Comparing two groups: categorical data

Tuan V. Nguyen
Professor and NHMRC Senior Research Fellow
Garvan Institute of Medical Research
University of New South Wales
Sydney, Australia

What we are going to learn …

- Examples (RCT, CC, Cohort)
- Two proportions
- Metrics of effect: d, RR, OR
- Applicability of d, RR, OR
- D and z-test
- NNT
- Measure of association: OR
- Small sample size: Fisher’s exact test

Zoledronate and fracture

Table 2. Rates of Fracture and Death in the Study Groups.*

Variable	Placebo	Zoledronic Acid	Hazard Ratio (95% CI)	P Value
Fracture — no. (cumulative %)				
Any	139 (13.9)	92 (8.6)	0.65 (0.50–0.84)	0.001
Nonvertebral	107 (10.7)	79 (7.6)	0.73 (0.55–0.98)	0.03
Hip	33 (3.5)	23 (2.0)	0.70 (0.41–1.19)	0.18
Vertebral	39 (3.8)	21 (1.7)	0.54 (0.32–0.92)	0.02
Death — no. (%)	141 (13.3)	101 (9.6)	0.72 (0.56–0.93)	0.01

* Rates of clinical fracture were calculated by Kaplan–Meier methods at 24 months and therefore are not simple percentages. There were 1062 patients in the placebo group, and 1065 in the zoledronic acid group. Because of variable follow-up, the number and percentage of patients who died are provided on the basis of 1057 patients in the placebo group and 1054 patients in the zoledronic acid group in the safety population.

Randomized controlled clinical trial

Placebo n = 1062, Zoledronate n = 1065

Length of follow-up: 3 years

Lyles KW, et al. Zoledronic acid and clinical fractures and mortality after hip fracture. *N Engl J Med* 2007;357. DOI: 10.1056/NEJMoa074941

Smoking and lung cancer

	Lung Cancer	Controls
Smokers	647	622
Non-smokers	2	27

R Doll and B Hill. BMJ 1950; ii:739-748

Sir Richard Doll (1912 – 2005)

http://en.wikipedia.org/wiki/Richard_Doll

Is there an association between smoking and lung cancer?

Mortality in the Titanic incident

Class	Dead	Survived	Total
I	123	200 (62%)	323
II	158	119 (43%)	277
III	528	181 (26%)	709
Total	809	500 (38%)	1309

http://lib.stat.cmu.edu/S/Harrell/data/descriptions/titanic3info.txt

Is there an association between passenger class and and death?

What are common characteristics of these data?

- Binary outcome: yes/no; dead / survived
- Proportion / percent / probability

Sample vs population

	Sample		Population	
	Group 1	Group 2	Group 1	Group 2
N	n_1	n_2	Infinite	Infinite
Probability of outcome	p_1	p_2	$\pi_1 = ?$	$\pi_2 = ?$
Difference	$d = p_1 - p_1$		$\delta = \pi_1 - \pi_2$	
Status	Known		Unknown	

Aim: use sample data *d* to estimate population parameter δ

Metrics of effect

- **Absolute difference (d)**
- **Relative risk (RR; risk ratio)**
- **Odds ratio (OR)**
- **Number needed to treat (NNT)**

The choice is dependent on study design

Absolute difference d

Outcome	Placebo	Treatment
Any fracture	139	92
Non-fracture	923	973
N	1062	1065

Outcome	Group 1	Group 2
Bad	a	b
Good	c	D
N	N_1	N_2

Absolute difference

$p_1 = 139 / 1062 = 0.131$

$p_2 = 92 / 1065 = 0.086$

$d = p_2 - p_1 = -0.044$

$p_1 = a / N_1$

$p_2 = b / N_2$

$d = p_2 - p_1$

Number needed to treat – NNT

Outcome	Placebo	Treatment
Any fracture	139	92
Non-fracture	923	973
N	1062	1065

Outcome	Group 1	Group 2
Bad	a	b
Good	c	D
N	N_1	N_2

Number needed to treat

$p_1 = 139 / 1062 = 0.131$

$p_2 = 92 / 1065 = 0.086$

$d = p_2 – p_1 = -0.044$

$NNT = 1 / d = 22$

$p_1 = a / N_1$

$p_2 = b / N_2$

$d = p_2 – p_1$

$NNT = 1 / d$

Relative risk - *RR*

Outcome	Placebo	Treatment
Any fracture	139	92
Non-fracture	923	973
N	1062	1065

Outcome	Group 1	Group 2
Bad	a	b
Good	c	D
N	N_1	N_2

Relative risk

$p_1 = 139 / 1062 = 0.131$

$p_2 = 92 / 1065 = 0.086$

$RR = p_2 / p_1 = 0.66$

$p_1 = a / N_1$

$p_2 = b / N_2$

$RR = p_2 / p_1$

Meaning of RR

- Risk of developing disease

  Treatment: $p_1 = a / N_1$

  Placebo: $p_2 = b / N_2$
- Relative risk

$$RR = p_1 / p_2$$

- Implications:

  RR = 1, there is no effect

  RR < 1, the treatment is beneficial.

