DATA 606: Statistical Methods in Data Science

—— Inference for contingency table

Wenjun Jiang

Department of Mathematics & Statistics
The University of Calgary

Lecture 9

Comparing the proportions

$\pi_1 - \pi_2$

- Given the following two-way contingency table

	Y_1	Y_2	
X_1	n_{11}	n_{12}	n_{1+}
x_2	n_{21}	n_{22}	n_{2+}

- We would like to know more about $\pi_1 - \pi_2$ where

$$\pi_1 = \mathbf{P}(Y = Y_1 | X = X_1), \quad \underset{\pi}{\cancel{p}}_2 = \mathbf{P}(Y = Y_1 | X = X_2).$$

- If X and Y are independent, then $\pi_1 - \pi_2 = 0$.

- The estimate of $\pi_1 - \pi_2$:

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

Comparing the proportions

$\pi_1 - \pi_2$

- The estimate of $\pi_1 - \pi_2$:

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

- The estimate of the variance:

$$\hat{\mathrm{Var}}(\hat{\pi}_1 - \hat{\pi}_2) = \frac{\hat{\pi}_1(1 - \hat{\pi}_2)}{n_{1+}} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_{2+}}.$$

- The estimate of $\pi_1 - \pi_2$:

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

- The estimate of the variance:

$$\hat{\mathrm{Var}}(\hat{\pi}_1 - \hat{\pi}_2) = \frac{\hat{\pi}_1(1 - \hat{\pi}_2)}{n_{1+}} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_{2+}}.$$

- Large-sample $1 - \alpha$ confidence interval for $\pi_1 - \pi_2$

$$\left[\hat{\pi}_1 - \hat{\pi}_2 - z_{\alpha/2}\sqrt{\hat{\mathrm{Var}}(\hat{\pi}_1 - \hat{\pi}_2)},\ \hat{\pi}_1 - \hat{\pi}_2 + z_{\alpha/2}\sqrt{\hat{\mathrm{Var}}(\hat{\pi}_1 - \hat{\pi}_2)}\right].$$

Comparing the proportions

$\pi_1 - \pi_2$

Example 1 (Aspirin and heart attack)

The effectiveness of different drugs are as follows

	Yes	No	
Placebo	189	10845	11034
Aspirin	104	10933	11037

Comparing the proportions

$\pi_1 - \pi_2$

Example 1 (Aspirin and heart attack)

The effectiveness of different drugs are as follows

	Yes	No	
Placebo	189	10845	11034
Aspirin	104	10933	11037

- The difference of curing probabilities is

$$\hat{\pi}_1 - \hat{\pi}_2 = \frac{189}{11034} - \frac{104}{11037} = 0.0077.$$

- Standard deviation is $\sqrt{\frac{0.0171(1-0.0171)}{11034} + \frac{0.0094(1-0.0094)}{11037}} = 0.0015.$
- Large-sample 95% CI for the difference is

$$[0.0077 - 1.96 \times 0.0015, 0.0077 + 1.96 \times 0.0015] \Rightarrow [0.0048, 0.0106].$$

Comparing the proportions

π_1/π_2

Example 2

When both π_1 and π_2 are close to zero, the difference between them may not be that meaningful:

1. $\pi_1 = 0.01$, $\pi_2 = 0.001$, then $\pi_1 - \pi_2 = 0.009$.
2. $\pi_1 = 0.41$, $\pi_2 = 0.401$, then $\pi_1 - \pi_2 = 0.009$.

Comparing the proportions

π_1/π_2

Example 2

When both π_1 and π_2 are close to zero, the difference between them may not be that meaningful:

1. $\pi_1 = 0.01$, $\pi_2 = 0.001$, then $\pi_1 - \pi_2 = 0.009$.
2. $\pi_1 = 0.41$, $\pi_2 = 0.401$, then $\pi_1 - \pi_2 = 0.009$.

