How is the Chi-square test calculated?

Authored by

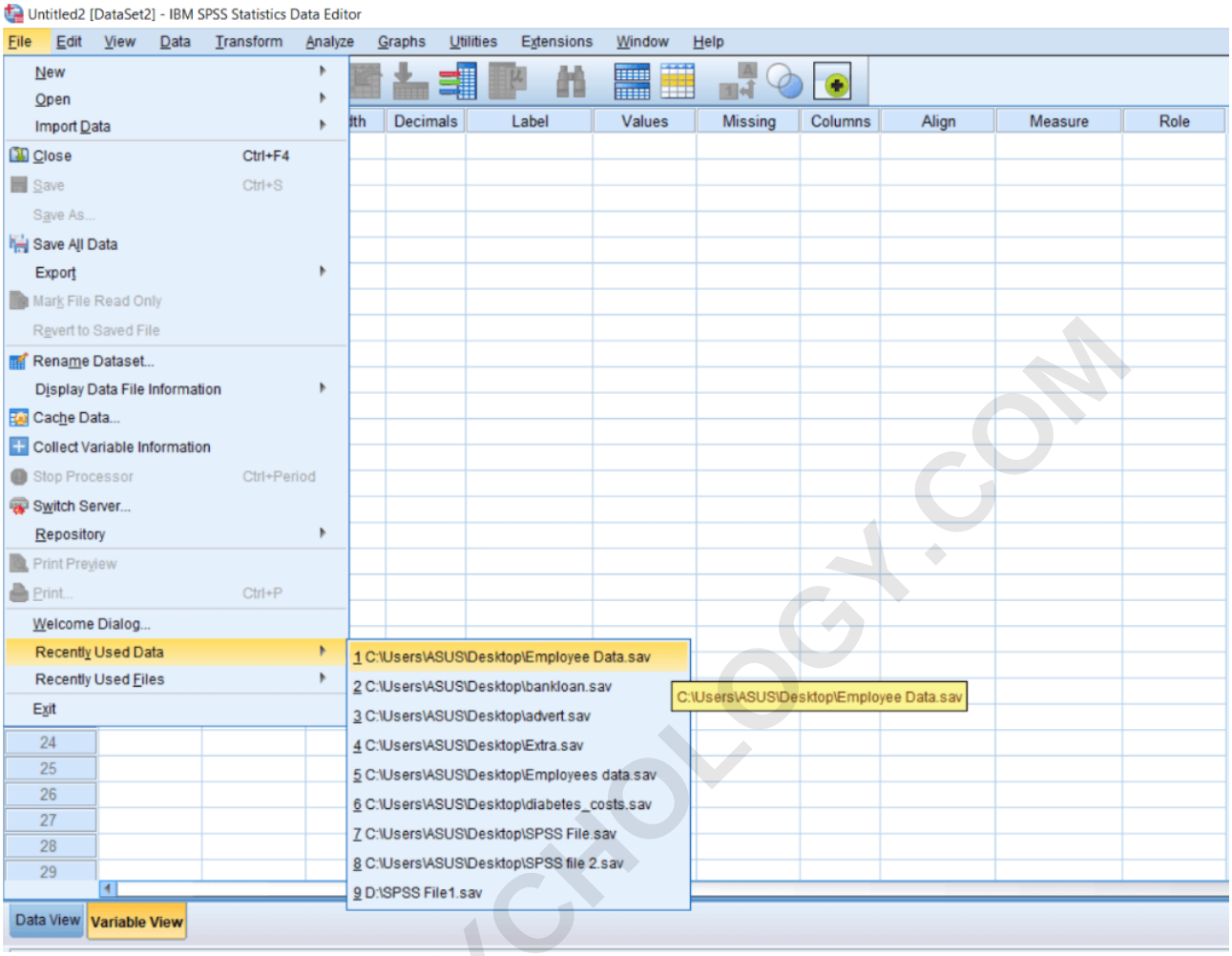
stats writer

June 24, 2024

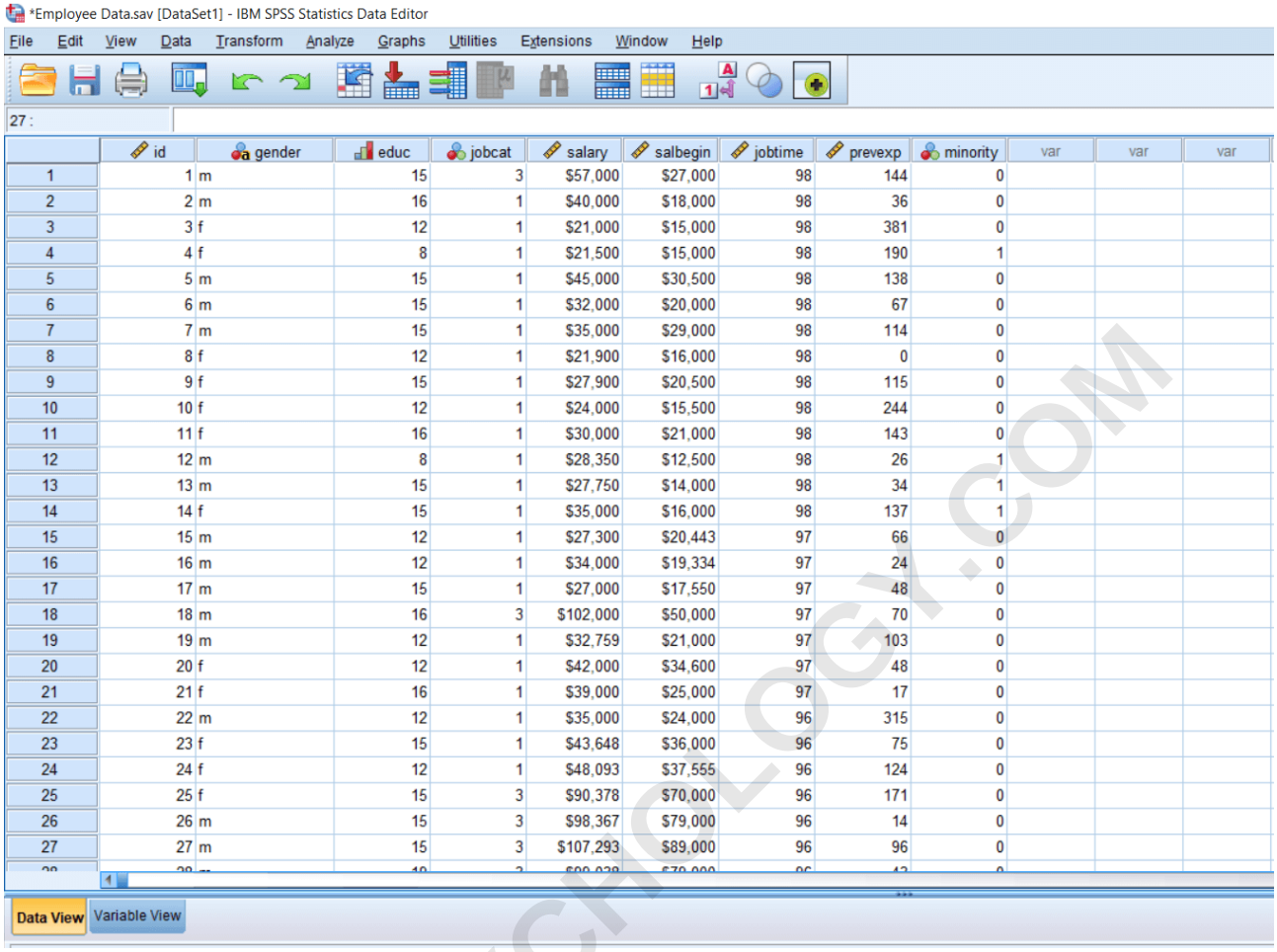
The Chi-square test is a statistical method used to determine the relationship between two categorical variables. It is calculated by comparing the observed frequencies of a particular variable to the expected frequencies, assuming there is no relationship between the variables. This is done through a formula that calculates the sum of the squared differences between the observed and expected frequencies, which is then compared to a critical value from a Chi-square table. If the calculated value is greater than the critical value, it indicates that there is a significant relationship between the variables. The Chi-square test is commonly used in various fields, such as social sciences, biology, and market research, to analyze data and draw conclusions about the relationship between variables.

Calculation of Chi-square test

In this section, we will learn how to calculate the chi-square test in SPSS. To calculate the chi-square test, we will open our Data set by going to the File menu, then go to Recently used Data as follows:



Now we will click on the above Employee Data option and see our Employee Data set as follows:


