# Contingency Table Probability Structure - II

Dipankar Bandyopadhyay

Department of Biostatistics,
Virginia Commonwealth University

BIOS 625: Categorical Data & GLM

# Recall from previous lecture

- For a contingency table resulting from a prospective study, we derived

$$E_{ij} = \frac{[i^{th} \text{ row total}] \cdot [j^{th} \text{ column total}]}{[\text{total sample size } (n_1 + n_2)\ ]}$$

- and the corresponding likelihood ratio test

$$G^2 = 2 \sum_{i=1}^{2} \sum_{j=1}^{2} O_{ij} \log \left( \frac{O_{ij}}{E_{ij}} \right)$$

- where $O_{ij}$ is the observed cell count in cell $i, j$

# PEARSON'S CHI-SQUARE

- Another Statistic which is a function of the $O_{ij}$'s and $E_{ij}$'s is **PEARSON'S CHI-SQUARE**.
- However, as we will see, Pearson's Chi-Square is actually just a $Z-$statistic for testing

$$\mathrm{H}_0: p_1 = p_2 = p \quad \text{versus} \quad \mathrm{H}_A: p_1 \neq p_2 \ ,$$

where the standard error is calculated under the null.

- Recall, the WALD statistic is

$$Z_W = \frac{(\widehat{p}_1 - \widehat{p}_2) - 0}{\sqrt{n_1^{-1}\widehat{p}_1(1-\widehat{p}_1) + n_2^{-1}\widehat{p}_2(1-\widehat{p}_2)}}$$

- Note that we used the variance of $(\widehat{p}_1 - \widehat{p}_2)$ calculated under the alternative $p_1 \neq p_2$.
- Under the null $p_1 = p_2 = p$, the variance simplifies to

$$\begin{aligned} Var(\widehat{p}_1 - \widehat{p}_2) &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \\ &= p(1-p)\left[\frac{1}{n_1} + \frac{1}{n_2}\right] \end{aligned}$$

- Then, we can use the following test statistic (with the variance estimated under the null),

$$Z_S = \frac{(\widehat{p}_1 - \widehat{p}_2) - 0}{\sqrt{\tilde{p}(1-\tilde{p})[n_1^{-1} + n_2^{-1}]}} \sim N(0,1)$$

where the pooled estimate is used in the variance

$$\tilde{p} = \left( \frac{Y_1 + Y_2}{n_1 + n_2} \right)$$

- If we square $Z_S$, we get

$$X^2 = Z_S^2 = \left( \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\tilde{p}(1-\tilde{p})[n_1^{-1} + n_2^{-1}]}} \right)^2 \sim \chi_1^2$$

  under the null hypothesis.
- After some algebra (i.e., pages), we can write $X^2$ in terms of the $O_{ij}$'s and $E_{ij}$'s .
- Instead of pages of algebra, how about an empirical proof?

- Consider the following example

| | Success | Failure | |
|---|---|---|---|
| Group 1 | 15 | 135 | 150 |
| Group 2 | 10 | 40 | 50 |
| Totals | 25 | 175 | 200 |

- Here

$$\tilde{p} = (15 + 10)/200 = 0.125$$

- With SE under the null as

$$\widehat{SE_0}(\widehat{p}_2 - \widehat{p}_1) = \sqrt{0.125 * (1 - 0.125) * (150^{-1} + 50^{-1})} = 0.054006$$

- Then

$$Z_s = \frac{(10/50 - 15/150)}{0.054006} = \frac{0.1}{0.054006} = 1.8516402$$

- and

$$Z^2 = 3.428571$$

# Pearson Chi Square

- Likewise, we can use the previous definition of the observed ($O_{ij}$) and expected ($E_{ij}$) to calculate

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

  which is known as **'Pearson's Chi-Square'** for a $(2 \times 2)$ table.

- Note, 'Pearson's Chi-Square' measures the discrepancy between the observed counts, and the estimated expected counts under the null; if they are similar, you would expect the statistic to be small, and the null not to be rejected.

- For our example, the matrix of expected counts is

| Expected | | Total |
|---|---|---|
| 18.75 | 131.25 | 150 |
| 6.25 | 43.75 | 50 |
| 25 | 175 | 200 |

- and

$$\begin{aligned} X^2 &= (15-18.75)^2/18.75 + (135-131.25)^2/131.25 + \\ &\quad (10-6.25)^2/6.25 + (40-43.75)^2/43.75 \\ &= 0.75 + 0.107142857 + 2.25 + 0.321428571 \\ &= 3.428571 \end{aligned}$$

- While not a true proof, this does indeed confirm that Pearson's Chi Square is simply the score test for the difference in proportions.

