

Chi-Squared Distribution



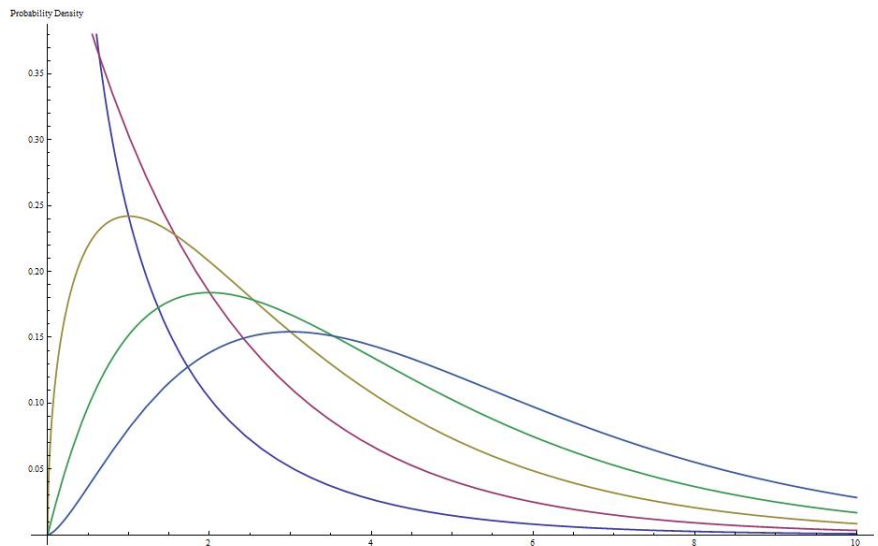
Michael Manser, Subhiskha Swamy, James Blanchard
Econ 7818 HW 5

1 What is it?

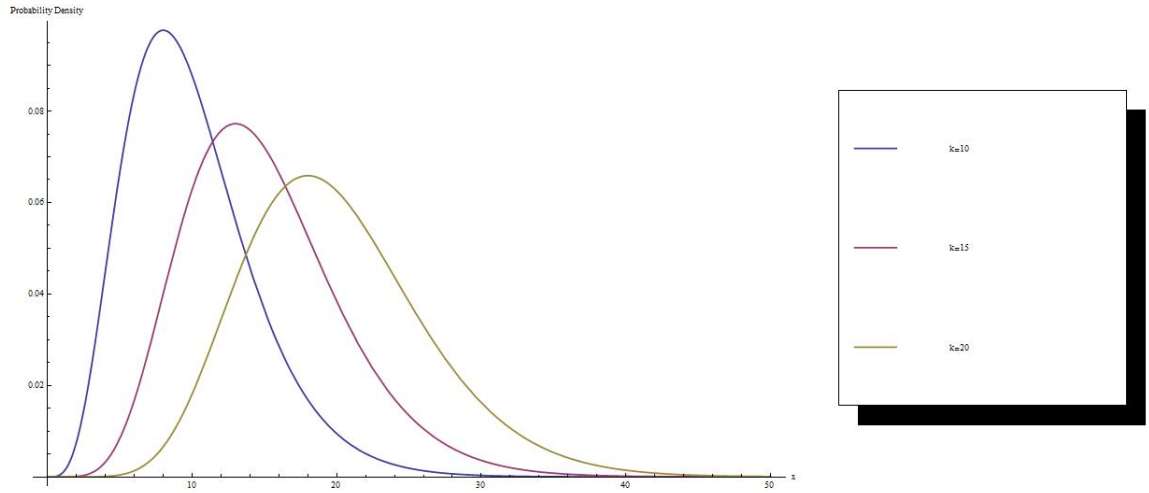
The Chi-Squared distribution is a one parameter distribution with density function:

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

where k is the parameter, x is a random variable with $x \in [0, \infty)$, and $\Gamma(x)$ is the Gamma function, defined as $\Gamma(x) = (x - 1)!$ for integers and $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ when x is not an integer. A few examples with the parameter value, k , varying. In this example k takes values of 1,2,3,4,5.



In the next example $k = 10, 15, 20$.

