yes	No	
Exposure	Yes	a	b	$a + b = n_1$
	No	c	d	$c + d = n_2$
		$a + c = m_1$	$b + d = m_2$	

Contd.

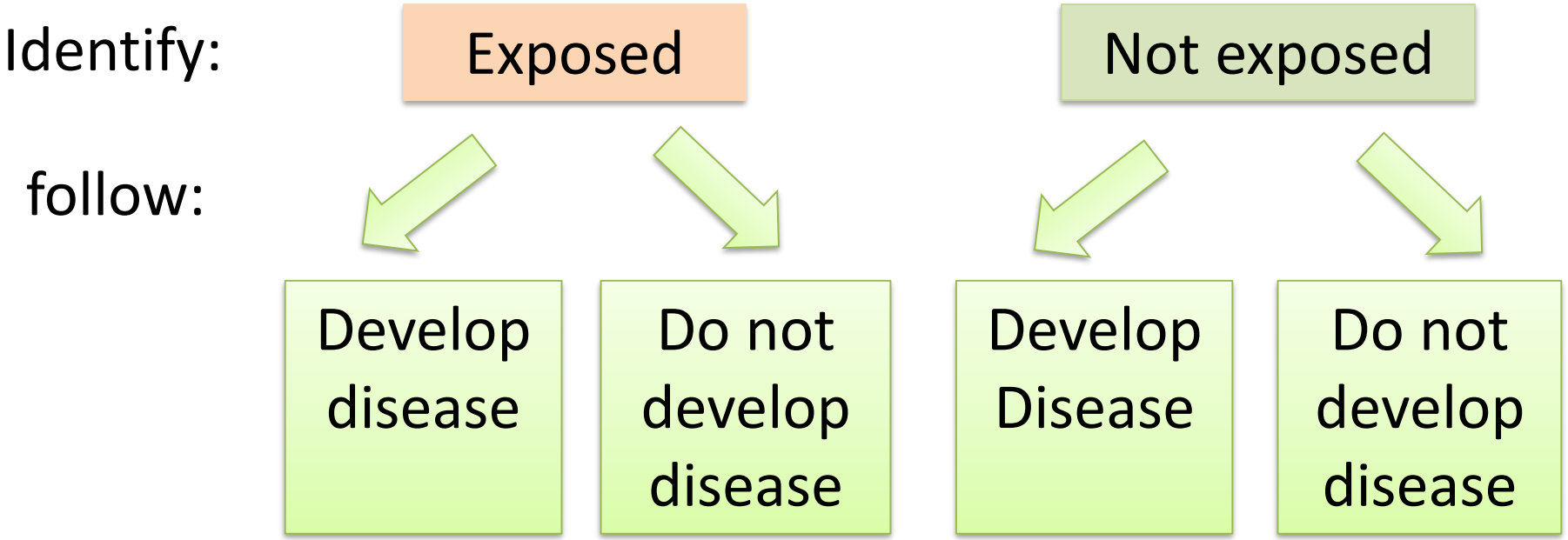
One of the Primary Goals of Epidemiological Investigation

Study design

1. prospective study, or cohort study
2. retrospective study, or case–control study
3. cross-sectional study, or prevalence study

prospective study

- Also called cohort study
- a group of disease-free individuals is identified at one point in time and are followed over a period of time until some of them develop the disease.



