1 Probability theory

Here we summarize some of the probability theory we need. If this is totally unfamiliar to you, you should look at one of the sources given in the readings.

In essence, for the major part of the work in econometrics, we need knowledge of only a few probability distributions, which "pop up" regularly. But to understand these distributions, we need *some* understanding of what probability is.

2 Random variables and probability theory.

The distribution function of a random variable X

$$F(x) = P(X \leq x)$$

A random variable is *discrete* if the set of outcomes is countable, *continuous* if the outcomes are continuous.

For a discrete random variable, each outcome x has an associated probability $p(x)$.

$$p(x) = P(X = x)$$

Example

x	$p(x)$
0	$\frac{1}{2}$
1	$\frac{1}{2}$

For a continuous random variable, each outcome has probability zero, but we can find probabilities for whether the variable will be in an interval $[a, b]$[1] by:

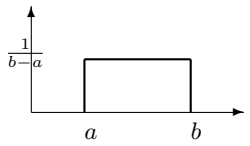
$$P(X \in [a, b]) = \int_a^b f(x)dx$$

$f(x)$ is the *probability density function.*

Example

The *uniform* distribution. The uniform distribution puts equal probability on any point in an interval $[a, b]$.

$$p(x) = \begin{cases} 0 & \text{if } x < a \\ 1 & \text{if } a \leq x \leq b \\ 0 & \text{if } b < x \end{cases}$$



The uniform distribution

For the uniform distrubution on the interval $[0, 1]$, calculate

$$P\left(x \in \left[0, \frac{1}{2}\right]\right) = \int_0^{\frac{1}{2}} p(x)dx = \int_0^{\frac{1}{2}} 1dx = x\|_0^{\frac{1}{2}} = \frac{1}{2} - 0 = \frac{1}{2}$$

[1] To be technical, can find probability of being in any Borel set $B(\mathbb{R})$ by $P(X \in B) = \int_B f(x)dx$

2.1 Expectation

The *expectation* (or mean) $E[X]$ of a random variable X is defined by

$$E[X] = \int_{-\infty}^{\infty} x dF(x) = \begin{cases} \int_{-\infty}^{\infty} x f(x) dx & \text{if } X \text{ is continuous} \\ \sum_x x p(x) & \text{if } X \text{ is discrete.} \end{cases}$$

The expectation of any function $h()$ of a random variable can be found as

$$E[h(x)] = \int_{-\infty}^{\infty} h(x) dF(x)$$

Example

For the case

x	$p(x)$
0	$\frac{1}{2}$
1	$\frac{1}{2}$

$$E[x] = \frac{1}{2}0 + \frac{1}{2}1 = \frac{1}{2}$$

Exercise 1.

The variable x is uniformly distributed on the interval $[0, 1]$.

1. Calculate the expected value of x, $E[x]$.

Solution to Exercise 1.

The uniform (0,1) distribution has expectation given by

$$E[x] = \int_0^1 x \cdot 1 dx = \int_0^1 x dx = \left[\frac{1}{2}x^2\right]_0^1 = \frac{1}{2}[1^2 - 0^2] = \frac{1}{2}$$

Some useful special cases is (let a, b be constants, X, Y be random variables.)

$$E[aX] = aE[X]$$

$$E[aX + bY] = aE[X] + bE[Y]$$

2.2 Variance

The *variance* of a random variable X is

$$\text{var}(X) = E\left[(X - E[X])^2\right]$$

Exercise 2.

The variable x is uniformly distributed on the interval $[0, 1]$.

1. Calculate the variance of x, var(x).

Solution to Exercise 2.

$$\text{var}(x) = E[(x - E[x])^2] = \int_{-\infty}^{\infty} (x - E[x])^2 p(x) dx = \int_0^1 \left(x - \frac{1}{2}\right)^2 dx = \int_0^1 \left(x^2 - x + \frac{1}{4}\right) dx$$

$$= \left[\frac{1}{3}x^3 - \frac{1}{2}x^2 + \frac{1}{4}x\right]_0^1 = \frac{1}{3} - \frac{1}{2} + \frac{1}{4} = \frac{4}{12} - \frac{6}{12} + \frac{3}{12} = \frac{1}{12}$$

Some useful properties

$$\text{var}(a) = 0$$

$$\text{var}(aX) = a^2\text{var}(X)$$

The *joint distribution* of two random variables X and Y is

$$F(x, y) = P(X \leq x, Y \leq y)$$

The distributions

$$F_X(x) = \lim_{y\to\infty} F(x, y)$$

$$F_Y(y) = \lim_{x\to\infty} F(x, y)$$

are called the *marginal distributions* of X and Y.

X and Y are *independent* if

$$F(x, y) = F_X(x)F_Y(y)$$

2.3 Covariance

The *covariance* of two random variables is

$$\text{cov}(X, Y) = E\left[(X - E[X])(Y - E[Y])\right] = E[XY] - E[X]E[Y]$$

If the covariance is zero, the random variables are *uncorrelated*.

Note that while independent variables are uncorrelated (show this), the opposite is not the case. It is possible to construct counterexamples, where one of the two random variables is a (deterministic) function of the other, but we still have a covariance of zero.

For continuous distributions, the calculation of joint probabilities is

$$P(X \in A, Y \in B) = \int_A \int_B f(x, y)dydx$$

Some useful properties:

$$\text{var}(X + Y) = \text{var}(X) + 2\text{cov}(X, Y) + \text{var}(Y)$$

$$\text{cov}(X + Y, Z) = \text{cov}(X, Z) + \text{cov}(Y, Z)$$

$$\text{cov}(aX, Y) = a\text{cov}(X, Y)$$

3 The Normal distribution

The most important distribution we will be using is the normal distribution.

