### Lecture 3

#### A: 2.4

#### B. Testing association in tables

1. Table entries

<table>
<thead>
<tr>
<th>Exp. cat.</th>
<th>Contr.</th>
<th>Cases</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$X_{00}$</td>
<td>$X_{01}$</td>
<td>$X_{0+}$</td>
</tr>
<tr>
<td>1</td>
<td>$X_{10}$</td>
<td>$X_{11}$</td>
<td>$X_{1+}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$K-1$</td>
<td>$X_{K-10}$</td>
<td>$X_{K-11}$</td>
<td>$X_{K-1+}$</td>
</tr>
<tr>
<td>Total</td>
<td>$X_{+0}$</td>
<td>$X_{+1}$</td>
<td>$X_{++}$</td>
</tr>
</tbody>
</table>

2. $J = K = 2$

   a. Stratified cohort study (ie., condition on row totals).

   i. Estimate $\pi_j$ under $H_0$ as $\hat{\pi}_j = X_{+j}/X_{++}$.

   ii. $Z^2 = T$ for $Z$ the standard normal theory test statistic and $T = \sum_{j,k}(X_{jk} - E_{kj})^2/E_{kj}$.

   iii. Since

   
   $Z^2 = [(\hat{\pi}_{10} - \hat{\pi}_{00}) / \sqrt{\hat{\pi}_{10}\hat{\pi}_{01}/X_{+0} + \hat{\pi}_{01}\hat{\pi}_{10}/X_{+1}}]^2 \sim \chi^2_1$

   
   $T = \sum_{j,k}(X_{jk} - E_{kj})^2/E_{kj}$

   iv. For $E_{kj} = X_{j+}X_{k+}/X_{++}$

   v. For $\nu = (X_{j+}X_{k+}/X_{++})^{-1}X_{j+}^2$

#### 3. Exact Inference for Various Designs

   a. As with approximate analysis,

   i. case–control approach is mathematically equivalent to the stratified cohort approach

   ii. conditionality principal justifies treating the unstratified cohort design as a stratified cohort design.

   b. Cohort inference is generated from distribution of $X_{00} \sim \text{Bin}(\pi_0, X_{0+})$, $X_{10} \sim \text{Bin}(\pi_1, X_{1+})$.

      i. $\pi_0 = P[\text{Control|Unexposed}]$

      ii. $\pi_1 = P[\text{Control|Exposed}]$

      c. $P[X_{00}, X_{10} | X_{0+}, X_{1+}] = P[X_{00}] P[X_{10} | X_{0+}, X_{1+}]$

      d. After rewriting $\pi_1$ in terms of $\pi_0$ and $\theta$, distribution of $T$ still depends on $\pi_0$

      i. $\pi_1 = \pi_0 \theta / (1 - \pi_0 + \pi_0 \theta)$

      ii. Then conditional table probabilities are

      $\frac{X_{++}}{X_{+0}X_{0+}} + \frac{X_{++}}{X_{+1}X_{1+}X_{0+}} + \frac{X_{++}}{X_{+1}X_{1+}X_{1+}}$

      $= \sum E_{kj}^{-1}$

   vi. Working backwards through the above calculations, $\nu$ is inverse of variance of $X_{00} - E_{00}$

   vii. Keep in mind that $E_{00}$ is random.

   viii. Note $(X_{kj} - E_{kj})^2$ is the same for all pairs $i, j$

   ix. Use $\chi^2$ test statistic as before:

   $T = \sum_{j,k=0}^1(X_{kj} - E_{kj})^2/E_{kj}$

   • Expectation satisfies $E_{j+} = X_{j+} + E_{k+} = X_{++}$

   • Under hypothesis for $\theta \neq 1$,

   • $E_{00}E_{11}/(E_{10}E_{01}) = \theta_0$, and $E_{j+} = X_{j+} + E_{k+} = X_{++}$

   • No closed-form solution.

   x. Note that $T$ and reference distribution do not depend on which variable you make rows, and which you make columns.

   xi. Here we approximate discrete distribution by continuous distribution

   • We need the lump sitting at the observed value

      i. On the raw obs scale, the lump has width 1

      ii. Again move upper corner by $1/2$ before calculating $T$

   B&D1: 4.3

b. Unconditional Poisson design

   i. $\nu$ above is same as approximation arising from stratified cohort formulation

---

**Fig. 8: Approximations to the Hypergeometric Distribution**

<table>
<thead>
<tr>
<th>CDF</th>
<th>Row margins are 10, 10 and column margins are 10, 10</th>
</tr>
</thead>
</table>