*Employee Data.sav [DataSet1] - IBM SPSS Statistics Data Editor

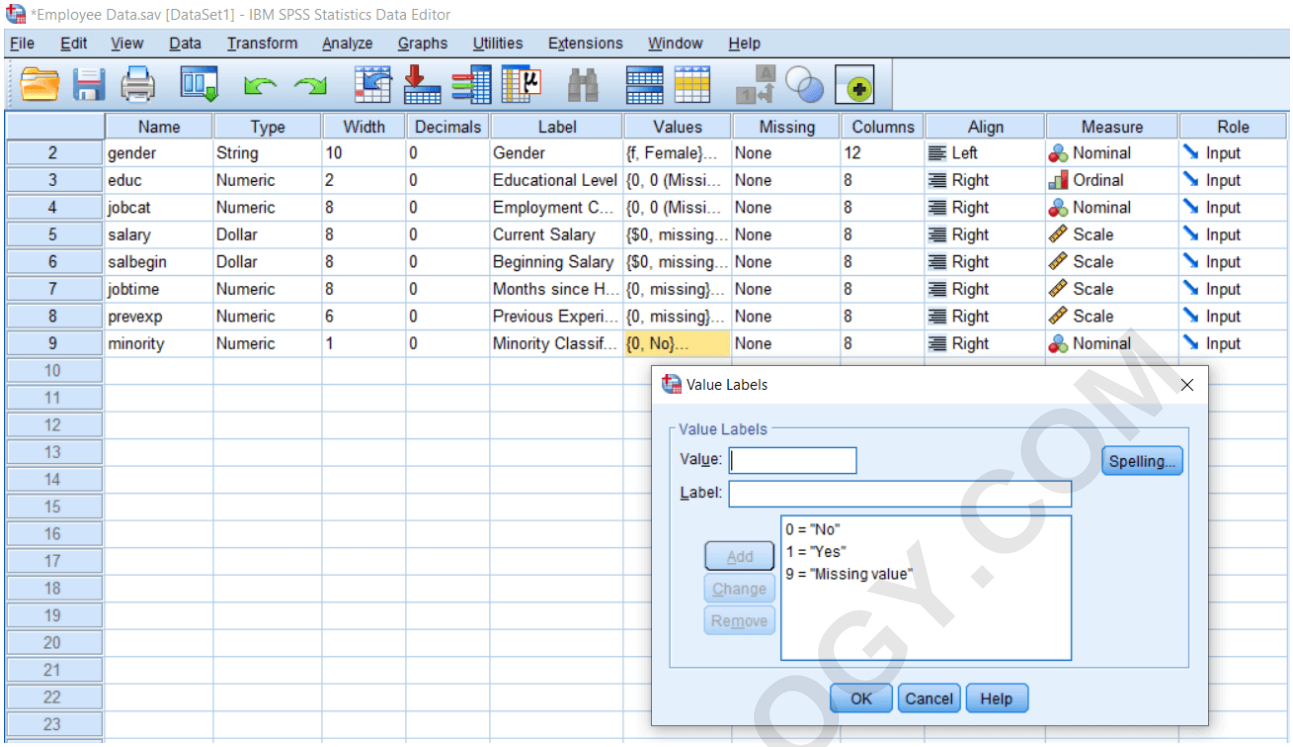
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

27 :

	id	gender	educ	jobcat	salary	salbegin	jobtime	prevexp	minority	var	var	var
1	1	m	15	3	$57,000	$27,000	98	144	0			
2	2	m	16	1	$40,000	$18,000	98	36	0			
3	3	f	12	1	$21,000	$15,000	98	381	0			
4	4	f	8	1	$21,500	$15,000	98	190	1			
5	5	m	15	1	$45,000	$30,500	98	138	0			
6	6	m	15	1	$32,000	$20,000	98	67	0			
7	7	m	15	1	$35,000	$29,000	98	114	0			
8	8	f	12	1	$21,900	$16,000	98	0	0			
9	9	f	15	1	$27,900	$20,500	98	115	0			
10	10	f	12	1	$24,000	$15,500	98	244	0			
11	11	f	16	1	$30,000	$21,000	98	143	0			
12	12	m	8	1	$28,350	$12,500	98	26	1			
13	13	m	15	1	$27,750	$14,000	98	34	1			
14	14	f	15	1	$35,000	$16,000	98	137	1			
15	15	m	12	1	$27,300	$20,443	97	66	0			
16	16	m	12	1	$34,000	$19,334	97	24	0			
17	17	m	15	1	$27,000	$17,550	97	48	0			
18	18	m	16	3	$102,000	$50,000	97	70	0			
19	19	m	12	1	$32,759	$21,000	97	103	0			
20	20	f	12	1	$42,000	$34,600	97	48	0			
21	21	f	16	1	$39,000	$25,000	97	17	0			
22	22	m	12	1	$35,000	$24,000	96	315	0			
23	23	f	15	1	$43,648	$36,000	96	75	0			
24	24	f	12	1	$48,093	$37,555	96	124	0			
25	25	f	15	3	$90,378	$70,000	96	171	0			
26	26	m	15	3	$98,367	$79,000	96	14	0			
27	27	m	15	3	$107,293	$89,000	96	96	0			

