BIOS 662 Fall 2018

Rates and Proportions

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

https://sakai.unc.edu/portal

Outline

- Prevalence/incidence
- Direct standardization
- Indirect standardization

Rates and Proportions

- Cf. chapter 15 of the text
- *Prevalence*: Proportion (π) of people with a particular disease at a fixed point in time
- *Rate*: Change in a variable over a specified time interval divided by the length of the time interval
- *Incidence*: The number of new cases of a disease in a period of time divided by the person-years at risk
- Incidence is a rate, prevalence is not

Prevalence

- Consider a random sample of size N from the population of interest
- Suppose n have the disease of interest ("cases")
- Estimator of prevalence:

$$\hat{p} = \frac{n}{N} = \frac{\text{number of cases}}{\text{sample size}}$$

- CIs and tests for prevalence are based on

$$n \sim \text{Binomial}(N, \pi)$$

where π is the population prevalence

Prevalence: Example

- A random sample of 1,717 injecting drug users in 6 major cities in the U.S. found that 206 were HIV positive.
- Estimated prevalence of HIV among injecting drug users

$$\hat{p} = 206/1717 = 0.120$$

- Large sample 95 % CI: (0.105, 0.135)

Incidence

- Estimator of incidence:

$$\hat{I} = \frac{\text{number of new cases}}{(\text{sample size}) \times (\text{time interval})}$$

(This is a simplified version; if we use the definition of incidence strictly, we should exclude time after a person has developed the disease)

- Example: Incidence of diabetes among Pima Indians

- $N = 1,728$, time $= 6$ years, new cases $= 346$

[Reference: *AJE* Oct 1, 2003, page 669]

$$\hat{I} = \frac{346}{1728 \times 6} = 0.033$$

- Thus estimated incidence is 0.033 cases per person-year

Incidence

- We usually multiply by some number, such as 1,000

$$\hat{I}_{1000} = 33.4$$

- Interpretation: estimated incidence is:

  33.4 cases per year per 1,000 persons

  or

  33.4 cases per 1,000 person years

Incidence

- Note general form

$$\hat{I}_{1000} = c \cdot \frac{\text{number of new cases}}{\text{sample size}}$$

where $c = 1000/\text{time interval}$

- Because

$$\frac{\text{number of new cases}}{\text{sample size}}$$

is a proportion, we can again use binomial principles for CIs and tests

Incidence

- Let

$$\hat{p} = \frac{n}{N} = \frac{\text{number of new cases}}{\text{sample size}}$$

so that

$$n \sim \text{Binomial}(N, \pi)$$

- Note that the π here is distinct from in the prevalence situation; now it is the probability of becoming a case in the follow-up interval

- Thus

$$\widehat{\text{Var}}(\hat{p}) = \hat{p}(1 - \hat{p})/N$$

implying

$$\widehat{\text{Var}}(\hat{I}_{1000}) = c^2 \hat{p}(1 - \hat{p})/N$$

Incidence CI

- Approximate $100(1-\alpha)\%$ CI

$$\hat{I}_{1000} \pm z_{1-\alpha/2}\sqrt{c^2\hat{p}(1-\hat{p})/N}$$

- Diabetes example:

$$\hat{p} = \frac{346}{1728} = 0.20; \quad c = \frac{1000}{6} = 166.67$$

- 95% CI

$$33.4 \pm 3.14 = (30.2, 36.5)$$

Direct Standardization

- We may need to adjust rates/proportions for possible confounders, e.g., age, gender
- Example: Study of smoking in China (1984)

  Urban women: 1,320 questioned, 330 current smokers

  Rural women: 1,338 questioned, 414 current smokers

$$\hat{p}_{\mathrm{u}} = 330/1320 = 0.25; \quad \hat{p}_{\mathrm{r}} = 414/1338 = 0.31$$

- Concern: Age may be a confounder

Direct Standardization

- Three steps
  1. Divide samples into K categories of the potential confounder
  2. Compute the proportion or rate in each confounder category
  3. Compute the weighted average of confounder-specific proportions/rates
- Choice of weights is based on a *standard or reference population*; e.g., aggregate of samples in hand, governmental population survey