---

**Table Corner**

<table>
<thead>
<tr>
<th>CDF</th>
<th>Row margins are 20, 10 and column margins are 10, 20</th>
</tr>
</thead>
</table>
4. Testing for \( J, K \)

a. Don't:
   - Distribution of \( c_i \) depends on \( \pi_0 \)
   - Assumes \( \pi_0 \) is fixed
   - Condition also on \( \pi_0 \)
   - Because of multiple comparisons problems.
   - Uses same statistic as before
   - Calculates expected values \( E_{kj} = X_{j+}X_{+k}/X_{++} \)

b. Use similar statistic as before

\[ \text{Hence CI for } \log(\theta) = \log(\hat{\theta}) \pm 1.96 \times \sqrt{\frac{X^{-1}_{00} + X^{-1}_{10} + X^{-1}_{01} + X^{-1}_{11}}{4}} \]

i. One-sided
   - \( H_0 : \theta = \theta_0 \) vs \( H_A : \theta > \theta_0 \)
   - Use \( T = \hat{\theta} \) or equivalently \( X_{00} \)
   - \( p \)-value is sum of probabilities for table with upper left corner at least observed

ii. For two-sided test
   - Order tables according to null probability
   - Implies something other than doubling smaller \( 1 \)-sided \( p \)-value
   - Result is called Fisher's Exact Test

i. Example:

\[
\begin{array}{cccccccc}
0 & 4 & 1 & 3 & 2 & 2 & 3 & 1 & 4 & 0 \\
1 & 4 & 1 & 4 & 1 & 4 & 1 & 4 & 1 & 4 \\
0 & 4 & 1 & 3 & 2 & 2 & 3 & 1 & 4 & 0 \\
2 & 3 & 2 & 3 & 2 & 3 & 2 & 3 & 2 & 3 \\
0 & 4 & 1 & 3 & 2 & 2 & 3 & 1 & 4 & 0 \\
3 & 2 & 3 & 2 & 3 & 2 & 3 & 2 & 3 & 2 \\
0 & 4 & 1 & 3 & 2 & 2 & 3 & 1 & 4 & 0 \\
4 & 1 & 4 & 1 & 4 & 1 & 4 & 1 & 4 & 1 \\
0 & 4 & 1 & 3 & 2 & 2 & 3 & 1 & 4 & 0 \\
5 & 0 & 5 & 0 & 5 & 0 & 5 & 0 & 5 & 0 \\
\end{array}
\]

i. Asymmetry intentional

ii. Tables with same marginals, and probabilities:

\[
\begin{array}{cccccccc}
0 & 4 & 1 & 3 & 2 & 2 & 3 & 1 & 4 & 0 \\
4 & 1 & 0.040 & 3 & 2 & 0.317 & 2 & 3 & 0.476 & 1 & 4 & 0.159 & 0 & 5 & 0.008 \\
\end{array}
\]

iii. Two-sided \( p \)-value: Sum of all \( p \)-values as small or smaller than the table we observe:

\[
0.159 + 0.008 + 0.404 = 0.207
\]

iv. Smaller than double one-sided \( p \)-value, since it avoids adding 32 (0.317)

v. \( p \)-value dominated by probability of observed table.

vi. Cf. \( p \)-value not conditioned on column totals: 30 tables:

<table>
<thead>
<tr>
<th>0</th>
<th>4</th>
<th>1</th>
<th>3</th>
<th>2</th>
<th>2</th>
<th>3</th>
<th>1</th>
<th>4</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

C. Estimation of Odds Ratios:

1. Substitute cell proportions for cell probabilities
2. Recall that conditional and unconditional procedures give the same result.

A: 2.3.1–2.3.2

3. Confidence Bounds for \( \theta \)
   a. Distribution of \( \hat{\theta} \)
      i. \( \hat{\theta} \approx N(\theta, \pi) \)
   ii. For stratified cohort study:
      - \( \log(\pi_0 / (1 - \pi_0)) = \log(X_{01}) - \log(X_{00}) \)
      - Under unknown \( \theta \), stratified cohort sampling,
        \[
        \frac{d}{dx} \log(\text{odds}) = \frac{x_{01} - x_{00} \theta}{x_{00} + x_{01}}
        \]
      - \( \operatorname{Var}[\log(\text{odds})] \approx \left( \text{Var}(X_{01} + x_{01}^{-1})(X_{01}^{-1} + x_{01}^{-1})^{-1} = (X_{01} + x_{01}^{-1}) \right) \)
      - \( \text{DF are same as number of odds ratios one could estimate.} \)

A: 2.3.3

4. Testing for \( J, K \) possibly \( > 2 \)
   a. Don't:
      i. Test pairwise
      ii. Because of multiple comparisons problems.
   b. Use same statistic as before
      i. Calculate expected values \( E_{kj} = X_{j+}X_{+k}/X_{++} \)
5. Exact Confidence intervals \((\theta_L, \theta_U)\) satisfies
\[
\begin{align*}
\Pr_{\theta_L}[X_{00} \geq x_{00}|X_{j^+, X_{+k}}] &= .025, \\
\Pr_{\theta_U}[X_{00} \leq x_{00}|X_{j^+, X_{+k}}] &= .025
\end{align*}
\]