Analytic Methods for Applied Epidemiology: Framework and Contingency Table Analysis

2014 Maternal and Child Health Epidemiology Training

Pre-Training Webinar: Friday, May 16 2-4pm Eastern

Kristin Rankin, PhD

The Epidemiologic Framework

Descriptive Epidemiology

- summarizes the distribution of risk markers, risk factors, and outcomes in a community without explicit causal or other hypotheses.

Analytic Epidemiology

- tests hypotheses about risk factors and their relationship to health outcomes. In addition, public health uses analytic epidemiology to test hypotheses about health services, health policies, and the social determinants of health, and conduct program evaluations

The Epidemiologic Framework

From Galea 2013:
“An Argument for a Consequentialist Epidemiology” (AJE)

“Epidemiology is the study of the causes and distributions of diseases in human populations so that we may identify ways to prevent and control disease”

“…In recent decades, our discipline's robust interest in identifying causes has come at the expense of a more rigorous engagement with the second part of our vision for ourselves—the intent for us to intervene—and this approach threatens to diminish our field's relevance.”

The Epidemiologic Framework

All epidemiologic analysis should support the MCH planning cycle from Needs Assessment to Evaluation

Epidemiologic analysis starts with establishing the research question(s) that will best support the specific planning cycle activity at hand (e.g. problem analysis for prioritization or planning, evaluation of the causal effect of a program on intended outcomes

A conceptual framework should guide the specification of the research question(s) and the subsequent analysis process

Research Questions

"Well-crafted questions guide the systematic planning of research. Fomulating your questions precisely enables you to design a study with a good chance of answering them."

-Light, Singer, Willett, By Design (1990)

Developing a Research Question

- Identify a topic that is relevant to your agency and your population of interest

- Review the scientific literature to find out the state of the science on that topic and what gaps are in the literature about that topic

- Discover what is novel or will advance scientific knowledge or influence health programs or policy

http://phs.ucdavis.edu/downloads/ClinEpi_Romano.pdf

Developing a Research Question

Assess the feasibility of the research question

- Retrospective study: Adequacy of sample size available, availability and quality of existing data, feasibility with regard to data access/data linkages

- Prospective study: Ethics, sample availability, affordability/funding, manageable in scope

Developing a Research Question

A good research question should have the following elements, which should all be defined precisely:

P: Population
I: Intervention or Indicator (“Exposure”)
C: Comparison/Control (“Unexposed”)
O: Outcome (must be measureable)

Note: For some more exploratory research questions, the “I” and the “C” may be a combined set of factors for which you are examining relationships with the outcome, so may not be defined separately

Conceptual Models

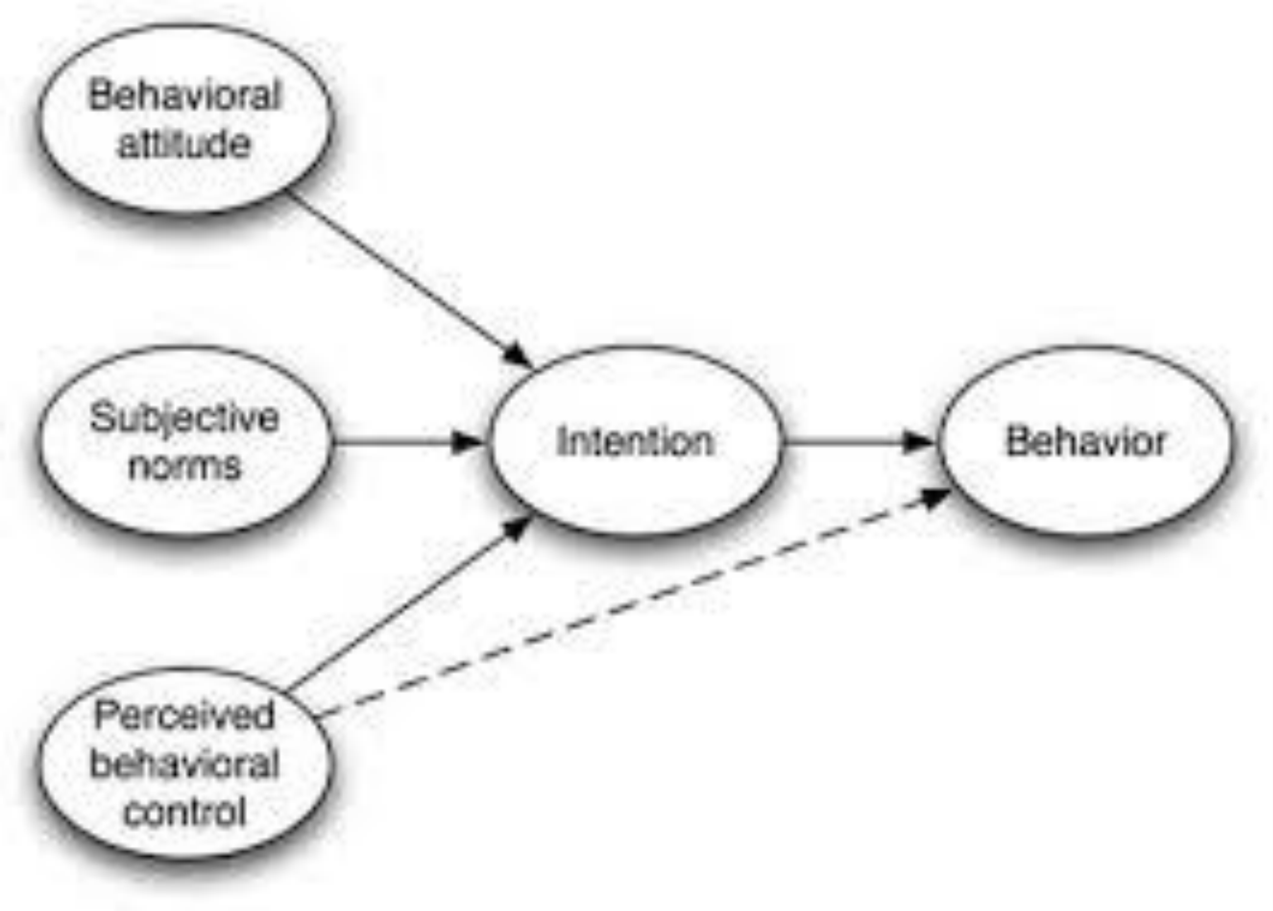
Conceptual models provide a framework to your research question, based on theory and prior research

Conceptual models can be broad (theoretical) or very specific (directed acyclic graphs), but should always be considered before launching into a data analysis

The analysis plan should flow from your conceptual model about how all of the variables (exposure, outcome, covariates) are related to one another

Conceptual Models - Examples

Theory of Planned Behavior (Ajzen)



Conceptual Models - Examples

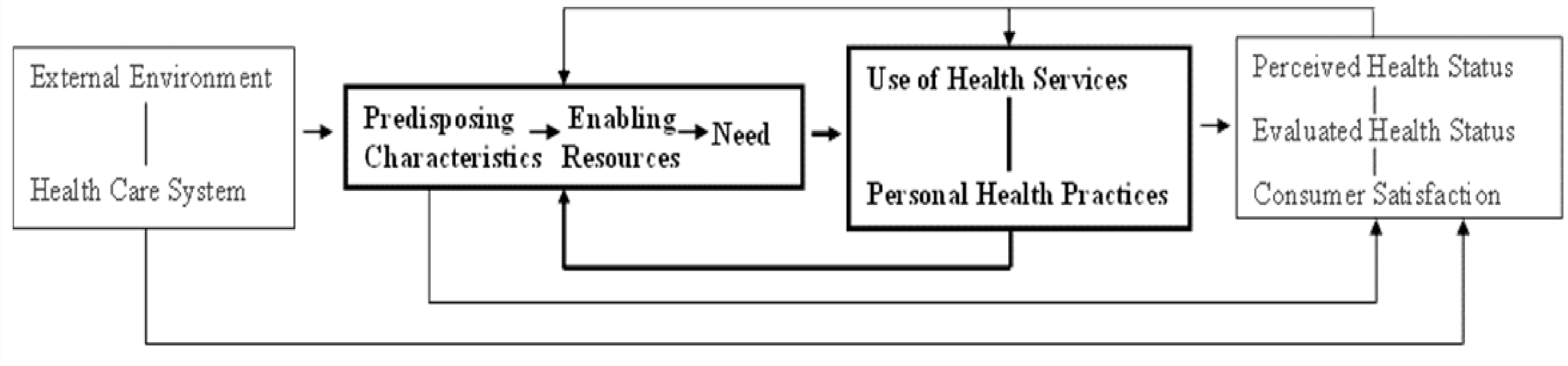
Behavioral Model of Healthcare Utilization (Anderson)

ENVIRONMENT

POPULATION CHARACTERISTICS

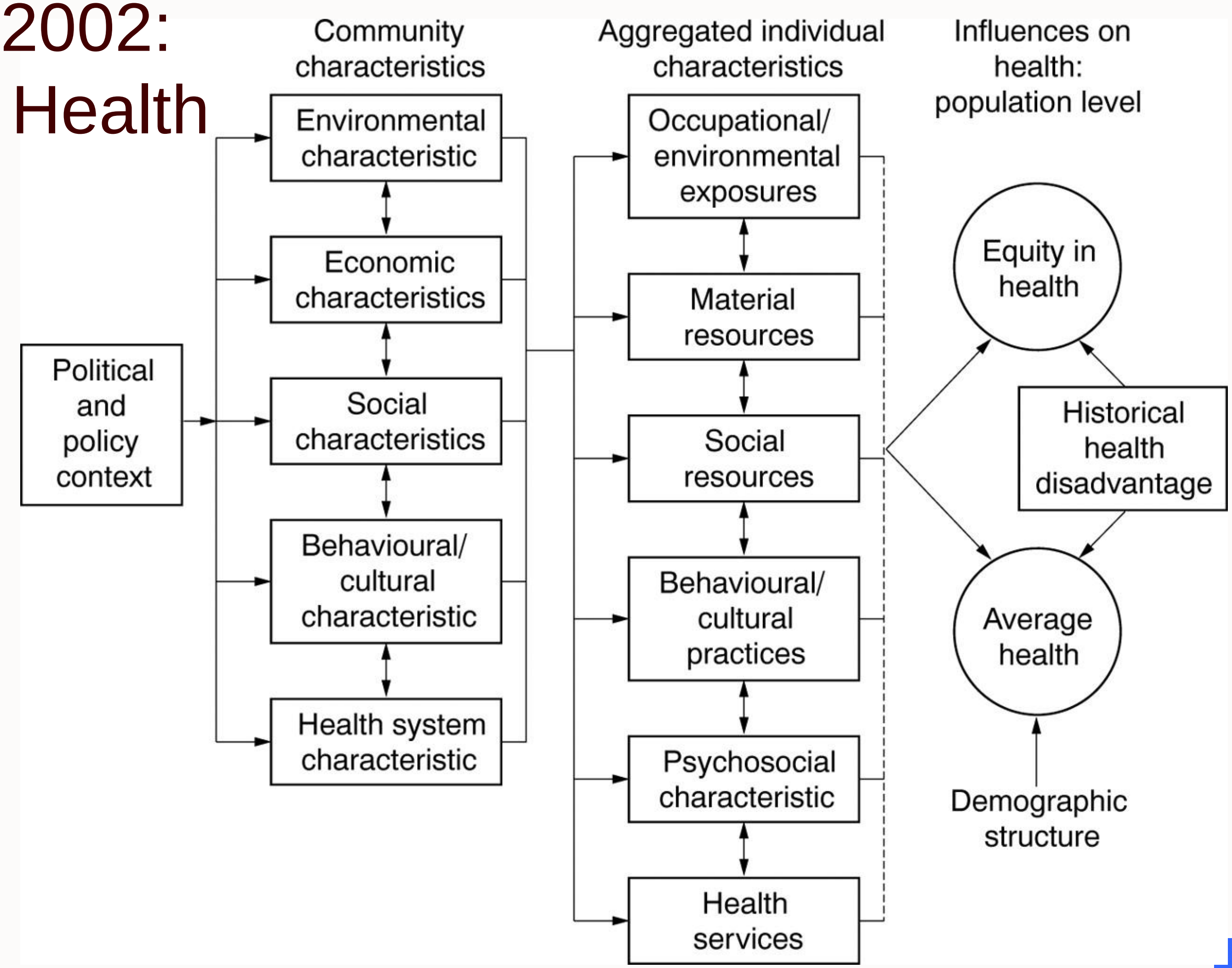
HEALTH BEHAVIOR

OUTCOMES



Conceptual Models - Examples

Starfield 2002:
Equity in Health


Community characteristics
Aggregated individual characteristics
Influences on health: population level
Political and policy context
Environmental characteristic
Economic characteristics
Social characteristics
Behavioural/ cultural characteristic
Health system characteristic
Occupational/ environmental exposures
Material resources
Social resources
Behavioural/ cultural practices
Psychosocial characteristic
Health services
Equity in health
Historical health disadvantage
Average health
Demographic structure

Paradies Y , and Stevens M J Epidemiol Community Health 2005;59:1012-1013

JECH

Defining Variables

Variable definition flows from the conceptual framework and research question

In this step, decisions are made about how to best operationalize concepts from the conceptual model to answer the research question

Types of variables:

- Outcome/Dependent Variable
- Exposure(s)/Risk Factor(s)/Program(s)/Intervention(s):

  - ***potentially modifiable***
- Covariates/Risk markers/Characteristics:

  - potential effect modifiers or confounders of exposure-outcome relationship

For each, decide how you will code variables for analysis to best answer your research question(s), balancing the need for sufficient sample size in categories with an approach that minimizes misclassification and preserves conceptual meaning (iterative process)

Univariate Statistics

- Describe the distribution of sample characteristics
- Estimate the occurrence of the outcome and exposure(s) in the target population

Measures of Occurrence in Epi:

- **Means** summarize continuous variables and are assumed to follow a *normal distribution*.
- **Proportions** summarize discrete variables and are assumed to follow the *binomial distribution*.
  - Cumulative incidence rates
  - Prevalence rates
- **Incidence Density Rates** also summarize discrete variables, but have person-time denominators and are assumed to follow the *Poisson distribution*.

