UNIVERSITÀ
DEGLI STUDI DELLA
TUSCIA

SAPIENZA
UNIVERSITÀ DI ROMA

Statistics for business and decision making course

Dr. Ilaria Benedetti

06. Association among variables

Two way table

zone	on line selling
center	yes
outskirt	yes
Semi central	no
outskirt	no
center	no
center	no
outskirt	no
Semi central	no
centre	yes

		on line selling		Tot
		yes	no	
Zone	Center	2	2	4
	Semi central	0	2	2
	outsk	1	2	3
Tot		3	6	9

N=9 selling points

2

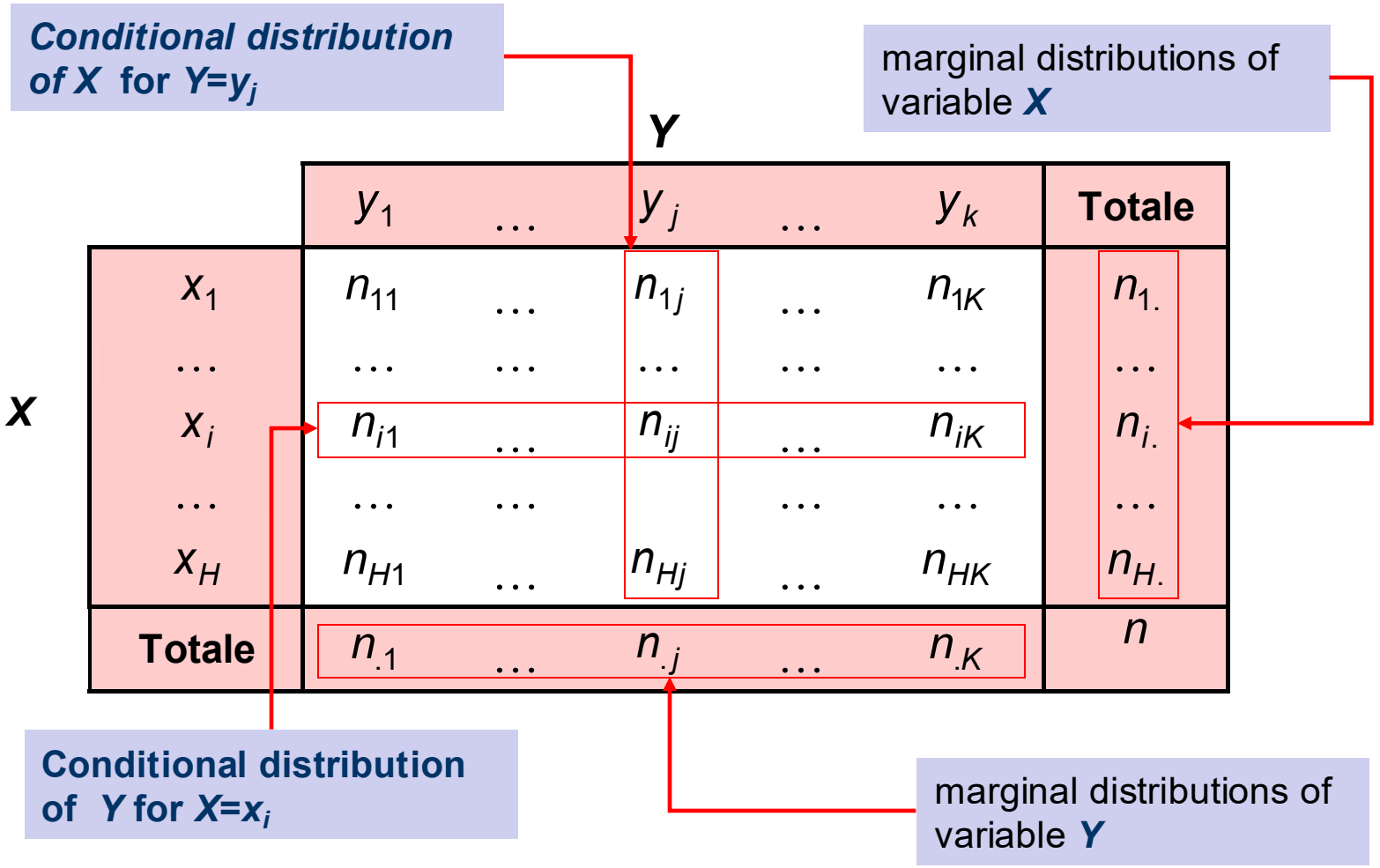
Two way table

		on line selling		Tot
		yes	no	
Zone	Center	2	2	4
	Semi central	0	2	2
	outsk	1	2	3
Tot		3	6	9

- **Identify the number of selling point located in the center**
- **Identify the number of selling point who sell on line**
- **Considering the selling point who sell on-line, identify the number of selling point located in the center**
- **Considering the selling point in the outskirt, identify the number of selling point who sell on line**

3

Two way table



Two way table

- $n_{i.} = \sum_{j=1}^{K} n_{ij}$ *per i=1,2,…,H*
- $n_{.j} = \sum_{i=1}^{H} n_{ij}$ *per j=1,2,…,K*
- $n = \sum_{i=1}^{H}\sum_{j=1}^{K} n_{ij} = \sum_{i=1}^{H} n_{i.} = \sum_{j=1}^{K} n_{.j}$

The distribution of a variable that is restricted to cases satisfying a condition is called a conditional distribution

Dependence and interdependence

Dependence: studies how the modalities of one character depend on those of another character according to a *unidirectional* link

Interdependence: It is assumed that the two characters have the same role and that the link is *bidirectional*