# SAS– MI Example

- Recall our MI example from the previous lecture

| | Myocardial Infarction | |
|---|---|---|
| | Fatal Attack or Nonfatal attack | No Attack |
| Placebo | 189 | 10845 |
| Aspirin | 104 | 10933 |

- We want to investigate whether or not Aspirin is beneficial in the prevention of an MI

# Using SAS

SAS code below:

```
data one;
input trt $ out $ y;
cards;
1(P) HA    189
1(P) NHA 10845
2(A) HA    104
2(A) NHA 10933
;
proc freq;
 table trt*out / expected chisq measures;
 weight y; /* tells SAS how many obs.   */
           /* in each cell of 2x2 table */
run;
```

TABLE OF TRT BY OUT

| TRT<br>Frequency<br>Expected<br>Percent<br>Row Pct<br>Col Pct | OUT<br>HA | <br>NHA | Total |
|---|---|---|---|
| 1(P) | 189 | 10845 | 11034 |
| | 146.48 | 10888 | |
| | 0.86 | 49.14 | 49.99 |
| | 1.71 | 98.29 | |
| | 64.51 | 49.80 | |
| 2(A) | 104 | 10933 | 11037 |
| | 146.52 | 10890 | |
| | 0.47 | 49.54 | 50.01 |
| | 0.94 | 99.06 | |
| | 35.49 | 50.20 | |
| Total | 293 | 21778 | 22071 |
| | 1.33 | 98.67 | 100.00 |

The second row in each cell is $E_{ij}$

# Estimated Expected Cell Counts

- If you work thru the $(2 \times 2)$ table, you will see

$$
\begin{aligned}
E_{11} &= 146.48 \\
&= \frac{[1^{st} \text{ row total}] \cdot [1^{st} \text{ column total}]}{[\text{total sample size } (n_1 + n_2)\ ]} \\
&= \frac{(11034)(293)}{22071}
\end{aligned}
$$

$$
\begin{aligned}
E_{12} &= 10888 \\
&= \frac{[1^{st} \text{ row total}] \cdot [2^{nd} \text{ column total}]}{[\text{total sample size } (n_1 + n_2)\ ]} \\
&= \frac{(11034)(21778)}{22071}
\end{aligned}
$$

$$
\begin{aligned}
E_{21} &= 146.52 \\
&= \frac{[2^{nd} \text{ row total}] \cdot [1^{st} \text{ column total}]}{[\text{total sample size } (n_1 + n_2)\ ]} \\
&= \frac{(11037)(293)}{22071}
\end{aligned}
$$

and

$$
\begin{aligned}
E_{21} &= 10890 \\
&= \frac{[2^{nd} \text{ row total}] \cdot [2^{nd} \text{ column total}]}{[\text{total sample size } (n_1 + n_2)\ ]} \\
&= \frac{(11037)(21778)}{22071}
\end{aligned}
$$

# More SAS PROC FREQ OUTPUT

STATISTICS FOR TABLE OF TRT BY OUT

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 25.014 | 0.000<=(Pearson's,Score) |
| Likelihood Ratio Chi-Square | 1 | 25.372 | 0.000<=LR STAT |
| Continuity Adj. Chi-Square | 1 | 24.429 | 0.000 |
| Mantel-Haenszel Chi-Square | 1 | 25.013 | 0.000 |
| Fisher's Exact Test (Left) | | | 1.000 |
| (Right) | | | 3.25E-07 |
| (2-Tail) | | | 5.03E-07 |
| Phi Coefficient | | 0.034 | |
| Contingency Coefficient | | 0.034 | |
| Cramer's V | | 0.034 | |

```
Estimates of the Relative Risk (Row1/Row2)
                                        95%
Type of Study          Value   Confidence Bounds
-----------------------------------------------------
Case-Control           1.832    1.440    2.331 <=(OR, using logOR)
Cohort (Col1 Risk)     1.818    1.433    2.306 <=(RR, using logRR)
Cohort (Col2 Risk)     0.992    0.989    0.995
Sample Size = 22071
```

# Comparing Test Statistics

- We want to compare test statistics for

$$H_0: p_1 = p_2 = p \quad \text{versus} \quad H_A: p_1 \neq p_2$$

- Recall our results from the previous lecture,

| Parameter | Estimate | Estimated Standard Error | $Z-$Statistic (Est/SE) |
|---|---|---|---|
| RISK DIFF | .0077 | .00154 | 5.00 |
| log(RR) | .598 (RR=1.818) | .1212 | 4.934 |
| log(OR) | .605 (OR=1.832) | .1228 | 4.927 |

