A: 2.4

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

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

xi. 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
\[ X_{00} \sim \text{Bin}(\pi_0, X_{0+}), \quad X_{10} \sim \text{Bin}(\pi_1, X_{1+}). \]

<table>
<thead>
<tr>
<th>Unexposed</th>
<th>Healthy</th>
<th>Diseased</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_{00} )</td>
<td>( X_{01} )</td>
<td>( X_{0+} )</td>
</tr>
<tr>
<td>( X_{10} )</td>
<td>( X_{11} )</td>
<td>( X_{1+} )</td>
</tr>
<tr>
<td>( X_{+0} )</td>
<td>( X_{+1} )</td>
<td>( X_{++} )</td>
</tr>
</tbody>
</table>

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

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

c. \[ P[X_{00}, X_{10}|X_{0+}, X_{1+}] = \left( \frac{X_{0+}}{X_{00}} \right) \left( \frac{X_{1+}}{X_{10}} \right) \pi_0 X_{00} (1 - \pi_0) X_{01} \pi_1 X_{10} (1 - \pi_1) X_{11} \]

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

i. \( \pi_1 = \frac{\pi_0 \theta}{1 - \pi_0 + \pi_0 \theta} \)

ii. Then conditional table probabilities are

\[ P[X_{00}, X_{10}|X_{+0}, X_{+1}] = \left( \frac{X_{0+}}{X_{00}} \right) \left( \frac{X_{1+}}{X_{10}} \right) (1 - \pi_1) X_{1+} \]
\[ \times \pi_0 X_{+0} (1 - \pi_0) X_{01} - X_{10} \theta X_{1+} \]
\[ = \left( \frac{X_{0+}}{X_{00}} \right) \left( \frac{X_{1+}}{X_{10}} \right) \left( \frac{1 - \pi_0}{1 - \pi_0 + \pi_0 \theta} \right) X_{1+} \]
\[ \times \pi_0 X_{+0} (1 - \pi_0) X_{+1} - X_{10} \theta X_{1+} \]
\[ = \left( \frac{X_{0+}}{X_{00}} \right) \left( \frac{X_{1+}}{X_{10}} \right) \frac{\pi_0 X_{+0} (1 - \pi_0) X_{+1} \theta X_{10}}{(1 - \pi_0 + \pi_0 \theta) X_{1+}} \]
Lecture 3

e. 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 $X_{+1}$.
   a. removes dependence on $\pi$
   b. Distribution is called hypergeometric
   c. If $\theta \neq 1$ called noncentral hypergeometric
   d. Indicate by $|X_{j+}, X_{+k}|$ conditional on $X_{0+}$ and $X_{1+}$ and $X_{+0}$ and $X_{+1}$.
   e. cuts number of tables to be examined.
      i. Both a blessing and a curse.
      ii. $\text{Var}_{\theta=1} [X_{00}|X_{j+}, X_{+k}] = \frac{X_{+1}X_{0+}X_{+0}X_{1+}}{X_{++}^2(X_{++}-1)}$
      iii. 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.
   iv. Here we approximate discrete distribution by continuous distribution
      - Probability of observed outcome must be added to the $p$ value
Lecture 3

- 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 \( E_{k,j} \geq 5 \forall j,k \). See Figure 7.
- Could have continuity correction described earlier.
  - Choice of cc and variance give 4 possible tests

g. Testing \( \theta \)
- One-sided, conditional on all margins:
  - \( H_0 : \theta = \theta_0 \) vs \( H_A : \theta > \theta_0 \)
  - Use \( T = \hat{\theta} = \frac{X_{00} \times X_{11}}{X_{01} \times X_{10}} \) 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 \( Fisher’s \ Exact Test \)

h. Example:

\[
\begin{array}{cc}
3 & 1 \\
1 & 4 \\
\end{array}
\]
Fig. 7: Approximations to the Hypergeometric Distribution
Lecture 3

i. Asymmetry intentional

ii. Tables with same marginals, and probabilities:

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

iii. Two-sided \( p \)-value: Sum of all table probabilities

as small or smaller than the table we observe:

\[
0.159 + 0.008 + 0.040 = .207
\]

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

\[
\begin{array}{cccc}
1 & 3 \\
3 & 2 & (0.317)
\end{array}
\]