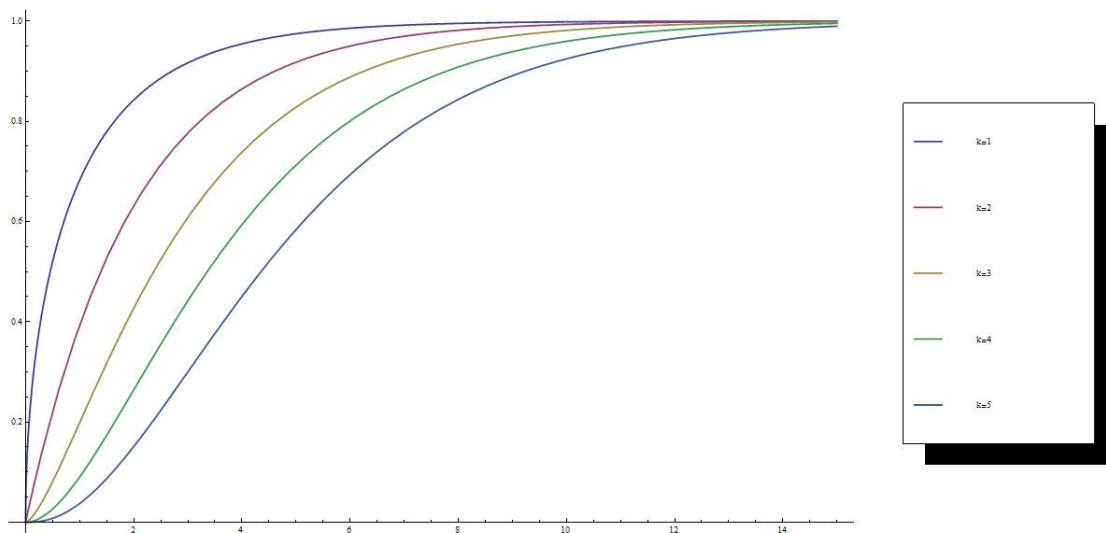


The cumulative distribution function for the Chi-Squared distribution is given by:

$$F(x; k) = \frac{\gamma(k/2, x/2)}{\Gamma(k/2)} = P(k/2, x/2)$$

Where $\gamma(a, x)$ is the lower incomplete gamma function defined by $\gamma(a, x) = \int_0^x t^{a-1} e^{-t} dt$.

Again, a few examples of the Chi-Squared CDF at $k = 1, 2, 3, 4, 5$.



2 Where does it come from and why is it useful?



This is Karl Pearson (1857-1936), the father of modern statistics (establishing the first statistics department in the world at University College London) and the man who came up with the Chi-Squared distribution. Pearson's work in statistics began with developing mathematical methods for studying the processes of heredity and evolution (leading to his aggressive advocacy of eugenics). The Chi-Squared distribution came about as Pearson was attempting to find a measure of the goodness of fit of other distributions to random variables in his heredity and evolutionary modelling. Also note that Ernst Abbe wrote his dissertation in 1863 deriving the Chi-Square distribution, although he switched fields soon after the publication of the paper to optics and astronomy.

It turns out that the Chi-Square is one of the most widely used distributions in inferential statistics. So understanding the Chi-Square distribution is important for hypothesis testing, constructing confidence intervals, goodness of fit, Friedman's analysis of variance by ranks, etc. The distribution is also important in discrete hedging of options in finance, as well as option pricing.

2.1 Measures of Central Tendency

The mean, median, mode and variance for the Chi-Squared distribution are:

$$\text{Mean} = k$$

$$\text{Median} = k\left(1 - \frac{2}{9k}\right)^2$$

$$\text{Mode} = \max(k - 2, 0)$$

2.2 Moment Generating Function

We know that for a continuous distribution the moment generating function takes the general form:

$$M(t) = \int_{-\infty}^{\infty} e^{xt} f(x) dx$$

Thus plugging in the Chi-Squared density function and integrating yields the moment generating function for the Chi-Squared distribution:

$$M(t) = (1 - 2t)^{-k/2}$$

From the moment generating function we can find out lots of information about the Chi-Squared distribution:

$$M'(t) = (k)(1 - 2t)^{-k/2-1} \Rightarrow M'(0) = k$$

which is the mean. Using the same procedure¹ one can find the variance, skewness, kurtosis, etc.