- An alternative measure is the *relative risk*: $RR = \frac{\pi_1}{\pi_2}$.
- Some properties:
  - $0 < RR < \infty$.
  - $\pi_1 > \pi_2 \iff RR > 1$.
  - $\pi_1 = \pi_2 \iff RR = 1$ (independence).
- An estimate for relative risk: $\frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}$.

Comparing the proportions

π_1/π_2

- Large-sample $1-\alpha$ CI for $\log RR$

$$\left[\log RR - z_{\alpha/2}\hat{\sigma}(\log RR), \log RR + z_{\alpha/2}\hat{\sigma}(\log RR)\right],$$

where

$$\hat{\sigma}^2(\log RR) = \frac{1-\hat{\pi}_1}{n_{11}} + \frac{1-\hat{\pi}_2}{n_{21}}.$$

Comparing the proportions

π_1/π_2

Example 3 (Aspirin revisited)

An estimate for the relative risk is $\frac{189/11034}{104/11037} \approx 1.818$. The approximate standard deviation is

$$\hat{\sigma}(\log RR) = \sqrt{\frac{1 - 189/11034}{189} + \frac{1 - 104/11037}{104}} \approx 0.121.$$

The 95% CI for $\underline{\log RR}$ is

$$[\log 1.818 - 1.96 \times 0.121, \log 1.818 + 1.96 \times 0.121] \Rightarrow [0.361, 0.835]$$

which gives the 95% CI for <u>RR</u>

$$\left[e^{0.361}, e^{0.835}\right] \Rightarrow [1.435, 2.305]\,.$$

Definition 4 (Odds ratio)

For a two-way contingency table, the odds ratio is defined as

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1}{\pi_2} \cdot \frac{1-\pi_2}{1-\pi_1} = RR \cdot \frac{1-\pi_2}{1-\pi_1}.$$

Properties of θ:

- $0 < \theta < \infty$.
- $\pi_1 > \pi_2 \iff \theta > 1$
- $\pi_1 = \pi_2 \iff \theta = 1$ (independence).

Comparing the proportions

Odds ratio

- An estimate for θ

$$\hat{\theta} = \frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_2/(1-\hat{\pi}_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}}.$$

- Large sample approximate variance for $\log \hat{\theta}$

$$\hat{\mathrm{Var}}(\log \hat{\theta}) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}.$$

- The approximate $1-\alpha$ CI for $\log \theta$ is

$$\left[\log \hat{\theta} - z_{\alpha/2}\sqrt{\hat{\mathrm{Var}}(\log \hat{\theta})}, \log \hat{\theta} + z_{\alpha/2}\sqrt{\hat{\mathrm{Var}}(\log \hat{\theta})}\right].$$

Comparing the proportions

Odds ratio

Example 5 (Aspirin revisited)

$$\hat{\theta} = \frac{189 \times 10933}{10845 \times 104} \approx 1.832.$$

$$\hat{\mathrm{Var}}(\log \hat{\theta}) = \frac{1}{189} + \frac{1}{10845} + \frac{1}{104} + \frac{1}{10933} \approx 0.015$$

The 95% CI for $\log \theta$

$$\left[\log 1.832 - 1.96 \times \sqrt{0.015}, 1.832 + 1.96 \times \sqrt{0.015}\right] \Rightarrow [0.365, 0.846]\,,$$

the 95% CI for θ is

$$\left[e^{0.365}, e^{0.846}\right] \Rightarrow [1.44, 2.33].$$

χ^2 test for independence

A joint table of X and Y is like

	Y_1	Y_2	$\cdots$	Y_J	
X_1	π_{11}	π_{12}	$\cdots$	π_{1J}	π_{1+}
X_2	π_{21}	π_{22}	$\cdots$	π_{2J}	π_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_I	π_{I1}	π_{I2}	$\cdots$	π_{IJ}	π_{I+}
	π_{+1}	π_{+2}	$\cdots$	π_{+J}	1

Goal: test the independence between X and Y, e.g.