Bivariate Statistics (*2 x 2 or k x k tables*)

- Prevalence or risk of outcome by categories of the exposure variable(s) - p1, p2, p3, etc (“row percents”)

- Crude measure(s) of effect (prevalence ratio, relative risk, odds ratio) for the relationship between the main exposure(s) and outcome and/or statistical tests comparing proportions, odds, means of the outcome for different exposure groups

- Above, repeated for relationship between covariates and outcome and covariates and exposure

Comparing Measures of Occurrence across Groups

Compare 2 or more proportions using chi-square tests

- Pearson’s chi-square (“General association”)
  - 2 x 2 table - 1 degree of freedom (df)
  - r x c table - (r-1) * (c-1) df *(unordered vars)*

- Mantel-Haenszel chi-square Test for Trend (“Nonzero correlation”)
  - r x 2 – 1 df (*ordered by dichotomous variable*)

Comparing Measures of Occurrence across Groups

- Compare 2 means – Independent sample t-test (Proc ttest)

- Compare 3 or more means – Analysis of Variance (ANOVA)
  - Continuous or ordinal variables: Proc ANOVA
  - Ordinal Variables: Proc freq – CMH statistics
  - (“Row Mean Scores Differ)

Review: Chi-Square and Other Tests for Proportions

		Outcome			
	Level of Measurement	**Continuous**	**Ordinal**	**Categorical/ Nominal (3+ categories)**	**Dichotomous**
Exposure	**Continuous**	--	--	--	--
	Ordinal	--	Trend – Chi-Square with 1 degree of freedom ***(Nonzero Correlation)***	Pearson Chi-Square Test with (r-1) * (c-1) df **(*General Association*)**	Trend – Mantel-Haenszel Chi-Square with 1 degree of freedom ***(Nonzero Correlation)***
	Categorical/ Nominal (3+categories)	--	ANOVA with r-1 degrees of freedom (***Row Mean Scores Differ***)	Pearson Chi-Square Test with (r-1)*(c-1) df (***General Association***)	Pearson Chi-Square Test with (r-1) * (c-1) df **(*General Association*)**
	Dichotomous	--			Pearson Chi-Square with 1 df (***General Association***);

Measures of Association from 2x2 Tables

	Outcome		
Exposure	Y	N	
Y	a	b	n_1
N	c	d	n_2
	m_1	m_2	

Difference Measures

- Between two or more proportions

  Risk Difference/Attributable Risk = $a/n_1 - c/n_2$

Ratio Measures

- Between two or more proportions

  Relative Risk (cohort)

  Relative Prevalence (cross-sectional)

  $$\left.\begin{array}{l}\\ \\ \end{array}\right\} \frac{a/n_1}{c/n_2}$$

- Between two odds

  Odds Ratio (case-control) = $\dfrac{a*d}{b*c}$

Using SAS for Measures of Association

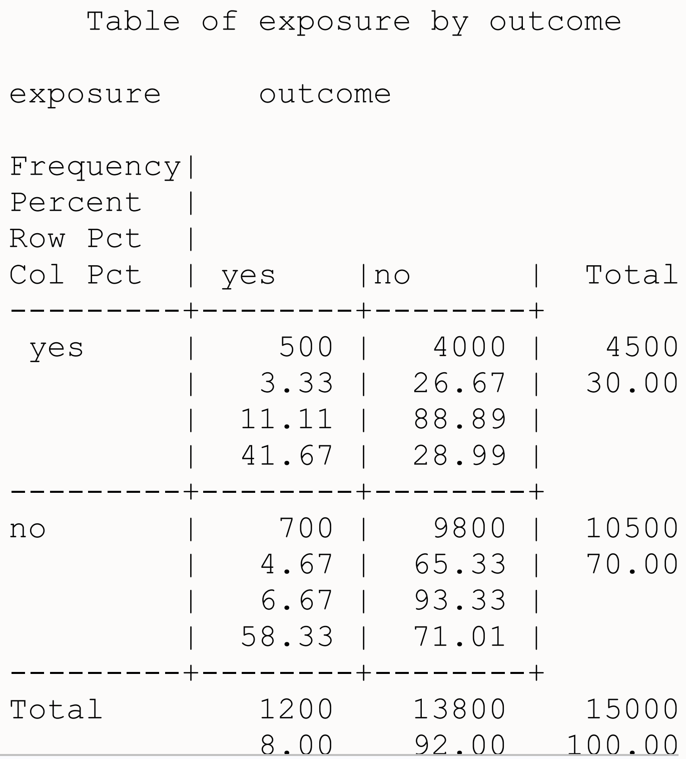
Crude Association: 2x2 Table

Crude RR = 1.67

Crude RD/AR = 0.044

Exposure Prevalence =30%

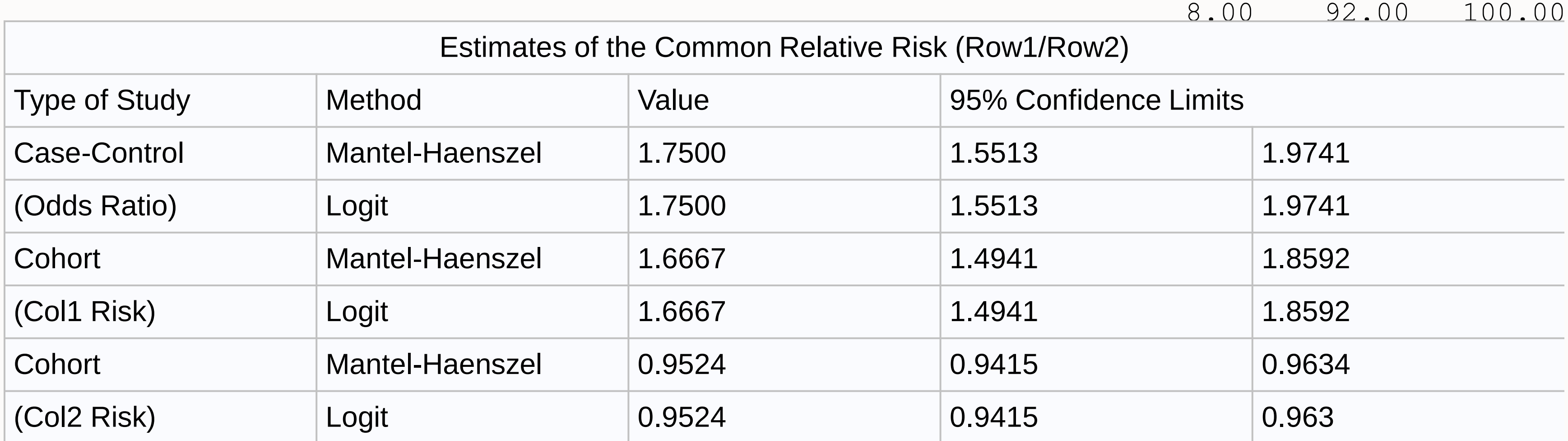
Disease Prevalence = 8%



```
          Table of exposure by outcome

exposure     outcome

Frequency|
Percent  |
Row Pct  |
Col Pct  | yes     |no       |  Total
---------+---------+---------+
 yes     |     500 |    4000 |   4500
         |    3.33 |   26.67 |  30.00
         |   11.11 |   88.89 |
         |   41.67 |   28.99 |
---------+---------+---------+
no       |     700 |    9800 |  10500
         |    4.67 |   65.33 |  70.00
         |    6.67 |   93.33 |
         |   58.33 |   71.01 |
---------+---------+---------+
Total         1200     13800    15000
              8.00     92.00   100.00
```



Estimates of the Common Relative Risk (Row1/Row2)				
Type of Study	Method	Value	95% Confidence Limits	
Case-Control	Mantel-Haenszel	1.7500	1.5513	1.9741
(Odds Ratio)	Logit	1.7500	1.5513	1.9741
Cohort	Mantel-Haenszel	1.6667	1.4941	1.8592
(Col1 Risk)	Logit	1.6667	1.4941	1.8592
Cohort	Mantel-Haenszel	0.9524	0.9415	0.9634
(Col2 Risk)	Logit	0.9524	0.9415	0.963

Using SAS Formatting to Control Output:

Suppose there is a variable called "marital" which contains information on marital status. The variable is coded "1" for individuals who are not married and "0" for those who are married.

```
proc format;
  value marital
  1 = 'not married'
  0 = 'married';

  value yn
  1 = ' yes'
  0 = 'no';
```

Sometimes proc format is used simply to attach descriptive labels to variable values. In addition, proc format can be used to **control the ordering of variable values in analysis (*yn format above*).**

Using SAS Formatting to Control Output:

The format must be associated with the variable, and this can be accomplished either in the data step or in a proc step. In the data step, the association is permanent; in a proc step it is only temporary.

The formatting can then be used to control the way that frequency tables are produced:

```
data one; set data.final;
   …other SAS statements…
   format exposure outcome yn.;
run;

proc freq order=formatted;
   tables exposure * outcome;
run;
```

```
proc freq order=formatted;
    tables exposure * outcome;
    format exposure outcome yn.;
 run;
```

Using SAS for Measures of Association

Crude Association: >2 X 2 Table

Race:

1 = Native American

2 = Asian / Pacific Islander

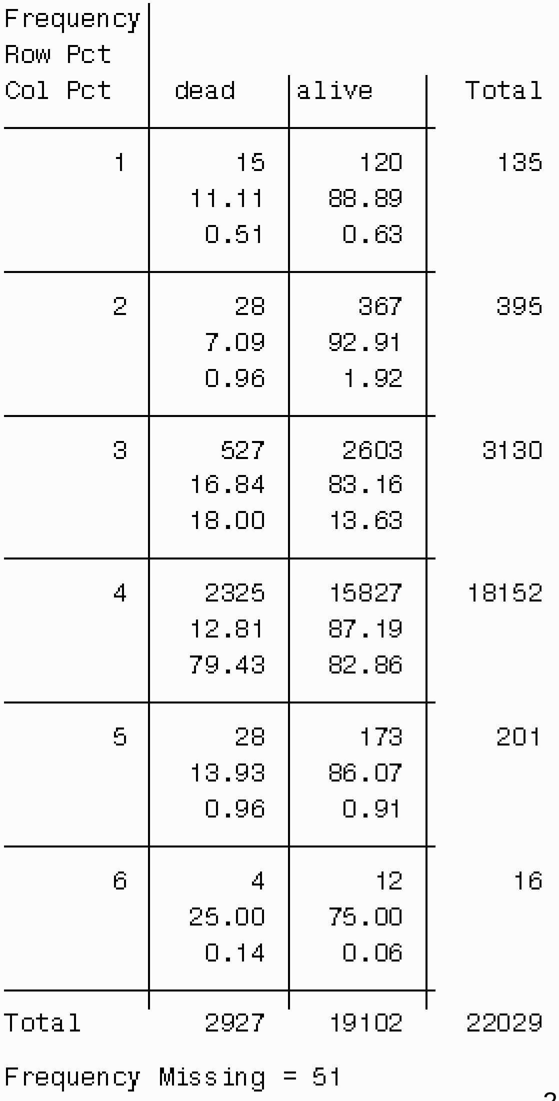
3 = Black

4 = White

5 = Other

6 = Multiple races

Choose common referent group to be the “unexposed” group and estimate RRs from multiple 2x2 tables



```
Frequency|
Row Pct  |
Col Pct  |  dead  |alive   |  Total
---------+--------+--------+
       1 |     15 |    120 |    135
         |  11.11 |  88.89 |
         |   0.51 |   0.63 |
---------+--------+--------+
       2 |     28 |    367 |    395
         |   7.09 |  92.91 |
         |   0.96 |   1.92 |
---------+--------+--------+
       3 |    527 |   2603 |   3130
         |  16.84 |  83.16 |
         |  18.00 |  13.63 |
---------+--------+--------+
       4 |   2325 |  15827 |  18152
         |  12.81 |  87.19 |
         |  79.43 |  82.86 |
---------+--------+--------+
       5 |     28 |    173 |    201
         |  13.93 |  86.07 |
         |   0.96 |   0.91 |
---------+--------+--------+
       6 |      4 |     12 |     16
         |  25.00 |  75.00 |
         |   0.14 |   0.06 |
---------+--------+--------+
Total        2927    19102    22029

Frequency Missing = 51
```

Categorical Exposures

You may be able to visually inspect patterns of row percents and estimate RRs in your head for different levels of ordinal or nominal variables, compared to a common referent group

If not, you want to explicitly produce these relative risks from a series of 2x2 tables to better understand the exposure – disease relationship for exposure and/or outcome variables with >2 categories:

```
proc freq order=formatted;
   tables race*outcome/ nopercent cmh relrisk;
   where race in (3,4); /*Switch 3 to 1,2,5,6
   in subsequent proc freqs to get RRs for all
   race groups*/
run;/*race = 4 - White is ref category*/
```

Stratified Analysis

Single and multiple factor stratified analysis are used to:

- Examine outcome prevalence/risk in finer subgroups
- Assess for small numbers in cells upon stratification and distribution of sample across finer subgroups
- Assess **effect modification** and **confounding** of exposure(s)/outcome relationship by single covariates and sets of covariates

Effect Modification

Interaction/effect modification is “a situation in which two or more risk factors modify the effect of each other with regard to the occurrence or level of a given outcome”

- Risk factor has different effect on outcome across subgroups of the population (e.g. by race/ethnicity, gender, age, insurance status, poverty level, etc)
- Program has differential effects across different subgroups of the population
- Risk factors jointly cause an outcome in a synergistic way

Effect Modification

Effect Modification is assessed by comparing **stratum-specific** measures of association to each other

- Effect Modification on Multiplicative Scale / Multiplicative Interaction: Stratum-specific RRs/ORs are different

- Effect Modification on Additive Scale / Additive Interaction: Stratum-specific RDs/ARs are different

Effect Modification on Multiplicative Scale

Statistical test for homogeneity of the stratum-specific effect measures – multiplicative scale

Breslow-Day Test for Homogeneity

$$\boxed{\chi^2_{\#\text{strata}-1} = \sum_{\text{i}=1}^{\#\text{strata}} \frac{\left(\text{stratum-specific measure}_{\text{i}} - \text{adjusted measure}\right)^2}{\text{Variance}\left(\text{stratum-specific measure}_{\text{i}}\right)}}$$

Effect Modification

Regardless of the method, if the stratum-specific estimates differ, statistically or qualitatively, then effect modification may be present.

