What is the Chi Square test and how is it used in statistical analysis?

Authored by

stats writer

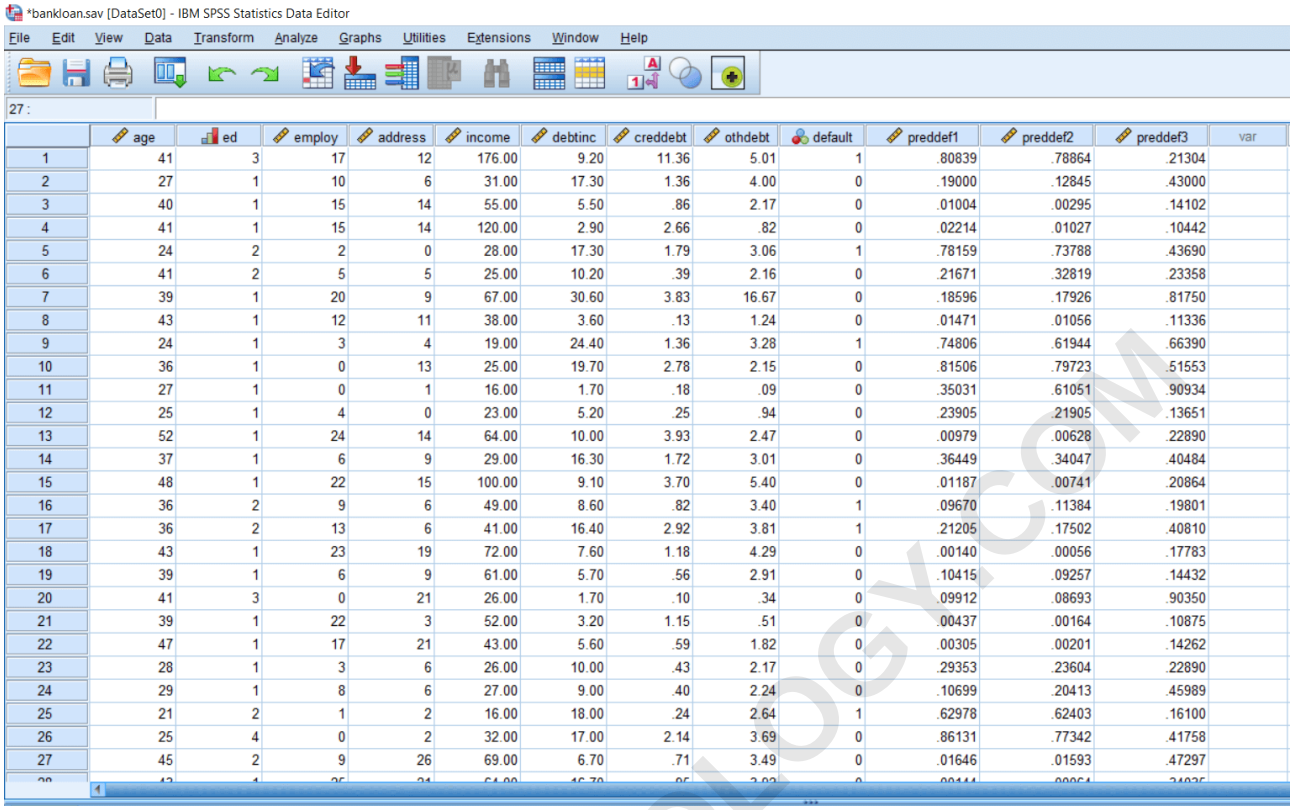
June 24, 2024

The Chi Square test is a statistical tool used to determine the association between two categorical variables. It compares the observed frequencies of the variables with the expected frequencies if there was no association between them. This test is commonly used in statistical analysis to determine whether there is a significant relationship between two variables, or if the differences in their frequencies are due to chance. It is particularly useful in analyzing survey data, and can also be applied to test the goodness of fit of a theoretical model to observed data. The Chi Square test provides a numerical value known as the Chi Square statistic, which is then compared to a critical value to determine the level of significance. This test is widely used in various fields such as social sciences, market research, and quality control, to name a few. Overall, the Chi Square test is an essential tool in statistical analysis for understanding the relationship between categorical variables and making informed decisions based on the results.

