Probability distributions in regression modeling

Contents

Disclaimer: sections and lines in brown are 'under construction'.

1 Motivation – Probability distributions in regression modeling

In regression modeling, we need random variables — and their probability distributions — because our models do not fit our data exactly. A regression model is composed of:

1. a deterministic model which captures as much of the data variation as possible;
2. an error term e characterized by a probability distribution which captures the unexplained variation (the variation that remains *after* predicting the population regression vector).

Regression models are therefore not deterministic. Whether they are used for parameter estimation or prediction, **we assess the uncertainty in our estimates/predictions using probability distributions.**

Notes

- **Notations**
  - In the context of probability theory (such as in this document), random variables are commonly written with capital letters. Indeed, as we rarely use vectors but rather individualize random variables, there is no risk of confusion by using capital letters.
  - In the modeling context, however, we are always manipulating *sequences of* random variables. E.g., in the classical linear regression model: $y_i = X_i'\beta + e_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + e_i, \quad e_i \overset{\text{iid}}{\sim} (0, \sigma^2), \quad i = 1, ..., n$, each individual y_i is an individual random variable.[1] As a result we manipulate many random variables in groups, and therefore vectors and matrices. Restricting capital letters to random variables would be problematic. Therefore, in the modeling context, lowercase letters usually refer to scalars, capital letters to matrices, and vectors are either or. In this document, we will use italic lower-case letters for scalars, roman lowercase letters for vectors (ex: $\mathrm{x}_i = \{x_{ik}\}_{k=1}^{p}$ is the p-dimensional vector of all dependent variables for unit i; $\mathrm{x}_k = \{x_{ik}\}_{i=1}^{n}$ is the n-dimensional vector of dependent variable x_k for all units), and roman capital letters for matrices (ex: X is the $n \times k$ matrix of all independent variables for all units).
- **Is x_i a random or a fixed variable?**

  In a regression model $y_i \sim f(\mathrm{x}_i, e_i|\theta) \ \forall i$, e_i is a random variable and therefore y_i, a transformation of e_i, is itself a random variable. However, the explanatory variables x_i may be considered random or fixed. The implications of that choice include which versions of convergence theorems (CLTs and LLNs) we will draw sampling-based inferences from, and the formulation of the model as y_i's marginal or conditional distribution. The details go beyond the objective of this document; we simply note:

  - In experimental studies, the researcher "sets" the values of x_i, therefore the X_is are typically considered fixed, i.e., real vectors $\mathrm{x}_i \in \mathbb{R}^p$.
    - model: a function $f()$ s.t. $y_i \sim f(\mathrm{x}_i, e_i|\theta) \ \forall i$
    - estimation: in OLS, we assume $\mathbb{E}[e_i] = 0$ and look for $\mathbb{E}[y_i]$.
  - In observational studies, we draw a sample $\{y_i, \mathrm{x}_i\}_{i=1}^{n}$ from the population, therefore the x_is are typically considered random.
    - model: $f()$ now gives the conditional distribution: $y_i|\mathrm{x}_i \sim f(\mathrm{x}_i, e_i|\theta)$
    - estimation: in OLS, we assume in addition $\mathbb{E}[e_i|X_i] = 0$ and look for $\mathbb{E}[y_i|\mathrm{x}_i]$.

[1] A common mathematical mistake is to consider $y_1, ..., y_n$ as different values or "realizations" of a single random variable y. Instead, we are dealing with a *sequence* of n random variables. If for example Y is a person's height, and there is a sample of n persons, then $Y_{i=1}$ is the random variable representing the height of the first person in sample, $Y_{i=2}$ is the random variable representing the height of the second person in sample, etc. Each Y_i of the sequence will then have a realization which we may or may not observe.

In general, we will assume the context of an observational study, and therefore consider x_i random variables.