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

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

\[
\begin{array}{cccccccccccccccccccccc}
0 & 4 & 1 & 3 & 2 & 2 & 3 & 1 & 4 & 0 \\
0 & 5 & 0 & 5 & 0 & 5 & 0 & 5 & 0 & 5 \\
0 & 4 & 1 & 3 & 2 & 2 & 3 & 1 & 4 & 0 \\
1 & 4 & 1 & 4 & 1 & 4 & 1 & 4 & 1 & 4 \\
0 & 4 & 1 & 3 & 2 & 2 & 3 & 1 & 4 & 0 \\
2 & 3 & 2 & 3 & 2 & 3 & 2 & 3 & 2 & 3 \\
0 & 4 & 1 & 3 & 2 & 2 & 3 & 1 & 4 & 0 \\
3 & 2 & 3 & 2 & 3 & 2 & 3 & 2 & 3 & 2 \\
0 & 4 & 1 & 3 & 2 & 2 & 3 & 1 & 4 & 0 \\
4 & 1 & 4 & 1 & 4 & 1 & 4 & 1 & 4 & 1
\end{array}
\]
vii. Probabilities depend on common null \( \pi \). [Mark B-sas] [Mark B-R]

i. Estimation of effect
   i. Pick one group as baseline
   ii. Calculate odds ratio compared to this group as before
   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

C. Estimation of Odds Ratios:
   1. Substitute cell proportions for cell probabilities
   2. Recall that conditional and unconditional procedures give the same result.
      A: 2.3.3

3. Confidence Bounds for \( \theta \)
   a. Distribution of \( \hat{\theta} \)?
      i. \( \hat{\theta} \approx \mathcal{N}(\theta, ?) \)
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,

\[
\frac{d}{dX_{00}} \log(\text{odds}) = X_{00}^{-1} + X_{01}^{-1}
\]

iii. \( \text{Var}[\log(\text{odds})] \approx (X_{00}^{-1} + X_{01}^{-1})^2(X_{00}^{-1} + X_{01}^{-1})^{-1} = (X_{00}^{-1} + X_{01}^{-1}) \)

iv. Bottom row is independent with same structure

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

b. Conditioning on all marginals?

i. No closed form expression for variance

4. Hence CI for \( \log(\theta) \) is \( \log(\hat{\theta}) \pm 1.96 \times \sqrt{X_{00}^{-1} + X_{10}^{-1} + X_{01}^{-1} + X_{11}^{-1}} \)

5. Exact Confidence intervals \((\theta_L, \theta_U)\) satisfies

\[
P_{\theta_L} [X_{00} \geq x_{00} | X_{j+}, X_{+k}] = .025, \\
P_{\theta_U} [X_{00} \leq x_{00} | X_{j+}, X_{+k}] = .025
\]

A: 2.5

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 \( T_j = \sum_{k=0}^{K-1} u_k (X_{k:j} - E_{k:j}) \)

a. \( \mathbb{E}[T_j] = 0 \).

3. To get variance, need conditional covariances of table entries.

a. Note \( \text{Var}[X_{k:j}] = X_{k+}(X_{++} - X_{+j})X_{+j}(X_{++} - X_{k+})/(X_{++}^2(X_{++} - 1)) \)

b. For two entries in the same column,

i. \( \text{Var}[X_{k:j} + X_{i:j}] = (X_{k+} + X_{i+})X_{0+}X_{+j}(X_{++} - X_{k+} - X_{i+})/(X_{++}^2(X_{++} - 1)) \)

ii. \( \text{Cov}[X_{k:j}, X_{i:j}] = (\text{Var}[X_{k:j} + X_{i:j}] - \text{Var}[X_{k:j}] - \text{Var}[X_{k:j}])/2 = -X_{k+}X_{i+}[X_{0+}X_{+j}]/(X_{++}^2(X_{++} - 1)) \)

c. Same tricks as before give \( \text{Cov}[X_{k:j}, X_{l:i}] \).

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

a. Rescaling is done using exact variance
b. Hence $\text{Var } [T_j] = \frac{(X_{++} - X_{+j})X_{+j}}{X_{++}(X_{++} - 1)} \left\{ \sum_{k=0}^{K-1} u_k^2 \frac{X_{k+}}{X_{++}} - \left( \sum_{k=0}^{K-1} u_k \frac{X_{k+}}{X_{++}} \right)^2 \right\}$

c. Tricks give variances and covariances of $T_j$

d. Properly rescaled, $S = \Sigma_j T_j^2 / c_j \sim \chi^2_{J-1}$

i. Since $\Sigma_j T_j = 0$, 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 betw. row and column scores (1 for column $j$, and 0 for all other cols):

5. Assign each of the columns a score $v_k$

6. Calculate $T = \Sigma_j T_j = \Sigma_{k=0}^{K-1} \Sigma_{j=0}^{J-1} v_j u_k (X_{k,j} - E_{k,j})$
7. Multiple of correlation between row and column scores

8. Squaring and rescaling makes it $\approx \chi^2$

   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, $\sum_j T_j^2 / c_j \sim \chi^2_{J-1}$