In-Class 3. Statistical Methods in Epidemiology

Suzer-Gurtekin

January 2025

Group Assignments

Group	Student
1	Einolf, Zach Scott
1	Fan, Zhaoyun
1	Mishra, Rohin Prem
1	DesJardins, Grace
2	Adeniyi, Kehinde
2	Lugu, Nicholas Reign
2	LU, Aria
2	Gunderson, Jeremy
3	Wenner, Theodore D
3	Zhou, Zhenjing
3	Kim, Jay
3	Bei, Rongqi
4	Beshaw, Yael Dejene
4	Hoglund, Quentin Michael
4	Jiang, Yujing
4	Jiang, Weishan
5	Popky, Dana
5	Sani, Jamila
5	O'Connell, Greg Al
5	Saucedo, Valeria Castaneda

Group	Student
6	Hussein, Aya Moham
6	Zou, Jianing
6	Wang, Zixin
6	Chakravarty, Sagnik
7	Valmidiano, Megan
7	Glidden, Sarah Acton
7	Sun, Yao
7	Blakney, Aaron
8	Xu, Kailin
8	Linares, Kevin
8	Odei, Doris
8	Nana Mba, Line
9	Zhou, Huan
9	Meng, Lingchen
9	Lin, Xinyu
9	Ge, Feiran
10	Liu, Xiaoqing
10	Lu, Angelina
10	Baez-Santiago, Felix
10	Ma, Ruisi

Group	Student
11	Ding, Yuchen
11	Shrivastava, Namit
11	Kakiziba, Johnia Johansen
11	Cranmer, Evan Koba

Expectations

Active participation in

- Reviewing question/data/method
- Code writing
- Computations
- Interpretation of results
- Select a spokesperson for group discussion

Action Plan

- Introduction to data collection and data presentation
  - Prevalence, incidence, relative risk, odds ratio
  - In-class Exercise 1
    - Group discussion: ~20 minutes
    - Class discussion: ~10 minutes
- Prospective vs. Retrospective studies
  - Prospective studies
  - Retrospective studies
  - In-class Exercise 3
    - Group discussion: ~20 minutes
    - Class discussion: ~5 minutes
- Attrition
  - Review
  - In-class Exercise 4
    - Group discussion: ~20 minutes
    - Class discussion: ~5 minutes
- Attributable risk
  - Review
  - In-class Exercise 5
    - Group discussion: ~20 minutes
    - Class discussion: ~5 minutes
  - In-class Exercise 6
    - Group discussion: ~20 minutes
    - Class discussion: ~5 minutes

- Class discussion on the key concepts from today's lecture
- Review of HW3 and Project
- Q&A

Overview

Notation

Table: General Classification of a Population by Risk Factor and Disease Status

Risk Factor Classification	Disease Classification		
	+(present)	-(absent)	Total at Risk
+(present)	A	B	A+B
-(absent)	C	D	C+D
Total	A+C	B+D	T

Prevalence

A key statistic from the two-way table is the **prevalence** rate. The prevalence is the proportion of the population that has the condition.

$$P_{Exposed} = \frac{A}{A+B}$$

$$P_{Unexposed} = \frac{C}{C+D}$$

$$P_{Pop} = \frac{A+C}{T}$$

Incidence

The **incidence proportion** is the proportion of the population that will develop a condition during a specified time period. The following are formulae for the incidence proportion:

$$I_{Exposed} = \frac{A}{A+B}$$

$$I_{Un\exp osed} = \frac{C}{C+D}$$

$$I_{Pop} = \frac{A+C}{T}$$

Relative Risk

The relative risk is often used to compare the incidence proportions across groups:

$$RR = \frac{I_{Exposed}}{I_{Unexposed}} = \frac{\frac{A}{A+B}}{\frac{C}{C+D}} = \frac{A(C+D)}{C(A+B)}$$

Relative risk is sometimes also used to compare incidence rates or even prevalence.

Odds Ratio

In this new notation:

$$OR = \frac{\frac{A}{A+B} \Big/ \frac{B}{A+B}}{\frac{C}{C+D} \Big/ \frac{D}{C+D}} = \frac{AD}{BC}$$

Example 1

Table: Data for Example 1 and 2

Smoker	Stroke: Yes	Stroke: No	Total
Yes	171	3,264	3,435
No	117	4,320	4,437
Total	288	7,584	7,872

Please calculate incidence among smokers, non-smokers, and the relative risk and odds ratio for smokers compared to non-smokers.

