DATA 606: Statistical Methods in Data Science

—— Introduction of contingency table

Wenjun Jiang

Department of Mathematics & Statistics
The University of Calgary

Lecture 8

Contingency table

Definition 1 (Contingency table)

A rectangular table having I rows of categories of X and J columns for categories of Y displays the IJ possible combinations of outcomes. A table that contains the frequency counts of these outcomes is called a contingency table.

Example 1

Contingency table

Definition 1 (Contingency table)

A rectangular table having I rows of categories of X and J columns for categories of Y displays the IJ possible combinations of outcomes. A table that contains the frequency counts of these outcomes is called a contingency table.

Example 1

	Fatal attach	Nonfatal attack	No attack
Placebo	18	171	10845
Aspirin	5	99	10933

Table 1: Whether aspirin intake reduces mortality from cardiovascular disease.

Distributions for contingency table

In some applications, both X and Y are response variables.

- π_{ij}: the probability that (X, Y) occurs in the cell in row i and column j.

	Y_1	Y_2	$\cdots$	Y_m
X_1	π_{11}	π_{12}	$\cdots$	π_{1m}
X_2	π_{21}	π_{22}	$\cdots$	π_{2m}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
X_n	π_{n1}	π_{n2}	$\cdots$	π_{nm}

- Marginal distribution:

$$\mathbf{P}(X = X_i) = \pi_{i+} = \pi_{i1} + \cdots + \pi_{im} = \sum_{j=1}^{m} \pi_{ij},$$

$$\mathbf{P}(Y = Y_j) = \pi_{+j} = \pi_{1j} + \cdots + \pi_{nj} = \sum_{i=1}^{n} \pi_{ij}.$$

Distributions for contingency table

- Conditional probability $\pi_{j|i}$: given the outcome is in row i, then the probability of the outcome appears in column j.

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}},$$
$$\pi_{i|j} = \frac{\pi_{ij}}{\pi_{+j}}.$$

- In the experiment, the probabilities π_{ij} are estimated via

$$\hat{\pi}_{ij} = \frac{s_{ij}}{s}, \quad s = \sum_{i=1}^{n} \sum_{j=1}^{m} s_{ij},$$

where s_{ij} is the frequency count of the outcome in row i and column j.

An example

Example 2 (Medical diagnose)

PSA blood test for prostate cancer, mammogram for breast cancer, etc. A diagnostic test for a condition is *positive* is the condition is present and negative if absent.

Cancer	Positive	Negative	Total
Yes	0.86	0.14	1.00
No	0.12	0.88	1.00

Independence

- Two random variables, e.g. X and Y, are independent if

$$\mathbf{P}(X = x, Y = y) = \mathbf{P}(X = x) \cdot \mathbf{P}(Y = y).$$

Independence

- Two random variables, e.g. X and Y, are independent if

$$\mathbf{P}(X = x, Y = y) = \mathbf{P}(X = x) \cdot \mathbf{P}(Y = y).$$

- In a joint table, we know

$$\mathbf{P}(X = X_i, Y = Y_j) = \pi_{ij},$$

$$\mathbf{P}(X = X_i) = \pi_{i+}, \quad \mathbf{P}(Y = Y_j) = \pi_{+j}.$$

Independence

- Two random variables, e.g. X and Y, are independent if

$$\mathbf{P}(X = x, Y = y) = \mathbf{P}(X = x) \cdot \mathbf{P}(Y = y).$$

- In a joint table, we know

$$\mathbf{P}(X = X_i, Y = Y_j) = \pi_{ij},$$

$$\mathbf{P}(X = X_i) = \pi_{i+}, \quad \mathbf{P}(Y = Y_j) = \pi_{+j}.$$

- X and Y are independent in a joint table if

$$\pi_{ij} = \pi_{i+} \cdot \pi_{+j}.$$

  In other words,

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} = \pi_{+j}.$$

An example

Example 3 (Lung cancer)

Smoker	With lung cancer	Without lung cancer
Yes	688	650
No	21	59

Population total $n = 688 + 650 + 21 + 59 = 1418$.

$$\mathbf{P}(\text{smoke and with lung cancer}) = \frac{688}{1418} \approx 0.485.$$

$$\mathbf{P}(\text{smoke}) = \frac{688 + 650}{1418} \approx 0.944,$$

$$\mathbf{P}(\text{with lung cancer}) = \frac{688 + 21}{1418} \approx 0.5.$$

Compare the proportions

- Many response variables are binary (success or failure), we suppose our explanatory variables are also binary, which yields a 2×2 contingency table.

Compare the proportions

- Many response variables are binary (success or failure), we suppose our explanatory variables are also binary, which yields a 2×2 contingency table.

- Let $\pi_{1|i}$ be the success probability for ith explanatory variable, which is shortened as π_i.

Compare the proportions

- Many response variables are binary (success or failure), we suppose our explanatory variables are also binary, which yields a 2×2 contingency table.
- Let $\pi_{1|i}$ be the success probability for ith explanatory variable, which is shortened as π_i.
- A direct way to compare two explanatory variables is $\pi_1 - \pi_2$.

