Week 2: Stratified Count Data

Video 12: Notation and Testing

EPI202 – Epidemiologic Methods II
Murray A. Mittleman, MD, DrPH
Department of Epidemiology, Harvard TH Chan School of Public Health

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Key Concepts

- Notation
- Hypothesis tests
- Point and interval estimates for the cumulative incidence ratio
- Point and interval estimates for the cumulative incidence difference

Notation for Count Data
(closed cohort and cross-sectional studies)

- Recall our notation for an unstratified table of count data in closed cohort or cross-sectional studies:

Notation

	E	$\bar{E}$	
Cases	a	b	M_1
Non-cases	c	d	M_0
	N_1	N_0	T

Notation for Count Data (closed cohort and cross-sectional studies)

- Now we stratify the data by one or more confounding variables, so that each stratum consists of subjects who have, on average, the same risk for disease, with the possible exception of the exposure effect.

- We have i=1,...,I of these strata, which are formed by each unique combination of levels of the confounding variables for which there are data.

Notation

	E	$\bar{E}$	
Cases	a_i	b_i	M_{1i}
Non-cases	c_i	d_i	M_{0i}
	N_{1i}	N_{0i}	T_i

Evans County Study Revisited

- 609 white men free of disease at baseline, are followed to investigate the association between the 7-year cumulative incidence of coronary heart disease (CHD) and endogenous catecholamines (CAT).
- We stratify by two potential confounders
  - □ age group (<55 / 55+)
  - □ electrocardiogram status (N=normal / A=abnormal).

	Age<55 ECG=N		Age<55 ECG=A		Age 55+ ECG=N		Age 55+ ECG=A	
Cat Level	**Hi**	**Low**	**Hi**	**Low**	**Hi**	**Low**	**Hi**	**Low**
Cases	1	17	3	7	9	15	14	5
Non-cases	7	257	14	52	30	107	44	27
Total	8	274	17	59	39	122	58	32
CI	12.5	6.2	17.6	11.9	23.1	12.3	24.1	15.6

Hypothesis Test for Unstratified Data

- Recall the hypothesis test statistic:

$$Z^2 = \frac{[X - E(X|H_0)]^2}{Var(X|H_0)}$$

- In closed cohorts (count data) with no confounding,
  - □ X = number of exposed cases = a
  - □ E(X|H₀) = number of exposed cases expected under H₀

    = total number of cases * Pr(E) = $M_1(N_1/T)$
  - □ $$Var(X|H_0) = \frac{M_1 M_0 N_1 N_0}{T^3}$$

Hypothesis Test for Stratified Data (1)

- In closed cohorts or cross-sectional studies (count data) with confounding, we stratify the data on all confounding variables to form I strata
- We then calculate the test statistic:

$$Z^2 = \frac{\left[\sum_{i=1}^{I} X_i - \sum_{i=1}^{I} E_i\left(X_i | H_0\right)\right]^2}{\sum_{i=1}^{I} \mathrm{Var}_i\left(X_i | H_0\right)} \sim \chi_1^2$$

Hypothesis Test for Stratified Data (2)

- In stratified analysis of closed cohort studies, as in case-control studies, the Mantel-Haenszel χ2 test statistic is used. This statistic has the formula:

$$Z^2 = \frac{\left[\sum_{i=1}^{I} a_i - \sum_{i=1}^{I} \frac{M_{1i} N_{1i}}{T_i} \right]^2}{\sum_{i=1}^{I} \frac{M_{1i} M_{0i} N_{1i} N_{0i}}{T_i^3}}$$

- Z^2 is distributed χ^2 with one degree of freedom

Hypothesis Test for Stratified Data
Null and Alternative Hypothesis

- H_0: There is no association between CAT level and CHD risk after stratifying by age and ECG status.
  - □ H_0: $CIR_{MH} = 1$ ⟷ $CID_{Summ} = 0$
- H_A: There is an association between CAT level and CHD risk after stratifying by age and ECG status.
  - □ H_A: $CIR_{MH} \neq 1$ ⟷ $CID_{Summ} \neq 0$