- Looking at the (square of the) WALD statistics from earlier, as well as the Likelihood Ratio and Pearson's Chi-Square calculated by SAS, we have

| STATISTIC | VALUE |
|---|---|
| WALD | |
| RISK DIFF | 25.00 |
| log(RR) | 24.34 |
| log(OR) | 24.28 |
| LR | 25.37 |
| Pearson's | 25.01 |

- We see that all of the statistics are almost identical. We would reject the null using any of them (the .05 quantile is $3.84 = 1.96^2$.
- All of the test statistics are approximately $\chi^2_1$ under the null, and are actually equivalent at $n_1 = \infty$ and $n_2 = \infty$.
- Under a given alternative, all will have high power (although not exactly identical).
- Note, the likelihood ratio and Pearson's Chi-Square statistic just depend on the predicted probabilities (i.e., the 'Estimated Expected Cell Counts'). and not how we measure the treatment difference.
- However, the WALD statistic does depend on what treatment difference (Risk Difference, log OR, or log RR) we use in the test statistic.
- In other words, the WALD test statistics using the Risk Difference, log OR, and log RR will usually be slightly different (as we see in the example).

# Empirical Logits

- Recall, we can write the estimated log-odds ratio as

$$\begin{aligned}
\log \widehat{OR} &= \log\left(\frac{\widehat{p}_1}{1-\widehat{p}_1}\right) - \log\left(\frac{\widehat{p}_2}{1-\widehat{p}_2}\right) \\
&= \log\left(\frac{y_1/n_1}{(n_1-y_1)/n_1}\right) - \log\left(\frac{y_2/n_2}{(n_2-y_2)/n_2}\right) \\
&= \log\left(\frac{y_1}{n_1-y_1}\right) - \log\left(\frac{y_2}{n_2-y_2}\right) \\
&= \log(y_1) - \log(n_1 - y_1) \\
&\qquad - \log(y_2) + \log(n_2 - y_2)
\end{aligned}$$

- Question: What happens if $y_1 = 0$, or $y_1 = n_1$, $(n_1 - y_1 = 0)$ or $y_2 = 0$, or $y_2 = n_2$, $(n_2 - y_2 = 0)$, so that $\log \widehat{OR}$ is indeterminate ?
- How will you adjust?

- Instead of

$$\left(\frac{y_t}{n_t - y_t}\right)$$

use

$$\left(\frac{y_t + a}{(n_t - y_t) + a}\right)$$

where the constant $a > 0$ is chosen so that, as nearly as possible,

$$E\left(\frac{y_t + a}{(n_t - y_t) + a}\right) = \frac{p_t}{1 - p_t}$$

- Haldane (1956) showed by a first order Taylor Series approximation,

$$a = .5$$

- The quantity

$$\log\left(\frac{y_t + .5}{(n_t - y_t) + .5}\right)$$

  is called an "empirical logit",
- The "empirical logit" has smaller finite sample bias than the usual logit.

- Using empirical logits is like adding .5 to each cell of the $(2 \times 2)$ table, and get

$$\widehat{OR}^{E} = \frac{(Y_1 + .5)(n_2 - Y_2 + .5)}{(Y_2 + .5)(n_1 - Y_2 + .5)}$$

and

$$\widehat{Var}\{\log[\widehat{OR}^{E}]\} = \frac{1}{y_1 + .5} + \frac{1}{n_1 - y_1 + .5} + \frac{1}{y_2 + .5} + \frac{1}{n_2 - y_2 + .5}$$

- The empirical logit was used more before exact computer methods became available (we will discuss these later).
- Not always liked because, some investigators feel that you are adding 'fake' data, even though, it does have smaller finite sample bias, and, is asymptotically the same as the usual estimate of the log odds ratio.

# Case-Control Studies: Probability Structure

- Alcohol Consumption and occurrence of esophageal cancer (Tuyms et al., Bulletin of Cancer, 1974)
- It is not ethical to randomize patients in a prospective study

| | | STATUS | | |
|---|---|---|---|---|
| | | CASE | CONTROL | Total |
| ALCOHOL | 80+ (gm/day) | 96 | 109 | 205 |
| | 0-79 (gm/day) | 104 | 666 | 770 |
| | Total | 200 (fixed by design) | 775 (fixed by design) | 975 |

- Cases in this study were 200 male esophageal cancer patients in regional hospitals; 775 controls were randomly sampled from the same regions.
- After being selected in the study, the subjects were then questioned about the consumption of alcohol (as well as other things) in the previous 10 years.

