

The F Distribution

Xin Geng
James Watson
Travis Weaver

Introduction

The PDF and CDF of the F distribution

$$f_{n,m}(x) = \frac{\left(\frac{n+m}{2}\right) n^{n/2} m^{m/2}}{\left(\frac{n}{2}\right) \left(\frac{m}{2}\right)} \frac{x^{\frac{n}{2}-1}}{(m+nx)^{\frac{(n+m)}{2}}} = \frac{m^{\frac{m}{2}} n^{\frac{n}{2}} x^{\frac{n}{2}-1}}{(m+nx)^{\frac{n+m}{2}} B\left(\frac{1}{2} n, \frac{1}{2} m\right)}$$

$$F_{n,m}(x) = I\left(\frac{nx}{m+nx}; \frac{1}{2} n, \frac{1}{2} m\right)$$

The Parameters

The F Distribution takes two parameters: m & n.

B, above, is not a parameter but the Beta Function:

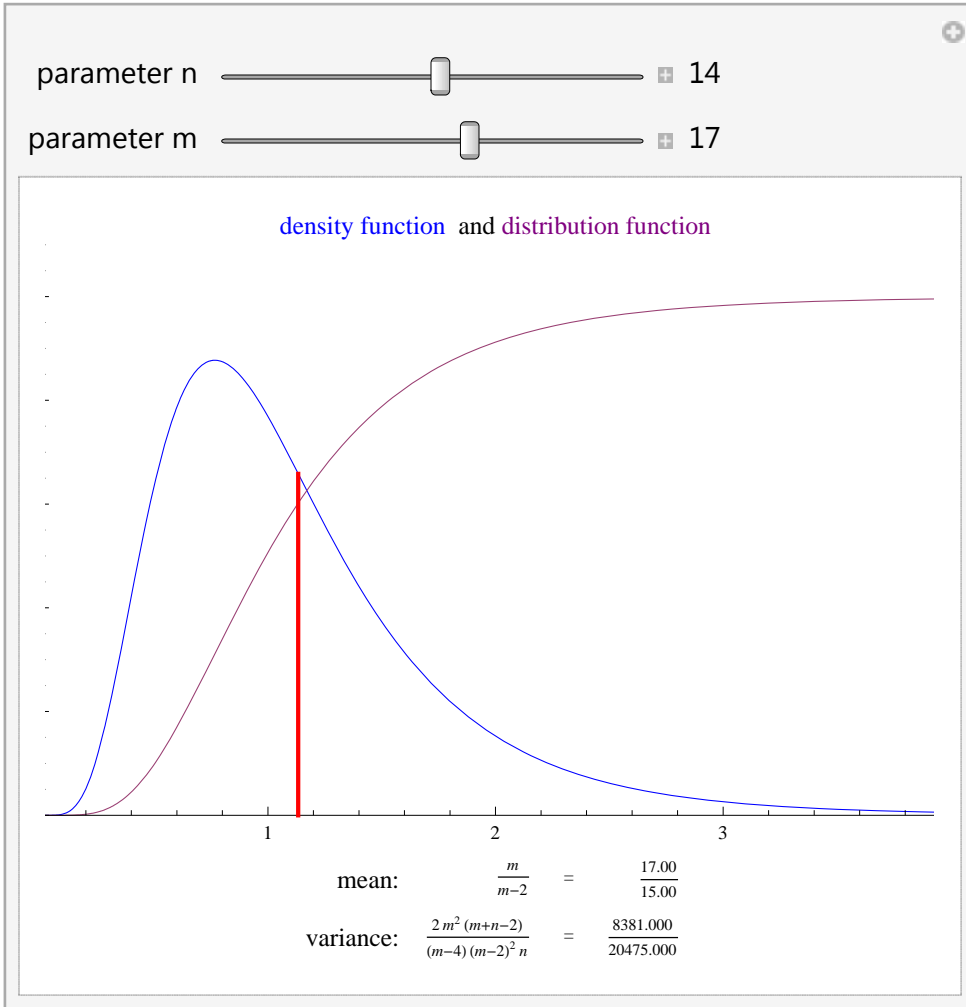
$$B(p, q) = \frac{\Gamma(p) \Gamma(q)}{\Gamma(p+q)} = \frac{(p-1)! (q-1)!}{(p+q-1)!}$$

where Γ is the Gamma Function

and I in the CDF is the Regularized Incomplete Beta function: $B(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt$

Properties of the F distribution

Shape: This depends on the two parameters, m and n. The curve is asymptotic to the x axis going towards infinity. Notice how the mean of the F distribution depends only on m.



