Econometrics 6027
Lecture 1
Basic Probability Concepts and Asymptotic Results

Instructor: Marcel Voia

Reference: Wooldridge (2002) Chapter 3, Greene Chapter 3, Chapter 4

1 Estimation and Testing

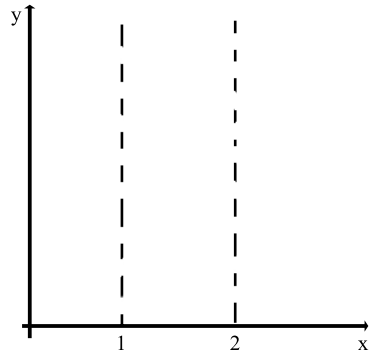
In this class, we'll be modelling individual behaviour. We could have:
Y_t for time series - discrete data ie. number of children
Y_i for cross section - continuous data ie. experience or marks
Y_{it} for panel - such as duration outcomes like times between transactions

Say we want to estimate this:

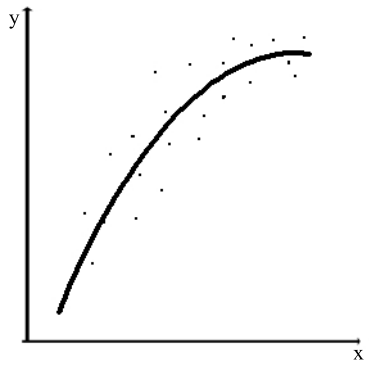
$$Y_i = \beta X_i + u_i$$

We need assumptions about the errors to estimate β. For example, assume:
$E(u_i|X) = 0$ no correlation between the error and X (no multicorrelation)
$var(u) = \sigma^2$ (no heteroskedasticity)
$E(u) = 0$ but this is implied by the first assumption.
u is iid (independent and identically distributed)

Can we estimate if we only have discrete variables? We can't fit a line if the data is discrete



We want to fit the population $E(Y_i|X) = \beta X_i$ which is represented by the line below



We can estimate in three ways: through OLS, through maximum likelihood methods, and method of moments.

1.1 OLS

Minimize with respect to β, the parameter of the model, the sum of squares:

$$\min_{\beta} \sum u_i^2$$

Would analyze using the objective function, that is,

$$\sum (Y_i - X_i\beta)^2$$

1.2 Maximum Likelihood Estimation

If u is iid $N(0, \sigma^2)$ then how do we estimate? We can use a maximum likelihood estimator. But we don't need the normal assumption: only use it if it holds. If it's normal, with a likelihood estimate, what will be the density in errors? We know from Masters' level courses that it is:

$$f(u) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-u_i^2}{2\sigma^2}}$$

So how do you get the likelihood? Estimate the joint likelihood, which is the product of individual densities.

$$L_T = \Pi f(u_i) = \big(\frac{1}{\sqrt{2\pi\sigma^2}}\big)^N e^{\frac{-u_i^2}{2\sigma^2}}$$

We want to maximize the likelihood or log likelihood, will have the same result

$$lnL_T = ln\Pi f(u_i) = \sum_i lnf(u_i)$$

If you put assumptions on the distribution, you get the variance. The restriction gets you a parameter you can estimate. There are many types of likelihood estimators, and we will explore them.

1.3 Method of Moments

There is another method based on moments (mean, standard deviation, skewness, kurtosis). Other moments can be created based on Euler Equations and other methods. More details about this approach will be provided in the lecture about this topic.

1.4 Hypothesis Testing

The difference between normal and t-distribution is that the t has higher kurtosis. With more observations, the t-distribution converges to the normal distribution. If we have a different sample size, we have a different t. Can we use this to estimate? No - because it is sample size dependent. But we can use it for testing, for example for constructing confidence intervals. The t-distribution is constructed as a ratio. Divide the difference by the standard deviation of your estimator.

$$\hat{\beta} \sim N(\beta, var(\hat{\beta}))$$

If we apply the Central Limit Theorem, we should have a distribution $E(\hat{\beta}) = \beta$ with $var(\hat{\beta}) = \frac{\sigma^2}{\sum x_i^2}$.

How can the variance be computed?

