Exercises: Contingency tables

Lecturer & author: Georgios Karagiannis georgios.karagiannis@durham.ac.uk

---

Exercise 1. Consider a $I \times J \times K$ contingency table, with classification variables X, Y, Z. Prove that if

1. $X$and Y are conditionally independent on Z; and
2. $X$and Z are conditionally independent on Y

then:
Y and Z are jointly independent from X

Hint: Write down the probability forms involved, and try to derive the result by using simple probability calculus.

Solution 2. I have

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}} \tag{1}$$

$$\pi_{ijk} = \frac{\pi_{ij+}\pi_{+jk}}{\pi_{+j+}} \tag{2}$$

By dividing those two, we get

$$\pi_{i+k}\pi_{+j+} = \pi_{ij+}\pi_{++k}$$

By summing with respect to j we get

$$\pi_{i+k} = \pi_{i++}\pi_{++k} \tag{3}$$

By substituting (3) in (1), we get

$$\pi_{ijk} = \pi_{i++}\pi_{+jk}$$

Hence Y and Z are jointly independent from X

The next exercise is from Homework 1

Exercise 3. Consider a $I \times J \times K$ contingency table, with classification variables X, Y, Z. Prove that if Y and Z are jointly independent from X,

then

1. $X$and Y are conditionally independent on Z; and
2. $X$and Z are conditionally independent on Y

Hint: Write down the probability forms involved, and try to derive the result by using simple probability calculus.

Solution 4.

The next exercise is from Homework 1

Exercise 5. [1] The 1988 General Social Survey compiled by the National Opinion Research Center asked: "Do you support or oppose the following measures to deal with AIDS? (1) Have the government pay all of the health care costs of AIDS patients; (2) Develop a government information program to promote safe sex practices, such as the use of condoms. Table 1 summarizes opinions about health care costs (H) and the information program (I), classified also by the respondent's gender (G).

Gender (G)	Information Opinion (I)	Health Opinion (H)	
		Support	Oppose
Male	Support	76	160
	Oppose	6	25
Female	Support	114	181
	Oppose	11	48

Table 1: Source: 1988 General Social Survey, National Opinion Research Center.

1. Compute the marginal GH-table
2. For the GH-table, compute the MLE of the marginal odds ratio, the confidence intervals. Interpret the result. (sig. level 5%)
3. Perform a hypothesis test, in order to test if the Information Opinion and the Health Opinion are independent at each level of the Gender. (sig. level 5%)
4. Compute the conditional IH odds ratio at each level of the Gender. Interpret the result.

[1]R-script is available to double check.

Solution 6.

Exercise from PJC' notes

Exercise 7. Let X and Y be discrete random variables with possible values $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ respectively and such that $p_{ij} > 0$ for all i and j. Here, p_{ij} denotes

$$\Pr(X = x_i \cap Y = y_j).$$

Let $p_{i|j}$ denote p_{ij}/p_j where p_j denotes $\sum_i p_{ij}$ so that $p_j = \Pr(Y = y_j)$ and $p_{i|j} = \Pr(X = x_i | Y = y_j)$. Also let p_i denote $\sum_j p_{ij}$.

Then the following statements are equivalent:

1. X and Y are independent;
2. $p_{i|j}$ does not depend on j for all i;
3. there exist $g_1, \ldots, g_m$ and $h_1, \ldots, h_n$ such that $p_{ij} = g_i h_j$ for all i and j;
4. the "odds"
$$\frac{p_{i|j}}{p_{i'|j}}$$
for values of X given $Y = y_j$ do not depend on j for all i and i';
5. the "odds ratios"
$$\left(\frac{p_{i|j}}{p_{i'|j}}\right) \Big/ \left(\frac{p_{i|j'}}{p_{i'|j'}}\right) = 1$$
for all i, i', j and j'.

Solution 8.