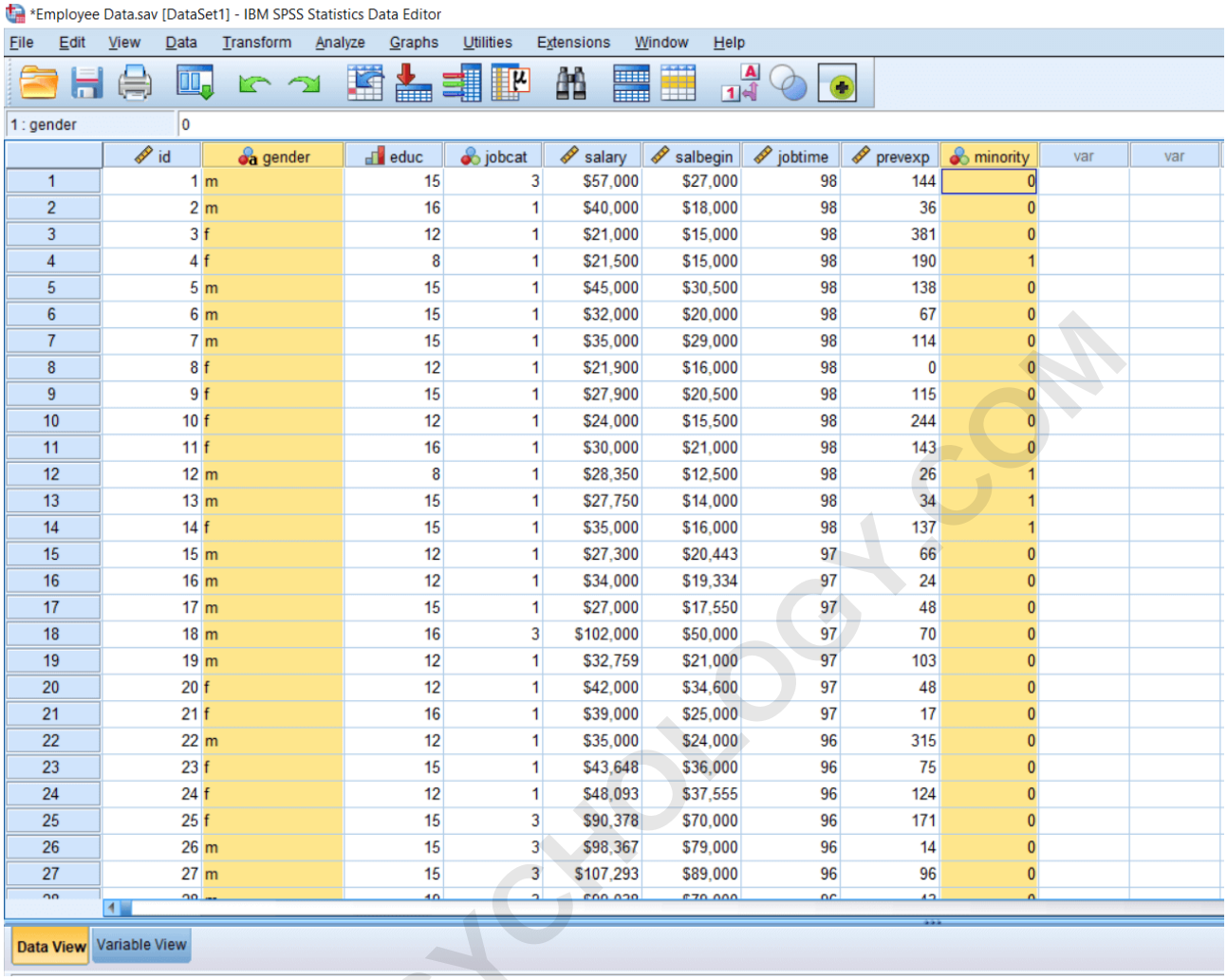
Data View Variable View

Look at this data set. In this data set, we have the recode of various employees with their ID, gender of the employee, education, job category, current salary, beginning salary, job time, previous experience and minority classification. In gender, m stands for males and f stands for females. In minority classification, 0 means No, which shows the person is not from a minority background, and 1 means Yes, which shows the person belongs to the minority background. 9 is for the missing value.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
2	gender	String	10	0	Gender	{f, Female}...	None	12	Left	Nominal	Input
3	educ	Numeric	2	0	Educational Level	{0, 0 (Missi...	None	8	Right	Ordinal	Input
4	jobcat	Numeric	8	0	Employment C...	{0, 0 (Missi...	None	8	Right	Nominal	Input
5	salary	Dollar	8	0	Current Salary	{$0, missing...	None	8	Right	Scale	Input
6	salbegin	Dollar	8	0	Beginning Salary	{$0, missing...	None	8	Right	Scale	Input
7	jobtime	Numeric	8	0	Months since H...	{0, missing}...	None	8	Right	Scale	Input
8	prevexp	Numeric	6	0	Previous Experi...	{0, missing}...	None	8	Right	Scale	Input
9	minority	Numeric	1	0	Minority Classif...	{0, No}...	None	8	Right	Nominal	Input
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											

Using this data set, we want to test if there is a significantly higher number of males as compared to females from minority communities. So the researcher is currently interested in the gender variable and its relationship with the minority classification.