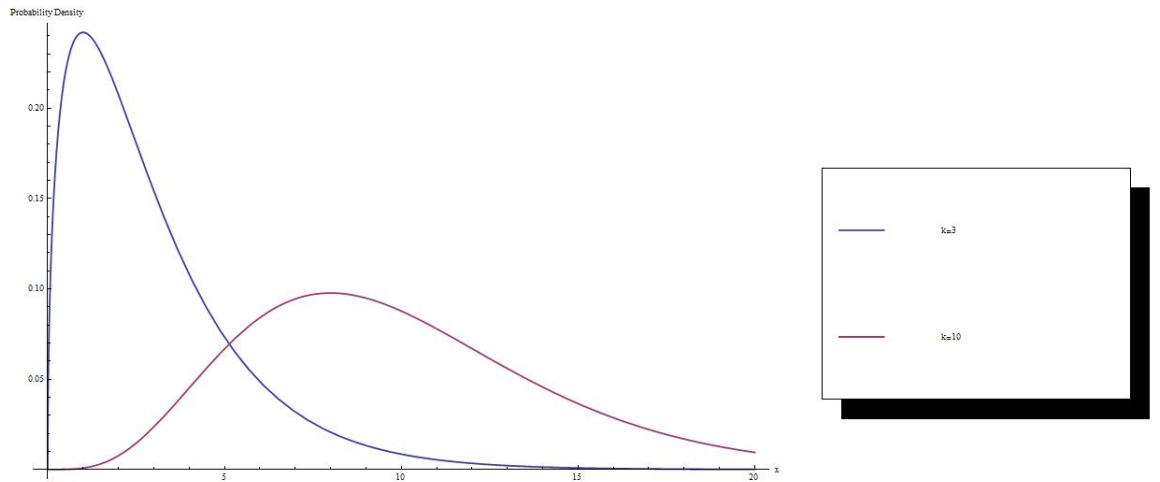
$$\text{Variance} = 2k$$

$$\text{Skewness} = \sqrt{8/k}$$

$$\text{Kurtosis} = 3 + 12/k$$

The graph below provides an illustration of the skewness and kurtosis of the Chi-Squared distribution as the parameter, k , changes. The blue graph represents $k = 3$, and the red graph represents $k = 10$.

¹Note that the first four moments (and perhaps more) can be found using Mathematica. The function for Mathematica are `Mean[ChiSquareDistribution[k]]`, `Variance[ChiSquareDistribution[k]]`, `Skewness[ChiSquareDistribution[k]]`, and `Kurtosis[ChiSquareDistribution[k]]`