# Case-Control Design

- Number of cases and controls (usually the outcomes) are fixed by design and exposures are random.
- Columns are independent binomials.
- Question of interest:
  Does alcohol exposure vary among cases and controls?
  Is alcohol exposure associated with esophageal cancer?

# Comparison to Prospective Design

- Suppose you use SAS as if the data were a prospective study.
- Would your analyses be OK ?

```
data one;
input exp $ ca $ count;
cards;
1 1  96
1 2 109
2 1 104
2 2 666
;
proc freq;
 table exp*ca / expected chisq measures;
 weight count; /* tells SAS how many obs.   */
               /* in each cell of 2x2 table */
run;
```

# Selected Results

| EXP / Frequency / Expected / Percent / Row Pct / Col Pct | CA 1 | CA 2 | Total |
|---|---|---|---|
| 1 | 96 | 109 | 205 |
| | 42.051 | 162.95 | |
| | 9.85 | 11.18 | 21.03 |
| | 46.83 | 53.17 | |
| | 48.00 | 14.06 | |
| 2 | 104 | 666 | 770 |
| | 157.95 | 612.05 | |
| | 10.67 | 68.31 | 78.97 |
| | 13.51 | 86.49 | |
| | 52.00 | 85.94 | |
| Total | 200 | 775 | 975 |
| | 20.51 | 79.49 | 100.00 |

STATISTICS FOR TABLE OF EXP BY CA

| Statistic | DF | Value | Prob | |
|---|---|---|---|---|
| Chi-Square | 1 | 110.255 | 0.000 | (Pearson's) |
| Likelihood Ratio Chi-Square | 1 | 96.433 | 0.000 | (G^2) |
| Continuity Adj. Chi-Square | 1 | 108.221 | 0.000 | |
| Mantel-Haenszel Chi-Square | 1 | 110.142 | 0.000 | |
| Fisher's Exact Test (Left) | | | 1.000 | |
| (Right) | | | 1.03E-22 | |
| (2-Tail) | | | 1.08E-22 | |
| Phi Coefficient | | 0.336 | | |
| Contingency Coefficient | | 0.319 | | |
| Cramer's V | | 0.336 | | |

Estimates of the Relative Risk (Row1/Row2)

| Type of Study | Value | 95% Confidence | Bounds | |
|---|---|---|---|---|
| Case-Control | 5.640 | 4.001 | 7.951 | (OR) |
| Cohort (Col1 Risk) | 3.467 | 2.753 | 4.367 | |
| Cohort (Col2 Risk) | 0.615 | 0.539 | 0.701 | |

# General Case Control Study

- Disease Status is known and fixed in advance:
- First, you go to a hospital and get patients with lung cancer (case) and patients without lung cancer (control)
- Conditional on CASE/CONTROL status, exposure is the response:
  Go back in time to find exposure, i.e., smoked (exposed) and didn't smoke (unexposed).

# Summary Counts

| | | DISEASE STATUS | | |
|---|---|---|---|---|
| | | Case | Control | |
| EXPOSED | YES | $Y_1$ | $Y_2$ | |
| | NO | $n_1 - Y_1$ | $n_2 - Y_2$ | |
| | | $n_1$ | $n_2$ | |

# Setting is similar to a prospective study

- $n_1$ and $n_2$ (columns) are fixed by design
- $Y_1$ and $Y_2$ are independent with distributions:

$$Y_1 \sim Bin(n_1, \pi_1) \quad \text{and} \quad Y_2 \sim Bin(n_2, \pi_2)$$

where

$$\pi_1 = P[\text{Exposed}|\text{Case}] \quad \text{and} \quad \pi_2 = P[\text{Exposed}|\text{Control}]$$

- The $(2 \times 2)$ table of probabilities are

| | | DISEASE 1 | DISEASE 2 | total |
|---|---|---|---|---|
| EXPOSE | 1 | $\pi_1$ | $\pi_2$ | $(\pi_1 + \pi_2)$ |
| | 2 | $(1 - \pi_1)$ | $(1 - \pi_2)$ | $[2 - (\pi_1 + \pi_2)]$ |
| total | | 1 | 1 | 2 |

