Lecture 3

2.4

- Under hypothesis for $\theta \neq 1$,
  - $E_{00}E_{11}/(E_{10}E_{01}) = \theta_0$, and $E_{j+} = X_{j+}$
  - $E_{+k} = X_{+k}$
  - No closed-form solution.

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

- Likelihood ratio
  - Write down probability for table as function of $\theta$
  - Compare value at 1 to highest value it takes
  - $2 \times \log(L) \sim \chi^2_1$

A: 2.6–2.6.3

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 $X_{00} \sim \text{Bin}(\pi_0, X_{0+})$, $X_{10} \sim \text{Bin}(\pi_1, X_{1+})$.

<table>
<thead>
<tr>
<th>Healthy</th>
<th>Diseased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unexposed</td>
<td>$X_{00}$</td>
</tr>
<tr>
<td>Exposed</td>
<td>$X_{10}$</td>
</tr>
</tbody>
</table>

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

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

2.7

2.8

- $\text{Var}_{\theta=1}[X_{00}X_{j+}, X_{+k}] = \frac{X_{j+}X_{+k}(X_{j+}X_{+k}-1)}{X_{++}(X_{j+}+X_{+k}-1)}$

- Conditioning is not suggested by conditionality principal.
  - $P[\text{disease}] = \pi_0(X_{0+} + X_{+1}\theta/(1 - \pi_0 + \pi_0\theta))$
  - Dependence is weak.

- Here we approximate discrete distribution by continuous distribution
  - Probability of observed outcome must be added to the $p$ value
  - On the raw obs scale, the lump has width 1
  - Again move upper corner by $\frac{1}{2}$ before calculating $T$

- Normal approx. works poorly unless $E_{kj} \geq 5\forall j,k$. See Figure 7.
  - Could have continuity correction described earlier.
  - Choice of $cc$ and variance give 4 possible tests

- Testing $\theta$
  i. One-sided, conditional on all margins:
    - $H_0 : \theta = \theta_0$ vs $H_A : \theta > \theta_0$
    - Use $T = \theta = X_{00} \times X_{11}/(X_{01}X_{10})$ 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

c. $P[X_{00}, X_{10}|X_{0+}, X_{1+}] = (X_{00}^+, X_{1+}^+)\pi_0^X_{00}(1 - \pi_0)\pi_1^X_{1+}(1 - \pi_1)^X_{1+}^{-1}$

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

\[
P[X_{00}, X_{10}|X_{0+}, X_{1+}] = \frac{X_{0+}X_{1+}}{X_{0+}X_{1+} - X_{0+}X_{1+}\theta X_{10}} \times \pi_0(X_{0+}X_{1+} - X_{0+}X_{1+}\theta X_{10})
\]

- Distribution of $T$ still depends on $\pi_0$
  i. $\pi_0$ contributes a constant factor to all tables with same $X_{0+}, X_{1+}$

4. Condition also on $X_{0+}$ and $X_{1+}$ as well as $X_{0+}$ and $X_{1+}$

- removes dependence on $\pi$

- Distribution is called hypergeometric

- If $\theta \neq 1$ called noncentral hypergeometric

- Indicate by $|X_{j+}, X_{+k}$ conditional on $X_{0+}$ and $X_{1+}$ and $X_{01}$ and $X_{10}$

- e. cuts number of tables to be examined.
  i. Both a blessing and a curse.

