Lesson 4: Measurements of Association and Agreement

Nicky Wakim

2025-04-07

Poll Everywhere Question 1

Learning Objectives

1. Identify cases when it is appropriate to use risk difference, relative risk, or odds ratios
2. Expand work on contingency tables to evaluate the agreement or reproducibility using Cohen's Kappa

Last class

- Used contingency tables to test and measure association between two variables
  - Categorical outcome variable (Y)
  - One categorical explanatory variable (X) with groups 1 and 2
- We looked at risk difference, risk ratio, and odds ratio to measure association

Measure	Formula for Estimate
Risk difference	$\widehat{RD} = \widehat{p}_1 - \widehat{p}_1 = \frac{n_{11}}{n_1} - \frac{n_{21}}{n_2}$
Relative risk / risk ratio	$\widehat{RR} = \frac{\widehat{p}_1}{\widehat{p}_2} = \frac{n_{11}/n_1}{n_{21}/n_2}$
Odds ratio	$\widehat{OR} = \frac{\widehat{\text{odds}}_1}{\widehat{\text{odds}}_2} = \frac{\widehat{p}_1/(1-\widehat{p}_1)}{\widehat{p}_2/(1-\widehat{p}_2)}$

- Discussed how OR will be an important measurement in logistic regression

Our example with Strong Heart Study

Glucose tolerance	Diabetes		Total
	No	Yes	
Impaired	334	198	532
Normal	1004	128	1132
Total	1338	326	1664

Measure	Formula	Interpretation
Risk difference	$\widehat{RD} = \widehat{p}_1 - \widehat{p}_1$	The diabetes diagnosis risk difference between impaired and normal glucose tolerance is 0.2591 (95% CI: 0.2141, 0.3041).
Relative risk / risk ratio	$\widehat{RR} = \dfrac{\widehat{p}_1}{\widehat{p}_2} = \dfrac{n_{11}/n_1}{n_{21}/n_2}$	The estimated risk of diabetes for American Indians with impaired glucose is 3.29 times the with normal glucose tolerance (95% CI: 2.70, 4.01).
Odds ratio	$\widehat{OR} = \dfrac{\widehat{\text{odds}}_1}{\widehat{\text{odds}}_2} = \dfrac{\widehat{p}_1/(1-\widehat{p}_1)}{\widehat{p}_2/(1-\widehat{p}_2)}$	The estimated odds of diabetes for American Indians with impaired glucose tolerance is 4.65 times the odds for American Indians with normal glucose tolerance.

Learning Objectives

1. Identify cases when it is appropriate to use risk difference, relative risk, or odds ratios

2. Expand work on contingency tables to evaluate the agreement or reproducibility using Cohen's Kappa

Relationship Between RR and OR (1/2)

- Notice that odds ratio is **not equivalent** to relative risk (or risk ratio)

- However, when the **probability of "success" is small** (e.g., rare disease), $\widehat{OR}$ is a nice approximation of $\widehat{RR}$

$$\widehat{OR} = \frac{\widehat{p}_1/(1-\widehat{p}_1)}{\widehat{p}_2/(1-\widehat{p}_2)} = \widehat{RR} \cdot \frac{1-\widehat{p}_2}{1-\widehat{p}_1}$$

  - The fraction in the last term of the above expression approximately equals to 1.0 if $\widehat{p}_1$ and $\widehat{p}_2$ BOTH quite small (< 0.1)

- The $\widehat{OR}$ and $\widehat{RR}$ are not very close to each other in SHS: diabetes not a rare disease
  - $\widehat{OR} = 4.65$
  - $\widehat{RR} = 3.29$

Relationship Between RR and OR (2/2)

- An example where a disease rare over the whole sample (~1%), but …
  - $\widehat{OR}$ is not a good estimate of $\widehat{RR}$ in "rare" disease

