R Project 1:
The Chi-Square Distribution

Data Analytics 2 – ADS523

James Sears

BAY PATH UNIVERSITY

Applied Data Science Program, Graduate School, Longmeadow, MA

Submitted To: Professor Aja Shabana

Spring 2021

Jim Sears
ADS5232 – Data Analytics 2
R Project 1

Introduction

The object of this project is to investigate how the chi-square distribution is created and what happens as the degrees of freedom increase. The chi-square probability distribution is one that enables hypotheses testing regarding the distribution of the possible outcomes of one or multiple qualitative (categorical or nominal) variables of a population or populations. The hypothesis tests involve checking if observed frequencies in one or more categories match expected frequencies with statistical significance. These possible types of tests include:

1. *Goodness-Of-Fit* Test: determining if an observed frequency distribution of a single population follows an expected frequency distribution or not.
2. Test for *Independence*: determining whether two qualitative variables of a single population are likely to be related (dependent) or not (independent).
3. Test for *Homogeneity*: determining whether different populations are the same, or not, regarding their proportions of multiple variables.

The chi-square distribution is used to determine how well an observed distribution of values compares to the theoretical one. The differences between theoretical – or expected – values and those that are actually observed are known as errors, and chi-square (χ^2) is the measurement of the accumulation of those errors. The number of values that the chi-square distribution - or any - statistical analysis can theorize are called the degrees of freedom. In this case, the degrees of freedom can represent the number of possible variable outcomes (minus one) in the *Goodness-Of-Fit Test* or the number of possible variable combinations (product of the number of outcomes of the first variable minus one and the number of outcomes of the second variable minus one) in the *Tests for Independence and Homogeneity*. With each additional variable and/or possible variable outcome, it is reasonable that there will be additional opportunities for error. For this reason, and as stated above, chi-square (χ^2) is the measurement of the accumulation of those errors. More precisely, it is the sum of the square of the errors.

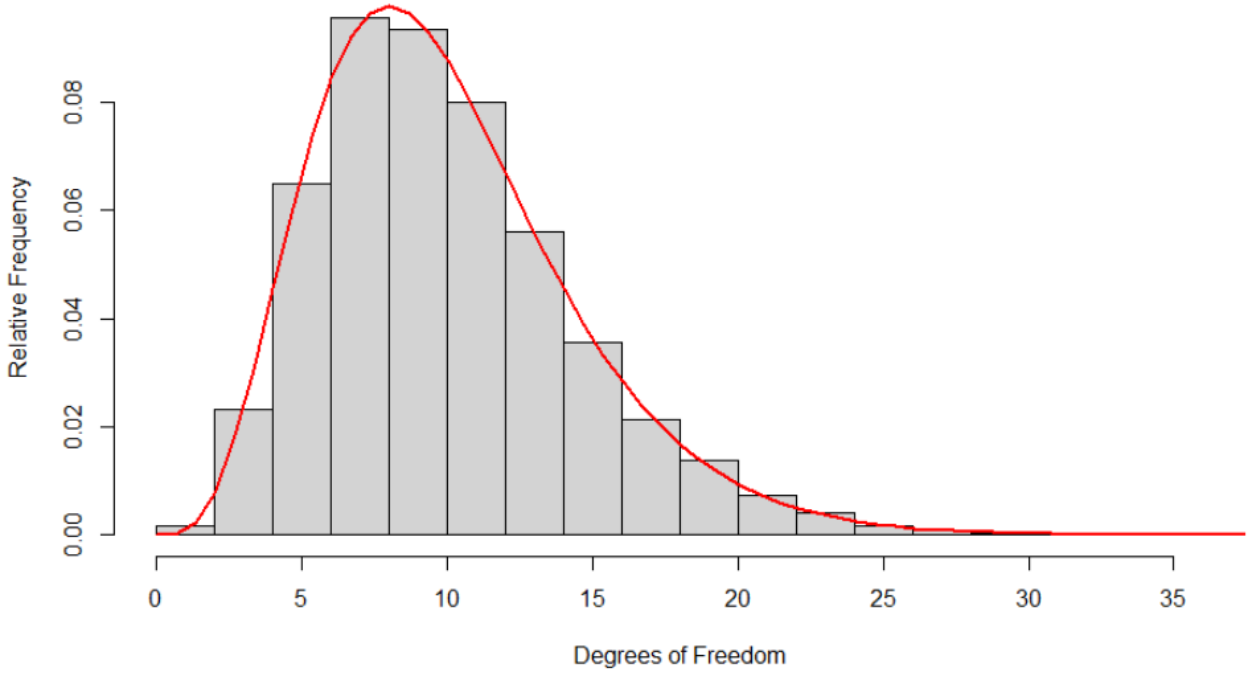
Creating and Describing a Chi-Square Distribution