6

Two way table

Freq. Observed: n_{ij}

Freq. Expected: nij' in case of independence

$$n'_{ij} = \frac{n_{i.} \times n_{.j}}{n}$$

7

		Y					Tot
		y_1	…	y_j	…	y_K	
X	X_1	n_{11}		n_{1j}		n_{1k}	$n_{1.}$
	…						
	X_i	n_{i1}		n_{ij}		n_{ik}	$n_{i.}$
	…						
	x_H	n_{H1}		n_{Hj}		n_{HK}	$n_{H.}$
Tot		$n_{.1}$		$n_{.j}$		$n_{.K}$	n

Freq. obs

Freq. Used to find the exp. freq.

$$n'_{ij} = \frac{n_{i.} \times n_{.j}}{n}$$

8

Obs. Freq.

		On-line sell		Tot
		yes	no	
Zone	Center	2	2	4
	Semi – Central	0	2	2
	outskirt	1	2	3
Tot		3	6	9

9

Exp. Freq

		on line selling		Tot
		yes	**no**	
Zone	**Centre**	$\frac{4 \times 3}{9}$	$\frac{4 \times 6}{9}$	**4**
	Semi central	$\frac{2 \times 3}{9}$	$\frac{2 \times 6}{9}$	**2**
	outsk.	$\frac{3 \times 3}{9}$	$\frac{3 \times 6}{9}$	**3**
Tot		**3**	**6**	**9**

In case of statistical independence

10

Observed. Freq. vs expected freq

f. obs		on line sell		Tot
		yes	no	
zone	Centre	2	2	4
	Semi centr	0	2	2
	outsk	1	2	3
Tot		3	6	9

Teoriche		on line selling		Tot
		si	no	
zone	Center	1,33	2,67	4
	Semi centr	0,67	1,33	2
	outsk.	1	2	3
Tot		3	6	9

- **What is it the degree of association among these two var?**

compare the observed data (original table) with the «artificial» table (expected frequencies in case of no association).

11

Chi-squared compares
the observed contingency table with an artificial – expected - table with the same marginal totals.

If the tables are similar, then there's not association.

The greater the difference between the tables becomes, the greater
the association.

