Chapter 9. Analyzing Simple Epidemiologic Data

December 10, 2020
Epidemiology (11)

Minato Nakazawa
<minato-nakazawa@umin.net>

- **In this chapter, the formulas to get confidence intervals and p values are given**
  - The equations are approximates and valid only for *large* samples (though threshold is difficult to determine)
  - More accurate measures are available by *exact* methods
  - Even for the studies with modest numbers, usually the results are same between approximate methods and exact methods
  - If the result is close to the border of statistical significance, the difference between approximate methods and exact methods may affect the result, but it may matter less if the general width and location of a confidence interval is considered.

- **Confidence intervals for measures of disease frequency**
  - Risk data and prevalence data
    - 20 among 100 become ill with flu during the winter season, the risk R=20/100 (=0.2)
    - For confidence interval, binomial model is applied: *a* denotes the number of cases, *N* denotes the population at risk, R=*a*/*N*.
    - Confidence interval can be obtained by equation [9-1]
    - Z is a fixed value taken from standard normal distribution. Z=1.645 for 90% confidence interval and Z=1.96 for 95% confidence interval

$$R_L, R_U = R \pm Z \cdot \mathrm{SE}(R) \quad [9-1]$$

$$\mathrm{SE}(R) = \sqrt{\frac{a(N-a)}{N^3}}$$

- Example: Confidence limits for a risk or prevalence
  - In the flu epidemic of 20 cases among 100 population at risk during a flu season, 90%CI is obtained as 0.13 to 0.27 by below

$$R_L = R - Z \cdot \mathrm{SE}(R) = 0.20 - 1.645 \cdot \sqrt{\frac{20 \cdot 80}{100^3}} = 0.13$$
$$R_U = R + Z \cdot \mathrm{SE}(R) = 0.20 + 1.645 \cdot \sqrt{\frac{20 \cdot 80}{100^3}} = 0.27$$

(Complementary info) see, https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/PASS/Confidence_Intervals_for_One_Proportion.pdf
This is very simple asymptotic formula.
In R software, somewhat improved Wilson Score CI (with/without continuity correction) is readily available. In this case,
`prop.test(20, 100, conf.level=0.9, correct=FALSE)`
gives 90%CI as [0.1425018, 0.2733038]
Exact method is also readily available by
`binom.test(20, 100, conf.level=0.9)`
It gives 90%CI as [0.1366613, 0.2772002].

- Incidence rate data
  - a denotes cases, PT denotes person-time.
  - Different from binomial model.
  - It's impossible to know how many people contributed time by the value of PT.
  - The IR obeys Poisson model.
  - The equation is given below.
- Example: Cancer incidence rate is estimated from a registry that reports 8 cases of astrocytoma among 85000 person-years at risk. 90%CI is 3.9/100000 person-years to 14.9/100000 person-years.
- By exact method, 90%CI I 4.7/100000 person-years to 17.0/100000 person-years.
- See, http://minato.sip21c.org/epispecial/CI-for-measures-of-disease-freq.R

$$\mathrm{IR} = \frac{a}{\mathrm{PT}}$$
$$\mathrm{IR}_L, \mathrm{IR}_U = \mathrm{IR} \pm Z \cdot \mathrm{SE}(\mathrm{IR})$$
$$\mathrm{SE}(\mathrm{IR}) = \sqrt{\frac{a}{\mathrm{PT}^2}}$$
$$a = 8, \mathrm{PT} = 85000$$
$$\mathrm{IR} = 8/85000 = 9.4/100000$$
$$\mathrm{IR}_L = 8/85000 - 1.645 \cdot \sqrt{\frac{8}{85000^2}} = 3.9/100000$$
$$\mathrm{IR}_U = 8/85000 + 1.645 \cdot \sqrt{\frac{8}{85000^2}} = 14.9/100000$$

Box: When whole population is measured instead of sample, there are two ways of consideration (up to context). (1) No sampling error, thus CI doesn't make sense, (2) It's possible by assuming hypothetical superpopulation.