Chi Square test

In this section, we will learn how to interpret and use the Chi-square test in SPSS. Chi-square test is also known as the Pearson chi-square test because it was given by one of the four most genius of statistics Karl Pearson.The Chi-square test is a non-parametric test for testing the significant differences between group frequencies. Often when we work with data, we get the data not always in continuous data format. Data might come in the format of frequencies. For example, how many males, how many females are working in an office. Do they significantly differ from each other, which means there are significantly more males working in a job or office as compared to females.For example, suppose we are a government servant, and we want to convey a message that in sectors that are

considered hazards like mining and all, there might be more male workers employed than female workers. So, in that case, we might compute a Pearson chi-square test and convey the result that there are significantly more male workers in mining as compared to female workers because the job is of a hazards nature.Suppose we are working in an IT sector or in some other sector where an equal number of males and females are supposed to be employed. In that case, we might be interested in proving that there is no significant difference between the number of males and females working in the office. So, in that case, we can use our data and calculate the Pearson chi-square test to covey our result. Since the number of males and females comes in the format of frequency so we cannot apply any parametric tests like ANOVA or t-test. So the chi-square test can be used only when we have our data in the format of frequencies.Frequencies can be as simple as the example I gave you as the number of males and the number of females. But if we take multiple groups, in that case, the frequencies might become complex. But chi-square makes our job easy by simply calculating the result. Let us look at this data:



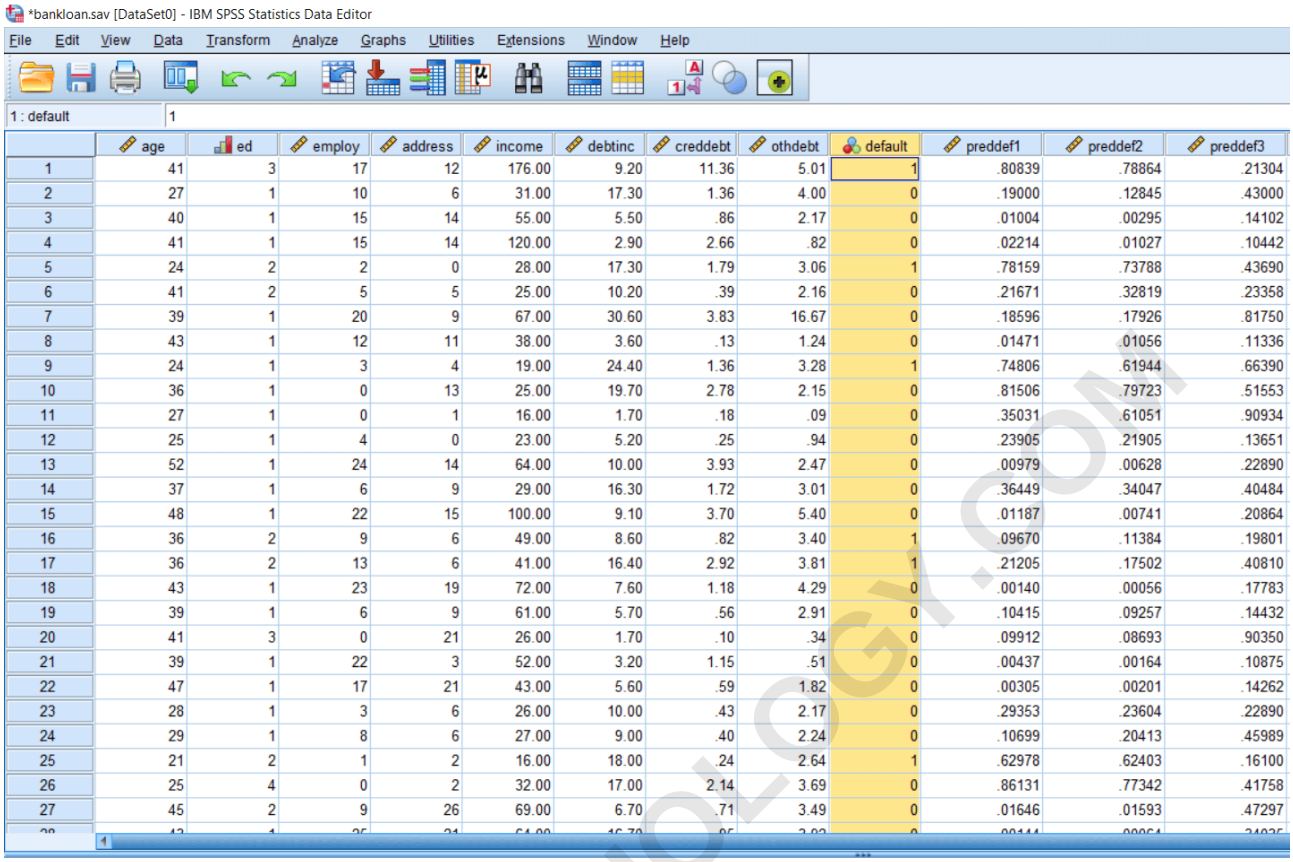
*bankloan.sav [DataSet0] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