Summary: steps to compute chi-square metrics

1. From the table of observed freq. distribution, prepare the table of **expected frequecy distribution** (with the same marginal distributions);
2. Subtract the cells of the expected table from those of the observed table and then square the differences.
3. Divide each squared differences in each cell by expected freq. Distribution.
4. Sum the normalized values obtained in point 3. over all of the cells.

Chi-square index

$$\chi^2 = \sum_{i=1}^{H} \sum_{j=1}^{K} \frac{c_{ij}^2}{n'_{ij}} \longrightarrow c_{ij} = n_{ij} - n'_{ij}$$

$\chi^2 = 0$ ➡ Statistical Independence

$\chi^2 > 0$ ➡ Statistical interdipendence

14

• Calculating The Chi-Square Statistic

The tables below show the observed and expected counts for the wine and music experiment. Calculate the chi-square statistic.

Observed Counts

	Music			
Wine	**None**	**French**	**Italian**	**Total**
French	30	39	30	**99**
Italian	11	1	19	**31**
Other	43	35	35	**113**
Total	**84**	**75**	**84**	**243**

Expected Counts

	Music			
Wine	**None**	**French**	**Italian**	**Total**
French	34.22	30.56	34.22	**99**
Italian	10.72	9.57	10.72	**31**
Other	39.06	34.88	39.06	**113**
Total	**84**	**75**	**84**	**243**

For the French wine with no music, the observed count is 30 bottles and the expected count is 34.22. The contribution to the χ^2 statistic for this cell is

$$\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \frac{(30 - 34.22)^2}{34.22} = 0.52$$

The χ^2 statistic is the sum of nine such terms:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \frac{(30 - 34.22)^2}{34.22} + \frac{(39 - 30.56)^2}{30.56} + \ldots + \frac{(35 - 39.06)^2}{39.06}$$

$$= 0.52 + 2.33 + \ldots + 0.42 = 18.28$$

V Cramer index

Relative index of association

$$V = \sqrt{\frac{\chi^2 / n}{\min[(H-1),(K-1)]}}$$

$$0 \leq V \leq 1$$

V=0 ⇨ statistical independence – no assoc
V=1 ⇨ perfect association

If V=0.25, we will say that
the association is weak.
If V=0.75, we will say that the association is strong.
In between, we will say there is moderate association

16

χ^2 and V

$$\chi^2 = \frac{(2-1{,}33)^2}{1{,}33} + \frac{(2-2{,}67)^2}{2{,}67} +$$

$$+ \frac{(0-0{,}67)^2}{0{,}67} + \frac{(2-1{,}33)^2}{1{,}33} +$$

$$+ \frac{(1-1)^2}{1} + \frac{(2-2)^2}{2}$$

$$\Rightarrow \quad \chi^2 = 0{,}33 + 0{,}17 + 0{,}33 + 0{,}67 = 1{,}5$$

H=3, K=2 min among H-1 and K-1 in equal to 1

$$\Rightarrow \quad V = \sqrt{\frac{1{,}50}{9}} = 0{,}41$$

17

A character Y perfectly depends on X when only one mode of Y is associated with each mode of X, that is, when in a double table for each i there is only one j for which nij ≠ 0

		on line		Tot
		yes	no	
	Centre	4	0	4
	Semi central	0	2	2
	Outsk	0	3	3
Tot		5	5	9

In this table the online sale (Y) depends perfectly on the location (X) Attention X does not depend on Y

Between two characters X and Y there is perfect interdependence if each modality of one of the two characters corresponds to one and only one modality of the other character and vice versa.

		Y			Tot
		y_1	y_2	y_3	
	x_1	4	0	0	4
	x_2	0	2	0	2
	x_3	0	0	3	3
Tot		4	2	3	9

ONLY for
Squared table!

Checklist:

Chi-squared and Cramer's V. Chi-squared and Cramer's V measure association between two categorical variables that define a contingency table.

Before you use these, verify that your data meet this prerequisite:

✓ Categorical variables. If a variable is numerical, there are better ways to measure association.