Mean (first raw moment): The mean of the F distribution is always equal to:

$$E[x] = \frac{m}{m-2}$$

Note that the mean is only defined for values of $m > 2$.

Variance (second central moment): The variance for this distribution is equal to:

$$v a r[x] = \frac{2 m^2(m+n-2)}{n (m-2)^2 (m-4)}$$

According to this formula, the variance is not defined for 1, 2, 3, and 4.

Skewness (third central moment): The skewness is equal to:

$$\frac{2 (m+2 n-2)}{m-6} \sqrt{\frac{2 (m-4)}{n(m+n-2)}}$$

Kurtosis (fourth central moment): The kurtosis is equal to:

$$\frac{12 (-16+20 m-8 m^2+m^3+44 n-32 m n+5 m^2 n-22 n^2+5 m n^2)}{n(m-6) (m-8) (n+m-2)}$$

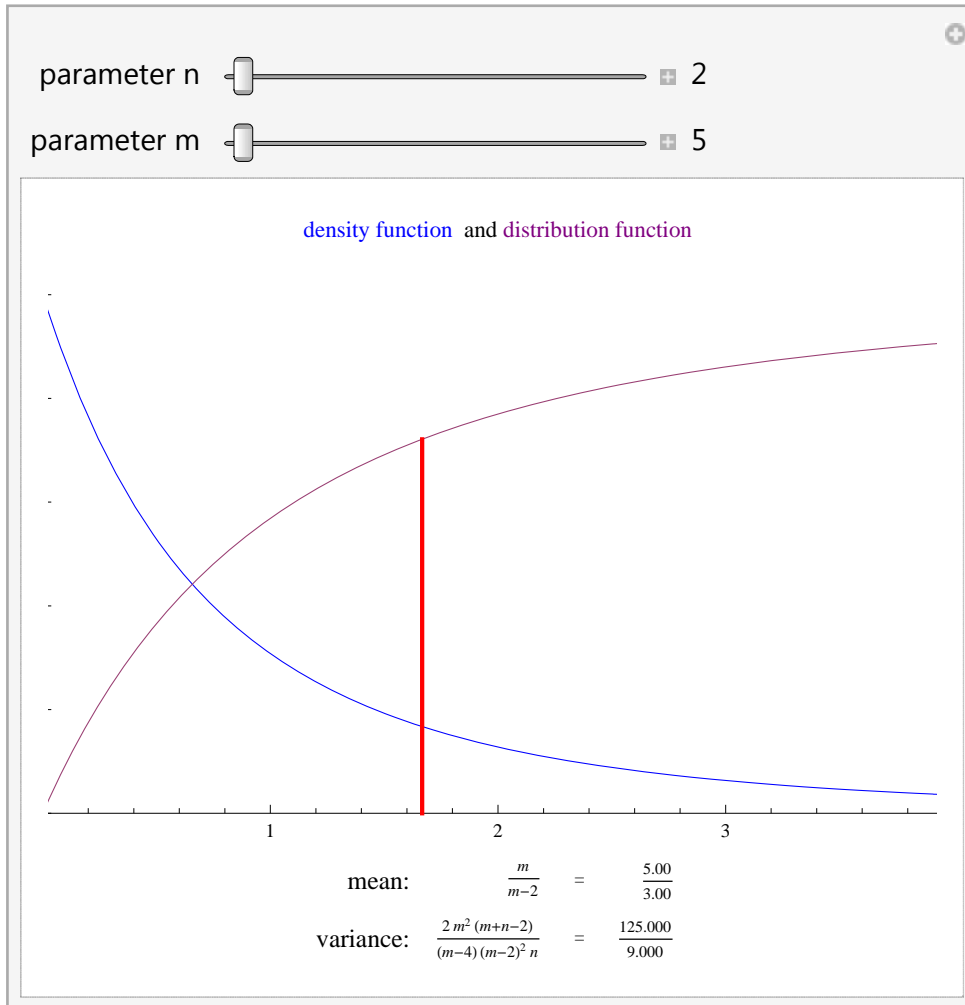
An Experiment that Generates an F Distribution



Example

Suppose we have just opened a new pub. We want to model the numbers of days that will pass before we get our first paying customer (we will give away drinks for free to our friends, but this is not a very successful business model). It may be reasonable to approximate the density function with an F distribution with parameter $n = 2$ and $m = 5$