- In a case-control study, $\pi_1$, $\pi_2$ and any parameters that can be expressed as functions of $\pi_1$ and $\pi_2$ can be estimated.
- However, the quantities of interest are not $\pi_1$, $\pi_2$ but, instead, are

$$p_1 = P[\text{Case}|\text{Exposed}] \quad \text{and} \quad p_2 = P[\text{Case}|\text{Unexposed}],$$

in the $(2 \times 2)$ table:

| | | DISEASE 1 | DISEASE 2 | |
|---|---|---|---|---|
| EXPOSE | 1 | $p_1$ | $(1-p_1)$ | 1 |
| | 2 | $p_2$ | $(1-p_2)$ | 1 |

- In the CASE-CONTROL study, we want to know:
  Does exposure affect the risk of (subsequent) disease ?
- **Problem**: $p_1$ and $p_2$ cannot be estimated from this type of design (i.e., neither can be expressed as functions of the quantities which can be estimated, $\pi_1$ and $\pi_2$).

- Since we are allowed to choose the number of cases and controls in the study, we could just as easily have chosen 775 cases and 200 controls.
- Thus, the proportion of cases is chosen by design, and could have nothing to do with the real world. Esophageal cancer is a rare disease. There is no way that the probability of Esophageal cancer in the population is

$$\widehat{P}[\text{Case}] = \frac{200}{975} = .205$$

- Further, the estimates

$$\widehat{p}_1 = \widehat{P}[\text{Case}|\text{Exposed}] = \frac{96}{205} = .47$$

and

$$\widehat{p}_2 = \widehat{P}[\text{Case}|\text{Unexposed}] = \frac{104}{770} = .14$$

are not even close to what they are in the real world.
- Bottom line: cannot estimate $p_1$ and $p_2$ with case-control data.

# ODDS RATIO

- However, we will now show that, even though $p_1$ and $p_2$ can not be estimated, the "odds ratio" as if the study were prospective, can be estimated from a case-control study, i.e., we can estimate

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)}$$

- We will use Baye's Rule to show that you can estimate the OR from a case-control study. Baye's rule states that

$$\begin{aligned} P[\text{A}|\text{B}] &= \frac{P[\text{AB}]}{P[\text{B}]} = \frac{P[\text{B}|\text{A}]P[\text{A}]}{P[\text{B}]} \\ &= \frac{P[\text{B}|\text{A}]P[\text{A}]}{P[\text{B}|\text{A}]P[\text{A}] + P[\text{B}|\text{not A}]P[\text{not A}]} \end{aligned}$$

- For example, applying Bayes's rule to

$$p_1 = P[\text{Case}|\text{Exposed}],$$

we get

$$\begin{aligned} p_1 &= P[\text{Case}|\text{Exposed}] \\ &= \frac{P[\text{Exposed}|\text{Case}]P[\text{Case}]}{P[\text{Exposed}]} \\ &= \pi_1 \left( \frac{P[\text{Case}]}{P[\text{Exposed}]} \right) , \end{aligned}$$

where, recall

$\pi_1 = P[\text{Exposed}|\text{Case}]$

- By applying Baye's rule to each of the probabilities in the odds ratio for a prospective study, $p_1$, $(1 - p_1)$, $p_2$ and $(1 - p_2)$, you can show that

# The odds ratio for a prospective study equals

$$\begin{aligned}
\frac{p_1/(1-p_1)}{p_2/(1-p_2)} &= \frac{\left(\frac{\pi_1}{\pi_2}\right)\left[\frac{P[\text{Case}]}{P[\text{Control}]}\right]}{\left(\frac{1-\pi_1}{1-\pi_2}\right)\left[\frac{P[\text{Case}]}{P[\text{Control}]}\right]} \\
&= \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \\
&= \text{OR from case-control } (2 \times 2) \text{ table}
\end{aligned}$$

where

$$\pi_1/(1-\pi_1)$$

is the "odds" of being exposed given a case, and

$$\pi_2/(1-\pi_2)$$

is the "odds" of being exposed given a control.

Thus, we can estimate $OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$ with an estimate of $OR = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$ since the OR can be equivalently defined in terms of the $p$'s or the $\pi$'s.

