A Brief Overview of Probability and Statistics

David Ubilava

August, 2017

Econometric methods are comprised of (and deal with) random variables. A characteristic feature of a *random variable* (**r.v.**) is that it takes on a set of possible numerical values determined probabilistically. Because **r.v.** is a collection of possibilities, it is not observed as such. Instead, a realization of a **r.v.** is observed. Conventionally, uppercase letters of Latin alphabet are reserved to denote random variables, and lowercase letters – their realizations.

Densities and Distributions

Probability Density Function

The *probability density function* (pdf) of a random variable X, denoted by $f(x)$, summarizes the information concerning the possible outcomes of X and the corresponding probabilities.

Random variables can be *discrete* or *continuous*. A discrete **r.v.** takes on a finite (or perhaps a countably infinite) number of values. A continuous **r.v.** takes on an infinite number of values, with zero probability of any individual value.

In the case of discrete random variables:

$$f(x) \equiv p_j = P(X = x_j), \quad j = 1, \dots, k$$

with $p_j = 0$, $\forall x \neq x_j$ for some j.

Thus, for any real number x, p_j is the probability that **r.v.** takes on the particular value x_j.

Properties of discrete **r.v.** density function are defined by:

$$p_j \geq 0, \quad j = 1, \dots, k$$

$$\sum_{j=1}^{k} p_j = 1$$

For the continuous **r.v.**, an expression equivalent to () is given by:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Cumulative Distribution Function

The *cumulative distribution function* (**cdf**) of a continuous random variable X, is an integral of the density function:

$$F(c) = P(X \leq c) = \int_{-\infty}^{c} f(x)dx$$

The equivalent **cdf** of a discrete random variable is the sum of probabilities over all values of x_j such that $x_j \leq c$.

Multivariate Distributions

We are typically interested in the joint variation of two or more variables. Let X and Y be discrete random variables. Then, (X, Y) have a *joint distribution* that is fully described by the joint **pdf** of (X, Y):

$$f_{X,Y}(x, y) = P(X = x, Y = y)$$

If X and Y are *independent random variables*, then the joint **pdf** of (X, Y) is a product of their respective **pdf**s:

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

In general, a set of random variables $\{X_1, X_2, \ldots, X_n\}$ are independent random variables if and only if their joint **pdf** is the product of the individual **pdf**s for any $(x_1, x_2, \ldots, x_n)$.

Often we would like to know how one random variable, Y, is related to another random variable, X. The most we can know about how X affects Y is contained in the *conditional distribution* of Y given X. The adequate conditional pdf is defined by:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

If X and Y are independent random variables, then:

$$f_{Y|X}(y|x) = f_Y(y)$$

Expectations

Mean

The expected value, or the mean of a random variable X, $\mathbb{E}(X) \equiv \mu_X$, is a weighted average of all possible values of X, where weights are determined by the pdf.

For a discrete random variable,

$$\mu_X = \sum_{j=1}^{n} x_j p_j$$

For a continuous random variable,

$$\mu_X = \int_{-\infty}^{\infty} x f(x) dx$$

Conditional Mean

In economics we often are required to explain one variable, say Y, in terms of another variable, say X. A single number will no longer suffice, since the distribution of Y given $X = x$ generally depends on the value of x. Nevertheless, we can summarize the relationship between Y and X by looking at conditional expectation, or conditional mean of Y given X. For a discrete random variable:

$$\mathbb{E}(Y|X = x) = \sum_{j=1}^{m} y_j f_{Y|X}(y_j|x)$$

For a continuous random variable:

$$\mathbb{E}(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dx$$

Variance

The variance of a random variable X, $Var(X) \equiv \sigma_X^2$, is an expected value of the squared deviations from the mean:

$$\sigma_X^2 = \mathbb{E}[(X - \mu_X)^2] \equiv \mathbb{E}(X^2) - \mu^2$$

For a discrete random variable:

$$\sigma_X^2 = \sum_j^n x_j^2 p_j - \mu_X^2$$

For a continuous random variable:

$$\sigma_X^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu_X^2$$

Standard Deviation

The standard deviation of a random variable, $SD(X) \equiv \sigma_X$ is a positive square-root of its variance:

$$\sigma_X = +\sqrt{\sigma_X^2}$$

Covariance

The covariance measure between two random variables X and Y is given by:

$$Cov(X, Y) \equiv \sigma_{XY} = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}(XY) - \mu_X \mu_Y$$

Note, that if either $\mu_X = 0$ or $\mu_Y = 0$, then $Cov(X, Y) = \mathbb{E}(XY)$.

