Analysis of Categorical Data
Chapter 3: Inference for Contingency Table

Shaobo Jin

Department of Mathematics

Intended Learning Outcome

Through this chapter, you should be able to

1. test independence in contingency table,
2. test monotone trend.

Odds Ratio

Suppose that we have observed a 2×2 table

	Y	
X	1	2
1	n_{11}	n_{12}
2	n_{21}	n_{22}

The sample odds ratio is

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \geq 0.$$

If $\hat{\theta} > 0$, then we can consider

$$\log \hat{\theta} = \log n_{11} + \log n_{22} - \log n_{12} - \log n_{21}.$$

Wald Confidence Interval

An estimated standard error of $\log \hat{\theta}$ is

$$\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

Hence, a Wald confidence interval for $\log \theta$ is

$$\log \hat{\theta} \pm z_{\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

However,

$$\hat{\theta} = \frac{n_{11} n_{22}}{n_{12} n_{21}}$$

can be 0 (if $n_{11}n_{22} = 0$), ∞ ($n_{12}n_{21} = 0$), or undefined (if $n_{11}n_{22} = n_{12}n_{21} = 0$). Consequently, the Wald interval may not exist.

- An ad-hoc approach is to add 0.5 to n_{ij}.
- Use other approaches such as the score interval or the likelihood ratio confidence interval.

Example: Aspirin Use and Myocardial Infraction

Compute $\hat{\theta}$ and find a 95% confidence interval for θ

	Myocardial Infraction	
	Yes	No
Placebo	28	656
Aspirin	18	658

Independence

We have an $I \times J$ contingency table from multinomial sampling with probabilities $\{\pi_{ij}\}$.We want to test

$$\begin{aligned} H_0 : &\quad \text{independence as } \pi_{ij} = \pi_{i+}\pi_{+j} \text{ for all } i, j, \\ H_1 : &\quad H_0 \text{ is not true.} \end{aligned}$$

The log-likelihood under H_1 is

$$\ell_0\left(\pi_{i+}, \pi_{+j}\right) = \log\left(\frac{n!}{n_{11}!\cdots n_{IJ}!}\right) + \sum_i \sum_j n_{ij} \log\left(\pi_{i+}\pi_{+j}\right).$$

The log-likelihood under H_1 is

$$\ell_1\left(\pi_{ij}\right) = \log\left(\frac{n!}{n_{11}!\cdots n_{IJ}!}\right) + \sum_i \sum_j n_{ij} \log\left(\pi_{ij}\right).$$

Likelihood Ratio Test

The MLE under H_0 is

$$\hat{\pi}_{i+} = \frac{n_{i+}}{n},\ \hat{\pi}_{+j} = \frac{n_{+j}}{n}.$$

The MLE under H_1 is

$$\hat{\pi}_{ij} \quad = \quad \frac{n_{ij}}{n}.$$

The likelihood ratio test statistic is

$$G^2 = -2\left[\ell_0\left(\hat{\pi}_{i+}, \hat{\pi}_{+j}\right) - \ell_1\left(\hat{\pi}_{ij}\right)\right] \quad = \quad -2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\left(\frac{n_{i+}n_{+j}/n}{n_{ij}}\right).$$

If H_0 holds, G^2 also converges in distribution to to chi-square with $(IJ-1)-(I-1)-(J-1) = (I-1)(J-1)$ degrees of freedom. A rule-of-thumb is that no more than 20% of $\hat{\mu}_{ij} < 5$.

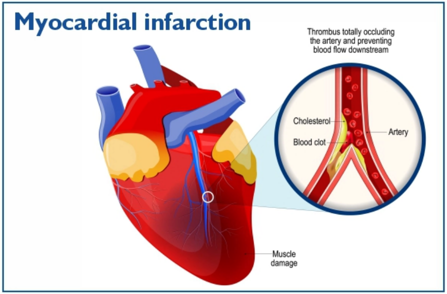
Pearson Chi-Square

The Pearson chi-square that tests H_0 : independence is

$$\begin{aligned} X^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(\text{observed frequency}_{ij} - \text{expected frequency}_{ij}\right)^2}{\text{expected frequency}_{ij}} \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij} - n\hat{\pi}_{i+}\hat{\pi}_{+j}\right)^2}{n\hat{\pi}_{i+}\hat{\pi}_{+j}}. \end{aligned}$$

If H_0 holds, X^2 converges in distribution to to chi-square with $(I-1)(J-1)$ degrees of freedom. A rule-of-thumb is still that no more than 20% of $\hat{\mu}_{ij} < 5$.

Aspirin Use and Myocardial Infarction



Test independence

	Myocardial Infarction	
	Yes	No
Placebo	28	656
Aspirin	18	658

Fisher's Exact Test

For 2×2 tables, regardless of sampling, under the independence assumption, conditioning on both sets of marginal totals, the only free cell is n_{11}. It follows the hypergeometric distribution

$$P\left(n_{11} = t\right) = \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{+1} - t}}{\binom{n}{n_{+1}}}.$$

- For $H_0 : \theta = 1$ (independence) versus $H_1 : \theta > 1$, the Fisher's exact test uses the p-value $P\left(n_{11} \geq t_o\right)$ where t_o is the observed value of n_{11}.
- For $H_0 : \theta = 1$ versus $H_1 : \theta \neq 1$, there are different ways of computing the p-value. They lead to different p-values.

