HWA1, Analysis of Categorical Data

September 22, 2025

1. (2pt) Suppose that we have observed a 2×2 table with independent binomial sampling with fixed row totals n_{i+} for $i = 1, 2$. Let $\pi_i = P(\text{column } 1 \mid \text{row } i)$ and n_i be the observed frequency in column 1 for $i = 1, 2$.

   (a) Find the MLE of π_1 and π_2.
   Solution: The likelihood function is

$$\begin{aligned} L(\pi_1, \pi_2) &= \binom{n_{1+}}{n_{11}} \pi_1^{n_{11}} (1-\pi_1)^{n_{1+}-n_{11}} \binom{n_{2+}}{n_{21}} \pi_2^{n_{21}} (1-\pi_2)^{n_{2+}-n_{21}} \\ &= \binom{n_{1+}}{n_{11}} \binom{n_{2+}}{n_{21}} \exp\{n_{11} \log \pi_1 + (n_{1+} - n_{11}) \log(1-\pi_1) + n_{21} \log \pi_2 \\ &\quad + (n_{2+} - n_{21}) \log(1-\pi_2)\} \end{aligned}$$

   The first order derivatives of the log-likelihood are

$$\begin{aligned} \frac{\partial \ell(\pi_1, \pi_2)}{\partial \pi_1} &= \frac{n_{11}}{\pi_1} - \frac{n_{1+} - n_{11}}{1-\pi_1} = \frac{n_{11} - n_{1+}\pi_1}{\pi_1(1-\pi_1)}, \\ \frac{\partial \ell(\pi_1, \pi_2)}{\partial \pi_1} &= \frac{n_{21}}{\pi_2} - \frac{n_{2+} - n_{21}}{1-\pi_2} = \frac{n_{21} - n_{2+}\pi_2}{\pi_2(1-\pi_2)}. \end{aligned}$$

   Hence, the MLEs are $\hat{\pi}_1 = n_{11}/n_{1+}$ and $\hat{\pi}_2 = n_{21}/n_{2+}$.

   (b) Approximate the distribution of $(\hat{\pi}_1, \hat{\pi}_2)$.
   Solution: The Fisher information is

$$\mathcal{I}(\pi_1, \pi_2) = \text{var}\begin{pmatrix} \frac{\partial \ell(\pi_1, \pi_2)}{\partial \pi_1} \\ \frac{\partial \ell(\pi_1, \pi_2)}{\partial \pi_2} \end{pmatrix} = \begin{bmatrix} \frac{n_{1+}}{\pi_1(1-\pi_1)} & 0 \\ 0 & \frac{n_{2+}}{\pi_2(1-\pi_2)} \end{bmatrix}.$$

   Hence,

$$\begin{bmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \end{bmatrix} - \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} \approx N\left(0, \begin{bmatrix} \frac{\pi_1(1-\pi_1)}{n_{1+}} & 0 \\ 0 & \frac{\pi_2(1-\pi_2)}{n_{2+}} \end{bmatrix}\right).$$

   (c) Consider the log odds ratio $\log \hat{\theta} = \log n_{11} - \log n_{12} - \log n_{21} + \log n_{22}$. Approximate the distribution of $\log \hat{\theta}$.
   Solution: By the delta method, we consider the function

$$g(x_1, x_2) = \log x_1 - \log(1-x_1) - \log(1-x_2) + \log x_2$$

   with

$$\frac{\partial g(x_1, x_2)}{\partial x} = \begin{bmatrix} \frac{1}{x_1} + \frac{1}{1-x_1} \\ \frac{1}{x_2} + \frac{1}{1-x_2} \end{bmatrix}.$$

Then,

$$\begin{aligned}
g(\pi_1,\pi_2) &= \log\pi_1 - \log(1-\pi_1) - \log(\pi_2) + \log(1-\pi_2) \\
&= \log\left(\frac{\pi_1(1-\pi_2)}{(1-\pi_1)\pi_2}\right) = \log\theta, \\
g\left(\frac{n_{11}}{n_{1+}},\frac{n_{21}}{n_{2+}}\right) &= \log\frac{n_{11}}{n_{1+}} - \log\left(1-\frac{n_{11}}{n_{1+}}\right) - \log\left(\frac{n_{21}}{n_{2+}}\right) + \log\left(1-\frac{n_{21}}{n_{2+}}\right) \\
&= \log\frac{n_{11}}{n_{1+}} - \log\left(\frac{n_{12}}{n_{1+}}\right) - \log\left(\frac{n_{21}}{n_{2+}}\right) + \log\left(\frac{n_{22}}{n_{2+}}\right) \\
&= \log n_{11} - \log n_{12} - \log n_{21} + \log n_{22} = \log\hat{\theta}, \\
\frac{\partial g(\pi_1,\pi_2)}{\partial\pi} &= \begin{bmatrix} \frac{1}{\pi_1} + \frac{1}{1-\pi_1} \\ \frac{1}{\pi_2} + \frac{1}{1-\pi_2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\pi_1(1-\pi)} \\ \frac{1}{\pi_2(1-\pi_2)} \end{bmatrix}.
\end{aligned}$$