Effect modification can take any of the following forms:

1. Stratum-specific estimates on the same side of the null but with meaningfully different magnitudes;
2. One null stratum-specific estimate and another significantly or meaningfully different than the null; or
3. Stratum-specific estimates on opposite sides of the null

Stratified Results by Covariate “A”

```
      Table 1 of exposure by outcome
      Controlling for covariate=yes
                                   A
exposure     outcome

Frequency|
Percent  |
Row Pct  |
Col Pct  | yes    |no      |  Total
---------+--------+--------+
 yes     |    250 |   1250 |   1500
         |   5.00 |  25.00 |  30.00
         |  16.67 |  83.33 |
         |  41.67 |  28.41 |
---------+--------+--------+
no       |    350 |   3150 |   3500
         |   7.00 |  63.00 |  70.00
         |  10.00 |  90.00 |
         |  58.33 |  71.59 |
---------+--------+--------+
Total         600     4400     5000
            12.00    88.00   100.00
```

```
      Table 2 of exposure by outcome
       Controlling for covariate=no
                                  A
exposure     outcome

Frequency|
Percent  |
Row Pct  |
Col Pct  | yes    |no      |  Total
---------+--------+--------+
 yes     |    250 |   2750 |   3000
         |   2.50 |  27.50 |  30.00
         |   8.33 |  91.67 |
         |  41.67 |  29.26 |
---------+--------+--------+
no       |    350 |   6650 |   7000
         |   3.50 |  66.50 |  70.00
         |   5.00 |  95.00 |
         |  58.33 |  70.74 |
---------+--------+--------+
Total         600     9400    10000
             6.00    94.00   100.00
```

-What is the prevalence of the outcome for the exposed and unexposed in each stratum of Covariate A?

-What are the stratum-specific RRs?

-Do they differ from each other?

Stratified Analysis: Is Covariate “A” an Effect Modifier of the Exposure-Outcome Relationship?

```
    Table 1 of exposure by outcome
    Controlling for covariate=yes
                                 A
exposure     outcome

Frequency|
Percent  |
Row Pct  |
Col Pct  | yes    |no      |  Total
---------+--------+--------+
 yes     |    250 |   1250 |   1500
         |   5.00 |  25.00 |  30.00
         |  16.67 |  83.33 |
         |  41.67 |  28.41 |
---------+--------+--------+
no       |    350 |   3150 |   3500
         |   7.00 |  63.00 |  70.00
         |  10.00 |  90.00 |
         |  58.33 |  71.59 |
---------+--------+--------+
Total         600     4400     5000
            12.00    88.00   100.00
```

```
    Table 2 of exposure by outcome
     Controlling for covariate=no
                                 A
exposure     outcome

Frequency|
Percent  |
Row Pct  |
Col Pct  | yes    |no      |  Total
---------+--------+--------+
 yes     |    250 |   2750 |   3000
         |   2.50 |  27.50 |  30.00
         |   8.33 |  91.67 |
         |  41.67 |  29.26 |
---------+--------+--------+
no       |    350 |   6650 |   7000
         |   3.50 |  66.50 |  70.00
         |   5.00 |  95.00 |
         |  58.33 |  70.74 |
---------+--------+--------+
Total         600     9400    10000
             6.00    94.00   100.00
```

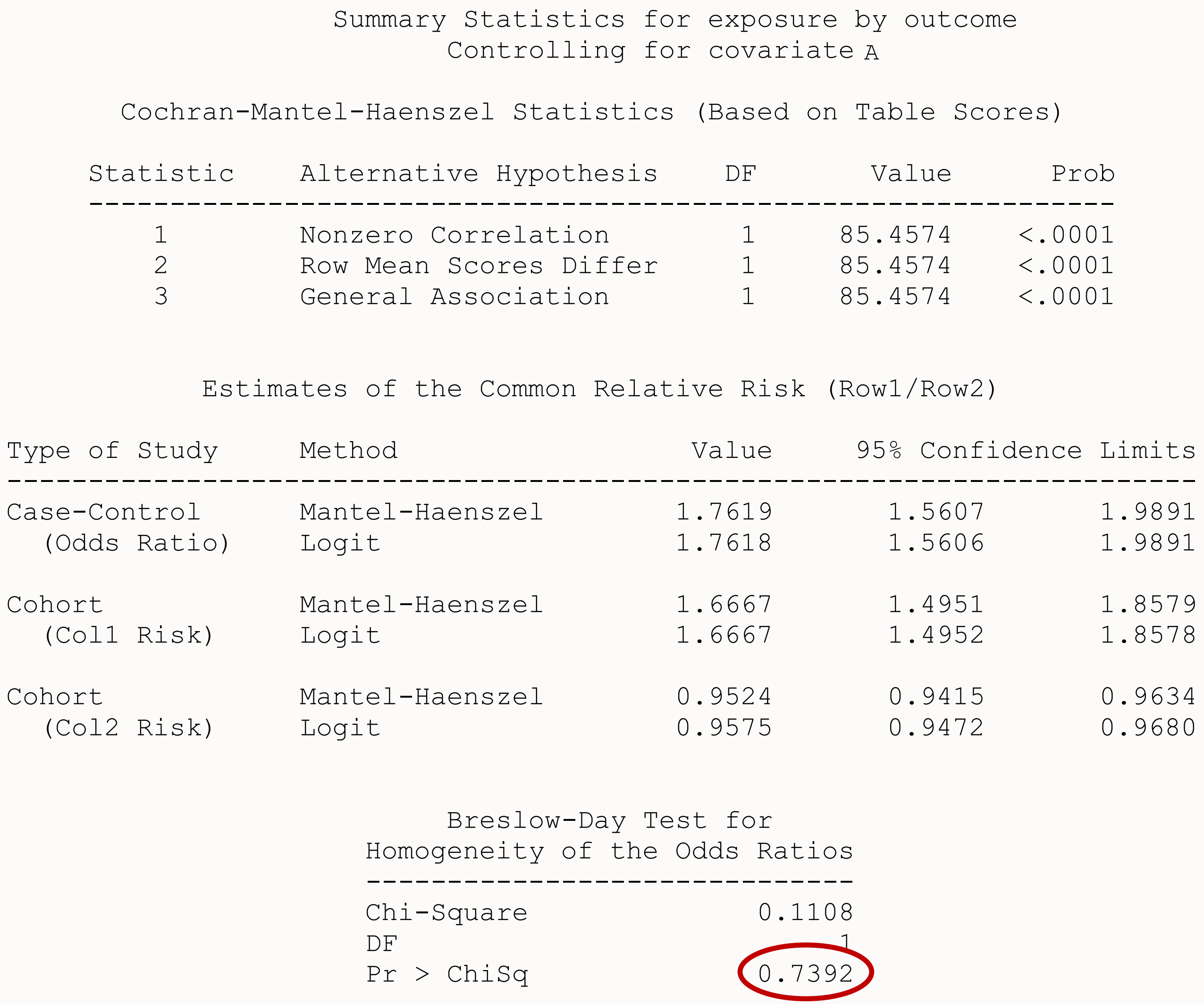
% with OC among Exp = 16.7
% with OC among UnExp = 10.0

% with OC among Exp = 8.3
% with OC among UnExp = 5.0

Prevalence Ratio$_{covA=y}$: 1.7

Prevalence Ratio$_{covA=n}$: 1.7

Stratified Analysis: Is Covariate "A" an Effect Modifier of the Exposure-Outcome Relationship?



```
                    Summary Statistics for exposure by outcome
                          Controlling for covariate A

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

          Statistic    Alternative Hypothesis    DF       Value      Prob
          ---------------------------------------------------------------
              1        Nonzero Correlation        1     85.4574    <.0001
              2        Row Mean Scores Differ     1     85.4574    <.0001
              3        General Association        1     85.4574    <.0001


                 Estimates of the Common Relative Risk (Row1/Row2)

     Type of Study     Method                   Value     95% Confidence Limits
     -------------------------------------------------------------------------
     Case-Control      Mantel-Haenszel         1.7619       1.5607       1.9891
       (Odds Ratio)    Logit                   1.7618       1.5606       1.9891

     Cohort            Mantel-Haenszel         1.6667       1.4951       1.8579
       (Col1 Risk)     Logit                   1.6667       1.4952       1.8578

     Cohort            Mantel-Haenszel         0.9524       0.9415       0.9634
       (Col2 Risk)     Logit                   0.9575       0.9472       0.9680


                              Breslow-Day Test for
                           Homogeneity of the Odds Ratios
                           ------------------------------
                           Chi-Square            0.1108
                           DF                         1
                           Pr > ChiSq            0.7392
```

Stratified Results by Covariate “C”

```
   Table 1 of exposure by outcome
   Controlling for covariate=yes
                    C
exposure     outcome

Frequency|
Percent  |
Row Pct  |
Col Pct  | yes    |no      |  Total
---------+--------+--------+
 yes     |    365 |   2135 |   2500
         |   7.30 |  42.70 |  50.00
         |  14.60 |  85.40 |
         |  67.59 |  47.87 |
---------+--------+--------+
no       |    175 |   2325 |   2500
         |   3.50 |  46.50 |  50.00
         |   7.00 |  93.00 |
         |  32.41 |  52.13 |
---------+--------+--------+
Total         540     4460     5000
            10.80    89.20   100.00
```

```
   Table 2 of exposure by outcome
    Controlling for covariate=no
                    C
exposure     outcome

Frequency|
Percent  |
Row Pct  |
Col Pct  | yes    |no      |  Total
---------+--------+--------+
 yes     |    135 |   1865 |   2000
         |   1.35 |  18.65 |  20.00
         |   6.75 |  93.25 |
         |  20.45 |  19.97 |
---------+--------+--------+
no       |    525 |   7475 |   8000
         |   5.25 |  74.75 |  80.00
         |   6.56 |  93.44 |
         |  79.55 |  80.03 |
---------+--------+--------+
Total         660     9340    10000
             6.60    93.40   100.00
```

-What is the prevalence of the outcome for the exposed and unexposed in each stratum of Covariate C?

-What are the stratum-specific RRs?

-Do they differ from each other?

Stratified Analysis: Is Covariate C an Effect Modifier?

```
   Table 1 of exposure by outcome
   Controlling for covariate=yes
                                   C
exposure      outcome

Frequency|
Percent  |
Row Pct  |
Col Pct  | yes     |no       |  Total
---------+--------+--------+
 yes     |    365 |   2135 |   2500
         |   7.30 |  42.70 |  50.00
         |  14.60 |  85.40 |
         |  67.59 |  47.87 |
---------+--------+--------+
no       |    175 |   2325 |   2500
         |   3.50 |  46.50 |  50.00
         |   7.00 |  93.00 |
         |  32.41 |  52.13 |
---------+--------+--------+
Total         540     4460     5000
            10.80    89.20   100.00
```

```
   Table 2 of exposure by outcome
    Controlling for covariate=no
                                   C
exposure      outcome

Frequency|
Percent  |
Row Pct  |
Col Pct  | yes     |no       |  Total
---------+--------+--------+
 yes     |    135 |   1865 |   2000
         |   1.35 |  18.65 |  20.00
         |   6.75 |  93.25 |
         |  20.45 |  19.97 |
---------+--------+--------+
no       |    525 |   7475 |   8000
         |   5.25 |  74.75 |  80.00
         |   6.56 |  93.44 |
         |  79.55 |  80.03 |
---------+--------+--------+
Total         660     9340    10000
             6.60    93.40   100.00
```

% with OC among Exp = 14.6
% with OC among UnExp = 7.0

Prevalence Ratio$_{covC=y}$: 2.1

% with OC among Exp = 6.8
% with OC among UnExp = 6.6

Prevalence Ratio$_{covC=n}$: 1.0

Stratified Analysis: Is Covariate C an Effect Modifier?

```
                    Summary Statistics for exposur by outcome
                           Controlling for covariate C

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

        Statistic    Alternative Hypothesis    DF       Value      Prob
        ---------------------------------------------------------------
            1        Nonzero Correlation        1     43.8370    <.0001
            2        Row Mean Scores Differ     1     43.8370    <.0001
            3        General Association        1     43.8370    <.0001


                 Estimates of the Common Relative Risk (Row1/Row2)

   Type of Study     Method                 Value     95% Confidence Limits
   -------------------------------------------------------------------------
   Case-Control      Mantel-Haenszel       1.5677       1.3756       1.7865
     (Odds Ratio)    Logit                 1.5498       1.3525       1.7759

   Cohort            Mantel-Haenszel       1.5091       1.3342       1.7068
     (Col1 Risk)     Logit                 1.4978       1.3218       1.6973

   Cohort            Mantel-Haenszel       0.9631       0.9524       0.9740
     (Col2 Risk)     Logit                 0.9723       0.9618       0.9829


                           Breslow-Day Test for
                       Homogeneity of the Odds Ratios
                       ------------------------------
                       Chi-Square               32.6678
                       DF                             1
                       Pr > ChiSq               <.0001
```

Confounding

Confounding refers to “a situation in which a non-causal association between a given exposure and an outcome is observed as a result of the influence of a third variable (or group of variables)”

Three criteria for confounder:

1. **Causally associated with the outcome**
2. **Non-causally or causally associated with exposure**
3. **Not in causal pathway between exposure and outcome**

Confounding

One way to assess for confounding by a third factor is by comparing the **crude** measure of association for the relationship between the exposure and outcome to the relative risk or odds ratio after **adjusting** for the suspected confounder

- There is no statistical test for confounding
- By convention, if the adjusted measure is > 10% different than the crude, confounding is considered to be present

Confounding

Confounding may be…

- **Positive** : Makes crude measure of association for E-D relationship stronger (*farther from the null*) than true relationship; in other words, after adjustment, the measure of association gets **closer to** the null
- **Negative**: Makes crude measure of association for E-D relationship weaker (*closer to the null*) than true relationship; in other words, after adjustment, the measure of association gets **farther from** the null

Confounding

Confounding may be:

Full: Completely explains the association between exposure and outcome, so that after adjustment, measure of association is 1 (null value)

Partial: **Incompletely** explains the association between exposure and outcome, so that after adjustment, measure of association is closer to, but not at the null value

Residual confounding may result from misclassification or measurement error for measured confounders or from unmeasured confounders

Confounding

Methods to Control Confounding

- Randomization
- Restriction
- Matching
- Direct and Indirect Standardization
- **Stratified Analysis**
- Regression Analysis

Randomization, restriction and matching occur in the study design phase, ***prior*** to analysis—the researcher assumes that confounding is present. The other methods are all used to assess confounding ***after*** data are collected, during analysis.