2 Definitions

2.1 Random variables & probability distributions

A **random variable** (r.v.) $\mathbf{X}$ is the formalization of the outcome of a random process. Mathematically, it is a function $X\colon S \mapsto E$ that maps a "sample space" S of all possible results of that process to a measurable space E, often the real numbers.[a] It can be:

- discrete — *Ex: the outcome of a coin toss:* $X \equiv \{1 \text{ if tails}, 0 \text{ if heads}\}$
- continuous — *Ex: a continuous uniform r.v.* $X \sim \mathcal{U}[a,b]$

Its **probability distribution** is a probability measure on E. It is described using:

- the **probability function $f_X(x)$**: the probability of occurrence of each value in E.
  - for a discrete r.v., $f_X(x) = \mathrm{P}(X = x)$ is called the probability mass function.
  - for a continuous r.v., $f_X(x)$ is called the probability density function.
- the **cumulative distribution function $F_X(x) \equiv \mathrm{P}(X<x)$**: the area under $f_X()$ over $]-\infty, x[$.

[a] As a function on S, X should be written $X()$, defined as $X(s)$, $\forall s \in S$. Similarly, in a statistical model with a sample of size N, we would write $\forall i \in N, X(i)$ or X_i. Typically, we will drop the notations $X(s)$ or X_i and just write X, leaving it implicit that it is a function defined on S.

A set of random variables $\{X, Y\}$ has a *joint* probability distribution $f_{XY}()$. Each r.v. also has an *individual* or *marginal* probability distribution, and a *conditional* probability distribution, s.t.:

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

2.2 Every probability distribution has its moments...

We often focus on a few moments of a random variable's probability distribution: its 1st, 2nd central, and 3rd standardized moments, and when looking at multiple random variables, their 2nd central mixed moment:[2]

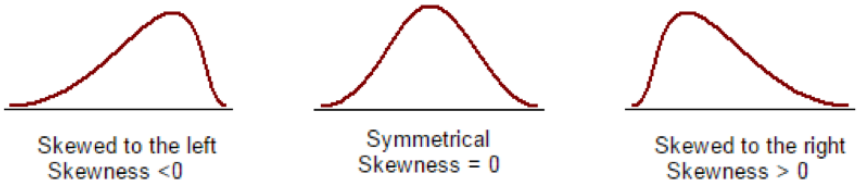
For X:

Expectation	$\mathbb{E}[X] \equiv \int x f_X(x)\ dx = \int x\ dF_X(x)$ is the distribution's center of mass or **mean**.
Variance	$\mathbb{V}[X] \equiv \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
Standard deviation	$\mathrm{SD}[X] \equiv \sqrt{\mathbb{V}[X]}$ is a commonly used measure of variability.[3]
Skewness	$\mathrm{Skew}[X] \equiv \dfrac{\mathbb{E}[(X-\mu_X)^3]}{\sigma_X^3}$ is a measure of the distribution's asymmetry.[4] For a unimodal continuous r.v.: • Skew < 0 = "skewed to the left" = "left-tailed" (fat left tail) • Skew > 0 = "skewed to the right" = "right-tailed" (fat right tail)

[2] The r^{th} moment of X's probability distribution is $\mathbb{E}[X^r]$. Its r^{th} *central* moment is $\mathbb{E}[(X-\mu_X)^r]$. Its r^{th} *standardized* moment is $\mathbb{E}[(X-\mu_X)^r]/\sigma_X^r$.

[3] A. Gelman advises against looking at the variance, as it is in the wrong units. One should look instead at the standard deviation, which represents the spread of the variable and is therefore in the right units.

[4] ⚠ A symmetric distribution has $Skew = 0$, but the reverse isn't true. E.g., a distribution with one long but thin tail, and one short but fat tail, will have $Skew = 0$.



