b. Distribution is called Fisher’s Exact Test.

c. If \( \theta \neq 1 \) called noncentral hypergeometric.

d. Indicate by \( |X_{1+}X_{+}10|1+X_{10}+X_{0+} \) conditional on \( X_{0+} \) and \( X_{1+} \)

\( X_{10}+X_{0+} \) and \( X_{1+} \).

e. Distribution of \( T \) still depends on \( \pi_0 \).

1. \( \pi_0 \) 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 \( X_{1+} \).

a. Removes dependence on \( \pi \).

b. Distribution is called hypergeometric.

c. If \( \theta \neq 1 \) called noncentral hypergeometric.

d. Indicate by \( |X_{1+}X_{+}10|1+X_{10}+X_{0+} \) conditional on \( X_{0+} \) and \( X_{1+} \)

\( X_{10}+X_{0+} \) and \( X_{1+} \).

e. Cuts number of tables to be examined.

i. Both a blessing and a curse.

---

**Lecture 3**

**Figure 7: Approximations to the Hypergeometric Distribution**

- **CDF**

- **Table Corner**

Row margins are 10, 10 and column margins are 10, 10

---

**Lecture 3**

ii. \( \text{Var}_{\theta=1} X_{00|X_{0+},X_{+}} = \frac{X_{1+}X_{0+}X_{0+}X_{1+}}{X_{++}(X_{+}-1)} \)

iii. Conditioning is not suggested by conditionality principal.

- \( P \) [disease] = \( \pi_0 (X_{0+} + X_{1+} \theta / (1 - \pi_0 + \pi_0 \theta)) \)

- Dependence is weak.

iv. 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 \( T \)

f. 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

---

**Lecture 3**

**Testing \( \theta \)**

- One-sided, conditional on all margins:

  - \( H_0 : \theta = \theta_0 \) vs \( H_A : \theta > \theta_0 \)
  
  - Use \( T = \theta = X_{00} X_{11} / (X_{01} X_{10}) \) or equivalently \( X_{00} \)
  
  - \( p \)-value is sum of probabilities for table with upper left corner at least observed

- 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
Lecture 3 29

D. Could also treat ordered row categories

\[ \log(4) \] Hence CI for

b. For two entries in the same column, Ridit scores

Another, one can use

b. Conditioning on all marginals?

i. No closed form expression for variance

4. Hence CI for \( \log(\theta) \) is

\[ \log(\theta) \pm 1.96 \times \sqrt{\frac{X_{00}^{-1} + X_{10}^{-1} + X_{01}^{-1} + X_{11}^{-1}}{X_{00}X_{11} - X_{01}X_{10}}} \]

5. Exact Confidence intervals \( (\hat{\theta}_L, \hat{\theta}_U) \) satisfies

\[ P_{\theta_L} \left[ X_{00} \geq x_00 | X_{j+}, X_{+,k} \right] = .025, \]

\[ P_{\theta_U} \left[ X_{00} \leq x_00 | X_{j+}, X_{+,k} \right] = .025 \]

\[ A: 2.5 \]

C. Estimation of Odds Ratios:

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

3. Confidence Bounds for \( \theta \)

a. Distribution of \( \hat{\theta} \)

i. \( \hat{\theta} \approx N(\theta, \gamma) \)

ii. For stratified cohort study?

- \( \log(\hat{\pi_0}/[1 - \hat{\pi_0}]) = \log(X_{01}) - \log(X_{00}) \)
- Under unknown \( \theta \), stratified cohort sampling,

\[ \frac{\sum d_i \log(\text{odds})}{\text{Var}[\log(\text{odds})]} = \frac{X_{01}^{-1} + X_{01}^{-1}}{(X_{00}^{-1} + X_{01}^{-1})^{-1}} \]

\[ \text{iv. Bottom row is independent with same structure} \]

\[ \text{v.} \quad \text{Var}[\log(\hat{\theta})] \approx X_{00}^{-1} + X_{10}^{-1} + X_{01}^{-1} + X_{11}^{-1} \]

4. Squaring and rescaling makes it \( \chi^2_1 \)

\[ \text{a. Rescaling is done using exact variance} \]

\[ \text{b. Hence } \text{Var}[T_j] = \]

\[ \text{c. Same tricks as before give } \text{Cov}[X_{kj}, X_{li}] \]

5. \( \text{Cf. } p \)-value not conditioned on column totals: 30 tables:

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

\[ \text{D. Could also treat ordered row categories} \]

1. Assign each of the categories a score \( u_k \)

- By default these are equally spaced

b. Alternatively, one can use Ridit scores

\[ u_k = \frac{\sum_{i<k} X_{ij} + (X_{kj} + 1)/2}{X_{+j}} \]

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 \( \text{Var}[X_{kj}] = X_{++}(X_{--} - X_{-+}X_{+-}/X_{++})/(X_{++}^2(X_{++} - 1)) \)

b. For two entries in the same column,

i. \( \text{Var}[X_{kj} + X_{ij}] = (X_{kj} + X_{ij})X_{00}X_{+j}(X_{++} - X_{kj} - X_{ij} + X_{++})/(X_{++}^2(X_{++} - 1)) \)

4. Multiple of correlation bewt. 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 \( T = \sum_{j=0}^{J-1} T_j = \sum_{k=0}^{K-1} \sum_{j=0}^{J-1} v_j u_k(X_{kj} - E_{kj}) \)

7. Multiple of correlation bewt. row and column scores
8. Squaring and rescaling makes it \( \approx \chi_1^2 \)
   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, \( \sum_j T_j^2/c_j \sim \chi_{J-1}^2 \)

This page intentionally left blank.