Confounding

Assessing Confounding

- Standardization: Does the standardized measure differ from the unstandardized measure?
- **Stratified Analysis**: Does the adjusted measure of association differ from the crude measure of association?
- Regression Analysis: Does the beta coefficient for a variable in a model that includes a potential confounder differ from the beta coefficient for that same variable in a model that does not include the potential confounder?

Confounding – Stratified Analysis

Typical Layout for Stratified Analysis

For cohort or cross-sectional data

Potential Confounder = 'Y'

Outcome

		Y	N
Exposure	Y	P_1	
	N	P_2	

Potential Confounder = 'N'

Outcome

		Y	N
Exposure	Y	P_3	
	N	P_4	

For case-control or other data when using the odds ratio

Potential Confounder = 'Y'

Disease

		Y	N
Exposure	Y	a	b
	N	c	d

Potential Counfounder = 'N'

Disease

		Y	N
Exposure	Y	e	f
	N	g	h

Confounding – Stratified Analysis

Typical data layout for stratified analysis:

- A table for each level (strata) of the potential confounder
- Each table displays the exposure-disease relationship within each stratum

Summary (adjusted) relative risk and odds ratio, and the corresponding statistical test using stratified methods:

$$\frac{\displaystyle\sum_{i=1}^{\#\text{ of strata}} \frac{a_i n_{2i}}{N_i}}{\displaystyle\sum_{i=1}^{\#\text{ of strata}} \frac{c_i n_{1i}}{N_i}} \qquad \frac{\displaystyle\sum_{i=1}^{\#\text{ of strata}} \frac{a_i d_i}{N_i}}{\displaystyle\sum_{i=1}^{\#\text{ of strata}} \frac{b_i c_i}{N_i}}$$

$$\chi^2 = \left(\frac{\displaystyle\sum_{i=1}^{\#\text{of strata}} a_i - \sum_{i=1}^{\#\text{of strata}} \frac{n_{1i} m_{1i}}{N_i}}{\sqrt{\displaystyle\sum_{i=1}^{\#\text{ of strata}} \frac{n_{1i} n_{2i} m_{1i} m_{2i}}{N_i{}^3}}} \right)^2$$

NOTE: This chi-square test estimates is for the relationship between the exposure and outcome, after adjustment for the confounder (*NOT a statistical test for confounding!*)

Using SAS for stratified analysis to Assess Confounding

Crude and single factor stratified analysis

```
proc freq order=formatted;
  tables exposure*outcome /*bivariate*/
         covariate*exposure*outcome / /*strat*/
         riskdiff relrisk cmh;
run;
```

The rightmost variable defines the columns; the second from the rightmost variable defines the rows; other variables define strata of covariates (potential effect modifiers or confounders)

Stratified Results by Covariate “A”

```
   Table 1 of exposure by outcome
   Controlling for covariate=yes
                             A
exposure     outcome

Frequency|
Percent  |
Row Pct  |
Col Pct  | yes    |no      |  Total
---------+--------+--------+
 yes     |    250 |   1250 |   1500
         |   5.00 |  25.00 |  30.00
         |  16.67 |  83.33 |
         |  41.67 |  28.41 |
---------+--------+--------+
no       |    350 |   3150 |   3500
         |   7.00 |  63.00 |  70.00
         |  10.00 |  90.00 |
         |  58.33 |  71.59 |
---------+--------+--------+
Total         600     4400     5000
            12.00    88.00   100.00
```

```
   Table 2 of exposure by outcome
    Controlling for covariate=no
                              A
exposure     outcome

Frequency|
Percent  |
Row Pct  |
Col Pct  | yes    |no      |  Total
---------+--------+--------+
 yes     |    250 |   2750 |   3000
         |   2.50 |  27.50 |  30.00
         |   8.33 |  91.67 |
         |  41.67 |  29.26 |
---------+--------+--------+
no       |    350 |   6650 |   7000
         |   3.50 |  66.50 |  70.00
         |   5.00 |  95.00 |
         |  58.33 |  70.74 |
---------+--------+--------+
Total         600     9400    10000
             6.00    94.00   100.00
```

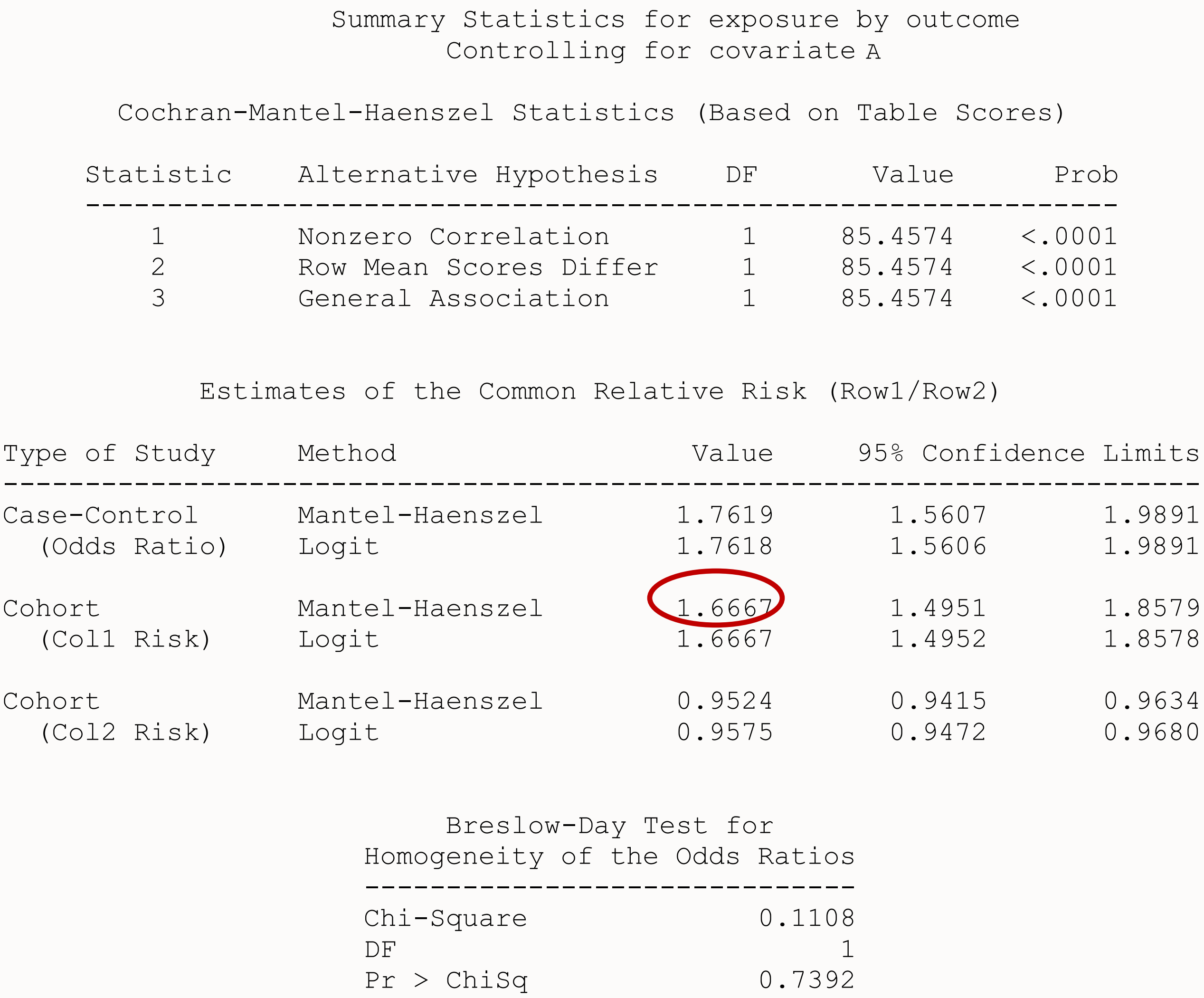
% with OC among Exp = 16.7
% with OC among UnExp = 10.0

% with OC among Exp = 8.3
% with OC among UnExp = 5.0

Crude RR = 1.7

What is the adjusted RR?

Stratified Analysis: Is Covariate “A” a Confounder?



```
                    Summary Statistics for exposure by outcome
                          Controlling for covariate A

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

        Statistic    Alternative Hypothesis    DF       Value      Prob
        ---------------------------------------------------------------
            1        Nonzero Correlation        1     85.4574    <.0001
            2        Row Mean Scores Differ     1     85.4574    <.0001
            3        General Association        1     85.4574    <.0001


                Estimates of the Common Relative Risk (Row1/Row2)

   Type of Study     Method                  Value     95% Confidence Limits
   -------------------------------------------------------------------------
   Case-Control      Mantel-Haenszel        1.7619       1.5607        1.9891
     (Odds Ratio)    Logit                  1.7618       1.5606        1.9891

   Cohort            Mantel-Haenszel        1.6667       1.4951        1.8579
     (Col1 Risk)     Logit                  1.6667       1.4952        1.8578

   Cohort            Mantel-Haenszel        0.9524       0.9415        0.9634
     (Col2 Risk)     Logit                  0.9575       0.9472        0.9680


                          Breslow-Day Test for
                      Homogeneity of the Odds Ratios
                      ------------------------------
                      Chi-Square              0.1108
                      DF                           1
                      Pr > ChiSq              0.7392
```

Stratified Results by Covariate “B”

```
   Table 1 of exposure by outcome
   Controlling for covariate=yes
                   B
exposure        outcome

Frequency|
Percent  |
Row Pct  |
Col Pct  | yes     |no       |  Total
---------+--------+--------+
 yes     |     350 |    2150 |   2500
         |    7.00 |   43.00 |  50.00
         |   14.00 |   86.00 |
         |   58.33 |   48.86 |
---------+--------+--------+
no       |     250 |    2250 |   2500
         |    5.00 |   45.00 |  50.00
         |   10.00 |   90.00 |
         |   41.67 |   51.14 |
---------+--------+--------+
Total          600      4400     5000
             12.00     88.00   100.00
```