- (2) trivially implies (4).
- From (3), $p_j = \sum_i (g_i h_j) = h_j \sum_i g_i$ and so $p_{i|j} = g_i / \sum_i' g_{i'}$ which does not depend on j. Thus (3) implies (2).
- From (4)
$$\frac{p_{i|j}}{p_{i'|j}} = k_{ii'}$$
for some matrix k and therefore $p_{ij} = k_{i1} p_{1j}$ and taking $g_i = k_{i1}$ and $h_j = p_{1j}$, we have (3). Thus (4) implies (3).

- From (1), $p_{ij} = p_i p_j$ and this is (3) if we take $g_i = p_i$ and $h_j = p_j$.

  Thus (1) implies (3).

- From (2), there exist k_i such that $p_{i|j} = k_i$ for all i and j. This implies that $p_{ij} = k_i p_j$ which implies that $p_i = k_i \sum_j p_j = k_i$ and so $p_{ij} = p_i p_j$ for all i and j.

  Thus (2) implies (1).

- (4) and (5) are trivially equivalent.

The first three bullet points show that (2), (3) and (4) are equivalent. The next two bullets show that they are also equivalent to (1) and the final bullet point that they are equivalent to (5).

The next exercise is from Problem Class 1

Exercise 9. [2] The 674 subjects classified in Table 2 were the defendants in indictments involving cases with multiple murders in Florida between 1976 and 1987. The variables in Table 2 are

Y: death penalty verdict, with categories ($j = 1$: Yes, 2: No)

X: race of Defendant, with categories ($i = 1$: White, 2: Black)

Z: race of Victim, with categories ($k = 1$: White, 2: Black)

		Death Penalty (Y)	
Victim's Race (Z)	Defendant's Race (X)	Yes	No
White	White	53	414
	Black	11	37
Black	White	0	16
	Black	4	139

Table 2: Death Penalty Verdict by Defendant's Race and Victims' Race

HINT: Regarding the computation of Odds ratio, if the matrix has ZERO cells we use a correction by adding $+0.5$ at each cell. Precisely,

$$\theta = \frac{(n_{11} + 0.5)(n_{22} + 0.5)}{(n_{21} + 0.5)(n_{12} + 0.5)}$$

if any of n_{11}, n_{12}, n_{21}, n_{22} is zero. This is just a treatment, so that we do not get in infinite numbers

E.g., in this exercise, as the $(1, 1, 2)$th cell is $n_{112} = 0$, the above remedy is applied for the computation of $\theta^{XY}_{(2)}$, $\theta^{ZY}_{(1)}$, etc...

[2] R-script is available to double check. `https://htmlpreview.github.io/?https://github.com/georgios-stats/Topics_in_Statistics_Michaelmas_2020/blob/master/Contigency_Tables/q7_R.nb.html`

1. Compute the XY-marginal table, and the marginal counts
2. Based on the XY-marginal table,
   (a) Compute the conditional proportion of the White defendants received a death penalty, and that of the Black defendants received a death penalty. What do you observe?
   (b) Compute the MLE of the odd ratio and the asymptotic confidence interval at 5% sig. level. What is the association between Death Penalty and the Defendant's Race?
   (c) Perform a Goodness of fit test to test if the Death Penalty and the Defendant's Race are independent at 5% sig. level.
3. Test the Hypothesis that the Death Penalty and the Defendant's Race are independent across the Victim's Race levels at sig. level 5%.
4. Based on the XY-partial table,
   (a) Compute the MLE of the conditional XY odds ratios at each level of Victim's Race (Z),
   (b) Test the hypothesis that the Death Penalty and Defendant's Race are independent when the Victim is White, and against the alternative hypothesis that it is more likely for the Death penalty to be imposed for a Black defendant when the Victim is White. (Sig. level 5%)
   (c) Test the hypothesis that the Death Penalty and Defendant's Race are independent when the Victim is Black, and against the general alternative hypothesis that they are dependent when the Victim in Black. (Sig. level 5%)
5. Test the hypothesis that the Death Penalty and the Victim's Race are independent across the Defendant's Race levels (Sig. level 5%).
6. Compute the ZY conditional odds ratios for each level of the Defendant's Race, and discuss what they imply.
7.
   (a) Compute the marginal YZ-table.
   (b) Compute the marginal YZ-odds ratios.
   (c) Test the hypothesis that the Death Penalty and the Victims Race are independent against the alternative that it is more likely for a Death Penalty to be imposed when the Victim is White, than what it is when the Victim is Black. (Sig. level 5%)
8.
   (a) Compute the marginal XZ-table.