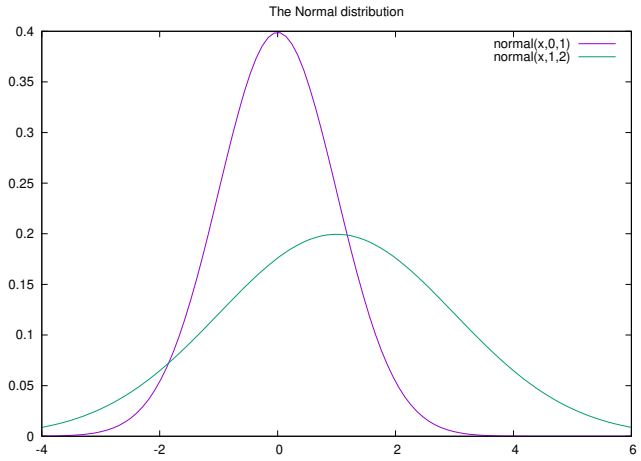
The standard normal distribution is

$$n(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

A normal distributed variable with variance σ^2 and mean μ has pd

$$n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(\mu-x)^2}{\sigma^2}}$$

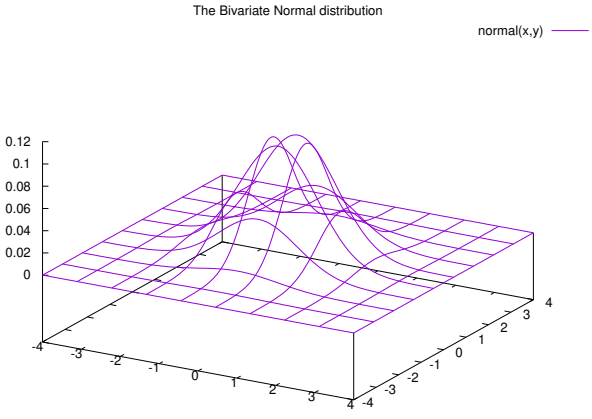
Let us look at a plot of these. We plot a standard normal $\mathcal{N}(0, 1)$ (mean=0, variance 1), and a normal with mean 1 and standard deviation 2. ($\mathcal{N}(\mu, \sigma)$)



The Normal Distributions generalizes to more than one dimension: The multinormal distribution.

$$n(x;\mu,\mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)'\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu)}$$

Here Σ is the covariance matrix. Let us plot the standard case, $\mathcal{N}(\mathbf{0},\mathbf{I})$.



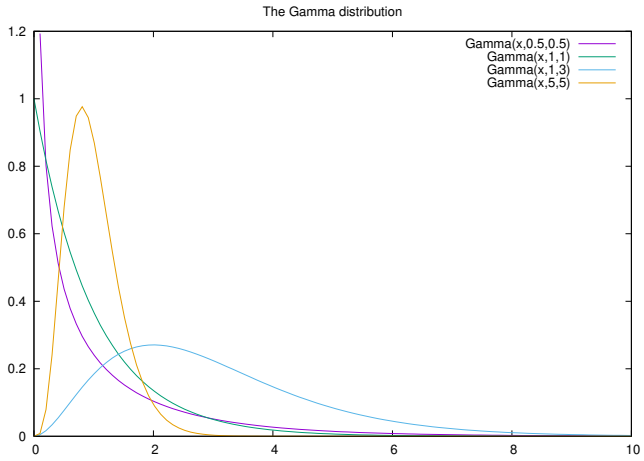
4 The chi-square distribution.

The general gamma distribution is given by:

$$f(x) = \frac{\lambda}{\Gamma(r)} e^{\lambda x}(\lambda x)^{r-1}$$

Here $r > 0$ and $\lambda > 0$ are parameters, and

$$\Gamma(n) = \int_0^\infty x^n e^{-n} dx$$



A special case of the gamma is the chi-square distribution, $\chi^2(x)$, the parameters $r = \frac{1}{2}$ and $\lambda = \frac{1}{2}$.

The chi-square distribution is important for reason of the following result:

Let $X_1, \cdots, X_k$ be independent, normally distributed normal variables with variance 1. Then their square sum $X_1^2 + \cdots + X_k^2$ has a chi-square distribution with k degrees of freedom.

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

5 Limit theorems.

We will be returning to these later, in various guises.

The Law of Large Numbers. If $X_1, X_2, \cdots$ are independent and identically distributed (iid) with common mean μ, then with probability one

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu \quad \text{as } n \to \infty$$

The Central Limit Theorem. If $X_1, X_2, \cdots$ are independent and identically distributed (iid) with common mean μ and variance σ^2, then

$$\lim_{n\to\infty} P\left(\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le a\right) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

(or, the limiting distribution is the normal distribution.)

6 Conditional Expectation

If X and Y are discrete, then the conditional probability that $Y = y$, given that $X = x$, is

$$P_{Y|X}(y|x) = P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)},$$

ie the joint probability of both x and y occurring, divided by the probability that x occurs.

A similar result hold for continuous probability distributions.

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

$$F_{Y|X}(y|x) = P(Y \leq y|X = x) = \int_{-\infty}^{y} f_{Y|X}(y|x)dy$$

The conditional expectation $E[Y|X = x]$ is important in econometrics.

$$E[Y|X = x] = \int_{-\infty}^{\infty} y f_{X|Y}(y|x)dy$$

We will often write this $E[Y|X]$.

An extremely useful result is the *Law of iterated expectations.*

$$E[Y] = E[E[Y|X]]$$

That is, to find the expectation of Y, we can first condition on the outcome X, and then take expectations over all possible outcomes of X.

This property of conditional expectation is useful for several reasons. One is that it can be very useful in calculating expectations:

Exercise 3.

A prisoner is placed in a cell with three doors. If he goes through the first door, he gains his freedom immediately. If he goes through the second door, he wanders in dark tunnels for one day, before he gets back into the cell. If he goes through the third door, he wanders in dark tunnels for three days before he is back in the cell. When he returns to the cell he is so confused that he can not recognise which door he went out through.

1. What is the expected length of time before the prisoner gets to freedom?

Solution to Exercise 3.

Let Y be the expected time to freedom, and X which door he goes through. We want to calculate $E[Y]$, the expected time before the prisoner gets out. The easy way to use this is to use conditional expectations, in particular the law of iterated expectations.

$$E[Y] = E[E[Y|X]]$$

The problem statement translates into

$$E[Y|X = 1] = 0$$
$$E[Y|X = 2] = 1 \text{ day } + E[Y]$$
$$E[Y|X = 3] = 3 \text{ days } + E[Y]$$

Note that $E[Y]$ reappears because when the prisoner returns to the cell, he will again choose doors at random, and he is back where he started.

