Exploratory Statistical Data Analysis With R Software (ESDAR)

Swayam Prabha

Lecture 38

Association of Discrete Variables with R Software

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

http://home.iitk.ac.in/~shalab/sp

Association between Two Discrete Variables

In general, let X and Y be two discrete variables

$x_1, x_2, \ldots, x_k$: k classes of X

$y_1, y_2, \ldots, y_l$: l classes of Y

n_{ij} : Frequency of $(i, j)^{th}$ cell corresponding to (x_i, y_j)

$$i = 1,2,\ldots,k; \quad j = 1,2,\ldots,l;$$

This frequencies can be presented in the following k x l contingency table.

Association between Two Discrete Variables

k x *l* Contingency Table

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

Marginal frequency

$$n_{i+} = \sum_{j=1}^{l} n_{ij}$$

Marginal frequency

$$n_{+j} = \sum_{i=1}^{k} n_{ij}$$

$$n = \sum_{i=1}^{k} n_{i+} = \sum_{j=1}^{l} n_{+j} = \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij}$$

Total frequency

Association between Two Discrete Variables

$f_{ij} = \frac{n_{ij}}{n}$: Relative frequency

: Represents joint relative frequency distribution of X and Y.

$f_{i|j}(X|\ Y = y_j) = \frac{n_{ij}}{n_{+j}}$: Conditional frequency distribution of X given $Y = y_j$

$f_{j|i}(Y|\ X = x_i) = \frac{n_{ij}}{n_{i+}}$: Conditional frequency distribution of Y given $X = x_i$

Conditional frequency distribution tells how the values of one variable behave when another variable is kept fixed.

Association between Two Discrete Variables

Example:

A soft drink was served to children, young persons and elder persons and its taste was recorded as good or bad. The following 2 X 3 contingency table was formed by compiling the data.

	Person	Children	Young persons	Elder persons	Total (Rows)
Taste	Good	20	30	10	60
	Bad	10	15	15	40
	Total (Columns)	30	45	25	100

Association between Two Discrete Variables

Example:

The same contingency table can also be formed by relative frequencies.

	Person	Children	Young persons	Elder persons	Total (Rows)
Taste	Good	20/100	30/100	10/100	60/100
	Bad	10/100	15/100	15/100	40/100
	Total (Columns)	30/100	45/100	25/100	1

Association between Two Discrete Variables

Example:

Interpretations

Joint frequency distribution tells how the values of both the variables behave jointly.

Marginal frequency distribution:

- **60 (or 60%) persons said that the drink is good.**
- **40 (or 40%) persons said that the drink is bad.**
- **Drink was tasted by 30 (or 30%) children, 45 (or 45%) young persons and 25 (or 25%) elder persons.**

Association between Two Discrete Variables

Example:

Interpretations

Conditional frequency distribution tells how the values of one variable behave when another variable is kept fixed.

- **20/60 = 33.3% children said that the drink is good.**
- **10/40 = 25% children said that the drink is bad.**
- **30/60 = 50% young persons said that the drink is good.**
- **15/40 = 37.5% young persons said that the drink is bad etc.**

Association between Two Discrete Variables

R command:

`x,y` : **Two data vectors**

`table(x,y)` **: uses the cross-classifying factors to build a contingency table of the counts at each combination of factor levels.**

`table(x,y)` **returns a contingency table with absolute frequencies.**

`table(x,y)/length(x)` **returns a contingency table with relative frequencies.**

Association between Two Discrete Variables

R command:

`addmargins` **is used with `table()` command to add the marginal frequencies to the contingency table.**

`addmargins(table(x,y))` **adds marginal frequencies to the contingency table with absolute frequencies.**

`addmargins(table(x,y)/length(x))` **adds marginal relative frequencies to the contingency table with relative frequencies.**

Association between Two Discrete Variables

Example

Following data on 20 persons has been collected on their age category and their response to the taste of a drink.

Person No.	Age Category	Taste of Drink
1	Child	Good
2	Young person	Good
3	Elder person	Bad
4	Child	Bad
5	Young person	Good
6	Young person	Bad
7	Elder person	Good
8	Elder person	Good
9	Elder person	Good
10	Elder person	Bad

Person No.	Age Category	Taste of Drink
11	Child	Good
12	Young person	Good
13	Elder person	Bad
14	Child	Bad
15	Young person	Good
16	Young person	Bad
17	Elder person	Good
18	Elder person	Good
19	Elder person	Good
20	Elder person	Bad

Association between Two Discrete Variables

Example

```
> person = c("Child", "Young person", "Elder
person", "Child", "Young person", "Young
person", "Elder person", "Elder person", "Elder
person", "Elder person", "Child", "Young
person", "Elder person", "Child", "Young
person", "Young person", "Elder person", "Elder
person", "Elder person", "Elder person")

> taste = c("Good", "Good", "Bad", "Bad",
"Good", "Bad",  "Good", "Good", "Good", "Bad",
"Good", "Good", "Bad", "Bad", "Good", "Bad",
"Good", "Good", "Good", "Bad")
```

Association between Two Discrete Variables

Example

Contingency table with absolute frequencies

```
> table(person, taste)
              taste
person         Bad Good
  Child          2    2
  Elder person   4    6
  Young person   2    4
```

Contingency table with marginal frequencies

```
> addmargins(table(person, taste))
              taste
person         Bad Good Sum
  Child          2    2   4
  Elder person   4    6  10
  Young person   2    4   6
  Sum            8   12  20
```