(b) Compute the marginal XZ-odds ratio.

(c) Test the hypothesis that the Defendant's Race and the Victims rate are independent, against the alternative that it is more likely for a Victim to be White when the defendant is White, than what it is when the Defendant is Black. (Sig. level 5%)

9. Any comments?

Solution 10.

1. The XY-marginal table is

	Death Penalty (Y)		
Defendant's Race (X)	Yes	No	Total
White	53	430	483
Black	15	176	191
Total	68	606	674

2.

(a) The conditional proportion of the White defendants received a death penalty is

$$p_{j=1|i=1} = \frac{53}{483} = 0.109$$

Namely, the 11% of the White defendants received death penalty

The conditional proportion of the Black defendants received a death penalty is

$$p_{j=1|i=2} = \frac{15}{191} = 0.079$$

Namely, the 7.9% of the White defendants received death penalty

(b) The MLE of the marginal odds ratio is

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{53 \times 176}{15 \times 430} = 1.446202$$

The asymptotic 95% CI for $\log(\theta)$ is

$$\begin{aligned} \{\log(\hat{\theta}) \pm z_{1-\frac{0.05}{2}} \sqrt{\sum_{i,j} \frac{1}{n_{i,j}}}\} \\ \{\log(1.446202) \pm 1.959964 \times \sqrt{0.09354202}\} \\ \{-0.2305073\, ,\, 0.9683883\} \end{aligned}$$

So the 95% CI for the marginal odds ratio θ is

$$(0.7941306\,,\,2.6336964)$$

The observed XY-marginal odds ratio implies that it is more likely a death penalty to be received given that the defender is White.

(c) The pair of hypothesis test is

$$\begin{aligned} &\mathrm{H}_0 : \text{X and Y are independent} \\ &\mathrm{H}_1 : \text{X and Y are not independent} \end{aligned}$$

The MLE of the counts under the null hypothesis are

$$\hat{\mu}_{i,j} = \frac{n_{i,+}n_{+,j}}{n_{+,+}}, \text{ for } i = 1, 2,\ j = 1, 2$$

I calculate

X	Y	$n_{i,j}$	$\hat{\mu}_{i,j}$	$\frac{(n_{i,j}-\hat{\mu}_{i,j})^2}{\hat{\mu}_{i,j}}$	$n_{i,j} \log \frac{n_{i,j}}{\hat{\mu}_{i,j}}$
White	Yes	53	48.72997	0.3741671	8.903753
White	No	430	434.27	0.04198575	-8.497935
Black	Yes	15	19.27003	0.9461923	-7.515025
Black	No	176	171.73	0.10617339	8.645364
Total				1.468519	1.536156

The chi-square Pearson's chi-square statistic

$$X^2 = \sum_{i,j} \frac{(n_{i,j} - \hat{\mu}_{i,j})^2}{\hat{\mu}_{i,j}}$$

The Rejection area at $a = 5\%$ sig. level is

$$R = \{X^{obs} \geq \chi^2_{df,1-a}\} = \{1.468519 \geq 3.841459\}$$

The likelihood ratio statistics is

$$G^2 = 2 \sum_{i,j} n_{i,j} \log \frac{n_{i,j}}{\hat{\mu}_{i,j}}$$

The Rejection area at $a = 5\%$ sig. level is

$$R = \{X^{obs} \geq \chi^2_{df,1-a}\} = \{1.536156 \geq 3.841459\}$$

It cannot reject the null hypothesis that the Death Penalty and the Dependence race are independent at 5% sig. level.