Now we can easily calculate the expected time to freedom, $E[Y]$.

$$\begin{aligned} E[Y] &= \tfrac{1}{3}E[Y|X = 1] + \tfrac{1}{3}E[Y|X = 2] + \tfrac{1}{3}E[Y|X = 3] \\ &= \tfrac{1}{3} \cdot 0 + \tfrac{1}{3}(1 + E[Y]) + \tfrac{1}{3}(3 + E[Y]) \end{aligned}$$

Solve for $E[Y]$:

$$\begin{aligned} 3E[Y] &= 1 + E[Y] + 3 + E[Y] \\ E[Y] &= 4 \end{aligned}$$

This example showed how conditional expectation can useful as a computational device.

But in economics, we have a more important use of conditional expectations. It is useful for modelling the revelation of *information*. We will often *condition on* the set of available information. In that case we call the random variable X that we condition on an *information set*.

Exercise 4.

Consider the tossing of a (fair) coin twice.

1. What are the possible outcomes?

2. Before you start tossing, what is the probability of observing at least one head?

3. Suppose you see the outcome of the first coin toss.
   (a) Calculate the conditional probabilities of observing one head after having observed the outcome of the first toss.
4. Calculate the expected number of heads.
   (a) Calculate the expected number of heads given that the first toss is a head.
   (b) Calculate the expected number of heads given that the first toss is a tail.

Solution to Exercise 4.
Before we start tossing, the possible outcomes are

$$I_0 = \{hh, ht, th, tt\}$$

This is all we know, that one of these outcomes will occur, each with probability a fourth (the coin is fair).

1. To calculate the expected numbers of heads we count

For example the probability of observing at least one tail is $\frac{3}{4}$

$$P(\text{Observing at least one tail}) = \frac{3}{4}$$

But after having seen the outcome of one coin toss, we have more information. Then we know only one of the following outcomes are feasible:

$$I_1 = \{\{hh, ht\}, \{th, tt\}\}$$

If the first toss is a tail, the probability of observing at least one tail is one, if the first toss is a head, this probability is one half.

$$P(\text{Observing at least one tail}|\text{First is } h) = \frac{1}{2}$$
$$P(\text{Observing at least one tail}|\text{First is } t) = 1$$

This example illustrates how we can think of the revelation of information as a random variable that can be used in forming a conditional expectation. The conditional expectation is changing as more information is being revealed. The conditional expectation can in economic settings for example be used to set prices.

The concept of rational expectations uses this conditional expectation concept.

7 Inequalities

Cauchy-Schwarz inequality

$$E[|XY|] \leq \sqrt{E[X^2]E[Y^2]}$$

8 Convergence

Convergence almost surely: X_n converges *almost surely* to X, ($X_n \overset{a.s.}{\to} X$), if either of the following hold

1.
$$\{\ \limsup X_n(\omega) = \liminf X_n(\omega)\ \forall\ \omega \notin A,\quad P(A) = 0\ \}$$

2.
$$P(\{\omega | X_n(\omega) \to X(\omega)\}) = 1$$

3.
$$\lim_{n\to\infty} P(\omega|\ |X_n(\omega) - X(\omega)| < \epsilon\ \forall\ n > m\}) = 1\ \forall\ \epsilon > 0$$

Convergence in probability: X_n converges *in probability* to X, ($X_n \overset{P}{\to} X$), if either of the following hold

1.
$$\forall \epsilon > 0 \lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$$

2.
$$\forall \epsilon > 0,\ \forall\ \delta > 0\ \exists N \ni \text{ if } n > N, P(|X_n - X| > 0) < \delta$$

Cauchy criterion

Let X_n be a sequence. If $|X_n - X_m|$ converges to 0, in either probability or a.s, then there is X such that $X_n \to X$, in either probability or a.s.

Exercise 5.

This problem is meant to illustrate the Difference between convergence a.s. and in probability. Let $(\Omega, \mathcal{B}, \mathcal{P})$ be $([0,1], \mathcal{B}_{\mathbb{R}\cap[0,1]}$, Lebesgue measure) and define k by $n = 2^k + \nu$ with $0 \leq \nu < 2^k$. Now define

$$X_n(\omega) = \begin{cases} 1 & \text{if } \omega \in \left[\frac{\nu}{2^k}, \frac{\nu+1}{2^k}\right] \\ 0 & \text{otherwise} \end{cases}$$

Note that $P(X_n \neq 0) = P(X_n = 1) = \frac{1}{2^k} \to 0$ as $n \to \infty$.

1. Show that X_n converges in probability, but not almost surely, $X_n \xrightarrow{P} 0$ but it is not the case that $X_n \xrightarrow{a.s.} 0$

Solution to Exercise 5.

Exercise 6.

A random variable x has a mean of 2 and a variance of 3.

1. Find the expected value for the random variable $y = 2x^2 + 5x + 4$.

Solution to Exercise 6.

$$E[x] = 2$$
$$\text{var}(x) = E[(x - E[x])^2] = E[x^2] - E[x]^2$$
$$3 = E[x^2] - 2^2$$
$$3 = E[x^2] - 4$$
$$E[x^2] = 3 + 4 = 7$$
$$E[y] = 2E[x^2] + 5E[x] + 4 = 2 \cdot 7 + 5 \cdot 2 + 4 = 14 + 10 + 4 = 28$$

9 Problems

10 Readings.

The material here is in Theil (1971). Sections 2.1, 2.2, 2.3, 2.4, (2.5), 2.6, (2.7), (2.8),(2.9) are relevant.

11 Further Readings.

If you do not know any probability theory, Ross (2009) is a good introduction.