Correlation

The correlation between two random variables, X and Y, is given by:

$$Corr(X, Y) \equiv \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Some Useful Distributions

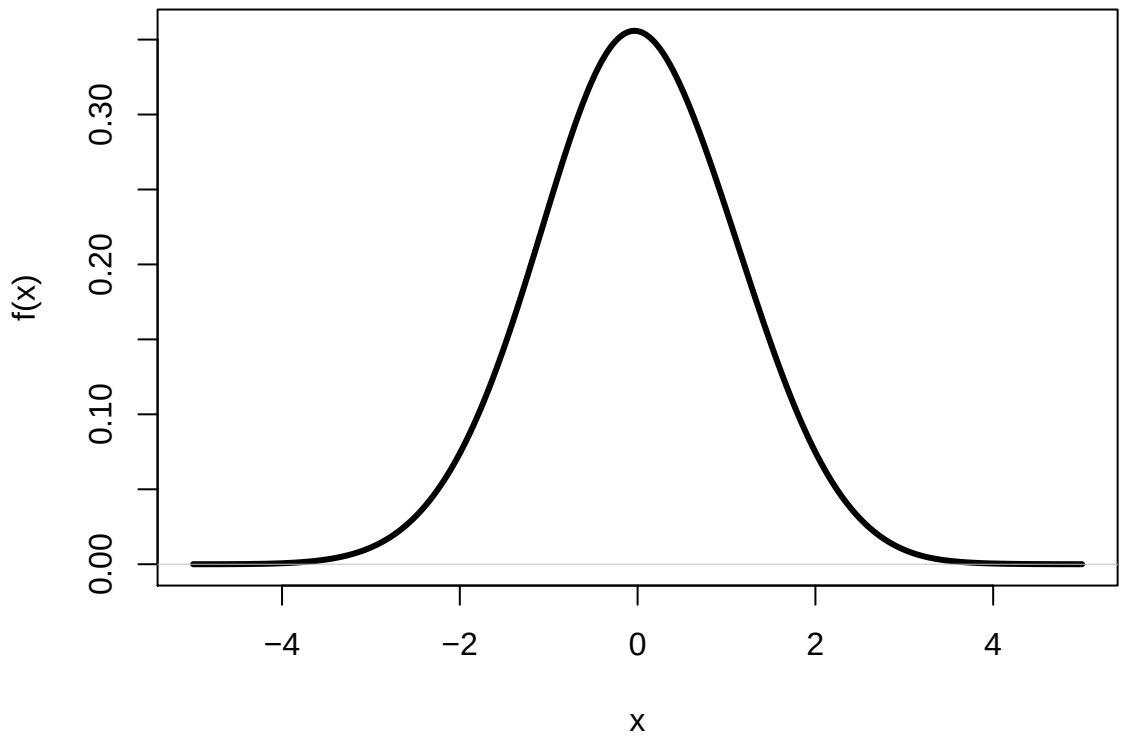
The Normal Distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \qquad -\infty < x < \infty$$

The Standard Normal Distribution:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right], \qquad -\infty < z < \infty$$

Any normal random variable can be transformed to a standard normal random variable. That is, if $X \sim N(\mu, \sigma^2)$, then $(X - \mu)/\sigma \equiv Z \sim N(0, 1)$.



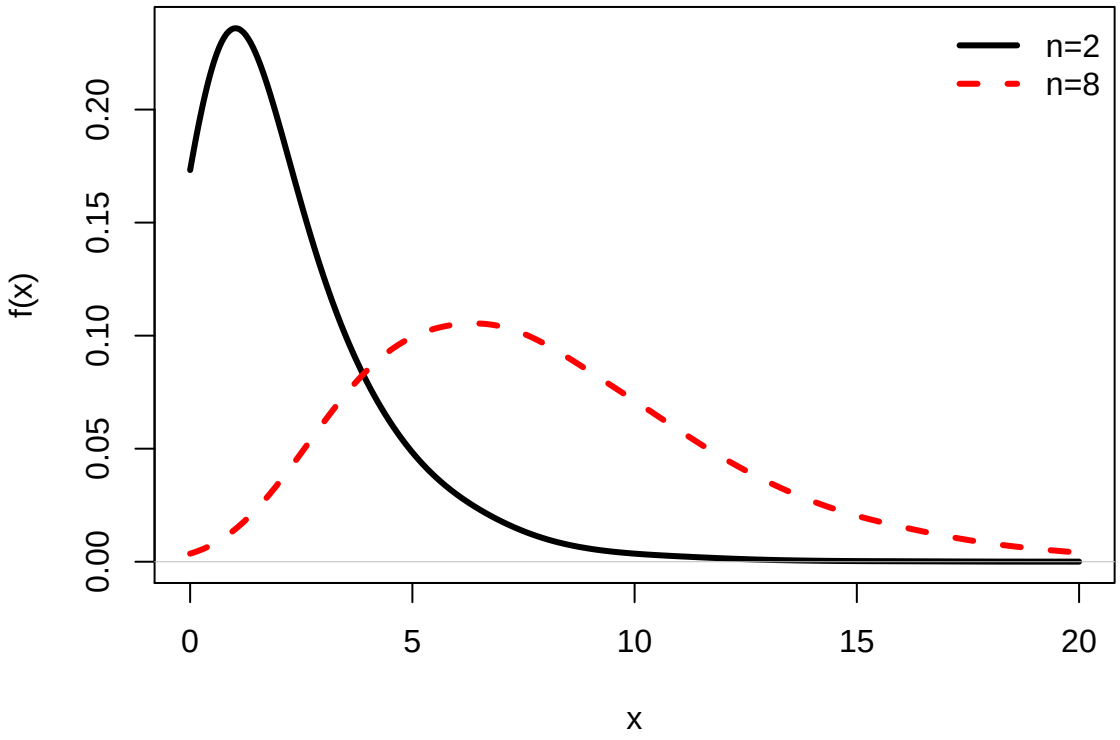
The Chi-Square Distribution:

Let $\{Z_1, \ldots, Z_n\}$ be a sequence of independent standard normal random variables. Define a new variable:

$$X = \sum_{i=1}^{n} Z_i^2.$$

Then, $X \sim \chi_n^2$.

A chi-square random variable is always nonnegative, and its distribution is not symmetric about any point.

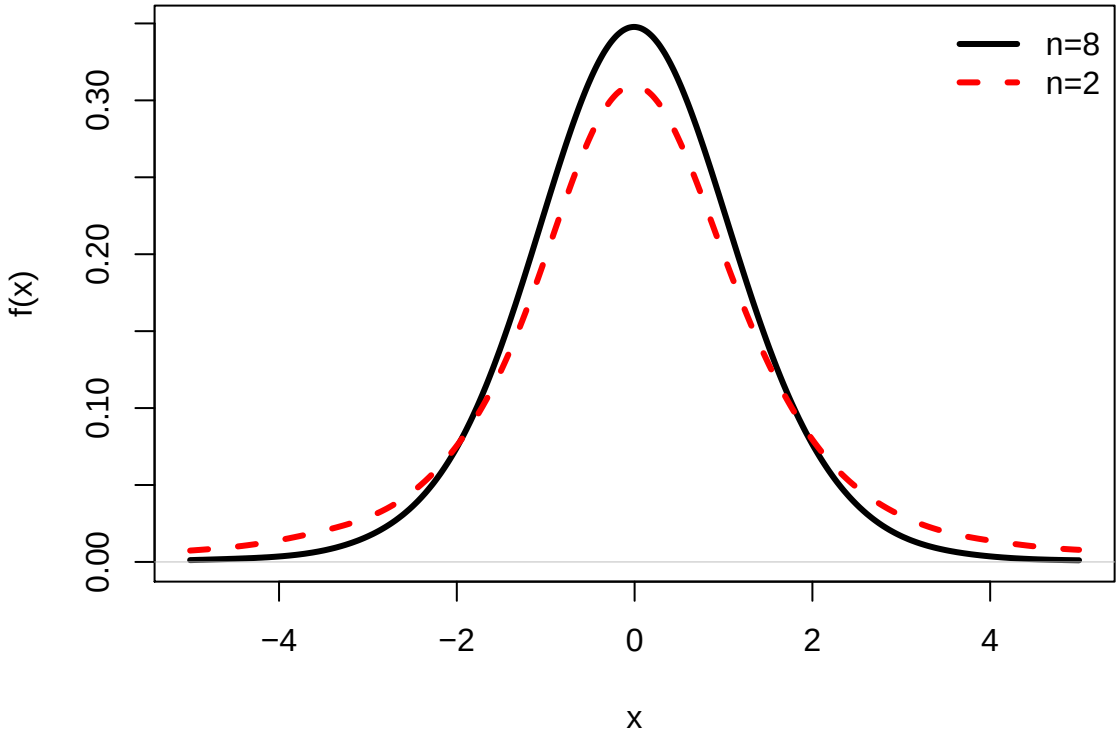


The t Distribution:

Let $Z \sim N(0,1)$ and $X \sim \chi_n$, and assume that Z and X are independent. Define a new variable:

$$T = \frac{Z}{\sqrt{X/n}}.$$

Then, $T \sim t_n$. The **pdf** of the t distribution has a shape similar to that of the standard normal distribution, except it is more spread out. The expected value of a t distributed random variable is zero, and the variance is $n/(n-2)$ for $n > 2$.



The F Distribution:

Let $X_1 \sim \chi_{n_1}$ and $X_2 \sim \chi_{n_2}$, and assume that X_1 and X_2 are independent. Define a new variable:

$$F = \frac{X_1/n_1}{X_2/n_2}.$$

Then, $F \sim F_{n_1,n_2}$.