A number selected from a normal distribution is expected to be close to the mean (μ). And as the Empirical Rule is applied to a normal distribution, there is a 68% probability the number will be within one standard deviation (σ, SD), a 95% probability it will be within 2 SDs, and a 99.7% probability it will be within 3 SDs. For example, if a number is randomly selected from a normal distribution that has a mean (μ) of zero and a standard deviation (σ) of 1, it is likely to be close to zero (the mean). More precisely, according to the Empirical Rule, there is a 68% probability it is between -1 and +1, 95% probability it is between -2 and +2, and 99.7% probability it is between -3 and +3.

The chi-square distribution is based on the normal distribution. It can be simulated by randomly drawing a set number of values from a normal distribution, squaring them, and then adding them together. Figure 1 demonstrates this simulation 10,000 times with ten numbers drawn. The expected chi-square distribution curve therefore would have ten degrees of freedom and is overlaid in red. The mean (μ) of a chi-square distribution is equal to the number of degrees of freedom (*df*) and the standard deviation (σ) is equal to the square root of two times *df*. It is noted that the distribution is skewed right. This is because the normally distributed values are squared before they are added together.

Jim Sears
ADS5232 – Data Analytics 2
R Project 1

Figure 1. Histogram of 10,000 repetitions of the sum of squared errors for 10 normally distributed values drawn randomly overlayed with a chi-square distribution curve with 10 degrees of freedom (red).



As noted above, normally distributed numbers are more likely to be close to zero, with 68% of the values expected to be between -1 and +1. Numbers in this range are even closer to zero when they are squared. As a result, the sum of a set of squared values that are likely to be close to zero is likely to be less than the number of items in that set (the corresponding degrees of freedom). This is why the median of the distribution is less than the mean, represented by the taller bars of the histogram and the left peak of the representative chi-square distribution curve in Figure 1. Conversely, values greater than one are less likely (32% probability), which can be seen in the shorter bars on the right side of the histogram and resulting right tail of the representative chi-square distribution curve. A distribution with these attributes is termed as right-skewed.

Degrees of Freedom and the Chi-Square Distribution

It was noted in the previous section that the sum of a set of squared values that are likely to be close to zero is likely to be less than the number of items in that set (also known as the degrees of freedom). If the degrees of freedom are changed, it is reasonable that the resulting distribution will change as well because there are more squared terms that are being added together. This is demonstrated by Figures 2 and 3 that show the theorhetical chi-square distribution curves of various increasing degrees of freedom, including one, two, five, ten, fifty, and one hundred. As the degrees of freedom increase, the distribution curve widens and its peak is shorter. This can be interpretted as the probabilty of each of the additional expected values is lower because it is being

distributed amongst an increasing number of variables and/or the number their possible outcomes (degrees of freedom).

Figure 2. Chi-square distributions of various degrees of freedom: one (black), two (orange), five (purple), and ten (red).