Fig. 7: Approximations to the Hypergeometric Distribution

- Exact
- Normal Approx.
- Normal Approx. Shifted by Half

Table Corner

Row margins are 10, 10 and column margins are 10, 10
### Asymmetry intentional

1. Assign each of the categories a score
2. Square and rescale makes it
3. Distribution of
4. When
5. Assign each of the columns a score
6. Smaller than double one-sided
7. Smaller than double one-sided

---

### C. Estimation of Odds Ratios:

1. Substitute cell proportions for cell probabilities
2. Recall that conditional and unconditional procedures give
3. Confidence Bounds for
   a. Distribution of \( \theta \)
     i. \( \hat{\theta} \approx N(\theta,?) \)
     ii. For stratified cohort study?
         - \( \log(\pi_0/1-\hat{\pi}_0) = log(X_{01}) - log(X_{00}) \)
         - Under unknown \( \theta \), stratified cohort sampling,
           \( \Delta = (\log(odds) - X_{01}^{-1} + X_{01}^{-1}) \)
   b. Recall that conditional and unconditional procedures give
     the same result.
     i. Estimate of odds ratio compared to this group as
     ii. Calculate odds ratio compared to this group as
     iii. Also can calculate CI
         - Via normal theory and same SE or exactly
     iv. Remember these things are NOT independent
       A: 2.3.1–2.3.2

---

### D. Could also treat ordered row categories

1. Assign each of the categories a score \( u_k \)
   a. By default these are equally spaced
   b. Alternatively, one can use Ridit scores
      \( u_k = \left( \sum_{i<k} X_{i+} + (X_{k+} + 1)/2 \right) / X_{++} \)
     i. Gives Mann–Whitney–Wilcoxon test
     ii. Test statistic has interpretation as estimated
         probability that a random individual from one group
         has a higher score than random individual from the
         other
2. Calculate
   \[ T_j = \sum_{k=0}^{K-1} u_k (X_{kj} - E_{kj}) \]
   a. \( E[T_j] = 0 \).
3. To get variance, need conditional covariances of table entries.
   a. Note
      \[ \text{Var}[X_{kj}] = X_{k+} (X_{++} - X_{kj}) X_{+j} (X_{++} - X_{kj}) / (X_{++}^2 (X_{++} - 1)) \]
   b. For two entries in the same column,
      i. \( \text{Var}[X_{kj} + X_{ij}] = (X_{k+} + X_{i+}) X_{00} X_{+j} (X_{++} - X_{k+} - X_{i+} - X_{ij}) / (X_{++}^2 (X_{++} - 1)) \)
4. Squaring and rescaling makes it
   \( \approx \chi_1^2 \)
   a. Rescaling is done using exact variance
   b. Hence \( \text{Var}[T_j] = \)
   \[ = \left( X_{++} - X_{+j} \right) X_{++} \left\{ \sum_{k=0}^{K-1} u_k^2 X_{k+} X_{++} - \left( \sum_{k=0}^{K-1} u_k X_{k+} X_{++} \right)^2 \right\} \]
   c. Tricks give variances and covariances of \( T_j \)
   d. Properly rescaled, \( S = \sum_j T_j^2 / c_j \sim \chi_1^{2-1} \)
   i. Since \( \sum_j T_j = 0 \), the \( T_j \) are not independent.
   e. \( S \) gives test of \( H_0 \) independence vs. \( H_A \): some
      rows have column probabilities putting more weight on
      higher columns than low rows

---

### E. Could also treat ordered row and column categories

1. Give scores for second dimension as well
3. When \( J = 2 \) or \( K = 2 \),
   a. this is the same as the previous example, with any
      second dimension scores
   b. Called Cochran–Armitage test.
4. Multiple of correlation between row and column scores (1
   for column \( j \), and 0 for all other columns)
5. Assign each of the columns a score \( v_k \)
6. Calculate
   \[ T = \sum_j T_j = \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} v_j u_k (X_{kj} - E_{kj}) \]
7. Multiple of correlation between row and column scores
8. Squaring and rescaling makes it $\approx \chi^2_1$
   a. $T$ gives test of $H_0$: independence vs. $H_A$: higher rows have column probabilities putting more weight on higher columns than low rows
      i. Since $\sum_j T_j = 0$, the $T_j$ are not independent.
      ii. Properly rescaled, $\sum_j T_j^2/c_j \sim \chi^2_{J-1}$ [Mark C sas]

This page intentionally left blank.