Introduction to Statistics and Data Science using *eStat*

Chapter 11 Testing Hypothesis for Categorical Data

11.2.1 Independence Test

Jung Jin Lee
Professor of Soongsil University, Korea
Visiting Professor of ADA University, Azerbaijan

11.1 Goodness of Fit Test

11.1.1 Goodness of Fit Test for Categorical Data
11.1.2 Goodness of Fit Test for Continuous Data

11.2 Testing Hypothesis for Contingency Table

11.2.1 Independence Test
11.2.2 Homogeneity Test

11.2 Testing Hypothesis for Contingency Table

11.2.1 Independence Test

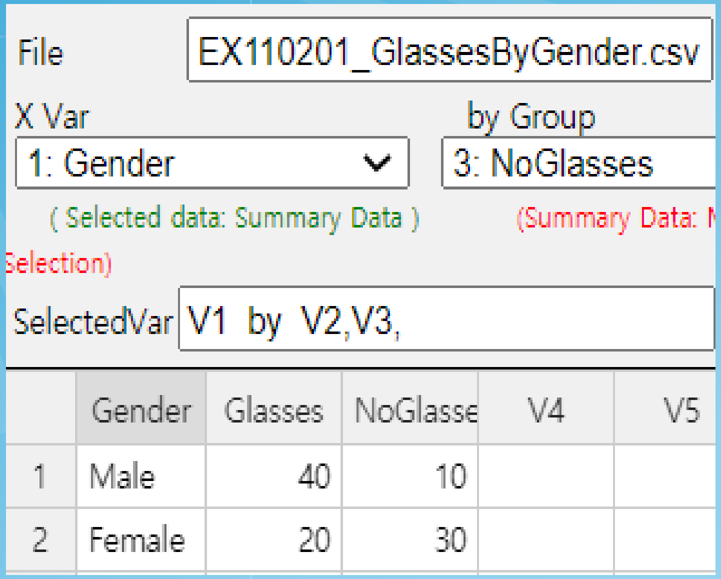
[Example 11.2.1] In order to investigate whether college students who are wearing glasses are independent by gender, a sample of 100 students was collected and its contingency table was prepared as follows:

1) Using 『eStat』, draw a line graph of use of eyeglasses by gender.
2) Test the hypothesis at 5% of the significance level to see if gender and wearing of glasses are independent or related to each other.
3) Check the result of the independence test using 『eStatU』.

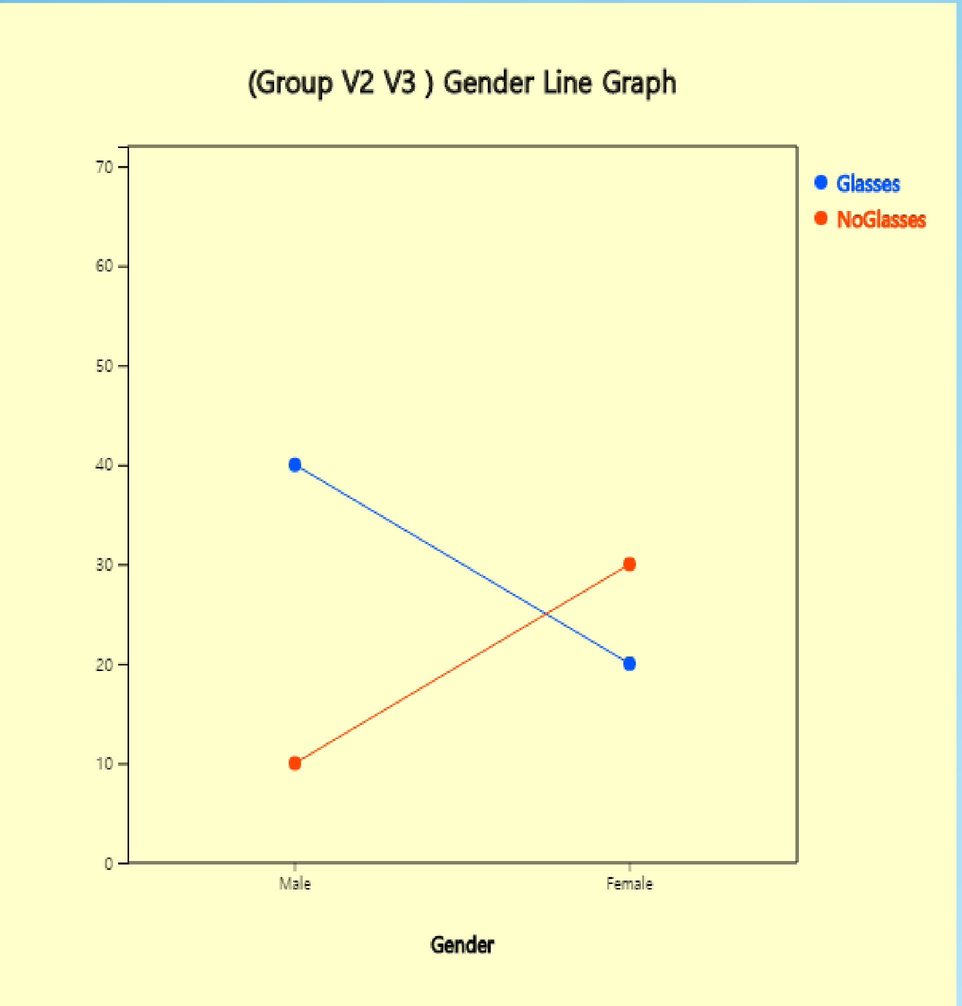
	Wear Glasses	No Glasses	Total
Men	40	10	50
Women	20	30	50
Total	60	40	100