prospective study

② Then, follow to see whether **③ Calculate and compare**

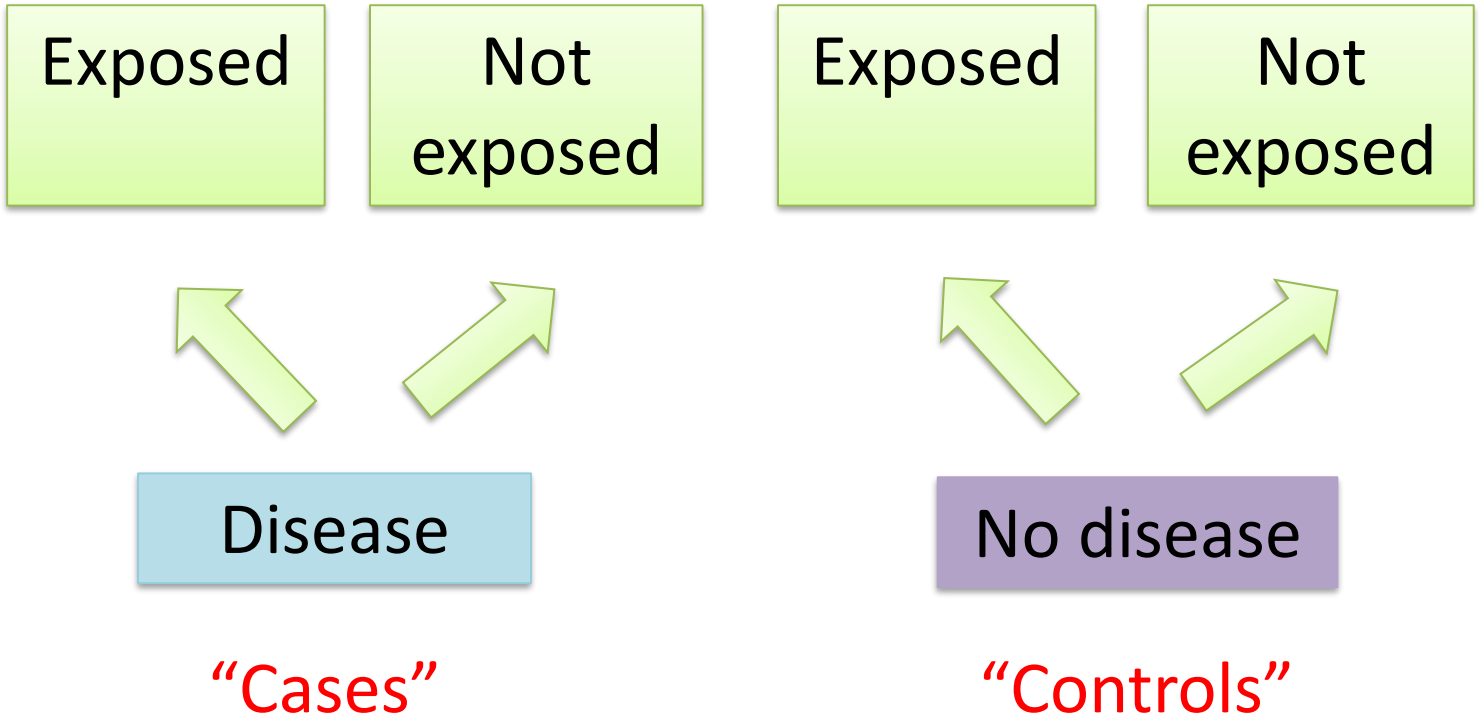
① First, identify

	Disease develops	Disease does not develop	Totals	Incidence of disease
Exposed	a	b	$a+b$	$\frac{a}{a+b}$
Not exposed	c	d	$c+d$	$\frac{c}{c+d}$

$\frac{a}{a+b}$ =incidence in exposed $\frac{c}{c+d}$ =incidence in not exposed

Retrospective study

- also sometimes called a case–control study.