# Proof

Using Baye's Rule, first, let's rewrite

$$\frac{p_1}{1-p_1} = \frac{P[\text{Case}|\text{Exposed}]}{P[\text{Control}|\text{Exposed}]}$$

Now,

$$\begin{aligned} p_1 &= P[\text{Case}|\text{Exposed}] \\ &= \frac{P[\text{Exposed}|\text{Case}]P[\text{Case}]}{P[\text{Exposed}]} \\ &= \pi_1 \left( \frac{P[\text{Case}]}{P[\text{Exposed}]} \right) \end{aligned}$$

and

$$\begin{aligned} 1-p_1 &= P[\text{Control}|\text{Exposed}] \\ &= \frac{P[\text{Exposed}|\text{Control}]P[\text{Control}]}{P[\text{Exposed}]} \\ &= \pi_2 \left( \frac{P[\text{Control}]}{P[\text{Exposed}]} \right) \end{aligned}$$

Then

$$\begin{aligned}\frac{p_1}{1-p_1} &= \left(\frac{\pi_1}{\pi_2}\right)\left[\frac{P[\text{Case}]/P[\text{Exposed}]}{P[\text{Control}]/P[\text{Exposed}]}\right] \\ &= \left(\frac{\pi_1}{\pi_2}\right)\left[\frac{P[\text{Case}]}{P[\text{Control}]}\right]\end{aligned}$$

Similarly, you can show that

$$\begin{aligned}\frac{p_2}{1-p_2} &= \left(\frac{1-\pi_1}{1-\pi_2}\right)\left[\frac{P[\text{Case}]/P[\text{Unexposed}]}{P[\text{Control}]/P[\text{Unexposed}]}\right] \\ &= \left(\frac{1-\pi_1}{1-\pi_2}\right)\left[\frac{P[\text{Case}]}{P[\text{Control}]}\right]\end{aligned}$$

Then, the odds ratio is

$$\begin{aligned}
\frac{p_1/(1-p_1)}{p_2/(1-p_2)} &= \frac{\left(\frac{\pi_1}{\pi_2}\right)\left[\frac{P[\text{Case}]}{P[\text{Control}]}\right]}{\left(\frac{1-\pi_1}{1-\pi_2}\right)\left[\frac{P[\text{Case}]}{P[\text{Control}]}\right]} \\
&= \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \\
&= \text{OR from case-control } (2 \times 2) \text{ table,}
\end{aligned}$$

where

$$\pi_1/(1-\pi_1)$$

is the "odds" of being exposed given a case, and

$$\pi_2/(1-\pi_2)$$

is the "odds" of being exposed given a control.

# Notes

- OR in terms of $(p_1, p_2)$ is the same as OR in terms of $(\pi_1, \pi_2)$
- OR, which measures how much $p_1$ and $p_2$ differ, can be estimated from a case-control study, even though $p_1$ and $p_2$ cannot.
- We can make inferences about OR, without being able to estimate $p_1$ and $p_2$.
- If we have additional information on $P[\text{Case}]$ or $P[\text{Exposed}]$, then we can estimate $p_1$ and $p_2$.
- Then for a case-control study, we usually are only interested in estimating the OR and testing if it equals some specified value (usually 1).

# Estimates

The likelihood is again product binomial (the 2 columns are independent binomials):

$$\begin{aligned} L(\pi_1, \pi_2) &= P(Y_1 = y_1|\pi_1)P(Y_2 = y_2|\pi_2) \\ &= \left( \begin{array}{c} n_1 \\ y_1 \end{array} \right) \left( \begin{array}{c} n_2 \\ y_2 \end{array} \right) \pi_1^{y_1}(1-\pi_1)^{n_1-y_1}\pi_2^{y_2}(1-\pi_2)^{n_2-y_2} \end{aligned}$$

# Question of interest

Are exposure and case control status associated?

Estimating the OR to look at this association is of the most interest, but to estimate the

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

we must first estimate

$$\pi_1 = P[\text{Exposed}|\text{Case}]$$

and

$$\pi_2 = P[\text{Exposed}|\text{Control}]$$

- Going thru the same likelihood theory as we did for estimating $(p_1, p_2)$ from two independent binomials in a prospective study, the MLE's of $(\pi_1, \pi_2)$ are the proportions exposed given case and control, respectively,

$$\widehat{\pi}_1 = \frac{Y_1}{n_1} \qquad \text{and} \qquad \widehat{\pi}_2 = \frac{Y_2}{n_2}$$

- Then,

$$\begin{aligned}
\widehat{OR} &= \frac{\widehat{\pi}_1/(1-\widehat{\pi}_1)}{\widehat{\pi}_2/(1-\widehat{\pi}_2)} \\
&= \frac{(y_1/n_1)/[1-(y_1/n_1)]}{(y_2/n_2)/[1-(y_2/n_2)]} \\
&= \frac{y_1(n_2-y_2)}{y_2(n_1-y_1)}
\end{aligned}$$