Example

Fisher's Tea Tasting Experiment

	Guess Poured First		
Poured First	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	8

Ordinality and Scoring

If our data are ordinal, using the above X^2 and G^2 are less ideal since they ignore ordinality of data.

To keep ordinality, many people choose to assign scores to the ordinal variables: $u_1 \leq u_2 \leq \cdots \leq u_I$ be the scores for the rows, and $v_1 \leq v_2 \leq \cdots \leq v_J$ be the scores for the columns. The scores are then treated as the values of the variables. However, this approach has several serious issues:

1. How shall we assign scores?
2. Are the distance between the assigned score actually reflect the "distance" between categories?

Ordinal Variables

Suppose that both X and Y are ordinal.

1. A pair of subjects is concordant if the subject ranked higher on X also ranks higher on Y.
2. A pair of subject is discordant if the subject ranking higher on X ranks lower on Y.

Age	Job satisfaction		
	1: Not satisfied	2: Satisfied	3: Very satisfied
1: < 30	34	53	88
2: $30 - 50$	80	174	304
3: > 50	29	75	172

Concordant/Discordant Pairs

Age	Job satisfaction		
	1: Not satisfied	2: Satisfied	3: Very satisfied
1: < 30	34	53	88
2: $30 - 50$	80	174	304
3: > 50	29	75	172

- Subject A belongs to $(1,1)$ and subject B belongs to $(2,2)$. The pair (A,B) is concordant.
- Subject A belongs to $(2,2)$ and subject B belongs to $(1,1)$. Also concordant.
- Subject A belongs to $(1,2)$ and subject B belongs to $(2,1)$. The pair (A,B) is discordant.
- Subject A belongs to $(2,1)$ and subject B belongs to $(1,2)$. Also discordant.

Probability of Concordant/Discordant

Suppose that we have two independent subjects A and B from a joint distribution $\{\pi_{ij}\}$.

1. The probability of a concordant pair is

$$\begin{aligned}\Pi_c &= \sum_{i,j}\{P[A=(i,j)]\,P[B=(h,k),\ h>i,\ k>j \mid A=(i,j)]\} \\ &\quad +\sum_{i,j}\{P[A=(i,j)]\,P[B=(h,k),\ h<i,\ k<j \mid A=(i,j)]\} \\ &= 2\sum_{i,j}\left\{\pi_{ij}\sum_{h>i}\sum_{k>j}\pi_{hk}\right\}.\end{aligned}$$

2. The probability of a discordant pair is

$$\Pi_d = 2\sum_{i,j}\left\{\pi_{ij}\sum_{h>i}\sum_{k<j}\pi_{hk}\right\}.$$

Gamma Coefficient

We define the Goodman-Kruskal's gamma as

$$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d}.$$

1. γ has the range $-1 \leq r \leq 1$. It work in a similar way as the Pearson correlation coefficient.
2. If $\gamma > 0$ $(\Pi_c > \Pi_d)$, then it is more likely to have concordant pairs than discordant pairs (positive trend).
3. If $\gamma < 0$ $(\Pi_c < \Pi_d)$, then it is less likely to have concordant pairs than discordant pairs (negative trend).
4. If $\gamma = 0$, then no trend.
5. If X and Y are independent, then $\gamma = 0$. But $\gamma = 0$ does not mean independence.

Alternative Method

For ordinal data, we use the sample Goodman-Kruskal's gamma is

$$\hat{\gamma} = \frac{C - D}{C + D}$$

to check whether they have a monotone trend, where C is the total number of concordant pairs of observations, and D is the total number of discordant pairs of observations.

- If $\gamma = 0$, there is no trend between X and Y.
- For a large sample size, $\hat{\gamma}$ is approximately normal.

Example Sample Gamma Coefficient

The sample version of γ is

$$\hat{\gamma} = \frac{C - D}{C + D},$$

where C is the total number of concordant pairs and D is the total number of discordant pairs.

Compute $\hat{\gamma}$

	Job satisfaction		
Age	1: Not satisfied	2: Satisfied	3: Very satisfied
1: < 30	34	53	88
2: $30 - 50$	80	174	304
3: > 50	29	75	172

Wilcoxon Test or Mann-Whitney Test

Suppose that we have two random variables Y_0 and Y_1. The Wilcoxon test or the Mann-Whitney test tests whether

$$P(Y_0 > Y_1) = P(Y_0 < Y_1).$$

In the special case where we have a $2 \times J$ table $(I = 2)$ and the scores for X are $\{0, 1\}$. We have two groups, one group with $X = 0$ and another group with $X = 1$. Then the general idea is that

1. Assign ranks to the whole sample of size $n_{0+} + n_{1+}$.
2. Compute the sum of ranks assigned to the group $X = 0$.
3. If H_0 is not true, the sum of ranks tends to be either small or large.

The Kruskal-Wallis test generalizes the Mann-Whitney test to more than 2 groups. The Kruskal-Wallis test can be viewed as a non-parametric version of one-way ANOVA.

Be Careful With Their Hypotheses

	Y			
X	1	2	3	4
0	0.05	0.5	0.35302019	0.09697981
1	0.1666553	0.2833447	0.5000000	0.0500000