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'Y \\
var(\hat{\beta}) &= var((X'X)^{-1}X'(X'\beta + u)) \\
&= var((X'X)^{-1}X'u) \\
&= (X'X)^{-1}X'var(u)X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}(X'X)(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}
\end{aligned}$$

$H_0 : \beta = \beta_0$ two-tailed test $H_0 : \beta >= \beta_0$ left side cut-off $H_0 : \beta <= \beta_0$ right side cut-off

If we know the σ^2, we can use the Z-test. What changes in different tests is the critical value, but you use the same test. The normalized distribution is:

$$Z = \frac{\hat{\beta} - \beta_0}{\sqrt{var(\hat{\beta})}}$$

This test does not change with a parameter, that is, it is standardized.
$E(Z) = 0$
$var(Z) = 1$

The first two-tailed test enables us to reject the hypothesis if test values are at extremes. If the population variance is unknown, use the t-test. The variance of the parameter will be chi-squared.

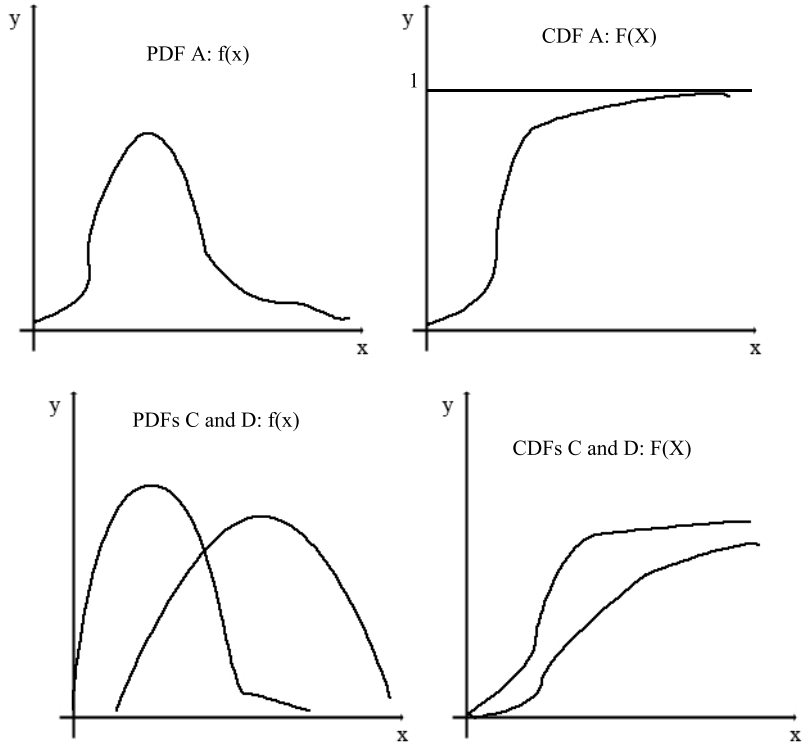
$$t = \frac{\hat{\beta} - \beta_0}{\sqrt{var(\hat{\beta})}}$$

This is non-normal because the variance is chi-squared. So we have a t-distribution that will converge to the normal. Be wary of using it for estimation because variance is a function of a sample size. The parameter is a function of the sample size too, it is not fixed.

2 Basic distributions

f(x) is the probability density function (PDF) and F(x) is the cumulative density function (CDF). All densities must have a positive support, above the zero line: the value for Y is positive. $\int_x f(x)dx = 1$ whatever the support of x, whether it is positive x, or the whole +/- domain of x.

The difference between a CDF and PDF can be seen visually as follows:



2.1 The Chi Squared Distribution

If $z \tilde{} N(0,1)$ and $x = z^2$ then x is said to have a chi-square distribution with one degree of freedom, denoted

$$z^2 \sim \chi^2_{(1)}$$

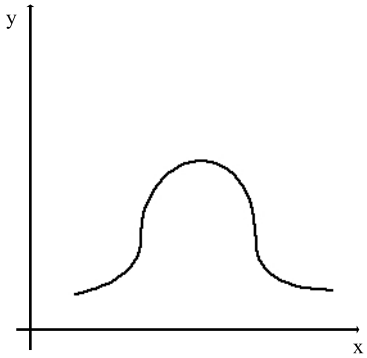
The mean of a chi square is 1 and the variance is 2. More generally,

If $x_1, x_2, ...x_n$ are n independent chi-squared [1] variables, then

$$p = \sum_{i=1}^{n} x_i \sim \chi^2_{(n)}$$

The mean of p is n and the variance is 2n.