11.2 Testing Hypothesis for Contingency Table

<Answer of Example 11.2.1>



	Gender	Glasses	NoGlasse	V4	V5
1	Male	40	10		
2	Female	20	30		



Independent contingency table	Wear Glasses	No Glasses	Total
Men	30	20	50
Women	30	20	50
Total	60	40	100

11.2 Testing Hypothesis for Contingency Table

<Answer of Example 11.2.1>

- **Hypothesis**

  H_0 : Row and column variables are independent of each other

  H_1 : Row and column variables are related

- **Test Statistic**

$$\chi^2_{obs} = \frac{(40-30)^2}{30} + \frac{(10-20)^2}{20} + \frac{(20-30)^2}{30} + \frac{(30-20)^2}{20} = 16.67$$

- **Decision Rule**

  'If $\chi^2_{obs} > \chi^2_{(r-1)(c-1);\,\alpha}$, reject H_0'

  Since $\chi^2_{(2-1)(2-1);\ 0.05}$ = 3.841, H_0 is rejected.

11.2 Testing Hypothesis for Contingency Table

<Answer of Example 11.2.1>

Testing Independence Menu

[Hypothesis] H_0 : *Row and column variables are independent*

H_1 : *Row and column variables are not independent*

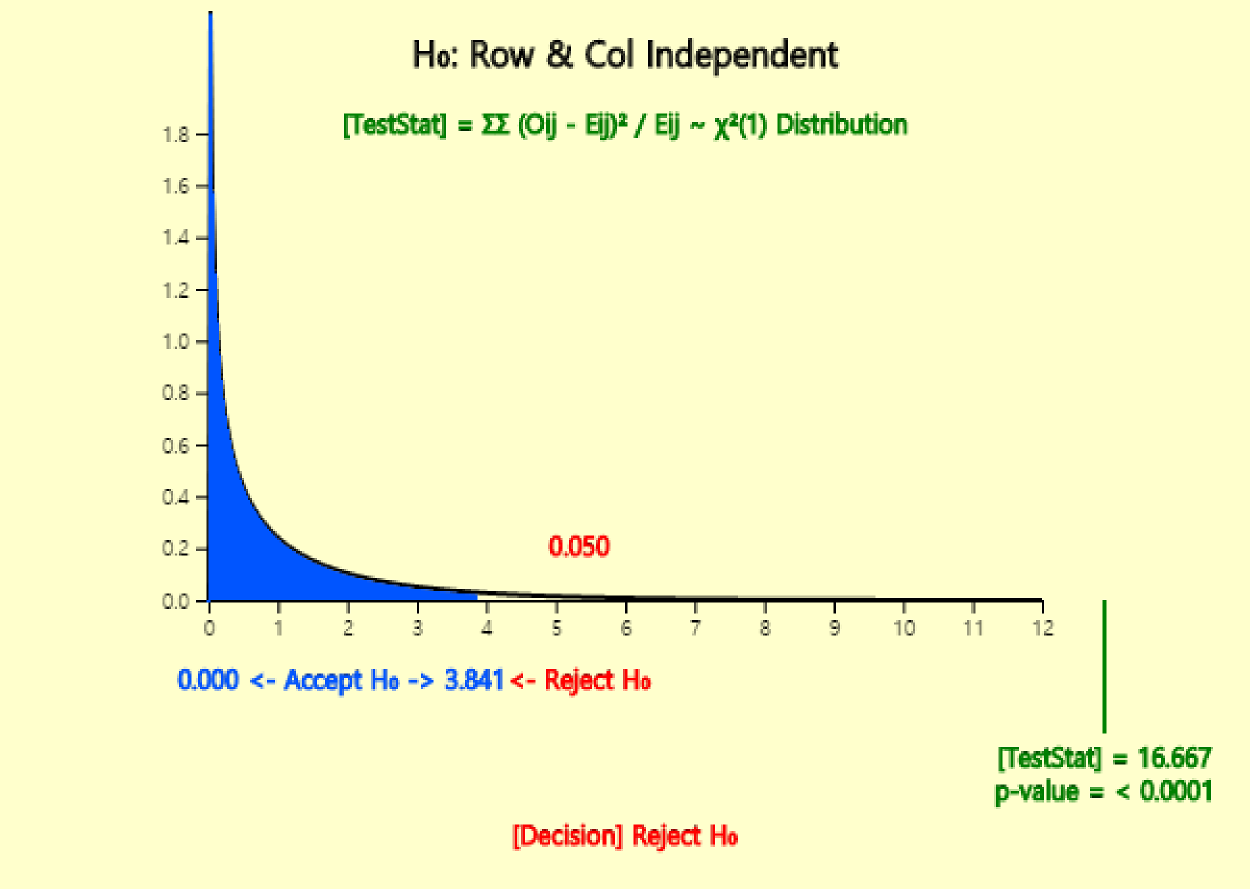
[Test Type] χ^2 *test*

Significance Level α = (●) 5% ( ) 1%

[Sample Data] *(Enter observation from upper left cell)*