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \quad \text{for } i \in \{1, 2, \ldots, I\},\ j \in \{1, 2, \ldots, J\}$$

.

χ^2 test for independence

- The estimate of π_{i+} and π_{+j} are

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

- Under H_0, we have

$$\hat{\mu}_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j} = \frac{n_{i+}n_{+j}}{n}.$$

- The Pearson χ^2 statistic is

$$\chi^2 = \sum_{i \in \{1,\ldots,I\}, j \in \{1,\ldots,J\}} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \sim \chi^2_{df}$$

  where $df = IJ - 1 - (I - 1 + J - 1) = (I - 1)(J - 1)$.
- **Reject** H_0 if $\chi^2 \geq \chi^2_{df}(\alpha)$.

χ^2 test for independence

Example 6 (Gender in party identification)

Suppose we are given the following table

	Democrat	Independent	Republican	Total
Female	762	327	468	1557
Male	484	239	477	1200
	1246	566	945	2757

- Then $\hat{\mu}_{11} = \frac{1577\times1246}{2757} = 703.7$, etc.
- Pearson's statistic $\chi^2 = \frac{(762-703.7)^2}{703.7} + \frac{(327-319.6)^2}{319.6} + \cdots = 30.1$.
- The degree of freedom (df) is $(2-1)(3-1) = 2$ and $\chi^2_2(0.05) = 5.99$.
- Reject H_0 as $30.1 > 5.99$.

Cell residuals

A limitation of significance test: these tests only tell us whether there is evidence for the association, but how strong this association is remains unclear.

Cell residuals

A limitation of significance test: these tests only tell us whether there is evidence for the association, but how strong this association is remains unclear.

Solution: a cell-by-cell comparison of the observed and estimated frequencies helps show more about the relation.

Cell residuals

- Under $H_0 : X$ and Y are independent, we have $\hat{\mu}_{ij} = \frac{n_{i+} n_{+j}}{n}$.
- We calculate the standardized Pearson residuals

$$e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - \hat{\pi}_{i+})(1 - \hat{\pi}_{+j})}}.$$

- Under H_0, e_{ij} behaves like a $N(0, 1)$ random variable.
- We can observe e_{ij} to check the departure from H_0.

Cell residuals

- In the gender and party identification example,

$$\hat{\pi}_{1+} = \frac{1557}{2757} = 0.565, \quad \hat{\pi}_{+1} = \frac{1246}{2757} = 0.452,$$

$$e_{11} = \frac{762 - 703.7}{\sqrt{703.7(1 - 0.565)(1 - 0.452)}} = 4.5.$$

- The cell-by-cell residuals are displayed as

	Democrat	Independent	Republican
Female	4.5	0.7	-5.3
Male	-4.5	-0.7	5.3
	1246	566	945

- There are significantly more democrat females (less males) than predicted by the independence model, there are significantly less republican females (more males) than predicted by the model.

Testing independence for ordinal data

- X has I categories: $X_1, \ldots, X_I$; Y has J categories: $Y_1, \ldots, Y_J$. We know $X_1 < \cdots < X_I$ and $Y_1 < \cdots < Y_J$.
- We want to test whether X is independent of Y.
- Assign scores $u_1 < \cdots < u_I$ to categories of X and $v_1 < \cdots < v_J$ to those of Y.
- See the following example, patients with two diseases X and Y. The categories are the level of symptoms(slight, medium and heavy).

