STAT475/675 TUT03

http://www.sfu.ca/~zza115/teaching.html
http://people.stat.sfu.ca/~joanh/stat475-675web.html

Zhiyang Zhou (zhiyang_zhou@sfu.ca)

2018-01-23

Two Bernoulli r.v.s X and Y

- For $i, j = 1, 2$,
  - joint probability:
    - $\Pr(X = i, Y = j) = \pi_{ij}$
  - marginal probablity:
    - $\Pr(X = i) = \pi_{i1} + \pi_{i2} = \pi_{i+}$
    - $\Pr(Y = j) = \pi_{1j} + \pi_{2j} = \pi_{+j}$
  - conditional probability:
    - $\Pr(X = i|Y = j) = \pi_{ij}/\pi_{+j}$
    - $\Pr(Y = j|X = i) = \pi_{ij}/\pi_{i+}$
  - $X \perp\!\!\!\perp Y$
    - $\Leftrightarrow \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i, j
    - $\Leftrightarrow \Pr(X = i|Y = j) = \pi_{i+}$ for all i, j
    - $\Leftrightarrow \Pr(Y = j|X = i) = \pi_{+j}$ for all i, j

Joint study on two Bernoulli r.v.s

- Example

Probabilities

	Disease (Y)		
Group (X)	Y (1)	N (2)	
M (1)	π_{11}	π_{12}	π_{1+}
F (2)	π_{21}	π_{22}	π_{2+}
	π_{+1}	π_{+2}	1

- Interested in
  - How to compare $p_1 = \Pr(\text{disease in M}|\text{M})$ and $p_2 = \Pr(\text{disease in F}|\text{F})$?
    - $p_1 = \Pr(\text{disease in M}|\text{M}) = \pi_{11}/\pi_{1+}$
    - $p_2 = \Pr(\text{disease in F}|\text{F}) = \pi_{21}/\pi_{2+}$

Difference

$$p_1 - p_2 = \frac{\pi_{11}}{\pi_{1+}} - \frac{\pi_{21}}{\pi_{2+}}$$

- Remarks
  - fail to indicate the magnitude of probabilities:
    - e.g., see $.02 - .01 = .99 - .98$

Relative risk (RR)

$$RR = \frac{p_1}{p_2} = \frac{\pi_{11}/\pi_{1+}}{\pi_{21}/\pi_{2+}}$$

- Remarks
  - highly depend on the magnitude of denominator
  - fail to reflect the magnitude of difference (esp. for rare disease)
    - e.g., .0002/.0001 = 2 > 1.25 = .25/.2
  - $X \perp\!\!\!\perp Y \Leftrightarrow RR = 1$

Odds ratio (OR)

- Odds in male-group

$$odds_1 = \frac{p_1}{1-p_1} = \frac{\pi_{11}/\pi_{1+}}{1-\pi_{11}/\pi_{1+}} = \frac{\pi_{11}}{\pi_{12}}$$

- Odds in female-group

$$odds_2 = \frac{p_2}{1-p_2} = \frac{\pi_{21}/\pi_{2+}}{1-\pi_{21}/\pi_{2+}} = \frac{\pi_{21}}{\pi_{22}}$$

- Odds ratio

$$OR = \frac{odds_1}{odds_2} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

- Remarks
  - $OR \approx RR$ when $\pi_{11} \ll \pi_{12}$ and $\pi_{21} \ll \pi_{22}$ (for rare disease)
  - $X \perp\!\!\!\perp Y \Leftrightarrow OR = 1$

Contingency table

- Outcomes of n iid trials: $\{(x_k, y_k) : k = 1, \dots, n\}$
- Tabulate the data by 2×2 (two-way) contingency table

	$y = 1$	$y = 2$	
$x = 1$	n_{11}	n_{12}	n_{1+}
$x = 2$	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	n

- I»J Contingency Table
  - a table with cells containing frequency counts of outcome according to 2 categorical variables with respective I and J levels
- Probability Models
  - binomial sampling (given row total $N_{i+} = n_{i+}$): $N_{i1} \sim \text{Binom}(n_{i+}, p_i)$ with $p_i = \pi_{i1}/\pi_{i+}$
  - multinomial sampling (given $N = n$): $(N_{11}, N_{12}, N_{21}, N_{22}) \sim \text{multinomial}(n; \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$
  - hypergeometric sampling (given n_{1+}, n_{2+}, n_{+1}, and n_{+2}): $N_{11} \sim \text{hypergeometric}(n, n_{+1}, n_{1+})$

Inference on probabilities

- MLE (with binomial or multinomial sampling)
  - $\hat{\pi}_{ij} = n_{ij}/n$ for all i, j
- Estimate other probabilities by plug-in (i.e., substituting $\hat{\pi}_{ij}$ for π_{ij})

Inference on difference

- MLE

$$\hat{p}_1 - \hat{p}_2 = \frac{\hat{\pi}_{11}}{\hat{\pi}_{1+}} - \frac{\hat{\pi}_{21}}{\hat{\pi}_{2+}} = \frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}$$

