Lecture 13

Design and Analysis Techniques for Epidemiologic Studies

10/27/2017
Learning Objectives

1. Define study designs
2. Measures of effects for categorical data
3. Confounders and effects modifications
4. Stratified analysis (Mantel Haenszel statistic, multiple logistic regression)
5. Use of Computer Program for Logistic regression (in Lab)
Study Design

To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination:

He may be able to say what the experiment died of.

- Sir Ronald A. Fisher
Study Design

- Designing a study is possibly the most important role for the statistical expert in a research team.

- Design of the study (one need to have a good knowledge about the exposures, disease of interest and study objectives and hypotheses)
  - Sampling design (selection of subjects)
  - Sample size calculation (new study) or power calculation (if study is from existing data)
  - Analysis plan
Study Design

Most clinical studies can be broadly classified into one of two categories, namely

– Experimental Studies (Clinical Trials): Experimental units are randomly assigned to a specific level of the exposure (intervention).

– Observational Studies: Data are collected in a given situation, without intentional interference (randomization) by the observer.
Experimental Study

- Gold-Standard for the proof of an effect of a treatment (required for registration-FDA)

- Four Phases in Drug research
  - Phase I (How the body copes with the drug, the safe dose range, the side effects, some therapeutic effect, few subjects)
  - Phase II (If the new treatment works well enough to test in phase 3, More about side effects and how to manage them, More about the most effective dose to use, more subjects)
  - Phase III (The new treatment or procedure is compared with the standard treatment, Different doses or ways of giving a standard treatment, sample size is large)
  - Phase IV (More about the side effects and safety of the drug, what the long term risks and benefits are, how well the drug works when it’s used more widely than in clinical trials)
Experimental Study

- Randomization protects against bias in assignment to groups.
- Blinding protects against bias in outcome assessment or measurement.
- Control for (major) sources of variability, although not necessarily reflecting real life conditions.
- Expensive in terms of time and money.
Experimental Study-Benefit of additional Stent in MI-Therapy

Patients with acute myocardial infarction → Randomization

- PTCA & Stent
  - healthy
  - reinfarction

- PTCA
  - healthy
  - reinfarction
Some definitions for the example

- **Percutaneous** means access to the blood vessel is made through the skin.
- **Transluminal** means the procedure is performed within the blood vessel.
- **Coronary** specifies that the coronary artery is being treated.
- **Angioplasty** means "to reshape" the blood vessel (with balloon inflation).
- **A stent** is a small, metal coil that helps to keep a “ballooned” artery open.
Observational Study

– Survey to characterize a target population with respect to specific parameters

– May include all population members (census)

– Typically includes only a part of the population (sample) because of time, cost and other practical constraints
Observational Study-Risk for MI and High Serum Cholesterol

Population at Risk

Cases
Myocardial Infarction

Exposed
Serum Cholesterol >200mg/dl

Unexposed
Serum Cholesterol <200mg/dl

Exposed
Serum Cholesterol >200mg/dl

Unexposed
Serum Cholesterol <200mg/dl

Controls
Blood Donors
Observational Study most likely used in Epidemiology

Types of study

- Cross-sectional study
- Case-control study (retrospective)
- Cohort study (Prospective)
Cross-Sectional Studies

→ Begin with “Cross-sectional” sample

→ Determine Exposure and Disease at same time
Cross-Sectional Studies

Exposure (Risk Factor)

<table>
<thead>
<tr>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>Random</td>
</tr>
<tr>
<td>Random</td>
<td>Random</td>
</tr>
</tbody>
</table>

Disease (Outcome)

<table>
<thead>
<tr>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>Random</td>
</tr>
<tr>
<td>Random</td>
<td>Random</td>
</tr>
</tbody>
</table>
Case-Control Studies

→ Begin with sample of “Cases and Controls”

→ Start with Disease status, then assess and compare Exposures in cases vs. controls.
Case-control Studies

Exposure (Risk Factor)

Disease (Outcome)

Fixed

Random

Random

Random

Random

Random
Cohort Studies

→ Begin with sample → “Healthy Cohort” (i.e., subjects without the outcome yet)