*Employee Data.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

1 : gender 0

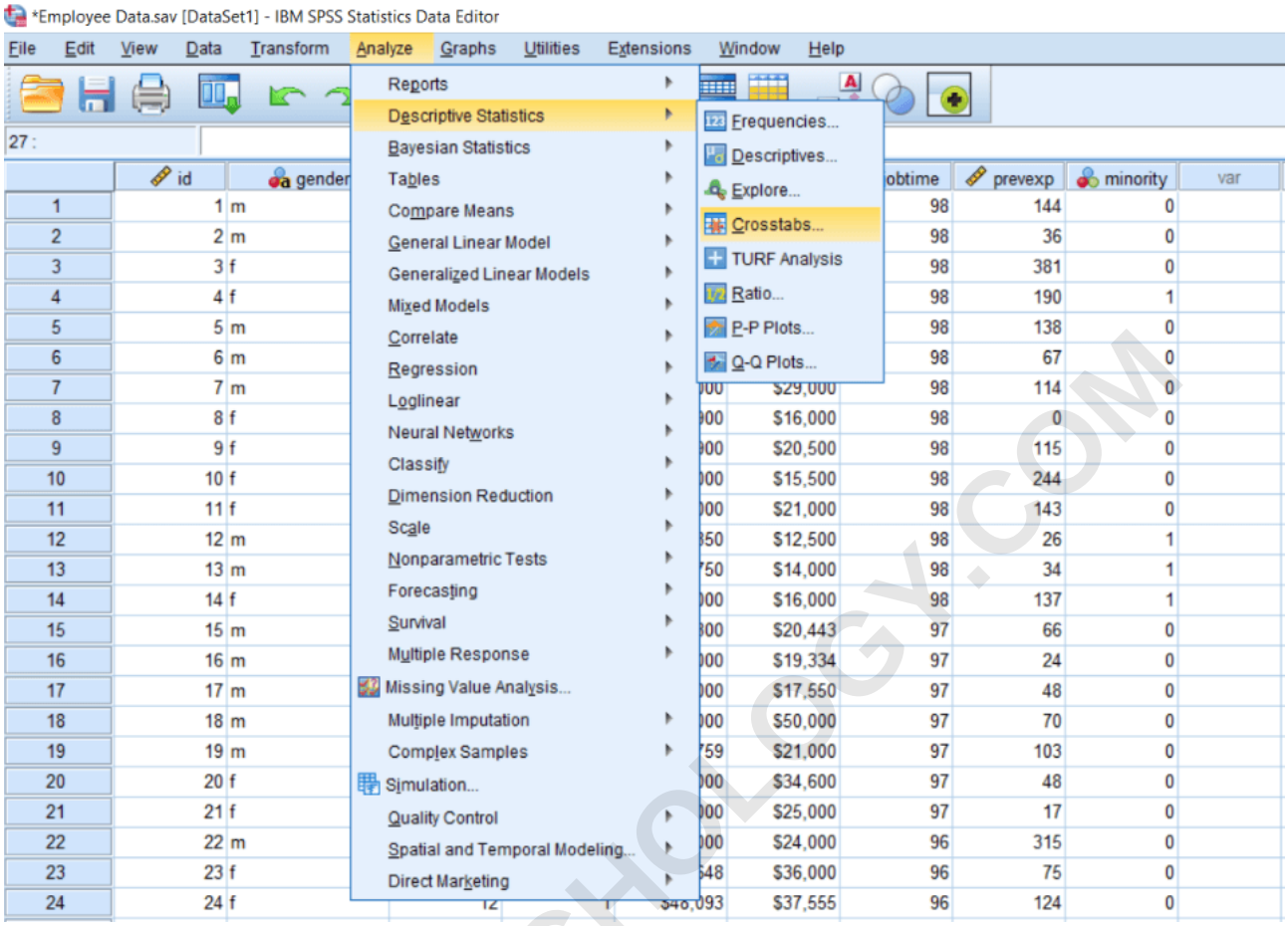
	id	gender	educ	jobcat	salary	salbegin	jobtime	prevexp	minority	var	var
1	1	m	15	3	$57,000	$27,000	98	144	0		
2	2	m	16	1	$40,000	$18,000	98	36	0		
3	3	f	12	1	$21,000	$15,000	98	381	0		
4	4	f	8	1	$21,500	$15,000	98	190	1		
5	5	m	15	1	$45,000	$30,500	98	138	0		
6	6	m	15	1	$32,000	$20,000	98	67	0		
7	7	m	15	1	$35,000	$29,000	98	114	0		
8	8	f	12	1	$21,900	$16,000	98	0	0		
9	9	f	15	1	$27,900	$20,500	98	115	0		
10	10	f	12	1	$24,000	$15,500	98	244	0		
11	11	f	16	1	$30,000	$21,000	98	143	0		
12	12	m	8	1	$28,350	$12,500	98	26	1		
13	13	m	15	1	$27,750	$14,000	98	34	1		
14	14	f	15	1	$35,000	$16,000	98	137	1		
15	15	m	12	1	$27,300	$20,443	97	66	0		
16	16	m	12	1	$34,000	$19,334	97	24	0		
17	17	m	15	1	$27,000	$17,550	97	48	0		
18	18	m	16	3	$102,000	$50,000	97	70	0		
19	19	m	12	1	$32,759	$21,000	97	103	0		
20	20	f	12	1	$42,000	$34,600	97	48	0		
21	21	f	16	1	$39,000	$25,000	97	17	0		
22	22	m	12	1	$35,000	$24,000	96	315	0		
23	23	f	15	1	$43,648	$36,000	96	75	0		
24	24	f	12	1	$48,093	$37,555	96	124	0		
25	25	f	15	3	$90,378	$70,000	96	171	0		
26	26	m	15	3	$98,367	$79,000	96	14	0		
27	27	m	15	3	$107,293	$89,000	96	96	0		