- Estimator for variance of MLE

$$\hat{\sigma}^2_{\hat{p}_1-\hat{p}_2} = \widehat{\text{var}}(\hat{p}_1 - \hat{p}_2) = \frac{\hat{p}_1(1-\hat{p}_1)}{n_{1+}} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_{2+}} = \frac{n_{11}n_{12}}{(n_{11}+n_{12})^3} + \frac{n_{21}n_{22}}{(n_{21}+n_{22})^3}$$

- Approximate $(1-\alpha)$ CI of $\ln(RR)$
  - Wald

$$\hat{p}_1 - \hat{p}_2 \pm Z_{1-\alpha/2}\hat{\sigma}_{\hat{p}_1-\hat{p}_2}$$

  - Agresti-Caffo

$$\tilde{p}_1 - \tilde{p}_2 \pm Z_{1-\alpha/2}\tilde{\sigma}_{\tilde{p}_1-\tilde{p}_2}$$

  with

$$\tilde{p}_i = \frac{n_{i1}+1}{n_{i+}+2}$$

$$\tilde{\sigma}^2_{\tilde{p}_1-\tilde{p}_2} = \widehat{\text{var}}(\tilde{p}_1 - \tilde{p}_2) = \frac{\tilde{p}_1(1-\tilde{p}_1)}{n_{1+}+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_{2+}+2} = \frac{(n_{11}+1)(n_{12}+1)}{(n_{11}+n_{12}+2)^3} + \frac{(n_{21}+1)(n_{22}+1)}{(n_{21}+n_{22}+2)^3}$$

- Remark
  - Wald confidence interval has a smaller true confidence level than what it states
  - Agresti-CaffoAdd performs well and avoid 0 cell count

Inference on RR

- MLE

$$\widehat{RR} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{\hat{\pi}_{11}/\hat{\pi}_{1+}}{\hat{\pi}_{21}/\hat{\pi}_{2+}} = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}$$

  - Interpretation: The estimated probability of sth. is $\widehat{RR}$ times as large as in group 1 than in group 2.

- Estimator for variance of MLE

$$\hat{\sigma}^2_{ln\widehat{RR}} = \widehat{\text{var}}(\ln(\widehat{RR})) = \frac{1-\hat{p}_1}{n_{1+}\hat{p}_1} + \frac{1-\hat{p}_2}{n_{2+}\hat{p}_2} = n_{11}^{-1} - n_{1+}^{-1} + n_{21}^{-1} - n_{2+}^{-1}$$

- Approximate $(1-\alpha)$ CI of $\ln(RR)$

$$\ln(\widehat{RR}) \pm Z_{1-\alpha/2}\hat{\sigma}_{ln\widehat{RR}}$$

- Approximate $(1-\alpha)$ CI of RR

$$\exp\{\ln(\widehat{RR}) \pm Z_{1-\alpha/2}\hat{\sigma}_{ln\widehat{RR}}\} = (\widehat{RR}\exp\{-Z_{1-\alpha/2}\hat{\sigma}_{ln\widehat{RR}}\}, \widehat{RR}\exp\{Z_{1-\alpha/2}\hat{\sigma}_{ln\widehat{RR}}\})$$

- Remark
  - Add a small constant, such as 0.5, to the 0 cell count and the corresponding row total

Inference on OR

- MLE

$$\widehat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}}$$

  - Interpretation: The estimated odds of sth. are $\widehat{OR}$ times as large as in group 1 than in group 2.
- Estimator for variance of MLE

$$\hat{\sigma}^2_{ln\widehat{OR}} = \widehat{\text{var}}(\ln(\widehat{OR})) = \sum_{i,j} n_{ij}^{-1}$$

- Approximate $(1-\alpha)$ CI of $\ln(OR)$

$$\ln(\widehat{OR}) \pm Z_{1-\alpha/2}\hat{\sigma}_{ln\widehat{OR}}$$

- Approximate $(1-\alpha)$ CI of OR

$$\exp\{\ln(\widehat{OR}) \pm Z_{1-\alpha/2}\hat{\sigma}_{ln\widehat{OR}}\} = (\widehat{OR}\exp\{-Z_{1-\alpha/2}\hat{\sigma}_{ln\widehat{OR}}\}, \widehat{OR}\exp\{Z_{1-\alpha/2}\hat{\sigma}_{ln\widehat{OR}}\})$$

- Remark
  - Add a small constant, such as 0.5, to the 0 cell count and the corresponding row total

Exercise (Textbook Ch1 Q17)

Before a placekicker attempts a field goal in a pressure simulation, the opposing team may call a time-out to give the kicker more time to think about it in the hopes that this extra time will cause him to become more nervous and lower the probability of his success. This strategy is called "icing the kicker".

	Success	Failure	Total
No time-out	22	4	26
Time-out	10	6	16
Total	32	10	42

- Compute 95% Wald and Agresti-Caffo CI for difference in probabilities of success conditioning on the strategy
- Estimate the relative risk and the corresponding CI.
- Estimate the odds ratio and the corresponding CI.
- Is there sufficient evidence to conclude that icing the kicker is a good strategy to follow? Explain.