12-1-2020

Using the Chi-Square Test to Analyze Voter Behavior

Bailey Fadden
Boise State University

USING THE CHI-SQUARE TEST TO ANALYZE VOTER BEHAVIOR

BAILEY FADDEN

Abstract. We explain the Chi-Square Test and how to use it to analyze voter behavior. Specifically we look at the behavior of U.S citizens and whether or not they voted in the 2016 U.S presidential election, and how this relates to income.

Some Americans choose to vote in presidential elections and some do not. Surely there are many factors that have an effect on each citizen's decision. Stereotypically, it is believed that the lower someone's income is, the less likely they are to vote in an election. To test this hypothesis, we can use a statistical test called the Chi-Square test.

The Chi-Square Test was conceived by Karl Pearson, referred to by some as the "father of modern statistics" (Ramana PV). Pearson first wrote about the Chi-Square test and distribution in 1900. The Chi-Square Test is a test commonly used when analyzing categorical data. Some examples of categorical data would be voters and non-voters, or below the poverty level and above the poverty level. Specifically, the Chi-Square Test is used to test the null hypothesis that there is "no association between two or more groups, population, or criteria." (Rana R, Singhal R.) It is also used to test the goodness of fit of the observed data with the expected data. When obtaining a Chi-Square Test statistic, you will need a formula. The Chi-Square Test formula is,

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{0.1}$$

To use this formula, we must obtain an observed frequency of a categorical event (O), and an expected frequency of the same categorical event (E). Once

Date: 10/16/2020.

we obtain a Chi-Square Test statistic (χ^2), we can look at a Chi-Square distribution table and find the significance of the value we have obtained. The Chi-Square distribution table will provide us with a significant P-value. In order to obtain the P-value, we need to obtain the degrees of freedom. This can be found with this equation,

$$df = (\#\text{of Rows} - 1) * (\#\text{of Columns} - 1) \tag{0.2}$$

Once we obtain the degrees of freedom, we will use a P-value chart or an online P-value calculator to get the final P-value.

On a Chi-Square distribution table, From left to right, the first value that we find that is larger that our Chi-Square value will be the value that corresponds with the appropriate P-value. The P-value tells us whether the correlation between the two or more groups is significant. Typically, the accepted significance level is a P-value of 0.05 or less. If the P-value is less than 0.05, we can assume that the correlation between the groups is significant, meaning we should reject the null hypothesis.

For this experiment, the null hypothesis would be that there is no correlation between income and voter turnout. Some might assume that low income correlates with a lack of voting, there is some reasoning behind this. People with a low income are usually surviving paycheck to paycheck. To pay the bills each month, they might be working multiple jobs. Election day is no national holiday. If one works all day on the first Tuesday of November, they may not be able to make it to their polling place while the polls are open. Voter suppression is a major problem in America. Forms of voter suppression include strict ID laws, no early voting, voter registration deadlines, disability accessibility, and a lack of polling locations to name a few. We can currently state that it seems like there might be a correlation between income and voting. In order to state whether or not the correlation is significant, we should use the Chi-Square test. If income had no correlation to whether or not someone voted, ideally a similar percentage of each income category would have confirmed voting. After looking at some data, this is not the case.

The data used for this test came from the United States Census Bureau. This data was collected from the 2016 November presidential election. Table 1

shows the number of U.S Citizens 18 and older that reported voting in the 2016 election. The data is separated into categories based on household income. The categories go from under $10,000 to over $150,000 in increments that are relevant to the data set. When looking at this table, it may be important to note that in 2016, the poverty level was $11,770 for someone living alone. The average family consists of about 3 people in a household. For a household size of 3, the poverty level was $20,090. As you can see on Table 1, the categories from under $10,000 to $19,000 have the lowest percentages of reported voters. As we previously discussed, there is a stereotypical relationship between these categories and reported voters. The Chi-Square test will help us to determine whether reported voters and household income are significantly related.

According to the article, "The Chi-Square Test of Independence", distribution free tests like the Chi-Square test can and should be used under a few conditions. We can use this test to analyze the census data because the level of measurement is ordinal, the "sample sizes of the study groups are unequal", and the "continuous data were collapsed into a small number of categories" (McHugh).