Data View Variable View

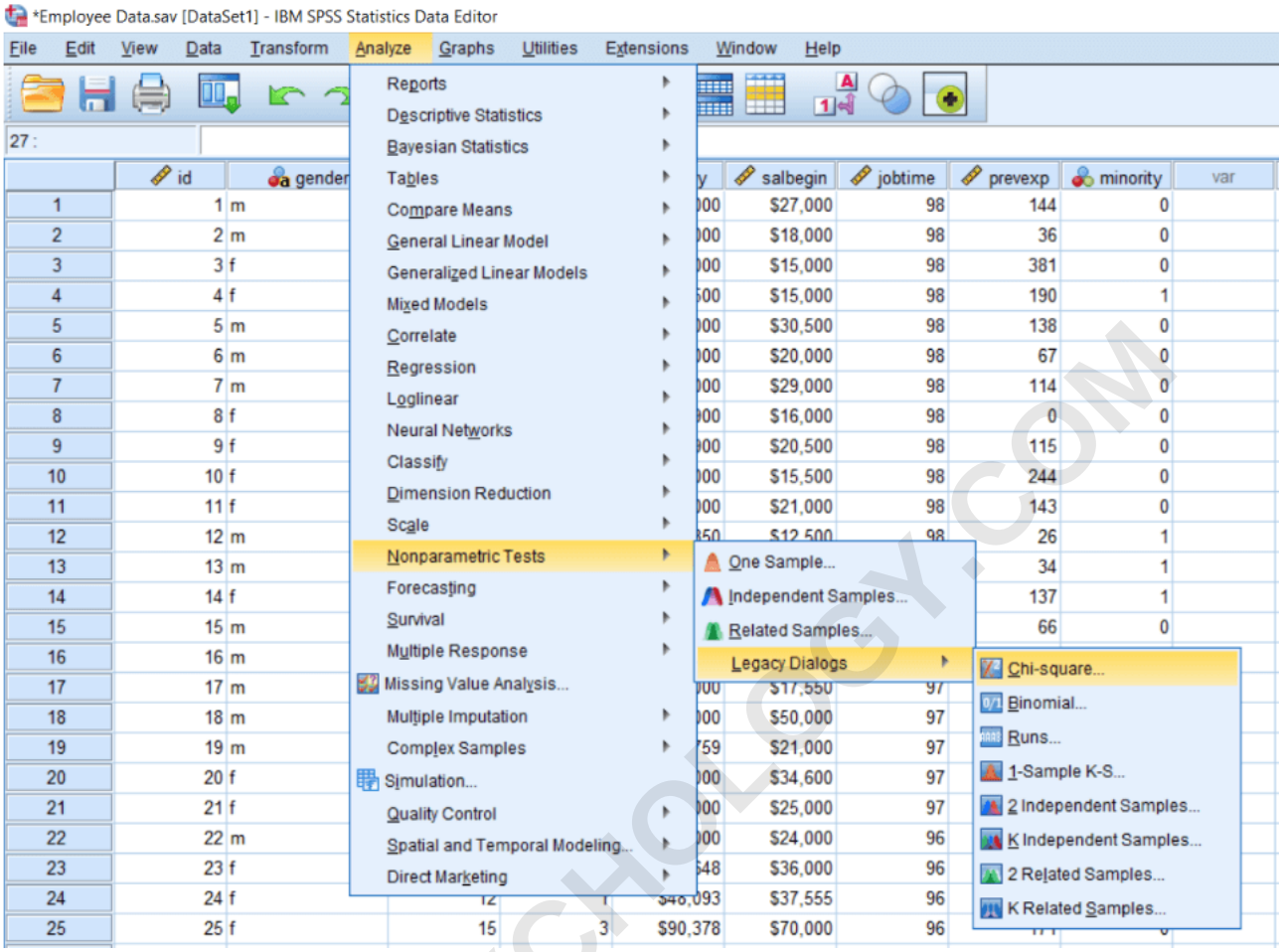
The question could be whether there is a significant difference between the number of males and females belonging to the minority classification. So, in this case, we will use the chi-square test. We will use the chi-square test because both variables, gender and minority classification are nominal variables or frequency variables. They are frequency variable, so any other parameter stats like t-test cannot be applied. The question is, should we consider one variable as an

independent variable and another variable as a dependent variable while calculating chi-square. The answer is yes. Yes, we can select a variable as an independent variable. We can see whether it impacts the dependent variable significantly by using the Crosstab option.