2.2 Student's t Distribution

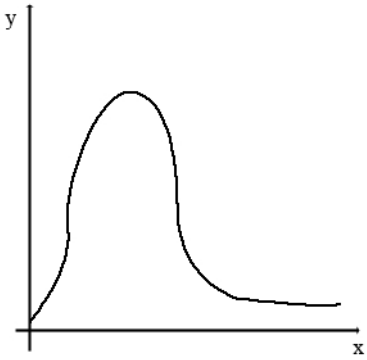


If z is an $N(0,1)$ variable and x is a $\chi^2_{(n)}$ and is independent of z, then the ratio

$$t_{(n)} = \frac{z}{\sqrt{x/n}}$$

has a t or student's t distribution with n degrees of freedom. The t distribution has the same shape as the normal distribution but has thicker tails.

2.3 The F Distribution



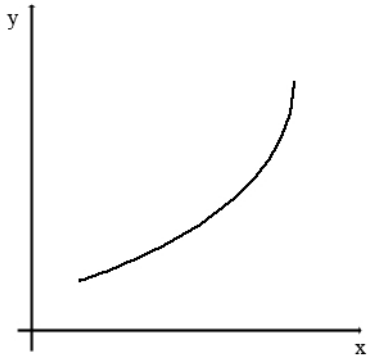
If x_1 and x_2 are two independent chi-square variables with degrees of freedom n_1 and n_2 respectively, then the ratio

$$F = \frac{x_1/n_1}{x_2/n_2} \sim F_{(n_1,n_2)}$$

has an F distribution.

It's a skewed distribution, and the degree of skewness depends on the sample size. We use the F distribution for testing. It is sample-dependent.

2.4 The Gamma and Exponential Distribution



Gamma Distribution is used on studies of income distribution and production function

$$\begin{aligned} f(x) &= \frac{\lambda^P}{\Gamma(P)} e^{-\lambda x} x^{P-1},\ x \geq 0,\ \lambda > 0,\ P > 0 \\ E(x) &= \frac{P}{\lambda},\ \mathrm{Var}(x) = \frac{P}{\lambda^2}. \end{aligned}$$

When $P = 1 \Rightarrow$ Exponential Distribution
When $\lambda = \frac{1}{2}$, $P = \frac{n}{2} \Rightarrow \chi_{(n)}$

2.5 The Beta Distribution

Used for variables constrained between 0 and $c > 0$,

$$\begin{aligned} f(x) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{x}{c}\right)^{\alpha-1} \left(1 - \frac{x}{c}\right)^{\beta-1} \left(\frac{1}{c}\right), \\ E(x) &= \frac{c\alpha}{(\alpha+\beta)} \\ Var(x) &= \frac{c^2\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2} \end{aligned}$$

very flexible in shape, if $\alpha = \beta \Rightarrow f(x)$ is symmetric

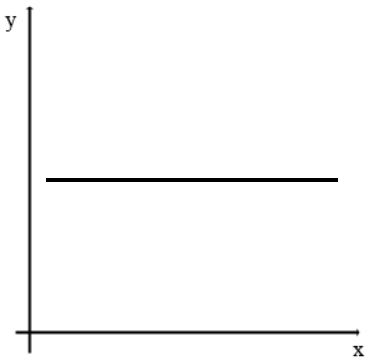
It is used in studies of labour force participation rates.

2.6 Logistic Distribution

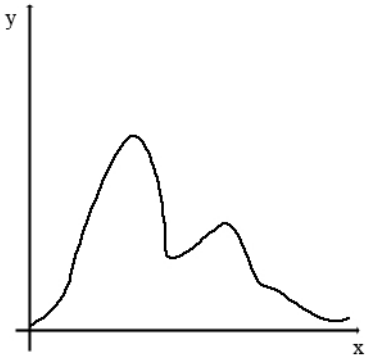
It is a thicker-tailed symmetric distribution, with the following CDF

$$\begin{aligned} F(x) &= \Lambda(x) = \frac{1}{1+e^{-x}}, \text{ and the density} \\ f(x) &= \Lambda(x)(1-\Lambda(x)) \\ E(x) &= 0 \\ Var(x) &= \frac{\pi^2}{3}. \end{aligned}$$

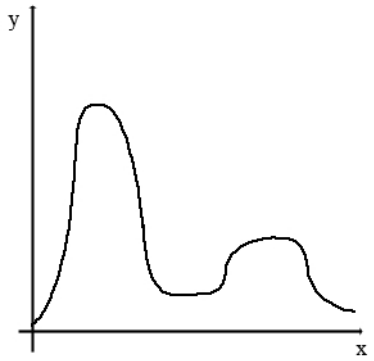
2.7 Uniform Distribution



2.8 A Mixture of Two Distributions



Log normal with normal:



3 Limiting and Asymptotic Properties

Introduce the ideas of the bias and efficiency of estimators. These are finite-sample concepts. We are often interested in the properties of estimators as we increase the sample size. This may be for two reasons:

First, we often find that we need to work with statistics (mainly test statistics such as the t and F) which are not independent of the sample size. If this is the case we need to be able to consider how the estimator will behave if the sample changes, most usually when it gets larger.