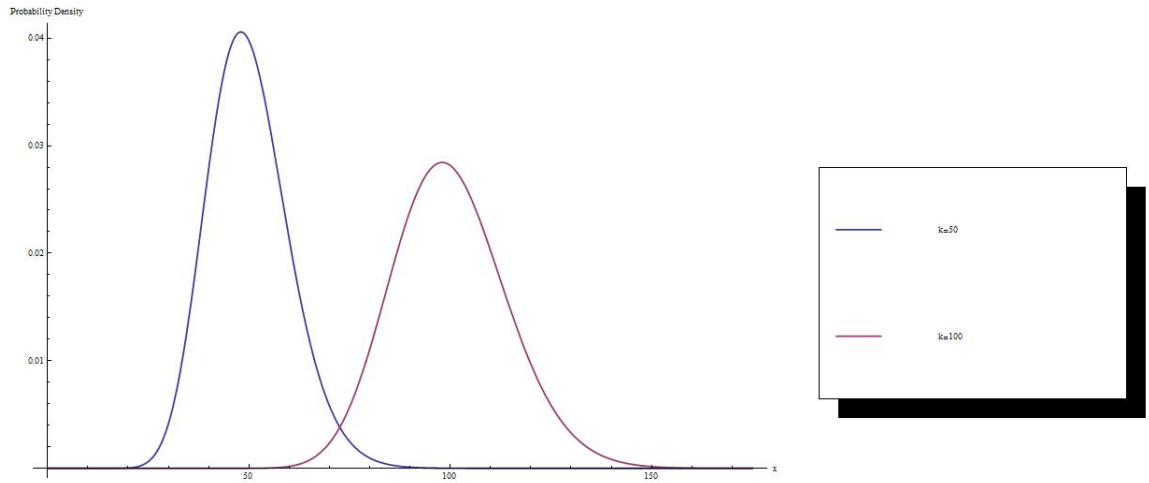


The mean and variance are positively related to k , so the red graph has a higher mean and variance than the blue graph. Skewness and kurtosis, on the other hand, are inversely related to k , so the red graph has a lower skewness and kurtosis than the blue graph².

If we think of skewness as how symmetric the distribution is around the mode, with perfect symmetry as skewness approaches zero, clearly the red PDF is much more symmetric than the blue PDF. Notice how the skewness of the density function changes with values of k , but the curve remains positively skewed. We can think of kurtosis as the "flatness" of the curve around the mode, so the high kurtosis means that the curve will be very steep around the mode. Again, the red and blue graphs illustrate this concept, with the red graph being much flatter than the blue graph around the mode. Another way to think of kurtosis is that the "tails are fatter" when the kurtosis is low.

²We can even calculate these to prove it. $\text{skewness}_{blue} = 1.633$ and $\text{skewness}_{red} = 0.894$, $\text{kurtosis}_{blue} = 7$ and $\text{kurtosis}_{red} = 4.2$

In the next example $k = 50$ and $k = 100$. Notice that as k increases, the PDF tends to flatten out.



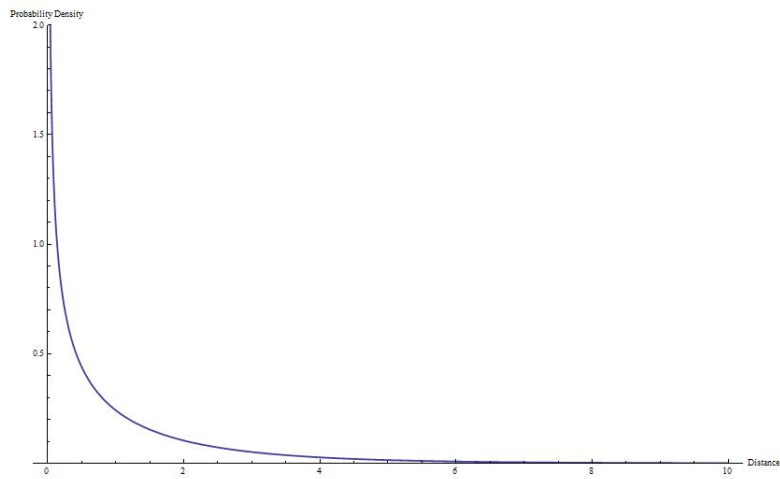
As k approaches ∞ the PDF tends to resemble the normal distribution, which is a result of the Central Limit Theorem³.

³We shall show by the Central Limit Theorem that the Chi-Squared distribution resembles the Normal distribution as k approaches ∞ . So let $Y_i \sim \chi_1^2$, thus $\sum_{i=1}^n Y_i \sim \chi_n^2$. Therefore, by the Central Limit Theorem, $\frac{\sum_{i=1}^n Y_i - nE(Y_i)}{\sigma} \rightarrow N(0, 1)$ which becomes $\frac{\chi_n^2 - n}{\sqrt{2n}\sqrt{n}} \rightarrow N(0, 1)$. Thus $\frac{\chi_n^2}{n\sqrt{2}} - \frac{1}{\sqrt{2}} \rightarrow N(0, 1)$ so $\chi_n^2 \rightarrow N(n, 2n^2)$

