Quantitative Reasoning

Chapter 2 : Working with categorical data

Kelvin Horia

pvokh@nus.edu.sg
Tutor
Provost office, QR
team

Where we left off

- Observational studies can ONLY establish **association between variables**
  - **For now, you can take "association between variables" to mean that there is a relationship/link between variables** → Intuitive but a little vague as well

- In this Chapter we will make precise what it means for 2 **categorical** variables to be associated in the context of observational studies.

- To do this we will need to acquire some basic understanding on rates.

****All variables in this Chapter are assumed to be categorical unless stated otherwise.**

How to compute rates?

- Suppose a researcher is investigating what is the relationship between smoking and heart disease via an observational study.

$$\text{rate (smokers } | \text{ NoHD)} = \frac{14962}{99918} \approx 0.15$$

	HD	No HD	Row Total
Smokers	38	14 962	15 000
Non-smokers	44	84 956	85 000
Column Total	82	99 918	100 000

- Compute rate(smokers), rate(smokers and HD), and rate(smokers|HD).

→ only look at B, count how many have A.

$= \frac{15000}{100000}$

$= \frac{38}{100000}$

$= \frac{38}{82} = 0.463.$

Measuring association using rates

	B	Not B	Row Total
A	w	x	$w+x$
Not A	y	z	$y+z$
Column Total	$w+y$	$x+z$	$w+x+y+z$

- Categorical variables A and B are associated to each other if

rate(A|B) ≠ rate(A | not B)

≡ rate(B|A) ≠ rate(B| not A)

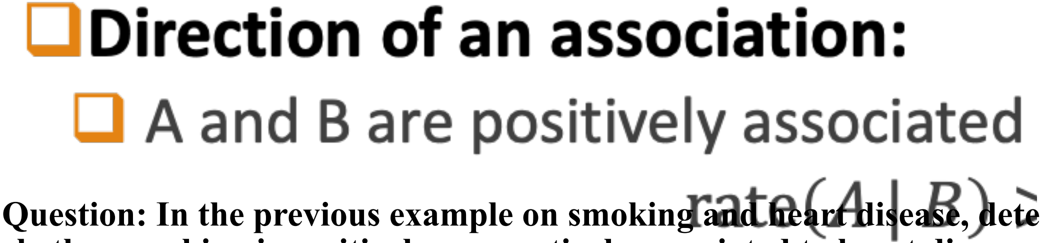
$$\text{rate}(A \mid B) \neq$$

OR

$$\text{rate}(B \mid A) \neq$$

To prove whether A and B are associated, we only need to show ONE of the inequalities holds. Why?

Positive and negative association.



❑ **Direction of an association:**

- ❑ A and B are positively associated

$\text{rate}(A \mid B) >$

- **Question: In the previous example on smoking and heart disease, determine whether smoking is positively or negatively associated to heart disease.**

Direction

- A and B positive

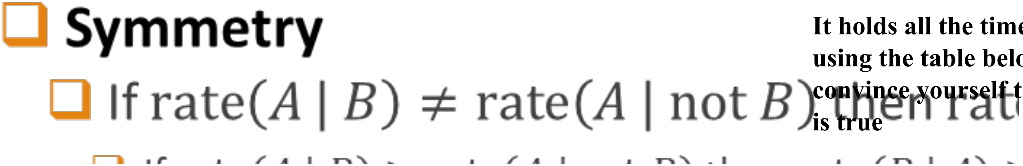
  rate(A|B) > rate(A| not B)

- negative if inverse. <.

Symmetry.

- If rate(A|B) ≠ rate(A|not B) then rate(B|A) ≠ rate(B|not A)
- If rate(A|B) > rate(A|not B) then rate(B|A) > rate(B|not A).

Symmetry of rates



□ **Symmetry**

- □ If rate(A | B) ≠ rate(A | not B) then rat
- □ If rate(A | B) > rate(A | not B) then rate(B | A) >

It holds all the time!! Try using the table below to convince yourself that it is true

	B	Not B	Row Total
A	w	x	$w + x$
Not A	y	z	$y + z$
Column Total	$w + y$	$x + z$	$w + x + y + z$

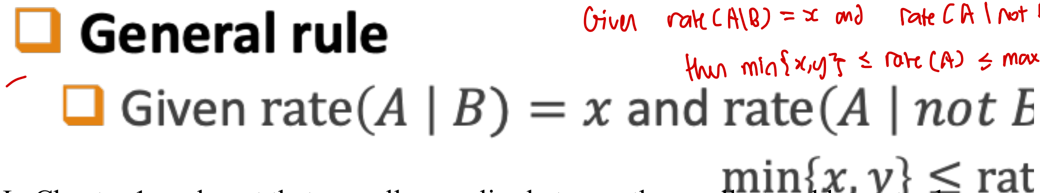
Basic rule of rates

	HD	No HD	Row Total
Smokers	38	14 962	15 000
Non-smokers	44	84 956	85 000
Column Total	82	99 918	100 000

Compare rate(H.D) with rate(H.D | smoker) and rate(H.D | non-smoker). What do you notice?

Given rate(A|B) = x and rate(A | not B) = y

then min{x,y} ≤ rate(A) ≤ max{x,y}



❑ **General rule**

❑ Given rate(A | B) = x and rate(A | not E

min{x, y} ≤ rat

- In Chapter 1 we learnt that overall mean lies between the smallest and largest subgroup means. Compare this to the basic rule of rates.

rate(A) = rate(A|B) × rate(B)

+ rate(A | not B) × rate(not B).

Confounders (revisited)

- A confounder is a third variable, **associated with both dependent and independent variables**. It must be a different variable from the dependent and independent variable.

	Males		Females		Row total
	H.D	No H.D	H.D	No H.D	
Smokers	25	9 582	13	5 380	15 000
Non-smokers	30	34 954	14	50 002	85 000
Column Total	55	40 334	27	59 584	100 000

- Using rates and the definition of association, are you able to determine whether gender is a confounder in this study?
- How does one control for confounders?

Simpson's paradox

- Relationship between rates in subgroups are reversed or "disappears" when subgroups are combined.

no association.

- SURE SIGN of confounding variable when you encounter Simpson's paradox.

- Counter by slicing.

Major	Males	Females
A	2000 applied. 40% selected	10 applied 20% selected
B	10 applied 80% selected	1000 applied 60% selected
C	10 applied 70% selected	1000 applied 50% selected

3 or more groups ⇒ look at majority's association.

Reminders about association

- When we say 2 variables are associated, it is **with reference to data that has been collected**. Whether the association can be generalized from the sample data to the population of interest goes back to what we learnt in Chapter 1 about sampling and generalizability criteria.

- Just showing association between 2 variables doesn't mean much. There are LOTS AND LOTS of variables to which we can do up 2 by 2 table and show association but remember **association does not prove causation.**

- This is why for observational studies, researchers must painstakingly control for as many confounders as possible. The more external variables they can control for, and if association is still present, it's regarded as having more "evidence" that there maybe a genuine relationship between the variables.