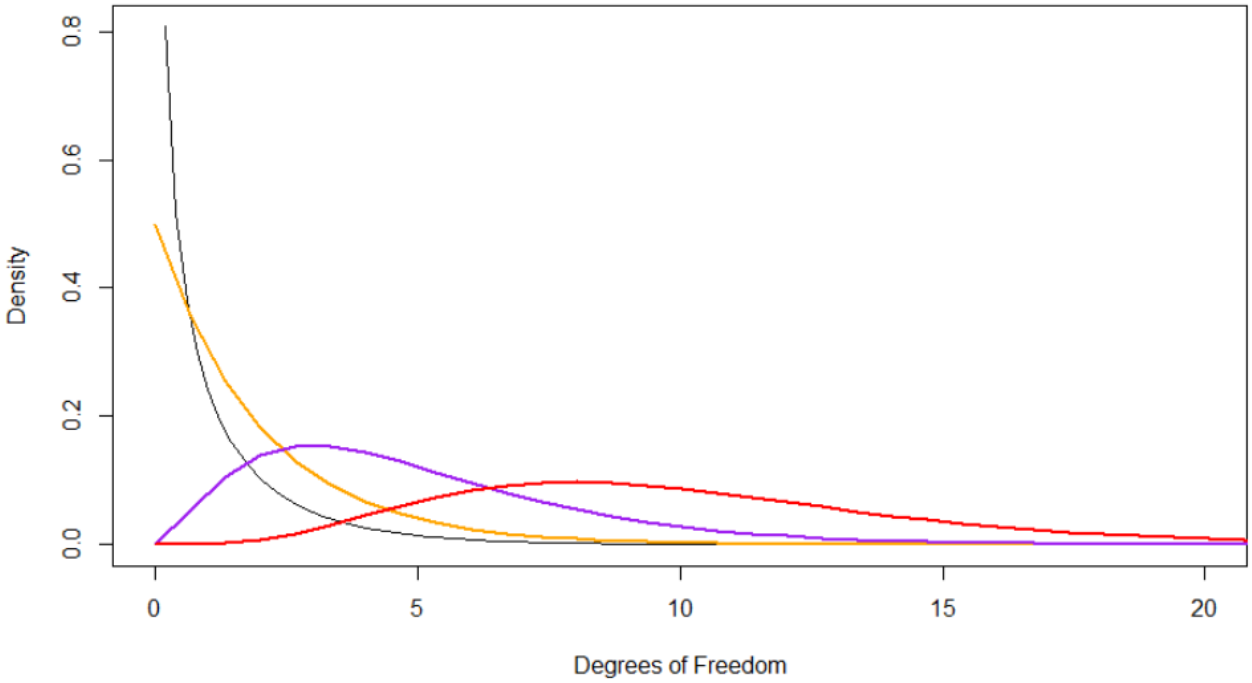
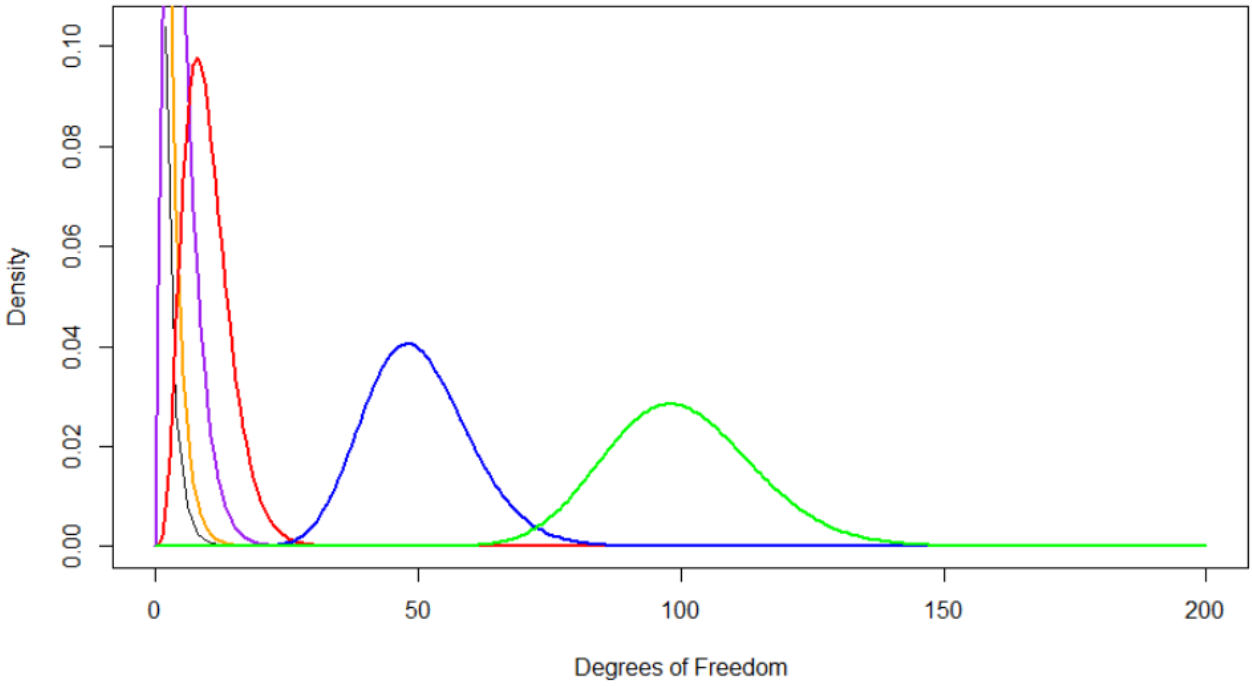