The delta method implies that

$$\begin{aligned}
\log\hat{\theta} - \log\theta &\approx N\left(0, \begin{bmatrix} \frac{1}{\pi_1(1-\pi)} \\ \frac{1}{\pi_2(1-\pi_2)} \end{bmatrix}^T \begin{bmatrix} \frac{\pi_1(1-\pi_1)}{n_{1+}} & 0 \\ 0 & \frac{\pi_2(1-\pi_2)}{n_{2+}} \end{bmatrix} \begin{bmatrix} \frac{1}{\pi_1(1-\pi)} \\ \frac{1}{\pi_2(1-\pi_2)} \end{bmatrix}\right) \\
&= N\left(0, \frac{1}{\pi_1(1-\pi)n_{1+}} + \frac{1}{\pi_2(1-\pi_2)n_{2+}}\right).
\end{aligned}$$

Since π's are unknown, the distribution can be also approximated by

$$\log\hat{\theta} - \log\theta \approx N\left(0, \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right),$$

replacing π by its estimate.

2. (2pt) We have shown in the lectures that, in a 2×2 table, odds ratio being 1 if and only if X and Y are independent. Suppose that we have a 2×3 table. Show that all pairwise odds ratio are 1 if and only if X and Y are independent.
Solution: Suppose that all pairwise odds ratio are 1. Then,

$$\pi_{12}\pi_{21} = \pi_{11}\pi_{22} = \pi_{11}(1-\pi_{1+}-\pi_{21}-\pi_{23}).$$

Hence,

$$\begin{aligned}
\pi_{11} &= \pi_{12}\pi_{21} + \pi_{11}\pi_{1+} + \pi_{11}\pi_{21} + \pi_{11}\pi_{23} \\
&= \pi_{11}\pi_{1+} + (\pi_{11}+\pi_{12}+\pi_{13})\pi_{21} + \pi_{11}\pi_{23} - \pi_{13}\pi_{21} \\
&= \pi_{11}\pi_{1+} + \pi_{1+}\pi_{21} + \pi_{11}\pi_{23} - \pi_{13}\pi_{21} \\
&= \pi_{1+}\pi_{+1} + \pi_{11}\pi_{23} - \pi_{13}\pi_{21},
\end{aligned}$$

where $\pi_{11}\pi_{23} - \pi_{13}\pi_{21} = 0$ since

$$\frac{\pi_{11}\pi_{23}}{\pi_{21}\pi_{13}}$$

is also an odds ratio. Likewise we can show that $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all i and j. Hence, X and Y are independent.
Suppose that X and Y are independent, then

$$\frac{\pi_{ab}\pi_{cd}}{\pi_{ab}\pi_{cd}} = \frac{\pi_{a+}\pi_{+b}\pi_{c+}\pi_{+d}}{\pi_{a+}\pi_{+b}\pi_{c+}\pi_{+d}} = 1.$$

Hence, all pairwise odds ratio are 1.

3. (1pt) In August 2023, Swedish Radio (Sveriges Radio Ekot) reported the following news. According to a survey study conducted by Arbetsmiljöverket (The Swedish Work Environment Authority), 12% of people who are born outside of Sweden have been exposed to mobbing, whereas 6% Swedish born have been exposed to mobbing in the Swedish job market. The news headline is "foreign-born been mobbed twice as often in their jobs". Suppose that the response rate of the survey is 100% and that all answers by the respondents are in line with reality. From the perspective of Simpson's paradox, name two confounders that may bias the conclusion.
Solution: One reason can be foreign-born applies more to the industry that is prone to mobbing. Another reason could be foreign-born have low-level positions whereas Swedish-born have higher level positions.

4. (1pt) We know conditional independence does not imply marginal independence. Does marginal independence imply conditional independence? If so, prove it. Otherwise give a counterexample.
Solution: Marginal independence does not imply conditional independence. One example could be Simpson's paradox for regression. In each group, we can have a downward trend, but when considered together, data from different groups are on the same horizontal line.

5. (2pt) A poll has been done regarding the hypothetical question: "if the election for president were being held today, and the candidates were Joe Biden the Democrat and Donald Trump the Republican, for whom would you vote". The poll has been done in three states that will be treated as a confounder. The results are tabulated in the following contingency table

		Gender	
State	Candidate	Male	Female
1	Biden	646	911
	Trump	828	580
2	Biden	518	820
	Trump	778	475
3	Biden	652	864
	Trump	728	485

   (a) Compute in software the conditional odds ratios and the marginal odds ratio. Do you think whether the table has homogeneous association?
   Solution: We can compute the conditional odds ratio by

```
library(vcd)

## Loading required package: grid

x <- array(data = c(646, 828, 911, 580,
                    518, 778, 820, 475,
                    652, 728, 864, 485),
           dim = c(2, 2, 3),
           dimnames = list(Candidate = c("Biden", "Trump"),
                           Gender = c("Male", "Female"),
                           State = c("1", "2", "3")))
oddsratio(x, log = FALSE)

##  odds ratios for Candidate and Gender by State
##
##         1         2         3
## 0.4967202 0.3856825 0.5027409
```