Evans County Study

Test Statistic (1)

$$Z^2 = \frac{\left[\sum_{i=1}^{I} X_i - \sum_{i=1}^{I} E_i\left(X_i \middle| H_0\right)\right]^2}{\sum_{i=1}^{I} Var_i\left(X_i \middle| H_0\right)}$$

$$\sum X_i = \sum a_i = 1 + 3 + 9 + 14 = 27$$

$$\sum E_i\left(X_i \middle| H_0\right) = \sum \left(\frac{M_{1i} N_{1i}}{T_i}\right) = \frac{18*8}{282} + \frac{10*17}{76} + \frac{24*39}{161} + \frac{19*58}{90} = 20.81$$

$$\sum_{i=1}^{I} \frac{M_{1i} M_{0i} N_{1i} N_{0i}}{T_i^3} = \frac{18*264*8*274}{282^3} + \frac{10*66*17*59}{76^3} + \frac{24*137*39*122}{161^3} + \frac{19*71*58*32}{90^3} = 9.16$$

Evans County Study

Test Statistic (2)

- Thus, the test statistic is:

$$Z^2 = \frac{\left[\sum_{i=1}^{I} X_i - \sum_{i=1}^{I} E_i\left(X_i \middle| H_0\right)\right]^2}{\sum_{i=1}^{I} \mathrm{Var}_i\left(X_i \middle| H_0\right)} = \frac{[27 - 20.81]^2}{9.16} = 4.18$$

Evans County Study
P-Value

- $\Pr[\chi_1^2 > 4.18] = 0.04$

- After conditioning on age and ECG status we still reject the null hypothesis at the 2-sided alpha of 0.05 and conclude that there is evidence of a statistically significant association between CHD risk and CAT level in these data (assuming no confounding by other variables, no residual confounding by age and ECG status, no selection bias, no information bias and no other source of bias).

- Recall that the crude test statistic had a value of 16.25, corresponding to a p-value of 0.00006. Due to confounding by age, ECG status or both in these data, the crude p-value computed is not valid, and the crude analysis greatly exaggerated the evidence against the null.

BREAK

Week 2: Stratified Count Data

Video 13: Estimation of Ratio Measures

EPI202 – Epidemiologic Methods II
Murray A. Mittleman, MD, DrPH
Department of Epidemiology, Harvard TH Chan School of Public Health

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Key Concepts

- Notation
- Hypothesis tests
- Point and interval estimates for the cumulative incidence ratio
- Point and interval estimates for the cumulative incidence difference

Cumulative Incidence Ratio
Point and Interval Estimates

- To calculate the estimate of the summary cumulative incidence ratio, we use a weighted sum of the stratum-specific estimates:

$$\hat{CIR} = \sum_{i=1}^{I} w_i \, \hat{CIR}_i = \sum_{i=1}^{I} \frac{w_i'}{\sum_{i=1}^{I} w_i'} \hat{CIR}_i = \frac{\sum_{i=1}^{I} w_i' \, \hat{CIR}_i}{\sum_{i=1}^{I} w_i'}$$

- Where:
  - □ w_i' = the weight for the estimated cumulative incidence ratio in stratum I
  - □ CIR_i = the estimated rate ratio in stratum i

$$Note: \quad w_i = \frac{w_i'}{\sum w_i'} \quad and \quad \sum w_i = 1$$

Choices of Weights
Mantel-Haenszel Weights

- To make the best use of our data, we use the Mantel-Haenszel weights:

$$w_i{}' = \frac{b_i\, N_{1i}}{T_i}$$

- Using the Mantel-Haenszel weights, the summary cumulative incidence ratio is:

$$\hat{CIR}_{MH} = \frac{\sum_{i=1}^{I} w_i{}'\, \hat{CIR}_i}{\sum_{i=1}^{I} w_i{}'} = \frac{\sum_{i=1}^{I} \frac{b_i\, N_{1i}}{T_i} \left(\frac{a_i}{N_{1i}}\right) \Big/ \left(\frac{b_i}{N_{0i}}\right)}{\sum_{i=1}^{I} \frac{b_i\, N_{1i}}{T_i}} = \frac{\sum_{i=1}^{I} \frac{a_i\, N_{0i}}{T_i}}{\sum_{i=1}^{I} \frac{b_i\, N_{1i}}{T_i}}$$

Evans County Study
CIR_{MH} Estimation

- The Mantel-Haenszel summary cumulative incidence ratio is:

$$\hat{\text{CIR}}_{\text{MH}} = \frac{\frac{1*274}{282} + \frac{3*59}{76} + \frac{9*122}{161} + \frac{14*32}{90}}{\frac{17*8}{282} + \frac{7*17}{76} + \frac{15*39}{161} + \frac{5*58}{90}} = \frac{15.10}{8.90} = 1.70$$

- After adjusting for confounding by age and ECG status, the estimated risk ratio changed from 2.45 to 1.70. Confounding by either or both of these variables led to an overestimation of the risk ratio.