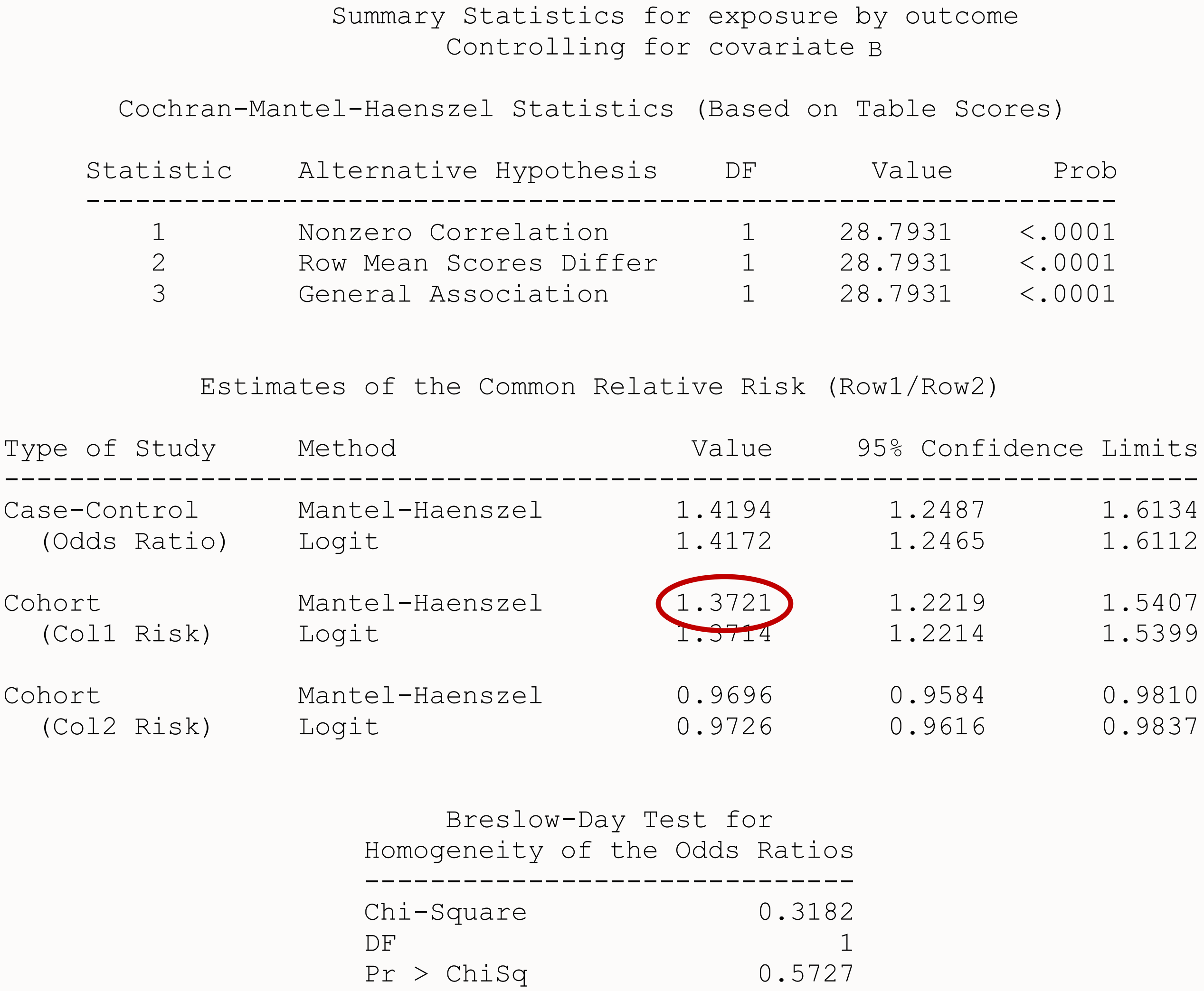
```
   Table 2 of exposure by outcome
   Controlling for covariate=no
               B
exposure        outcome

Frequency|
Percent  |
Row Pct  |
Col Pct  | yes     |no       |  Total
---------+--------+--------+
 yes     |     150 |    1850 |   2000
         |    1.50 |   18.50 |  20.00
         |    7.50 |   92.50 |
         |   25.00 |   19.68 |
---------+--------+--------+
no       |     450 |    7550 |   8000
         |    4.50 |   75.50 |  80.00
         |    5.63 |   94.38 |
         |   75.00 |   80.32 |
---------+--------+--------+
Total          600      9400    10000
              6.00     94.00   100.00
```

Crude RR = 1.7

What is the adjusted RR?

Stratified Analysis: Is Covariate B a Confounder?



```
                    Summary Statistics for exposure by outcome
                           Controlling for covariate B

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

         Statistic    Alternative Hypothesis    DF       Value      Prob
         ---------------------------------------------------------------
             1        Nonzero Correlation        1     28.7931    <.0001
             2        Row Mean Scores Differ     1     28.7931    <.0001
             3        General Association        1     28.7931    <.0001


                Estimates of the Common Relative Risk (Row1/Row2)

    Type of Study     Method                    Value     95% Confidence Limits
    -------------------------------------------------------------------------
    Case-Control      Mantel-Haenszel          1.4194       1.2487       1.6134
      (Odds Ratio)    Logit                    1.4172       1.2465       1.6112

    Cohort            Mantel-Haenszel          1.3721       1.2219       1.5407
      (Col1 Risk)     Logit                    1.3714       1.2214       1.5399

    Cohort            Mantel-Haenszel          0.9696       0.9584       0.9810
      (Col2 Risk)     Logit                    0.9726       0.9616       0.9837


                              Breslow-Day Test for
                          Homogeneity of the Odds Ratios
                          ------------------------------
                          Chi-Square              0.3182
                          DF                           1
                          Pr > ChiSq              0.5727
```

Confounding vs Effect Modification (Multiplicative)

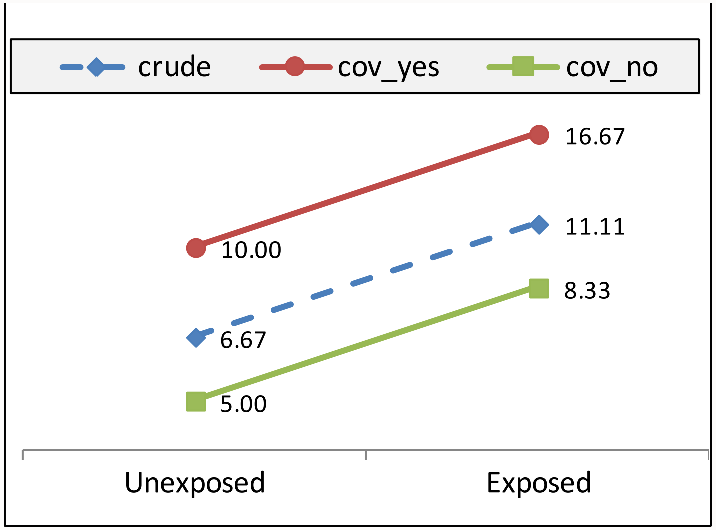
Confounding	Effect Modification
Compare crude v. adjusted OR/RR	Compare stratum-specific OR/RR
No statistical testing	Statistical testing

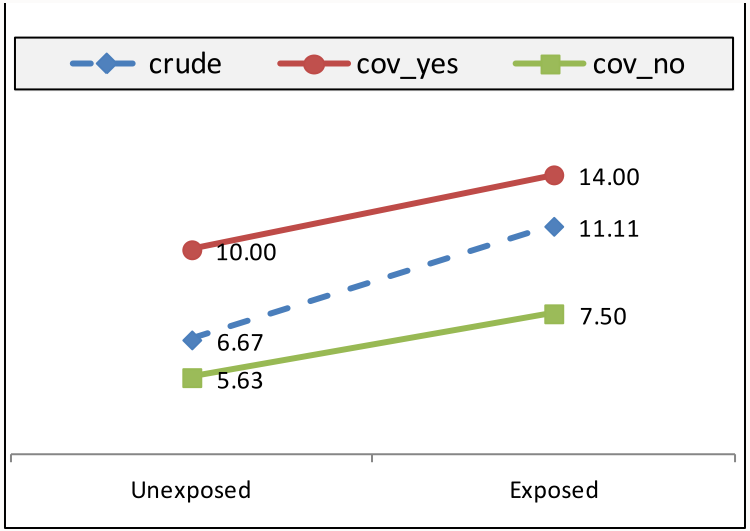
With **confounding**, the association between a risk factor and an outcome is the same (or close to the same) in each stratum, but different from the crude

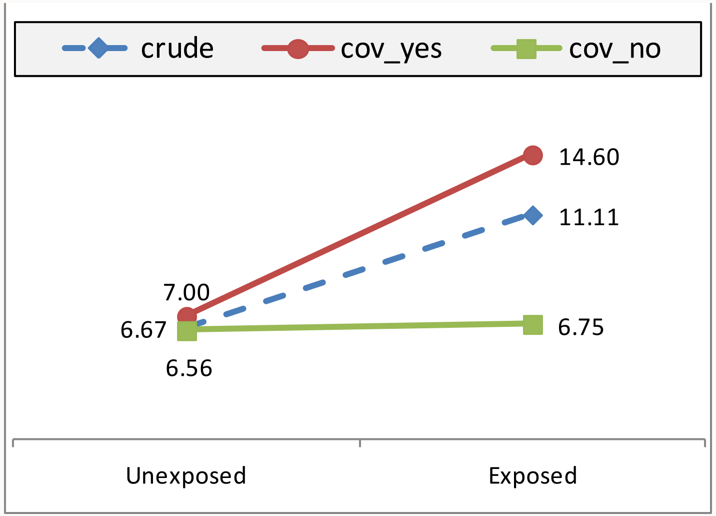
With **effect modification**, the association between a risk factor and a health outcome is different across strata.

Effect modification is ***always*** assessed first; confounding is meaningless in the presence of meaningfully and statistically significant effect modification

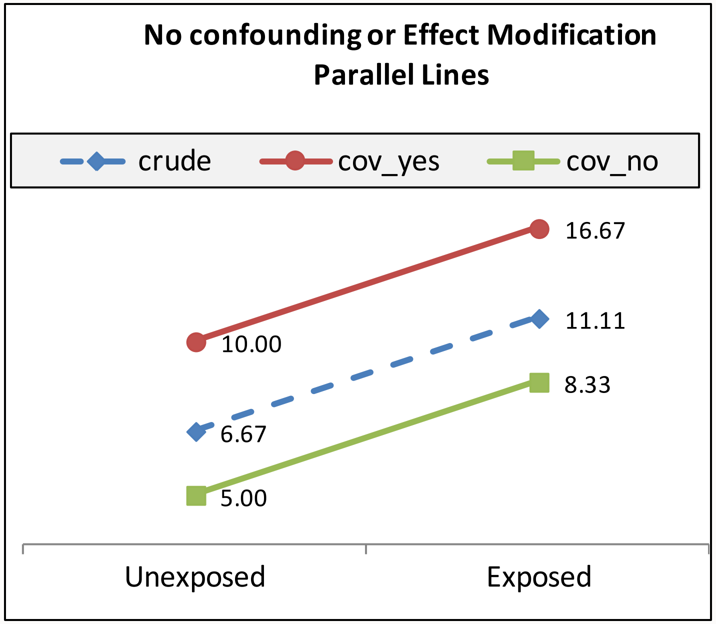
Confounding vs Effect Modification (Multiplicative)

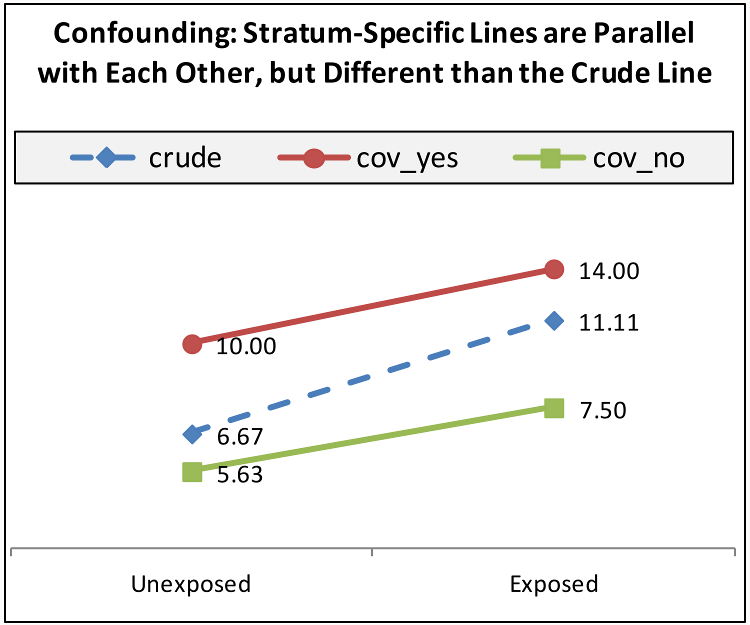


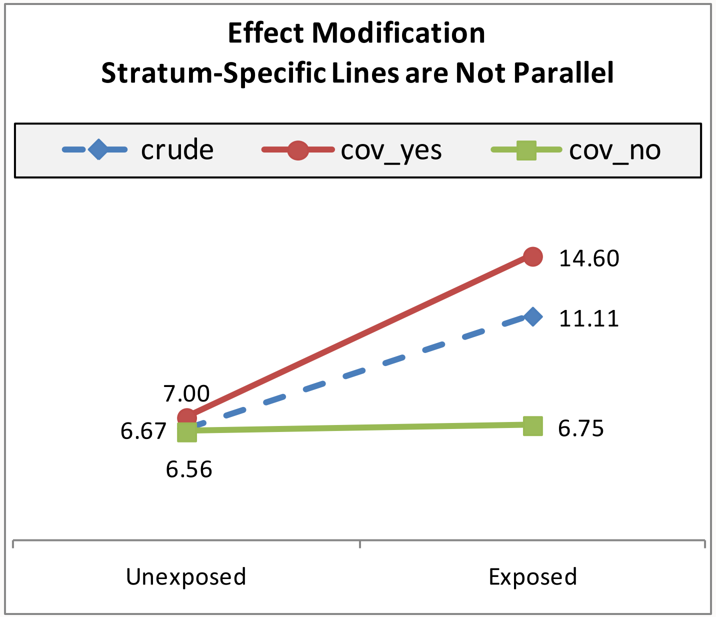




Confounding vs Effect Modification (Multiplicative)


No confounding or Effect Modification
Parallel Lines


Confounding: Stratum-Specific Lines are Parallel with Each Other, but Different than the Crude Line


Effect Modification
Stratum-Specific Lines are Not Parallel

Multiplicative Interaction

For statistically testing ***multiplicative*** interaction, the null hypothesis can be stated in two equivalent ways:

1. Equality (homogeneity) of the stratum-specific measures of association
2. Equality of the joint effect and the ***product*** of the set of separate effects

Multiplicative Model of Interaction

- The joint effect is the effect of exposure to both of the factors of interest compared to non-exposure to both
- The separate effects are the effect of exposure to each factor in the absence of the other compared to non-exposure to both
- The group that is unexposed to either factor is the "common referent group"

$$\mathrm{OR}_{\text{stratum 1}} = \mathrm{OR}_{\text{stratum 2}}$$

$$\mathrm{OR}_{\textbf{joint}} = \mathrm{OR}_{x_1+,x_2-} \times \mathrm{OR}_{x_1-,x_2+}$$

Multiplicative Model of Interaction

Rearranged Data Layout for Contingency Tables: Alternative Perspective on Effect Modification

- Each table displays the exposure – disease relationship for either a joint or separate effect
- There is a table for each of the (# strata-1) joint effects, plus a table for each of the # strata separate effects
- The joint effects are combinations of covariates and the exposure variable; the separate effects are combinations of covariates in the absence of exposure plus exposure alone