For X, Y:

Covariance $\text{cov}[X,Y] \equiv \mathbb{E}\big[\big(X - \mathbb{E}[X]\big)\big(Y - \mathbb{E}[Y]\big)\big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

Correlation $\rho_{X,Y} \equiv \dfrac{\text{cov}[X,Y]}{\sigma_X \sigma_Y} \quad \in [-1,1]$

Pearson's correlation coefficient is the normalized covariance. The covariance is not easy to interpret as its value depends on the values of the variables. Instead, $\rho_{X,Y}$ normalizes the variables, s.t. its magnitude shows the *strength* of the *linear* relation.

⚠ Covariance and correlation are measures of *linear* dependence only. $\rho_{X,Y}$ is the slope of the regression of standardized Y on standardized X, i.e., the strength of the *linear* relation. A correlation of 0 would only indicate that there is no *linear* relationship between the variables. They may have a nonlinear relationship; always check using a scatterplot.

2.3 ... of which we can compute only sample equivalents

Given a sample of observations $\{x_i, y_i\}_{i=1}^n$, we can define sample equivalents of the moments and properties of X and Y's individual and joint probability distributions. It will be important to correct for bias (from reduced degrees of freedom) when necessary:[5]

	Population parameter	*Bias-adjusted* sample equivalent
Expectation	$\boldsymbol{\mu_X} = \mathbb{E}[X]$	$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$
Variance	$\boldsymbol{\sigma_X^2} = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big]$	$s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$
Standard deviation	$\boldsymbol{\sigma_X}$	$s_x = \frac{1}{\sqrt{n-1}}\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}$
Covariance	$\boldsymbol{\sigma_{X,Y}^2} = \mathbb{E}\big[\big(X - \mathbb{E}[X]\big)\big(Y - \mathbb{E}[Y]\big)\big]$	$s_{x,y}^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$

2.4 Robust measures: *median, interquantile range...*

The mean (as a measure of central tendency) and the standard deviation (as a measure of statistical dispersion) are heavily influenced by the magnitude of values. An outlier can therefore change the sample mean and sample standard deviation drastically, and these measures can also be misleading with skewed data.[6]

We can use instead measures that are rank-based and therefore *robust* to outliers, for example to measure:

[5] For example: by definition, an *unbiased* estimator of σ^2 is a random variable s^2 s.t. $\mathbb{E}[s^2] = \sigma^2$. We can show that $\mathbb{E}[\sum_{i=1}^n (x_i - \bar{x})^2] = \ldots = (n-1)\sigma^2$. So we must divide this sum by $n-1$ for it to be an *unbiased* estimate of σ^2. The reason for this $n-1$ bias has to do with *degrees of freedom*. Informally: the bias is caused by our using the mean calculated from the sample $\bar{x}$ instead of the mean of the population $\mathbb{E}[X]$, and manifests itself as $n-1$ instead of n because we lost one degree of freedom by calculating this sample mean (remember that every time we calculate a statistic, we lose a degree of freedom).

[6] In the standard deviation, the distances from the mean are squared, so large deviations are weighted more heavily, and thus outliers can heavily influence it.

- central tendency: the **median** $\tilde{\boldsymbol{X}}$. It has a breakdown point[7] of 50% (while the mean has a breakdown point of $\frac{1}{n}$: a single large observation can throw it off).
- statistical dispersion
  - The **interquantile range (IQR)** is the difference between the upper and lower quartiles: $\text{IQR}[X] \equiv \mathbb{Q}_X(.75) - \mathbb{Q}_X(.25) \equiv F_X^{-1}(.75) - F_X^{-1}(.25)$. It has a breakdown point of 25%.
  - The difference between the 90^{th} and the 10^{th} percentile divided by the mean.
  - The **median absolute deviation (MAD)** is the median of the absolute deviations from the median: $\text{MAD}[X] \equiv \text{median}\big[|X_i - \tilde{X}|\big]$

2.5 Standardizing: *z-score*

A standardized variable or "z-score" is a variable that has been centered and scaled to have mean 0 and standard deviation 1. The standardized z measures how many standard deviations the raw y is from the mean.