Second, in many cases we may not be able to say anything precise about the properties of an estimator in a small sample but we may be able to derive (approximate) results about the estimator as the sample size gets large.

This section describes some of the tools from statistical analysis.

3.1 Pointwise Convergence

A sequence of functions $\{h_n(x)\}$ converges pointwise to a limit function $h(x)$, $h_n(x) \to h(x)$, if for each x

$$\lim_{n\to\infty} h_n(x) = h(x).$$

3.2 Uniform Convergence

A sequence of functions $\{h_n(x)\}$ is said to converge uniformly to a limit function $h(x)$, *if*

$$\lim_{n\to\infty} \sup_x |h_n(x) - h(x)| = 0.$$

3.3 Convergence in Probability.

A random variable x_n is said to converge in probability if:

$$\lim_{n\to\infty} x_n = p\lim x_n = x$$

where x is some arbitrary value is read as "the probability limit of x_n is x" . The probability limit can be rewritten as

$$\lim_{n\to\infty} \Pr\left(x-\varepsilon \leq x_n \leq x+\varepsilon\right) = 1, \text{ with } \varepsilon > 0.$$

3.4 Convergence in Distribution and Limiting Distributions

The limiting distribution of a random variable x is denoted F(x) and we say that the random variable x_n converges in distribution to x, denoted $x_n \xrightarrow{d} x$. The limiting mean and limiting variance are simply the mean and variance of the limiting distribution, $F(x)$. A key result is :

Theorem 1 *The Slutsky Theorem*
If g(x) is a continuous function of x then

$$p\lim g\left(x\right) = g\left(p\lim\left(x\right)\right)$$

which states that "the limit of the function is the function of the limit".

Definition 2 *Jensen's Inequality:*
which states that, in general, $E\left[g\left(x\right)\right] \neq g\left[E\left(x\right)\right]$.
More specifically, when g(x) is a convex function, $E\left[g\left(x\right)\right] \geq g\left[E\left(x\right)\right]$.
Only in the case where g(.) is a linear function does the equality hold.

The general case
If $y = \alpha x^{\beta}$ then $E\left(y\right) = E\left(\alpha x^{\beta}\right) \neq \alpha E\left(x^{\beta}\right)$
However, in the linear case
If $y = \alpha + \beta x$ then $E\left(y\right) = E\left(\alpha + \beta x\right) = \alpha + \beta E\left(x\right)$
another reason why linear models are easy to deal with!
The Slutsky Theorem applies to random vectors (matrices) as well as to random scalars. For example
If $p\lim W_n = \Omega$ then $p\lim W_n^{-1} = \Omega^{-1}$
The key result are therefore as follows:

- $x_n \xrightarrow{p} x \Rightarrow x_n \xrightarrow{d} x$
- If $x_n \xrightarrow{d} x$ and $g(x)$ is a function then the Slutsky Theorem says $g\left(x_n\right) \xrightarrow{d} g\left(x\right)$.
- (Delta Method) If $\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \Sigma\right) \Rightarrow$

$$\sqrt{n}\left(g\left(\widehat{\theta}\right) - g\left(\theta_0\right)\right) \xrightarrow{d} N\left(0, \frac{\partial g\left(\theta\right)}{\partial \theta'}|_{\theta_0} \sum \frac{\partial g\left(\theta\right)}{\partial \theta'}|_{\theta_0}^{'}\right)$$

- If $x_n \xrightarrow{d} x$ and $y_n \xrightarrow{p} \alpha$, where α is a constant, then

$$(i)\ \ x_n + y_n \xrightarrow{d} x + a$$
$$(ii)\ \ x_n y_n \xrightarrow{d} xa$$
$$(iii)\ \ \frac{x_n}{y_n} \xrightarrow{d} \frac{x}{a},\ a \text{ different from } 0.$$

3.5 Orders of Magnitude

Let $\{a_n\}$ be a sequence of real numbers, $\{X_n\}$ be a sequence of random variables and $\{g_n\}$ be a sequence of positive real numbers. Then,

1. a_n is of smaller order (in magnitude) than g_n, denoted $a_n = o(g_n)$, if $\lim_{n\to\infty} \frac{a_n}{g_n} = 0$.