→ Start with Exposure status, then compare subsequent disease experience in exposed vs. unexposed.
Cohort Studies

Exposure (Risk Factor)

Disease (Outcome)

<table>
<thead>
<tr>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>random</td>
<td>random</td>
</tr>
<tr>
<td>random</td>
<td>random</td>
</tr>
</tbody>
</table>

fixed fixed
Measures of effects for categorical data

- Depends on study design

- Prospective study: Incidence of disease (risk difference, relative risk, odds ratio of disease)

- Cross-sectional: Prevalence of disease (risk difference, relative risk, odds ratio of disease)

- Case-cohort: study of exposure (odds ratio of exposure)
2x2 tables notations

Exposure (Risk Factor)

<table>
<thead>
<tr>
<th></th>
<th>E</th>
<th>~E</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>a</td>
<td>c</td>
</tr>
<tr>
<td>~D</td>
<td>b</td>
<td>d</td>
</tr>
</tbody>
</table>

Disease (Outcome)

N = n₁ + n₂ = m₁ + m₂

a + b = n₁

a + c = m₁

b + d = n₂

c + d = m₂

N = m₁ + m₂ = n₁ + n₂
Risk difference

Only for cross-sectional and cohort studies
Measured the attributable risk due to exposure

\[ RD = P(D \mid E) - P(D \mid \bar{E}) \]

\[ \hat{p}_1 = \frac{a}{n_1} \quad \hat{p}_2 = \frac{c}{n_2} \]

\[ \hat{RD} = \hat{p}_2 - \hat{p}_1 \]

\[ se(\hat{RD}) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{ab}{n_1^3} + \frac{cd}{n_2^3}} \]
Risk Difference

- A Confidence interval for the true risk difference can be easily constructed

- If the confidence interval contains 0, there is no evidence from data to suggest that the probability of the disease differs for the exposed and the unexposed groups
Relative Risk

Only for cross-sectional and cohort studies: Ratio of the probability that the outcome characteristic is present for one group, relative to the other.

\[ RR = \frac{P(D|E)}{P(D|\overline{E})} \]

The range of RR is \([0, \infty)\). By taking the logarithm, we have \((-\infty, +\infty)\) as the range for \(\ln(\hat{RR})\) and a better approximation to normality for the estimated \(\ln(\hat{RR})\):

\[
\ln(\hat{RR}) = \ln\left( \frac{\hat{P}(D|E)}{\hat{P}(D|\overline{E})} \right) \\
= \ln\left( \frac{a/n_1}{c/n_2} \right)
\]

\[
\ln(\hat{RR}) \sim N\left( \ln\left( \frac{p_1}{p_2} \right), \frac{1-p_1}{p_1n_1} + \frac{1-p_2}{p_2n_2} \right)
\]
Relative Risk

<table>
<thead>
<tr>
<th></th>
<th>Cold - Y</th>
<th>Cold - N</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vitamin C</td>
<td>17</td>
<td>122</td>
<td>139</td>
</tr>
<tr>
<td>Placebo</td>
<td>31</td>
<td>109</td>
<td>140</td>
</tr>
<tr>
<td>Total</td>
<td>48</td>
<td>231</td>
<td>279</td>
</tr>
</tbody>
</table>

The estimated relative risk is:

\[
\hat{RR} = \frac{\hat{P}(D | E)}{\hat{P}(D | \overline{E})} = \frac{17/139}{31/140} = 0.55
\]

We can obtain a confidence interval for the relative risk by first obtaining a confidence interval for the log-RR: and exponentiating the endpoints of the CI.

\[
\ln(\hat{RR}) \pm Z\frac{\alpha}{2} \times \sqrt{\frac{1-\hat{p}_1}{\hat{p}_1 n_1} + \frac{1-\hat{p}_2}{\hat{p}_2 n_2}}
\]
Odds Ratio

- Odds of an event is the probability that disease occurs divided by the probability it does not occur.
- Can be computed for all study designs
- In cohort studies, we have the odds ratio for disease (fixed # of exposed and non exposed)
- In case-control studies, we have the odds ratio for exposure (fixed # of cases and controls)
- In cross-sectional, we have both the odds ratio for exposure and disease (random margins)
Odds Ratio - Disease

- Odds ratio is the odds of the event for exposed divided by the odds of the event for unexposed.