Association between Two Discrete Variables

Example

R Console

```
> person
 [1] "Child"        "Young person" "Elder person" "Child"
 [5] "Young person" "Young person" "Elder person" "Elder person"
 [9] "Elder person" "Elder person" "Child"        "Young person"
[13] "Elder person" "Child"        "Young person" "Young person"
[17] "Elder person" "Elder person" "Elder person" "Elder person"
> taste
 [1] "Good" "Good" "Bad"  "Bad"  "Good" "Bad"  "Good" "Good" "Good" "Bad"
[11] "Good" "Good" "Bad"  "Bad"  "Good" "Bad"  "Good" "Good" "Good" "Bad"
> table(person, taste)
              taste
person         Bad Good
  Child          2    2
  Elder person   4    6
  Young person   2    4
```

Association between Two Discrete Variables

Example

```
> length(person)
[1] 20
```

Contingency table with relative frequencies

```
> table(person, taste)/length(person)
              taste
person         Bad Good
  Child        0.1  0.1
  Elder person 0.2  0.3
  Young person 0.1  0.2
```

Contingency table with marginal relative frequencies

```
> addmargins(table(person, taste)/length(person))
              taste
person         Bad Good Sum
  Child        0.1  0.1 0.2
  Elder person 0.2  0.3 0.5
  Young person 0.1  0.2 0.3
  Sum          0.4  0.6 1.0
```

Association between Two Discrete Variables

Example

R Console

```
> length(person)
[1] 20
> table(person, taste)/length(person)
              taste
person         Bad Good
  Child        0.1  0.1
  Elder person 0.2  0.3
  Young person 0.1  0.2
> addmargins(table(person,taste)/length(person))
              taste
person         Bad Good Sum
  Child        0.1  0.1 0.2
  Elder person 0.2  0.3 0.5
  Young person 0.1  0.2 0.3
  Sum          0.4  0.6 1.0
> 
```

Association between Two Discrete Variables

Pearson's Chi-squared (χ2) statistic

Used to measure the association between variables in a contingency table. The χ^2 statistic for *k* x *l* contingency table is given by

$$\chi^2 = \sum_{i=1}^{k} \sum_{i=1}^{l} \left[\frac{\left(n_{ij} - \dfrac{n_{i+} n_{+j}}{n} \right)^2}{\dfrac{n_{i+} n_{+j}}{n}} \right] \quad ; \quad 0 \le \chi^2 \le n\left[\min(k,l) - 1 \right]$$

where $n_{i+} = \sum_{j=1}^{l} n_{ij}, \; n_{+j} = \sum_{i=1}^{k} n_{ij}, \; n = \sum_{i=1}^{k} n_{i+} = \sum_{j=1}^{l} n_{+j} = \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij}.$

n_{ij} : Absolute frequencies

n_{i+} and n_{+j} : Marginal frequencies of *X* and *Y* respectively.

n : Total frequency

Association between Two Discrete Variables

Pearson's Chi-squared ($\chi 2$) statistics

- **Value of χ^2 close to 0 $\Rightarrow$ weak association between the two variables.**
- **Value of χ^2 close to $n[\min(k, l) - 1]$ $\Rightarrow$ strong association between the two variables.**
- **Other values will suitably indicate the degree of association between the two variables to be low-moderate-high.**

χ^2 statistc is symmetric in the sense that its value does not depend on which variable is defined as *X* and which as *Y*.

Association between Two Discrete Variables

Pearson's Chi-squared ($\chi 2$) statistics

For example:

For a 2 x 2 contingency table

		Y		Total (Rows)
		y_1	y_2	
X	x_1	a	b	$a + b$
	x_2	c	d	$c + d$
Total (Columns)		$a + c$	$b + d$	n

$$\chi^2 = \left[\frac{n\left(ad - bc\right)^2}{(a+b)(c+d)(a+c)(b+d)} \right]$$

Association between Two Discrete Variables

Example: Pearson‘s Chi-squared (χ2) statistics

Following data on 20 persons has been collected on their age category and their response to the taste of a drink.

Person No.	Age Category	Taste of Drink
1	Child	Good
2	Young person	Good
3	Elder person	Bad
4	Child	Bad
5	Young person	Good
6	Young person	Bad
7	Elder person	Good
8	Elder person	Good
9	Elder person	Good
10	Elder person	Bad

Person No.	Age Category	Taste of Drink
11	Child	Good
12	Young person	Good
13	Elder person	Bad
14	Child	Bad
15	Young person	Good
16	Young person	Bad
17	Elder person	Good
18	Elder person	Good
19	Elder person	Good
20	Elder person	Bad

Association between Two Discrete Variables

Example: Pearson‘s Chi-squared ($\chi 2$) statistic

Contingency table with absolute frequencies

```
> table(person, taste)
              taste
person         Bad Good
  Child          2    2
  Elder person   4    6
  Young person   2    4
```

Pearson‘s Chi-square ($\chi 2$) statistic

```
> chisq.test(table(person, taste))$statistic
X-squared
0.2777778
Warning message:
In chisq.test(table(person, taste)) :
  Chi-squared approximation may be incorrect
```

Association between Two Discrete Variables

Cramer's *V* Statistics

Range of Pearson's χ^2 statistic depends on sample size and size of contingency table. These values depends on the situations.

This is modified in following Cramer's *V* Statistic for a *k* x *l* contingency table.

$$V = \sqrt{\frac{\chi^2}{n[\min(k,l)-1]}} \quad ; 0 \leq V \leq 1$$

Association between Two Discrete Variables

Cramer‘s *V* Statistics

- **Value of *V* close to 0 $\Rightarrow$ low association between the variables.**
- **Value of *V* close to 1 $\Rightarrow$ high association between the variables.**
- **Other values indicates the moderate association between the variables.**

For earlier example, χ^2 = 0.2777778. So

$$V = \sqrt{\frac{0.2777778}{20[\min(3,2) - 1]}} = 0.08333334$$

This again shows a low association.