F. 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 *concordant*.
         • Pairs with different response levels for exposed and unexposed are called *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 *concordant*.
         • Pairs with different exposure levels for case and control are called *discordant*.

A: 8–8.1.2, 8.2.5
2. Assumption (exposed–unexposed pairs):
   a. Let $\pi_k^i$ be the probability of event in exposure group $k$ for pair $i$
   b. Assume $\pi_1^i (1 - \pi_0^i) / [\pi_0^i (1 - \pi_1^i)] = \theta \forall i$

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
Lecture 5

4. Called McNemar’s Test
   a. Test where units are pairs
   b. Each pair has two measurements
   c. 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.

A: 2.4.3

6. Estimation for Matched pairs

a. Pairs have probabilities

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

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

i. \[\omega = \theta/(1 + \theta); \theta = \omega/(1 - \omega).\]

c. Hence \[\hat{\theta} = n_{01}/n_{10}\]

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

e. This is also Mantel–Haenszel estimator

7. 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

8. Many models we will employ later will allow us to adjust for confounders without matching.
9. When matched groups are larger than 2
   a. and not necessarily all the same size
   b. still use Mantel-Haenszel procedure
   c. exact binomial results no longer hold
   d. Returns in efficiency from many control matches to a single case diminish

   A: 4.3

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
Lecture 5

1. ie., same (or similar) values for all covariates

3. Need some structure betw. rates at different exposure levels
   a. Interpretability
   b. Stability of estimates

4. We will assume linearity on log scale
   \[ B&D2: \text{4.3a} \]

C. Assume that

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
E. Casting current models in the regression framework

1. One dimension:
   a. $\lambda_k = \exp(\alpha_k)$
   b. $\beta = (\alpha_0, \ldots, \alpha_{K-1})$, $z_k = (0, \ldots, 0, 1, 0, \ldots, 0)$, with the 1 in position $k$.
   c. Model now has one parameter for every observation: saturated
   d. $L(\alpha) = \prod_{k=0}^{K-1} \exp([\omega_k + \alpha_k]X_k - \exp([\omega_k + \alpha_k])/X_k!$
   e. $l(\alpha) = \sum_{k=0}^{K-1} \{\alpha_k + \omega_k\}X_k - \exp(\alpha_k + \omega_k) - \log(X_k!])$
   f. $l^k(\alpha) = X_k - \exp(\alpha_k + \omega_k)$
   g. Maximizer satisfies $\hat{\alpha}_k = \log(X_k) - \omega_k$
   h. For the submodel with all $\alpha$'s equal, $l(\alpha) =$
      $\alpha X_+ + \sum_{k=0}^{K-1} \omega_k X_k - \exp(\alpha) \sum_{k=0}^{K-1} \exp(\omega_k) - \sum_{k=0}^{K-1} \log(X_k!)$
   i. $l'(\alpha) = X_+ - \exp(\alpha) \sum_{k=0}^{K-1} \exp(\omega_k)$
   ii. $\hat{\alpha} = \log(X_+/ \sum_{k=0}^{K-1} \exp(\omega_k))$
   iii. Profile score statistic is $l^k(\hat{\alpha}) = X_k -$
      $X_+ \exp(\omega_k)/ \sum_{k=0}^{K-1} \exp(\omega_k)$
   i. After conditioning on $X_+$,distribution is now multinomial with probabilities
\[ \pi_k = \exp(\omega_k + \alpha_k) / \sum_{m=0}^{K-1} \exp(\omega_m + \alpha_m) \]

ii. Increasing or decreasing all of the \( \alpha_k \) by the same amount gives the same probabilities.

iii. Hence one can not identify all of the \( \alpha_k \).

iv. Pick one of these (ie., \( \alpha_0 = 0 \)), or set sum to zero (PROC CATMOD)

F. Model contains log of time at risk as an offset

1. Fit component is added to every log rate

2. If you know something that rates might be proportional to, log of this could be added to the offset as well

a. For ex, rate in unexposed population by age

\[ \text{Mark C sas} \] \[ \text{Mark C R} \]