In-Class 1

$$I_{Smo\ker} = \frac{171}{3435} \approx 0.0498$$

$$I_{Nonsmo\ker} = \frac{117}{4437} \approx 0.0264$$

$$RR = \frac{I_{Smoker}}{I_{Nonsmoker}} = \frac{\frac{171}{3,435}}{\frac{117}{3,347}} \approx 1.89$$

$$\hat{OR} = \frac{AD}{BC} = \frac{171 \times 4320}{117 \times 3264} = 1.93$$

Prospective Studies

For these studies, the following estimators are used for the incidence rate and its variance:

$$E\left\{\frac{a}{a+b}\right\} = \frac{A}{A+B}$$

$$V\left\{\frac{a}{a+b}\right\} = \frac{AB}{(a+b)\,(A+B)^2} \approx \frac{p(1-p)}{n}$$

Note that the variance estimator incorporates population values. These are for the group with the risk factor, also called the exposed group. There are similar estimators for the unexposed group.

Prospective Studies

Approximate confidence intervals for the RR and OR can be constructed on the logarithmic scale:

$$\hat{V}\{\ln \hat{R}\} = \frac{b}{a(a+b)} + \frac{d}{c(c+d)}$$

$$\hat{V}\{\ln \hat{O}\} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

Prospective Studies

These quantities are then used to construct confidence intervals using the following steps:

- $\ln(\hat{O}) \pm 1.96\sqrt{\hat{V}(\ln(\hat{O}))} = (L, U)$
- (e^L, e^U)

Retrospective or Case-Control Studies

Under this design, $\frac{a}{a+b}$ is *not* an unbiased estimator the population incidence.

This should make sense as those are set by the *design*, and not by their rate of occurrence in the population.

Retrospective or Case-Control Studies

We can estimate slightly different quantities:

$$E\left\{\frac{a}{a+c}\right\} = \frac{A}{A+C}$$

$$E\left\{\frac{b}{b+d}\right\} = \frac{B}{B+D}$$

For example, the proportion of persons with cancer (*cases*) that smoke. And the proportion of persons without cancer (*controls*) that smoke.

Retrospective or Case-Control Studies

If the sample sizes are large, then the estimated odds ratio

$$\hat{O} = \frac{ad}{bc}$$

is a *consistent* estimator. The variance can be estimated on the logarithmic scale:

$$\hat{V}\{\ln \hat{O}\} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

We also know that for rare conditions, the odds ratio and relative risk are approximately equal. This gives us a way to estimate these quantities in case-control studies.

In-Class Variance Calculations

Using the data on the next slide, and with the knowledge that this is a prospective study, please calculate $\hat{V}\{\ln \hat{R}\}$ and $\hat{V}\{\ln \hat{O}\}$. Once you have these variances, please compute 95% confidence intervals for each.

Example 2

Table: Data for Example 1 and 2

Smoker	Stroke Yes	Stroke No	Total
Yes	171	3,264	3,435
No	117	4,320	4,437
Total	288	7,584	7,872

Please calculate incidence among smokers, non-smokers, and the relative risk and odds ratio for smokers compared to non-smokers.

In-Class Variance Calculations

$$\hat{V}\{\ln \hat{R}\} = \frac{3264}{171\,(3435)} + \frac{4320}{117\,(4437)} = 0.0139$$

$$\hat{V}\{\ln \hat{O}\} = \frac{1}{171} + \frac{1}{3264} + \frac{1}{117} + \frac{1}{4320} = 0.0149$$

In-Class CI Formation

$SE(ln\hat{R}) = 0.1178$ and $SE(ln\hat{O}) = 0.1222$.

We also need $ln\hat{R} = 0.6355$ and $ln\hat{O} = 0.6598$.

Then, $LL = ln\hat{R} - 1.96 \times SE(ln\hat{R}) = 0.4046$ and
$UL = ln\hat{R} + 1.96 \times SE(ln\hat{R}) = 0.8664$.

Finally, back to the scale of $\hat{R}$. $LL = e^{0.4046} = 1.4986$ and
$UL = e^{0.8664} = 2.3782$.

For $\hat{O}$, $LL = e^{0.4203} = 1.5224$ and $UL = e^{0.8993} = 2.4579$.

Attrition

- Clinical trials with follow-up after initial recruitment often lose participants
- This process is known as attrition
- In some circumstances, it may create biased estimates
- Akin to nonresponse bias

Attrition

Table: Example Table with Notation

		Outcome		
		1	0	
Exposure	1	$r_{11}p_1n_{1+}$	$r_{12}(1-p_1)n_{1+}$	n_{1+}
	0	$r_{21}p_2n_{2+}$	$r_{22}(1-p_2)n_{2+}$	n_{2+}

If $r_{11} = r_{12} = r_{21} = r_{22}$, then the Odds Ratio is not biased.