	Column 1	Column 2	Column 3	Column 4	Column 5
Row 1	40	10			
Row 2	20	30			
Row 3					
Row 4					

Execute



11.2 Testing Hypothesis for Contingency Table

Observed frequency		Column Variable B B_1	B_2	$\cdots$	B_c	Total
Row Variable A	A_1	O_{11}	O_{12}	$\cdots$	O_{1c}	$T_{1\cdot}$
	A_2	O_{21}	O_{22}	$\cdots$	O_{2c}	$T_{2\cdot}$
	$\vdots$		$\vdots$			$\vdots$
	A_r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	$T_{r\cdot}$
Total		$T_{\cdot 1}$	$T_{\cdot 2}$	$\cdots$	$T_{\cdot c}$	n

- **Test Statistics:**

$$\chi^2_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

Observed frequency		Column Variable B B_1	B_2	$\cdots$	B_c
Row Variable A	A_1	$E_{11} = T_{1\cdot}\frac{T_{\cdot 1}}{n}$	$E_{12} = T_{1\cdot}\frac{T_{\cdot 2}}{n}$	$\cdots$	$E_{1c} = T_{1\cdot}\frac{T_{\cdot c}}{n}$
	A_2	$E_{21} = T_{2\cdot}\frac{T_{\cdot 1}}{n}$	$E_{22} = T_{2\cdot}\frac{T_{\cdot 2}}{n}$	$\cdots$	$E_{2c} = T_{2\cdot}\frac{T_{\cdot c}}{n}$
	$\vdots$		$\vdots$		
	A_r	$E_{r1} = T_{r\cdot}\frac{T_{\cdot 1}}{n}$	$E_{r2} = T_{r\cdot}\frac{T_{\cdot 2}}{n}$	$\cdots$	$E_{rc} = T_{r\cdot}\frac{T_{\cdot c}}{n}$