Retrospective study

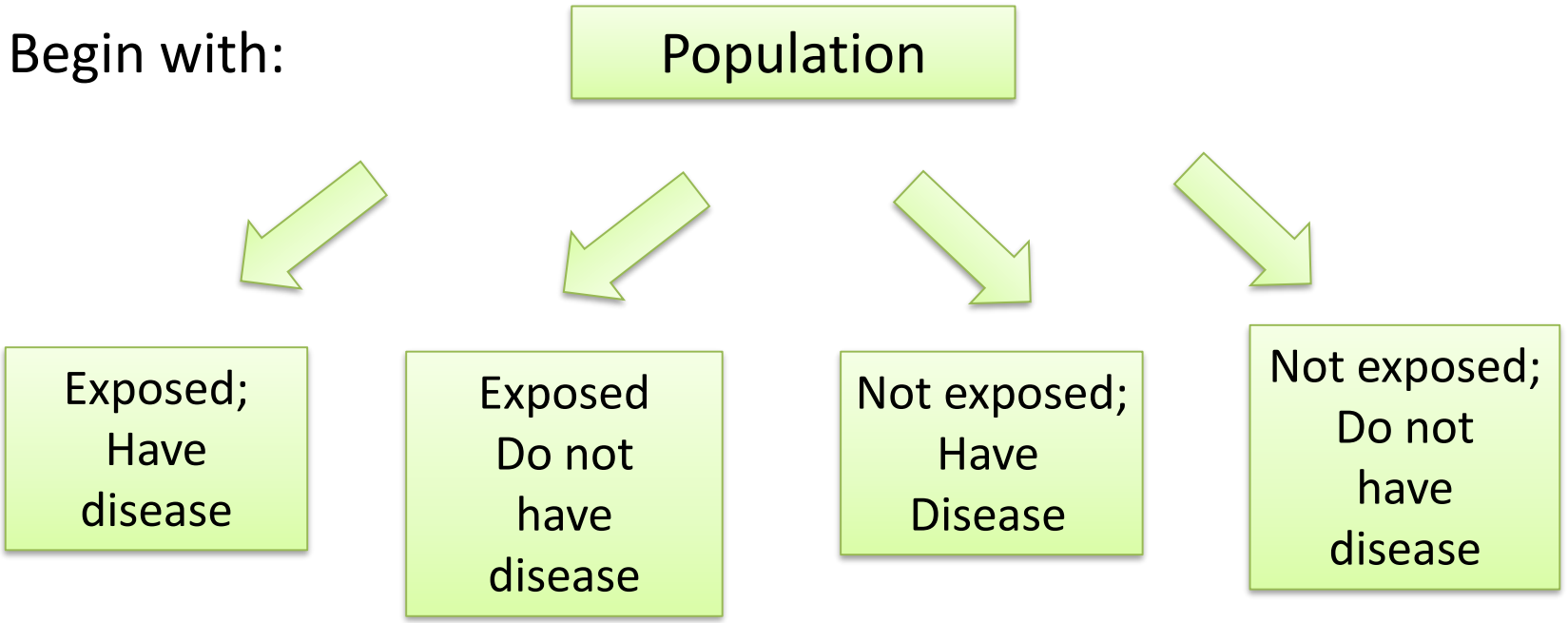
① **First, select**

② Then, measure past exposure	Cases (with disease)	Controls (without disease)
Were exposed	a	b
Were not exposed	c	d
Totals	$a+c$	$a+d$

③ Proportion exposed $\frac{a}{a+c}$ $\frac{b}{b+d}$

cross-sectional study

- Sometimes called a prevalence study
- All participants are asked about their current disease status and their current or past exposure status.



cross-sectional study

	Disease	No Disease
Exposed	a	b
Not Exposed	c	d

- It assesses the prevalence of exposures and/or of diseases in the population rather than incidence

Nested Case-Control Study

- The principal aims of the study were to investigate the effect of aspirin use on CHD and the effect of beta-carotene use on cancer incidence.

- randomized to one of four treatment groups

Arm	Aspirin	beta-carotene
1	placebo	placebo
2	active	placebo
3	placebo	active
4	active	active

Contd.

- The goal of the second part of the study was to relate lipid (油脂) abnormalities identified in the blood samples to the occurrence of CHD.
- It would have been prohibitively expensive to analyze all the blood samples that were collected.
- instead…

Group	Sample size
Case (develop CHD)	≈ 300
Control (matched age)	≈ 600

=> a case–control study nested within a prospective study

Comparison

- prospective study
  - The gold standard of designs
  - Control confounders
  - But
    - expensive
    - long time (i.e., × rare diseases)
- retrospective study
  - Inexpensive
  - Save time
  - But
    - recall bias
    - selection bias

Measures of effect for categorical data

The Risk Difference (RD)

- The risk difference (RD) is defined as $p_1 - p_2$

- $p_1 =$ probability of developing disease for exposed individuals
- $p_2 =$ probability of developing disease for unexposed individuals

- Then

$$\frac{|p_1 - p_2 - (\hat{p}_1 - \hat{p}_2)| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0,1)$$

The Risk Ratio (RR)

- The risk ratio(RR), or relative ratio, is defined as p_1/p_2, with a point estimate $\widehat{RR} = \hat{p}_1/\hat{p}_2$

- p_1 = probability of developing disease for exposed individuals
- p_2 = probability of developing disease for unexposed individuals

- Then how can we obtain an interval estimate?
- Note that the sampling distribution of $\ln(\widehat{RR})$ more closely follows a normal distribution than $\widehat{RR}$ itself.

- But how can we get the SE of $\ln(RR)$? ⇒ delta method

$$Var\left(\ln \widehat{RR}\right) = Var(\ln(\hat{p}_1/\hat{p}_2))$$
$$= Var(\ln \hat{p}_1) + Var(\ln \hat{p}_2)$$
$$\approx \frac{(1-\hat{p}_1)}{\hat{p}_1 n_1} + \frac{(1-\hat{p}_2)}{\hat{p}_2 n_2}$$

- Then a approximate two-sided $100\% \times (1-\alpha)$ CI for $\ln(RR)$ is

$$\ln\left(\widehat{RR}\right) \pm z_{1-\alpha/2}\sqrt{\frac{\hat{q}_1}{\hat{p}_1 n_1} + \frac{\hat{q}_2}{\hat{p}_2 n_2}}$$

- Taking antilog(i.e., exponential) of each end of this interval provides a two-sided $100\% \times (1-\alpha)$ CI for RR.