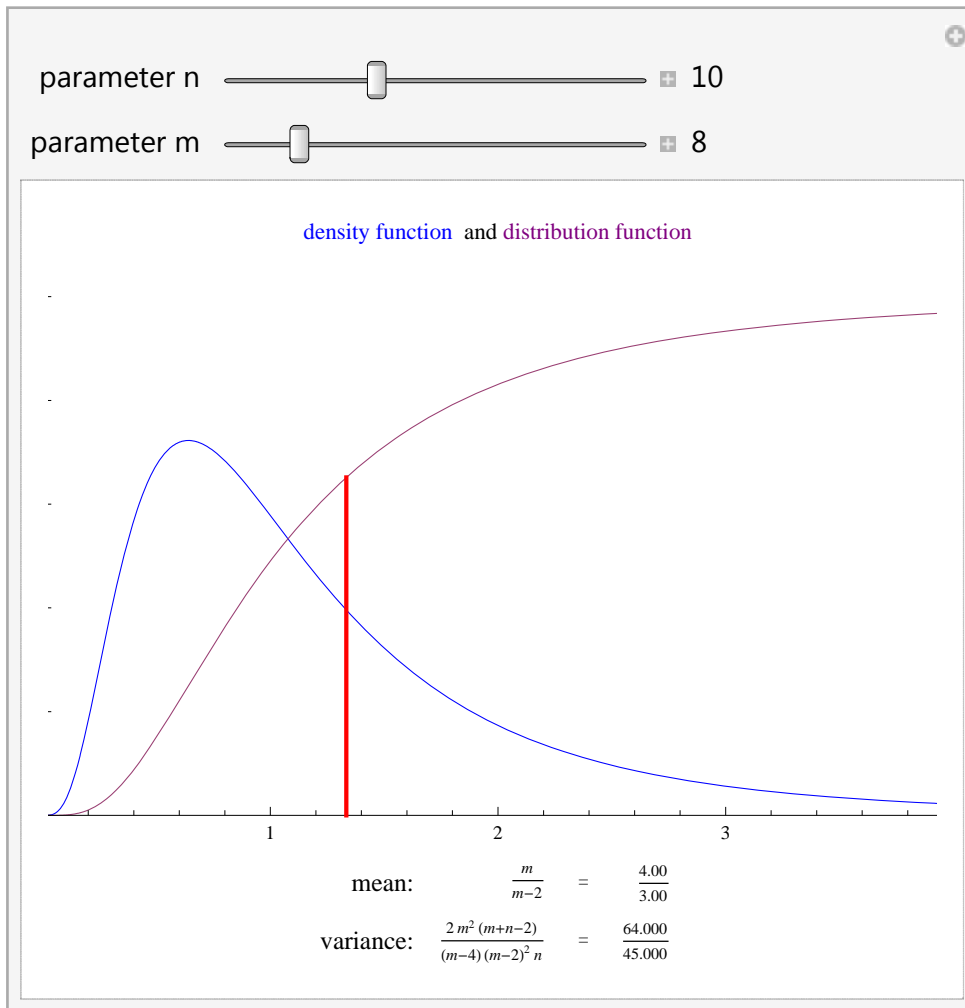
The distribution would look like this:



We might pick this model if we think it is very likely that our first paying customer will come sometime during day one. However, if we have not had a paying customer by day four, it might be reasonable to conclude that we won't be getting customers anytime soon. But we are persistent and intend to keep the pub open for eternity; which is convenient, since the F distribution converges to zero going towards infinity.

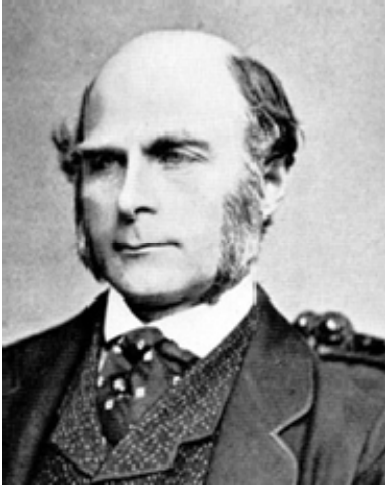


Assume you owned a department store with, for security purposes, a revolving door entrance. It is black Friday. Your security detail wants to know how fast customers will be storming through the revolving door to enter the store. Perhaps, you could use the F distribution, with parameter $n = 10$ and $m = 8$ to predict the number of seconds in between customers entering through the revolving door. Since there is a minimum amount of time that must elapse for another person to enter the revolving door, the distribution might look something like this:



Of course, since the F distribution still converges to 0 as the amount of time goes towards infinity, there would still be a infinitesimally small probability that a billion years would pass without someone entering the store.