11.2 Testing Hypothesis for Contingency Table

[Independence Test]

- **Hypothesis:**

  H_0 : Row and column variables are independent. (i.e., $p_{ij} = p_{i\cdot}\, p_{\cdot j}$)

  H_1 : Row and column variables are not independent

- **Decision Rule:**

  'If $\chi^2_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} > \chi^2_{(r-1)(c-1);\,\alpha}$, reject H_0'

  where r and c are the number of attributes of row and column variable

❖ In order to use the chi-square distribution for the independence test, all expected frequencies are at least 5 or more.

❖ If an expected frequency of a cell is smaller than 5, the cell is combined with adjacent cell for analysis.

11.2 Testing Hypothesis for Contingency Table

[Example 11.2.2] A market research institute surveyed 500 people on how three beverage products (A, B and C) are preferred by region and obtained the following contingency table.

1) Draw a line graph of beverage preference by region using 『eStat』.
2) Test whether the beverage preference by the region is independent of each other at the significance level of 5%.
3) Check the result of the independence test using 『eStatU』.

Region	Beverage A	B	C	Total
New York	52	64	24	140
Los Angels	60	59	52	171
Atlanta	50	65	74	189
Total	162	188	150	500

11.2 Testing Hypothesis for Contingency Table

<Answer of Example 11.2.2>

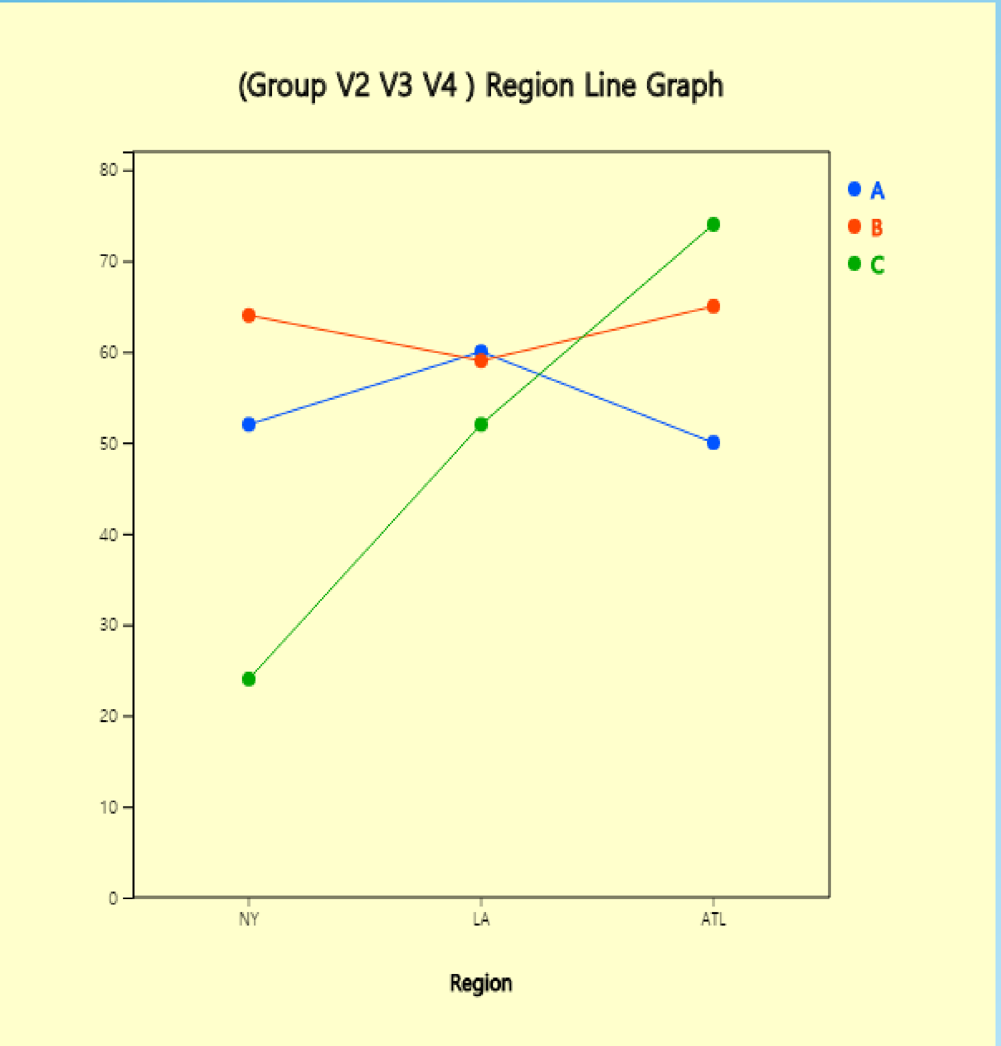
File: EX110202_BeverageByRegion.csv   EditVar

