# Project 1: The Chi-Square Distribution
# PHSX815

David Coria

February 22, 2021

## 1 Introduction

The chi-square distribution with $k$ degrees of freedom, or $\chi^2(k)$, is the distribution of the sum of the squares of k independent Gaussian random variables. The chi-square distribution has a mean $= k$ and a variance $= 2k$ [1]. This distribution is a special case of the gamma distribution, and it is most notably known (to me, at least) for its use in the chi-square test for the goodness of fit between observed data and hypothetical distributions [2].

Remember the parameters of the Gaussian Distribution are the mean $\mu$, the standard deviation $\sigma$ and the variance $\sigma^2$. In a standard normal distribution, as required for the chi-square distribution, mean $\mu = 0$, the standard deviation $\sigma = 1$.

Probability Distribution Function (Gaussian) [3]:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{1}$$

Probability Distribution Function ($\chi^2$) [1]:

$$f(x) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} \tag{2}$$

## 2 Significance: Chi-Squared Tests

I wanted to do something for this project using the chi-square distribution in an attempt to understand it better since it sets the foundation for the chi-squared statistic that is used extensively in "goodness of fit" tests. This statistic proved useful in my own research by helping me calculate isotopic abundances in several solar twin stars by comparing the observed spectra to Solar models with varying abundances of the same isotopologue. Here, the chi-squared statistic is defined as the weighted sum of squared deviations [2] or:

$$\chi^2 = \sum_i \frac{(O_i - M_i)^2}{\sigma_i^2} \tag{3}$$

where $O_i$ represents observations (in my case: the flux at a certain point of the observed spectrum), $M_i$ represents the model/calculated data (in my case: the flux at a certain point of the model spectrum),

and $\sigma_i$ represents the corresponding uncertainties (of the observed spectrum, in my case). This $\chi^2$ statistic asymptotically approaches a $\chi^2$ distribution and can then, in turn, be used to calculate a p-value by comparing the statistic to a $\chi^2$ distribution [2]. If you calculate the $\chi^2$ value between the observed $^{13}$CO line and the each of the model lines corresponding to a different quantity of $^{13}$CO, you obtain four points in $\chi^2$ space. These four points can be fit with a parabola, and the minimum of this parabola marks the interpolated "best fit" of the data. This corresponds to a xSolar abundance. This process is pictured/explained a bit more in Figure 1.

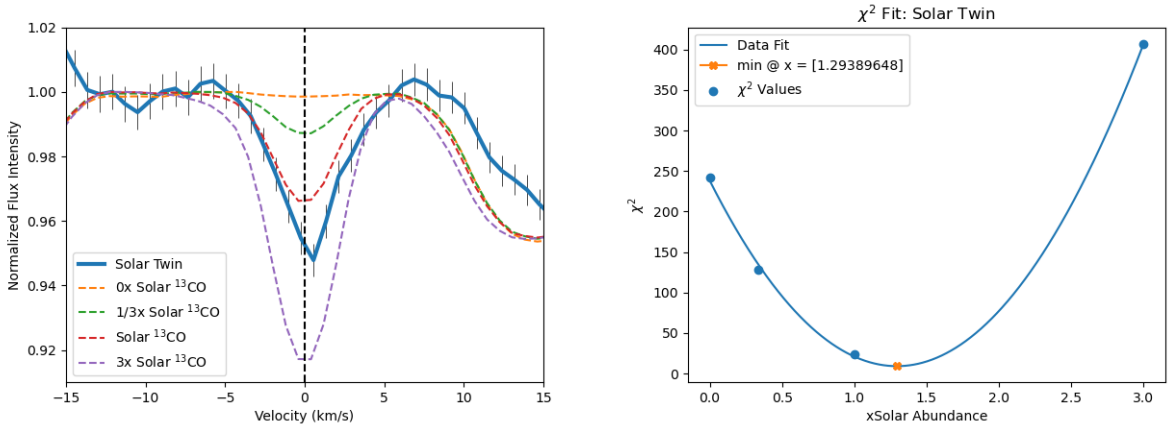


Figure 1: On the left: This is a plot of "stacked $^{13}$CO absorption lines". This process (not super important for you to understand) is done for the observed spectrum of a solar twin and each of the four Solar models representing 0x, $\frac{1}{3}$x, 1x, and 3x Solar $^{13}$CO abundances. This plot is used to compare the amount of a certain species in the observed star to the varying amounts in a set of models. Just taking a quick look, you might say that the observed Solar Twin has a $^{13}$CO abundance between 1x and 3x Solar value. We need to perform a series of chi-squared "goodness of fit" tests to determine an actual value.

On the right: This is a plot of the four $\chi^2$ values calculated between the observed spectrum of a solar twin and each of the four Solar models representing 0x, $\frac{1}{3}$x, 1x, and 3x Solar $^{13}$CO abundances. The $\chi^2$ values are fit with a parabola which is then minimized. This minimum value by definition represents the interpolated best abundance fit of the observed line. This means that this solar twin star has a $^{13}$CO abundance of $\approx 1.47$ xSolar.

## 3 Code and Experimental Simulation

The goal for this project is to use built-in numpy/scipy functions to produce a set of n independent standard Gaussian random variables, output them to a .txt file and then use these to produce a chi-square distribution with k degrees of freedom.