27 :

	age	ed	employ	address	income	debtinc	creddebt	othdebt	default	preddef1	preddef2	preddef3	var
1	41	3	17	12	176.00	9.20	11.36	5.01	1	.80839	.78864	.21304	
2	27	1	10	6	31.00	17.30	1.36	4.00	0	.19000	.12845	.43000	
3	40	1	15	14	55.00	5.50	.86	2.17	0	.01004	.00295	.14102	
4	41	1	15	14	120.00	2.90	2.66	.82	0	.02214	.01027	.10442	
5	24	2	2	0	28.00	17.30	1.79	3.06	1	.78159	.73788	.43690	
6	41	2	5	5	25.00	10.20	.39	2.16	0	.21671	.32819	.23358	
7	39	1	20	9	67.00	30.60	3.83	16.67	0	.18596	.17926	.81750	
8	43	1	12	11	38.00	3.60	.13	1.24	0	.01471	.01056	.11336	
9	24	1	3	4	19.00	24.40	1.36	3.28	1	.74806	.61944	.66390	
10	36	1	0	13	25.00	19.70	2.78	2.15	0	.81506	.79723	.51553	
11	27	1	0	1	16.00	1.70	.18	.09	0	.35031	.61051	.90934	
12	25	1	4	0	23.00	5.20	.25	.94	0	.23905	.21905	.13651	
13	52	1	24	14	64.00	10.00	3.93	2.47	0	.00979	.00628	.22890	
14	37	1	6	9	29.00	16.30	1.72	3.01	0	.36449	.34047	.40484	
15	48	1	22	15	100.00	9.10	3.70	5.40	0	.01187	.00741	.20864	
16	36	2	9	6	49.00	8.60	.82	3.40	1	.09670	.11384	.19801	
17	36	2	13	6	41.00	16.40	2.92	3.81	1	.21205	.17502	.40810	
18	43	1	23	19	72.00	7.60	1.18	4.29	0	.00140	.00056	.17783	
19	39	1	6	9	61.00	5.70	.56	2.91	0	.10415	.09257	.14432	
20	41	3	0	21	26.00	1.70	.10	.34	0	.09912	.08693	.90350	
21	39	1	22	3	52.00	3.20	1.15	.51	0	.00437	.00164	.10875	
22	47	1	17	21	43.00	5.60	.59	1.82	0	.00305	.00201	.14262	
23	28	1	3	6	26.00	10.00	.43	2.17	0	.29353	.23604	.22890	
24	29	1	8	6	27.00	9.00	.40	2.24	0	.10699	.20413	.45989	
25	21	2	1	2	16.00	18.00	.24	2.64	1	.62978	.62403	.16100	
26	25	4	0	2	32.00	17.00	2.14	3.69	0	.86131	.77342	.41758	
27	45	2	9	26	69.00	6.70	.71	3.49	0	.01646	.01593	.47297	