- Sample odds of the outcome for each group:

\[
\text{odds}_E = \frac{a}{b} \quad \text{and} \quad \text{odds}_E = \frac{c}{d}
\]

\[
\text{OR}(\text{disease}) = \frac{\frac{P(D|E)}{1-P(D|E)}}{\frac{P(D|\overline{E})}{1-P(D|\overline{E})}} = \frac{\text{odds}_E}{\text{odds}_E} = \frac{ad}{bc}
\]
Odds Ratio-Exposure

we fixed the number of cases and controls then ascertained exposure status. The relative risk is therefore not estimable from these data alone. Instead of the relative risk we can estimate the exposure OR which Cornfield (1951) showed equivalent to the disease OR:

\[
\frac{P(E|D)/(1-P(E|D))}{P(E|D)/(1-P(E|D))} = \frac{P(D|E)/(1-P(D|E))}{P(D|E)/(1-P(D|E))}
\]

In other words, the odds ratio can be estimated regardless of the sampling scheme.

\[
OR(\text{disease}) = OR(\text{exposure}) = \frac{ad}{bc}
\]
Odds Ratio-Relative risk

For rare diseases, the disease odds ratio approximates the relative risk:

\[
\frac{P(D|E)}{P(D|\bar{E})} \approx \frac{P(D|E)}{P(D|\bar{E})}
\]

Since with case-control data we are able to effectively estimate the exposure odds ratio we are then able to equivalently estimate the disease odds ratio which for rare diseases approximates the relative risk.
Odds Ratio-Relative risk

![Graph showing the relationship between Disease prevalence and Odds Ratio vs. Relative Risk.](image)

- **Odds Ratio**
- **Relative Risk**

Disease prevalence

- 0
- 0.1
- 0.2
- 0.3
- 0.4

- 2
- 4
- 6
Odds Ratio

The odds ratio has \([0, \infty)\) as its range. The log odds ratio has \((-\infty, +\infty)\) as its range and the normal approximation is better as an approximation to the estimated log odds ratio.

\[
\ln(\hat{OR}) \sim N\left(\ln(\text{OR}), \frac{1}{n_1 p_1} + \frac{1}{n_1 q_1} + \frac{1}{n_2 p_2} + \frac{1}{n_2 q_2}\right)
\]

Confidence intervals are based upon:

\[
\ln\left(\frac{ad}{bc}\right) \pm Z_{1-\alpha/2} \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}
\]

Therefore, a \((1 - \alpha)\) confidence interval for the odds ratio is given by exponentiating the lower and upper bounds.
Example - NSAIDs and PAIN

Case-Control Study (Retrospective)

- Cases: 137 Self-Reporting Patients with back pain reduction
- Controls: 401 Population-Based Individuals matched to cases wrt demographic factors

<table>
<thead>
<tr>
<th></th>
<th>Pain reduction</th>
<th>No pain reduc</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSAID User</td>
<td>32</td>
<td>138</td>
<td>170</td>
</tr>
<tr>
<td>NSAID Non-User</td>
<td>105</td>
<td>263</td>
<td>368</td>
</tr>
<tr>
<td>Total</td>
<td>137</td>
<td>401</td>
<td>538</td>
</tr>
</tbody>
</table>

Example - NSAIDs and PAIN

\[ OR = \frac{32(263)}{138(105)} = \frac{8416}{14490} = 0.58 \]

\[ \text{var}[\ln(OR)] = \frac{1}{32} + \frac{1}{138} + \frac{1}{105} + \frac{1}{263} = 0.0518 \]

95% CI : \( (0.58e^{-1.96\sqrt{0.0518}}, 0.58e^{1.96\sqrt{0.0518}}) \equiv (0.37, 0.91) \)