Population: $z = \dfrac{y - \mu}{\sigma_y}$

Sample: $z = \dfrac{y - \bar{y}}{s_y}$, where s_y = the sample standard deviation

2.6 Independence $\Longrightarrow$ uncorrelation = orthogonality

Independence and correlation are statistical concepts, whereas orthogonality is a linear algebra concept.

- 2 random variables X and Y are **independent** $(\boldsymbol{X \perp\!\!\!\perp Y})$ iff $f_{XY}(x,y) = f_X(x) f_Y(y),\ \forall (x,y)$.
- 2 random variables X and Y are **uncorrelated** iff $\text{cov}[X,Y] = 0$, i.e., $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.
- 2 vectors u and v are **orthogonal** iff their inner product $<u,v> = 0$
  A space of random variables can be considered a vector space. We can therefore define an inner product in that space, in different ways. One common choice is to define it as the covariance: $<X,Y> \equiv \text{cov}[X,Y]$. 2 r.v. X and Y are therefore **orthogonal** iff $\text{cov}[X,Y] = 0$.

independent	$\Longrightarrow$	uncorrelated, orthogonal
$\shortparallel$		$\shortparallel$
$X \perp\!\!\!\perp Y$		$\text{cov}[X,Y] = 0$

[7] The breakdown point of an estimator is the proportion of incorrect observations (e.g. arbitrarily large observations) it can handle before giving an incorrect (e.g., arbitrarily large) result.

3 Common families of probability distributions

3.1 Discrete

Many discrete probability distributions are built from the concept of Bernouilli(p) trials. A Bernouilli(p) trial is a random success/failure experiment, where the probability of success is p.
But we can also think of a successful outcome of a Bernouilli trial as the occurrence of an event. With this view then, for example, the number of *occurrences* of an event in a sequence of observations, is also the number of successes in a sequence of n *iid* Bernoulli(p) trials. The table below uses interchangeably "event" and "success".

Name	**Description & Support**	**PDF** $\mathbf{P(X{=}x \mid \theta) = ...}$	**Moments** $\mathbb{E}[X\|\theta], \mathbb{V}[X\|\theta]$
Bernoulli $X \sim Ber(p)$	X is the outcome of a single Bernouilli trial with probability p of success: $X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$	$= p^x(1-p)^{1-x}$	$\mathbb{E} = p$ $\mathbb{V} = p(1-p)$
Binomial $X \sim \mathcal{B}(n,p)$	X is the number of successes in a sequence of n *iid* Bernoulli(p) trials. $X = 0, 1, ..., n$	$= \binom{n}{x} p^x (1-p)^{n-x}$	$\mathbb{E} = np$ $\mathbb{V} = np(1-p)$
Poisson $X \sim Pois(\lambda)$	X is the number of events occurring in a fixed interval, where each occurs randomly and independently with a constant mean rate λ. $X = 0, 1, ...$	$= \frac{\lambda^x e^{-\lambda}}{x!}$	$\mathbb{E} = \lambda$ $\mathbb{V} = \lambda$
Negative Binomial $X \sim NB(r,p)$	X is the number of successes in a sequence of *iid* Bernoulli(p) trials before a number r of failures occurs. $X = 0, 1, ...$	$= \binom{x+r-1}{x} p^x (1-p)^r$	$\mathbb{E} = \frac{pr}{1-p}$ $\mathbb{V} = \frac{pr}{(1-p)^2} = \mathbb{E} + \frac{\mathbb{E}^2}{r}$

Note on the count distributions $NB(r,p)$ and $Pois(\lambda)$:

- $X \sim Pois(\lambda)$ is equidispersed: $\mathbb{V}[X] = \mathbb{E}[X]$. This assumption is not often realistic: count data are usually overdispersed, i.e., their variance is higher than their mean.
- $X \sim NB(r,p)$ is overdispersed: $\mathbb{V}[X] > \mathbb{E}[X]$. In the case of overdispersed count data, we may therefore want to choose the NB distribution over Poisson. $\frac{1}{r}$ is referred to as the dispersion parameter. As it gets smaller, the variance converges to the mean, and the negative binomial turns into a Poisson distribution.

