6. Variances and Covariances

a. \( \text{Var} \left[ X_k | X_+ \right] = X_+ \pi_k (1 - \pi_k) \) for \( \pi_k \) equal to the proportion of person years in group \( k \)

b. \( \text{Var} \left[ X_k + X_j | X_+ \right] = X_+ (\pi_k + \pi_j) (1 - \pi_k - \pi_j) \)

c. \( \text{Cov} \left[ X_k, X_j | X_+ \right] = -X_+ \pi_k \pi_j \)

7. Proof that Pearson’s statistic has \( \chi^2_{K-1} \) distribution:

a. Let

i. \( Y \) be the \( K - 1 \) by 1 matrix \( (X_1, \ldots, X_{K-1}) \) (note excluding \( X_0 \)),

ii. \( \nu \) be the \( K - 1 \) by 1 matrix \( (\pi_1, \ldots, \pi_{K-1}) \).

iii. \( \mathbf{1} \) be the \( K - 1 \) by 1 matrix whose entries are all 1.

iv. \( V = \left( \text{diag} \left( \frac{1}{\nu} \right) + \mathbf{1} \mathbf{1}^\top / \pi_0 \right) / X_+ \).

b. Hence \( \text{Var} \left[ Y \right] = X_+ (\text{diag}(\nu) - \nu \nu^\top) \)

c. Then \( V = \text{Var} \left[ Y \right]^{-1} \), as seen by noting that

\[
\text{Var} \left[ Y \right] V = (\text{diag}(\nu) - \nu \nu^\top) (\text{diag}(1/\nu) + \mathbf{1} \mathbf{1}^\top / \pi_0)
= \text{diag}(\nu) \text{diag}(1/\nu) + \text{diag}(\nu) \mathbf{1} \mathbf{1}^\top / \pi_0 -
\nu \nu^\top \text{diag}(1/\nu) - \nu \nu^\top \mathbf{1} \mathbf{1}^\top / \pi_0
= I + \nu \mathbf{1}^\top / \pi_0 - \nu \mathbf{1}^\top - \nu \mathbf{1}^\top (1 - \pi_0) / \pi_0 = I,
\]
Lecture 2

using

i. \( \text{diag}(\nu) \mathbf{1} = \nu \),

ii. \( \nu^\top \text{diag}(1/\nu) = \mathbf{1} \),

iii. \( \nu^\top \mathbf{1} = 1 - \pi_0 \).

d. Recall from 580 or 582 that 
\[
T = (\mathbf{Y} - X_+\nu)^\top V (\mathbf{Y} - X_+\nu) \sim \chi^2_{K-1}
\]

e. Now I show that \( T \) is Pearson’s statistic:
\[
T = (\mathbf{Y} - X_+\nu)^\top V (\mathbf{Y} - X_+\nu)
\]
\[
= (\mathbf{Y} - X_+\nu)^\top (\text{diag}(1/\nu) + \mathbf{1}\mathbf{1}^\top /\pi_0)(\mathbf{Y} - X_+\nu)/X_+
\]
\[
= (\mathbf{Y} - X_+\nu)^\top \text{diag}(1/(X_+\nu))(\mathbf{Y} - X_+\nu) +
\]
\[
(\mathbf{Y} - X_+\nu)^\top \mathbf{1}\mathbf{1}^\top (\mathbf{Y} - X_+\nu)/(X_+\pi_0)
\]

i. First term is \( \sum_{j=1}^{K-1} (Y_i - X_+\pi_i)/(X_+\pi_i) \)

ii. \( \mathbf{1}^\top (\mathbf{Y} - X_+\nu)/(X_+\pi_0) = \sum_{j=1}^{K-1} (Y_i - X_+\pi_i) =
\]
\[
- (X_0 - X_+\pi_0)
\]

iii. Hence second term is \( (X_0 - X_+\pi_0)^2/(X_+\pi_0) \). [Mark A sas]

[Mark A R]