Interval is entirely below 1, NSAID use appears to be lower among cases than controls.
Summary

\[ RD = p_1 - p_2 = \text{risk difference (null: } RD = 0) \]

• also known as \textit{attributable risk} or \textit{excess risk}

• measures \textit{absolute effect} – the proportion of cases among the exposed that can be attributed to exposure

\[ RR = \frac{p_1}{p_2} = \text{relative risk (null: } RR = 1) \]

• measures \textit{relative effect} of exposure

• bounded above by \( \frac{1}{p_2} \)

\[ OR = \frac{[p_1(1-p_2)]}{[p_2 (1-p_1)]} = \text{odds ratio (null: } OR = 1) \]

• range is 0 to \( \infty \)

• approximates \( RR \) for rare events

• invariant of switching rows and cols

• key parameter in logistic regression
Two main complications of analysis of single exposure effect

(1) Effect modifier - useful information

(2) Confounding factor - bias
Effect modifier

• Variation in the magnitude of measure of effect across levels of a third variable.

• Effect modification is not a bias but useful information

Happens when RR or OR is different between strata (subgroups of population)
Effect modifier

- To study interaction between risk factors
- To identify a subgroup with a lower or higher risk
- To target public health action
Confounding

- Distortion of measure of effect because of a third factor

- Should be prevented or Needs to be controlled for
Confounding

Be associated with exposure - without being the consequence of exposure

Be associated with outcome - independently of exposure
Confounding
(Simpson’s Paradox)

“Condom Use increases the risk of STD”

<table>
<thead>
<tr>
<th>Condom Use</th>
<th>Yes</th>
<th>55/95</th>
<th>(61%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>45/105</td>
<td>(43%)</td>
</tr>
</tbody>
</table>
Confounding
(Simpson’s Paradox)

BUT ...

<table>
<thead>
<tr>
<th># Partners &lt; 5</th>
<th>STD rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condom Use</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>5/15</td>
</tr>
<tr>
<td>No</td>
<td>30/82</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th># Partners &gt; 5</th>
<th>STD rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condom Use</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>50/80</td>
</tr>
<tr>
<td>No</td>
<td>15/23</td>
</tr>
</tbody>
</table>

Explanation: Individuals with more partners are more likely to use condoms. But individuals with more partners are also more likely to get STD.
Confounding - Causal Diagrams

E = Exposure
D = Disease
C = Potential Confounder

An apparent association between E and D is completely explained by C. C is a confounder.

An association between E and D is partly due to variations in C. C is a confounder.

C is in the causal path between E and D, a confounder.
Confounding - Causal Diagrams

C has an independent effect on D.

C is not a confounder.

The effect of C on D is completely contained in E. C is not a confounder.
Example – Genetic Association study

- Idiopathic Pulmonary Fibrosis (IPF) is known to be associated with age and gender (older and male are more likely)
- One study had 174 cases and 225 controls found association of IPF with one gene genotype COX2.8473 (C → T).

<table>
<thead>
<tr>
<th>Genotype</th>
<th>CC</th>
<th>CT</th>
<th>TT</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>88</td>
<td>72</td>
<td>14</td>
<td>174</td>
</tr>
<tr>
<td>Control</td>
<td>84</td>
<td>113</td>
<td>28</td>
<td>225</td>
</tr>
<tr>
<td>Total</td>
<td>172</td>
<td>185</td>
<td>42</td>
<td>399</td>
</tr>
</tbody>
</table>

- P-value by Pearson Chi-squares test: $p = 0.0241$.
- Q: Is this association true?
Example – Genetic Association study – continued

Stratify by sex or age

<table>
<thead>
<tr>
<th>Sex</th>
<th>male</th>
<th>female</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>108</td>
<td>66</td>
<td>174</td>
</tr>
<tr>
<td>Control</td>
<td>72</td>
<td>153</td>
<td>225</td>
</tr>
<tr>
<td>Total</td>
<td>180</td>
<td>219</td>
<td>399</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th>&lt;29</th>
<th>30-49</th>
<th>50-64</th>
<th>65-74</th>
<th>75+</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>0</td>
<td>10</td>
<td>42</td>
<td>68</td>
<td>54</td>
<td>174</td>
</tr>
<tr>
<td>Control</td>
<td>104</td>
<td>77</td>
<td>35</td>
<td>7</td>
<td>2</td>
<td>225</td>
</tr>
<tr>
<td>Total</td>
<td>104</td>
<td>87</td>
<td>77</td>
<td>75</td>
<td>56</td>
<td>399</td>
</tr>
</tbody>
</table>
Confounding