Origins



Sir Francis Galton



Karl Pearson

Sir Francis Galton (1822-1911) originally conceived the statistical concept of correlation (or dependence). The most familiar measure of dependence is Karl Pearson's (1857-1936) product-moment correlation coefficient:

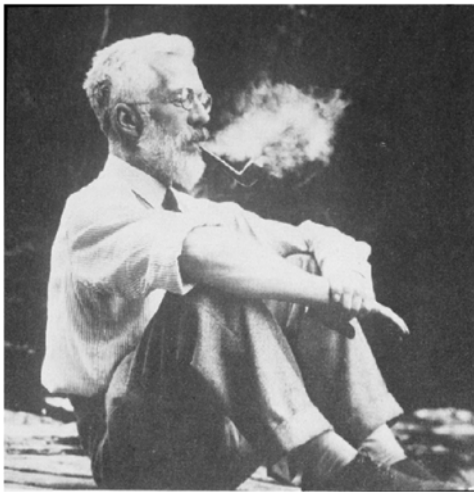
$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \bar{X})(Y - \bar{Y})]}{\sigma_X \sigma_Y}$$

Denoted as ρ , the linear population correlation between two random variables X and Y is defined as the covariance divided by the product of the standard deviations. If we have a series of n observations of X and Y then the sample correlation coefficient is defined as:



$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

However, for high levels of positive correlation, the distribution of Pearson's r is negatively skewed because it cannot take a value greater than 1. **Because it is not normally distributed,** it is not possible to calculate the probability that r takes a particular value simply by knowing the mean and standard error of r .



R.A. Fisher



George W. Snedecor

In order to remedy this difficulty, R.A. Fisher suggested the z -transformation in 1915:

$$z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$$

This transformation is approximately normal for sample sizes of 10 or more from a bivariate normal distribution. The z distribution has a mean and variance of:

$$\frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right) + \frac{\rho}{2(n-1)} \quad \text{and} \quad \frac{1}{n-3}$$

The F distribution was created by George W Snedecor (1881-1974) in his paper *Calculation and Interpretation of Analysis of Variance and Covariance* (1934). This distribution was meant to improve upon the presentation of Fisher's analysis of variance, which uses the natural log. It is interesting to note Fisher's reaction to the F distribution, which was named in homage to Fisher, in some comments contained within a letter written to another statistician, H.W. Heckstall-Smith:



“I have always regarded the F-test and the z-test as the same, and indeed in 1924 I calculated the variance-ratio as a means of getting their natural logarithms. ¹ Of course if everyone had a computing machine at hand on their desk the variance ratios are the quicker, but those without computing machines, and who can use four-figure tables, are about five times more numerous.”

¹Heckstall-Smith had suggested that in an article he was writing for a medical journal he would need to use the F-test rather than the z-test because natural logarithms would not be tolerated. Also, we didn't mean to include anything about the F-test, it was just in the quote.

“I think it was only an afterthought that led Snedecor to say that the capital F he had used was intended as a compliment to myself.”

While we cannot comment on Snedecor's intentions in the naming of the F distribution, it is fortunate for us all that computing machines have become somewhat more widely adopted and user-friendly since the 1930's.

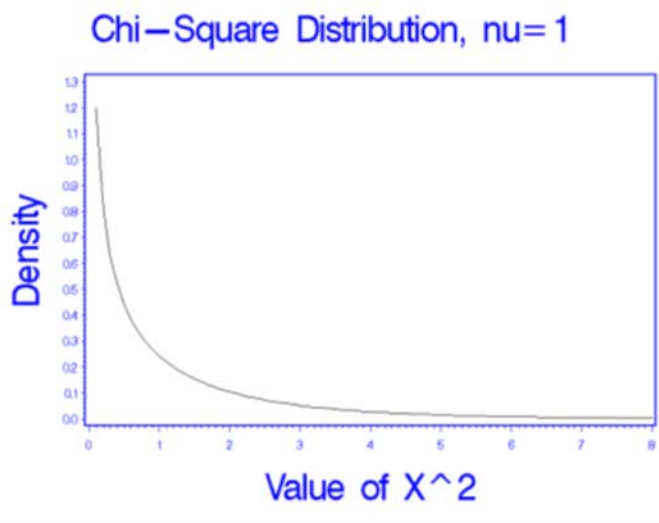
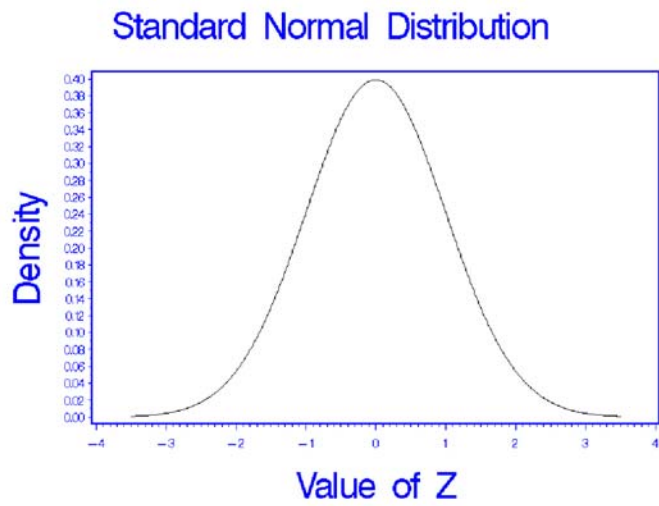
Appendix (Chi-squared Distribution)