Multiplicative Model of Interaction

Rearranged Layout: Observed and Expected Joint Effects
The Simple Case of 3 Dichotomous Variables

Joint Effect of
Exposure and Covariate

	Disease Y	 N
E=Y and C=Y	P_1	
E=N and C=N	P_4	

Effect of the Covariate
in the Absence of Exposure

	Disease Y	 N
E=N and C=Y	P_2	
E=N and C=N	P_4	

Effect of Exposure
in the Absence of the Covariate

	Disease Y	 N
E=Y and C=N	P_3	
E=N and C=N	P_4	

Compare to tables on slide 41

Multiplicative Model of Interaction

Rearranged Layout: Observed and Expected Joint Effects

Joint Effect of
Exposure and Covariate

	Disease	
	Y	N
E=Y and C=Y	a	b
E=N and C=N	g	h

Effect of the Covariate
in the Absence of Exposure

	Disease	
	Y	N
E=N and C=Y	c	d
E=N and C=N	g	h

Effect of Exposure
in the Absence of the Covariate

	Disease	
	Y	N
E=Y and C=N	e	f
E=N and C=N	g	h

Case Control Data

Ex: Multiplicative Effect Modification

(from before, **slide 33**)

```
     Table 1 of exposure by outcome
     Controlling for covariate=yes
                                C
  exposure     outcome

  Frequency|
  Percent  |
  Row Pct  |
  Col Pct  | yes    |no      |  Total
  ---------+--------+--------+
   yes     |    365 |   2135 |   2500
           |   7.30 |  42.70 |  50.00
           |  14.60 |  85.40 |
           |  67.59 |  47.87 |
  ---------+--------+--------+
  no       |    175 |   2325 |   2500
           |   3.50 |  46.50 |  50.00
           |   7.00 |  93.00 |
           |  32.41 |  52.13 |
  ---------+--------+--------+
  Total         540     4460     5000
              10.80    89.20   100.00
```

```
     Table 2 of exposure by outcome
      Controlling for covariate=no
                                C
  exposure     outcome

  Frequency|
  Percent  |
  Row Pct  |
  Col Pct  | yes    |no      |  Total
  ---------+--------+--------+
   yes     |    135 |   1865 |   2000
           |   1.35 |  18.65 |  20.00
           |   6.75 |  93.25 |
           |  20.45 |  19.97 |
  ---------+--------+--------+
  no       |    525 |   7475 |   8000
           |   5.25 |  74.75 |  80.00
           |   6.56 |  93.44 |
           |  79.55 |  80.03 |
  ---------+--------+--------+
  Total         660     9340    10000
               6.60    93.40   100.00
```

RR=2.1 Breslow-Day p <0.0001 RR=1.0

Ex: Multiplicative Effect Modification

Rearranged Layout:
Joint and Separate Effects

```
The FREQ Procedure

Table of ExpCovC by outcome

ExpCovC           outcome

Frequency        ,
Percent          ,
Row Pct          ,
Col Pct          ,       0,       1,  Total
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
exp y, cov y     ,   2135 ,    365 ,   2500
                 ,  14.23 ,   2.43 ,  16.67
                 ,  85.40 ,  14.60 ,
                 ,  15.47 ,  30.42 ,
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
exp n, cov y     ,   2325 ,    175 ,   2500
                 ,  15.50 ,   1.17 ,  16.67
                 ,  93.00 ,   7.00 ,
                 ,  16.85 ,  14.58 ,
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
exp y, cov m     ,   1865 ,    135 ,   2000
                 ,  12.43 ,   0.90 ,  13.33
                 ,  93.25 ,   6.75 ,
                 ,  13.51 ,  11.25 ,
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Neither          ,   7475 ,    525 ,   8000
                 ,  49.83 ,   3.50 ,  53.33
                 ,  93.44 ,   6.56 ,
                 ,  54.17 ,  43.75 ,
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Total               13800     1200    15000
                    92.00     8.00   100.00


Summary Statistics for ExpCovC by outcome

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

Statistic    Alternative Hypothesis    DF       Value      Prob
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
    1        Nonzero Correlation        1    117.0014    <.0001
    2        Row Mean Scores Differ     3    178.0537    <.0001
    3        General Association        3    178.0537    <.0001
```

Ex: Multiplicative Effect Modification

Rearranged Layout: Joint Effect

```
Joint         outcome

Frequency     ,
Percent       ,
Row Pct       ,
Col Pct       , yes    ,no      ,  Total
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
exp  y, cov y ,    365 ,   2135 ,   2500
              ,   3.48 ,  20.33 ,  23.81
              ,  14.60 ,  85.40 ,
              ,  41.01 ,  22.22 ,
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
exp n, cov n  ,    525 ,   7475 ,   8000
              ,   5.00 ,  71.19 ,  76.19
              ,   6.56 ,  93.44 ,
              ,  58.99 ,  77.78 ,
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
Total                890     9610    10500
                    8.48    91.52   100.00
```

```
Type of Study                   Value       95% Confidence Limits
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
Case-Control (Odds Ratio)      2.4341        2.1120        2.8055
Cohort (Col1 Risk)             2.2248        1.9618        2.5230
```

Ex: Multiplicative Effect Modification

Rearranged Layout: Separate Effects

```
 sep_exp        outcome

 Frequency     ,
 Percent       ,
 Row Pct       ,
 Col Pct       , yes    ,no       ,  Total
 ƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 exp  y, cov n ,    135 ,   1865 ,   2000
               ,   1.35 ,  18.65 ,  20.00
               ,   6.75 ,  93.25 ,
               ,  20.45 ,  19.97 ,
 ƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 exp n, cov n  ,    525 ,   7475 ,   8000
               ,   5.25 ,  74.75 ,  80.00
               ,   6.56 ,  93.44 ,
               ,  79.55 ,  80.03 ,
 ƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 Total               660     9340    10000
                    6.60    93.40   100.00

Type of Study                Value          95% CI
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
Case-Control (Odds Ratio)   1.0306      0.8473   1.2536
Cohort (Col1 Risk)          1.0286      0.8568   1.2347
```

```
 sep_cov        outcome

 Frequency     ,
 Percent       ,
 Row Pct       ,
 Col Pct       , yes    ,no       ,  Total
 ƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 exp n, cov  y ,    175 ,   2325 ,   2500
               ,   1.67 ,  22.14 ,  23.81
               ,   7.00 ,  93.00 ,
               ,  25.00 ,  23.72 ,
 ƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 exp n, cov n  ,    525 ,   7475 ,   8000
               ,   5.00 ,  71.19 ,  76.19
               ,   6.56 ,  93.44 ,
               ,  75.00 ,  76.28 ,
 ƒƒƒƒƒƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆƒƒƒƒƒƒƒƒˆ
 Total               700     9800    10500
                    6.67    93.33   100.00

Type of Study                Value          95% CI
ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
Case-Control (Odds Ratio)   1.0717      0.8976   1.2796
Cohort (Col1 Risk)          1.0667      0.9043   1.2581
```

Multiplicative Model of Interaction

The null hypothesis of no multiplicative interaction

Joint and Separate Effects

$$\frac{p_1}{p_4} = \frac{p_2}{p_4} \times \frac{p_3}{p_4} \qquad \frac{ah}{bg} = \frac{ch}{dg} \times \frac{eh}{fg}$$

$$p_1 = \frac{p_2 p_3}{p_4} \qquad \frac{ad}{bg} = \frac{ceh}{fg^2}$$

$$\frac{p_1}{p_2} = \frac{p_3}{p_4} \qquad \frac{ad}{bc} = \frac{eh}{fg}$$

Stratum-specific Effects

$$RR_1 = RR_2 \qquad OR_1 = OR_2$$

Multiple Factor Stratified Analysis

Examine potential joint confounding—whether crude and adjusted estimates differ and whether there is interaction across strata.

Several scenarios when controlling simultaneously for several factors:

1. **No confounding or interaction:** the crude and adjusted estimates of effect are the same, regardless of whether factors are adjusted for singly or jointly

Multiple Factor Stratified Analysis

2. **Confounding present with one sufficient confounder:** crude and adjusted estimates of effect differ, but the magnitude of confounding does not change—the multiple-factor adjusted estimate is the same as that obtained with control for one of the factors by itself—control of multiple factors simultaneously does not yield a "better" estimate than control of a single factor

Multiple Factor Stratified Analysis

3. **Confounding present only when two or more covariates are considered or magnitude of confounding is different when more than one confounder is considered:** crude and adjusted estimates of effect differ, and the magnitude of confounding changes—the multiple-factor adjusted estimate is different from that obtained with control for one of the factors by itself—control of multiple factors simultaneously yields a "better" estimate than control by any single factor by itself

Multiple Factor Stratified Analysis

How do we know which
combinations of variables to consider?

“A sufficient confounder group is a minimal set of one or more risk factors whose simultaneous control in the analysis will correct for **joint confounding** in the estimation of the effect of interest. Here, 'minimal' refers to the property that, for any such set of variables, no variable can be removed from the set without sacrificing validity.”

Kleinbaum, DG, Kupper, LL., Morgenstern,H. *Epidemiologic Research: Principles and Quantitative Methods,* Nostrand Reinhold Company, New York, 1982, p 276.

Multiple Factor Stratified Analysis

What about interaction?

What if there was interaction between the exposure and one of the other factors alone, but not after looking at two factors jointly?

What if there was no interaction when looking at each factor alone, but there is after considering both factors jointly?

Epidemiologic analysis is an art as well as a science!

Stratified Analysis—Summary

At each step of the stratification process, look at numbers and percents (measures of occurrence) in the cells. Depending on the study design, look at row and/or column totals and percents, or person time totals.

When you are at the **bivariate** stage, look at:

- Appropriate test statistic (chi-square test)
- Crude measures of association (relative risk, prevalence ratio, odds ratio, risk difference/attributable risk)
- 95% confidence interval for the measure of association

Stratified Analysis—Summary

When you are doing **single or multiple factor stratified analysis**, also look at:

- Outcome risk/prevalence in each cell of the stratified tables
- Sample sizes in cells
- Stratum-specific measures of association and 95% CIs
- Adjusted measures of association and 95% CIs
- Test statistic and p-value for homogeneity of the measure of association across strata

In addition, if the exposure variable is ordinal rather than dichotomous—a k x 2 table, look at:

- Test for trend at the bivariate, single, and multiple factor stratified levels

Stratified Analysis—Summary

Some issues to think about when doing stratified analysis:

Categories:

- Are my categories adequate—for the exposure, the outcome, and the covariates (stratification variables)?
- Are there too many or too few categories?
- Could there be misclassification and how might it be affecting my results?

Stratified Analysis—Summary

Some issues to think about when doing stratified analysis:

Sample size:

- Are there any sparse cells?
- How are the sample sizes in each stratum affecting results?
- Are some stratum-specific point estimates (OR/RR) equivalent, but only significant in some strata?
- Could categories be changed to improve stratum-specific sample size without imposing misclassification?
- What is the impact of missing values on sample size and representativeness of sample?

Stratified Analysis—Summary

Some issues to think about when doing stratified analysis:

Assessing Effect Modification:

- Do I agree with the Breslow-Day results?
- Are there patterns I think are important regardless of the statistical results?
- What might be reasons for seemingly different stratum-specific results? For example, why might an exposure be protective in some strata, but confer excess risk in other strata?

Overall:

- How will I use the stratified results as I move into regression modeling?

Analytic Framework - Summary

All of the following steps should be taken prior to any multivariable modeling:

1. Establish research question
2. Articulate a conceptual framework
3. Select and define variables of interest/available to address question
4. Define categories, sometimes more than once, for a given variable
5. Examine univariate distributions
6. Examine bivariate distributions

Analytic Framework - Summary

BEFORE any multivariable modeling (cont’d):

7. Perform single factor stratified analysis for the primary association(s) of interest, with each potential confounder / effect modifier

6. Rethink variables and categories

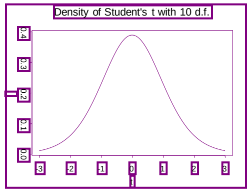
7. Perform multiple factor stratified analysis for the primary association of interest with different combinations of potential confounders / effect modifiers

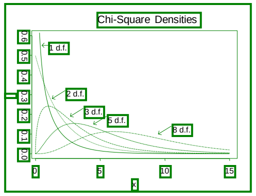
Overview of Multivariable Models

Webinar, Friday, May 16, 2014

Training Course in MCH Epidemiology

Deb Rosenberg, PhD
Research Associate Professor
Division of Epidemiology and Biostatistics
University of IL School of Public Health



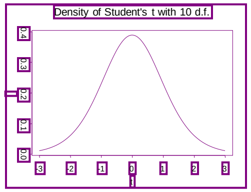


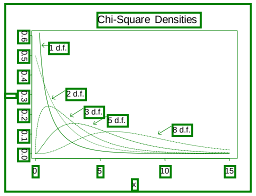
Linear Models: General Considerations

		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

- Multivariable analysis implies acknowledging and accounting for the intricacies of the real world as reflected in the relationships among a set of variables

- The accuracy of statistics from multivariable analysis and therefore the accuracy of conclusions drawn and the appropriateness of any subsequent public health action taken is dependent on using appropriate methods.





Linear Models: General Considerations

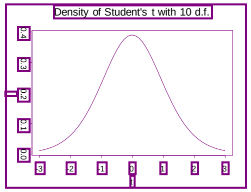
		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

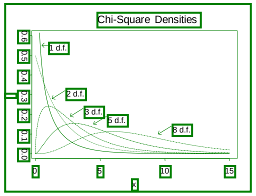
Why Use Regression Modeling Approaches?

Why not just do stratified analysis?

Unlike stratified analysis, regression approaches:

1. more efficiently handle many variables and the sparse data that stratification by many factors may imply
2. can accommodate both continuous and discrete variables, *both as outcomes and as independent variables*.