• Positive confounding
  - positively or negatively related to both the disease and exposure

• Negative confounding
  - positively related to disease but is negatively related to exposure or the reverse
How to prevent/control confounding?

Prevention (Design Stage)
- Restriction to one stratum
- Matching

Control (Analysis Stage)
- Stratified analysis
- Multivariable analysis
Choosing Confounders for Statistical Adjustment

• One school says choice should be based on a priori considerations

- Confounders selected based on their role as known risk factors for the disease

- Selection on basis of statistical significance of association with disease can leave residual confounding effect
Choosing Confounders for Statistical Adjustment

• Others say choice of confounders should be based on how much they affect RR (OR, RD) when included/ excluded from the model.

Compare crude measure of effect (RR or OR) to adjusted (weighted) measure of effect (Mantel Haenszel RR or OR)
To analyse effect modification
To control confounding

• Solution

Stratification (stratified analysis) Create strata according to categories inside the range of values taken by the effect modifier or the confounder
Mantel Haenszel Methods
Mantel Haenszel Methods - Notations

Assess association between disease and exposure after controlling for one or more confounding variables.

\[
\begin{array}{|c|c|c|}
\hline
D & E & E^- \\
\hline
a_i & b_i & (a_i + b_i) \\
\hline
\bar{D} & c_i & d_i & (c_i + d_i) \\
\hline
(a_i + c_i) & (b_i + d_i) & n_i \\
\hline
\end{array}
\]

where \( i = 1, 2, \ldots, K \) is the number of strata.
Cochran Mantel Haenszel Chi-square tests

(1) Correlation Statistic (Mantel-Haenszel statistic) has 1 df and assumes that either exposure or disease are measured on an ordinal (or interval) scale, when you have more than 2 levels.

(2) ANOVA (Row Mean Scores) Statistic has k-1 df and disease lies on an ordinal (or interval) scale when you have more than 2 levels.

(3) General Association Statistic has k-1 df and all scales accepted
(1) The Mantel-Haenszel estimate of the odds ratio assumes there is a common odds ratio:

$$\text{OR}_{\text{pool}} = \text{OR}_1 = \text{OR}_2 = \ldots = \text{OR}_K$$

To estimate the common odds ratio we take a weighted average of the stratum-specific odds ratios:

$$\hat{OR} = \frac{\sum_{i=1}^{K} a_i d_i / n_i}{\sum_{i=1}^{K} b_i c_i / n_i}$$
Mantel Haenszel Methods

(2) **Test of common odds ratio**

H₀: common OR is 1.0 vs. Hₐ: common OR ≠ 1.0

- A standard error is available for the MH common odds
- Standard CI intervals and test statistics are based on the standard normal distribution.

(3) **Test of effect modification** (heterogeneity, interaction)

H₀: OR₁ = OR₂ = … = ORₖ
Hₐ: not all stratum-specific OR’s are equal

Breslow-Day (SAS) homogeneity test can be used
Computing Cochran-Mantel-Haenszel Statistics for a Stratified Table

- The data set Migraine contains hypothetical data for a clinical trial of migraine treatment. Subjects of both genders receive either a new drug therapy or a placebo. Assess the effect of new drug adjusting for gender.