3.2 Continuous

Name	Description & Support	PDF $\mathbf{P(X=x \mid \theta) = ...}$	Moments $\mathbb{E}[X\|\theta], \mathbb{V}[X\|\theta]$
Uniform $X \sim \mathcal{U}(a,b)$	X is the outcome from a trial that is limited between two bounds: $X \in [a,b]$	$= \frac{1}{b-a}$	$\mathbb{E} = \frac{1}{2}(a+b)$ $\mathbb{V} = \frac{1}{12}(b-a)^2$
Beta $X \sim Beta(\alpha,\beta)$	X is a process limited to intervals of finite length, such as a percentage or a proportion.[8] $\alpha, \beta > 0$ are two shape parameters. $X \in [0,1]$	$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$\mathbb{E} = \frac{\alpha}{\alpha+\beta}$ $\mathbb{V} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Gamma[9] $X \sim \Gamma(\alpha,\beta)$	$\alpha > 0$ is a shape parameter and $\beta > 0$ a rate or 'inverse scale' parameter. $X \in (0,\infty)$	$= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\mathbb{E} = \frac{\alpha}{\beta}$ $\mathbb{V} = \frac{\alpha}{\beta^2}$
Chi-squared $X \sim \chi^2_k$	X is a sum of the squares of k independent standard normal random variables. The χ^2 distribution is used primarily in hypothesis testing. It is a special case of the gamma distribution: $\Gamma\left(\frac{k}{2}, \frac{1}{2}\right)$. $X \in (0,\infty)$	$= \frac{2^{-\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$	$\mathbb{E} = k$ $\mathbb{V} = 2k$
Logistic $X \sim Logistic(\mu,s)$	X's cumulative distribution function is the logistic function (which appears in logistic regression). It resembles the normal distribution in shape but has heavier tails (higher kurtosis). $X \in (-\infty,\infty)$	$= \frac{e^{\frac{\mu-x}{s}}}{s\left(1+e^{\frac{\mu-x}{s}}\right)^2}$	$\mathbb{E} = \mu$ $\mathbb{V} = \frac{s^2\pi^2}{3}$
Normal[10] $X \sim \mathcal{N}(\mu,\sigma^2)$	$X \in (-\infty,\infty)$	$= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$	$\mathbb{E} = \mu$ $\mathbb{V} = \sigma^2$
Student's $\mathcal{T}$ $X \sim \mathcal{T}_k$	$X \in (-\infty,\infty)$	$= \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)} \left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}$	$\mathbb{E} = 0$ for $k > 1$ $\mathbb{V} = \frac{k}{k-2}$ for $k > 2$

3.3 Conjugate prior probability distributions in Bayesian inference

In Bayesian inference, if the prior and posterior probability distributions $f(\theta)$ and $f(\theta|data)$ are in the same family, this family is called a conjugate prior distribution for the distribution that is the likelihood function $f(x|\theta)$, i.e., for the sampling model. Having an conjugate prior is convenient: it gives an algebraic expression for the posterior.

[8] Say we want a bell-shape distribution, defined by its mean and standard deviation, but that needs to be bounded to a given interval — such that the normal distribution is not appropriate. Ex: test scores data are bounded to [0-100]. Should we use a Beta distribution or a truncated normal? A. Gelman recommends using a truncated normal. I.e., using a normal, and if we get some simulated data at 104, transform them to 100. This is sort of representing the underlying process: individuals whose ability would truly take them above 100, but the test isn't able to account for such ability levels, so they get 100. He doesn't recommend using a Beta distribution, which is very abstract and does not represent any type of underlying process.