Linear Models: General Considerations

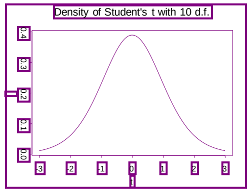
		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

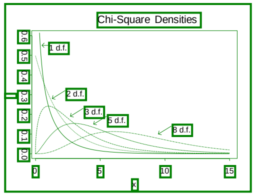
Unlike stratified analysis, regression approaches:

3. allow for examination of multiple factors (independent variables) ***simultaneously*** in relation to an outcome (dependent variable)

4. allow variables to take on different roles depending on the focus of analysis—a variables might be considered an "exposure" or it might be considered a "covariate" or both

5. provide more flexibility in assessing effect modification and controlling confounding.





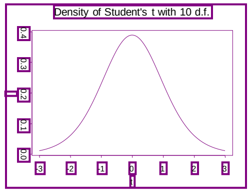
Linear Models: General Considerations

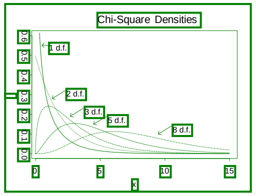
		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

Regression Modeling is Used for Multiple Purposes

Sometimes, regression modeling is carried out in order to assess **one association;** other variables are included to adjust for confounding or account for effect modification. In this scenario, the focus is on obtaining the ‘best’ estimate of the single association.

Sometimes, regression modeling is carried out in order to assess **multiple, competing exposures**, or to identify a **set of variables** that together predict the outcome.



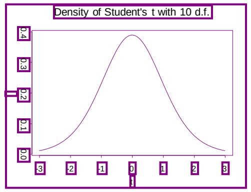


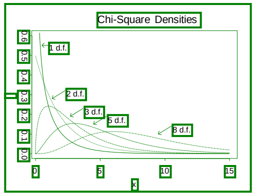
Linear Models: General Considerations

		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

Regression analysis is an alternative to and an extension of simpler methods used to estimate incidence and prevalence and to test hypotheses about associations:

- Regression analysis yields estimates of **means, proportions, or rates,**
- Regression analysis tests differences between **means,** and is an extension of t-tests and analysis of variance.
- Regression analysis tests differences between or ratios of **proportions or rates,** and is an extension of chi-square tests from contingency tables – crude & stratified analysis.





Linear Models: General Considerations

		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

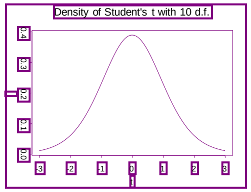
Confidence Intervals

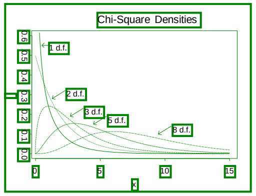
$$\boxed{\text{CI} = \text{Obs. Meas. of Occ.} \pm \text{Critical Value} \times \text{s.e. of the Meas. of Occ.}}$$

$$\boxed{\text{CI} = \text{Predicted Value} \pm \text{Critical Value} \times \text{s.e. of Predicted Value}}$$

$$\boxed{\text{CI} = \text{Obs. Assoc.} \pm (\text{Critical Value} \times \text{s.e. of the Assoc.})}$$

$$\boxed{\text{CI} = \text{Beta Coef.} \pm \text{Critical Value} \times \text{s.e. of Beta Coef.}}$$





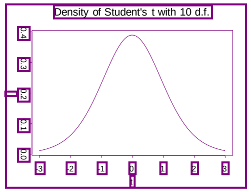
Linear Models: General Considerations

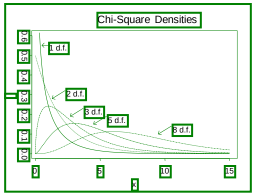
		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

Hypothesis Testing

$$\boxed{\text{Test Statistic} = \frac{\text{Observed Association - Expected Association}}{\text{Standard Error of the Association}}}$$

$$\boxed{\text{Test Statistic} = \frac{\text{Obs. Beta Coeff.} - \text{Exp. Beta Coef.}}{\text{Standard Error of the Beta Coef.}}}$$



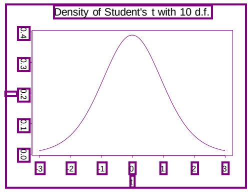


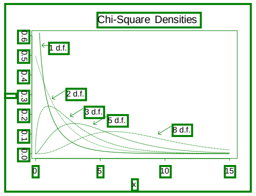
Assessing Effect Modification

- Stratified Analysis: Are the stratum-specific measures of association different (heterogeneous)?
- Regression Analysis: Is the beta coefficient resulting from the multiplication of two variables large?

Regardless of the method, if the stratum-specific estimates differ, then reporting a weighted average will mask the important stratum-specific differences.

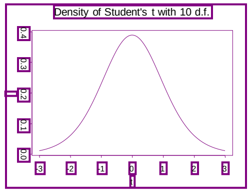
Stratum-specific differences can be statistically tested.

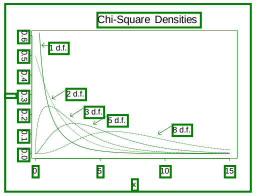




Assessing Confounding

- Standardization: Does the standardized measure differ from the unstandardized measure?
- Stratified Analysis: Does the adjusted measure of association differ from the crude measure of association?
- Regression Analysis: Does the beta coefficient for a variable in a model that includes a potential confounder differ from the beta coefficient for that same variable in a model that does not include the potential confounder?





Linear Models: General Considerations

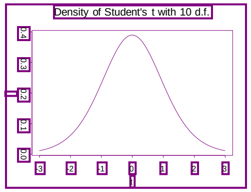
Disease or Other Health Outcome

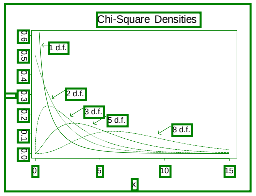
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n$_1$)
	No	c	d	c + d (n$_2$)
		a + c (m$_1$)	b + d (m$_2$)	a + b + c + d N

Multivariable modeling should be the culmination of an analytic strategy that includes articulating a conceptual framework and carrying out preliminary analysis.

BEFORE any multivariable modeling—

- Select variables of interest
- Define levels of measurement, sometimes more than once, for a given variable
- Examine univariate distributions
- Examine bivariate distributions





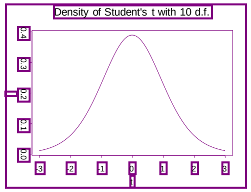
Linear Models: General Considerations

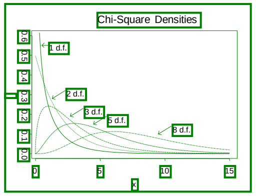
		Disease or Other Health Outcome Yes	 No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

BEFORE any multivariable modeling—

- Perform single factor stratified analysis to assess confounding and effect modification
- Rethink variables and levels of measurement
- Perform multiple factor stratified analysis with different combinations of potential confounders / effect modifiers

These steps should never be skipped!





Linear Models: General Considerations

		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

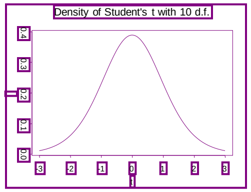
The most common regression models used to analyze health data express the hypothesized association between risk or other factors and an outcome as a linear (straight line) relationship:

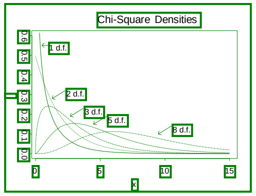
$$\mathbf{Outcome}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

Dependent Var. = ------Independent Variables------

This equation is relevant to any ***linear*** model; **what differentiates one modeling approach from another is**

- ***the structure of the outcome variable, and***
- ***the corresponding structure of the errors.***





Linear Models: General Considerations

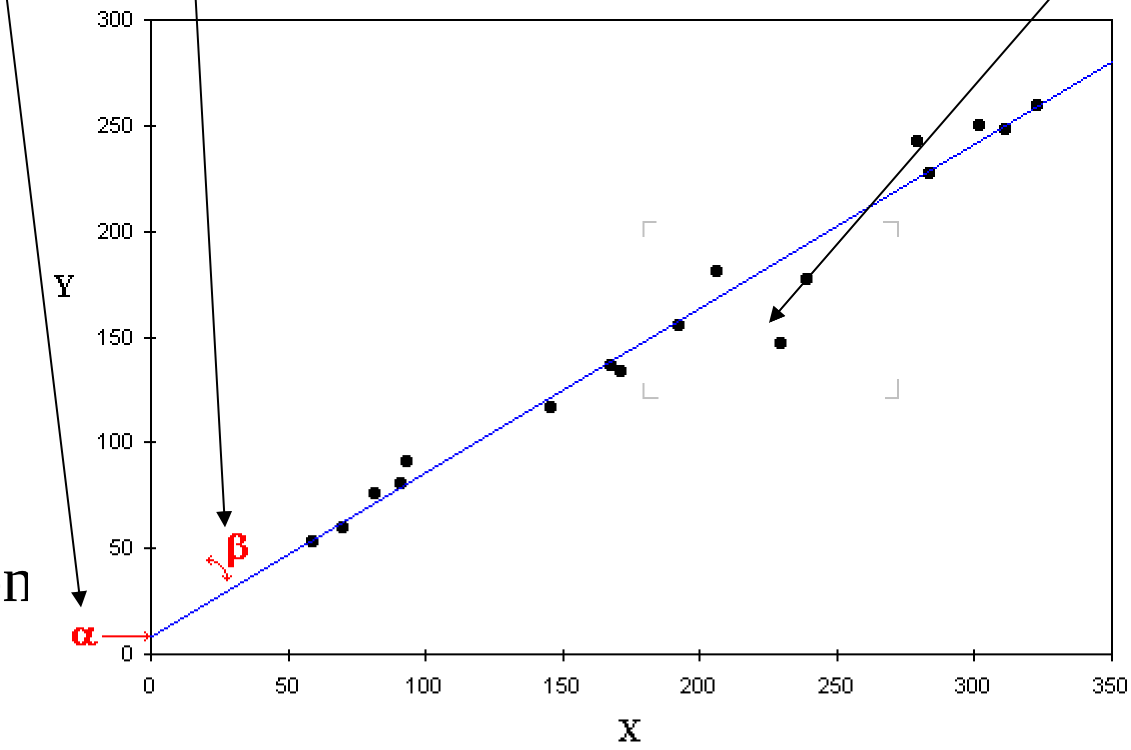
Disease or Other Health Outcome

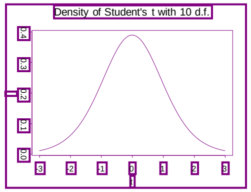
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

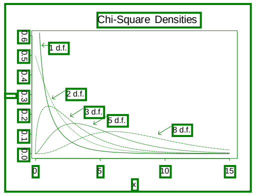
$$\textbf{Outcome}_i = \boxed{\beta_0} + \boxed{\beta_1 X_{i1}} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \boxed{\varepsilon_i}$$

The straight line relationship includes an intercept and one or more slope parameters.

The differences between the actual data points and the regression line are the errors.







Linear Models: General Considerations

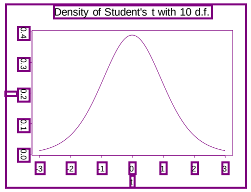
		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

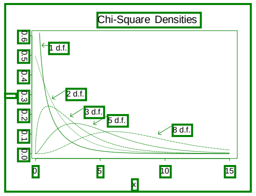
The Traditional, 'Normal' Regression Model

$$\boxed{Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i}$$

This model has the following properties:

- The outcome "Y" is continuous & normally distributed.
- The Y values are independent.
- The errors are independent, normally distributed; their sum equals 0, with constant variance across levels of X.
- The expected value (mean) of the Y's is linearly related to X (a straight line relationship exists).





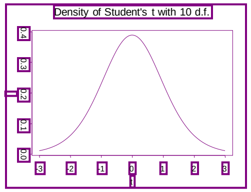
Linear Models: General Considerations

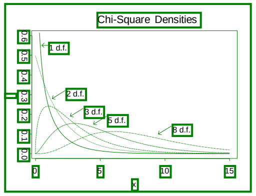
		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

When the outcome variable is **not** continuous and normally distributed, a linear model cannot be written in the same way, and the properties listed above no longer pertain.