- Assuming there is no confounding by other variables, no residual confounding by these variables, no selection bias and no information bias, these data indicate that men with high CAT levels have a 70% higher 7-year cumulative incidence of CHD than men with low CAT levels.

95% CI for *ln*(CIR$_{MH}$)

To construct the 95% confidence interval for *ln*(CIR$_{MH}$), we use the usual formula:

$$ln\,\hat{\text{CIR}}_{\text{MH}} \pm 1.96\sqrt{\text{Var}\left(ln\,\hat{\text{CIR}}_{\text{MH}}\right)}$$

Where

$$\text{Var}\left(ln\,\hat{\text{CIR}}_{\text{MH}}\right) = \frac{\displaystyle\sum_{i=1}^{I} \frac{M_{1i}N_{1i}N_{0i} - a_i b_i T_i}{T_i^2}}{\left[\displaystyle\sum_{i=1}^{I} \frac{a_i N_{0i}}{T_i}\right]\left[\displaystyle\sum_{i=1}^{I} \frac{b_i N_{1i}}{T_i}\right]}$$

Evans County Study

CIR_{MH} Confidence Interval Estimation

- In this example, $\hat{\text{V}}\text{ar}\,(ln\,\text{C}\hat{\text{I}}\text{R}) = 0.067$
- Thus, the 95% confidence interval for ln (CIR_{MH}) is:

$$ln\,(1.70) \pm 1.96\,\sqrt{0.067} = 0.528 \pm 1.96\,\sqrt{0.067} = (0.020, 1.036)$$

- and the 95% confidence interval for CIR_{MH} is:

$$e^{(0.020,\ 1.036)} = (1.02,\ 2.82)$$

Evans County Study
CIR_{MH} Confidence Interval Interpretation

- Assuming no residual confounding by age or ECG status, no confounding by other variables, no selection bias, no information bias or any other source of bias, we conclude that there is evidence of a significant elevation in the 7-year cumulative incidence of CHD associated with high CAT level compared with lower levels, ranging in magnitude from 2% to nearly a three-fold higher cumulative incidence with 95% confidence.

- The crude 7-year cumulative incidence ratio 2.45 (1.58, 3.79) was incorrectly centered and far too narrow due to confounding by age and/or ECG status.

BREAK

Week 2: Stratified Count Data

Video 14: Estimation of Difference Measures

EPI202 – Epidemiologic Methods II
Murray A. Mittleman, MD, DrPH
Department of Epidemiology, Harvard TH Chan School of Public Health

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Key Concepts

- Notation
- Hypothesis tests
- Point and interval estimates for the cumulative incidence ratio
- Point and interval estimates for the cumulative incidence difference

Cumulative Incidence Difference
Point and Interval Estimates

- To calculate the estimated summary cumulative incidence difference, we use a weighted average of the stratum-specific differences:

$$C\hat{I}D = \sum_{i=1}^{I} w_i * C\hat{I}D_i = \sum_{i=1}^{I} \frac{w_i'}{\sum_{i=1}^{I} w_i'} C\hat{I}D_i = \frac{\sum_{i=1}^{I} w_i' \, C\hat{I}D_i}{\sum_{i=1}^{I} w_i'}$$

- Where:
  - □ w_i' = the weight for the estimated cumulative incidence difference in the ith stratum
  - □ $C\hat{I}D_i$ = the estimated cumulative incidence difference in the ith stratum

Choice of Weights
Summary Cumulative Incidence Difference

- As shown in the roadmap, the weights for the summary cumulative incidence difference is $w_i = \frac{N_{1i}N_{0i}}{T_i}$.

- The computational form of the summary cumulative incidence difference is $\frac{\sum\left(\frac{a_i N_{0i} - b_i N_{1i}}{T_i}\right)}{\sum \frac{N_{1i}N_{0i}}{T_i}}$.