When will the odds ratio be biased?

Attrition - In-Class Exercise

Table: Example Table with Notation

		Outcome		
		1	0	
Exposure	1	$r_{11}p_1n_{1+}$	$r_{12}(1-p_1)n_{1+}$	n_{1+}
	0	$r_{21}p_2n_{2+}$	$r_{22}(1-p_2)n_{2+}$	n_{2+}

When will the odds ratio be biased?

When $r_{11} = r_{22} \neq r_{21} = r_{12}$.

1. Explain this in words?

In-Class Exercise

Table: Substance Abuse Treatment Program Evaluation

		Abusing Substances After 6 Months		
		1	0	
Program	1	$r_{11} * 10$	$r_{12} * 90$	$n = 100$
	0	$r_{21} * 20$	$r_{22} * 80$	$n = 100$

2. Calculate the odds ratio for the full sample, i.e. $r_{11} = 1.0, r_{12} = 1.0, r_{21} = 1.0, r_{22} = 1.0$.
3. Calculate the odds ratio when $r_{11} = .8, r_{12} = .6, r_{21} = .6, r_{22} = .8$.
4. Calculate the odds ratio when $r_{11} = .8, r_{12} = .8, r_{21} = .6, r_{22} = .6$.

In-Class Exercise

Table: Substance Abuse Treatment Program Evaluation

		Abusing Substances After 6 Months		
		1	0	
Program	1	$r_{11} * 10$	$r_{12} * 90$	$n = 100$
	0	$r_{21} * 20$	$r_{22} * 80$	$n = 100$

1. Calculate the odds ratio for the full sample, i.e. $r_{11} = 1.0, r_{12} = 1.0, r_{21} = 1.0, r_{22} = 1.0$. $OR = 0.444$
2. Calculate the odds ratio when $r_{11} = .8, r_{12} = .6, r_{21} = .6, r_{22} = .8$. $OR = 0.790$
3. Calculate the odds ratio when $r_{11} = .8, r_{12} = .8, r_{21} = .6, r_{22} = .6$. $OR = 0.444$

Attributable Risk in Exposed Group

Attributable Risk in Exposed Group. Conceptually, this is the proportion of risk that is related to exposure.

The "excess incidence rate in the exposed group"
$= I_{Exposed} - I_{Unexposed}$.

The attributable risk in exposed group is:

$$A_{Exposed} = \frac{I_{Exposed} - I_{Un\exp osed}}{I_{Exposed}} = 1 - \frac{I_{Un\exp osed}}{I_{Exposed}} = \frac{R-1}{R}$$

Attributable Risk in Population

Conceptually, the reduction in incidence in the population that would occur in the absence of the risk factor.

$$A_{Pop} = \frac{I_{Pop} - I_{Unexposed}}{I_{Pop}} = \frac{P(R-1)}{1+P(R-1)}$$

where $P = \frac{A+B}{T}$ is the proportion of the population exposed to the risk factor.

Estimators of Attributable Risk: Prospective Studies

Estimators for Prospective Studies (Jewell, 2004):

$$\hat{R} = \frac{a(c+d)}{c(a+b)}$$

$$\hat{A}_{Exposed} = \frac{\hat{R}-1}{\hat{R}}$$

$$\hat{P} = \frac{a+b}{t}$$

$$\hat{A}_{Pop} = \frac{\hat{P}(\hat{R}-1)}{1+\hat{P}(\hat{R}-1)} = \frac{ad-bc}{(a+c)(c+d)}$$

$$V\left(\ln(1-\hat{A}_{Pop})\right) = \frac{b+\hat{A}_{Pop}(a+d)}{tc}$$

In-Class Exercise 5

Assume the following data were collected from a Prospective study. Please estimate the relative risk($\hat{R}$), odds ratio ($O\hat{}R$), the attributable risk in population ($\hat{A}_{pop}$), and $V(\ln(1-\hat{A}_{Pop}))$.