When we use the Chi-Square test, we are going to need observed data and expected data. For the observed data, we can use Table 1. In this table, totals have been included in order to use them later for expected data calculations. It is easy to see, even by the naked eye, that the number of people who voted grows disproportionately to the number of people who did not vote. For the expected data, we will have to come up with what we should expect if there was no significant relationship between yearly income and whether or not people vote. We should expect that the same percentage of each income category would be reported voters. We cannot assume that 100% of people will be voting because that is unrealistic. The total percentage of reported voters is 63.5%. We could assume that 63.5% of each income category would report voting. To be even more accurate, we can use an expected data equation. We need to calculate an expected data value for each individual result from Table 1. For each value in the table, you would calculate,

$$\text{Expected} = \frac{\text{Row Total} * \text{Column Total}}{\text{Grand Total}} \tag{0.3}$$

For example, we can use equation (0.3) to calculate the expected number of voters in the "Under 10,000" category. The row total would be 4,142 and the column total would be 102,840. The grand total for all of these calculations will be 162,061. If you put these numbers into the equation,

$$\frac{4,142 * 102,840}{162,061} = 2,628 \tag{0.4}$$

As you can see, this value matches the first value in the expected data table. After all of the expected values are calculated, the Chi-Square Test can be used to calculate a Chi-Square test statistic by using equation (0.1),

$$\begin{aligned} \chi^2 &= \frac{(1,713 - 2,628)^2}{2,628} + \frac{(1,934 - 935)^2}{935} + \cdots \\ &= 318.58 + 1,067.38 + \cdots \\ &= 23,887 \end{aligned} \tag{0.5}$$

Now that we have a χ^2, we need to find the degrees of freedom in order to get a p-value. Using equation (0.2),

$$(11 - 1) * (3 - 1) = 20 \tag{0.6}$$

We find that our χ^2 statistic has 20 degrees of freedom. With a χ^2 value this large and 20 degrees of freedom, our P-value is virtually zero. Zero is less than our level of significance 0.05 so we can reject our null hypothesis and assume that there is a relationship between income and whether or not people are voting.

TABLE 1. Observed Data (In 1,000's)

Yearly Income($)	Voted	Did Not Vote	No Response	Total
Under 10,000	1,713	1,934	495	4,142
10,000 - 14,999	1,898	1,683	402	3,982
15,000 - 19,999	1,652	1,363	371	3,386
20,000 - 29,999	5,167	3,837	991	9,996
30,000 - 39,999	7,340	3,914	1,207	12,461
40,000 - 49,999	6,323	2,885	927	10,135
50,000 - 74,999	16,761	6,128	1,888	24,777
75,000 - 99,999	13,755	3,903	1,595	19,253
100,000 - 149,000	16,601	3,545	1,576	21,722
150,000 +	15,777	2,357	1,572	19,646
N/A	15,853	5,016	11,692	32,561
Total	102,840	36,566	22,655	162,061

TABLE 2. Expected Data (In 1,000's)

Yearly Income	Voted	Did Not Vote	No Response
Under 10,000	2,628	935	569.023
10,000 - 14,999	2,527	898	556.656
15,000 - 19,999	2,149	764	473
20,000 - 29,999	6,343	2,255	1,397
30,000 - 39,999	7,907	2,812	1,742
40,000 - 49,999	6,431	2,287	1,416.80
50,000 - 74,999	15,723	5,590	3,464
75,000 - 99,999	12,217	4,344	2,691
100,000 - 149,999	13,784	4,901	3,037
150,000 +	12,467	4,433	2,746
N/A	20,662	7,347	4,552

REFERENCES

[1] Singhal R. Rana R., *Chi-square test and its application in hypothesis testing* (2015), 69-71. ↑

[2] Mary L. Mchugh, *The Chi-Square Test of Independence*, posted on 2013, 143–149, DOI 10.11613/bm.2013.018. ↑

[3] Congress United States Voting and Registration, *Voting and Registration in the Election of November 2016* (2017). ↑

[4] Park View MC, *2016 Federal Poverty Level Chart* (2016). ↑

[5] Erin Duffin, *Average Family Size in the US 1960-2019* (28 Novm 2019). ↑