Risk factor status	Disease Status		Total	Risk
	Disease	No disease		
Exposed	9	9	18	0.5
Not exposed	1	981	982	0.00102
Total	10	990	1000	0.01010

- $\widehat{p}_1$ is 0.5: thus $\widehat{OR}$ and $\widehat{RR}$ are very different

$$\widehat{RR} = \frac{0.5}{0.00102} = 490 \text{ and } \widehat{OR} = \frac{0.5(1-0.5)}{0.00102(1-0.00102)} = 981$$

Poll Everywhere Question 2

RR in retrospective case-control study (1/3)

- In retrospective case-control studies: we **identify cases** (patients with the outcome), then select a number of controls (patients without the outcome)
  - Case-control study to require **much smaller sample size** than equivalent cohort studies
  - So we pick out the cases and controls first, **then see if there is exposure**

- However, the **proportion of cases in the sample** **does not** represent the **proportion of cases in the population**
  - RR compares probability of the outcome (case) for exposed and unexposed groups
  - Number of outcomes has been artificially inflated for case-control study

RR in retrospective case-control study (2/3)

- Assume a 1:2 case-control study summarized in below table:

Group	Outcome		Total
	Case	Control	
Exposed	40 (n_{11})	40 (n_{12})	80 (n_{1+})
Not exposed	60 (n_{21})	160 (n_{22})	220 (n_{2+})
Total	100 (n_{+1})	200 (n_{+2})	300 (n)

- Assume we compute the RR as if it is from a cohort study:

$$\widehat{RR} = \frac{\widehat{p_1}}{\widehat{p_2}} = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}} = \frac{40/80}{60/220} = 1.8333$$

RR in prospective cohort study

- In real world, the proportion of controls (not diseased) is typically much higher. Assume the table below shows the **proportion in the population in a cohort study**

Group	Outcome		Total
	Case	Control	
Exposed	400	4000	4400
Not exposed	600	16000	16600
Total	1000	20000	21000

- The estimated RR for the patient population is:

$$\widehat{RR} = \frac{\widehat{p_1}}{\widehat{p_2}} = \frac{400/4400}{600/16600} = 2.5152$$

Notes for Odds Ratios

- The OR is valid for
  - Case-control studies (where the RR is not appropriate)
  - Prospective cohort studies
  - Cross-sectional studies

- It can be interpreted either as…
  - Odds of **event** for exposed vs. unexposed individuals, or
  - Odds of **exposure** for individuals with vs. without the event of interest

- Pay attention to the numerator and denominator for the OR

OR in retrospective case-control study

- While we cannot estimate RR from a case-control study, we can still estimate OR for case-control study
  - OR does not require us to distinguish between the outcome variable and explanatory variable in the contingency table
    - AKA: Odds ratio of disease comparing exposed to not exposed is same as odds ratio of being exposed comparing diseased and not diseased

- For case-control study where the probability of having outcome is small, the $\widehat{OR}$ is a nice approximation to $\widehat{RR}$
  - For the 1:2 case-control table: $\widehat{OR} = \frac{40 \cdot 160}{40 \cdot 60} = 2.667$
  - Population cohort study: $\widehat{RR} = 2.5152$

Group	Outcome		Total
	Case	Control	
Exposed	40 (n_{11})	40 (n_{12})	80 (n_{1+})
Not exposed	60 (n_{21})	160 (n_{22})	220 (n_{2+})
Total	100 (n_{+1})	200 (n_{+2})	300 (n)

Which measurement should one use?

Measurement	Pros and Cons
Risk difference	• Can provide additional information, but can be misleading on its own • Not the preferred measurement
Risk ratio	• Easy to interpret because is a ratio of probabilities • Cannot use in retrospective, case-control studies
Odds ratio	• Adequate for all studies • Good estimate of RR for rare diseases • Not as easy to interpret or translate to clinical setting as RR • **Most preferred by statisticians because integrated into logistic regression**

Learning Objectives

1. Identify cases when it is appropriate to use risk difference, relative risk, or odds ratios

2. Expand work on contingency tables to evaluate the agreement or reproducibility using Cohen’s Kappa

Measuring Agreement

- Still within the realm of contingency tables
- What if we are NOT looking at the association between two variables?