[9] The Gamma distribution is a common choice to model right-skewed continuous data. It also has a specific mean-variance relationship: the variance of the data increases with the square of the mean. It can therefore be a suitable distribution for data whose standard deviation might be approximately proportional to the mean.

[10] The normal distribution is ubiquitous in statistics notably because summary statistics (differences, regression slope estimates...) can be expressed mathematically as weighted averages of many independent samples with finite mean and variance, which the Central Limit Theorem says are approximately normally distributed (they converge to a normal distribution as the number of samples increases).

Examples:

- The Beta distribution, for a $\theta \in [0,1]$, is a conjugate prior for the Bernoulli, binomial, negative binomial and geometric distributions.

  Ex:

  $$\begin{cases} \text{sampling model:} & Y_i|\theta \sim \text{binomial}(n,p) \\ \text{prior:} & \theta \sim \text{Beta}(a,b) \end{cases} \implies \text{posterior:} \quad \theta \mid Y_i = y_i \sim \text{Beta}(a+y_i, b+n-y_i)$$

  – Generalization: The Dirichlet distribution, for a vector of probabilities that must sum to 1, is a conjugate prior for the categorical and multinomial distributions;

- The Gamma distribution, for a rate (inverse scale) parameter, is a conjugate prior for a Poisson or exponential distribution.

  Ex:

  $$\begin{cases} \text{sampling model and prior:} \\ Y_i|\theta \sim \text{Poisson}(\theta) \\ \theta \sim \text{Gamma}(a,b) \end{cases} \implies \begin{cases} \text{posterior predictive distrib and posterior:} \\ \tilde{Y}_i \mid Y_i = y_i \sim \text{NB}(a+\sum y_i, b+n) \\ \theta \mid Y_i = y_i \sim \text{Gamma}(a+\sum y_i, b+n) \end{cases}$$

- The Gamma distribution, for a precision (inverse variance) parameter, is a conjugate prior for a normal distribution with known mean.

  Ex:

  $$\begin{cases} \text{sampling model and priors:} \\ Y_i \mid \theta, \sigma^2 \sim \mathcal{N}(\theta, \sigma^2) \\ \theta \mid \sigma^2 \sim \mathcal{N}\left(\mu_0, \dfrac{\sigma^2}{\kappa_0}\right) \\ \dfrac{1}{\sigma^2} \sim \text{Gamma}\left(\dfrac{\nu_0}{2}, \dfrac{\nu_0\sigma_0^2}{2}\right) \end{cases} \implies \begin{cases} \text{posteriors:} \\ \theta \mid y_i, \sigma^2 \sim \mathcal{N}\left(\mu_n, \dfrac{\sigma^2}{\kappa_n}\right) \\ \dfrac{1}{\sigma^2} \mid y_i \sim \text{Gamma}\left(\dfrac{\nu_n}{2}, \dfrac{\nu_n\sigma_n^2}{2}\right) \end{cases}$$

  – Generalization: The Wishart distribution, for a symmetric non-negative definite matrix, is a conjugate prior for a multivariate normal distribution (specifically: for the inverse of its covariance matrix).

4 Convergence theorems (Probability Theory)

Many estimators and tests statistics are made of sample averages. We are therefore interested in how sequences of sample averages behave as a sample size $n \to \infty$. Two theorems come into play:

- **Laws of Large Numbers (LLNs)** say they converge in probability;
- **Central Limit Theorems (CLTs)** say they converge in distribution.