3 Examples

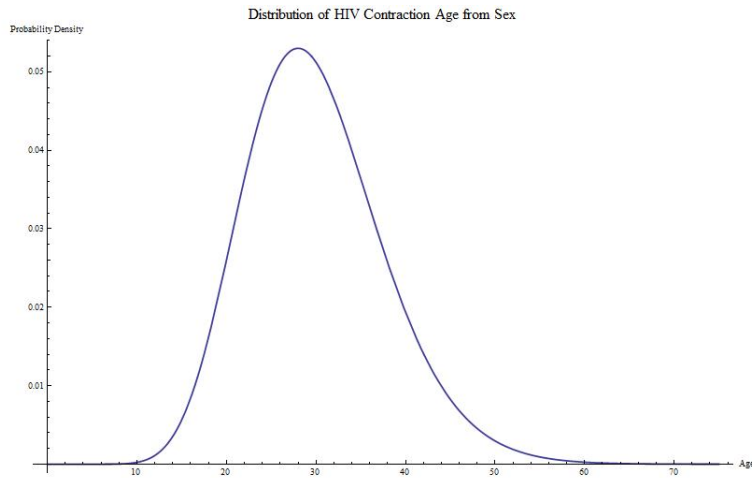
Example 1

Suppose we are trying to assess the property damage that results from the flooding of a river. It is reasonable to suppose that the damage will be worse closer to the river. In this case it is reasonable to expect that if damage is a function of distance from the river, it may be Chi-Squared distributed with the parameter $k=1$. The following graph illustrates how property damage may be distributed according to distance. Other scenarios possibly represented by this distribution include the value of a Boulder home as a function of distance to official "open space".



Example 2

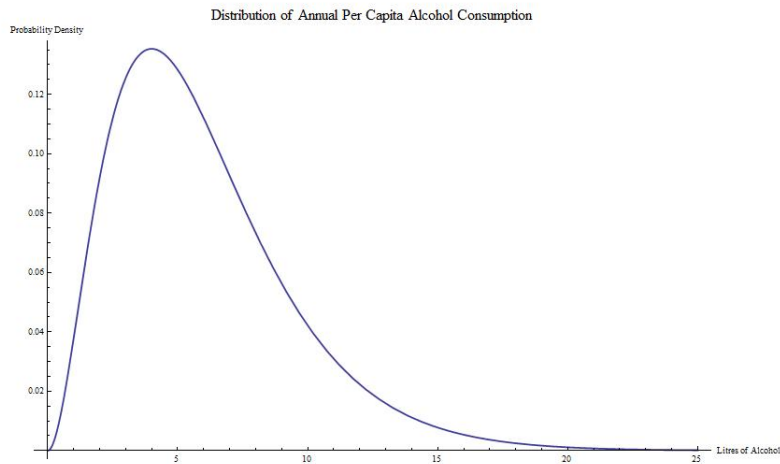
Now suppose that we are trying to find a distribution that represents the age that an individual contracts HIV from sexual activity. We go out and ask every person with HIV at what age they contracted the virus. With our data we then randomly draw samples from the uniform distribution. After enough draws we should see the mean tending to 30 years. Thus our distribution may be Chi-Squared distributed with $k = 30$. This seems reasonable since it is very unlikely that many children will have gotten HIV through sexual activity, and presumably the older someone gets the less promiscuous they are. Thus the risk of contracting HIV declines as an individual ages.



Example 3

Say we were interested in modelling the amount of alcohol (measured in litres) that the average American who does, in fact, drink consumes per year. Since we are excluding all Americans who do not drink (children, certain religious groups, etc.) it is reasonable to assume that the probability of one of these individuals consuming no alcohol per year is zero, or else they would be excluded from our population. We also know that as the litres consumed increases to infinity, the probability decreases, since alcohol in vast quantities has some peculiar side effects, most importantly, death. Perhaps we would want to use the Chi-Square distribution for this model. As long as k is not one or two, the probability at zero is zero, and the probability tends to zero as x approaches infinity. Since the average annual consumption of alcohol per capita is around 6 litres, but this population includes non-drinkers, so perhaps we decide to increase the number of litres by 1. So we believe that the mean of our population is 7 litres per

year, per person who drinks. The data generating process for this distribution is similar to the HIV data generating process. We simply draw individuals who drink from a uniform distribution and record the amount of alcohol they consume per year. After enough draws the distribution should tend towards the Chi-Squared. Since the mean of the Chi-Squared is equal to k , we know our parameter value. Plotting the PDF for $k = 6$ yields:



A few more examples

The PDF for one and three dimensional fields in a mode-stirred chamber is Chi-Squared distributed. "Statistical methods for a mode-stirred chamber", Kostas and Boverie, IEEE Transactions on Electromagnetic Compatibility, 2002.