SAS manual
## Example - Migraine

### Response

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Better</th>
<th>Same</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>28</td>
<td>27</td>
<td>55</td>
</tr>
<tr>
<td>Placebo</td>
<td>12</td>
<td>39</td>
<td>51</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>66</td>
<td>106</td>
</tr>
</tbody>
</table>

Pearson Chi-squares test $p = 0.0037$

But after stratify by sex, it will be different for male vs female.
## Example – Migraine

### Male Response

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Better</th>
<th>Same</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>12</td>
<td>16</td>
<td>28</td>
</tr>
<tr>
<td>Placebo</td>
<td>7</td>
<td>19</td>
<td>26</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>19</td>
<td>35</td>
<td>54</td>
</tr>
</tbody>
</table>

p = 0.2205

### Female Response

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Better</th>
<th>Same</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>16</td>
<td>11</td>
<td>27</td>
</tr>
<tr>
<td>Placebo</td>
<td>5</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>21</td>
<td>31</td>
<td>52</td>
</tr>
</tbody>
</table>

p = 0.0039

10/27/2017
SAS- codes

data Migraine;
    input Gender $ Treatment $ Response $ Count @@;
datalines;
    female Active Better 16   female Active Same 11
    female Placebo Better 5   female Placebo Same 20
    male   Active Better 12   male   Active Same 16
    male   Placebo Better 7   male   Placebo Same 19
;

proc freq data=Migraine;
    weight Count;
    tables Gender*Treatment*Response / cmh noprint;
    title1 'Clinical Trial for Treatment of Migraine Headaches';
run;

************* In SAS, Need to put Exposure BEFORE Disease to generate right results for CMH results;
The FREQ Procedure

Summary Statistics for Treatment by Response
Controlling for Gender

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

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Alternative Hypothesis</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Nonzero Correlation</td>
<td>1</td>
<td>8.3052</td>
<td>0.0040</td>
</tr>
<tr>
<td>2</td>
<td>Row Mean Scores Differ</td>
<td>1</td>
<td>8.3052</td>
<td>0.0040</td>
</tr>
<tr>
<td>3</td>
<td>General Association</td>
<td>1</td>
<td>8.3052</td>
<td>0.0040</td>
</tr>
</tbody>
</table>

Estimates of the Common Relative Risk (Row1/Row2)

<table>
<thead>
<tr>
<th>Type of Study</th>
<th>Method</th>
<th>Value</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-Control</td>
<td>Mantel-Haenszel</td>
<td>3.3132</td>
<td>1.4456 7.5934</td>
</tr>
<tr>
<td></td>
<td>Logit</td>
<td>3.2941</td>
<td>1.4182 7.6515</td>
</tr>
<tr>
<td>Cohort (Col1 Risk)</td>
<td>Mantel-Haenszel</td>
<td>2.1636</td>
<td>1.2336 3.7948</td>
</tr>
<tr>
<td></td>
<td>Logit</td>
<td>2.1059</td>
<td>1.1951 3.7108</td>
</tr>
<tr>
<td>Cohort (Col2 Risk)</td>
<td>Mantel-Haenszel</td>
<td>0.6420</td>
<td>0.4705 0.8761</td>
</tr>
<tr>
<td></td>
<td>Logit</td>
<td>0.6613</td>
<td>0.4852 0.9013</td>
</tr>
</tbody>
</table>

Breslow-Day Test for Homogeneity of the Odds Ratios

<table>
<thead>
<tr>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.4929</td>
<td>1</td>
<td>0.2218</td>
</tr>
</tbody>
</table>

Total Sample Size = 106
Comments

- The significant $p$-value (0.004) indicates that the association between treatment and response remains strong after adjusting for gender.

- The probability of migraine improvement with the new drug is just over two times the probability of improvement with the placebo.

- The large $p$-value for the Breslow-Day test (0.2218) indicates no significant gender difference in the odds ratios.
Limitations of the Stratified Methods

- Can study only one independent variable at a time
- Problematic when there are too many variables to adjust for (too many strata)
- Limited to categorical variables (if continuous, can categorize, which may result in residual confounding)
Multiple Logistic Regression
Logistic regression

- Often response variables in health studies are binary, eg. disease vs no disease, damage vs no damage, death vs live, etc.