Following the example used in the above, we define the Chi-squared distribution:

$$X_{1H}^2 = z_H^2$$

So what would the sampling distribution of X_{1H}^2 look like?



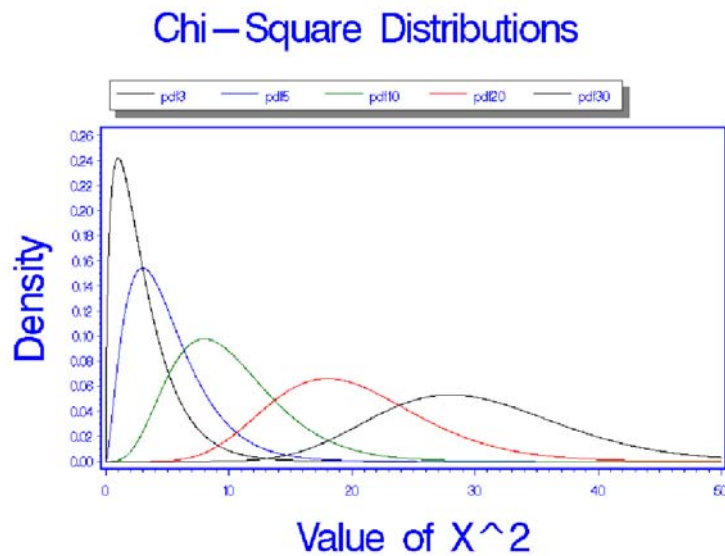
Similarly, if we repeatedly draw independent samples of $n=2$ from $N(\mu_H, \sigma_H^2)$

Compute $z_{1H}^2 = \frac{(H_1 - \mu_H)^2}{\sigma_H^2}$ and $z_{2H}^2 = \frac{(H_2 - \mu_H)^2}{\sigma_H^2}$

Then compute the sum: $X_{2H}^2 = z_{1H}^2 + z_{2H}^2$

Generalize:

- Chi-square is the distribution of a sum of squared deviation taken from unit normal. For n independent observations from a $N(\mu_H, \sigma_H^2)$, the sum of the squared standard heights has a Chi-square distribution with n degrees of freedom.
- Chi-squared distribution only depends on degrees of freedom, which in turn depends on sample size n .
- The standard heights are computed using population μ_H and σ_H^2 ; however, we usually don't know what μ_H and σ_H^2 equal to. When μ_H and σ_H^2 are estimated from the sampled data, the degrees of freedom are less than n .