Testing independence for ordinal data

		Y		
		v_1	v_2	v_3
X	u_1	2	1	3
	u_2	1	2	1
	u_3	1	1	2

$\Rightarrow$

Patient	X	Y
1	u_1	v_1
2	u_1	v_1
3	u_1	v_2
4	u_1	v_3
5	u_1	v_3
6	u_1	v_3
7	u_2	v_1
8	u_2	v_2
9	u_2	v_2
10	u_2	v_3
11	u_3	v_1
12	u_3	v_2
13	u_3	v_3
14	u_3	v_3

Testing independence for ordinal data

- Pearson correlation coefficient describes the *linear relationship* between X and Y:

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \frac{1}{n-1}\sum_{i=1}^{n}(y - I - \bar{y})^2}}$$

where

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{n}\sum_{i=1}^{I} n_{i+}u_i = \bar{u},$$

$$\bar{y} = \bar{v}.$$

- In that case, we have

$$r = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\pi_{ij}(u_i - \bar{u})(v_j - \bar{v})}{\sqrt{\sum_{i=1}^{I}\pi_{i+}(u_i - \bar{u})^2 \cdot \sum_{j=1}^{J}\pi_{+j}(v_j - \bar{v})^2}}.$$

Testing independence for ordinal data

- Under H_0: X and Y are independent, large sample theory gives

$$\sqrt{n-1} \cdot r \sim N(0,1),$$

$$M^2 = (n-1)r^2 \sim \chi_1^2.$$

This test is named **Mantel-Haenszel** test.

Testing independence for ordinal data

- Under H_0: X and Y are independent, large sample theory gives

$$\sqrt{n-1} \cdot r \sim N(0,1),$$

$$M^2 = (n-1)r^2 \sim \chi_1^2.$$

  This test is named **Mantel-Haenszel** test.

- How to choose scores $\{u_i\}$ and $\{v_i\}$?

Testing independence for ordinal data

- Under H_0: X and Y are independent, large sample theory gives

$$\sqrt{n-1} \cdot r \sim N(0,1),$$

$$M^2 = (n-1)r^2 \sim \chi_1^2.$$

  This test is named **Mantel-Haenszel** test.

- How to choose scores $\{u_i\}$ and $\{v_i\}$? **Answer:** any increasing/decreasing sequence is ok.

Testing independence for ordinal data

- Under H_0: X and Y are independent, large sample theory gives

$$\sqrt{n-1} \cdot r \sim N(0,1),$$

$$M^2 = (n-1)r^2 \sim \chi^2_1.$$

  This test is named **Mantel-Haenszel** test.

- How to choose scores $\{u_i\}$ and $\{v_i\}$? **Answer:** any increasing/decreasing sequence is ok.

- In the gender and party identification example, $M^2 = 28.98 > \chi^2_1(0.05) = 3.381$, therefore reject H_0.

Testing independence for ordinal data

Example 7 (Mother's alcohol consumption and infant malformation)

	Present (Y=1)	Absent (Y=0)
0	48	17066
< 1	38	14464
1–2	5	788
3–5	1	126
> 6	1	37

- Pearson's test: $\chi^2 = 12.1 > \chi^2_4(0.95) = 9.49$.
- Assign scores $0, 0.5, 1.5, 4, 7$ to alcohol consumption and $0, 1$ to absent/present. We have

$$M^2 = 6.6 > \chi^2_1(0.95) = 3.84.$$

Conclusion: there exists relationship between mother's alcohol consumption and infant malformation.

Tests for nominal-ordinal data

- X-nominal, Y-ordinal, such that

$$
\begin{array}{cc|c|c|c|c}
 & & \multicolumn{3}{c}{Y} & \\
 & & v_1 & v_2 & v_3 & \\
\cline{3-5}
 & 1 & n_{11} & n_{12} & n_{13} & n_{1+} \\
\cline{3-5}
X & 2 & n_{21} & n_{22} & n_{23} & n_{2+} \\
\cline{3-5}
 & 3 & n_{31} & n_{32} & n_{33} & n_{3+} \\
\cline{3-5}
\end{array}
$$

- $H_0 : X$ and Y are independent $\Longrightarrow$ the conditional distributions $(\mathbf{P}(Y|X = i))$ of Y given X are the same across all levels of $X \Longrightarrow$ the conditional means remain unchanged $\mathbf{E}[Y|X = i]$.