We have taken this data from the SPSS folder. There is bank loan data, and we are having variables like age, education, years with current employer, address, income and other such variables along with the loan default. The researcher might be interested in finding out whether highly educated persons default less loan as compared to persons who are less educated. So we are having default in the format of frequency whether a person defaulted or not. In our case, not default is measured as 0 and default is measured as 1.



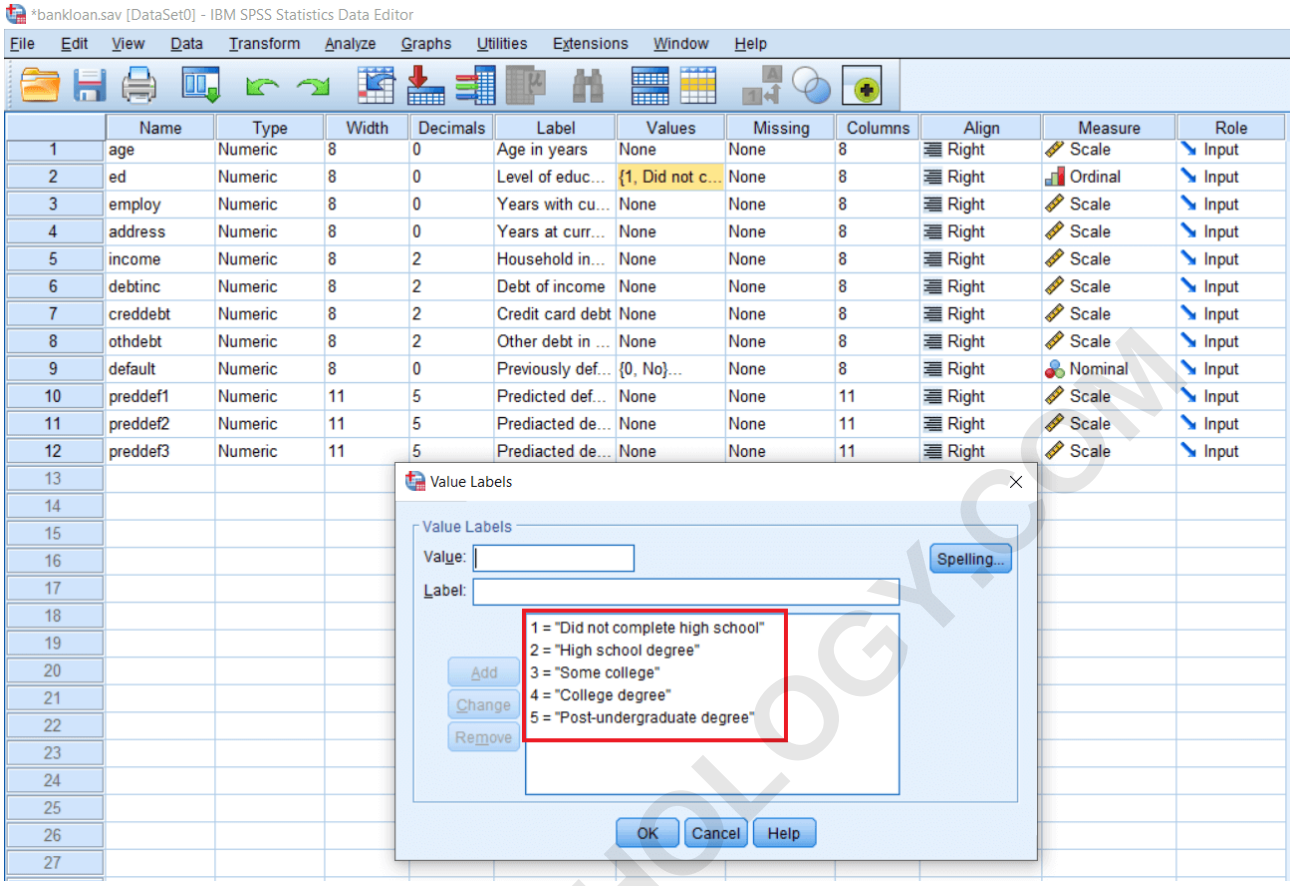
*bankloan.sav [DataSet0] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