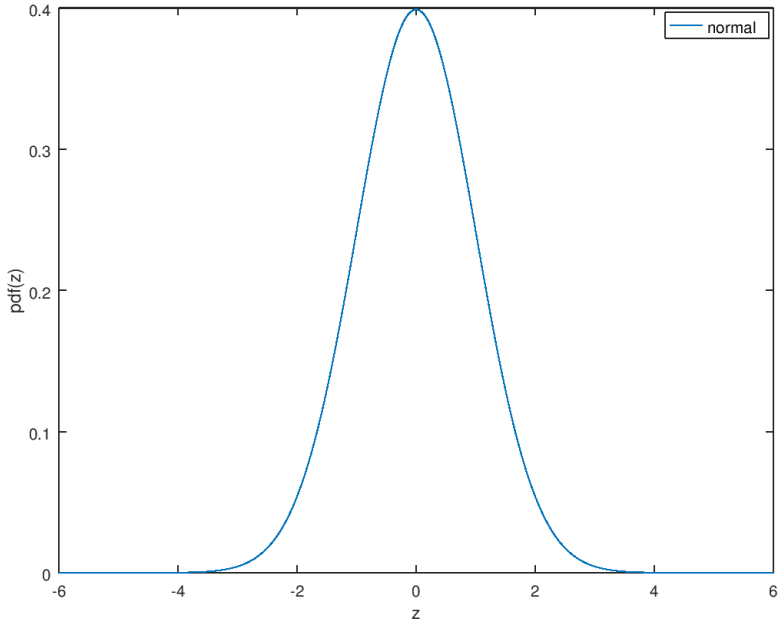
(Spanos, 1986, Part II) is a rigorous presentation of the material, but only if you have a good background.

Rao (1973) is the ultimate source for econometricians.

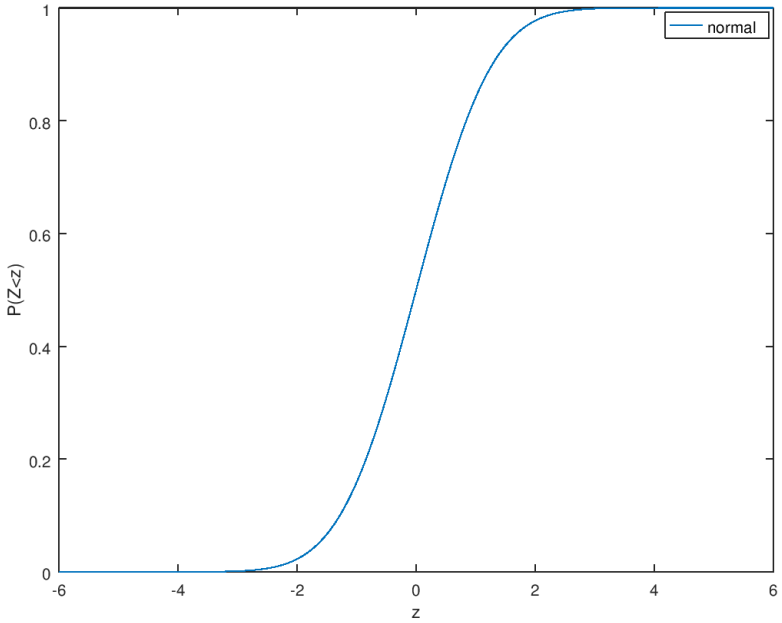
12 Probability distributions typically encountered in statistics

The unit normal distribution Is defined over $(-\infty, \infty)$, but most of the probability mass sentered at zero

The probability density function of a normal distribution

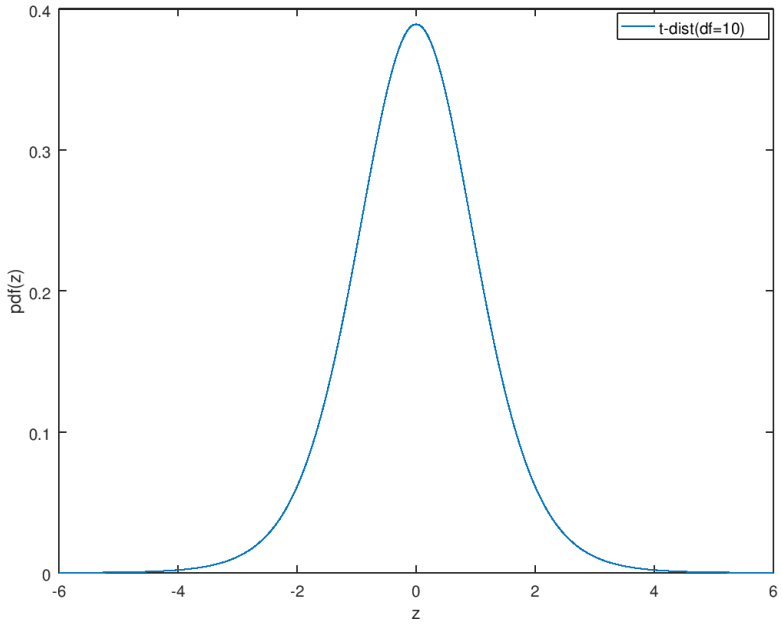


The cumulative probability function of a normal distribution

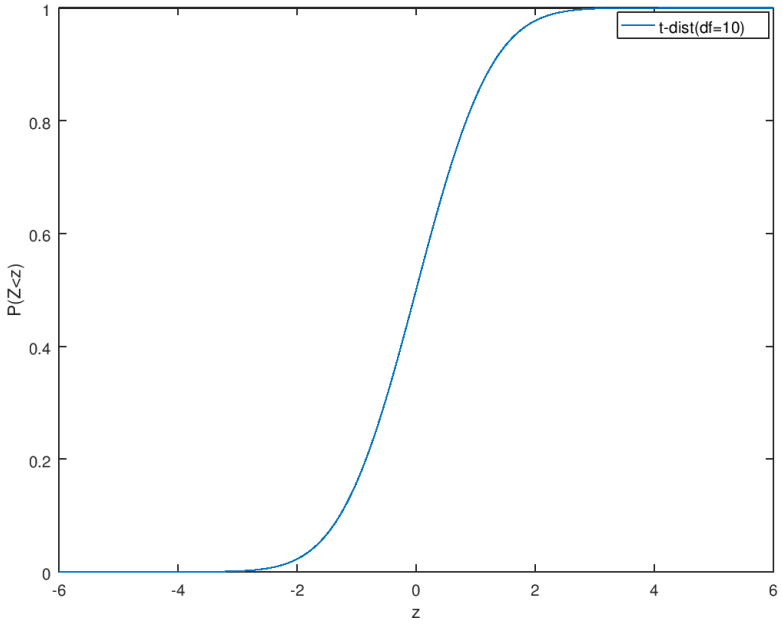


The T distribution Is similar to the normal, defined over $(-\infty, \infty)$, but most of the probability mass sentered at zero

