KBA-8 - Chi-Squared Test

M. Degonish

March 26, 2023

1 Chi-Squared Test

There are all large number of inferential statistics that assume the data involved are scores on some measure calculated from sampels drawn from pouplations that are normally distributed (e.g., Regression, ANOVA, correlation, and t tests). Of course, these conditions are often not met in practice. The violation of these assumptions represent a sort of good news, bad new situation. The bad news is that if the assumptiosn are violated to an alarming degree, the results of these statistics can be difficult to interpret, even meaningless. The good news is that "to an alarming degree" is an imprecise phrase and is open to interpretation. In many situations, violating assumptions of normally distributed data do not make the results invalid, or event alter them very much. Another pieced of good news is that even when the assumptions of these statistics are horrifically violated, there is a whole batch of statistics that researchers can use that do not have the same assumptions of normality and random seleciton: **nonparameteric statistics**.

There are a number of nonparametric tests available, however, this KBA will focus on one of the most commonly used nonparameteric tests: the chi-squared test of independence. This test is appropriate for use when you have data from two categorial variables. When you have two (or more) categorial variables, you may want to know whether the division of cases in one variable is independent of the other categorical variable. For example, suppose you have a sample of component failures from a set of U.S. nuclear power plants. You may want to know whether the component failure rates depend on the site or if the component failure rate is independent of individual sites. That is the type of question that the Chi-Squared test of independnce (from herein referred to as the Chi-Squared test) was designed to answer.

2 Leading by Example

Over the years 2015-2020, the following set of data related to portable combustion turbine generators failing to run was collected:

Plant	**No. of Failures**	**Exposure (hours)**
Plant 1	3	36
Plant 2	3	74
Plant 3	0	24
Plant 4	0	14
Plant 5	0	48
Plant 6	0	4
Plant 7	1	26

What we are interested in here is whether the failure rate for combustion turbine generators is *dependent* or *independent* on individual plants. In other words, we want to determine if a single failure rate can accurately model the failure rate of combustion turbine generators at the industry level, or if failure rates are necessary at the plant level. To determine this, the Chi-Squared test can be used. The formula for calculating the Chi-Squared statisticis is:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Where, **O** is the observed failure rate for each plant, and **E** is the expected failure rate for each plant.
So, in our example, we have the observed number of failures at each plant.

Plant	**No. of Failures (O)**	**Exposure (hours)**
Plant 1	3	36
Plant 2	3	74
Plant 3	0	24
Plant 4	0	14
Plant 5	0	48
Plant 6	0	4
Plant 7	1	26

The next piece of information that is required is the expected number of failures for each plant. To generate this value, we look first at our expectation of the failure rate. The most likely expectation is that the failure rates for all plants are equal, and thus this will be our null hypothesis. If this is case, we'd expect the mean failure rate to be able to be determined from the sample data. In this case, we simply take the ratio of total number of failures, to the total exposure time. Once we have this estimate, we can calculate the failure estimate for each plant by multiplying this estimate by the exposure time for each plant.

Plant	**No. of Failures (O)**	**Exposure (hours)**	**No. of Failures (E)**
Plant 1	3	36	1.12
Plant 2	3	74	2.29
Plant 3	0	24	0.74
Plant 4	0	14	0.43
Plant 5	0	48	1.49
Plant 6	0	4	0.12
Plant 7	1	26	0.81
Sum	**7**	**226**	

From this table, we can calculate the Chi-Squared statistic for each plant, and then sum the results for each plant to get the total Chi-Squared statistic.

Plant	**No. of Failures (O)**	**Exposure (hours)**	**No. of Failures (E)**	χ^2
Plant 1	3	36	1.12	3.19
Plant 2	3	74	2.29	0.22
Plant 3	0	24	0.74	0.74
Plant 4	0	14	0.43	0.43
Plant 5	0	48	1.49	1.49
Plant 6	0	4	0.12	0.12
Plant 7	1	26	0.81	0.05
Sum	**7**	**226**		**6.24**

For the data in question, we have produced an observed χ^2 value of 6.24. We must compare this value to a critical χ^2 value to determine whether the failure rate between plants is *statistically significant*. To determine a critical χ^2 value we must: (1) determine the degress of freedom for the problem and (2) determine the confidence level (i.e., the alpha level). The degrees of freedom can be calculated as follows:

$$DOF = (R-1)(C-1)$$

Where, **R** is the number of rows and **C** is the number of columns in our contingency table. In our case we have 7 rows (7 plants) and 2 columns (observed number of failures, and exposures). Therefore, the degrees of freedom for our example problem is equal to 6. We select an alpha level of 0.05 which corresponds to a confidence level of 95 percent. The coresponding critical χ^2 value is 12.59, based on standard tables.
From here, we need to determine whether or not we have enough evidence to reject the null hypothesis. Given that our critical value is greater than our test statistic, we do not have enough evidence to reject the null hypothesis.

3 References

1. NUREG/CR-6823, *Handbook of Parameter Estimation for Probabilistic Risk Assessment*, U.S. Nuclear Regulatory Commission, September 2003.