	Success	Failure
X_1	π_1 $(\pi_{1\|1})$	$1-\pi_1$ $(\pi_{2\|1})$
X_2	π_2 $(\pi_{1\|2})$	$1-\pi_2$ $(\pi_{2\|2})$

Table 2: The conditional table.

- Relative risk: $\frac{\pi_1}{\pi_2}$.

Compare the proportions

- For success probability π, the odds are defined as

$$\Omega = \frac{\pi}{1-\pi}.$$

In a 2×2 table, we have two odds Ω_1 and Ω_2. The *odds ratio* is defined as

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}.$$

Compare the proportions

- For success probability π, the odds are defined as

$$\Omega = \frac{\pi}{1-\pi}.$$

In a 2×2 table, we have two odds Ω_1 and Ω_2. The *odds ratio* is defined as

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}.$$

- For a joint distribution table $\{\pi_{ij}\}$, the odds ratio is defined as

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

	Y_1	Y_2
X_1	π_{11}	π_{12}
X_2	π_{21}	π_{22}

Compare the proportions

- In the case of independence, the odds ratio for a joint distribution table is

$$\theta = \frac{\pi_{1+}\pi_{+1}\pi_{2+}\pi_{+2}}{\pi_{1+}\pi_{+2}\pi_{2+}\pi_{+1}} = 1.$$

- Conversely, if $\theta = 1$ then X and Y are independent (proof ignored).

- If we only have cell counts $\{s_{ij}\}$, then an estimate for θ is

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

Conditional association

In order to study the relationship between X and Y, we need to control other covariates that would influence that relationship.

- A three-way contingency table cross-classifies X, Y and Z. We control Z to study $X - Y$ relationship.
- The partial table refers to the sub-table cross-classify X and Y at separate categories of Z.
- The two-way contingency table obtained by combining the partial tables is called the XY *marginal table*.
- The associations in partial tables are called *conditional associations*.

Example 4 (Racial characteristics and the death penalty)

Table 2.6 Death Penalty Verdict by Defendant's Race and Victims' Race

Victims' Race	Defendant's Race	Death Penalty Yes	Death Penalty No	Percent Yes
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8
Total	White	53	430	11.0
	Black	15	176	7.9

Source: M. L. Radelet and G. L. Pierce, *Florida Law Rev.* **43**: 1–34, 1991. Reprinted with permission from the *Florida Law Review*.

Conditional and marginal odds ratios

We use μ_{ijk} to denote the cell expected frequencies where i refers to the category of X, j refers to the category of Y and k refers to the category of Z.

- Within a fixed category k of Z, the odds ratio

$$\theta_{XY(k)} = \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}}$$

  describes conditional XY association in partial table k.
- The marginal XY table has expected frequencies $\{\mu_{ij+} = \sum_k \mu_{ijk}\}$. Then XY marginal odds ratio is

$$\theta_{XY} = \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}}.$$

An example (cont.)

- If the victim's race is white, then

$$\hat{\theta}_{XY(1)} = \frac{53 \times 37}{414 \times 11} = 0.43.$$

- If the victim's race is black, then

$$\hat{\theta}_{XY(2)} = \frac{0 \times 139}{16 \times 4} = 0.$$

- In the marginal table,

$$\hat{\theta}_{XY} = \frac{53 \times 176}{430 \times 15} = 1.45.$$

Marginal independence v.s. conditional independence

		Y_1	Y_2		
Z_1	X_1	π_{111}	π_{121}	π_{1+1}	π_{++1}
	X_2	π_{211}	π_{221}	π_{2+1}	
		π_{+11}	π_{+21}		
Z_2	X_1	π_{112}	π_{122}	π_{1+2}	π_{++2}
	X_2	π_{212}	π_{222}	π_{2+2}	
		π_{+12}	π_{+22}		
Total	X_1	π_{11+}	π_{12+}	π_{1++}	1
	X_2	π_{21+}	π_{22+}	π_{2++}	
		π_{+1+}	π_{+2+}		

Table 3: Three-way joint distribution table.

Marginal independence v.s. conditional independence

- *Conditional independence*

  X and Y are said to be conditionally independent at level K of Z if

  $$\mathbf{P}(Y = j | X = i, Z = k) = \mathbf{P}(Y = j | Z = k).$$

  Conditional independence is equivalent to (proof ignored)

  $$\pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k}.$$

- *Marginal independence*: $\pi_{ij+} = \pi_{i++}\pi_{+j+}$.

An example

Example 5 (Clinic treatment)

Table 2.7 Expected Frequencies Showing that Conditional Independence Does Not Imply Marginal Independence

		Response	
Clinic	Treatment	Success	Failure
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32
Total	A	20	20
	B	20	40

An example

- First partial table

$$\hat{\theta}_{XY(1)} = \frac{18 \times 8}{12 \times 12} = 1.$$

- Second partial table

$$\hat{\theta}_{XY(1)} = \frac{2 \times 32}{8 \times 8} = 1.$$

- Marginal table

$$\hat{\theta}_{XY} = \frac{20 \times 40}{20 \times 20} = 2.$$