Quantitative Methods – I
A.Y. 2020-21

Practice 2

Lorenzo Cavallo

For any clarification/meeting: cavallo@istat.it

THEME #1

—

Contingency Tables

Contingency Table

A contingency table, sometimes called a two-way frequency table, is a tabular mechanism with at least two rows and two columns used in statistics to present categorical data in terms of frequency counts. More precisely, an r×c contingency table shows the observed frequency of two variables, the observed frequencies of which are arranged into r rows and c columns.

The intersection of a row and a column of a contingency table is called a cell.

gender	*cup*	*cone*	*sundae*	*sandwich*	*other*	**Total**
male	592	300	204	24	80	**1200**
female	410	335	180	20	55	**1000**
Total	**1002**	**635**	**384**	**44**	**135**	**2200**

For example, the above contingency table has two rows and five columns (not counting header rows/columns) and shows the results of a random sample of 2200 adults classified by two variables, namely gender and favorite way to eat ice cream.

In a general framework to summarize the absolute frequencies of two discrete variables in contingency tables we use the following notations:
let $x_1, x_2,\ldots, x_k$ be the k classes of a variable X and let $y_1, y_2,\ldots, y_l$ be the l classes of a variable Y

It is possible to summarize the absolute frequencies n_{ij} related to (x_i , y_j), i = 1, 2, . . . , k, j = 1, 2, . . . , l, in a k × l **contingency table**.

		Y					
		y_1		y_j		y_l	Total (rows)
X	x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1l}	n_{1+}
	x_2	n_{21}	$\cdots$	n_{2j}	$\cdots$	n_{2l}	n_{2+}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{il}	n_{i+}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_k	n_{k1}	$\cdots$	n_{kj}	$\cdots$	n_{kl}	n_{k+}
	Total (columns)	n_{+1}	$\cdots$	n_{+j}	$\cdots$	n_{+l}	n

The frequencies n_{ij} represent the **joint frequency distribution** of X and Y
The frequencies n_{i+} represent the **marginal frequency distribution of X**.
The frequencies n_{+j} represent the **marginal frequency distribution of Y**

		Y					
		y_1		y_j		y_l	Total (rows)
X	x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1l}	n_{1+}
	x_2	n_{21}	$\cdots$	n_{2j}	$\cdots$	n_{2l}	n_{2+}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{il}	n_{i+}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_k	n_{k1}	$\cdots$	n_{kj}	$\cdots$	n_{kl}	n_{k+}
	Total (columns)	n_{+1}	$\cdots$	n_{+j}	$\cdots$	n_{+l}	n

The frequencies n_{ij} represent the **joint frequency distribution** of X and Y (gender="male", ice-cream="cone", 300 male that prefers the cone)

gender	*cup*	*cone*	*sundae*	*sandwich*	*other*	**Total**
male	592	300	204	24	80	**1200**
female	410	335	180	20	55	**1000**
Total	**1002**	**635**	**384**	**44**	**135**	**2200**

The frequencies n_{+j} represent the **marginal frequency distribution of Y** (favourite way to eat the ice-cream)

The frequencies n_{i+} represent the **marginal frequency distribution of X** (adult by gender)

The frequencies f_{ij} represent the **joint relative frequency distribution of X and Y**

gender	*cup*	*cone*	*sundae*	*sandwich*	*Other*	**Total**
male	$\frac{592}{2200}$	$\frac{300}{2200}$	$\frac{204}{2200}$	$\frac{24}{2200}$	$\frac{80}{2200}$	$\frac{1200}{2200}$
female	$\frac{410}{2200}$	$\frac{335}{2200}$	$\frac{180}{2200}$	$\frac{20}{2200}$	$\frac{55}{2200}$	$\frac{1000}{2200}$
Total	$\frac{1002}{2200}$	$\frac{635}{2200}$	$\frac{384}{2200}$	$\frac{44}{2200}$	$\frac{135}{2200}$	$\frac{2200}{2200}$

The **marginal frequency distributions** are displayed in the last column and last row, respectively

The **conditional frequency distributions** give us an idea about the behaviour of one variable when the other one is kept fixed.