2. a_n is at most of order (in magnitude) g_n, denoted $a_n = O(g_n)$, if there exists a real number B such that

$$\frac{|a_n|}{g_n} \leq B \text{ for all } n.$$

3. X_n is of smaller order (in probability) than g_n, denoted $X_n = o_p(g_n)$, if $\frac{X_n}{g_n} \xrightarrow{p} 0$.

4. X_n is at most of order (in probability) g_n, denoted $X_n = O(g_n)$, if there exists a nonstochastic sequence $\{c_n\}$ such that $c_n = O(1)$ and

$$\left(\frac{X_n}{g_n} - c_n\right) \xrightarrow{p} 0.$$

3.6 Stabilizing Transformations

Finally, it is often the case that the limiting distribution, $F(x)$, of a random variable is a point (often zero). There is very little information in this point and we often want to analyze the properties of a random variable before it collapses to a spike. This can be achieved using a stabilizing transformation.

Consider the following statistic

$$x \sim \left(\mu, \frac{\sigma^2}{n}\right),$$

which says that x has a mean of μ and a variance of $\frac{\sigma^2}{n}$. We know that

$$p\lim x = \mu$$

but as the sample size increases the distribution collapses to a point. This is known as a degenerate distribution.

Imagine we had another statistic with a distribution

$$x' \sim \left(\mu, \theta\frac{\sigma^2}{n}\right)$$

This has the same property that $p\lim x' = \mu$ but even though it has a different variance it too will have the same degenerate distribution. As the sample size gets large enough we cannot distinguish between the two distributions. In other words, different distributions may be indistinguishable in the limit.

However we may be able to define a transformation such that

$$z = h(x) \xrightarrow{d} f(z)$$

but

$$z' = h(x') \xrightarrow{d} f(z')$$

where f(z) is a well-defined limiting distribution. We now have a basis for comparison. This is a so-called stabilizing transformation.

This allows us to introduce our next desirable property of an estimator:

4 Consistency

An estimator is said to be consistent if its probability limit is equal to the true population parameter. In other words

$$p\lim\widehat{\theta} = \theta$$

Thus if, as the sample size increases, the bias and the variance decline and if they both converge on zero when the sample is infinite, then the estimator is consistent.

Note: all unbiased estimators are consistent; but not all consistent estimators are unbiased. An estimator may be biased in finite sample (i.e. $E\left(\widehat{\theta}\right) \neq \theta$) but consistent ($p\lim\widehat{\theta} = \theta$). There is a large class of estimators with this property, the most important being IV and GMM estimators.

For data to be unbiased, we have to make a lot of assumptions. But often these conditions are not satisfied. In infinite samples, $\hat{\theta} = \theta + bias$. If you look only at the asymptotic, you will ignore the bias. With infinite samples, this is a problem. Need an associated measure of bias. Maybe the bias will converge to a distribution and then you can make inference.

$$\hat{\theta} = \overline{X} + \frac{1}{n}$$
$$E(\hat{\theta}) = E(\overline{X})$$
$$\frac{1}{n} = \mu + \frac{1}{n}$$
$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$
$$\hat{\theta_3} = X + \frac{1}{n^2}\hat{\theta}$$
$$E(\hat{\theta_3}) = \mu + \frac{1}{n^2}$$

This is better, because it converges to zero quicker as sample size increases. σ is the order of the error term. If $\frac{1}{n}$ does not converge to zero, we have a biased estimator with a small order σ.

$$\hat{\theta_4} = \overline{X} + \frac{1+n}{1+1000n}$$

The plim is $\frac{1}{1000}$ the bias is small. In this case, it is a big order bias because it is a constant. 0 is a big order bias. This is a very important feature. The order, and the rate of convergence to the zero.

Say we have $\sigma(n)$ where n is the rate of convergence; compared to that, $\sigma(n^2)$ is a faster rate of convergence, because the n is squared. This can tell us how big the sample should be.