Evans County Study

$\text{CID}_{\text{summary}}$ Estimation

- The summary cumulative incidence difference:

$$C\hat{I}D_{Summary}\frac{\sum\left(\frac{a_iN_{0i}-b_iN_{1i}}{T_i}\right)}{\sum\frac{N_{1i}N_{0i}}{T_i}}$$

$$=\frac{\frac{1*274-17*8}{282}+\frac{3*59-7*17}{76}+\frac{9*122-15*39}{161}+\frac{14*32-5*58}{90}}{\frac{8*274}{282}+\frac{17*59}{76}+\frac{39*122}{161}+\frac{58*32}{90}}$$

= 0.0871

- Assuming no confounding by additional variables, no residual confounding by age and ECG status, no selection bias, no information bias or any other source of bias, these data indicate that men with high CAT levels have an 8.7% higher cumulative incidence of CHD than men with low CAT levels at baseline.

- Recall that the crude CID was 13.1%. Adjustment for age and ECG status led to a considerable reduction in the estimated cumulative incidence difference by (8.5 - 13.1)/8.5 = 54%.

Evans County Study

$\text{CID}_{\text{inv var}}$ Estimation

- The summary cumulative incidence difference:

$$\hat{\text{CID}} = \frac{\sum_{i=1}^{I} w_i{}' \, \hat{\text{CID}}_i}{\sum_{i=1}^{I} w_i{}'}$$

$$= \frac{(72.02)\,(0.063) + (96.89)\,(0.058) + (183.97)\,(0.108) + (137.42)\,(0.085)}{72.02 + 96.89 + 183.97 + 137.42} = 0.085$$

- Assuming no confounding by additional variables, no residual confounding by age and ECG status, no selection bias, and no information bias, these data indicate that men with high CAT levels have an 8.5% higher cumulative incidence of CHD than men with low CAT levels.

- Recall that the crude CID was estimated to be 13.1%. Adjustment for age and ECG status led to a reduction in the estimated cumulative incidence difference by (8.5 - 13.1)/8.5 = 54%.

95% CI for CID$_{\text{summary}}$

The 95% confidence interval for the summary incidence rate difference is calculated in the usual way:

$$C\hat{I}D_{summary} \pm 1.96\sqrt{var\left(C\hat{I}D_{summary}\right)}$$

Where the variance of the summary cumulative incidence difference is:

$$var\left(C\hat{I}D_{summary}\right) = \frac{\sum\left(\frac{a_i c_i N_{0i}^2}{T_i^2\left(N_{1i}-1\right)} + \frac{b_i d_i N_{1i}^2}{T_i^2\left(N_{0i}-1\right)}\right)}{\left(\sum\frac{N_{1i}N_{0i}}{T_i}\right)^2}$$

Evans County Study

$\text{CID}_{\text{summary}}$ CI Estimation and Interpretation

- In this example using statistical software, we find that the variance of the $\hat{\text{C}}\hat{\text{I}}\text{D}$ is:

$$\widehat{\text{V}}\text{ar}\ (\text{C}\hat{\text{I}}\text{D}) = 0.0020811$$

- Thus, the 95% confidence interval for CID is:

$$0.0871\ \pm\ 1.96\sqrt{0.0020811} = (\,-0.0023, 0.176\,)$$

- After adjusting for age and ECG status, the 7-year cumulative incidence of CHD is consistent with a slight deficit in risk to an absolute excess of 17.6% with 95% confidence (assuming no residual confounding by age or ECG status, no confounding by other factors and no other sources of bias).

Evans County Study
Crude Vs. Stratified CID Estimates

- Recall that the confidence interval for the crude CID, (5.3%, 20.9%), was incorrectly centered and much too narrow.

- Note that the p-value from the hypothesis test indicated that there was a statistically significant association between CAT level at baseline and the 7-year cumulative incidence of CHD. Furthermore, the 95% CI for the CIR excluded the null value of 1. However, the 95% CI for the cumulative incidence difference included the null value of 0.

- This can occur when the p-value is close to the threshold cut-point. In this instance, there is a significant association, but the estimation procedure for the difference parameter is less statistically efficient than for the ratio parameter.

Stratified Count Data

- Note that all formulas and strategies discussed in this lecture apply without modification for testing the null hypothesis and for the analysis of difference and ratio measures when analyzing prevalence data in cross-sectional studies.

BREAK