   To compute the marginal odds ratio, we use

```
MarginTable <- margin.table(x, margin = c("Candidate", "Gender"))
oddsratio(MarginTable, log = FALSE)

##  odds ratios for Candidate and Gender
##
## [1] 0.4617409
```

To test homogenenous association in a $2 \times 2 \times K$ table, we can use the Breslow-Day test as

```
library(DescTools)
BreslowDayTest(x)

##
##  Breslow-Day test on Homogeneity of Odds Ratios
##
## data:  x
## X-squared = 7.0934, df = 2, p-value = 0.02882
```

At $\alpha = 0.05$, we can reject the null hypothesis of homogeneous association.

(b) Compute also the 95% confidence intervals for the conditional odds ratios and the marginal odds ratio.
Solution: To compute the confidence intervals for the conditional odds ratio, we use

```
confint(loddsratio(x), level = 0.95, log = FALSE)

##       2.5 %    97.5 %
## 1 0.4290832 0.5750188
## 2 0.3291469 0.4519289
## 3 0.4312392 0.5860979
```

To compute the confidence intervals for the marginal odds ratio, we use

```
confint(loddsratio(MarginTable), level = 0.95, log = FALSE)

##                               2.5 %    97.5 %
## Biden:Trump/Male:Female 0.4228635 0.5041926
```

6. (2pt) A study has been done to evaluate students' level of academic writing expertise in two majors. Results are summarized in the following contingency table

	Level of Academic Writing Expertise			
Major	None	Beginner	Intermediate	Expert
1	8	7	11	39
2	11	41	13	9

(a) Test independence of Major and Level of Academic Writing Expertise under multinomial sampling using Pearson chi-square and likelihood ratio test. You can use software for this task.
Solution: To test independence,

```
Data <- matrix(c(8, 11, 7, 41, 11, 13, 39, 9), 2, 4)
n <- sum(Data)
## Pearson chi-square
chisq.test(Data)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Data
## X-squared = 43.072, df = 3, p-value = 2.376e-09
```

Hence, we reject the null hypothesis of independence. The LRT is done by

```
px <- matrix(rowSums(Data) / n, ncol = 1)
py <- matrix(colSums(Data) / n, nrow = 1)
## LRT statistic
-2 * sum(Data * log((px %*% py) / (Data / n)))

## [1] 46.93656

## Critical value
qchisq(0.95, (2 - 1) * (4 - 1))

## [1] 7.814728
```

Hence, the LRT also rejects the null hypothesis of independence.

(b) Write your own code to compute Goodman-Kruskal's gamma for the table. You can compare your results with the R function GoodmanKruskalGamma().
Solution: We compute the concordant and discordant pairs as follows.

```
ConPair <- matrix(0, 2, 4)
ConPair[1, 1] <- Data[1, 1] * sum(Data[2, 2 : 4])
ConPair[1, 2] <- Data[1, 2] * sum(Data[2, 3 : 4])
ConPair[1, 3] <- Data[1, 3] * sum(Data[2, 4 : 4])
C <- sum(ConPair)
DisPair <- matrix(0, 2, 4)
DisPair[1, 2] <- Data[1, 2] * sum(Data[2, 1 : 1])
DisPair[1, 3] <- Data[1, 3] * sum(Data[2, 1 : 2])
DisPair[1, 4] <- Data[1, 4] * sum(Data[2, 1 : 3])
D <- sum(DisPair)
## Gamma
(C - D) / (C + D)

## [1] -0.6158335
```

We can compare our results with the package

```
library(DescTools)
GoodmanKruskalGamma(Data, conf.level = .95)

##      gamma     lwr.ci     upr.ci
## -0.6158335 -0.8007850 -0.4308821
```

(c) Test whether students in major 1 tends to have a higher level than students in major 2. You can use software for solve this task.
Solution: We will use the Mann-Whitney test here.

```
ExData <- rbind(matrix(c(1, 1), 8, 2, byrow = TRUE),
                matrix(c(1, 2), 7, 2, byrow = TRUE),
                matrix(c(1, 3), 11, 2, byrow = TRUE),
                matrix(c(1, 4), 39, 2, byrow = TRUE),
                matrix(c(2, 1), 11, 2, byrow = TRUE),
```

```
                matrix(c(2, 2), 41, 2, byrow = TRUE),
                matrix(c(2, 3), 13, 2, byrow = TRUE),
                matrix(c(2, 4), 9, 2, byrow = TRUE))
colnames(ExData) <- c("Major", "Writing")
ExData <- data.frame(ExData)
wilcox.test(Writing ~ Major, data = ExData, alternative = "greater")

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Writing by Major
## W = 3618.5, p-value = 3.985e-08
## alternative hypothesis: true location shift is greater than 0
```

Hence, we reject the null hypothesis that students in major 1 tends to have a lower level than students in major 2.