5 Using Distribution Theory: The Sampling Distribution of the Sample Mean

Imagine drawing a random sample of n observations from a population and computing a statistic B for example, the sample mean. If we drew another sample we would, obviously, get another value for the statistic. The sample mean is thus a random variable.

We are interested in deriving the sampling distribution of this sample mean in cases where the variable can assume any value and can come from any kind of distribution.

Proposition 3 *If $X_1 ... X_n$ are a random sample and can be regarded as identically and independently distributed draws from a population with mean, μ and variance σ^2 then whatever the form of the distribution of X, the sampling distribution of the random variable $\overline{X}$ will have mean μ and variance $\frac{\sigma^2}{n}$.*

Proof. We define the sample mean as

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\left(X_1 + X_2 + ... + X_n\right)$$

where $X_1, X_2, ..., X_n$ are n variables drawn from the same sample. It is assumed that the X_i are identically and independently distributed. Since n is constant,

$$E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(X_1 + X_2 + ... + X_n\right)$$

We know from Jensen's Inequality that in the case of a linear function, the expectation of a sum is equal to the sum of the expectations, and that the mean of each X_i ,$\overline{X}$, is, by definition, μ:

$$E\left(\overline{X}\right) = \frac{1}{n}E\left(X_1 + X_2 + ... + X_n\right) = \frac{1}{n}\left(\mu + \mu + ... + \mu\right) = \frac{n\mu}{n} = \mu$$

Thus the mean of the sampling distribution of $\overline{X}$ is equal to the population mean. The variance of the sample mean is given by

$$\sigma^2_{\overline{X}} = Var\left(\overline{X}\right) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right)$$

Since the variables are independent, their covariances are zero, then we know that $Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var\left(X_i\right)$ and thus

$$\sigma^2_{\overline{X}} = \frac{1}{n^2}\left(Var\left(X_1\right) + Var\left(X_2\right) + ... + Var\left(X_n\right)\right) = \frac{1}{n^2}\left(\sigma^2 + \sigma^2 + ... + \sigma^2\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

To summarize: If $X \sim \left(\mu, \sigma^2\right)$ then $\overline{X} \sim \left(\mu, \frac{\sigma^2}{n}\right)$. Thus the sample mean is an unbiased estimator of the population mean and as the sample size increases the distribution of $\overline{X}$ converges to a point. Taken together this implies that $\overline{X}$ is a consistent estimator. ■

6 The link with estimation theory

Theorem 4 *The Central Limit Theorem*

If $x_1 ... x_n$ are a random sample from any probability distribution with finite mean μ and variance σ^2 then

$$\sqrt{n}\left(\overline{x} - \mu\right) \xrightarrow{d} N\left(0, \sigma^2\right),$$

which states that the limiting distribution of the mean of the sample is Normal.

There are a number of representations of this expression. For example is we standardize the random variable then the limiting distribution is given as

$$\frac{\sqrt{n}\left(\overline{x} - \mu\right)}{\sigma} \xrightarrow{d} N\left(0, 1\right).$$

If each random variable has a common mean μ, but non-similar variances σ_i^2 then the limiting distribution is given as

$$\sqrt{n}\left(\overline{x} - \mu\right) \xrightarrow{d} N\left(0, \overline{\sigma}^2\right).$$

Finally the central limit theorem applies to a multivariate setting

Theorem 5 *Lindberg-Levy Central Limit Theorem*

If $x_1 ... x_n$ are a random sample from any multivariate probability distribution with finite mean vector μ and finite positive definite covariance matrix $\mathbf{Q}$, then

$$\sqrt{n}(\overline{x} - \mu) \xrightarrow{d} N(\mathbf{0}, \mathbf{Q})$$

which states that the mean of the sample distribution is multivariate Normal.

Central limit theorems give us an indication of the properties of the limiting distribution of the sample mean. The final key theorem is

Theorem 6 *Asymptotic Distribution of the Sample Mean*

If

$$\frac{\sqrt{n}\left(\overline{x} - \mu\right)}{\sigma} \xrightarrow{d} N\left(0, 1\right)$$

then, asymptotically

$$\overline{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

which states that the mean of the random variable x is asymptotically (i.e. in very large samples) distributed normally with a mean of μ (the population mean) and a variance $\frac{\sigma^2}{n}$ (which goes to zero as the sample size goes to infinity).

The central limit theorem is absolutely crucial to econometrics as it allows us to base our inference on the properties of the sample mean on the assumption that its distribution can be approximated by a normal distribution whatever the true (but unknown population distribution).