The probability density function of a t-distribution (df=10)

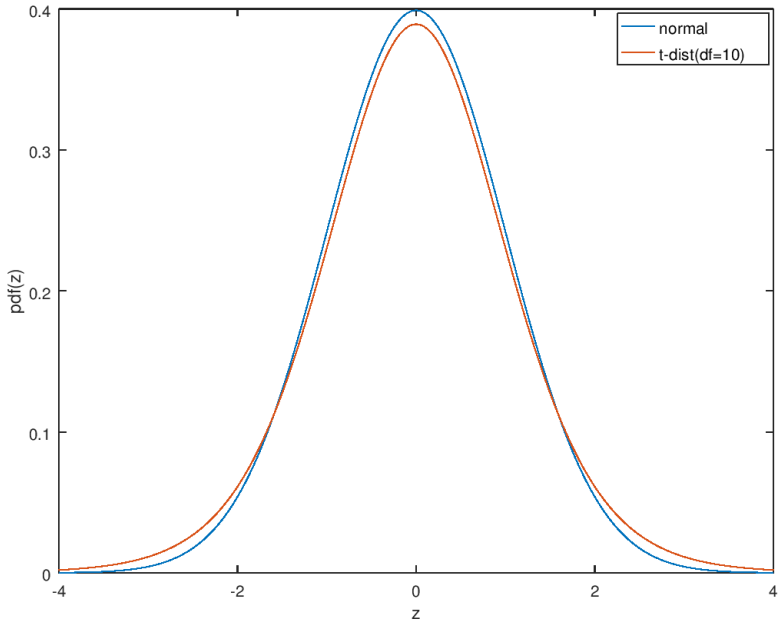


The cumulative probability function of a t-distribution (df=10)

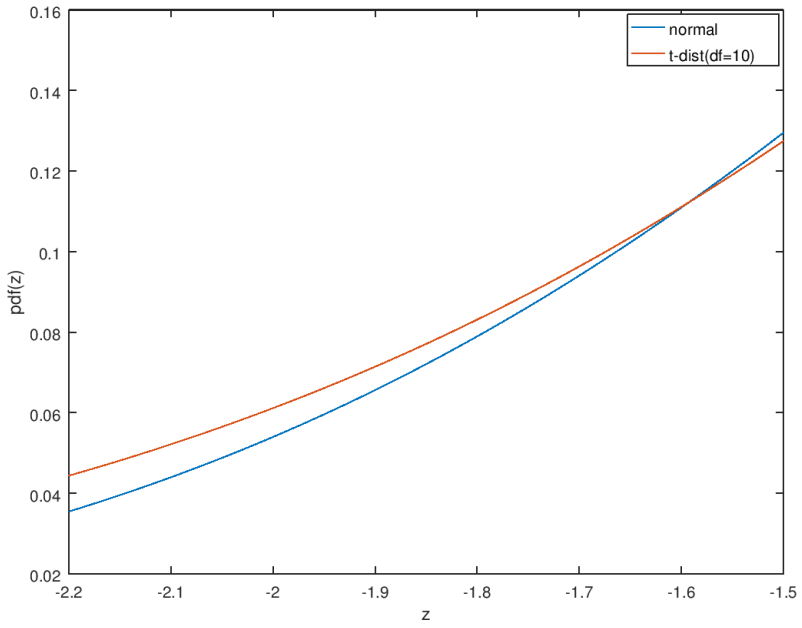


Comparing normal distribution and t distribution What is the difference betwen the normal and the t-distribution?

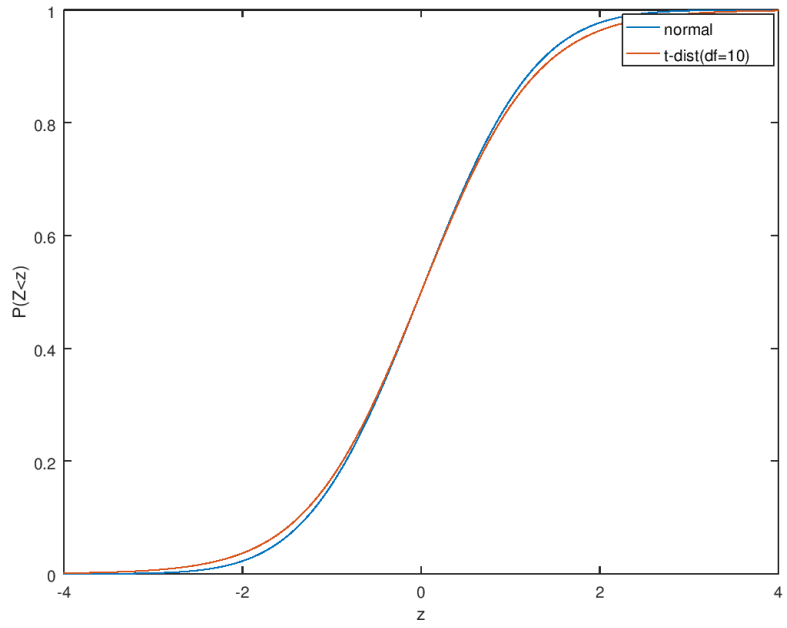
Comparing probability density functions of a normal distribution and a t-distribution (df=10)