Direct Standardization

- China smoking example

	Urban			Rural		
Age	N_{1k}	n_{1k}	$\hat{p}_{1k}$	N_{2k}	n_{2k}	$\hat{p}_{2k}$
35-39	129	8	0.062	387	44	0.114
40-44	243	53	0.218	441	138	0.313
45-49	478	135	0.282	300	130	0.433
50-54	470	134	0.285	210	102	0.486

Direct Standardization

- Combined age distribution

Age	N_k	w_k
35-39	516	0.194
40-44	684	0.257
45-49	778	0.293
50-54	680	0.256
Total	2658	1.000

Direct Standardization

- Adjusted prevalence estimator

$$\hat{p}_{j_{\text{adj}}} = \frac{\sum_{k=1}^{K} w_k \, \hat{p}_{jk}}{\sum_{k=1}^{K} w_k}$$

- Estimator of prevalence in the reference (i.e., standard) population is based on the observed rates from the study population

Direct Standardization

- Example: (Urban=1, Rural=2)

$$\hat{p}_{1_{\text{adj}}} = (0.194 \times 0.062 + \cdots + 0.256 \times 0.285)/1 = 0.224$$

$$\hat{p}_{2_{\text{adj}}} = 0.354$$

- Crude difference, ratio:

$$\hat{p}_1 - \hat{p}_2 = 0.25 - 0.31 = -0.06$$

$$\hat{p}_2/\hat{p}_1 = 0.31/0.25 = 1.24$$

- Age adjusted difference, ratio:

$$\hat{p}_{1_{\text{adj}}} - \hat{p}_{2_{\text{adj}}} = 0.224 - 0.354 = -0.13$$

$$\hat{p}_{2_{\text{adj}}}/\hat{p}_{1_{\text{adj}}} = 0.354/0.224 = 1.58$$

Direct Standardization

- World Health Organization Standard Weights

Age	w_i	Age	w_i
<1	2.4	45-49	6
1-4	9.6	50-54	5
5-9	10	55-59	4
10-14	9	60-64	4
15-19	9	65-69	3
20-24	8	70-74	2
25-29	8	75-79	1
30-34	6	80-84	0.5
35-39	6	>84	0.5
40-44	6		

Direct Standardization

- China-smoking example using WHO standard:

$$\hat{p}_{1_{\text{adj}}} = \frac{6 \times 0.062 + \cdots + 5 \times 0.285}{6 + 6 + 6 + 5} = 0.209$$

$$\hat{p}_{2_{\text{adj}}} = 0.330$$

$$\hat{p}_{1_{\text{adj}}} - \hat{p}_{2_{\text{adj}}} = -0.121$$

$$\frac{\hat{p}_{2_{\text{adj}}}}{\hat{p}_{1_{\text{adj}}}} = 1.58$$

Direct Standardization

	Crude	Combined	WHO
Difference	-0.06	-0.13	-0.12
Ratio	1.24	1.58	1.58

- Note: Combined and WHO estimates are further from null than the crude estimates
- The confounder, age, partially masks difference in smoking between urban and rural
- Intuition: Rural, older people smoke more; urban sample has greater proportion of older people

Direct Standardization

- $\hat{p}_{j_{\text{adj}}}$ is a weighted average of independent random variables (the $\hat{p}_{jk}$)
- Because $n_{jk} \sim \text{Binomial}(N_{jk}, \pi_{jk})$, we know that

$$\text{Var}(\hat{p}_{jk}) = \pi_{jk}(1 - \pi_{jk})/N_{jk}$$

and

$$\widehat{\text{Var}}(\hat{p}_{jk}) = \hat{p}_{jk}(1 - \hat{p}_{jk})/N_{jk}$$

Direct Standardization

- Thus

$$\widehat{\text{Var}}(\hat{p}_{1_{\text{adj}}} - \hat{p}_{2_{\text{adj}}}) = \frac{\sum_{k=1}^{K} w_k^2 \big(\widehat{\text{Var}}(\hat{p}_{1k}) + \widehat{\text{Var}}(\hat{p}_{2k})\big)}{\big(\sum_{k=1}^{K} w_k\big)^2}$$

- Large sample tests and CIs are obtained from the CLT

Direct Standardization

- Revisiting the smoking example (using combined weights)

$$\widehat{\text{Var}}(\hat{p}_{1_{\text{adj}}} - \hat{p}_{2_{\text{adj}}}) = 0.000318$$