Table: Servings of Vegetables Per Day and Heart Disease

		Heart Disease	
		Yes	No
Avg Servings	0-2	23	125
	3+	13	150

In-Class Prospective Solution

$$\hat{R} = \frac{a(c+d)}{c(a+b)} = 1.949$$

$$\hat{O} = \frac{ad}{bc} = 2.123$$

$$\hat{A}_{Pop} = \frac{\hat{P}(\hat{R}-1)}{1+\hat{P}(\hat{R}-1)} = \frac{ad-bc}{(a+c)(c+d)} = 0.3110$$

$$V\left(\ln(1-\hat{A}_{Pop})\right) = \frac{b+\hat{A}_{Pop}(a+d)}{tc} = 0.0442$$

Estimators of Attributable Risk: Retrospective Studies

Estimators for Retrospective Studies. In this case, we need the "rare" disease assumption. Can't do relative risk, so substitute odds ratio:

$$\hat{R} = \frac{ad}{bc}$$

$$\hat{A}_{Exposed} = \frac{\hat{R}-1}{\hat{R}}$$

$$\hat{A}_{Pop} = \frac{(ad-bc)}{d(a+c)}$$

$$V\left(\ln(1-\hat{A}_{Pop})\right) = \frac{a}{c(a+c)} + \frac{b}{d(b+d)}$$

In-Class Exercise 6

Table: Servings of Vegetables Per Day and Heart Disease

		Heart Disease	
		Yes	No
Avg Servings	0-2	23	125
	3+	13	150

If these data were collected from a retrospective study:

Would we have adequate estimates of $\hat{R}$?

What is the estimate of $\hat{R}$?

What is the estimate of $\hat{A}_{Exposed}$?

What is the estimate of $\hat{A}_{Pop}$?

In-Class Retrospective Solution

$$\hat{R} = \frac{ad}{bc} = 2.124$$

$$\hat{A}_{Exposed} = \frac{\hat{R}-1}{\hat{R}} = 0.5290$$

$$\hat{A}_{Pop} = \frac{(ad-bc)}{d(a+c)} = 0.3380$$

$$V\left(\ln(1 - \hat{A}_{Pop})\right) = \frac{a}{c(a+c)} + \frac{b}{d(b+d)} = 0.0522$$

Confidence Intervals

For both prospective and retrospective studies, we estimated the variance of $\ln(1 - \hat{A}_{Pop})$.

Form confidence intervals using the following back-transformation to scale of $\hat{A}_{Pop}$:

$$LCL = 1 - \exp\left(\ln(1 - \hat{A}_{Pop}) + 1.96\sqrt{V\left(\ln(1 - \hat{A}_{Pop})\right)}\right)$$

$$UCL = 1 - \exp\left(\ln(1 - \hat{A}_{Pop}) - 1.96\sqrt{V\left(\ln(1 - \hat{A}_{Pop})\right)}\right)$$

Example 6

Remember the *epiR* package for attributable risk:

```
library(epiR)

(bp<-
matrix(data=c(23,13,125,150),nrow=2))

bp<-as.table(bp)

epi.2by2(bp)
```

Example 6

```
> epi.2by2(bp,method="cross.sectional")

             Outcome +    Outcome -      Total                Prev risk *
Exposed +           23          125        148     15.54 (10.11 to 22.40)
Exposed -           13          150        163       7.98 (4.31 to 13.25)
Total               36          275        311      11.58 (8.24 to 15.66)

Point estimates and 95% CIs:
-------------------------------------------------------------------
Prev risk ratio                                 1.95 (1.02, 3.71)
Prev odds ratio                                 2.12 (1.03, 4.36)
Attrib prev in the exposed *                    7.57 (0.40, 14.73)
Attrib fraction in the exposed (%)             48.68 (2.41, 73.01)
Attrib prev in the population *                 3.60 (-1.87, 9.07)
Attrib fraction in the population (%)          31.10 (-4.04, 54.37)
-------------------------------------------------------------------
Uncorrected chi2 test that OR = 1: chi2(1) = 4.337 Pr>chi2 = 0.037
Fisher exact test that OR = 1: Pr>chi2 = 0.050
 Wald confidence limits
 CI: confidence interval
 * Outcomes per 100 population units
```

Example 6

```
 Outcomes per 100 population units
> epi.2by2(bp,method="case.control")
             Outcome +    Outcome -      Total                 Odds
Exposed +           23          125        148      0.18 (0.11 to 0.28)
Exposed -           13          150        163      0.09 (0.04 to 0.14)
Total               36          275        311      0.13 (0.09 to 0.18)

Point estimates and 95% CIs:
-------------------------------------------------------------------
Exposure odds ratio                          2.12 (1.03, 4.36)
Attrib fraction (est) in the exposed (%)    52.79 (-1.83, 78.96)
Attrib fraction (est) in the population (%) 33.80 (-3.59, 57.69)
-------------------------------------------------------------------
Uncorrected chi2 test that OR = 1: chi2(1) = 4.337 Pr>chi2 = 0.037
Fisher exact test that OR = 1: Pr>chi2 = 0.050
 Wald confidence limits
 CI: confidence interval
```