In detecting a deterministic signal in white Gaussian noise by means of an energy measuring device, the decision statistic is chi-squared distributed when the signal is absent, and noncentral Chi-Squared distributed when the signal is present. "Energy detection of unknown deterministic signals", Urkowitz, Proceedings of the IEEE, 2005.

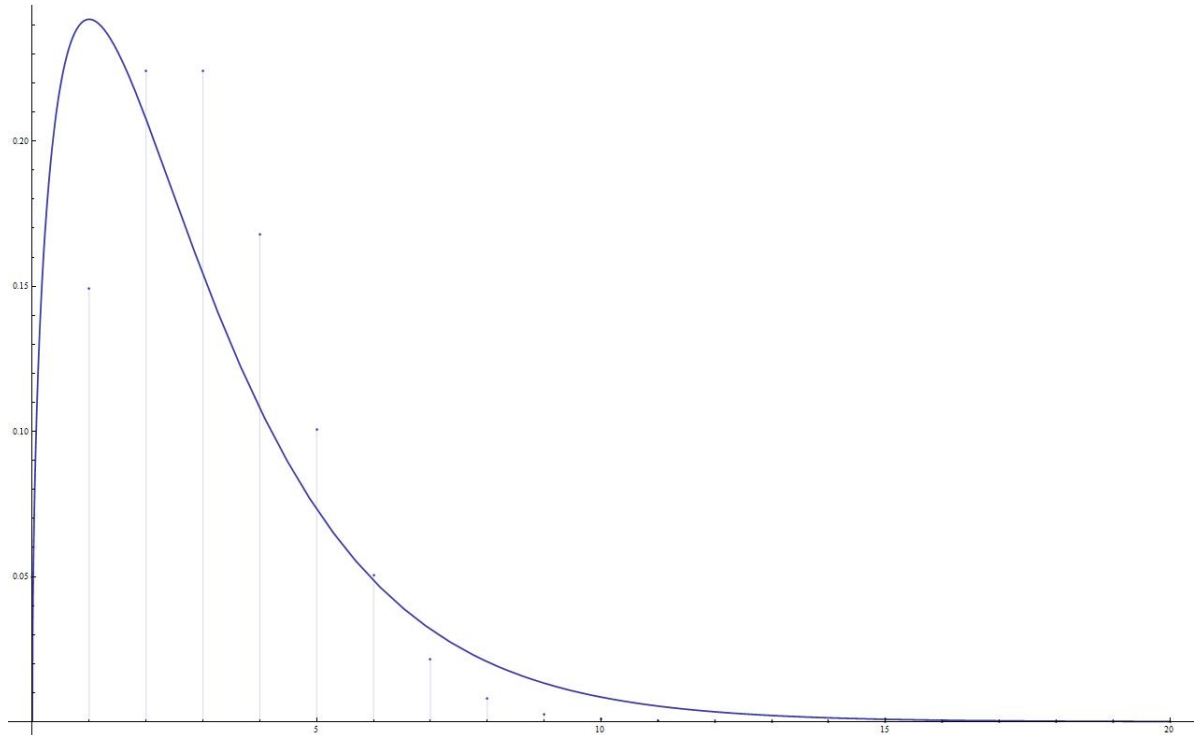
The constant elasticity of variance options pricing formula can be represented by the noncentral Chi-Squared distribution. "Computing the constant elasticity of variance options pricing formula", Schroeder, Journal of Finance, 1989.