- What if we want to look at the agreement between two things?
  - Answers of same subjects for same survey taken at different times
  - Two different radiologists' assessment of the same X-ray

- Cohen's Kappa statistics: widely used as a measure of agreement
  - Example: Reliability studies, interobserver agreement

Poll Everywhere Question 3

Let’s get our mood data down!

Monday’s response	Wednesday’s response		Total
	Good	Bad	
Good			
Bad			
Total			

Measuring Agreement

- If **perfect agreement** among the two raters/surveys:
  - We would expect nonzero entries only in the diagonal cells of the table

- p_o is the observed proportion of complete agreement (concordance)
- p_E is the expected proportion of complete agreement if the agreement is just due to chance
- If the p_o is much greater than p_E, then the agreement level is high.
  - Otherwise, the agreement level is low

- **Cohen's Kappa** is based on the difference between p_o and p_E:

$$\widehat{\kappa} = \frac{p_o - p_E}{1 - p_E}$$

- $\widehat{\kappa} = 0$: No agreement between surveys/raters other than what would be expected by chance
- $\widehat{\kappa} = 1$: Complete agreement

Measuring Agreement: Cohen's Kappa

- Point estimate:

$$\widehat{\kappa} = \frac{p_o - p_E}{1 - p_E}$$

  - With $p_o = \frac{\sum_i n_{ii}}{n}$ (sum of diagonals divided by total)
  - With $p_E = \sum_i a_i b_i$
  - With range of point estimate from $[-1, 1]$

> **What's $\sum_i a_i b_i$?**
>
> For i responses (row/columns), a_i is proportion of i response category in first survey and b_i is proportion of i response category in second survey (we'll show this in the example)

- Approximate standard error:

$$SE_{\widehat{\kappa}} = \sqrt{\frac{1}{n\left(1 - p_e\right)^2}\left\{p_e^2 + p_e - \sum_i \left[a_i b_i \left(a_i + b_i\right)\right]\right\}}$$

- 95% Wald confidence interval for $\widehat{\kappa}$:

$$\widehat{\kappa} \pm 1.96 \cdot SE_{\widehat{\kappa}}$$

Example: Our moods (1/3)

Agreement of surveys

Compute the point estimate and 95% confidence interval for the agreement between our Monday and Wednesday moods.

Monday's response	Wednesday's response		Total
	Good	Bad	
Good			
Bad			
Total			

Needed steps:

1. Compute the kappa statistic
2. Find confidence interval of kappa
3. Interpret the estimate

Example: Our moods (2/3)

Agreement of surveys

Compute the point estimate and 95% confidence interval for the agreement between our Monday and Wednesday moods.

Monday's response	Wednesday's response		Total
	Good	Bad	
Good			
Bad			
Total			

Needed steps:

1/2. Compute the kappa statistic and find confidence interval of kappa

```
1 library(epiR)
2 moods = matrix(c(100, 40, 10, 30), nrow = 2, byrow = T)
3 moods
```

```
     [,1] [,2]
[1,]  100   40
[2,]   10   30
```

```
1 epi.kappa(moods, method = "cohen")$kappa
```

```
        est         se     lower     upper
1 0.3661972 0.07617362 0.2168996 0.5154947
```

Example: Our moods (3/3)

Agreement of surveys

Compute the point estimate and 95% confidence interval for the agreement between our Monday and Wednesday moods.

Monday's response	Wednesday's response		Total
	Good	Bad	
Good			
Bad			
Total			

Needed steps:

3. Interpret the estimate

The kappa statistic is _____ (95% CI: _____, _____), indicating _____ agreement.