3. We perform the Mantel-Haenszel Chi-Squared Test

$$\begin{cases} \text{H}_0: & X, Y \text{are independent accross the partial tables at each level of } Z \\ \text{H}_1: & X, Y \text{are not independent accross the partial tables at each level of } Z \end{cases}$$

$$\Longrightarrow \begin{cases} \text{H}_0: & \theta^{xy}_{(1)} = \theta^{xy}_{(2)} = 1 \\ \text{H}_1: & \theta^{xy}_{(1)} \neq \theta^{xy}_{(2)} \end{cases}$$

The statistic is

$$T_{MH} = \frac{[\sum_k (n_{11k} - \mu_{11k})]^2}{\sum_k \sigma^2_{11k}} \xrightarrow{D} \chi^2_{df}$$

where

$$\mu_{11k} = \frac{n_{1+k} n_{+1k}}{n_{++k}}$$

and

$$\sigma^2_{11k} = \frac{n_{1+k} n_{2+k} n_{+1k} n_{+2k}}{n^2_{++k}(n_{++k} - 1)}$$

We reject the Null hypothesis for large values. The Rejection area at a sig. level is

$$R = \{T^{obs}_{MH} \geq \chi^2_{df,1-a}\}$$

Calculations:

$n_{i,+,k}$	1	2
1	467	16
2	48	155

$n_{+,j,k}$	1	2
1	64	4
2	451	155

$n_{+,+,k}$	515	159

$\hat{\mu}_{1,1,k}$	58.03495	0.4025157

$\sigma^2_{1,1,k}$	4.746102	0.3551374

The observed statistic is

$$T^{obs}_{MH} = \frac{29.56605}{5.101239} = 5.7959$$

The degrees of freedom are

$$df = 1$$

and the critical value at $a = 5\%$ sig. level is

$$\chi^2_{1,0.95} = 3.841459$$

Because the Rejection area at $a = 5\%$ sig. level is

$$R = \{T^{obs}_{MH} \geq \chi^2_{df,1-a}\}$$

I reject the null hypothesis that the the Death Penalty and the Defendant's Race are independent across the Victim's Race levels at sig. level 5%

4.

(a) When the Victim is White the MLE of the partial odds ratio is

$$\hat{\theta}^{XY}_{(1)} = \frac{n_{111}n_{221}}{n_{121}n_{211}} = 0.4306105$$

showing negative dependence between X and Y given Z=1.

When the Victim is Black the MLE of the marginal odds ratio is

$$\hat{\theta}^{XY}_{(2)} = \frac{(n_{112} + 0.5)(n_{222} + 0.5)}{(n_{122} + 0.5)(n_{212} + 0.5)} = 0.9393939$$

implying a slight negative dependence between X and Y given Z=2.

(b) The pair of hypothesis tests is

$$\begin{cases} \mathrm{H}_0 : & \theta^{xy}_{(1)} = 1 \\ \mathrm{H}_1 : & \theta^{xy}_{(1)} < 1 \end{cases}$$

The statistic is

$$Z = \frac{\log(\hat{\theta}^{xy}_{(1)}) - \log(1)}{\sqrt{\sum_{i,j} \frac{1}{n_{i,j,1}}}} \xrightarrow[D]{H_0} \mathrm{N}(0,1)$$

and because I reject for small values the rejection area at a sig. level is

$$R = \{Z^{obs} \leq -z_a\}$$

The observed statistic is

$$Z^{obs} = \frac{-0.8425513 - 0}{0.3731213} = -2.258117$$

and the critical value is $z_{0.05} = -1.644854$. Therefore, at sig. level 5%, I reject the Null hypothesis against the alternative that it is more likely for the Death penalty to be imposed for a Black defendant than a White one when the Victim is White.