CONFIDENCE INTERVALS FOR MEASURES OF EFFECT (1)

- The **effect** of exposure is compared between two (or more) groups
  - Cohort studies (as **difference** or **ratio**)
    - Direct comparison of **risks** of the exposed and unexposed groups with same follow-up period for all individuals
    - Comparison of **incidence rates** between the exposed and unexposed groups with different follow-up periods by person
  - Case-control studies (as **ratio**)
    - Usually analysis of **odds ratio** is done
  - Surveys or cross-sectional studies
    - Usually the **prevalence** data, treated as risk data because those are expressed as proportions (though the effect measure is often **odds ratio**)
  - Case-fatality rates
    - Also usually treated as risk data because those are proportions

- **Cohort Studies with Risk Data** or **Prevalence Data**
  - Assume the dichotomous exposure (exposed, unexposed), all subjects were followed for a fixed period, no important competing risk, no confounding
  - RD (risk difference) and RR (risk ratio) with SE (standard error) can be estimated by the formula below

	Exposed	Unexposed
Cases	a	b
People at risk	N_1	N_0

$$\mathrm{RD} = \frac{a}{N_1} - \frac{b}{N_0}$$

$$\mathrm{RR} = \frac{a/N_1}{b/N_0}$$

$$\mathrm{SE(RD)} = \sqrt{\frac{a(N_1-a)}{{N_1}^3} + \frac{b(N_0-b)}{{N_0}^3}} \quad [9-2]$$

$$\mathrm{SE(\ln(RR))} = \sqrt{\frac{1}{a} - \frac{1}{N_1} + \frac{1}{b} - \frac{1}{N_0}} \quad [9-3]$$

CONFIDENCE INTERVALS FOR MEASURES OF EFFECT (2)

- Example (Table 9-1)
  - RD is 321/686 – 411/689 = 0.47 – 0.60 = -0.13, 90%CI is -0.17 to -0.08
  - 17% to 8% lower in absolute terms for women receiving combined tamoxifen and radiotherapy
  - RR is 0.47/0.60=0.78, 90%CI is 0.72 to 0.85
  - 28% to 15% lower risk in relative term, compared to tamoxifen alone.

Table 9-1. Risk of recurrence of breast cancer in a randomized trial of women treated with tamoxifen and radiotherapy or tamoxifen alone

	Tamoxifen and radiotherapy	Tamoxifen alone
Women with recurrence	321	411
Total women treated	686	689

Data from Overgaard M et al., 1999 (https://www.ncbi.nlm.nih.gov/pubmed/10335782)

- Confidence intervals vs confidence limits
  - "Interval" is a range indicating the degree of statistical precision that describes the estimate
    - Level of confidence is set arbitrarily
    - Width of the interval expresses the precision: Wider interval implies less precision, narrower interval implies more precision
  - The upper and lower boundaries of the interval are the "limits"

- (Complementary info)
  - In R with fmsb package (including the formula given here), it's easy to calculate by

```
library(fmsb)
riskdifference(321, 411, 686, 689, conf.level=0.9)
riskratio(321, 411, 686, 689, conf.level=0.9)
```

  - Exact confidence intervals can be obtained by Santner-Snell method or Z-pooled method. Getting exact confidence intervals of RD by Z-pooled method is possible using R with Exact package such as

```
library(Exact)
T <- matrix(c(321, 411, 365, 278), 2)
exact.test(T, conf.int=TRUE, conf.level=0.9)
```

CONFIDENCE INTERVALS FOR MEASURES OF EFFECT (3)

	Exposed	Unexposed
Cases	a	b
People-time at risk	PT_1	PT_0

- Cohort studies with **incidence rate** (IR) data
  - IR among exposed
    $IR_1 = a/PT_1$
  - IR among unexposed
    $IR_0 = b/PT_0$
  - $IRD = IR_1 - IR_0 = a/PT_1 - b/PT_0$
  - $IRR = IR_1 / IR_0 = (a/PT_1)/(b/PT_0)$
  - Standard errors can be obtained by the following formula

$$\mathrm{SE}(\mathrm{IRD}) = \sqrt{\frac{a}{\mathrm{PT_1}^2} + \frac{b}{\mathrm{PT_0}^2}} \quad [9-4]$$

$$\mathrm{SE}(\ln(\mathrm{IRR})) = \sqrt{\frac{1}{a} + \frac{1}{b}} \quad [9-5]$$