Figure 3. Chi-square distributions of various degrees of freedom: one (black), two (orange), five (purple), ten (red), fifty (blue), and one hundred (green).



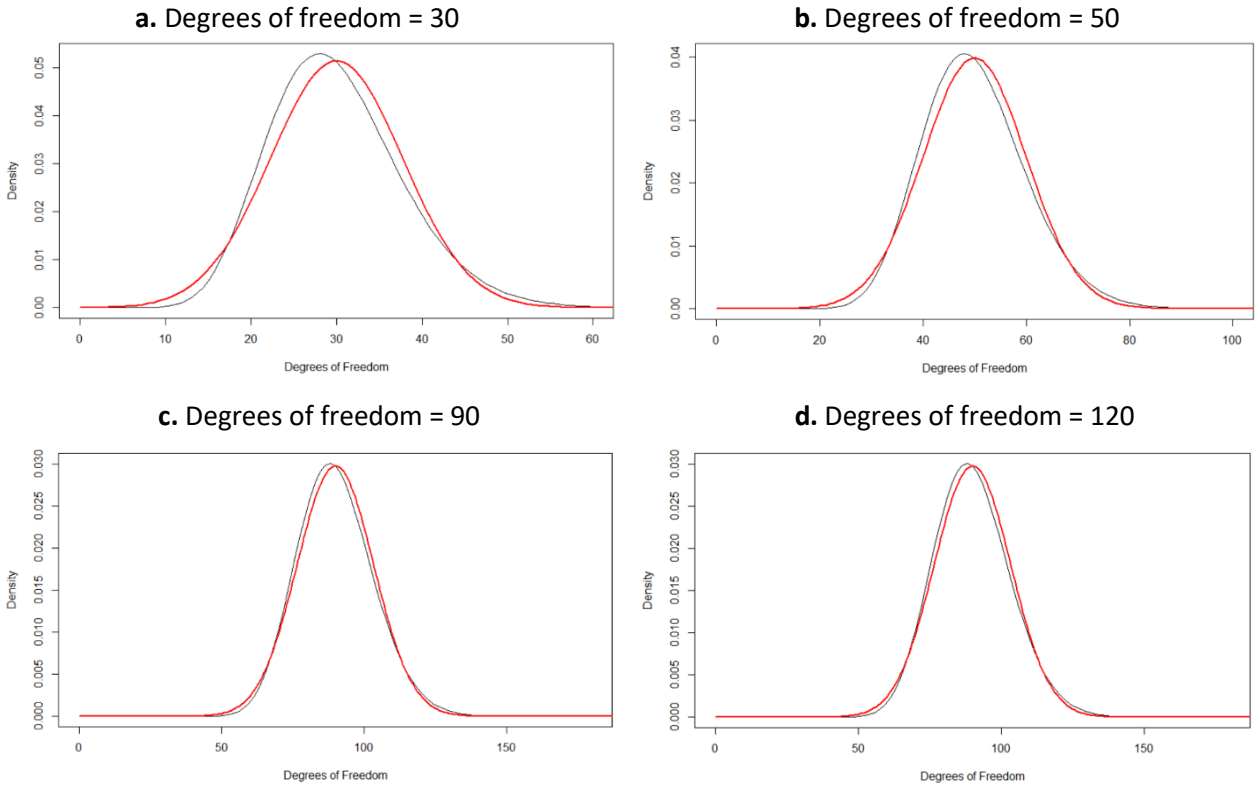
Jim Sears
ADS5232 – Data Analytics 2
R Project 1