The conditional distribution of the short rate in the Cox-Ingersoll-Ross process can be represented by the noncentral Chi-Squared distribution. "Evaluating the noncentral Chi-Squared distribution for the Cox-Ingersoll-Ross process",

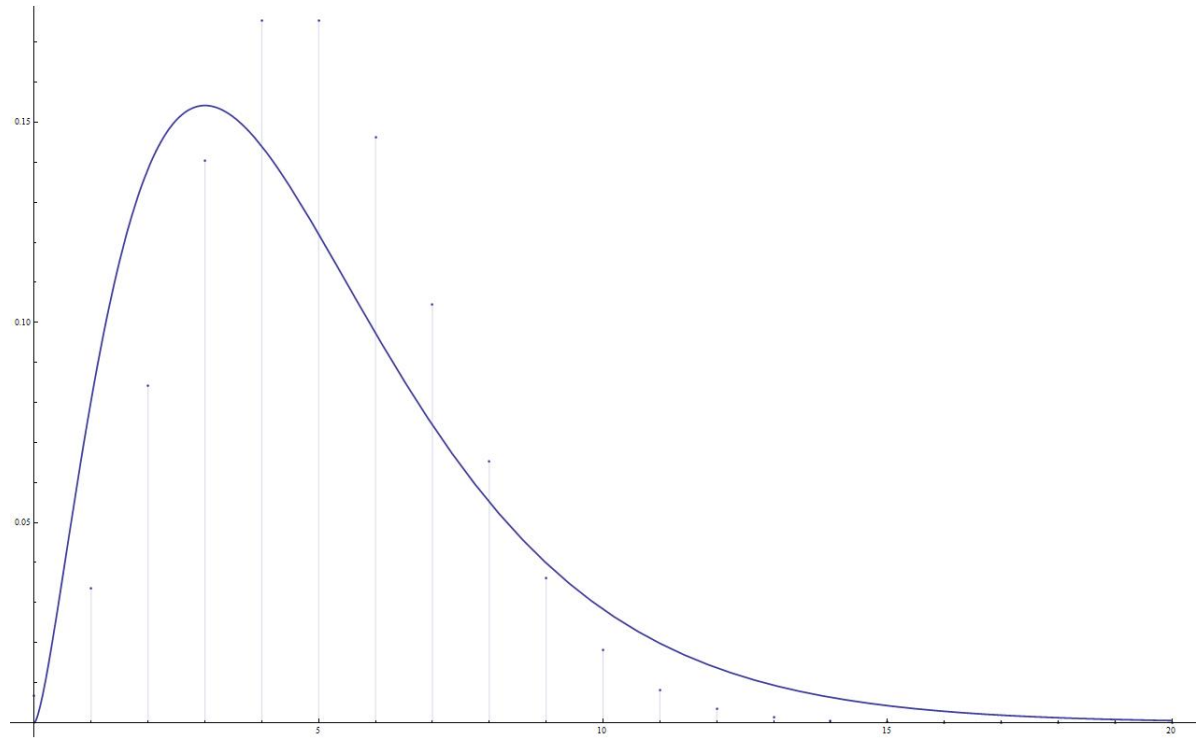
Dyrting, Computational Economics, 2004.

