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

ii. \( \mathbf{\nu} \) 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. \( V = \left( \text{diag}(1/\mathbf{\nu}) + \mathbf{11}^\top / \pi_0 \right) / X_+ \).

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

c. Then \( V = \text{Var} \left[ \mathbf{Y} \right]^{-1} \), as seen by noting that
\[
\text{Var} \left[ \mathbf{Y} \right] V = (\text{diag}(\mathbf{\nu}) - \mathbf{\nu \nu}^\top)(\text{diag}(1/\mathbf{\nu}) + \mathbf{11}^\top / \pi_0)
\]
\[
= \text{diag}(\mathbf{\nu}) \text{diag}(1/\mathbf{\nu}) + \text{diag}(\mathbf{\nu}) \mathbf{11}^\top / \pi_0 -
\]
\[
\mathbf{\nu \nu}^\top \text{diag}(1/\mathbf{\nu}) - \mathbf{\nu \nu}^\top \mathbf{11}^\top / \pi_0
\]
\[
= I + \mathbf{\nu}^\top / \pi_0 - \mathbf{\nu}^\top - \mathbf{\nu}^\top (1 - \pi_0) / \pi_0 = I,
\]
Lecture 2

using

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

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

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

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

e. Now I show that \( T \) is Pearson’s statistic:

\[
T = (\mathbf{Y} - X_+\boldsymbol{\nu})^\top V (\mathbf{Y} - X_+\boldsymbol{\nu}) = (\mathbf{Y} - X_+\boldsymbol{\nu})^\top (\text{diag}(1/\boldsymbol{\nu}) + \mathbf{1}\mathbf{1}^\top /\pi_0)(\mathbf{Y} - X_+\boldsymbol{\nu})/X_+
\]

\[
= (\mathbf{Y} - X_+\boldsymbol{\nu})^\top \text{diag}(1/(X_+\boldsymbol{\nu}))(\mathbf{Y} - X_+\boldsymbol{\nu}) + (\mathbf{Y} - X_+\boldsymbol{\nu})^\top \mathbf{1}\mathbf{1}^\top (\mathbf{Y} - X_+\boldsymbol{\nu})/(X_+p\pi_0)
\]

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

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

iii. Hence second term is \( (X_0 - X_+\pi_0)^2/(X_+\pi_0) \).

8. Ordered exposures; see Fig. 7.

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

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 \( \text{Var} \left[ \sum_k u_k X_k | X_+ \right] = \sum_k u_k^2 X_+ \pi_k (1 - \pi_k) - \sum_{k \neq j} u_j u_k X_+ \pi_k \pi_j = X_+ \left[ \sum_k u_k^2 \pi_k - \sum_{k,j} u_j u_k \pi_k \pi_j \right] = \sum_k u_k^2 E_k - (\sum_k u_k E_k)^2 / X_+ \)

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. Exact 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 \( E_1 = X_+ - E_0 \).

b. \( E_1 = X_+ \pi \)

c. \( T = (X_0 - E_0)^2 / E_0 + (X_1 - E_1)^2 / E_1 = (X_0 - E_0)^2 [1/E_0 + 1/E_1] = (X_0 - E_0)^2 X_+^{-1} [1/\pi + 1/(1 - \pi)] = (X_0 - E_0)^2 X_+^{-1} / (\pi(1 - \pi)) \)
d. Hence \( \chi^2 \) statistic is square of \( Z \) statistic
III. Association in tables

A. Some models:

1. $J \times K$ variables $X_{k,j} \sim \mathcal{P}(\lambda_{k,j})$
   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, $\pi_{0j} = \frac{\pi_{00}\pi_{+j}}{\pi_{+0}}$ and $\pi_{k0} = \frac{\pi_{k+}\pi_{00}}{\pi_{0+}}$
      iii. Hence $\pi_{k,j} = \frac{\pi_{k+}\pi_{+j}}{\pi_{++}}$
   d. No loss of information

A: 2.1.5

3. $J$ multinomials each with $K$ categories,
Lecture 2

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,

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

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:

\[
x_{++!} \prod_{kj}^{x_{kj}} \pi_{kj}^{x_{kj}} = \left[ x_{++!} \prod_{j}^{x_{+j}} \pi_{+j}^{x_{+j}} \right] \left[ \prod_{j}^{x_{+j}} \psi_{kj}^{x_{kj}} \right]
\]

\[ \text{A: 2.3.4} \]

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 $\pi_1 = .020$, $\pi_0 = .005$.
- Hence $q = 4$
- Analyze cohort with some other high–risk factor and get $\pi_0 = .3$
- Hence $\pi_1 = 1.2$: invalid.

4. $K$ multinomials each with $J$ categories:

a. Condition on row totals $X_{k+}$

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. Easy case: \( J = K = 2 \), \( x_{+j} = x_{k+} = 1 \) for \( j = 1, 0 \), \( k = 1, 0 \).

b. Then

i. Conditional cell probabilities are \( \psi_{00} \psi_{01} \psi_{10} \psi_{11} \)

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 \( P[X_{1+} = 0|X_{+j} = 1] = \psi_{00}\psi_{10} \)
\( P[X_{1+} = 1|X_{+j} = 1] = \psi_{01}\psi_{10} + \psi_{00}\psi_{11} = \psi_{10}\psi_{01}(1+\theta_{11}) \)
\( P[X_{1+} = 2|X_{+j} = 1] = \psi_{11}\psi_{01} \)

c. Marginal probabilities in terms of \( \theta_{11} \) and \( \psi_{00} \):
\[
\begin{align*}
\psi_{00}^2 &\quad (1 + \theta_{11})(1 - \psi_{00}) \psi_{00} \\
\theta_{11} + \psi_{00} - \theta_{11} \psi_{00} &\quad \theta_{11} + \psi_{00} - \theta_{11} \psi_{00} \\
(\theta_{11} + \psi_{00} - \theta_{11} \psi_{00})^2 &\quad (\theta_{11} + \psi_{00} - \theta_{11} \psi_{00})^2 \\
\end{align*}
\]
\( \propto (\psi_{00}, (1 + \theta_{11})(1 - \psi_{00}), \frac{\theta_{11}(1 - \psi_{00})}{\theta_{11} + \psi_{00} - \theta_{11} \psi_{00}}) \)

d. 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
Lecture 3

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

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 $\kappa_{k,j}$

ii. Natural probabilities for our results are conditional on disease status total $\psi_{k,j}$

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.