1 : default 1

	age	ed	employ	address	income	debtinc	creddebt	othdebt	default	preddef1	preddef2	preddef3
1	41	3	17	12	176.00	9.20	11.36	5.01	1	.80839	.78864	.21304
2	27	1	10	6	31.00	17.30	1.36	4.00	0	.19000	.12845	.43000
3	40	1	15	14	55.00	5.50	.86	2.17	0	.01004	.00295	.14102
4	41	1	15	14	120.00	2.90	2.66	.82	0	.02214	.01027	.10442
5	24	2	2	0	28.00	17.30	1.79	3.06	1	.78159	.73788	.43690
6	41	2	5	5	25.00	10.20	.39	2.16	0	.21671	.32819	.23358
7	39	1	20	9	67.00	30.60	3.83	16.67	0	.18596	.17926	.81750
8	43	1	12	11	38.00	3.60	.13	1.24	0	.01471	.01056	.11336
9	24	1	3	4	19.00	24.40	1.36	3.28	1	.74806	.61944	.66390
10	36	1	0	13	25.00	19.70	2.78	2.15	0	.81506	.79723	.51553
11	27	1	0	1	16.00	1.70	.18	.09	0	.35031	.61051	.90934
12	25	1	4	0	23.00	5.20	.25	.94	0	.23905	.21905	.13651
13	52	1	24	14	64.00	10.00	3.93	2.47	0	.00979	.00628	.22890
14	37	1	6	9	29.00	16.30	1.72	3.01	0	.36449	.34047	.40484
15	48	1	22	15	100.00	9.10	3.70	5.40	0	.01187	.00741	.20864
16	36	2	9	6	49.00	8.60	.82	3.40	1	.09670	.11384	.19801
17	36	2	13	6	41.00	16.40	2.92	3.81	1	.21205	.17502	.40810
18	43	1	23	19	72.00	7.60	1.18	4.29	0	.00140	.00056	.17783
19	39	1	6	9	61.00	5.70	.56	2.91	0	.10415	.09257	.14432
20	41	3	0	21	26.00	1.70	.10	.34	0	.09912	.08693	.90350
21	39	1	22	3	52.00	3.20	1.15	.51	0	.00437	.00164	.10875
22	47	1	17	21	43.00	5.60	.59	1.82	0	.00305	.00201	.14262
23	28	1	3	6	26.00	10.00	.43	2.17	0	.29353	.23604	.22890
24	29	1	8	6	27.00	9.00	.40	2.24	0	.10699	.20413	.45989
25	21	2	1	2	16.00	18.00	.24	2.64	1	.62978	.62403	.16100
26	25	4	0	2	32.00	17.00	2.14	3.69	0	.86131	.77342	.41758
27	45	2	9	26	69.00	6.70	.71	3.49	0	.01646	.01593	.47297

For education, we are having various categories ranging from 1 to 5 like this:



So it could be very interesting to find out if people across different educational categories significantly differ from each other in terms of loan default. In our case, the independent variable is educational qualification or educational category and loan default is our dependent variable. The independent variable and dependent variable are both coming in the format of frequency. So we cannot apply any parametric tests like t-test or ANOVA. In this case, a suitable test will be the Pearson chi-square test.