Table 9-2. Incidence rate of cancer among a blind population and a population that is visually severely impaired but not blind

	Totally blind	Visually severely impaired but not blind
Cancer cases	136	1709
Person-years	22050	127650

Data from Feychting M et al., 1998 (https://www.ncbi.nlm.nih.gov/pubmed/9730026)

- Example (Table 9-2)
  - Feychting et al. calculated standardized rate ratio with exact 95%CI based on national data and Poisson distribution
  - IRD = 136/22050 – 1709/127650 = -7.2/1000 person-years (pyrs), 90%CI is -8.2/1000 pyrs to -6.2/1000 pyrs
  - By R with fmsb package, ratedifference(136, 1709, 22050, 127650, conf.level=0.9)
  - IRR = (136/22050) / (1709/127650) = 0.46, 90%CI is 0.40 to 0.53
  - By R with fmsb package rateratio(136, 1709, 22050, 127650, conf.level=0.9)

CONFIDENCE INTERVALS FOR MEASURES OF EFFECT (4)

	Exposed	Unexposed
Cases	a	b
Controls	c	d

- Case-Control Studies (for density case-control study or cumulative case-control study)
  - Analysis of case-cohort studies and case-crossover studies is slightly different
  - As the estimate of IRR or RR (depending on how the controls were sampled), OR is used.
  - OR = ad/bc
  - Standard errors can be obtained by the following formula

$$\mathrm{SE}(\ln(\mathrm{OR})) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \quad [9-6]$$

Table 9-3. Frequency of recent amphetamine use among stroke cases and controls among women between 15 and 44 years old

	Amphetamine users	No Amphetamine use
Stroke cases	10	337
Controls	5	1016

Data from Petitti et al., 1998 (https://www.ncbi.nlm.nih.gov/pubmed/9799166)

- Example (Table 9-3)
  - OR = (10/337)/(5/1016) = 6.0, 90%CI is 2.4 to 14.9
  - By R with fmsb package, oddsratio(10, 337, 5, 1016, conf.level=0.9)
  - The point estimate is the geometric mean between the lower limit and upper limit of the CI. This relation applies whenever CI is set on the log scale.

CALCULATION OF *P* VALUES

- Though CI is better than p-values, the basic formula to calculate p-values is given for completeness. Testing the null hypothesis that exposure is not related to disease.
- Risk Data
  - χ-statistics is used to get p-value (eg. Table 9-1 data for [9-7] using standard normal distribution given in Appendix, whereas it's easy to get the p-value using R function pnorm().
  - χ= -4.78, p≈0.0000009 (Assuming one-sided, pnorm(-4.78) → 8.76 x 10^{-7}; By riskratio(), p=1.785e-06, it's two-sided.)

	Exposed	Unexposed	Total
Cases	*a*	*b*	M_1
Noncases	*c*	*d*	M_0
People at risk	N_1	N_0	*T*

$$\chi = \frac{a - \frac{N_1 M_1}{T}}{\sqrt{\frac{N_1 N_0 M_1 M_0}{T^2 (T-1)}}} \quad [9-7]$$

$$\chi = \frac{a - \frac{PT_1}{T} M}{\sqrt{M \frac{PT_1}{T} \frac{PT_0}{T}}} \quad [9-8]$$

- Incidence rate data
  - χ-statistics is used to get p-value (eg. Table 9-2 data for [9-8] using standard normal distribution)
  - χ= -8.92, p<10^{-20} (Assuming one-sided, pnorm(-8.92) → 2.3 x 10^{-19}; By rateratio(), p<2.2e-16)

	Exposed	Unexposed	Total
Cases	*a*	*b*	*M*
Person-time	PT_1	PT_0	*T*

- Case-control data
  - [9-7] can be used, because the null hypothesis (as 2x2 table, exposure and disease are independent) is same for risk data and case-control data (It's the answer to Question 5).
  - Eg. Table 9-3 ata for [9-7] using standard normal distribution
  - χ= 3.70, p=0.00022 (Two-sided test, 2*(1-pnorm(3.7))=0.0002155…; By oddsratio(), p=0.0002196; Difference due to rounding error)
- All those were easily obtained by fmsb package's functions riskratio(), rateratio() and oddsratio().