		Y					
		y_1		y_j		y_l	Total (rows)
X	x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1l}	n_{1+}
	x_2	n_{21}	$\cdots$	n_{2j}	$\cdots$	n_{2l}	n_{2+}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{il}	n_{i+}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_k	n_{k1}	$\cdots$	n_{kj}	$\cdots$	n_{kl}	n_{k+}
	Total (columns)	n_{+1}	$\cdots$	n_{+j}	$\cdots$	n_{+l}	n

The relative frequencies of variable X, conditional on value

$$f_{i|j}^{X|Y} = \frac{n_{ij}}{n_{+j}} = \frac{f_{ij}}{f_{+j}}, \quad i = 1, 2, \ldots, k.$$

The relative frequencies of variable Y , conditional on value

$$f_{j|i}^{Y|X} = \frac{n_{ij}}{n_{i+}} = \frac{f_{ij}}{f_{i+}}, \quad j = 1, 2, \ldots, l.$$

Notice that the conditional frequency is obtained by dividing the joint frequency by the marginal frequency of the conditioning variable.

$$f_{i|j}^{X|Y} = \frac{n_{ij}}{n_{+j}} = \frac{f_{ij}}{f_{+j}}, \quad i = 1, 2, \ldots, k.$$

gender	cup	cone	sundae	sandwich	Other	Total
male	$\frac{592}{1002}$	$\frac{300}{635}$	$\frac{204}{384}$	$\frac{24}{44}$	$\frac{80}{135}$	$\frac{1200}{2200}$
female	$\frac{410}{1002}$	$\frac{335}{635}$	$\frac{180}{384}$	$\frac{20}{44}$	$\frac{55}{135}$	$\frac{1000}{2200}$
Total	$\mathbf{\frac{1002}{1002}=1}$	$\mathbf{\frac{635}{635} = 1}$	$\mathbf{\frac{384}{384} = 1}$	$\mathbf{\frac{44}{44} = 1}$	$\mathbf{\frac{135}{135} = 1}$	$\mathbf{\frac{2200}{2200} = 1}$

$$f_{j|i}^{Y|X} = \frac{n_{ij}}{n_{i+}} = \frac{f_{ij}}{f_{i+}}, \quad j = 1, 2, \ldots, l.$$

gender	cup	cone	sundae	sandwich	Other	Total
male	$\frac{592}{1200}$	$\frac{300}{1200}$	$\frac{204}{1200}$	$\frac{24}{1200}$	$\frac{80}{1200}$	$\mathbf{\frac{1200}{1200} = 1}$
female	$\frac{410}{1000}$	$\frac{335}{1000}$	$\frac{180}{1000}$	$\frac{20}{1000}$	$\frac{55}{1000}$	$\mathbf{\frac{1000}{1000} = 1}$
Total	$\frac{1002}{2200}$	$\frac{635}{2200}$	$\frac{384}{2200}$	$\frac{44}{2200}$	$\frac{135}{2200}$	$\mathbf{\frac{2200}{2200} = 1}$

THEME #2

—

Indipendence

Independence between variables

Two variables are said to be independent if the conditional distribution of either does not vary with the value of the other.

The conditional relative frequency distributions of one variable are identical and equal to the marginal distribution. In other words, the distribution of one variable is unaffected by the other variable.

Independence is a symmetric concept

The independence condition:

$$n_{ij} = \frac{n_{i+} n_{+j}}{n}, \quad f_{ij} = f_{i+} f_{+j}$$

Measuring association in a two-way table

Starting from a bivariate distribution in the form of a two-way table, we define the theoretical independence frequencies (or expected):

$$\tilde{n}_{ij} = \frac{n_{i+}n_{+j}}{n} = Theoretical\ (or\ Expected)\ frequency = \frac{(Row\ Total)(Column\ Total)}{Total}$$

We define an index of association called Pearson's Chi-square.

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l}\left[\frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}}\right]$$

Where $0 \leq \chi^2 \leq n(\min(k, l) - 1)$

A value of χ^2 close to zero indicates a weak association and a value of χ^2 close to $n(\min(k, l) - 1)$ indicates a strong association between the two variables.

Exercise 2

A random sample of 300 adults was selected, and these adults were asked if they favor giving more freedom to their children. The two-way classification of the responses of these adults is presented in the following table.

	In favor	Against	No opinion	tot
Men (M)	93	70	12	175
Women (W)	87	32	6	125
tot	180	102	18	300

Calculate the expected frequencies for this table, assuming that the two attributes, gender and opinions on the issue, are independent.

Does the sample provide sufficient evidence to conclude that the two attributes, gender and opinions of adults, are dependent?