The Odds Ratio (OR)

- Let $p =$ the probability of a success
- The odds in favor of success$= p/(1-p)$
- Then

$$\text{OR}= \frac{p_1/q_1}{p_2/q_2} = \frac{p_1 q_2}{p_2 q_1}$$

- OR is estimated by $\widehat{OR} = \frac{\hat{p}_1 \hat{q}_2}{\hat{p}_2 \hat{q}_1}$, for 2×2 contingency table, the sample OR is $\widehat{OR} = \frac{ad}{bc}$

	Disease	No Disease
Exposed	a	b
Not Exposed	c	d

- Similar with the RR, how to obtain interval estimates for the OR?

The Woolf method

- The Woolf method (based on the delta method)

$$Var\left(\ln \widehat{OR}\right) \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

- Then a approximate two-sided $100\% \times (1-\alpha)$ CI for $\ln(OR)$ is

$$\ln\left(\widehat{OR}\right) \pm z_{1-\alpha/2}\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

- Taking antilog(i.e., exponential) of each end of this interval provides a two-sided $100\% \times (1-\alpha)$ CI for OR.

Summary

- the main effect measures used in epidemiologic studies: the RD, RR, and OR.

	Effect measures		
Designs	RD	RR	OR
Prospective	√	√	√
Case-control	×	×	√

- In case-control studies, we can not estimate the probability of disease as each category of exposed group.

- In case-control studies with a rare disease outcome, the *OR* provides an indirect estimate of the *RR*.

Confounding and Standardization

Motivation example
Cancer

- the relationship between lung-cancer incidence and heavy drinking (defined as ≥2 drinks per day)

		Lung cancer: Yes	Lung cancer: No	
Drinking status	Heavy drinker	33	1667	1700
	Nondrinker	27	2273	2300
		60	3940	4000

- Because lung cancer is relatively rare, we estimate the *RR* by the *OR* = (33 × 2273)/(27 × 1667) = 1.67. Thus it appears heavy drinking is a risk factor for lung cancer.

Is that true?

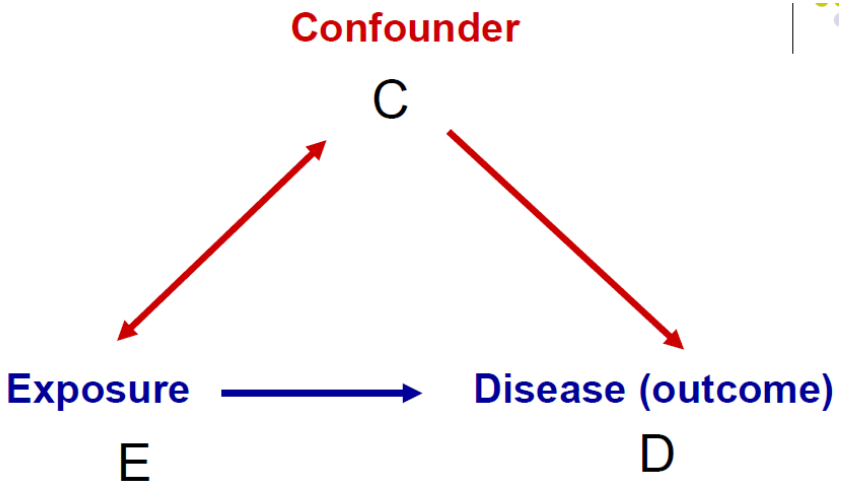
Confounding

- A confounding variable is a variable that is associated with both the disease and the exposure variable.

- i.e., smoking
  - related to drinking status, of heavy drinkers:
    - 800 of the 1000 smokers (80%)
    - 900 of the 3000 nonsmokers (30%)
  - related to lung cancer, of those developed lung cancer:
    - 30 of the 1000 smokers (3%)
    - 30 of the 3000 nonsmokers (1%)

- Such a variable must usually be controlled for before looking at a disease–exposure relationship.