Since the 95% confidence interval does/does not contain 0, we have/do not have sufficient evidence that there is _____ agreement between our mood on Monday and our mood on Wednesday.

Measuring Agreement: Oberved Kappas

- Guidelines for evaluating Kappa (Rosner TB)
  - **Excellent agreement** if $\widehat{\kappa} \geq 0.75$
  - **Fair to good agreement** if $0.4 < \widehat{\kappa} < 0.75$
  - **Poor agreement** if $\widehat{\kappa} \leq 0.4$

If $\widehat{\kappa} < 0$, suggest agreement less than by chance

Learning Objectives

1. Identify cases when it is appropriate to use risk difference, relative risk, or odds ratios
2. Expand work on contingency tables to evaluate the agreement or reproducibility using Cohen's Kappa

Measurement of Association So Far

- Used contingency tables to test and measure association between two variables
  - Categorical outcome variable (Y)
  - One categorical explanatory variable (X)
- We looked at risk difference, risk ratio, and odds ratio to measure association
- Such an association is called crude association
  - No adjustment for possible confounding factors
  - Also called marginal association
- But we cannot expand analysis based on contingency tables past 3 variables
  - We can get into stratified contingency tables to bring in a 3rd variable
  - But I don't think it's worth it because regression can bring in (adjust for) many variables

Extra example in case the mood example fails beautifully

Just in case our data doesn't work out: Beef Consumption in Survey

A diet questionnaire was mailed to 537 female American nurses on two separate occasions several months apart. The questions asked included the quantities eaten of more than 100 separate food items. The data from the two surveys for the amount of beef consumption are presented in the below table. How can reproducibility of response for the beef-consumption data be quantified?

Survey 1	Survey 2		Total
	≤ 1 serving/week	> 1 serving/week	
≤ 1 serving/week	136	92	228
> 1 serving/week	69	240	309
Total	205	332	537

Example: Beef Consumption in Survey (1/3)

Agreement of surveys

Compute the point estimate and 95% confidence interval for the agreement between beef consumption surveys. Similar to question: Are results reproducible for the beef-consumption in the survey?

Survey 1	Survey 2		Total
	≤ 1 serving/week	> 1 serving/week	
≤ 1 serving/week	136	92	228
> 1 serving/week	69	240	309
Total	205	332	537

Needed steps:

1. Compute the kappa statistic
2. Find confidence interval of kappa
3. Interpret the estimate

Example: Beef Consumption in Survey (2/3)

Agreement of surveys

Compute the point estimate and 95% confidence interval for the agreement between beef consumption surveys. Similar to question: Are results reproducible for the beef-consumption in the survey?

Survey 1	Survey 2		Total
	≤ 1 serving/week	> 1 serving/week	
≤ 1 serving/week	136	92	228
> 1 serving/week	69	240	309
Total	205	332	537

Needed steps:

1/2. Compute the kappa statistic and find confidence interval of kappa

```
1 library(epiR)
2 beef = matrix(c(136, 92, 69, 240), nrow = 2, byrow = T)
3 epi.kappa(beef, method = "cohen")$kappa
```

```
        est         se     lower     upper
1 0.3781906 0.04100635 0.2978196 0.4585616
```

Example: Beef Consumption in Survey (3/3)

Agreement of surveys

Compute the point estimate and 95% confidence interval for the agreement between beef consumption surveys. Similar to question: Are results reproducible for the beef-consumption in the survey?

Survey 1	Survey 2		Total
	≤ 1 serving/week	> 1 serving/week	
≤ 1 serving/week	136	92	228
> 1 serving/week	69	240	309
Total	205	332	537

Needed steps:

3. Interpret the estimate

The kappa statistic is 0.378 (95% CI: 0.298, 0.459), indicating fair agreement.

Since the 95% confidence interval does not contain 0, we have sufficient evidence that there is fair agreement between the surveys for beef consumption. The survey is not reliably reproducible since we did not achieve excellent agreement.