Approximating the Chi-Square Distribution with the Normal Distribution

It is noted in Figures 2 and 3 that chi-square distribution curve becomes increasingly symmetrical as the number of degrees of freedom increase. This is expected as the Central Limit Theorem states that regardless of the shape of the underlying population, the distribution of independent standard normal variables becomes approximately normal as their quantity increases. Further, the Central Limit Theorem is invoked when the minimum number of values, samples, degrees of freedom, etc. is "large enough"; as a rule of thumb for a skewed distribution, "large enough" is greater than or equal to 30.

Testing at which number of degrees of freedom that the chi-square distribution appears approximately normal can be done graphically and algebraically. Generating chi-square distribution curves and corresponding normal distribution curves (mean = *df*, standard deviation = √2*df*) with various degrees of freedom (Figure 4) reveals that the former better approximates the latter with degrees of freedom greater than 30. How well one curve "fits" the other is not only demonstrated by how closely they overlap, but by the difference of their densities at the same points of the curves (or at specific *n* observations).

Figure 4. Chi-square distribution with various degrees of freedom (black) and the corresponding normal distribution with a mean = *df*, standard deviation = √2*df* (red).



Jim Sears
ADS5232 – Data Analytics 2
R Project 1

Table 1 contains the calculated densities for a selection of vales of *n* of the four distribution curve pairings that correspond with those of Figure 4. The actual values of *n* are not important except that they represent where the chi-square and normal distribution curves vary the most from each other. It is noted that increasing the degrees of freedom greatly decreases the variance between the two curves. This is important as the more normal a chi-square distribution is, the more statistically significant a conclusion of a hypothesis test on a p-value (which is derived from a normal distribution) will be. Therefore, hypothesis tests of higher significance should be conducted with approximately 90 degrees of freedom.

Table 1. Comparison of the chi-square distribution of various degrees of freedom (30, 50, 90, 120) and the corresponding normal distribution (mean = *df*, standard deviation = √2*df*) with a selection of probabilities for *n* observations at (bold) and around where the difference between the two distributions is greatest.

df, mean	n	Chi-Square	Normal Dist	Difference
30	23.0	0.04111	0.03424	0.00687
	23.5	0.04326	0.03622	0.00705
	24.0	**0.04524**	**0.03815**	**0.00709**
	24.5	0.04703	0.04003	0.00700
	25.0	0.04860	0.04182	0.00678
50	40.5	0.02924	0.02541	0.00384
	41.0	0.03057	0.02661	0.00396
	41.5	**0.03185**	**0.02780**	**0.00405**
	42.0	0.03306	0.02897	0.00409
	42.5	0.03421	0.03011	0.00409
90	78.5	0.02276	0.02059	0.00217
	79.0	0.02344	0.02125	0.00220
	79.5	**0.02410**	**0.02189**	**0.00221**
	80.0	0.02473	0.02252	0.00221
	80.5	0.02533	0.02314	0.00219
120	107.0	0.01973	0.01811	0.00162
	107.5	0.02023	0.01860	0.00163
	108.0	**0.02071**	**0.01908**	**0.00164**
	108.5	0.02119	0.01955	0.00164
	109.0	0.02164	0.02001	0.00163