Confounders

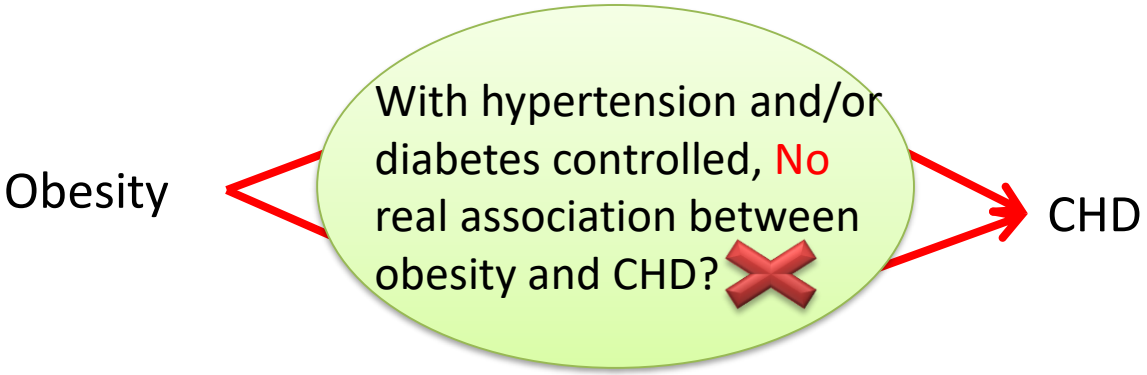


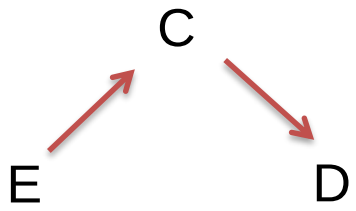
- When is it reasonable to control for a confounder when exploring the relationship between an exposure and disease?
- It depends …

Causal pathway

- It depends on whether or not C is in the causal pathway between E and D
  - the exposure is causally related to the confounder
  - the confounder is causally related to disease.

- The causal pathway should be made on the basis of biological rather than purely statistical considerations.





Revisited
Cancer

- Heavy drinkers are more likely than nondrinkers to be smokers, and smokers are more likely to develop lung cancer than nonsmokers.

		(a) Smokers at baseline			(b) Nonsmokers at baseline		
		Lung cancer			Lung cancer		
		Yes	No		Yes	No	
Drinking status	Heavy drinker	24	776	800	9	891	900
	Nondrinker	6	194	200	21	2079	2100
		30	970	1000	30	2970	3000

- OR(among smokers) = $(24 \times 194)/(6 \times 776) = 1.0$
- OR(among nonsmokers) = $(9 \times 2079)/(21 \times 891) = 1.0$.

$\Rightarrow$ *no* relationship between lung cancer and drinking status.

Stratification

- The analysis of disease–exposure relationships in separate subgroups of the data, in which the subgroups are defined by one or more potential confounders.
- The subgroups themselves are called strata.
  - i.e., stratification by age

R function: epi.2by2

Association between MI and OC use by age

Age	Recent OC use	Cases (MI)	Controls	$\widehat{OR}$	Proportion OC user	Proportion MI
25–29	Yes	4	62	7.2	23	2
	No	2	224			
30–34	Yes	9	33	8.9	9	5
	No	12	390			
35–39	Yes	4	26	1.5	8	9
	No	33	330			
40–44	Yes	6	9	3.7	3	16
	No	65	362			
45–49	Yes	6	5	3.9	3	24
	No	93	301			
Total	Yes	29	135	1.7		
	No	205	1607			

Age-stratified Analysis

- It is often routine to control for age when assessing disease–exposure relationships. Then how estimate the RR?
  - Age-standardized risk of disease among the exposed

$$\hat{p}_1^* = \sum n_i \hat{p}_{i1} / \sum n_i$$

  - Age-standardized risk of disease among the unexposed

$$\hat{p}_2^* = \sum n_i \hat{p}_{i2} / \sum n_i$$

  - (Age-) Standardized RR

$$\text{RR} = \hat{p}_1^* / \hat{p}_2^*$$

- But…
- How to control for confounding in assessing disease–exposure relationships in a hypothesis-testing framework?

The Mantel-Haenszel Test

Methods of Inference for Stratified Categorical Data

Motivation example

Cancer

- main purpose: look at the effect of passive smoking on cancer risk.
- One potential confounding variable: smoking by the participants themselves (i.e., personal smoking)

		(a) Smokers			(b) Nonsmokers		
		Passive smoker			Passive smoker		
		Yes	No	Total	Yes	No	Total
Status	Case	120	111	231	161	117	278
	Control	80	155	235	130	124	254
	Total	200	266	466	291	241	532

- The key question is how to combine the results from the two tables to obtain an overall estimated OR and test of significance for the passive-smoking effect.