B&D2: 3.4c

8. Ordered exposures:

a. Give scores \( u_k \) to groups (often 1, 2, \ldots )
Lecture 2

b. Add up \( \sum_{k=1}^{K} u_k (X_k - E_k) \)

c. Conditional expectation zero.

d. Need standard error
   i. If independent, variance would be \( \sum_{k=1}^{K} u_k^2 E_k \)
   ii. Here \( \text{Var} [\sum_k u_k X_k | X+] = \sum_k u_k^2 X_+ \pi_k (1 - \pi_k) - \sum_{k \neq j} u_j u_k X_+ \pi_k \pi_j = X_+ [\sum_k u_k^2 \pi_k - \sum_{k,j} u_j u_k \pi_k \pi_j] = \sum_k u_k^2 E_k - (\sum_k u_k E_k)^2 / X_+ \)

e. Sampling distribution approximately standard normal

9. Why not CI?
   a. CI can give test when we have one parameter to test
   b. Here we need \( K - 1 \) parameters
   c. CI becomes confidence region: more complicated.

10. Conditional methods?
    a. Same test statistic
    b. Distribution in cells is given by sequence of binomials
    c. Hard to calculate

11. When \( K = 2 \):
    a. \( X_1 = X_+ - X_0 \) and \( E_1 = X_+ - E_0 \).

[Mark B] [Mark B R]
b. $E_1 = X + \pi$

c. $T = \frac{(X_0 - E_0)^2}{E_0} + \frac{(X_1 - E_1)^2}{E_1} = (X_0 - E_0)^2 \frac{1}{E_0} + \frac{1}{E_1} = (X_0 - E_0)^2 \frac{1}{\pi} + 1 \frac{1}{1 - \pi} = (X_0 - E_0)^2 \frac{1}{\pi (1 - \pi)}$

d. Hence $\chi^2$ statistic is square of $Z$ statistic

e. Hence inference is the same.

A: 2.3

III. Association in tables

A. Some models:

1. $J \times K$ variables $X_{k,j} \sim P(\lambda_{k,j})$
   a. Generally interested in comparisons between rates like $\theta_{k,j} = \frac{\lambda_{k,j} \lambda_{00}}{\lambda_{k0} \lambda_{0j}}$

2. One multinomial with $J \times K$ categories, probabilities $\pi_{k,j} = \frac{\lambda_{k,j}}{\lambda_{++}}$
   a. Get from full Poisson by conditioning on $X_{++}$
   b. Interest parameters are now $\theta_{k,j} = \frac{\pi_{k,j} \pi_{00}}{(\pi_{k0} \pi_{0j})}$
   c. Null hypothesis $\theta_{k,j} = 1 \iff \pi_{k,j} = \frac{\pi_{k0} \pi_{0j}}{\pi_{00}}$
      i. In this case, $\pi_{+j} = \frac{\pi_{+0} \pi_{0j}}{\pi_{00}}$ and $\pi_{k+} = \frac{\pi_{k0} \pi_{0+}}{\pi_{00}}$ and $\pi_{++} = \frac{\pi_{+0} \pi_{0+}}{\pi_{00}}$
ii. In this case, $\pi_{0j} = \pi_{00} \pi_{+j} / \pi_{+0}$ and $\pi_{k0} = \pi_{k+} \pi_{00} / \pi_{0+}$

iii. Hence $\pi_{kj} = \pi_{k+} \pi_{+j} / \pi_{++}$

d. No loss of information

A: 2.1.5

3. $J$ multinomials each with $K$ categories,

   a. Get from full multinomial by conditioning on column totals $X_{+j}$

   b. Probabilities $\psi_{kj} = \pi_{kj} / \pi_{+j}$

   c. Interest parameters are now $\theta_{kj} = \psi_{kj} \psi_{00} / (\psi_{k0} \psi_{0j})$

      i. Quantities $\psi_{1j} / \psi_{0j}$ are odds

      ii. Quantity $\theta_{11}$ is odds ratio

   d. With $\psi_{00}$ fixed, increasing $\theta_{kj}$ is equivalent to pushing $\pi_{kj}$ near 1.

