Rates and Proportions
Bios 662

Michael G. Hudgens, Ph.D.

mhudgens@bios.unc.edu

http://www.bios.unc.edu/∼mhudgens

2008-11-12 17:19

Outline

- Prevalence/incidence
- Direct standardization
- Indirect standardization

Rates and Proportions

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

Prevalence

- Random sample of size N from population of interest
- n have disease ("cases")
- Estimator of prevalence:

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

- CIs and tests for prevalence based on

$$n \sim Bin(N, \pi)$$

where π is the population prevalence

Prevalence: Example

- Ex: Random sample of 1717 injection drug users in 6 major cities in U.S. found 206 were HIV positive.
- Estimated prevalence of HIV among IDUs

$$\hat{p} = 206/1717 = .120$$

- Large sample 95 % CI (0.105, 0.135)

Incidence

- Estimator of incidence:

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

- Example: incidence of diabetes among Pima Indians. $N = 1728$, time=6 years, new cases=346 [Ref: AJE Oct 1, 2003, page 669]

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

- Thus estimated incidence is 0.033 cases per person-year

Incidence

- Usually multiply by some number, say 1000

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

- Interpretation: estimated incidence is

  33.4 cases per year per 1000 persons

  or

  33.4 cases per 1000 person years

Incidence

- Note general form

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

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

- Since

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

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

Incidence

- Let

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

such that

$$n \sim Bin(N, \pi)$$

- Note π here is distinct from earlier slide; probability of becoming a case in interval of follow-up
- Thus

$$\hat{V}(\hat{p}) = \hat{p}(1 - \hat{p})/N$$

implying

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

Incidence CI

- Approximate $(1-\alpha)$x100% CI

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

- Diabetes example:

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

- 95% CI

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

Direct Standardization

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

  Urban women: 1320 questioned, 330 current smokers

  Rural women: 1338 questioned, 414 current smokers

$$\hat{p}_u = 330/1320 = .25; \hat{p}_r = 414/1338 = .31$$

- Concern: age may be a confounder

Direct Standardization

- Three steps

  1. Divide samples into K categories of the potential confounder

  2. Compute the confounder category-specific proportions/rates

  3. Compute the weighted average of confounder-specific proportions/rates

- Choice of weights based on *standard or reference population*; e.g. aggregate of samples in hand, governmental population survey

Direct Standardization

- China smoking example

	Urban			Rural		
Age	N_{1i}	n_{1i}	$\hat{p}_{1i}$	N_{2i}	n_{2i}	$\hat{p}_{2i}$
35-39	129	8	.062	387	44	.114
40-44	243	53	.218	441	138	.313
45-49	478	135	.282	300	130	.433
50-54	470	134	.285	210	102	.486

Direct Standardization

- Combined age distribution

Age	N_i	w_i
35-39	516	.194
40-44	684	.257
45-49	778	.293
50-54	680	.256
Total	2658	1

Direct Standardization

- Adjusted prevalence estimator

$$\hat{p}_{j_{adj}} = \frac{\sum_{i=1}^{K} w_i \hat{p}_{ji}}{\sum_{i=1}^{K} w_i}$$

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

Direct Standardization

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

$$\hat{p}_{1_{adj}} = (.194 * .062 + \cdots + .256 * .285)/1 = .224$$

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

- Crude difference, ratio:

$$\hat{p}_1 - \hat{p}_2 = .25 - .31 = -.06$$

$$\hat{p}_2/\hat{p}_1 = .31/.25 = 1.24$$

- Age adjusted difference, ratio:

$$\hat{p}_{1_{adj}} - \hat{p}_{2_{adj}} = .224 - .354 = -.13$$

$$\hat{p}_{2_{adj}}/\hat{p}_{1_{adj}} = .354/.244 = 1.58$$

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	.5
35-39	6	> 84	.5
40-44	6		

Direct Standardization

- China-smoking example with WHO standard:

$$\hat{p}_{1_{adj}} = 0.209; \hat{p}_{2_{adj}} = 0.330$$

$$\hat{p}_{1_{adj}} - \hat{p}_{2_{adj}} = -.121$$

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

Direct Standardization

	Crude	Combined	WHO
Difference	-.06	-.13	-.12
Ratio	1.24	1.58	1.58

- Note Combined/WHO farther from null than Crude
- 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