Mantel-Haenszel Test

- In general, the data are stratified into *k* subgroups according to one or more confounding variables to make the units within a stratum as homogeneous as possible.

- **Procedure:**

1. Form *k* strata, based on the level of the confounding variable(s), and construct a 2 × 2 table relating disease and exposure within each stratum, as shown as follows

		Exposure		
		Yes	No	Total
Disease	Yes	a_i	b_i	$a_i + b_i$
	No	c_i	d_i	$c_i + d_i$
	Total	$a_i + c_i$	$b_i + d_i$	n_i

Contd.

2. the total observed number of units (O) in the (1, 1) cell over all strata,
$$O = \sum O_i = \sum a_i$$

3. the total expected number of units (E) in the (1, 1) cell over all strata,
$$E = \sum_i E_i = \sum_i \frac{(a_i + b_i)(a_i + c_i)}{n_i}$$

4. Compute the variance (V) of O under H_0, where
$$V = \sum_i V_i = \sum_i \frac{(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)}{n_i^2 (n_i - 1)}$$

5. The test statistic is then given by (only $V > 5$)

$$X^2_{MH} = \frac{(|O - E| = 0.5)^2}{V} \sim \chi^2_1 \text{ under } H_0$$

Stratified Data

- Assuming that the underlying *OR* is the same for each stratum, an estimate of the common underlying *OR* is provided by

$$\widehat{OR}_{MH} = \frac{\sum_i a_i d_i / n_i}{\sum_i b_i c_i / n_i}$$

- Interval Estimate

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

- Where

$$Var(\ln \hat{O}R_{MH}) = \frac{\sum_{i=1}^{k} P_i R_i}{2\left(\sum_{i=1}^{k} R_i\right)^2} + \frac{\sum_{i=1}^{k} (P_i S_i + Q_i R_i)}{2\left(\sum_{i=1}^{k} R_i\right)\left(\sum_{i=1}^{k} S_i\right)} + \frac{\sum_{i=1}^{k} Q_i S_i}{2\left(\sum_{i=1}^{k} S_i\right)^2} = A + B + C$$

and

$$P_i = \frac{a_i + d_i}{n_i},\ Q_i = \frac{b_i + c_i}{n_i},\ R_i = \frac{a_i d_i}{n_i},\ S_i = \frac{b_i c_i}{n_i}$$

- What if the underlying *OR* is different in the various strata?
- In general, it is important to test for homogeneity of the stratum-specific *OR*s.
- If the true *OR*s are significantly different, then it makes no sense to obtain a pooled-*OR* estimate such as given by the Mantel-Haenszel estimator.
- Instead, separate *OR*s should be reported.

Chi-Square Test for Homogeneity of *OR*s

- Chi-Square Test for Homogeneity of *OR*s over Different Strata (Woolf Method)

- To test $H_0: OR_1 = \cdots = OR_k$ vs. H_1: at least two of the OR_i are different (MH test $H_0: OR_1 = \cdots = OR_k = 1$)

- $\ln \widehat{OR}_i = \ln[a_i d_i/(b_i c_i)]$
- $w_i = \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)^{-1}$, $\overline{\ln OR} = \frac{\sum w_i \ln \widehat{OR}_i}{\sum w_i}$

- Then the test statistic is given by

$$X^2_{HOM} = \sum_i w_i \left(\ln \widehat{OR}_i - \overline{\ln OR}\right)^2$$

$$= \sum_i w_i \left(\ln \widehat{OR}_i\right)^2 - \left(\sum_i w_i \ln \widehat{OR}_i\right)^2 / \sum w_i$$

$$\sim \chi^2_{k-1} \text{ under } H_0$$

Revisited
Cancer

```
## input data
> cancer <-
+ array(c(120, 80, 111, 155,
+          161, 130, 117, 124),
+       dim = c(2, 2, 2),
+       dimnames = list(
+              status = c("case", "control"),
+              pass.smok = c("yes", "no"),
+              smok = c("yes","no")))
> mantelhaen.test(cancer)

Mantel-Haenszel chi-squared test with continuity
correction

data: cancer
```

