Descriptive Statistics With R Software

Association of Variables

::

Measures of Association for Discrete and Counting Variables: Contingency Table with R, Chi-Squared Statistic, Cramer‘s V Statistic and Contingency Coefficient

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

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 ($\chi 2$) statistics

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

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

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 ⇒ weak association between the two variables.**

- **Value of χ^2 close to n[min(k, l) - 1] ⇒ 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*

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

Association between Two Discrete Variables

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

Example: A sample of 100 students was chosen and divided into two groups – Weak and strong - in academics. Some of the students are given tuition. We would like to see if tuition was helpful in improving the academic performance of the student or not. The data is compiled in the following contingency table:

		Students	
		Weak Students	Strong Students
Tuition	Tuition given	30	10
	Tuition not given	20	40

Association between Two Discrete Variables

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

Example:

		Students		Total (Rows)
		Weak Students	Strong Students	
Tuition	Tuition given	30	10	40
	Tuition not given	20	40	60
	Total (Columns)	50	50	100

It indicates moderate association.

Association between Two Discrete Variables

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

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 ($\chi 2$) statistics

```
> 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 statistics 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.

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

Association between Two Discrete Variables

Cramer's *V* Statistics

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

For earlier example, χ^2 = 16.66. So

$$V = \sqrt{\frac{16.66}{100[\min(2,2)-1]}} = 0.40$$

This again shows a moderate association.

Association between Two Discrete Variables

R Command

We need a package `lsr`

```
> install.packages("lsr")
```

```
> library(lsr)
```

Contingency table with absolute frequencies

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

Association between Two Discrete Variables

R Command

```
> cramersV(table(person, taste))
[1] 0.1178511
Warning message:
In chisq.test(...) : Chi-squared approximation
may be incorrect
```

R Console

```
> table(person, taste)
              taste
person         Bad Good
  Child          2    2
  Elder person   4    6
  Young person   2    4
> cramersV(table(person, taste))
[1] 0.1178511
Warning message:
In chisq.test(...) : Chi-squared approximation may be incorrect
>
```

Association between Two Discrete Variables

Contingency Coefficient *C*

Corrected version of Pearson's contingency coefficient is

$$C_{corr} = \frac{C}{C_{\max}} \quad ; 0 \le C_{corr} \le 1$$

$$\text{where } C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \;, C_{\max} = \sqrt{\frac{\min(k,l) - 1}{\min(k,l)}}$$

- **Value of *C* close to 0 ⇒ lower association between the two variables.**
- **Value of *C* close to 1 ⇒ higher association between the two variables.**
- **Other values of *C* indicates the moderate association between the two variables.**

Association between Two Discrete Variables

Contingency Coefficient *C*

For earlier example, $\chi^2 = 16.66$. So

$$C = \sqrt{\frac{16.66}{16.66 + 100}} = 0.38$$

$$C_{\max} = \sqrt{\frac{\min(2,2) - 1}{\min(2,2)}} = \sqrt{\frac{1}{2}} = 0.71$$

$$C_{corr} = \frac{0.38}{0.71} = 0.54$$

This again shows moderate association.