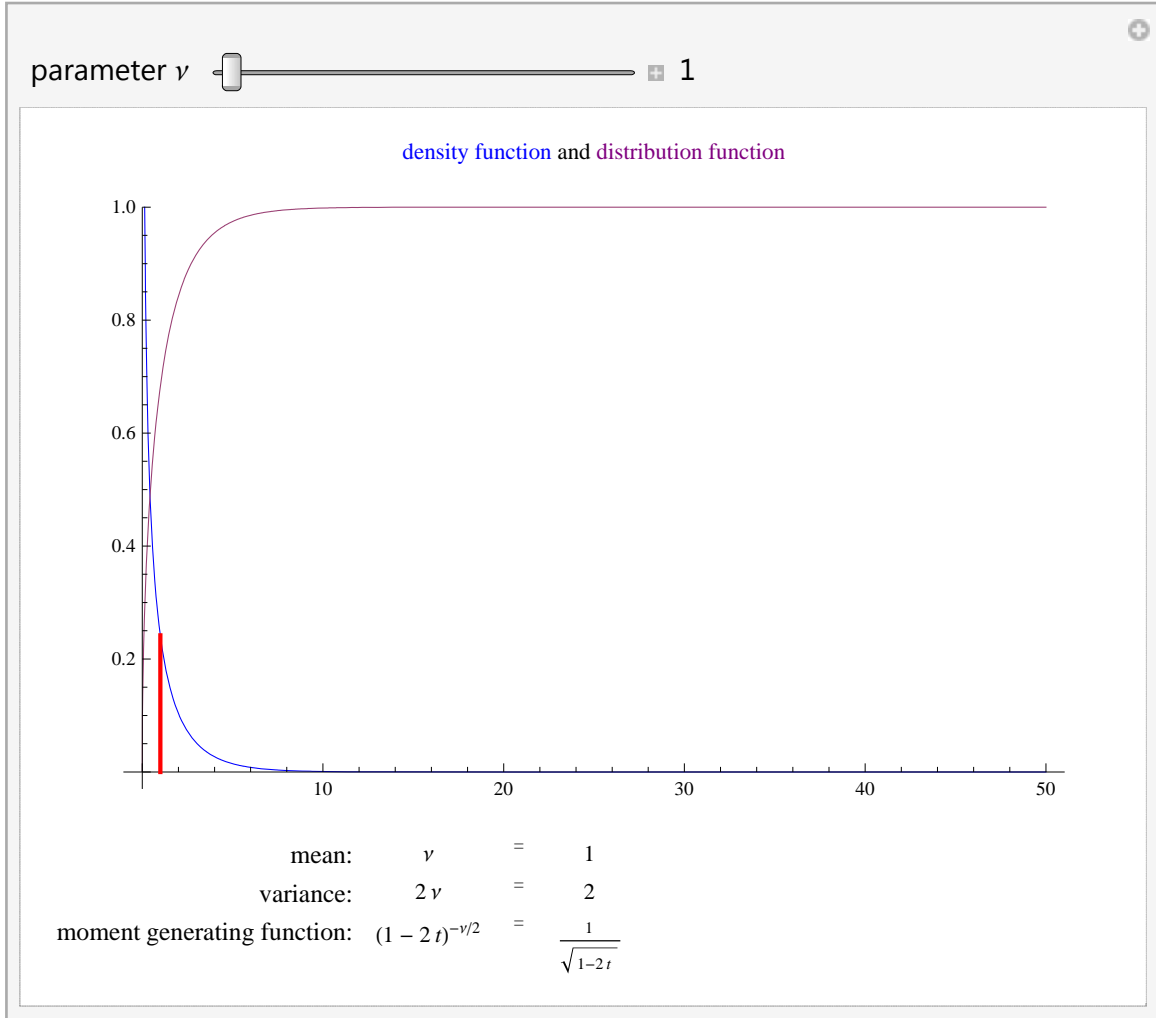


Properties of Family of Chi-squared Distributions

- $E(X_n^2) = \text{mean} = n = \text{degrees of freedom}$
- $\text{Var}(X_n^2) = 2n$
- **Mode of X_n^2 is at value $(n - 2)$ (for $n \geq 2$)**
- **Median is approximately $= \frac{3n-2}{3}$ (for $n \geq 2$)**
- $X_{n_1+n_2}^2 = X_{n_1}^2 + X_{n_2}^2$

(the sum has a chi – squared distribution with $n_1 + n_2$ degrees of freedom)

χ^2 Distribution



References:

Selected correspondence of R.A. Fisher

http://digital.library.adelaide.edu.au/coll/specialfisher/stat_inf/index.html

Sampling Distribution of Pearson's r

<http://davidmlane.com/hyperstat/A98696.html>

Wikipedia: F Distribution

<http://en.wikipedia.org/wiki/F-distribution>

Earliest Known Uses of Some of the Words of Mathematics

<http://jeff560.tripod.com/f.html>

Fisher's z transformation

<http://www.answers.com/topic/fisher-s-z-transformation-1>

Chi-Square & F Distributions, Carolyn J. Anderson

http://www.ed.uiuc.edu/courses/EdPsy496/lectures/7ChiSq_Fdist_05_online.pdf

Chi-Square & F Distributions

<http://www.docin.com/p-61359261.html>

Wolfram Demonstrations Project

<http://www.demonstrations.wolfram.com>