X Var: 1: Region   by Group: 4: C

(Selected data: Summary Data)   (Summary Data: Multiple Selection)

SelectedVar: V1 by V2,V3,V4,   Cancel

	Region	A	B	C	V5	V
1	NY	52	64	24		
2	LA	60	59	52		
3	ATL	50	65	74		



11.2 Testing Hypothesis for Contingency Table

<Answer of Example 11.2.2>

- Hypothesis

H_0 : Region and beverage preference are independent of each other

H_1 : Region and beverage preference are not independent

- Expected frequency

$$\left(\frac{T_{\cdot 1}}{n}, \frac{T_{\cdot 2}}{n}, \frac{T_{\cdot 3}}{n}\right) = \left(\frac{162}{500}, \frac{88}{500}, \frac{50}{500}\right)$$

$$E_{11} = T_{1\cdot}\frac{162}{500} \qquad E_{12} = T_{1\cdot}\frac{88}{500} \qquad E_{13} = T_{1\cdot}\frac{50}{500}$$

$$E_{21} = T_{2\cdot}\frac{162}{500} \qquad E_{22} = T_{2\cdot}\frac{88}{500} \qquad E_{23} = T_{2\cdot}\frac{50}{500}$$

$$E_{31} = T_{3\cdot}\frac{162}{500} \qquad E_{32} = T_{3\cdot}\frac{88}{500} \qquad E_{33} = T_{3\cdot}\frac{50}{500}$$

11.2 Testing Hypothesis for Contingency Table

<Answer of Example 11.2.2>

- **Hypothesis**

  H_0 : Region and beverage preference are independent of each other

  H_1 : Region and beverage preference are not independent

- **Test Statistic**

$$\chi^2_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(52-45.36)^2}{45.36} + \frac{(60-55.40)^2}{55.40} + \cdots + \frac{(74-56.70)^2}{56.70} = 18.825$$

- **Decision Rule**

  'If $\chi^2_{obs} > \chi^2_{(r-1)(c-1);\,\alpha}$, reject H_0'

  Since $\chi^2_{(3-1)(3-1);\ 0.05} = 9.488$, H_0 is rejected.

