IV. Controlling for the presence of additional variables

A. Notation:
   1. Add superscript $i$ to tell which table

B. Additional variable provides an alternative explanation for association between disease and exposure: *confounding*
   1. Definition: distortion of disease/exposure association by other factor
      a. Other factor related to exposure
         \[ C \rightarrow D \]
         \[ \downarrow \]
      b. Other factor causally related to disease $E$
   2. Can change direction of relationship: *Simpson’s Paradox* (See example)

A: 2.7.4–2.7.6

3. Rational Define the effect of exposure to be that with everything else held constant

4. Definitions
   a. Split contingency table into separate tables defined by
confounder

b. Separate odds ratios are called \textit{conditional odds ratios}

c. Over-all odds ratio is called \textit{marginal odds ratio}

d. If distribution of exposure and disease are independent in each separate table, they are \textit{conditionally independent} \iff conditional odds ratios are all 1.

e. If conditional odds ratios are all the same, association between disease and exposure is \textit{homogeneous}, even if the common odds ratio is not 1.

5. Example

a. Aspirin is associated with stomach upset

b. Does aspirin cause stomach upset?

c. Alternative explanation: stress causes
   i. stomach upset
   ii. diseases like headaches for which aspirin is likely treatment.

d. Direction of causation not indicated in an observational study

   A: 2.7–2.7.3, 4.3.4

C. Testing whether common odds ratio is 1

1. Use \[ T = \sum_{i=1}^{I} w_i (X_{11}^i - E_{11}^i) \]
a. Intuition might suggest \( w_i = 1/\sqrt{\text{Var} [X_{00}^i | X_{j+}, X_{+k}]} \)

b. We will use \( w_i = 1 \)

c. Use as standard error sum of exact variances.
   i. Implies assumption that tables are independent.

2. Called Mantel–Haenszel test. [Mark B sas] [Mark B R]

A1: 3.2.3

D. Estimation of the common odds ratio

1. Mantel–Haenszel estimator
   \[
   \frac{\sum_{i=1}^I X_{00}^i X_{11}^i / X_{++}^i}{\sum_{i=1}^I X_{10}^i X_{01}^i / X_{++}^i}
   \]
   a. \( \infty \) only if all bottom products are 0

2. logit estimator
   \[
   \hat{\theta} = \exp \left( \frac{\sum_{i=1}^I w_i \log(X_{00}^i X_{11}^i / [X_{10}^i X_{01}^i])}{\sum_{i=1}^I w_i} \right)
   \]
   a. \( w_i = (1/X_{00}^i + 1/X_{01}^i + 1/X_{10}^i + 1/X_{11}^i)^{-1} \)
   b. Omit term \( i \) if \( X_{jk}^i = 0 \) for some \( j, k \)
      i. \( w_i = 0 \)
      ii. Corresponding logit will be \( \infty \)
      iii. Acceptable since \( \lim_{x \to 0} x \log(x) = 0 \)
      iv. Alternative method is to add a bit to zero counts.
   c. This \( w_i \) minimizes variance
Lecture 4

d. SE of $\log(\hat{\theta})$ is $1/\sqrt{\sum_j w_j}$

E. When do you need to stratify?

1. Heuristically: when stratifier is a confounder
   a. That is, it is related to both exposure and disease
   b. Empirically, the odds ratio will change if both row and column proportions differ according to stratifier. [Mark C sas] [Mark C R]

   A: 4.3.5

2. If $\theta$ for the various strata are different, there is an interaction between the confounder and exposure.
   a. Use Breslow and Day statistic to test homogeneity of odds ratio in a series of $I \times 2 \times 2$ tables:

   \[
   \sum_{i,j,k} (X_{k,j}^i - \hat{E}_{k,j}^i)^2 / \hat{E}_{k,j}^i - C \sim \chi^2_{I-1}
   \]

   i. $\hat{E}_{k,j}^i$ satisfy $\hat{E}_{+j}^i = X_{+j}^i \forall j, i$, $\hat{E}_{k,+}^i = X_{k,+}^i \forall k, i$, $(\hat{E}_{11}^i \hat{E}_{00}^i) / (\hat{E}_{10}^i \hat{E}_{01}^i) = \theta \forall i$

   ii. $C = \sum_i (X_{00}^i - \hat{E}_{00}^i)^2 / \sum_i (1/\hat{E}_{00}^i + 1/\hat{E}_{10}^i + 1/\hat{E}_{01}^i + 1/\hat{E}_{11}^i)^{-1}$

   iii. Agresti says that that generally $C$ is small

   iv. SAS appears to ignore $C$.

3. Checking for confounding via hypothesis test
Lecture 5

a. Procedure
   i. test for association betw. $C$ and $D$ and betw. $C$ and $E$,
   ii. adjust if these are significant

b. Uses significance as a proxy for strength of effect

c. To make it work at all, typically make very loose criteria for significance confounders

d. Fails to control Type 1 error [Mark D sas] [Mark D R]