- To explain the variability of the binary variable by other variables, either continuous or categorical, such as age, sex, BMI, marriage status, socio-economic status, etc. use a statistical model to relate the probability of the response event to the explanatory variables.
Logistic regression

- Prob of event labeled as binary outcome

- Event \((Y = 1)\), no event \((Y = 0)\) model the mean:

\[
E(Y) = P(Y=1) \cdot 1 + P(Y=0) \cdot 0 = P(Y=1)
\]

but \(\pi = P(Y=1)\) is between 0 and 1

while \(\beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \ldots\) is a linear combination and may take any value.

A transformation is required.
Logistic regression

- Logistic regression model:
- logit function.

\[
\log \left( \frac{\pi}{1 - \pi} \right) = \beta_0 + x_1\beta_1 + \ldots + x_p\beta_p = \eta
\]

Equivalent to:
\[
p = \frac{\exp(\eta)}{1 + \exp(\eta)}
\]
Multiple Logistic Regression

Assumptions of Logistic Regression

- The independent variables are linear in the logit which may contain interaction and power terms
- The dependent variable is binary $Y=0$ or $1$
- The independent variables may be binary, categorical, continuous
Multiple Logistic Regression - Formulation

\[ E(Y \mid x) = P(Y = 1 \mid x) = \pi(x) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}} \]

\[ \ln \left[ \frac{\pi(x)}{1 - \pi(x)} \right] = \beta_0 + \beta_1 x + \cdots + \beta_p X_p \]

The relationship between \( \pi \) and \( x \) is S shaped
The logit (log-odds) transformation (link function)
Multiple Logistic Regression

Assess risk factors

- Individually \( H_0: \beta_k = 0 \)

- Globally \( H_0: \beta_m = \cdots \beta_{m+t} = 0 \)

while controlling for confounders and other important determinants of the event
Interpretation of the parameters

- If $\pi$ is the probability of an event and $O$ is the odds for that event then

$$Odds = \frac{\pi(x)}{1-\pi(x)} = \frac{\text{probability of event}}{\text{probability of no event}}$$

- The link function in logistic regression gives the log-odds

$$g(x) = \ln \left[ \frac{\pi(x)}{1-\pi(x)} \right] = \beta_0 + \beta_1 x + \cdots + \beta_p X_p$$
# Interpretation of parameter $\beta$ in logistic regression

Model: $\text{logit} (\pi) = \beta_0 + x_1 \beta_1$

<table>
<thead>
<tr>
<th></th>
<th>$Y=1$</th>
<th>$Y=0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X=1$</td>
<td>$\pi(1 \mid x = 1) = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}$</td>
<td>$1 - \pi(1 \mid x = 1) = \frac{1}{1 + e^{\beta_0 + \beta_1}}$</td>
</tr>
<tr>
<td>$X=0$</td>
<td>$\pi(1 \mid x = 0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$</td>
<td>$1 - \pi(1 \mid x = 0) = \frac{1}{1 + e^{\beta_0}}$</td>
</tr>
</tbody>
</table>

$$OR = \frac{\pi(1 \mid x = 1) \times [1 - \pi(1 \mid x = 0)]}{[1 - \pi(1 \mid x = 1)] \times \pi(1 \mid x = 0)} = e^{\beta_1}$$
An epidemiologic study surveyed 2484 subjects to examine whether snoring was a possible risk factor for heart disease.

<table>
<thead>
<tr>
<th>Heart Disease</th>
<th>Snoring</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Never</td>
<td>Occasional</td>
<td>Nearly every night</td>
<td>Every night</td>
</tr>
<tr>
<td>Yes</td>
<td>24</td>
<td>35</td>
<td>21</td>
<td>30</td>
</tr>
<tr>
<td>No</td>
<td>1355</td>
<td>603</td>
<td>192</td>
<td>224</td>
</tr>
<tr>
<td>Prop(yes)</td>
<td>.017</td>
<td>.055</td>
<td>.099</td>
<td>.118</td>
</tr>
</tbody>
</table>
Constructing Indicator variables