Contd.

```
Mantel-Haenszel X-squared = 13.9423, df = 1, p-value =
0.0001885
alternative hypothesis: true common odds ratio is not
equal to 1
95 percent confidence interval:
       1.263955 2.090024
sample estimates:
common odds ratio
             1.625329

> library("vcd")
> woolf_test(cancer)
Woolf-test on Homogeneity of Odds Ratios (no 3-Way assoc.)

data: cancer
X-squared = 3.2697, df = 1, p-value = 0.07057
```

Recall (McNemar's test for matched pair data, lecture 5)

- Matched pairs are a special case of stratification in which each matched pair corresponds to a separate stratum of size 2.
- McNemar's test is a special case of the Mantel-Haenszel test for strata of size 2??

discordant pair

		Case		
		Exposed	Not exposed	Total
Control	Exposed	$n_{11}\ (\pi_{11})$	$n_{12}\ (\pi_{12})$	$n_{1+}\ (\pi_{1+})$
	Not exposed	$n_{21}\ (\pi_{21})$	$n_{22}\ (\pi_{22})$	$n_{2+}\ (\pi_{2+})$
	Total	$n_{+1}\ (\pi_{+1})$	$n_{+2}\ (\pi_{+2})$	n

type B discordant pair

type A discordant pair

McNemar's test (Revisited)

- $n_D = n_{12} + n_{21}$
- Under H_0, $n_{21} \sim Binomial(n_D, 1/2)$
- Test statistic

$$\frac{(n_{21} - n_D/2)^2}{n_D/4} \text{ or } \frac{(n_{21} - n_{12})^2}{n_{12} + n_{21}}$$

follows an asymptotic χ^2 distribution with 1 *df*

- Four possible cases in each stratum

	1 0 0 1	0 1 1 0	1 0 1 0	0 1 0 1
O_i	1	0	1	0
E_i	½	½	2*1/2=1	0*1/2=0
V_i	¼	¼	0	0

Matched-Pair Studies

- want to study the relationship between a dichotomous disease and exposure variable, in a case–control design.
- control for confounding by forming matched pairs of subjects with disease (cases) and subjects without disease (controls).

- MH estimator $\widehat{OR} = n_A/n_B$ (page 34)
- $\text{var}\left(\ln \widehat{OR}\right) = 1/(n\hat{p}\hat{q}), \hat{p} = \frac{n_A}{n_A+n_B}, \hat{q} = 1 - \hat{p}$
- Then

$$\frac{\ln \widehat{OR} - \ln OR}{\sqrt{1/(n\hat{p}\hat{q})}} \sim N(0,1)$$

- only be used if n = number of discordant pairs is ≥20.

Example

Cancer

- want to compare two different chemotherapy regimens for breast cancer after mastectomy (from example 10.21 in McNemar's test)

		Outcome of treatment B patient		
		Survive for 5 yrs	Die within 5 yrs	Total
Outcome of treatment A patient	Survive for 5 yrs	510	16	526
	Die within 5 yrs	5	90	95
	Total	515	106	621

R codes

```
> a5=rep(1,526) # A 方法存活大于5年(暴露)
> a4=rep(0,95) # A 方法存活小于5年(非暴露)

> b5=a5
> b4=a4
> b5[511:526]=0 # A>5年的B有16个小于5年
> b4[1:5]=1 # A<5年的B有5个大于5年

> a=c(a5,a4) # A方法(case)的所有结果
> b=c(b5,b4) # B方法(control)的所有结果

> caseControl=c(rep(0,621),rep(1,621))
> expose=c(a,b)
> pair=c(1:621,1:621) # 配对
> library(epicalc)
```

R codes(contd.)

```
> matchTab(caseControl,expose,pair)

Exposure status: expose = 1

Total number of match sets in the tabulation = 621

Number of controls = 1
                      No. of controls exposed
No. of cases exposed    0   1
                     0 90  16
                     1  5 510

Odds ratio by Mantel-Haenszel method = 0.312

Odds ratio by maximum likelihood estimate (MLE) method =
0.313 95%CI= 0.114 , 0.853
```

Trend test with confounder

- Suppose we have *s* strata. In each stratum, we have a 2 $\times$ *k* table relating disease (2 categories) to exposure (*k* ordered categories) with score for the *j*th category = x_j as shown

Relationship of disease to exposure in the i th stratum, $i = 1, \ldots, s$						
		Exposure				
		1	2	…	k	
Disease	+	n_{i1}	n_{i2}	…	n_{ik}	n_i
	−	m_{i1}	m_{i2}	…	m_{ik}	m_i
Score		t_{i1}	t_{i2}	…	t_{ik}	N_i
		x_1	x_2	…	x_k	