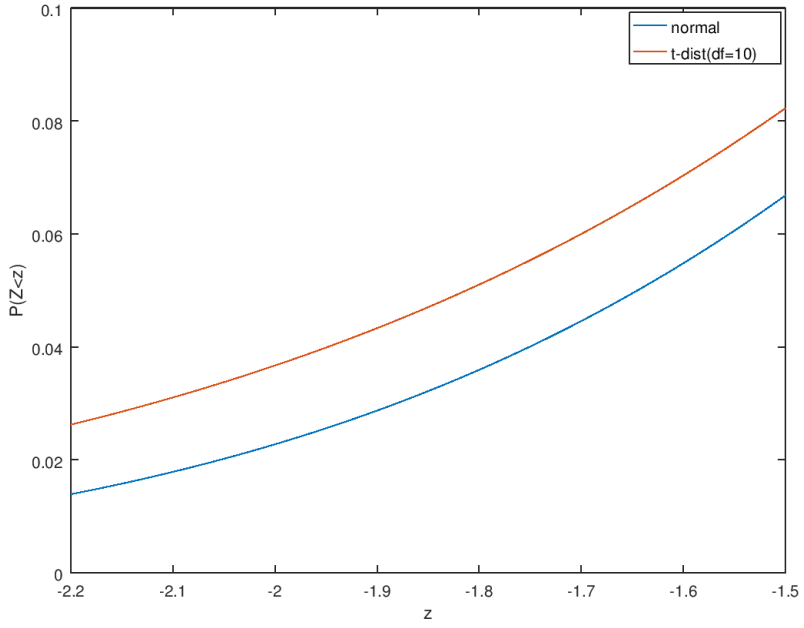
Comparing probability density functions of a normal distribution and a t-distribution (df=10) - Detail



Comparing cumulative probability functions of a normal distribution and a t-distribution (df=10)



Comparing cumulative probability functions of a normal distribution and a t-distribution (df=10) - Detail

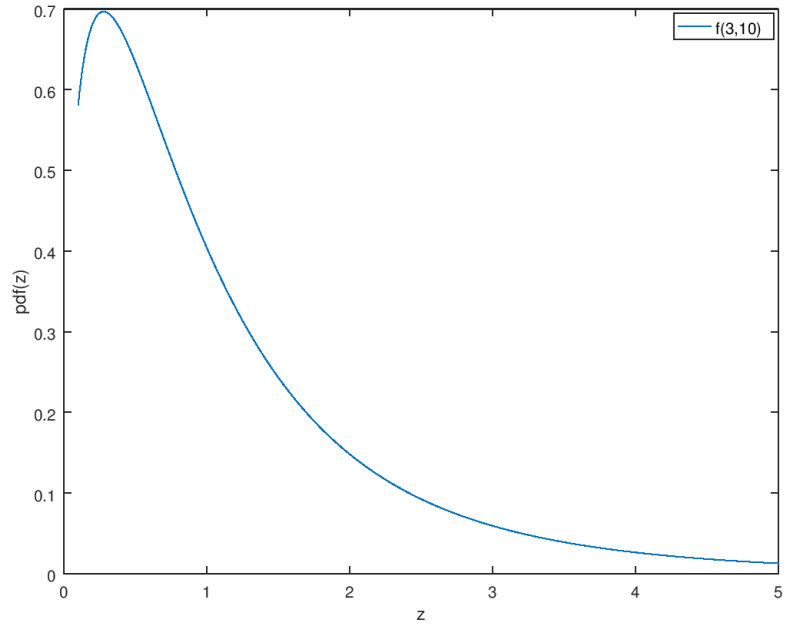


When do you use one or the other?

The t-distribution corrects for the fact that you do not have an infinite sample. It is more conservative regarding rejecting the null, and will result in wider confidence bands.

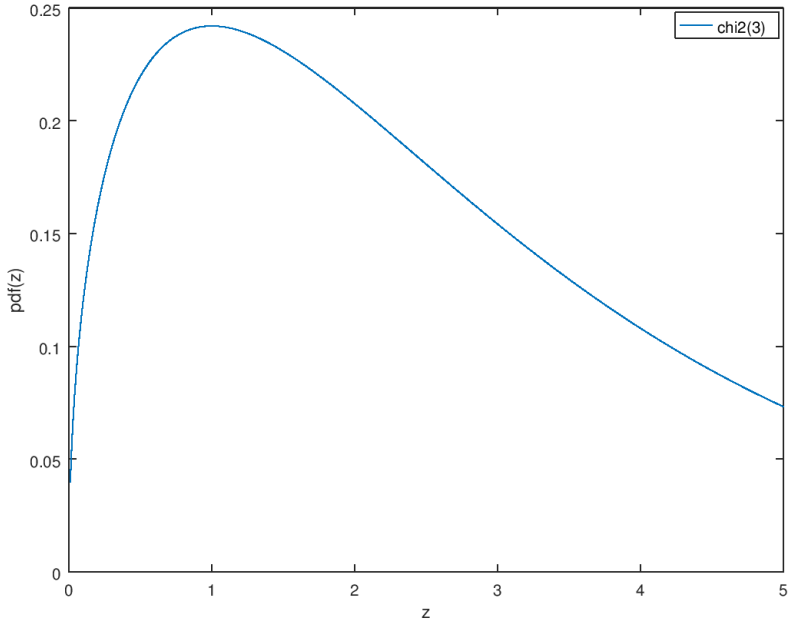
F dist

The probability density function of a F-distribution (df=3,10)



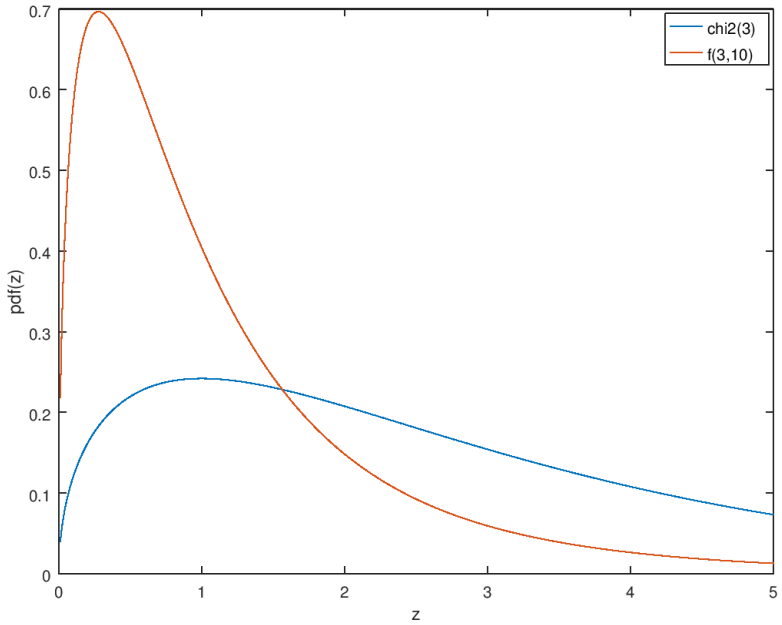
χ^2 dist

The probability density function of a χ^2-distribution (df=3)

