Descriptive Statistics With R Software

Association of Variables
::
Measures of Association for Discrete and Counting Variables: Bivariate Frequency and Contingency Tables

Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur

Association between Two Discrete Variables

Example: Suppose we want to know if boys and girls have any inclination to choose between mathematics and biology.

If there is no discrimination, we expect that the total number of boys and girls opting for mathematics and biology should be nearly the same.

Data on such issues are obtained as frequency.

A measure based on frequency data or summarized frequency data is needed to study the association between two such variables.

Association between Two Discrete Variables

Suppose the data is obtained as follows:

Student number	1	2	3	4	5	6	7	8	9	10
Gender M: male F: female	M	F	M	M	F	F	F	M	M	F
Subject Math: Mth Biology: Bio	Bio	Bio	Mth	Mth	Mth	Bio	Bio	Mth	Mth	Mth

Association between Two Discrete Variables

Data can be summarized as follows

	Male Students	Female Students	Total (Rows)
Math	$n_{11} = 4$	$n_{12} = 2$	$n_{1+} = 6$
Biology	$n_{21} = 1$	$n_{22} = 3$	$n_{2+} = 4$
Total (Columns)	$n_{+1} = 5$	$n_{+2} = 5$	$n = 10$

Students preferring maths

Students preferring biology

Male Students preferring maths and biology

Female Students preferring maths and biology

This is a 2 x 2 contingency table.

Association between Two Discrete Variables

n_{ij} : Frequency in $(i, j)^{th}$ cell

$n_{1+} = n_{11} + n_{12}$: Row total (1st row of data)

$n_{2+} = n_{21} + n_{22}$: Row total (2nd row of data)

$n_{+1} = n_{11} + n_{21}$: Column total (1st column of data)

$n_{+2} = n_{12} + n_{22}$: Column total (2nd column of data)

$n = n_{11} + n_{12} + n_{21} + n_{22} = n_{1+} + n_{2+} = n_{+1} + n_{+2}$ = Total frequency

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

When the data on two variables are summarized in a contingency table, there are several characteristics of the data can be studied.

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

$\boldsymbol{n_{ij}}$ **: Absolute frequencies**

: Represents <u>joint frequency distribution</u> of *X* and *Y*

<u>Joint frequency distribution</u> tells how the values of both the variables behave jointly.

Association between Two Discrete Variables

n_{i+} : Represents marginal frequency distribution of X

n_{+j} : Represents marginal frequency distribution of Y

Marginal frequency distribution tells how the values of one variable behave in the joint distribution.

If relative frequency is used instead of absolute frequency, then the similar information is provided by the

- joint relative frequency distribution,
- marginal relative frequency distribution, and
- conditional relative frequency distribution.

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 <u>frequency distribution</u> of X given $Y = y_j$

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

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

Association between Two Discrete Variables

$$f_{i+} = \sum_{i=1}^{l} f_{ij}$$: **Marginal relative frequency distribution of *X***

$$f_{+j} = \sum_{i=1}^{k} f_{ij}$$: **Marginal relative frequency distribution of *Y***

$$f_{i|j}(X \mid Y)$$: **Conditional relative frequency distribution of *X* given $Y = y_j$**

$$f_{j|i}(Y \mid X)$$: **Conditional relative frequency distribution of *Y* given $X = x_i$**

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.**