11.2 Testing Hypothesis for Contingency Table

<Answer of Example 11.2.2>

Testing Independence

Menu

[Hypothesis] H_0 : *Row and column variables are independent*

H_1 : *Row and column variables are not independent*

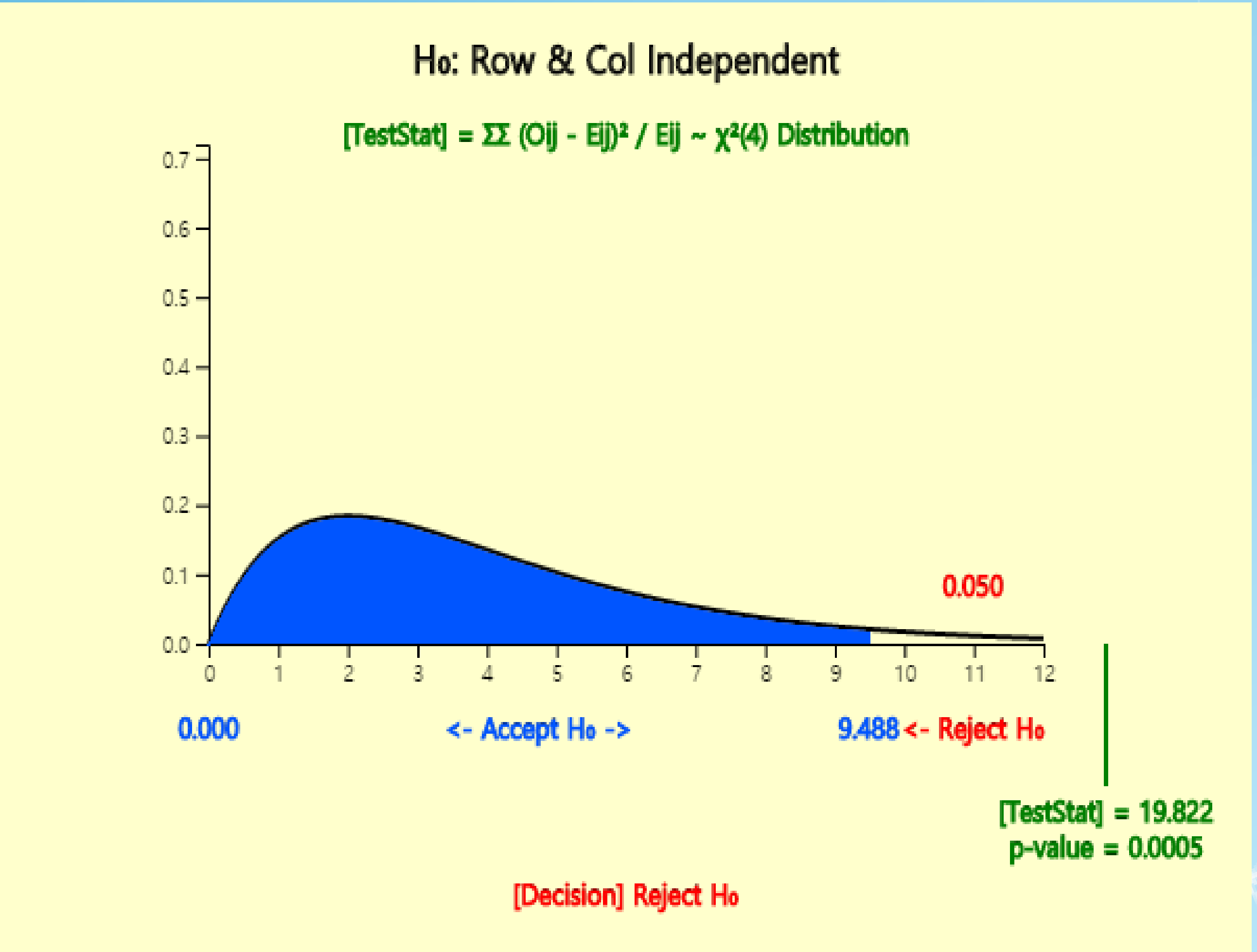
[Test Type] χ^2 *test*

Significance Level α = ◉ 5% ○ 1%

[Sample Data] *(Enter observation from upper left cell)*

	Column 1	Column 2	Column 3	Column 4	Column 5
Row 1	52	64	24		
Row 2	60	59	52		
Row 3	50	65	74		
Row 4					

Execute



11.2 Testing Hypothesis for Contingency Table

<Answer of Example 11.2.2>

File: EX040201_Categorical_MaritalBy | EditVar

Analysis Var: 2: Marital | by Group: 1: Gender

(Selected data: Raw Data) (Summary Data: Multiple Selection)

SelectedVar: V2 by V1, | Cancel

	Gender	Marital	V3	V4	V5	V
1	1	1				
2	2	2				
3	1	1				
4	2	1				
5	1	2				
6	1	1				
7	1	1				
8	2	2				
9	1	3				
10	2	1				

Cross Table	Col Variable	(Gender)			
Row Variable (Marital)	1	2	Total		
Group 1	4 66.7%	2 33.3%	6 100%		
Group 2	1 33.3%	2 66.7%	3 100%		
Group 3	1 100.0%	0 0.0%	1 100%		
Total	6 60.0%	4 40.0%	10 100%		
	Missing Observations	0			
Independence Test					
Sum of χ^2 value	1.667	deg of freedom	2	p-value	0.4346

Thank you