  RR > 1, the treatment is harmful.

Odds ratio - *OR*

Outcome	Placebo	Treatment
Any fracture	139	92
Non-fracture	923	973
N	1062	1065

Outcome	Group 1	Group 2
Bad	a	b
Good	c	D
N	N_1	N_2

Odds ratio

$odds_1 = 139 / 923 = 0.140$

$odds_2 = 92 / 973 = 0.094$

$OR = odds_2 / odds_1 = 0.68$

$odds_1 = a / c$

$odds_2 = b / d$

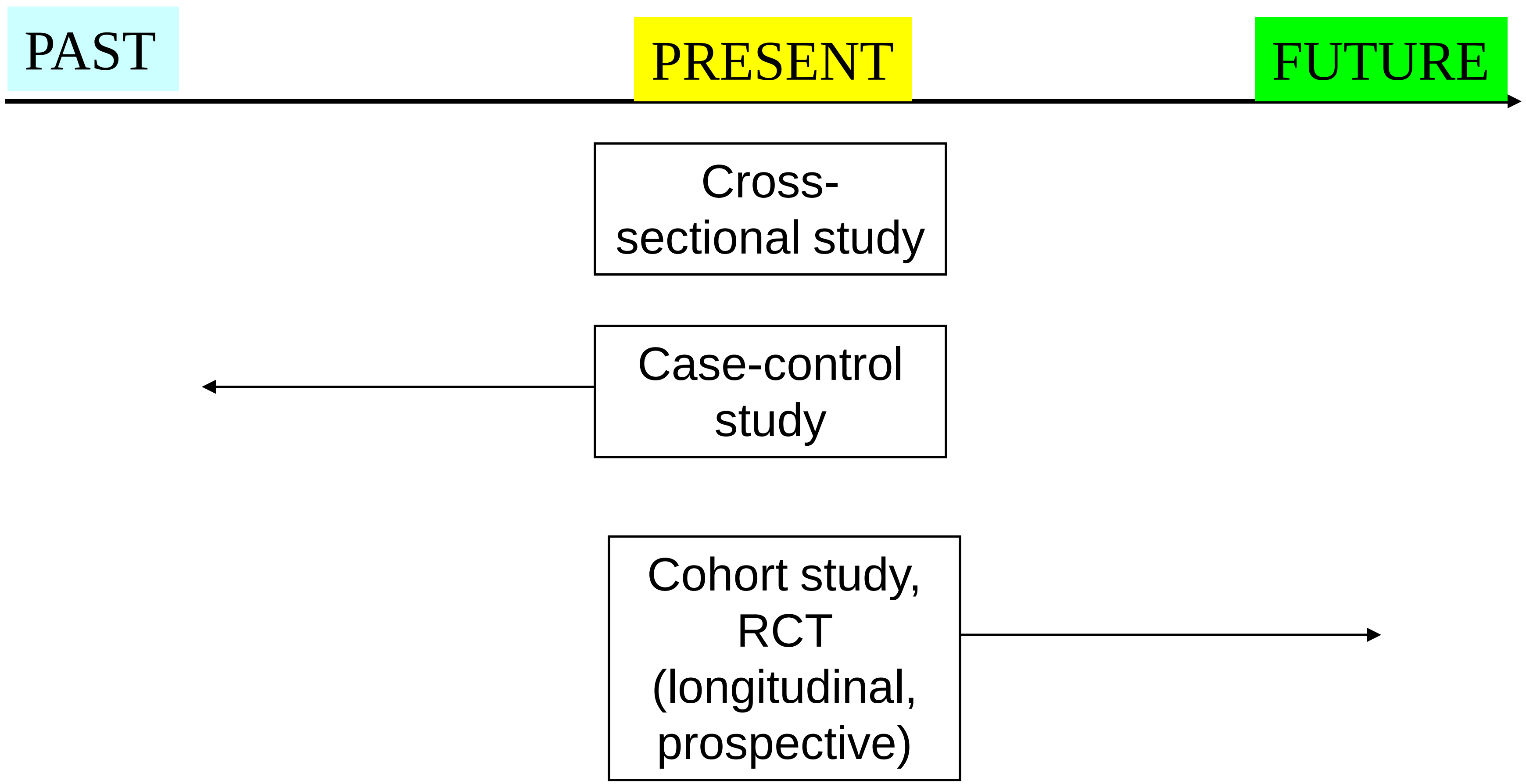
$OR = odds_2 / odds_1$

$OR = (a \times d) / (b \times c)$

Meaning of OR

- **OR = 1, there is no association**
- **OR < 1, the risk factor is associated with *reduced* disease risk**
- **OR > 1, the risk factor is associated with *increased* disease risk**