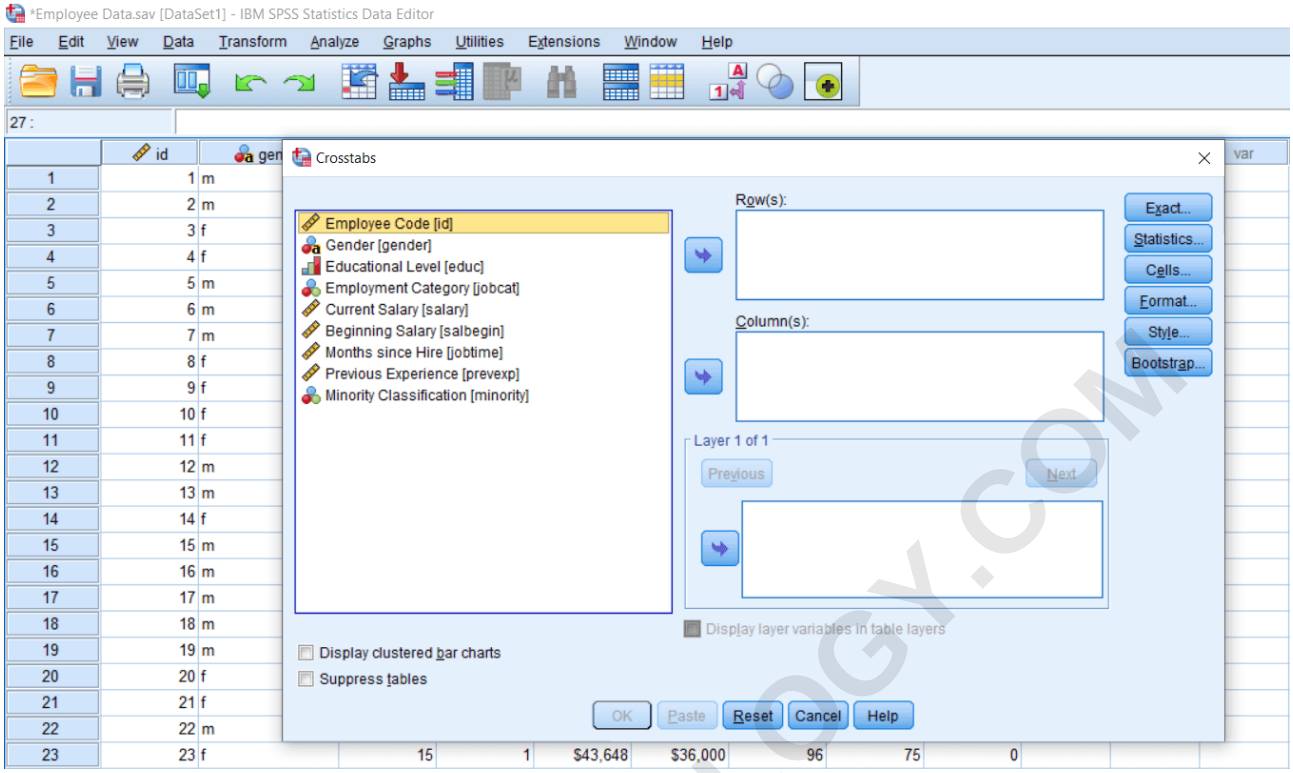
But all in all, while calculating the chi-square test, IBM SPSS does not make a distinction between the independent and dependent variables. It calculates an interaction between the two frequency variables that we choose and derive inference based on our interpretation. So before moving into that, let us see how we can calculate the chi-square test. So we can calculate the chi-square test in many ways. For example, we can calculate the chi-square test using Descriptive statistics and the Crosstabs option that we are going to use currently.



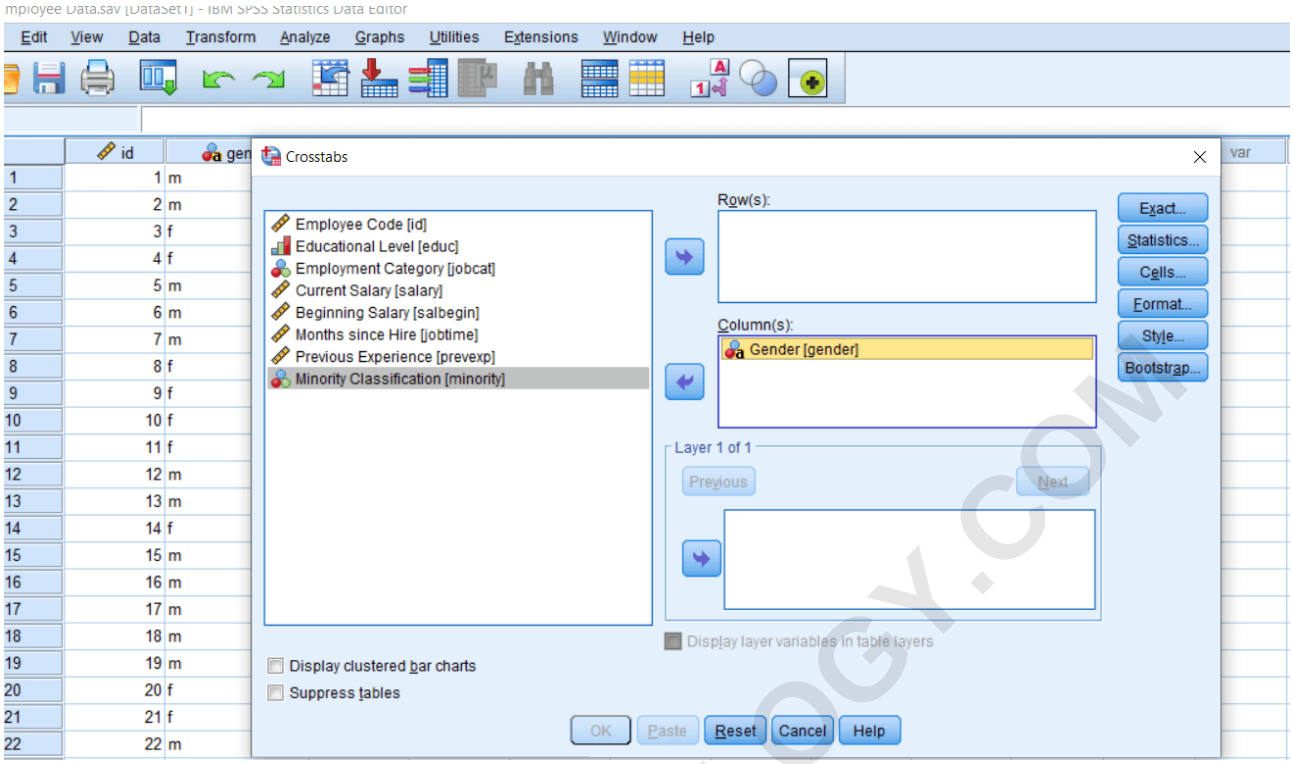
Alternatively, since we know, chi-square is a non-parametric test. So we can also calculate the chi-square test by going to Non-parametric tests, and then we can go to Legacy Dialogs, and then we can calculate the Chi-square test.