Chi-Square Generator: This file takes the given sample size n from the user and produces k randomly generated standard normal (Gaussian) distributions. Then, the script takes the sum of the squares of the k standard normal distributions to construct a chi-square distribution. The file also creates a histogram of the constructed chi-square distribution and outputs the data to a .txt file if told to do so via user input flags. The script then proceeds to plot a chi-square probability distribution function

with k degrees of freedom to show see how well the $\chi^2$ sample constructed from k Gaussian distributions matches the true $\chi^2(k)$ distribution. It also calculates the mean (and plots it on the histogram) and the variance and tells the user what each value should be theoretically.

# 4 Analysis

The analysis for this project was done for three different randomly generated $\chi^2(k)$ distributions, each with a different k value i.e. degrees of freedom. Below, we can see the histograms created for each data set by the Chi Square Generator.py script. Each plot includes a continuous $\chi^2$ probability distribution function with a k corresponding to the randomly generated data set. The script also calculated the mean and variance of each data set. The calculated values are written in the figure captions.

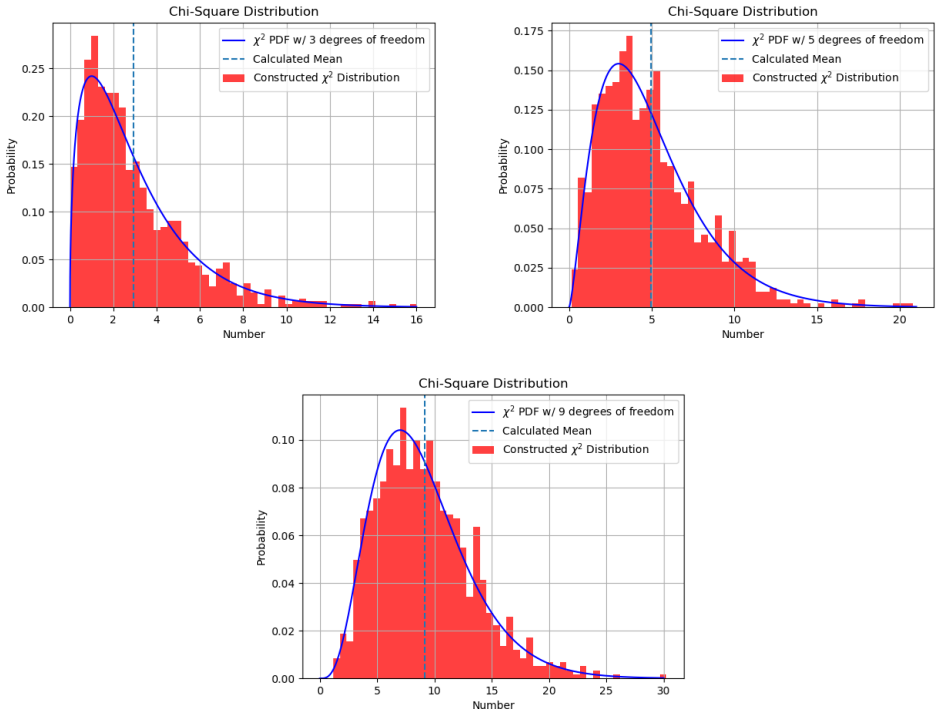


Figure 2: Left: A $\chi^2$ distribution, in red, constructed from a set of three (3) randomly generated standard normal (Gaussian) distributions. The blue line corresponds to the continuous probability distribution function of $\chi^2(3)$. The mean for this set is 2.93. The variance is 5.83.

Center: A $\chi^2$ distribution, in red, constructed from a set of three (5) randomly generated standard normal (Gaussian) distributions. The blue line corresponds to the continuous probability distribution function of $\chi^2(5)$. The mean for this set is 4.95. The variance is 9.58.

Right: A $\chi^2$ distribution, in red, constructed from a set of three (9) randomly generated standard normal (Gaussian) distributions. The blue line corresponds to the continuous probability distribution function of $\chi^2(9)$. The mean for this set is 9.18. The variance is 17.84.

## 5 Conclusion

The continuous $\chi^2$ distributions seem to agree with the $\chi^2$ distributions constructed from a set of k- randomly generated standard normal (Gaussian) distributions. Furthermore, the mean calculated for each data set agrees well with the theoretical mean of a $\chi^2$ distribution–that is, the mean should be equal to k, the number of degrees of freedom (or the number of independent Gaussian distributions used to produce the final $\chi^2$ distribution). Similarly, the variance calculated for each data set agrees with the theoretical variance of a $\chi^2$ distribution which should be equal to $2k$, or double the number of degrees of freedom. I was able to successfully reproduce a $\chi^2$ distribution using a set of independent, randomly generated standard normal (Gaussian) distributions using only built-in numpy and scipy functions.

## References


[1] N/A, *Chi-square distribution*, *Wikipedia* (Feb, 2021) .

[2] N/A, *Pearson's chi-squared test*, *Wikipedia* (Jan, 2021) .

[3] N/A, *Normal distribution*, *Wikipedia* (Feb, 2021) .