# Estimated Odds ratio

Looking at the $(2 \times 2)$ table of observed counts,

| | DISEASE STATUS 1 | 2 | |
|---|---|---|---|
| EXPOS 1 | $Y_1$ | $Y_2$ | $Y_1 + Y_2$ |
| EXPOS 2 | $(n_1 - Y_1)$ | $(n_2 - Y_2)$ | $[(n_1 + n_2) - (Y_1 + Y_1)]$ |
| total | $n_1$ | $n_2$ | $(n_1 + n_2)$ |

and again letting $O_{ij}$ be the count in the $ij^{th}$ cell of the $(2 \times 2)$ table, we can rewrite the table as

| | DISEASE STATUS 1 | 2 | |
|---|---|---|---|
| EXPOS 1 | $O_{11}$ | $O_{12}$ | $O_{11} + O_{12}$ |
| EXPOS 2 | $O_{21}$ | $O_{22}$ | $O_{21} + O_{22}$ |
| total | $O_{11} + O_{21}$ | $O_{12} + O_{22}$ | |

The estimated odds ratio equals

$$\begin{aligned}\widehat{OR} &= \frac{y_1(n_2 - y_2)}{y_2(n_1 - y_1)} \\ &= \frac{O_{11}O_{22}}{O_{12}O_{21}},\end{aligned}$$

which is the same thing we would get if we treated the case-control data as if it was prospective data.

The null hypothesis of no association is or, usually,

$$\mathrm{H}_0: OR = 1$$

and the alternative is

$$\mathrm{H}_A: OR \neq 1$$

Where,

$$\widehat{OR} = \frac{y_1(n_2 - y_2)}{y_2(n_1 - y_1)} = \frac{O_{11}O_{22}}{O_{12}O_{21}},$$

(which is the same as if the study was a prospective study.)

# Wald Statistic based on estimated OR

- Again, the $\log(\widehat{OR})$ is often used is test statistics since it goes from $-\infty$ to $\infty$ and is more approximately normal than the $OR$, which is strictly positive.
- The MLE of $\log OR$ is $\log \widehat{OR}$
- Similar to a prospective study,

$$Var[\log(\widehat{OR})] = \frac{1}{n_1\pi_1} + \frac{1}{n_1(1-\pi_1)} + \frac{1}{n_2\pi_2} + \frac{1}{n_2(1-\pi_2)}$$

- which is estimated by

$$\begin{aligned}
\widehat{Var}[\log(\widehat{OR})] &= \frac{1}{n_1\widehat{\pi}_1} + \frac{1}{n_1(1-\widehat{\pi}_1)} + \frac{1}{n_2\widehat{\pi}_2} + \frac{1}{n_2(1-\widehat{\pi}_2)} \\
&= \frac{1}{y_1} + \frac{1}{n_1-y_1} + \frac{1}{y_2} + \frac{1}{n_2-y_2} \\
&= \frac{1}{O_{11}} + \frac{1}{O_{12}} + \frac{1}{O_{21}} + \frac{1}{O_{22}},
\end{aligned}$$

which is identical to what we would get if we had assumed the study was a prospective study.

- The WALD statistic for $H_0 : OR = 1$, i.e., $H_0 : \log(OR) = 0$, is

$$Z = \frac{\log(\widehat{OR}) - 0}{\sqrt{\widehat{Var}(\log(\widehat{OR}))}},$$

- Also, a 95% confidence interval for the odds ratio is

$$\exp\{\log(\widehat{OR}) \pm 1.96\sqrt{\widehat{Var}[\log(\widehat{OR})]}\}$$

- The bottom line here is that you could treat case-control data as if it came from a prospective study and get the same test statistic and confidence interval described here.

# Double Dichotomy or Cross-sectional

| | | Job Satisfaction | | |
|---|---|---|---|---|
| | | Dissatisfied | Satisfied | |
| Income | $< \$15,000$ | 104 | 391 | 495 |
| | $\geq \$15,000$ | 66 | 340 | 406 |
| | | 170 | 731 | 901 |

- Neither margin is fixed by design, although the total sample size $n$ (901) is fixed
- **Study Design**–Randomly select $n$ (fixed) independent subjects and classify each subject on 2 variables, say $X$ and $Y$, each with two levels
- For example,

$$X = \text{Income} = \begin{cases} 1 \text{ if } < \$15,000 \\ 2 \text{ if } \geq \$15,000 \end{cases}$$

$$Y = \text{JOB SATISFACTION} = \begin{cases} 1 \text{ if Not Satisfied} \\ 2 \text{ if Satisfied} \end{cases}$$