	In favor	Against	No opinion	tot
Men (M)	93	70	12	175
Women (W)	87	32	6	125
tot	180	102	18	300

$$Expected\ value = \tilde{n}_{ij} = \frac{(Row\ Total)(Column\ Total)}{Sample\ size}$$

E for *Men* and *In Favor* cell = (175)(180)/300 = **105.00**

E for *Men* and *Against* cell = (175)(102)/300 = **59.50**

E for *Men* and *No Opinion* cell = (175)(18)/300 = **10.50**

E for *Women* and *In Favor* cell = (125)(180)/300 = **75.00**

E for *Women* and *Against* cell = (125)(102)/300 = **42.50**

E for *Women* and *No Opinion* cell = (125)(18)/300 = **7.50**

	In Favor (F)	**Against (A)**	**No Opinion (N)**	**Row Totals**
Men (M)	93 (105.00)	70 (59.50)	12 (10.50)	175
Women (W)	87 (75.00)	32 (42.50)	6 (7.50)	125
Column Totals	180	102	18	300

	In Favor (F)	Against (A)	No Opinion (N)	Row Totals
Men (M)	93 (105.00)	70 (59.50)	12 (10.50)	175
Women (W)	87 (75.00)	32 (42.50)	6 (7.50)	125
Column Totals	180	102	18	300

$$\chi^2 = \frac{\sum\left(n_{ij} - \tilde{n}_{ij}\right)^2}{\tilde{n}_{ij}} = \frac{(93 - 105.00)^2}{105.00} + \frac{(70 - 59.50)^2}{59.50} + \frac{(12 - 10.50)^2}{10.50} + \frac{(87 - 75.00)^2}{75.00} + \frac{(32 - 42.50)^2}{42.50} + \frac{(6 - 7.50)^2}{7.50}$$

$$= 1.371 + 1.853 + .214 + 1.920 + 2.594 + .300 = 8.252$$

$$0 \leq \chi^2 \leq n(\min(k, l) - 1) = 300 \cdot (\min(3, 2) - 1) = 300 \cdot (2 - 1) = 300$$

$\chi^2 = 8.252$ indicates a moderate association between “Gender” and “Opinion”

THEME #3

Covariance and Correlation

Covariance

Covariance is the average cross-product of the values of the two variables in deviation from their mean.

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

It is a measure of linear association between two variables.

Correlation

The **correlation coefficient** $r_{xy} = r$ measures the degree of linear relationship between X and Y.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}, \qquad -1 \leq r_{xy} \leq 1.$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}, \qquad -1 \leq r_{xy} \leq 1.$$

$$s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n}} = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$$

$$s_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n}} = \sqrt{\frac{\sum y^2}{n} - \bar{y}^2}$$

$$s_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n} = \frac{1}{n}\sum xy - \bar{x}\bar{y}$$

$$r = \frac{\frac{1}{n}\sum xy - \bar{x}\bar{y}}{\sqrt{\frac{\sum x^2}{n} - \bar{x}^2}\sqrt{\frac{\sum y^2}{n} - \bar{y}^2}}$$

The time x in years that an employee spent at a company and the employee's hourly pay, y, for 5 employees are listed in the table below. Calculate and interpret the correlation coefficient r. Include a plot of the data in your discussion.

x	y
5	25
3	20
4	21
10	35
15	38

The time x in years that an employee spent at a company and the employee's hourly pay, y, for 5 employees are listed in the table below. Calculate and interpret the correlation coefficient r. Include a plot of the data in your discussion.

x	y
5	25
3	20
4	21
10	35
15	38

$$r = \frac{\frac{1}{n}\sum xy - \bar{x}\bar{y}}{\sqrt{\frac{\sum x^2}{n} - \bar{x}^2}\sqrt{\frac{\sum y^2}{n} - \bar{y}^2}}$$

x	y	x^2	y^2	xy
5	25	25	625	125
3	20	9	400	60
4	21	16	441	84
10	35	100	1225	350
15	38	225	1444	570
$\sum x = 37$	$\sum y = 139$	$\sum x^2 = 375$	$\sum y^2 = 4135$	$\sum xy = 1189$

$n = 5$ Calculate the means

$$\bar{x} = \frac{\sum x}{n} = \frac{37}{5} = 7.4 \qquad \bar{y} = \frac{\sum y}{n} = \frac{139}{5} = 27.8$$

Calculate the correlation

$$r = \frac{\frac{1}{n}\sum xy - \bar{x}\bar{y}}{\sqrt{\frac{\sum x^2}{n} - \bar{x}^2}\sqrt{\frac{\sum y^2}{n} - \bar{y}^2}} = \frac{\frac{1}{5}\cdot 1189 - 7.4 \cdot 27.8}{\sqrt{\frac{375}{5} - 7.4^2}\sqrt{\frac{4135}{5} - 27.8^2}} = \frac{237.8 - 205{,}72}{\sqrt{75 - 54.76}\sqrt{827 - 772.84}} = \frac{32.08}{4.5 \cdot 7.36} =$$

$= 0.9686$

There is a strong positive correlation between the number of years and employee has worked and the employee's salary, since r is very close to 1.