- Testing $H_0 : \pi_{1_{\text{adj}}} = \pi_{2_{\text{adj}}}$ versus $H_A : \pi_{1_{\text{adj}}} \neq \pi_{2_{\text{adj}}}$,

$$Z = \frac{\hat{p}_{1_{\text{adj}}} - \hat{p}_{2_{\text{adj}}}}{\sqrt{\widehat{\text{Var}}(\hat{p}_{1_{\text{adj}}} - \hat{p}_{2_{\text{adj}}})}} = \frac{-0.13}{\sqrt{0.000318}} = -7.28$$

- Conclude that there is a significant difference in the prevalence of smoking between rural and urban women after adjusting for age

Standardization

- *Direct standardization*: Estimate rate or proportion in the reference population using the observed rate or proportion from the study sample
- *Indirect standardization*: Estimate the rate or proportion in the study population using the rate or proportion from the reference population

Indirect Standardization

- Suppose we observe stratum-specific prevalences (or incidences)
  - Reference population: m_k/M_k for $k = 1, \ldots, K$
  - Study population: n_k/N_k for $k = 1, \ldots, K$
- Observed prevalence in the study population

$$\hat{p}_{\text{study}} = \frac{\sum_{k=1}^{K} n_k}{\sum_{k=1}^{K} N_k}$$

- Expected prevalence in the study population assuming stratum-specific prevalences from the reference population

$$\hat{p}_{\text{ref}} = \frac{\sum_{k=1}^{K} N_k m_k/M_k}{\sum_{k=1}^{K} N_k}$$

Indirect Standardization

- *Standardized mortality ratio* (SMR)

$$s = \frac{\hat{p}_{\text{study}}}{\hat{p}_{\text{ref}}} = \frac{\sum_{k=1}^{K} n_k}{\sum_{k=1}^{K} N_k m_k / M_k} = \frac{O}{E}$$

- Note: Calculation of s requires knowing just $\sum_k n_k$ for the study population, that is, we do not need to know the number of events for each level of the confounder

- *Standardized incidence ratio* (SIR) is defined analogously

Indirect Standardization

- The variance of s can be estimated by

$$\widehat{\text{Var}}(s) = \frac{\widehat{\text{Var}}(O) + s^2\widehat{\text{Var}}(E)}{E^2}$$

where $\widehat{\text{Var}}(O) = \sum_k n_k$

and $\widehat{\text{Var}}(E) = \sum_k \left(\frac{N_k}{M_k}\right)^2 m_k$

- To test $H_0 : \pi_{\text{study}}/\pi_{\text{ref}} = 1$ vs. $H_0 : \pi_{\text{study}}/\pi_{\text{ref}} \neq 1$,

$$Z = \frac{s-1}{\sqrt{\widehat{\text{Var}}(s)}} \sim N(0,1)$$

Indirect Standardization

- Revisit smoking example
- Let's compute standardized prevalence ratio for rural women using urban women as the reference population, adjusting for age
- For rural women $O = 414$,

$$E = \frac{387 \times 8}{129} + \frac{441 \times 53}{243} + \frac{300 \times 135}{478} + \frac{210 \times 134}{470}$$

$$= 264.79$$

- Therefore $s = 414/264.79 = 1.56$

Indirect Standardization

- Now $\widehat{\text{Var}}(O) = O = 414$ and

$$\widehat{\text{Var}}(E) = 8\left(\frac{387}{129}\right)^2 + 53\left(\frac{441}{243}\right)^2 + 135\left(\frac{300}{478}\right)^2 + 134\left(\frac{210}{470}\right)^2$$

$$= 326.49$$

- Therefore

$$\widehat{\text{Var}}(s) = \frac{\widehat{\text{Var}}(O) + s^2\widehat{\text{Var}}(E)}{E^2} = \frac{414 + 1.56^2(326.49)}{264.79^2}$$

$$= 0.0173$$

implying

$$Z = \frac{s-1}{\sqrt{\widehat{\text{Var}}(s)}} = 4.29$$

Indirect Standardization

- When computing standardized rates or proportions, inspect observed and expected cells (if feasible) to facilitate understanding

Age	O_k	E_k	O_k/E_k
35-39	44	24.0	1.83
40-44	138	96.2	1.43
45-49	130	84.7	1.53
50-55	102	59.9	1.70