      i. Taking logs,

\[
\log (\psi_{11} / \psi_{01}) = \log (\psi_{10} / \psi_{00}) + \log (\theta).
\]

      ii. Function $\logit(x) = \log(x/(1 - x)) = \log(x) - \log(1 - x)$ is called logistic function.

   e. Factorization of joint probability into marginal and conditional probabilities:
\[
x^{++}! \prod_{k,j} \frac{x_{kj}}{\prod_{k,j} x_{kj}!} = \left[ x^{++}! \prod_{j} \frac{\pi_j}{x^j} \right] \left[ \prod_{j} x^j! \prod_{k} \frac{\psi_{kj}}{\prod_{k} x_{kj}!} \right]
\]

A: 2.3.4

f. Special Case: \( J = K = 2 \), \( \psi_{k1} \) small
   
i. Interpretation: Rows represent exposure to a carcinogen, columns represent healthy and sick.
   
ii. Alternate comparator \( \psi_{11}/\psi_{01} \) relative risk
   
iii. Approximately odds ratio if \( \psi_{10} \) and \( \psi_{00} \approx 1 \).
   
iv. Multiplicative relationship between \( \psi_{11} \) and \( \psi_{01} \) fails when one (and hence both) are not really small.
   
v. Example
      
   • Do a cohort study and find \( \pi_1 = 0.020 \), \( \pi_0 = 0.005 \).
   
   • Hence \( q = 4 \)
   
   • Analyze cohort with some other high–risk factor and get \( \pi_0 = 0.3 \)
   
   • Hence \( \pi_1 = 1.2 \): invalid.
   
4. \( K \) multinomials each with \( J \) categories:
   
a. Condition on row totals \( X_{k+j} \)
   
b. like previous.
c. Conditional probabilities $\kappa_{kj} = \lambda_{kj} / \lambda_{k+}$

d. Note that $\kappa_{kj} \kappa_{00} / (\kappa_{0j} \kappa_{k0}) = \lambda_{kj} \lambda_{00} / (\lambda_{0j} \lambda_{k0}) = \psi_{kj} \psi_{00} / (\psi_{0j} \psi_{k0})$

e. Odds ratios are the same.

5. Distribution conditional on both row and column totals:

a. Testing for $J, K$ possibly $> 2$

i. Don’t:
   • Test pairwise
   • because of multiple comparisons problems.

ii. Use same statistic as before
   • Calculate expected values $E_{kj} = X_{j+} X_{+k} / X_{++}$
   • $T = \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} (X_{kj} - E_{kj})^2 / E_{kj}$.
   • $T \sim \chi^2_{(K-1)(J-1)}$ (approximately)
   ▶ Same requirement of $> 5$ expected

iii. DF are same as number of odds ratios one could estimate.

b. Easy case: $J = K = 2$, $x_{+j} = x_{k+} = 1$ for $j = 1, 0$, $k = 1, 0$.

c. Then

i. Conditional cell probabilities are $\psi_{00}$ $\psi_{01}$ $\psi_{10}$ $\psi_{11}$
Lecture 2

ii. Probabilities satisfy \( \psi_j^+ = 1, (\psi_{11}\psi_{00})/(\psi_{10}\psi_{01}) = \theta_{11} \).

iii. Solution is \( \psi_{10} = \psi_{00}/(-\psi_{00}\theta_{11} + \theta_{11} + \psi_{00}) \).

iv. Then \( P[X_{1+} = 0|X_{+j} = 1] = \psi_{00}\psi_{10} \)

\[
P[X_{1+} = 1|X_{+j} = 1] = \psi_{01}\psi_{10} + \psi_{00}\psi_{11} = \psi_{10}\psi_{01}(1 + \theta_{11})
\]

\[
P[X_{1+} = 2|X_{+j} = 1] = \psi_{11}\psi_{01}
\]

d. Marginal probabilities in terms of \( \theta_{11} \) and \( \psi_{00} \):