Study design – time aspect



Appropriateness of effect size

RCT / prospective study	Cross-sectional study	Case-control study
Relative risk	Odds ratio	Odds ratio
Odds ratio	Prevalence ratio	
NNT	D	
D		

Problem and solution

- Finding an estimate for d, OR, RR is easy
- Finding the 95% confidence interval is harder
- We can however use R

Example of *d*

	Treatment	Control
Disease	*a*	*b*
No disease	*c*	*d*
Sample size	N_1	N_2

	Zole	Placebo
Fracture	92	139
No fracture	973	923
Sample size	1065	1062

$$p_1 = \frac{a}{N_1} \qquad p_2 = \frac{b}{N_2}$$

$$d = p_1 - p_2$$

$$SE(d) = \sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_1(1-p_2)}{N_2}}$$

$$95\%CI = d \mp 1.96 SE(d)$$

$$d = \frac{92}{1065} - \frac{139}{1062} = 0.131 - 0.086 = 0.044$$

$$SE(d) = \sqrt{\frac{0.131(0.869)}{1065} + \frac{0.044(0.956)}{1062}} = 0.0134$$

$$95\%CI(d) = 0.044 \mp 1.96 \times 0.0134$$

$$95\%CI(d) = 0.018, 0.081$$

Example of NNT

$$d = \frac{92}{1065} - \frac{139}{1062} = 0.131 - 0.086 = 0.044$$

$$SE(d) = \sqrt{\frac{0.131(0.869)}{1065} + \frac{0.044(0.956)}{1062}} = 0.0134$$

$$95\%CI(d) = 0.044 \mp 1.96 \times 0.0134$$

$$95\%CI(d) = 0.018, 0.081$$

- **NNT = 1 / 0.044 = 22**
- **95% CI for NNT:**
  - **1 / 0.018 = 55**
  - **1 / 0.081 = 14**

Example of RR

	Treatment	Control
Disease	*a*	*b*
No disease	*c*	*d*
Sample size	N_1	N_2

	Zole	Placebo
Fracture	92	139
No fracture	973	923
Sample size	1065	1062

$$RR = \frac{a / N_1}{b / N_2}$$

$$LRR = \log(RR)$$

$$SE(LRR) = \sqrt{\frac{1}{a} - \frac{1}{N_1} + \frac{1}{b} - \frac{1}{N_2}}$$

$$95\%CI(LRR) = LRR \mp 1.96 SE(LRR)$$

$$95\%CI(RR) = e^{LRR \mp 1.96 SE(LRR)}$$

$$RR = \frac{92/1065}{139/1062} = \frac{0.086}{0.131} = 0.66$$

$$LRR = \log(0.66) = -0.4155$$

$$SE(LRR) = \sqrt{\frac{1}{92} - \frac{1}{1065} + \frac{1}{139} - \frac{1}{1062}} = 0.127$$

$$95\%CI(LRR) = -0.416 \mp 1.96 \times 0.127$$

$$95\%CI(RR) = e^{-0.416 \mp 1.96 \times 0.127}$$

= 0.514 to 0.847

Example of OR

	Disease	No disease
Risk +ve	*a*	*b*
Risk –ve	*c*	*d*

	Lung K	Control
Smoking	647	622
No smoking	2	27

$$OR = \frac{ad}{bc}$$

$$LOR = \log(OR)$$

$$SE(LOR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

$$95\%CI(LOR) = LOR \mp 1.96SE(LOR)$$

$$95\%CI(OR) = e^{LOR \mp 1.96SE(LOR)}$$

$$OR = \frac{647 \times 27}{622 \times 2} = 14.04$$

$$LOR = \log(14.04) = 2.64$$

$$SE(LOR) = \sqrt{\frac{1}{647} + \frac{1}{622} + \frac{1}{2} + \frac{1}{27}} = 0.735$$

$$95\%CI(LOR) = 2.642 \mp 1.96 \times 0.735$$

$$95\%CI(OR) = e^{2.64 \mp 1.96 \times 0.735}$$

= 3.32 to 59.03

Introducing epiR package

	Disease	No disease
Exposed (treatment)	*a*	*b*
Not exposed (control)	*c*	*d*

epi.2by2(a, b, c, d, method = “xxx", conf.level = 0.95)

Where method = "cohort.count"
"case.control"
"cross.sectional"

