STAT475/675 TUT04

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

Zhiyang Zhou (zhiyang_zhou@sfu.ca)

2018-01-30

Recall

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

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

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

- Odds ratio (OR)

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

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

- Probability Models
  - binomial sampling (given row total $N_{i+} = n_{i+}$)
  - multinomial sampling (given $N = n$)
- MLE
  - $\widehat{RR} = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}$
  - $\widehat{OR} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}}$
- Estimating $\text{var}(\ln \widehat{RR})$ and $\text{var}(\ln \widehat{OR})$
  - $\hat{\sigma}^2_{\ln \widehat{RR}} = \widehat{\text{var}}(\ln \widehat{RR}) = n_{11}^{-1} - n_{1+}^{-1} + n_{21}^{-1} - n_{2+}^{-1}$
  - $\hat{\sigma}^2_{\ln \widehat{OR}} = \widehat{\text{var}}(\ln \widehat{OR}) = \sum_{i,j} n_{ij}^{-1}$
- Approximate $(1-\alpha)$ Wald CIs of $\ln RR$ and $\ln OR$
  - $\ln \widehat{RR} \pm Z_{1-\alpha/2}\hat{\sigma}_{\ln \widehat{RR}}$
  - $\ln \widehat{OR} \pm Z_{1-\alpha/2}\hat{\sigma}_{\ln \widehat{OR}}$

- Approximate $(1-\alpha)$ CIs of RR and OR
  - $(\widehat{RR}\exp\{-Z_{1-\alpha/2}\hat{\sigma}_{\ln \widehat{RR}}\}, \widehat{RR}\exp\{Z_{1-\alpha/2}\hat{\sigma}_{\ln \widehat{RR}}\})$
  - $(\widehat{OR}\exp\{-Z_{1-\alpha/2}\hat{\sigma}_{\ln \widehat{OR}}\}, \widehat{OR}\exp\{Z_{1-\alpha/2}\hat{\sigma}_{\ln \widehat{OR}}\})$

Hypothesis testing on independence

- $H_0 : X \perp\!\!\!\perp Y$ vs $H_1 : X \not\perp\!\!\!\perp Y$
  - or, equivalently, $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i, j vs $H_1 : \pi_{ij} \neq \pi_{i+}\pi_{+j}$ for some i, j
- 2×2
  - $X \perp\!\!\!\perp Y \Leftrightarrow RR = 1 \Leftrightarrow OR = 1$
  - $H_0 : RR = 1$ vs $H_1 : RR \neq 1$ (or $H_0 : OR = 1$ vs $H_1 : OR \neq 1$)
  - estimate RR (or OR) and corresponding $(1-\alpha)$ CI
  - reject H_0 if CI doesn't cover 1 at the significance level α and otherwise do not reject H_0 at the significance level of α
- $I \times J$
  - RR and OR are unapplicable
  - general testing procedures
    * Pearson's chi-squared (χ^2-) test (K. Pearson, 1900)
    * likelihood ratio test (LRT)

χ^2-test (with purposive or multinomial sampling)

- $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i, j vs $H_1 : \pi_{ij} \neq \pi_{i+}\pi_{+j}$ for some i, j
- Test statistic:
$$\chi^2_{\text{obs}} = \sum_{i,j} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$
  with $\hat{\mu}_{ij} = n_{i+}n_{+j}/n$
- Caculate p-value $= \Pr_{H_0}(\chi^2 \geq \chi^2_{\text{obs}})$ based on $\chi^2 \sim \chi^2(df)$ approximately under H_0 with $df = (I-1)(J-1)$
- Draw conclusion
  - p-value $< \alpha$: there is a strong evidence against H_0, i.e., there is a signficant association between X and Y
- Remark
  - The χ^2-approximation is good usually when $\mu_{ij} \geq 5$ for all i, j.

LRT (with purposive or multinomial sampling)

- $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i, j vs $H_1 : \pi_{ij} \neq \pi_{i+}\pi_{+j}$ for some i, j
- Test statistic:
$$G^2_{\text{obs}} = 2\sum_{i,j} n_{ij} \ln \frac{n_{ij}}{\hat{\mu}_{ij}}$$
  with $\hat{\mu}_{ij} = n_{i+}n_{+j}/n$
- Caculate p-value $= \Pr_{H_0}(G^2 \geq G^2_{\text{obs}})$ based on $G^2 \sim \chi^2(df)$ approximately under H_0 with $df = (I-1)(J-1)$
- Draw conclusion

    - p-value $< \alpha$: there is a strong evidence against H_0, i.e., there is a signficant association between X and Y

- Remark
    - The χ^2-approximation is good usually when $\mu_{ij} \geq 5$ for all i, j.

Exercise

The table below was taken from the 2002 General Social Survey.

		Party Identification			
		democrat	independent	republican	Total
	White	871	444	873	2188
Race	Black	302	80	43	425
	Total	1173	524	916	2613

- Conduct the Pearson's χ^2- test for independence bewtween party identificaiton and race and interpret the result.
- Conduct the LRT for independence bewtween party identificaiton and race and interpret the result.