\[
\left( \frac{\psi_{00}^2}{\theta_{11} + \psi_{00} - \theta_{11}\psi_{00}}, \frac{(1 + \theta_{11})(1 - \psi_{00})\psi_{00}}{\theta_{11} + \psi_{00} - \theta_{11}\psi_{00}} \right),
\]

\[
(\theta_{11} (1 - \psi_{00})\psi_{00})^2)
\]

\[
\propto (\psi_{00}, (1 + \theta_{11})(1 - \psi_{00}), \frac{\theta_{11}(1 - \psi_{00})}{\theta_{11} + \psi_{00} - \theta_{11}\psi_{00}})
\]

e. Even in easiest case, marginal distribution crucially involves \( \theta_{11} \).

A: 2.3.5–2.3.6

6. Case–Control Study

a. Individuals are distributed into

i. disease groups (columns of table) and

ii. exposure groups (rows of table)

iii. according to Poisson model conditional.

b. Condition on disease status totals
Lecture 2

i. I.e., choose a certain number in each disease status

- Exposure proportions proportional to those in the overall population
- Because disease may be rare enough that sampling based on exposure group would yield almost no diseased individuals

c. Randomness backwards:

i. We want to explore probabilities of disease conditional on exposure \( \kappa_{kj} \)

ii. Natural probabilities for our results are conditional on disease status total \( \psi_{kj} \)

iii. Odds ratios are the same

d. Conditioning on row and column margins makes analysis of case-control study exactly the same as a cohort study.

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>Exp. cat. 0</td>
<td>( X_{00} )</td>
<td>( X_{01} )</td>
<td>( X_{0+} )</td>
</tr>
<tr>
<td>Exp. cat. 1</td>
<td>( X_{10} )</td>
<td>( X_{11} )</td>
<td>( X_{1+} )</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>Exp. cat. ( 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).

**A: 2.2.1**

i. Estimate \( \pi_j \) under \( H_0 \) as \( \hat{\pi}_j = \frac{X+j}{X++} \).

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

iii. since 
\[
Z^2 = \left[ \frac{(\hat{\pi}_{10} - \hat{\pi}_{00})}{\sqrt{\hat{\pi}_{00}\hat{\pi}_{10}/X_{0+} + \hat{\pi}_{00}\hat{\pi}_{10}/X_{1+}}} \right]^2 \sim \chi^2_1
\]
\[
= \frac{X_{++}X_{0+}X_{1+}}{X_{0+}X_{1+}} \left( \frac{X_{00}}{X_{0+}} - \frac{X_{10}}{X_{1+}} \right)^2
\]
\[
= \frac{X_{++}X_{0+}X_{1+}}{X_{0+}X_{1+}} \left( \frac{X_{00}}{X_{++}} \frac{X_{++}}{X_{0+}} - \frac{X_{10}}{X_{++}} \frac{X_{++}}{X_{1+}} \right)^2
\]
\[
= \left( \frac{X_{00} - E_{00}}{X_{++}} \right)^2 v
\]

iv. For \( E_{kj} = \frac{X_j + X_k}{X++} \)

v. For \( v = \left( \frac{X_{+1}X_{0+}X_{+0}X_{1+}}{X_{+0}X_{0+}X_{+1}X_{0+}} \right)^{-1} \frac{X_{++}^3}{X_{++}} \)
\[
= \frac{X_{++}}{X_{+0}X_{0+}} + \frac{X_{++}}{X_{+0}X_{1+}} + \frac{X_{++}}{X_{+1}X_{0+}} + \frac{X_{++}}{X_{+1}X_{1+}}
\]
\[
= \sum E_{kj}^{-1}
\]

vi. Working backwards through the above calculations, \( v \) is inverse of variance of \( X_{00} - E_{00} \)
vii. Keep in mind that $E_{00}$ is random.

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

ix. Use $\chi^2$ test statistic as before: $T = \sum_{j,k=0}^{1} (X_{k,j} - E_{k,j})^2 / E_{k,j}$

- Expectation satisfies $E_{j+} = X_{j+} \quad E_{+k} = X_{+k}, \ (3$ equations, 4 unknowns)