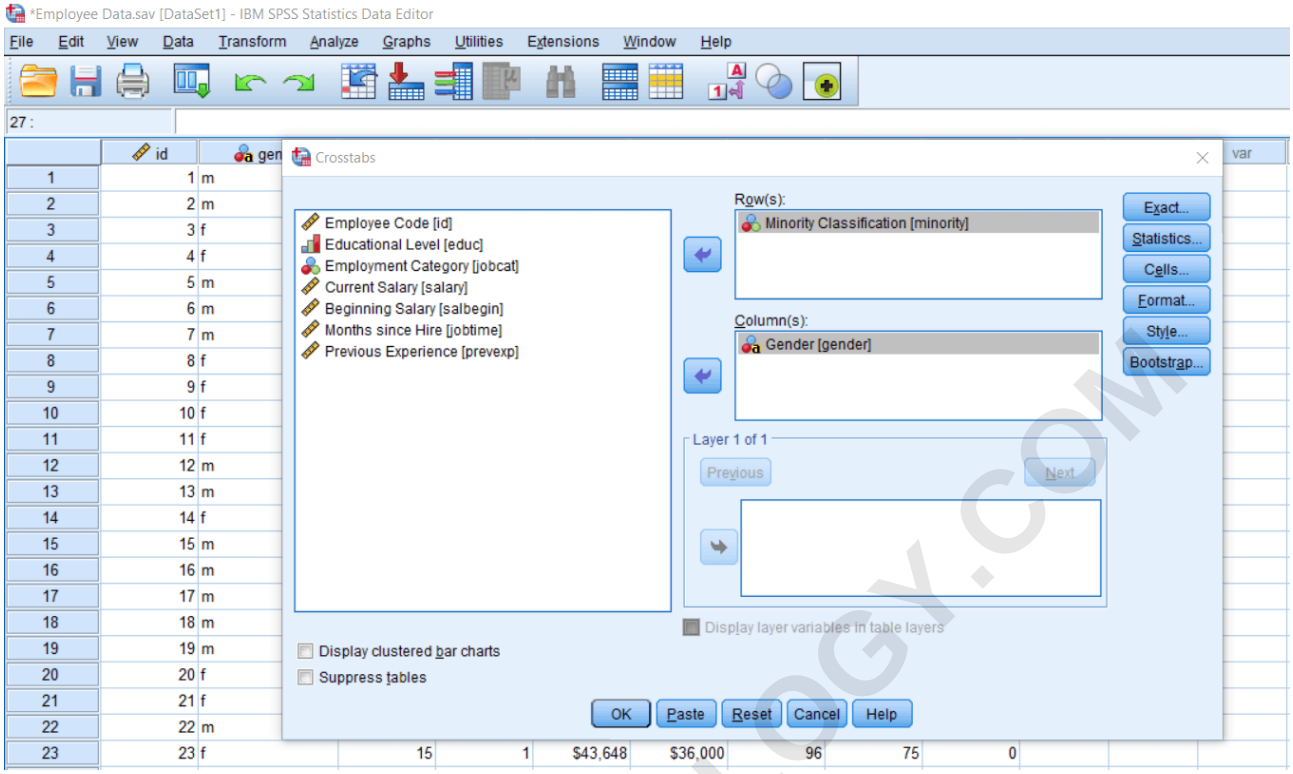
So currently, we are going to demonstrate a chi-square test using the Crosstab option. To calculate chi-square using cross tab options, we will go to Descriptive statistics and then go to crosstabs. After clicking on the crosstab option, we can see all our variables populated like this:



Now we want to see whether males and females are significantly different in number in reference to their minority classification. So we will take gender as an independent variable for all practical purposes in this study. So if we want to take a variable as an independent variable, we can take it in a column dialog box. So we will take gender in our column dialog box.



Our dependent variable is minority classification. So take minority classification in our row dialog box. But practically it's going to make no difference in the result either we choose gender in a column dialog box or row dialog box or visa versa. It's only going to tell us the interaction between frequencies.



Note: Chi-square is just a test of association, and it does not matter whether we are taking our independent variables in row or column. But as a convention, if we hypothesize an independent variable, then take it in the column. The dependent variable will be taken in a row.