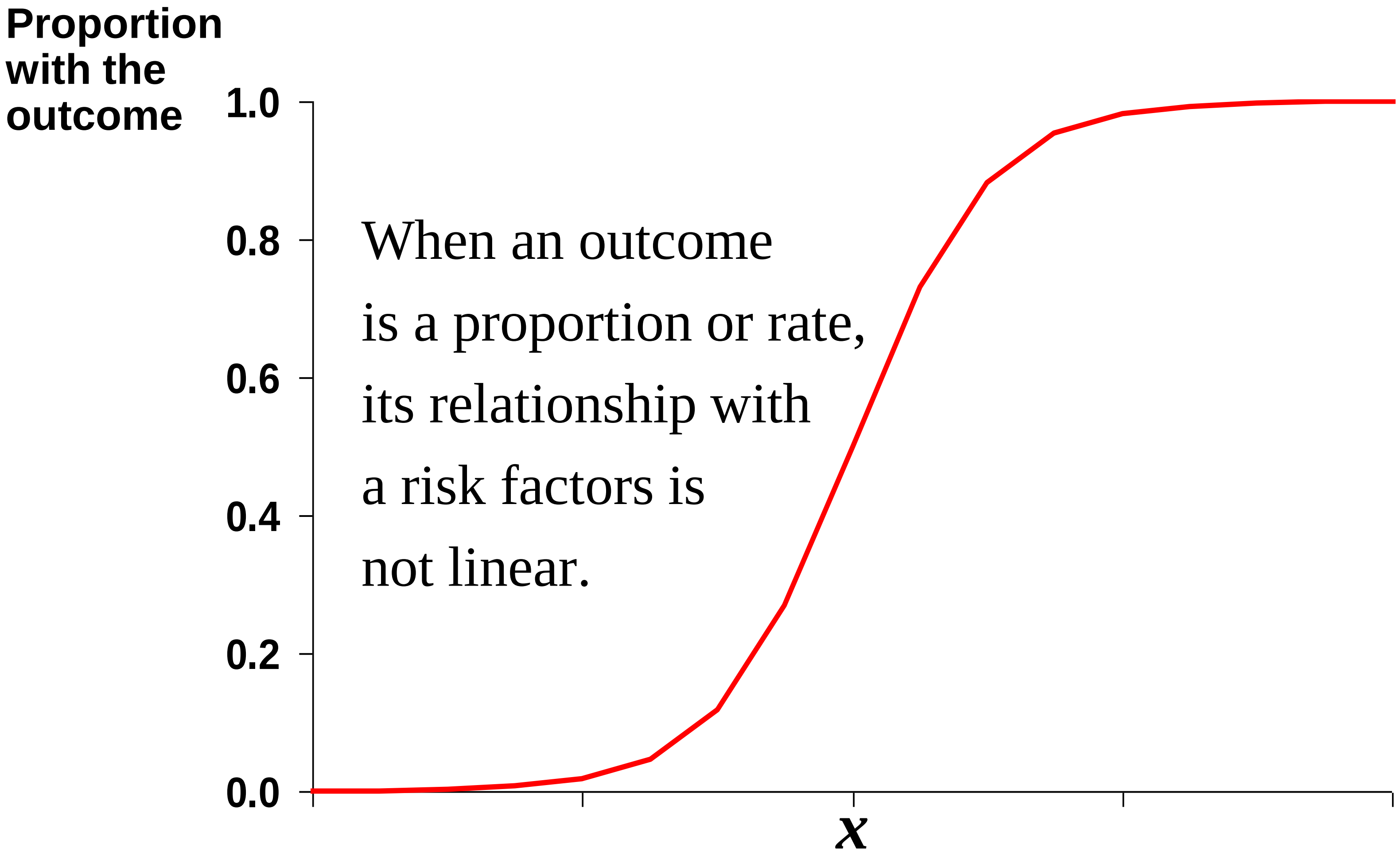
For example, if the outcome variable is a proportion or rate:

- The errors are **not** normally distributed
- The variance across levels of X is **not** constant. (By definition, **p(1-p)** changes with **p** and **r** changes with **r**).
- The expected value (proportion or rate) is **not** linearly related to X (**a straight line relationship does not exist**).

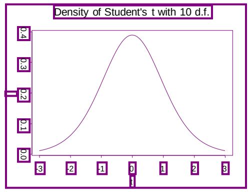


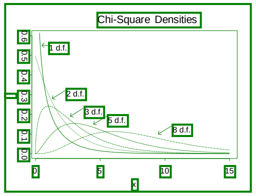


Linear Models: General Considerations



When an outcome is a proportion or rate, its relationship with a risk factors is not linear.





Linear Models: General Considerations

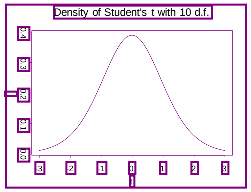
		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

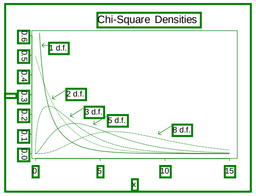
General Linear Models

How can a linear modeling approach be applied to the many health outcomes that are proportions or rates?

The normal, binomial, Poisson, exponential, chi-square, and multinomial distributions are all in the **exponential family.**

Therefore, it is possible to define a **“link function”** that **transforms** an outcome variable from any of these distributions so that it is linearly related to a set of independent variables; the error terms can also be defined to correspond to the form of the outcome variable.





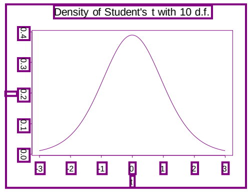
		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

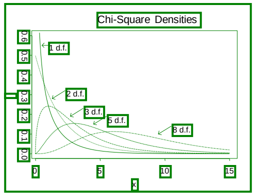
General Linear Models

Some common link functions:

- identity (untransformed)
- natural log
- logit
- cumulative logit
- generalized logit

The interpretation of the parameter estimates—the beta coefficients—changes depending on whether and how the outcome variable has been transformed (which link function has been used).





Linear Models: General Considerations

		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

Linear equation →

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 X$$

$$\frac{p}{1-p} = e^{b_0 + b_1 X}$$

$$p = e^{b_0 + b_1 X}(1-p)$$

$$p = e^{b_0 + b_1 X} - pe^{b_0 + b_1 X}$$

$$p + pe^{b_0 + b_1 X} = e^{b_0 + b_1 X}$$

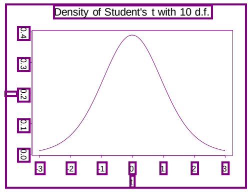
$$p\left(1 + e^{b_0 + b_1 X}\right) = e^{b_0 + b_1 X}$$

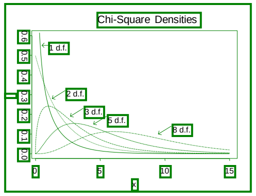
$$p = \frac{e^{b_0 + b_1 X}}{1 + e^{b_0 + b_1 X}}$$

The logit link function:

(logistic regression)

Non-linear equation →





Linear Models: General Considerations

		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

The natural log link function:

log-binomial or Poisson regression with count data

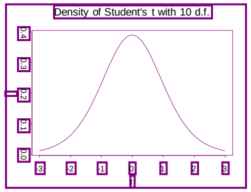
Non-linear model →

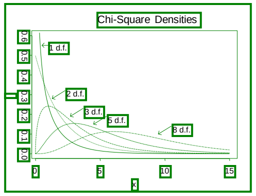
$$\text{count} = ne^{b_0+b_1X}$$

$$\frac{\text{count}}{n} = r = e^{b_0+b_1X}$$

The linear model →

$$\ln(r) = b_0 + b_1X$$





Linear Models: General Considerations

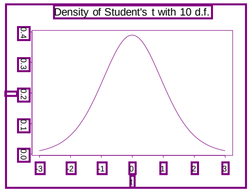
		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

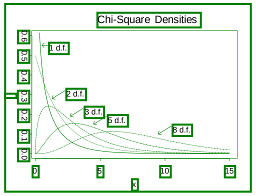
'Normal' Regression—Link=Identity, Dist=Normal

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

Binomial Regression—Link=Identity, Dist=Binomial

$$\pi = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i$$





Linear Models: General Considerations

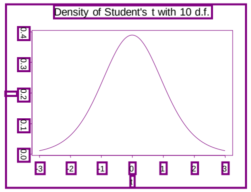
		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

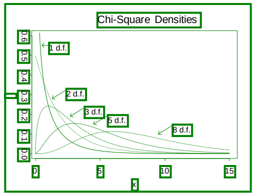
Logistic Regression—Link=Logit, Dist=Binomial

$$\ln\left(\frac{\pi}{1-\pi}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k+\varepsilon$$

Log-Binomial or Poisson Regression with Count Data—Link=Log, Dist=Binomial or Dist=Poisson

$$\ln(\pi)\text{ or }\ln(\lambda)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k+\varepsilon$$





Ordinal and Nominal Model

For an ordinal outcome with k+1 categories

$$\ln \text{Odds}_1 = \ln \frac{p_1}{1-p_1}$$

$$\ln \text{Odds}_{1+2} = \ln \frac{p_{1+2}}{1-p_{1+2}}$$

$$\ln \text{Odds}_{1+2+\ldots+k} = \ln \frac{p_{1+2+\ldots+k-1}}{1-p_{1+2+\ldots+k-1}}$$

Both the numerator and denominator change

For a nominal outcome with k+1 categories

$$\ln \text{Odds}_1 = \ln \frac{p_1}{1-p_{1+2+\ldots+k-1}}$$

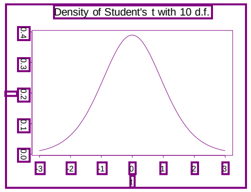
$$\ln \text{Odds}_2 = \ln \frac{p_2}{1-p_{1+2+\ldots+k-1}}$$

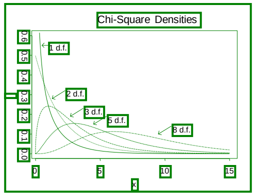
$$\ldots$$

$$\ln \text{Odds}_k = \ln \frac{p_k}{1-p_{1+2+\ldots+k-1}}$$

Fixed denominator (reference) category

http://www.indiana.edu/%7Estatmath/stat/all/cat/2b1.html



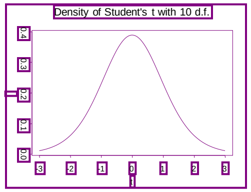


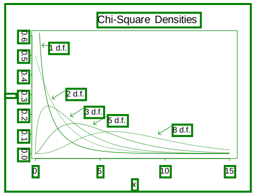
Some Models with Correlated Errors

Mixed Models

$$\boxed{\hat{Y} = \underset{\text{fixed effects}}{\left(b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k\right)} + \underset{\text{random effects}}{\left(c_0 + c_1Z_1 + c_2Z_2 + \ldots + c_kZ_k\right)}}$$

- Multilevel/clustered data
- Repeated measures/longitudinal data
- Matched data
- Time series analysis
- Spatial analysis





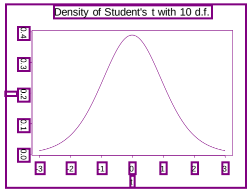
Regression Modeling Results

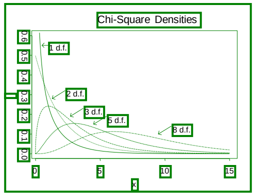
Measures of Occurrence

Predicted Values: Crude, Adjusted, or Stratum-Specific

The predicted values are ***points*** on the regression line given particular values of the set of independent variables

- **‘Normal’ model yields means**
- **Binomial models yields proportions**
- **Logistic model yields ln(odds)**
- **Binomial / Poisson models yield ln(proportions / rates)**





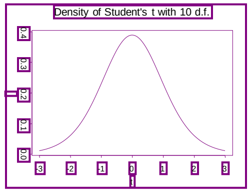
		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

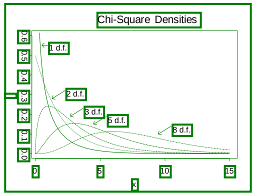
Measures of Association

Beta coefficients: Crude, Adjusted, or Stratum-Specific

The measures of association are ***comparisons of points*** on the regression line at differing values of the independent variables

- **‘Normal’ model yields differences between means**
- **Binomial model yields differences in proportions**
- **Logistic model yields differences in ln(odds)**
- **Log Binomial / Poisson models yield differences in ln(proportions / rates)**





Linear Models: General Considerations

		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

Regression Modeling Approaches

Measures of Association

‘Normal” regression

Differences between means

Binomial Regression

Differences between proportions:

Risk Differences / Attributable Risks

Logistic regression

(binary, cumulative, generalized)

Differences between log odds:

Odds Ratio(s) for—

- a single binary outcome
- a ***set*** of binary outcomes
- an ordinal outcome

Log-Binomial or Poisson regression

Differences between log proportions:

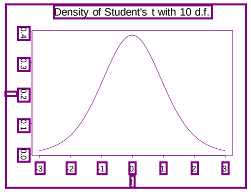
Relative Risk / Relative Prevalence

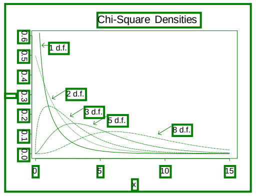
Poisson regression

(person-time data)

Differences between log rates:

Rate Ratio



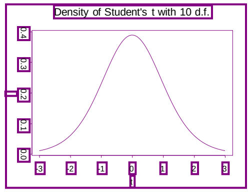


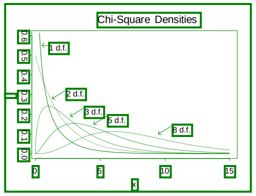
Linear Model: General Considerations

		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

- Multivariable models may be quite complex, including both continuous and discrete measures, and measures at the individual level and/or at an aggregate level such as census tract, zip code, or county.

- Interpretation of the slopes or "beta coefficients" can be equally complex as these reflect the measures of occurrence (means, proportions, rates) or measures of association (odds ratios, relative risks, rate ratios, risk differences) when used singly or in combination.





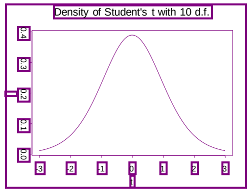
Linear Models: General Considerations

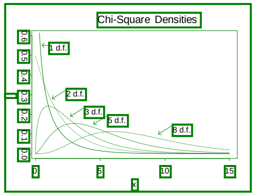
		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

The challenge for an MCH epidemiologist goes beyond carrying out complex multivariable analysis to include:

advocating for and facilitating the more ***routine*** incorporation of complex multivariable methods into the work of public health agencies, and

- guiding interpretation of findings
- working to design reporting templates
- working to build dissemination strategies
- working to link findings with action plans or policy recommendations





Linear Models: General Considerations

		Disease or Other Health Outcome		
		Yes	No	
Exposure or Person, Place, or Time Variable	Yes	a	b	a + b (n_1)
	No	c	d	c + d (n_2)
		a + c (m_1)	b + d (m_2)	a + b + c + d N

As we move into sophisticated statistical methods, it's important to keep our perspective:

"...technical expertise and methodology are not substitutes for conceptual coherence. Or, as one student remarked a few years ago, public health spends too much time on the "p" values of biostatistics and not enough time on "values."

Jonathan M. Mann in **Medicine and Public Health, Ethics and Human Rights**
The Hastings Center Report , Vol. 27, No. 3 (May - Jun., 1997), pp. 6-13
Published by: The Hastings Center