Application of epiR – RCT study

	Fracture	No frcture
Zoleronate	92	973
Placebo	139	923

```
library(epiR)
epi.2by2(92, 973, 139, 923, method="cohort.count",
conf.level=0.95)
```

```
> epi.2by2(92, 973, 139, 923, method = "cohort.count", conf.level = 0.95)
             Disease +    Disease -      Total        Inc risk *        Odds
Exposed +           92          973       1065              8.64      0.0946
Exposed -          139          923       1062             13.09      0.1506
Total              231         1896       2127             10.86      0.1218

Point estimates and 95 % CIs:
---------------------------------------------------------
Inc risk ratio                          0.66 (0.51, 0.85)
Odds ratio                              0.63 (0.48, 0.83)
Attrib risk *                           -4.45 (-7.09, -1.81)
Attrib risk in population *             -2.23 (-4.65, 0.19)
Attrib fraction in exposed (%)          -51.51 (-94.42, -18.08)
Attrib fraction in population (%)       -20.52 (-33.15, -9.08)
---------------------------------------------------------
 * Cases per 100 population units
```

Application of epiR – Case-control study

	K	Not K
Smoking	647	622
No smoking	2	27

```
> epi.2by2(647,622,2,27, method="case.control", conf.level=0.95)
          Disease +    Disease -      Total        Prevalence *        Odds
Exposed +       647          622       1269                51.0       1.040
Exposed -         2           27         29                 6.9       0.074
Total           649          649       1298                50.0       1.000

Point estimates and 95 % CIs:
---------------------------------------------------------------
Odds ratio                                 14.04 (3.33, 59.3)
Attrib prevalence *                        44.09 (34.46, 53.71)
Attrib prevalence in population *          43.1 (33.49, 52.72)
Attrib fraction (est) in exposed  (%)      92.88 (69.93, 98.31)
Attrib fraction (est) in population (%)    92.59 (68.98, 98.23)
---------------------------------------------------------------
```

Application of epiR – Titanic accident

Passenger class	Dead	Survived
Economy	528	181
Not economy	281	319

```
> epi.2by2(528,181,281,319, method="cross.sectional", conf.level=0.95)


Point estimates and 95 % CIs:
-----------------------------------------------------------------
Prevalence ratio                             1.59 (1.45, 1.75)
Odds ratio                                   3.31 (2.62, 4.18)
Attrib prevalence *                          27.64 (22.51, 32.76)
Attrib prevalence in population *            14.97 (10.19, 19.75)
Attrib fraction in exposed (%)               37.11 (30.81, 42.84)
Attrib fraction in population (%)            24.22 (19.25, 28.88)
------------------------------------------------------------------
```

Summary

RCT / prospective study	Cross-sectional study	Case-control study
Relative risk	Odds ratio	Odds ratio
Odds ratio	Prevalence ratio	
NNT	D	
D		

Optional – Bayesian analysis of 2 proportions

	Side effects	None
Drug A	11	9
Drug B	5	15

- **Are the effects the same for the 2 groups?**

Frequentist analysis

- **Let X $\sim$ Binomial(n_1, π_1) and p_1 = X / n_1**
- **Let Y $\sim$ Binomial(n_2, π_2) and p_2 = Y / n_2**
- **Consider the hypothesis $\pi_1 = \pi_2$**
- **The score statistic is:**

$$TS = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

where $\hat{p} = \frac{X+Y}{n_1+n_2}$ is the estimate of the common proportion under the null hypothesis

This statistic is normally distributed for large n_1 and n_2.

Frequentist analysis

- **p_1 = 0.55, p_2 = 5/20 = 0.25, p = 16/40 = 0.4**

Test statistic

$$\frac{.55 - .25}{\sqrt{.4 \times .6 \times (1/20 + 1/20)}} = 1.61$$

Bayesian analysis

- **Consider putting independent Beta(α_1, β_1) and Beta(α_2, β_2) priors on p_1 and p_2 respectively**
- **Then the posterior is**

$$\pi(p_1, p_2) \propto p_1^{x+\alpha_1-1}(1-p_1)^{n_1+\beta_1-1} \times p_2^{y+\alpha_2-1}(1-p_2)^{n_2+\beta_2-1}$$

- **Hence under this (potentially naive) prior, the posterior for p_1 and p_2 are independent betas**
- **The easiest way to explore this posterior is via Monte Carlo simulation**

R analysis

```
x = 11; n1 = 20; alpha1 = 1; beta1 = 1

y = 5;  n2 = 20; alpha2 = 1; beta2 = 1

p1 = rbeta(1000, x + alpha1, n - x + beta1)
p2 = rbeta(1000, y + alpha2, n - y + beta2)
rd = p2 - p1

plot(density(rd))

quantile(rd, c(.025, .975))

mean(rd)

median(rd)
```