- Let $Z_1 = 1$ if occasional, 0 otherwise
- Let $Z_2 = 1$ if nearly every night, 0 otherwise
- Let $Z_3 = 1$ if every night, 0 otherwise
SAS Codes

data hd;
input hd $ snoring $ count;
Z1=(snoring="occa");
Z2=(snoring="nearly");
Z3=(snoring="every");
cards;
yes never  24
yes occa  35
yes nearly 21
yes every  30
no never 1355
no occa  603
no nearly 192
no every  224
;
run;

proc logistic data=hd
descending;
model hd (event="yes") =Z1 Z2 Z3;
freq count;
run;
The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-4.0335</td>
<td>0.2059</td>
<td>383.6641</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Z1</td>
<td>1</td>
<td>1.1869</td>
<td>0.2695</td>
<td>19.3959</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Z2</td>
<td>1</td>
<td>1.8205</td>
<td>0.3086</td>
<td>34.8027</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Z3</td>
<td>1</td>
<td>2.0231</td>
<td>0.2832</td>
<td>51.0313</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Odds Ratio Estimates

<table>
<thead>
<tr>
<th>Effect</th>
<th>Point Estimate</th>
<th>95% Wald Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z1</td>
<td>3.277</td>
<td>1.932</td>
</tr>
<tr>
<td>Z2</td>
<td>6.175</td>
<td>3.373</td>
</tr>
<tr>
<td>Z3</td>
<td>7.561</td>
<td>4.341</td>
</tr>
</tbody>
</table>
Calculating Probabilities

- The fitted logistic regression function is
  \[ \text{Logit}(\pi) = -4.0335 + 1.1869 Z_1 + 1.8205 Z_2 + 2.0231 Z_3 \]

- So, the probability of heart disease if never snore is
  \[ \frac{\exp(-4.0335)}{1 + \exp(-4.0335)} = 0.0174 \]

- If snore occasionally,
  \[ \frac{\exp(-4.0335 + 1.1869)}{1 + \exp(-4.0335 + 1.1869)} = 0.0549 \]
Calculating Odds Ratios

- If $Z_1=Z_2=Z_3=0$, then odds are $\exp(-4.0335)$

- If $Z_2=Z_3=0$, but $Z_1=1$, then odds are $\exp(-4.0335+1.1869)$

- The ratio of odds is then $\exp(1.1869)$

- What is the odds ratio for comparing those who snore nearly every night with occasional snorers?

- What is the odds ratio for comparing those who snore every night with those who snore nearly every night?
Idiopathic Pulmonary Fibrosis (IPF) is known to be associated with age and gender (older and male are more likely).

One study had 174 cases and 225 controls found association of IPF with one gene genotype COX2.8473 (C → T).

<table>
<thead>
<tr>
<th>Genotype</th>
<th>CC</th>
<th>CT</th>
<th>TT</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>88</td>
<td>72</td>
<td>14</td>
<td>174</td>
</tr>
<tr>
<td>Control</td>
<td>84</td>
<td>113</td>
<td>28</td>
<td>225</td>
</tr>
<tr>
<td>Total</td>
<td>172</td>
<td>185</td>
<td>42</td>
<td>399</td>
</tr>
</tbody>
</table>

P-value by Pearson Chi-squares test: $p = 0.0241$. 
Q: Is this association true?
Old example on genetic effect of SNP COX2.8473 on IPF

- Logistic regression model
  \[
  \text{logit } [\Pr(\text{IPF})] = \text{intercept} + \text{snp} + \text{sex} + \text{age}
  \]

- Results: Wald
  
<table>
<thead>
<tr>
<th>Effect</th>
<th>DF</th>
<th>Chi-square</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP</td>
<td>2</td>
<td>2.7811</td>
<td>0.2489</td>
</tr>
<tr>
<td>sex</td>
<td>1</td>
<td>9.1172</td>
<td>0.0025</td>
</tr>
<tr>
<td>age</td>
<td>1</td>
<td>100.454</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
Conclusion: What we have learned

1. Define study designs
2. Measures of effects for categorical data (unstratified analysis)
3. Confounders and effects modifications
4. Stratified analysis (Mantel-Haenszel statistic, multiple logistic regression)
5. Use of SAS Proc FREQ and Proc Logistic
Conclusion: Further readings

Read textbook for

1. Power and sample size calculations

2. Tests in matched pair studies