(c) The pair of hypothesis tests is

$$\begin{cases} \mathrm{H}_0: & \theta^{xy}_{(2)} = 1 \\ \mathrm{H}_1: & \theta^{xy}_{(2)} \neq 1 \end{cases}$$

The statistic is

$$Z = \frac{\log(\hat{\theta}^{xy}_{(2)}) - \log(1)}{\sqrt{\sum_{i,j} \frac{1}{n_{i,j,1}}}} \xrightarrow[D]{H_0} \mathrm{N}(0,1)$$

and because I reject for small values the rejection area at a sig. level is

$$R = \{|Z^{obs}| \geq z_{1-\frac{a}{2}}\}$$

The observed statistic is

$$Z^{obs} = \frac{-0.0625204 - 0}{1.513274} = -0.04131464$$

and the critical value is $z_{0.975} = 1.959964$. Therefore, at sig. level 5%, I cannot reject the null hypothesis that the Death Penalty and Defendant's Race are independent when the Victim is Black.

5. We perform the Mantel-Haenszel Chi-Squared Test

$$\begin{cases} \mathrm{H}_0: & Z, Y \text{ are indipendent accross the partial tables at each level of } X \\ \mathrm{H}_1: & Z, Y \text{ are not indipendent accross the partial tables at each level of } X \end{cases}$$

$$\Longrightarrow \begin{cases} \mathrm{H}_0: & \theta^{ZY}_{(1)} = \theta^{ZY}_{(2)} = 1 \\ \mathrm{H}_1: & \theta^{ZY}_{(1)} \neq \theta^{ZY}_{(2)} \end{cases}$$

The statistic is

$$T_{MH} = \frac{\sum_i (n_{i11} - \mu_{i11})^2}{\sum_k \sigma^2_{i11}} \xrightarrow{D} \chi^2_{df}$$

where

$$\mu_{i11} = \frac{n_{i+1} n_{i1+}}{n_{i++}}$$

and

$$\sigma^2_{i11} = \frac{n_{i+1} n_{i+2} n_{i1+} n_{i2+}}{n^2_{i++}(n_{i++} - 1)}$$

and we reject the Null hypothesis for large values. The Rejection area at a sig. level is

$$R = \{T^{obs}_{MH} \geq \chi^2_{df,1-a}\}$$

Calculations:

$n_{i,+,k}$	1	2
1	467	16
2	48	143

$n_{i,j,+}$	1	2
1	53	430
2	15	176

$n_{i,+,+}$	483	191

$\hat{\mu}_{i,1,+}$	51.24	3.769

$\sigma^2_{i,1,1}$	1.514398	2.614333

The observed statistic is

$$T_{MH}^{obs} = \frac{80.74928}{4.12873} = 19.5579$$

The degrees of freedom are

$$df = 1$$

and the critical value at $a = 5\%$ sig. level is

$$\chi^2_{1,0.95} = 3.841459$$

Because the Rejection area at $a = 5\%$ sig. level is

$$R = \{T_{MH}^{obs} \geq \chi^2_{df,1-a}\}$$

I reject the null hypothesis that the Death Penalty and the Victim's Race are independent across the Defendant's Race levels at sig. level 5%.

6. When the Defendant is White the MLE of the marginal odds ratio is

$$\hat{\theta}^{ZY}_{(1)} = \frac{(n_{111} + 0.5)(n_{122} + 0.5)}{(n_{112} + 0.5)(n_{121} + 0.5)} = 4.259349$$

showing positive dependence between Z and Y given X=1.

When the Defendant is Black the MLE of the marginal odds ratio is

$$\hat{\theta}^{ZY}_{(2)} = \frac{n_{211}n_{222}}{n_{212}n_{221}} = 10.33108$$

showing positive dependence between Z and Y given X=2.

- regardless of defendant's race, the death penalty was considerably more likely when the victims were white than when the victims were black.