4 Other Distributions?

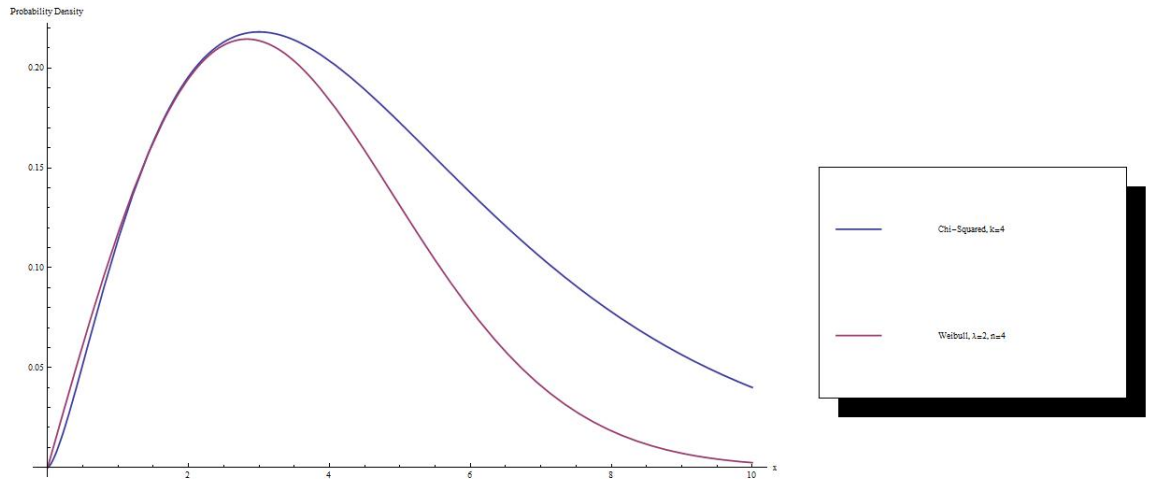
There are many distributions that are similar to the Chi-Squared. The Poisson distribution resembles a discrete version of the Chi-Squared. Like the Chi-Squared, the Poisson is a one parameter distribution and has the same shape as the Chi-Squared. The following example is a Poisson distribution with $k = 3$.



Another example where $k = 5$. Notice how the distribution takes the same shape as the Chi-Squared.



Another distribution that is similar to the Chi-Squared is the Weibull distribution, which is a two parameter continuous distribution. In the following example $\lambda = 2$ and $n = 2$.



5 Proof That Square of Normal is Chi-Squared

The following proof is of interest since it shows the direct relationship between the Normal distribution and the Chi-Squared distribution. If $X \sim N(0, 1)$, we will show that $X^2 \sim$ Chi-Squared distribution with $k = 1$.

To find the distribution of X^2 , we write the CDF of X^2 as:

$$\begin{aligned} F_{X^2}(a) &= P(x^2 \leq a) \\ &= \begin{cases} P(-\sqrt{a} \leq x \leq \sqrt{a}) & \text{if } a > 0; \\ 0 & \text{if } a \leq 0. \end{cases} \end{aligned}$$

for all $a \geq 0$,

$$F_{X^2}(a) = F_X(\sqrt{a}) - F_X(-\sqrt{a})$$

This result is derived in the sampling distribution notes. Differentiation with respect to a , we get the density function:

$$\begin{aligned} f_{X^2}(a) &= f_X(\sqrt{a}) \frac{a^{-1/2}}{2} - f_X(-\sqrt{a}) \frac{-a^{-1/2}}{2} \\ &= \frac{a^{-1/2}}{2} (f_X(\sqrt{a}) + f_X(-\sqrt{a})) \end{aligned}$$

$$\begin{aligned}
&= \frac{a^{-1/2}}{2} \frac{1}{\sqrt{2\pi}} e^{-a/2} \\
&= \frac{(\frac{1}{2})^{1/2} a^{-1/2} e^{-a/2}}{\sqrt{1/2}}
\end{aligned}$$

6 Chi-Squared Tests

A Chi-Squared test is any hypothesis test where the sampling distribution of the test statistic is Chi-Squared distributed when the null hypothesis is true. Perhaps the most well known example of a Chi-Squared test is Pearson's Chi-Squared test, which tests the null that the distribution of the sample is consistent with a specific theoretical distribution. Other examples of Chi-Squared tests include Ljung-Box, Box-Pierce, and other portmanteau tests which test the stationarity of time-series data.

References

- [1] Introduction to the Theory of Statistics - Mood, Graybill, Boes
- [2] Alcohol Consumption - <http://www.greenfacts.org/en/alcohol/figtableboxes/figure3.htm>
- [3] Chi-Squared, WolframMathworld - <http://mathworld.wolfram.com/Chi-SquaredDistribution.html>
- [4] Wikipedia - Chi-Squared Distribution, Chi-Squared Tests, Gamma Function, Karl Pearson, Ernst Abbe
- [5] The Chi-Squared Distribution, Harpreet Bedi, Chad Garnett, Wooyoung Park
- [6] A bit on sampling distributions, Edward Morey