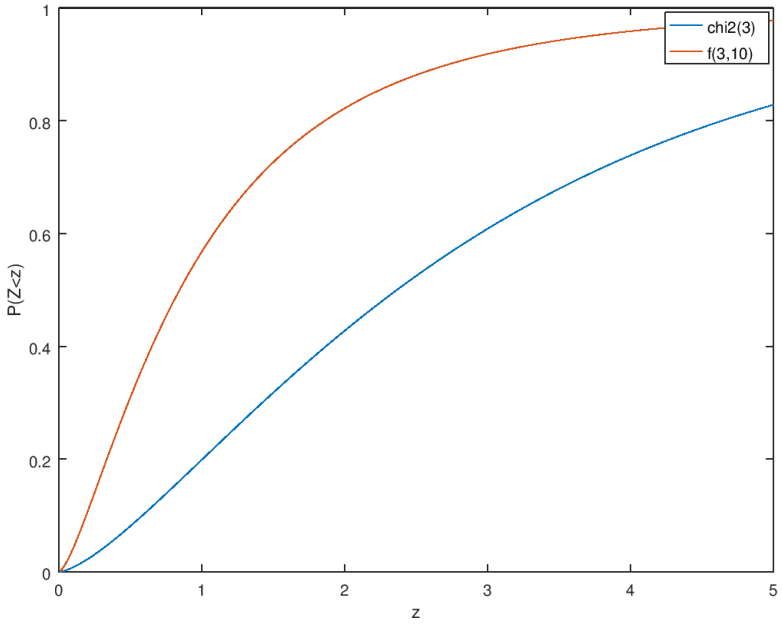


Comparing F and χ^2 distributions

The probability density functions of a F and χ^2-distribution (df=3)



The cumulative probability functions of a F and χ^2-distribution (df=3)



12.1 Looking at large samples by simulation

So far we have discussed test statistics under the assumption of normality

$$e \sim \mathcal{N}(0, \sigma^2 I)$$

But assuming normality is usually too strong. Instead, we will want to make assumptions that allow us to make statements about the behaviour of statistics in large samples.

The main concepts we will use are

- Convergence in probability.
- Convergence in distribution.

Let x_T be a sequence of random variables

Definition 1 *x_T converges to a constant c if*

$$\lim_{T \to \infty} P(|x_T - c| > \epsilon) = 0 \ \forall \ \epsilon > 0$$

will often write this as

$$x_T \xrightarrow{P} x$$

or

$$\text{plim } x_T = x$$

Example

Consider estimating the mean of a normally distributed random variable.

$$x_T = \sum_{t=1}^{T} x_t$$

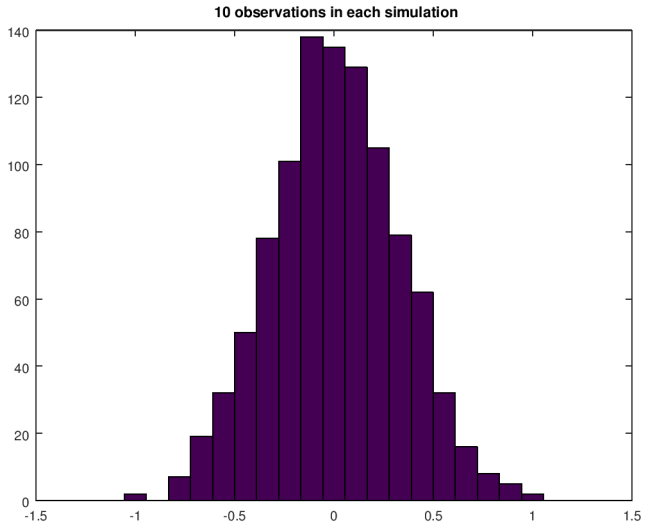
A property most will agree is desirable is that as the numbers of observations increase, the mean gets closer to the true mean of the distribution.

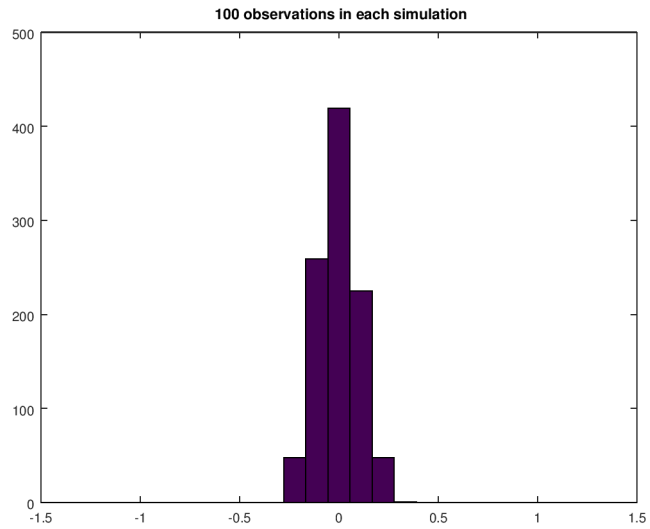
Let us illustrate that by simulating a drawing of T random variables, and calculate the mean these T observations of each simulation.

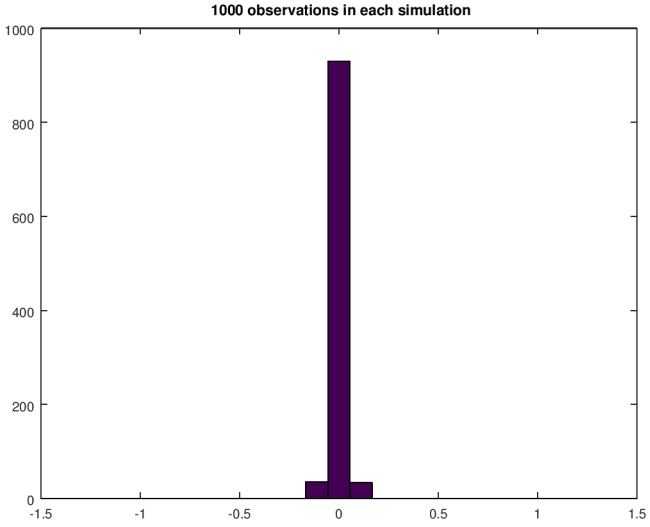
We repeat this experiment, and make a histogram of where the mean is

We hope that the estimate of the mean gets to be more accurate as the number of observations increase.