Matched pairs data

		Method I		
		Success	Failure	Total
	Success	4	6	10
Method II	Failure	3	3	6
	Total	7	9	16

- Probability model: $(N_{11}, N_{12}, N_{21}, N_{22}) \sim \text{multinom}(n; \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$
- Interested in $\pi_{1+} - \pi_{+1}$
    - $\pi_{1+} = \pi_{11} + \pi_{12}$: probability of success of Method I
    - $\pi_{+1} = \pi_{11} + \pi_{21}$: probability of success of Method II
- MLE for $\pi_{1+} - \pi_{+1}$

$$\hat{\pi}_{1+} - \hat{\pi}_{+1} = \frac{n_{12} - n_{21}}{n}$$

- Estimating $\text{var}(\hat{\pi}_{1+} - \hat{\pi}_{+1})$

$$\widehat{\text{var}}(\hat{\pi}_{1+} - \hat{\pi}_{+1}) = \frac{\hat{\pi}_{12} + \hat{\pi}_{21} + (\hat{\pi}_{12} - \hat{\pi}_{21})^2}{n} = \frac{n_{12} + n_{21}}{n^2} + \frac{(n_{12} - n_{21})^2}{n^3}$$

- Confidence interval
    - Wald:

$$\hat{\pi}_{1+} - \hat{\pi}_{+1} \pm Z_{1-\alpha/2} \sqrt{\widehat{\text{var}}(\hat{\pi}_{1+} - \hat{\pi}_{+1})}$$

        * diffpropci.Wald.mp() from the PropCIs package

  - Agresti-Min: add 0.5 to each cell in the table and compute a Wald confidence interval using these adjusted counts
    * diffpropci.mp() from the PropCIs package

- Test on $H_0 : \pi_{1+} = \pi_{+1}$ vs $H_1 : \pi_{1+} \neq \pi_{+1}$
  - Wald
    * test statistic
$$Z_{\text{obs}} = \frac{\hat{\pi}_{1+} - \hat{\pi}_{+1}}{\sqrt{\widehat{\text{var}}(\hat{\pi}_{1+} - \hat{\pi}_{+1})}}$$
    * reject H_0 if $|Z_{\text{obs}}| > Z_{1-\alpha/2}$
  - McNemar
    * simplify the denominator of Z_{obs} to $n^{-1}\sqrt{n_{12} + n_{21}}$
    * mcnemar.test()

Simple logistic regression

- binary response Y (e.g. success (1)/failure (0)) with one explanatory variable X
- $Y|X = x \sim \text{Bernoulli}(\pi(x))$
- Model
$$\text{logit}(\pi(x)) = \ln \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x$$
  or, equivalently,
$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \in [0, 1]$$
  - $\pi(x) \approx \alpha + \frac{\pi(x)}{x} x = \alpha + \beta x \pi(x)(1 - \pi(x))$
  - $X \perp\!\!\!\perp Y$ if $\beta = 0$
  - $\beta = \ln OR$ if X is binary, e.g male(1)/female (0), and OR is the odds ratio of success with male vs female
- Inference
  - estimate α, β and $\pi(x)$
    * MLE $\hat{\alpha}$, $\hat{\beta}$ and
$$\hat{\pi}(x) = \text{logit}^{-1}(\hat{\alpha} + \hat{\beta}x) = \frac{\exp(\hat{\alpha} + \hat{\beta}x)}{1 + \exp(\hat{\alpha} + \hat{\beta}x)}$$
  - CI for α, β and $\pi(x)$
    * estimate asymptotic variances $AV_{\hat{\alpha}}$ and $AV_{\hat{\beta}}$
    * 95% Wald CI
      · $\hat{\alpha} \pm 1.96 ASE_{\hat{\alpha}}$ with $ASE_{\hat{\alpha}} = \sqrt{AV_{\hat{\alpha}}}$
      · $\hat{\beta} \pm 1.96 ASE_{\hat{\beta}}$ with $ASE_{\hat{\beta}} = \sqrt{AV_{\hat{\beta}}}$
      · $\text{logit}^{-1}((\hat{\alpha} + \hat{\beta}x) \pm 1.96 ASE_{\hat{\alpha}+\hat{\beta}x})$ with $ASE_{\hat{\alpha}+\hat{\beta}x} = \sqrt{AV_{\hat{\alpha}} + x^2 AV_{\hat{\beta}} + 2x\widehat{\text{Cov}}(\hat{\alpha}, \hat{\beta})}$
  - test on on $H_0 : \beta = \beta_0$ vs $H_1 : \beta \neq \beta_0$
    * Wald: $Z = (\hat{\beta} - \beta_0)/ASE_{\hat{\beta}} \sim N(0, 1)$

Exercise

The failure of an O-ring on the space shuttle Challenger's booster rockets led to its destruction in 1986.

- Use logistic regression to model the probability of an O-ring failure as a function of temperature at launch. Data is included in "orings" in package faraway.

Review for Midterm I

- One Bernoulli r.v.
  - inference on probability of success π
    - MLE for π
    - CIs for π
      - Wald, Wlison (score type), Agresti-Coull and Clopper-Pearson (exact)
      - interpretation
      - comparison: true confidence level
- Two Bernoulli r.v.s and 2×2 contingency table
  - interested in RR (or OR) (with binomial or multinomial sampling)
    - MLE
      - interpretation
    - $\widehat{\text{var}}(\ln \widehat{RR})$ (or $\widehat{\text{var}}(\ln \widehat{OR})$)
    - approximate $(1-\alpha)$ Wald CI of $\ln RR$ (or $\ln OR$)
    - approximate $(1-\alpha)$ Wald CI of RR (or OR)
      - interpretation
  - interested in $\pi_{1+} - \pi_{+1}$ (for matched pairs data, with multinomial sampling)
    - MLE
    - $\widehat{\text{var}}(\hat{\pi}_{1+} - \hat{\pi}_{+1})$
    - CI
      - Wald
      - Agresti-Min
    - Test on $H_0 : \pi_{1+} = \pi_{+1}$ vs $H_1 : \pi_{1+} \neq \pi_{+1}$
      - Wald
      - McNemar
- Testing on independence of two r.v.s
  - 2×2 contingency table
    - test on whether RR (or OR) equals 1
  - $I \times J$ contingency table (with purposive or multinomial sampling)
    - Pearson's χ^2-test
    - likelihood ratio test
- Simple logistic regression