Notation

- Let p_{ij} = proportion of subjects with disease among subjects in the ith stratum and jth exposure category

- $O = \sum_{i=1}^{s} O_t = \sum_{i=1}^{s} \sum_{j=1}^{k} n_{ij} x_j$
- $E = \sum_{i=1}^{s} E_t = \sum_{i=1}^{s} \left[\left(\sum_{j=1}^{k} t_{ij} x_j \right) n_i / N_i \right]$
- $V = \sum_{i=1}^{s} V_t = \sum_{i=1}^{s} \frac{n_i m_i \left(N_i s_{2i} - s_{1i}^2 \right)}{N_i^2 (N_i - 1)}$
- $s_{1i} = \sum_{j=1}^{k} t_{ij} x_j \, , i = 1, \ldots, s$
- $s_{2i} = \sum_{j=1}^{k} n_{ij} x_j^2 , i = 1, \ldots, s$

Mantel Extension Test

- Chi-Square Test for Trend-Multiple Strata

- To test the hypothesis H_0: $\beta = 0$ vs. H_1: $\beta \neq 0$, where $p_{ij} = \alpha_i + \beta x_j$
- We compute the test statistic

$$X_{TR}^2 = (|O - E| - 0.5)^2 / V \sim \chi_1^2 \text{ under } H_0$$

- only be used if $V \geq 5$.

Power and Sample-Size Estimation

for Stratified Categorical Data

Example
Cancer

- A study was performed based on a sample of 106,330 women enrolled in the Nurses' Health Study (NHS) relating ever use of OCs at baseline (in 1976) to breast-cancer incidence from 1976 to 1980.

- Because both OC use and breast cancer are related to age, the data were stratified by 5-year age groups and the Mantel-Haenszel test was employed to test for this association.

- The estimated OR $\left(\widehat{OR}_{MH}\right)$ was 1.0 with 95% CI = (0.8, 1.3).
-
- What power did the study have to detect a significant difference if the underlying *OR* = 1.3?

Power Estimation

- Power Estimation for a Collection of 2 × 2 Tables Based on the Mantel-Haenszel Test with the common underlying *OR*

- Suppose we wish to relate a dichotomous disease variable *D* to a dichotomous exposure variable *E* and want to control for a categorical confounding variable *C*.

		Exposure		
		+	−	Total
Disease	+	a_i	b_i	N_{1i}
	−	c_i	d_i	N_{2i}
	Total	M_{1i}	M_{2i}	N_i

Contd.

- To test $H_0: OR = 1$ vs. $H_1: OR = \exp(\gamma)$ for $\gamma \neq 0$

- $N =$ size of the total study population
- $r_i =$ proportion of exposed subjects in stratum i
- $s_i =$ proportion of diseased subjects in stratum i
- $t_i =$ proportion of total study population in stratum i

- With a significance level of α, the power is given by

$$\text{Power} = \Phi\left[\frac{\sqrt{N}\left(\gamma B_1 + \frac{\gamma^2}{2}B_2\right) - z_{1-\alpha/2}\sqrt{B}}{(B_1 + \gamma B_2)^{1/2}}\right]$$

- Where

$$B_1 = \sum B_{1i}, \quad B_{1i} = r_i s_i t_i (1 - r_i)(1 - s_i)$$
$$B_2 = \sum B_{2i}, \quad B_{2i} = B_{1i}(1 - 2r_i)(1 - 2s_i)$$

Sample-Size Estimation

- Sample-Size Estimation for a Collection of 2 × 2 Tables Based on the Mantel-Haenszel Test
- $N = \left(z_{1-\alpha/2}\sqrt{B} + z_{1-\beta}\sqrt{B_1 + \gamma B_2}\right)^2 / \left(\gamma B_1 + \frac{\gamma^2}{2} B_2\right)^2$

- Where
  - $\alpha =$ type I error,
  - $1 - \beta =$ power,
  - $\gamma = \ln OR$ under H_1,
  - and B_1, B_2 are defined before.

R package
for Epidemiologic Studies

- stats
  - (prop.test, prop.trend.test, fisher.test)
- epicalc
  - (cs, cc, ci, mhor, matchTab)
- rateratio.test
  - (rateratio.test)
- epiR
  - (epi.2by2, epi.kappa)