# Question of interest

- Are $X$ and $Y$ associated or are they independent ?
- Under independence,

$$P[(X = i), (Y = j)] = P[X = i] \cdot P[Y = j],$$

i.e.,

$$p_{ij} = p_{i\cdot}.p_{\cdot j}$$

- Then, the null hypothesis is

$$\mathrm{H}_0 : p_{ij} = p_{i\cdot}.p_{\cdot j} \qquad \text{for } i, j = 1, 2.$$

and the alternative is

$$\mathrm{H}_A : p_{ij} \neq p_{i\cdot}.p_{\cdot j}$$

# Parameters of interest

- We are interested in the association between $X$ and $Y$.
- We may ask: Are $X$ and $Y$ independent ?
- In the Double Dichotomy, if one variable is thought of as an outcome (say $Y$), and the other as a covariate, say $X$, then we can condition on $X$, and look at the risk difference, the relative risk and the odds ratio, just as in the prospective study.
- In the prospective study, $p_1$ was the probability of outcome 1 ($Y = 1$) given treatment 1 ($X = 1$), which, in terms of the probabilities for the Double Dichotomy, is

$$p_1 = P[Y = 1|X = 1] = \frac{P[(X = 1), (Y = 1)]}{P[X = 1]} = \frac{p_{11}}{p_{1\cdot}}$$

- Similarly,

$$p_2 = P[Y = 1|X = 2] = \frac{P[(X = 2), (Y = 1)]}{P[X = 2]} = \frac{p_{21}}{p_{2\cdot}}$$

# The RELATIVE RISK

- Then, the RELATIVE RISK is

$$RR = \frac{p_1}{p_2} = \frac{[p_{11}/p_{1\cdot}]}{[p_{21}/p_{2\cdot}]}$$

- Now, suppose $X$ and $Y$ are independent, i.e.,

$$p_{ij} = p_{i\cdot}p_{\cdot j}$$

then

$$\begin{aligned} \frac{p_1}{p_2} &= \frac{[p_{11}/p_{1\cdot}]}{[p_{21}/p_{2\cdot}]} \\ &= \frac{[p_{1\cdot}p_{\cdot 1}/p_{1\cdot}]}{[p_{2\cdot}p_{\cdot 1}/p_{2\cdot}]} \\ &= \frac{p_{\cdot 1}}{p_{\cdot 1}} \\ &= 1 \end{aligned}$$

- Then, when $X$ and $Y$ are independent (the null), the relative risk is

$$RR = 1.$$

# The Odds Ratio

- In general, if $X$ and $Y$ are not independent, the odds ratio, in terms of $p_1$ and $p_2$, is

$$\begin{aligned} OR &= \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \\ &= \frac{(p_{11}/p_{1\cdot})/(1-(p_{11}/p_{1\cdot})}{(p_{21}/p_{2\cdot})/(1-(p_{21}/p_{2\cdot})} \\ &= \frac{(p_{11}/p_{1\cdot})/(p_{12}/p_{1\cdot})}{(p_{21}/p_{2\cdot})/((p_{22}/p_{2\cdot})} \\ &= \frac{p_{11}p_{22}}{p_{21}p_{12}} \end{aligned}$$

- Similarly, if we instead condition on the columns, as would result from a case-control study,

$$\pi_1 = P[X = 1|Y = 1] = \frac{P[(X = 1), (Y = 1)]}{P[Y = 1]} = \frac{p_{11}}{p_{\cdot 1}}$$

and

$$\pi_2 = P[X = 1|Y = 2] = \frac{P[(X = 1), (Y = 2)]}{P[Y = 2]} = \frac{p_{12}}{p_{\cdot 2}},$$

then

$$\begin{aligned} OR &= \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \\ &= \frac{(p_{11}/p_{\cdot 1})/(1-(p_{11}/p_{\cdot 1}))}{(p_{12}/p_{\cdot 2})/(1-(p_{12}/p_{\cdot 2}))} \\ &= \frac{(p_{11}/p_{\cdot 1})/(p_{21}/p_{\cdot 1})}{(p_{12}/p_{\cdot 2})/((p_{22}/p_{\cdot 2})} \\ &= \frac{p_{11}p_{22}}{p_{21}p_{12}} \end{aligned}$$

- Thus, if we condition on the rows or columns, we get the same odds ratio (as seen in prospective and case-control studies).
- If we do not make the analogy to the prospective or case-control studies, then the odds ratio can be thought of as a 'measure of association' for a cross-sectional, and is sometimes called a 'cross-product ratio', since it is formed from the cross products of the $(2 \times 2)$ table.

$$OR = \frac{p_{11}p_{22}}{p_{21}p_{12}}$$