1. If $\theta$ for the various strata are different, there is an interaction between the confounder and exposure.

   a. Use Breslow and Day statistic

   $$\sum_{i,j,k} (X_{k,j}^{i} - \hat{E}_{k,j}^{i})^2 / \hat{E}_{k,j}^{i} - C$$

   i. $\hat{E}_{k,j}^{i}$ satisfy $\hat{E}_{+j}^{i} = X_{+j}^{i} \forall j, i$, $\hat{E}_{k+}^{i} = X_{k+}^{i} \forall k, i$,

   $$(\hat{E}_{11}^{i} \hat{E}_{00}^{i}) / (\hat{E}_{10}^{i} \hat{E}_{01}^{i}) = \hat{\theta} \forall i,$$

   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 $C$.

2. Checking for confounding via hypothesis test

   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

   d. Should not be used for factors that are not confounders

   e. Adjust even if effect mitigated by matching.
G. Extreme case of stratification: Each has two elements

1. AKA matching
   a. Can either be case–control pairs or exposed–unexposed pairs
   b. Exposed-Unexposed
      i. Let $n_{il} =$ number of pairs with unexposed at response level $i$, exposed at response level $l$
         • Pairs with the same response levels for exposed and unexposed are called \textit{concordant}.
         • Pairs with different response levels for exposed and unexposed are called \textit{discordant}.
   c. Case-Control
      i. Let $n_{il} =$ number of pairs with case at exposure level $i$, control at exposure level $l$
         • Pairs with the same exposure levels for case and control are called \textit{concordant}.
         • Pairs with different exposure levels for case and control are called \textit{discordant}.

A: 8–8.1.2, 8.2.5
2. Assumption (exposed–unexposed pairs):
   a. Let $\pi^i_k$ be the probability of event in exposure group $k$ for pair $i$
   b. Assume $\pi^i_1(1 - \pi^i_0)/[\pi^i_0(1 - \pi^i_1)] = \theta \forall i$

3. Use Mantel–Haenszel test
   a. For concordant pairs
      i. Expected values are exactly observed
      ii. Variance is zero
      iii. Hence contribution is zero
   b. For discordant pairs
      i. Expected is all $\frac{1}{2}$
      ii. Obsd-expected is
          - $(1 - \frac{1}{2}) = \frac{1}{2}$ for pairs with + association
          - $(0 - \frac{1}{2}) = -\frac{1}{2}$ for pairs with - association
      iii. Using hypergeometric distribution, null variance contribution for pair is $(1 \times 1 \times 1 \times 1)/(2 \times 2 \times (2 - 1)) = \frac{1}{4}$
      • Total variance is $\frac{1}{4}(n_{10} + n_{01})$.
   c. Test statistic is $(n_{10} - n_{01})/\sqrt{n_{10} + n_{01}}$
      i. same as test that binomial proportion equals $\frac{1}{2}$
      ii. Compare to standard normal
4. Called McNemar’s Test
   a. Test where units are pairs
   b. Each pair has two measurements
   c. Note that this is NOT a test of whether the two pairs agree
5. What should we match on?
   a. Often match on traits that are expected to impact disease
   b. Matching is to remove effect of something associated with both putative cause and effect
   c. Matching can reduce efficiency:
      i. Matching on something correlated to exposure, $E \rightarrow D$
         \[ \downarrow \]
         \[ C \]
         • you get pairs with similar exposure
         • that don’t give much info about effect of exposure on disease
      ii. Matching on an intermediate step in causal chain, $E \rightarrow C \rightarrow D$
         • make exposed more similar to non-exposed.
         • artificially deflate effect of exposure
      iii. Both are known as over-matching
iv. Sometimes matched pairs are multiple observations on one individual.

Se: 9 pp. 285–287

6. Estimation for Matched pairs

a. Pairs have probabilities

\[
\begin{array}{cc}
0 & 1 \\
0 & \psi_{00} \psi_{10} \psi_{00} \psi_{11} \\
1 & \psi_{01} \psi_{10} \psi_{01} \psi_{11}
\end{array}
\]

b. \( n_{01} | n_{10} + n_{10} \sim \text{Bin} \left( \frac{\psi_{00} \psi_{11}}{\psi_{00} \psi_{11} + \psi_{01} \psi_{10}} , n_{10} + n_{01} \right) = \text{Bin} \left( \frac{\theta}{1 + \theta} , n_{10} + n_{01} \right) \) after conditioning on \( n_{10} + n_{01} \).

i. \( \omega = \frac{\theta}{1 + \theta} ; \theta = \frac{\omega}{1 - \omega} \).

c. Hence \( \hat{\theta} = \frac{n_{01}}{n_{10}} \)

d. And get CI for \( \theta \) by getting binomial CI and transforming.

Se: 9 pp. 282–285

7. This is also Mantel–Haenszel estimator

8. Sometimes it is hard to make matched pairs,

a. because collection of subjects doesn’t contain pair

b. or setting up pairs is a lot of work

9. Many models we will employ later will allow us to adjust for confounders without matching.
V. Modeling rates in terms of covariates

A. Before

1. Exposure dichotomous, or categorical with few levels
2. Simple model allowed disease rates to vary from exposure group to exposure group

B. Now

1. want covariate with more levels
   a. Suppose \( L \) covariates
      i. Includes constant 1
      ii. Includes dichotomous “response”, if present.
2. Identify \( K \) relatively homogeneous groups
   a. i.e., same (or similar) values for all covariates
3. Need some structure between rates at different exposure levels
   a. Interpretability
   b. Stability of estimates
4. We will assume linearity on log scale

B&D2: 4.3a

C. Assume that
Lecture 6

1. numbers of events in an interval are Poisson
   a. \( P [X_j = x_j] = \exp(-\lambda_j) \frac{\lambda_j^{x_j}}{x_j!} \)
   b. Implies that each person has chance \( \exp(-\Delta \lambda_j) \) of surviving interval \( \Delta \) without an event.
   c. As before, assume individuals act independently.

2. Log linear model for effect of covariates
   a. Suppose that \( z_{kl} \) is covariate \( l \) in group \( k \)

3. Fit model that says
   \[
   \log(\lambda_k) = + \sum_{l=1}^{L} z_{kl} \beta_l = z_k \beta
   \]
   a. Bold faced quantities are vectors
   b. Multiplication in last expression is inner product.

D. Model is an example of a generalized linear model.
   1. More specifically, Poisson regression