- This is an ANOVA problem !

Tests for nominal-ordinal data

- We have $SSTO = SSW + SSB$:

$$\sum_{i=1}^{I}\sum_{j=1}^{j} n_{ij}(v_j - \bar{v})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(v_j - \bar{v}_i)^2 + \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(\bar{v}_i - \bar{v})^2,$$

  where

$$\bar{v} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}v_j}{n}$$

$$\bar{v}_i = \frac{\sum_{j=1}^{J} n_{ij}v_j}{n_{i+}}$$

- The F-test is

$$F = \frac{SSB/(I-1)}{SSW/(n-I)} \sim F_{I-1,n-I}.$$

Exact inference

As you can see, most aforementioned methods require large sample. **What if the sample size is small**?

Example 8 (Fisher's tea)

Fisher's colleague, Muriel Bristol claims she could tell whether or not tea (or milk) was added to the cup first.

		Muriel's Guess		
		Milk	Tea	
True	Milk	3	1	4
	Tea	1	3	4
		4	4	

Exact inference

- We want to test whether X (the true order) and Y (the guessed order) are independent, then we test

$$H_0 : \theta = 1 \quad v.s. \quad H_1 : \theta \neq 1.$$

Exact inference

- We want to test whether X (the true order) and Y (the guessed order) are independent, then we test

$$H_0 : \theta = 1 \quad v.s. \quad H_1 : \theta \neq 1.$$

- Because of the small sample, the Pearson test performs poorly.

Exact inference

- We want to test whether X (the true order) and Y (the guessed order) are independent, then we test

$$H_0 : \theta = 1 \quad v.s. \quad H_1 : \theta \neq 1.$$

- Because of the small sample, the Pearson test performs poorly.

- In total, there are $4+4=8$ trials. We know there are 4 times each for milk first ($n_{1+}=4, n_{2+}=4$) and tea first and there are 4 trials each for milk guessed and tea guessed ($n_{+1}=4, n_{+2}=4$).

Exact inference

- We want to test whether X (the true order) and Y (the guessed order) are independent, then we test

$$H_0 : \theta = 1 \quad v.s. \quad H_1 : \theta \neq 1.$$

- Because of the small sample, the Pearson test performs poorly.

- In total, there are $4+4=8$ trials. We know there are 4 times each for milk first ($n_{1+}=4, n_{2+}=4$) and tea first and there are 4 trials each for milk guessed and tea guessed ($n_{+1}=4, n_{+2}=4$).

- **What is the probability that $n_{11}=3$ under H_0?**

Exact inference

- We want to test whether X (the true order) and Y (the guessed order) are independent, then we test

$$H_0 : \theta = 1 \quad v.s. \quad H_1 : \theta \neq 1.$$

- Because of the small sample, the Pearson test performs poorly.

- In total, there are $4+4=8$ trials. We know there are 4 times each for milk first ($n_{1+}=4, n_{2+}=4$) and tea first and there are 4 trials each for milk guessed and tea guessed ($n_{+1}=4, n_{+2}=4$).

- **What is the probability that $n_{11}=3$ under H_0?**
Answer: a hyper-geometric distribution

$$\mathbf{P}(n_{11}=3) = \frac{\binom{n_{1+}}{3} \cdot \binom{n_{2+}}{n_{+1}-3}}{\binom{n}{n_{+1}}} = \frac{\binom{4}{3} \cdot \binom{4}{1}}{\binom{8}{4}}.$$

Exact inference

- $\theta = 1 \iff n_{11} = 2$. In other words, if $n_{11} \neq 2$, then $\theta \neq 1$.
- The probability distribution table

	$n_{11} = 0$	$n_{11} = 1$	$n_{11} = 2$	$n_{11} = 3$	$n_{11} = 4$
Prob	0.014	0.229	0.514	0.229	0.014

- The P-value of this exact test is

$$0.014 + 0.229 + 0.229 + 0.014 = 0.486.$$