- $\hat{p}_{j_{adj}}$ is a weighted average of independent RVs (the $\hat{p}_{ji}$'s)
- Since $n_{ji} \sim Bin(N_{ji}, \pi_{ji})$, we know that

$$V(\hat{p}_{ji}) = \pi_{ji}(1 - \pi_{ji})/N_{ji}$$

and

$$\hat{V}(\hat{p}_{ji}) = \hat{p}_{ji}(1 - \hat{p}_{ji})/N_{ji}$$

Direct Standardization

- Thus

$$\hat{V}(\hat{p}_{1_{adj}} - \hat{p}_{2_{adj}}) = \frac{\sum_{i=1}^{K} w_i^2 [\hat{V}(\hat{p}_{1i}) + \hat{V}(\hat{p}_{2i})]}{(\sum_{i=1}^{K} w_i)^2}$$

- Large sample tests and CIs are obtained from the CLT

Direct Standardization

- Revisiting the smoking example (using combined weights)

$$\hat{V}(\hat{p}_{1_{adj}} - \hat{p}_{2_{adj}}) = 0.000318$$

- Testing $H_0 : \pi_{1_{adj}} = \pi_{2_{adj}}$,

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

- Conclude significant difference in prevalence of smoking between rural and urban women after adjusting for age

Standardization

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

Indirect Standardization

- Suppose observe stratum specific prevalences
  - from reference population: m_i/M_i for $i = 1, \ldots, K$
  - from study population: n_i/N_i for $i = 1, \ldots, K$
- Observed prevalence from study population

$$\hat{p}_{study} = \frac{\sum_{i=1}^{K} n_i}{\sum_{i=1}^{K} N_i}$$

- Expected prevalence for study population assuming stratum-specific prevalences from reference population

$$\hat{p}_{ref} = \frac{\sum_{i=1}^{K} N_i m_i / M_i}{\sum_{i=1}^{K} N_i}$$

Indirect Standardization

- *Standardized mortality ratio* (SMR)

$$s = \frac{\hat{p}_{study}}{\hat{p}_{ref}} = \frac{\sum_{i=1}^{K} n_i}{\sum_{i=1}^{K} N_i m_i / M_i} = \frac{O}{E}$$

- Note: calculation of s requires only $\sum_i n_i$, i.e., we do not need to know number of events for each level of confounder in study

- *Standardized incidence ratio* (SIR) defined analogously

Indirect Standardization

- The variance of s can be estimated by

$$\hat{V}(s) = \frac{\hat{V}(O) + s^2\hat{V}(E)}{E^2}$$

where $\hat{V}(O) = \sum_i n_i$ and

$$\hat{V}(E) = \sum_i \left(\frac{N_i}{M_i}\right)^2 m_i$$

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

$$Z = \frac{s-1}{\sqrt{\hat{V}(s)}} \sim N(0,1)$$

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 $O = 414$,

$$E = 8\left(\frac{387}{129}\right)+53\left(\frac{441}{243}\right)+135\left(\frac{300}{478}\right)+134\left(\frac{210}{470}\right) = 264.79$$

- Therefore $s = 414/264.79 = 1.56$

- Now $V(O) = O = 414$ and

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

$$= 326.49$$

- Therefore

$$\hat{V}(s) = \frac{V(O) + s^2 V(E)}{E^2} = \frac{414 + 1.56^2(326.49)}{264.79^2} = 0.0173$$

implying

$$Z = \frac{s-1}{\sqrt{\hat{V}(s)}} = 4.28$$

Indirect Standardization

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

Age	O_i	E_i	O_i/E_i
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