The following diagrams shows histograms of the distribution of the mean for $T = 10, 100$ and 1000.







It is clear that as we get more observtions, the sample mean becomes a more and more accurate description of the true mean.

This is of course a desirable property of any estimator

We formalize this by the concept of *consistency*

Definition 2 *An estimator* $\widehat{\theta}$*. of a parameter* θ *is a* consistent *estimator if*

$$plim\ \widehat{\theta} = \theta$$

The property of consistency is similar to the unbiasedness property, but more general.

What about distributions?

In the previous we have assumed normality of the error term. We want to make probability statements about the behaviour of the test statistics in large samples.

That is what Central Limit Theorems does

Example

Again look at drawings from a normal distribution, and properties of the sample mean.

When we looked at

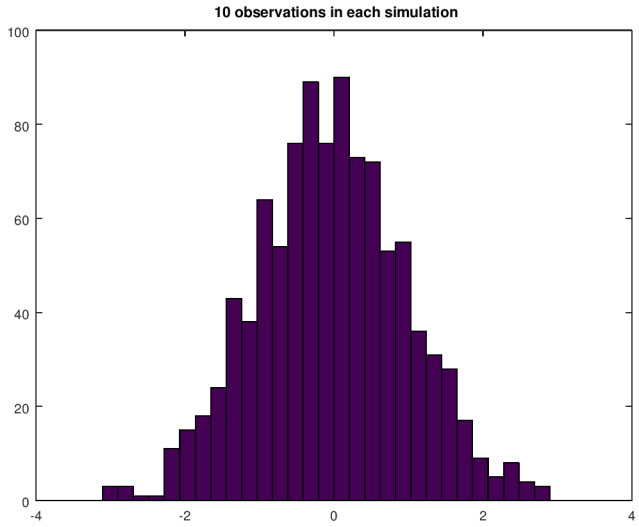
$$x_T = \sum_{t=1}^{T} x_t$$

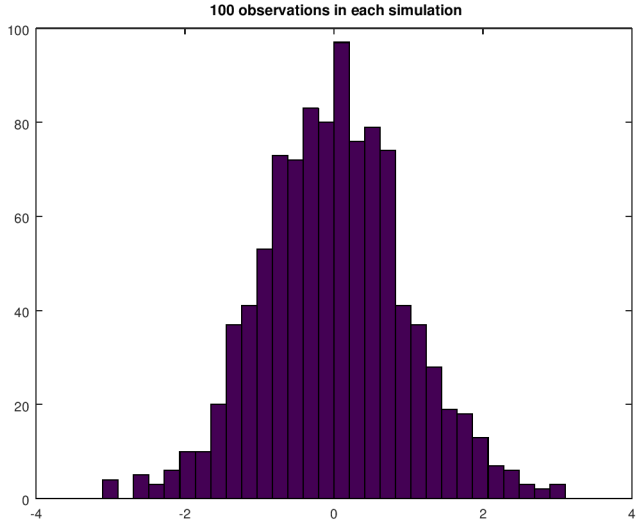
we found all probability mass concentrated at the mean. That is of course not that useful if we want to make probability statements. Let us look at

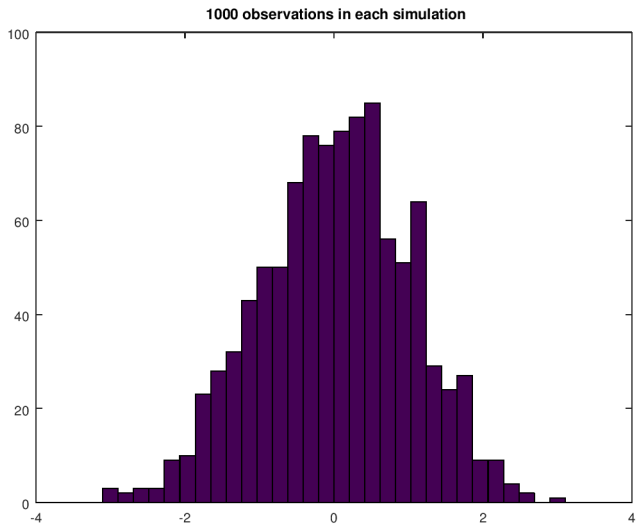
$$\sqrt{T}x_T$$

It turns out that *this* will have a "proper" distribution.

In the following I have plotted outcomes of $\sqrt{T}x_T$ for $T = 10, 100$ and 1000.







As you see, $\sqrt{T}x_T$ does not concentrate in a "spike" it stays "spread out"
Now you see why we multiply with $\sqrt{T}$

Definition 3 *Central Limit Theorem: If $x_1, \ldots, x_t$ are random samples from any probability distribution with mean μ and finite variance σ^2. Let*

$$\overline{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$$

Then

$$\sqrt{T}(\overline{x}_T - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

Here $\xrightarrow{D}$ means "convergence in distribution."

This concept is harder to formalize than convergence in probability.

Definition 4 *x_t converges in distribution to a random variable x with cdf F if*

$$\lim_{T\to\infty} |F_T(x) - F(x)| = 0$$

at all continuity points of F.

We also talk about *asymptotic distribution*

Definition 5 *An asymptotic distribution is a distribution that is used to approximate the tru finite sample distribution of a random variable.*

The form we will see these results in is usually that an estimator $\widehat{\theta}$ satisfies

$$\sqrt{T}(\widehat{\theta} - \theta_0) \xrightarrow{D} \mathcal{N}(0, V)$$

wher θ_0 is the "true" expected values, and V a covariance matrix.

References

C Radhakrisna Rao. *Linear Statistical Inference and its applications.* Wiley, Second edition, 1973.

Sheldon M Ross. *Introduction to Probability Models.* Academic Press, Tenth edition, 2009.

Aris Spanos. *Statistical foundations of econometric modelling.* Cambridge University Press, 1986.

Henri Theil. *Principles of econometrics.* Wiley, 1971.