7.

(a) The YZ marginal table is

	Death Penalty (Y)	
Victim's Race (Z)	Yes	No
White	64	451
Black	4	155

(b) The observed marginal odds ratio θ^{YZ} is

$$\hat{\theta}^{YZ} = 5.498891$$

The odds ratio shows that the death penalty was considerably more likely when the victims were white than when the victims were black.

(c) The pair of hypothesis tests is

$$\begin{cases} \mathrm{H}_0: & \theta^{YZ} = 1 \\ \mathrm{H}_1: & \theta^{YZ} > 1 \end{cases}$$

The statistic is

$$Z = \frac{\log(\hat{\theta}^{YZ}) - \log(1)}{\sqrt{\sum_{i,j} \frac{1}{n_{i,j,1}}}} \xrightarrow[D]{H_0} \mathrm{N}(0,1)$$

and because I reject for large values the rejection area at a sig. level is

$$R = \{Z^{obs} \geq z_{1-a}\}$$

The observed statistic is

$$Z^{obs} = \frac{1.704546 - 0}{0.5237308} = 3.25$$

and the critical value is $z_{0.95} = 1.644854$. Therefore, at sig. level 5%, I reject the null hypothesis against the alternative that it is more likely for a Death Penalty to be imposed when the Victim is White, than what it is when the Victim is Black.

8.

(a) The XZ marginal table is

	Victim's Race (Z)	
Defendant's Race (X)	White	Black
White	467	16
Black	48	143

(b) The observed marginal odds ration θ^{XZ} is

$$\hat{\theta}^{XZ} = 86.95443$$

The odds that a white defendant had white victims are estimated to be 87.0 times the odds that a black defendant had white victims.

(c) The pair of hypothesis tests is

$$\begin{cases} \mathrm{H}_0: & \theta^{XZ} = 1 \\ \mathrm{H}_1: & \theta^{XZ} > 1 \end{cases}$$

The statistic is

$$Z = \frac{\log(\hat{\theta}^{XZ}) - \log(1)}{\sqrt{\sum_{i,j} \frac{1}{n_{i,j,1}}}} \xrightarrow[D]{H_0} \mathrm{N}(0,1)$$

and because I reject for large values the rejection area at a sig. level is

$$R = \{Z^{obs} \geq z_{1-a}\}$$

The observed statistic is

$$Z^{obs} = \frac{4.465384 - 0}{0.304085} = 14.68466$$

and the critical value is $z_{0.95} = 1.644854$. Therefore, at sig. level 5%, I reject the null hypothesis against the alternative that it is more likely for a Victim to be white when the defendant is White.

9. This is called 'Simpson's Paradox'; namely, when the X, Y marginal association has a different direction from X,Y conditional associations at each each level of Z.

   Here we observed that:

   - According to (Q. 2): overall, the death penalty was appeared to be imposed more often for White Defendants than for Black Defendants. However, we cannot reject the hypothesis that the Defendant's Race and the Death Penalty are marginally independent at sig. level 5%.
   - According to (Q. 4): when the Victim was White, the death penalty was imposed more often for Black Defendants than for White Defendants, and this association is sig. at sig. level 5%.
   - According to (Q. 4): when the Victim was Black, the death penalty was imposed more often for Black Defendants than for White Defendants. However, at sig. level 5%, the Death Penalty and Defendant's Race are independent when the Victim is Black.

- According to (Q. 7), overall, at sig. level 5%, it is more likely for a Victim to be White when the defendant is White.
- According to (Q. 6), regardless of defendant's race, the death penalty was considerably more likely when the Victims were White than when the Victims were Black.
- The explanation is the following. The data imply that whites are tending to kill whites, and killing whites is more likely to result in the death penalty. This suggests that the marginal association should show a greater tendency for white defendants to receive the death penalty than do the conditional associations.
  - [Agresti (2007) An Introduction to Categorical Data Analysis, Wiley]