Convergence

- Consider a sequence of real numbers: $(a_n)_{n\in\mathbb{N}} \equiv a_1, a_2, ...$, succinctly noted a_n. For example, $a_n = 4 + \frac{5}{n}$. This a_n has a limit, to which it converges with certainty: $a_n \xrightarrow[n\to\infty]{} a_\infty = 4$, $\lim_{n\to\infty} a_n = a_\infty$
- Consider a sequence of random variables: $X_n \equiv X_1, X_2, ...$ For example, a stochastic extension of a_n. This X_n also has a limit: the random variable X. As X_n is stochastic, we are only *almost certain* that it will converge to it, s.t. $\boldsymbol{X_n}$ **converges in probability to** $\boldsymbol{X}$:[a]

$$X_n \xrightarrow[n\to\infty]{p} X \iff \operatorname*{plim}_{n\to\infty} X_n = X$$

- A weaker statement is **convergence in distribution:** $X_n \xrightarrow[n\to\infty]{d} X \iff \lim_{n\to\infty} F_{X_n} = F_X$

[a] Formally, a series converges iff it will eventually be within any small distance ε of its limit:
- a_n converges to a_∞ iff $\forall \varepsilon > 0, \exists\ N$ s.t. $\forall n \geqslant N, |a_n - a_\infty| < \varepsilon$.
- X_n converges *in probability* to X iff $\forall \varepsilon > 0,\ \mathrm{P}(|X_n - X| > \varepsilon) \xrightarrow[n\to\infty]{} 0$.

Consider a sequence of random variables $(X_i)_{i\in\mathbb{N}} \equiv X_1, X_2, ...$, succinctly noted X_i. Any sample average $\bar{X}_N \equiv \frac{1}{N}(X_1 + ... + X_N)$ is also a random variable. We are interested in the sequence of sample averages $(\bar{X}_n)_{n\in\mathbb{N}} \equiv \bar{X}_{N_1}, \bar{X}_{N_2}, ...$, succinctly noted $\bar{X}_n$.[11] We distinguish three situations:[12]

a. X_i are **not independent** over i

b. X_i are **independent** over i but not identically distributed: $X_i \sim (\mu_i, \sigma_i^2)$

c. X_i are **independent and identically distributed (iid)**: $X_i \overset{\text{iid}}{\sim} (\mu, \sigma^2)$

Law of Large Numbers (LLN)

a. The average $\bar{X}_n$ of n random variables converges in probability:

$$\bar{X}_n - \mathbb{E}[\bar{X}_n] \xrightarrow[n\to\infty]{p} 0 \iff \operatorname*{plim}_{n\to\infty} \bar{X}_n = \lim_{n\to\infty} \mathbb{E}[\bar{X}_n]$$

b. If the X_i are independent, the probability limit simplifies to: $\operatorname*{plim}_{n\to\infty} \bar{X}_n = \lim_{n\to\infty} \frac{1}{n} \sum_i \mathbb{E}[X_i]$

c. If the X_i are iid: $\operatorname*{plim}_{n\to\infty} \bar{X}_n = \lim_{n\to\infty} \frac{1}{n} n\mu = \mu \iff \bar{X}_n \xrightarrow[n\to\infty]{p} \mu$

[11] For example: consider X_i the result of a coin flip, s.t. $X_i \equiv 1$ for heads and 0 for tails. The sample average $X_n \equiv \frac{1}{n}(X_1 + ... + X_n)$ is the proportion of heads in the n coin flips. Intuitively, we know it will converge *most probably* to $\frac{1}{2}$.

[12] This section draws heavily from Colin Cameron's lecture notes "Asymptotic Theory for OLS".

Central Limit Theorem (CLT)

a. The *normalized* average Z_n of n random variables is a random variable that converges to a normal distribution, *even if the original variables are not normally distributed*:

$$Z_n \equiv \frac{\bar{X}_n - \mathbb{E}[\bar{X}_n]}{\sqrt{\mathbb{V}[\bar{X}_n]}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

b. If the X_i are independent:

$$\frac{\frac{1}{n}\sum_i \left(X_i - \mathbb{E}[X_i]\right)}{\frac{1}{n}\sqrt{\sum_i \mathbb{V}[X_i]}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

c. If the X_i are iid:[a]

$$\frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

To express results in terms of $\bar{X}_n$, we say that $\bar{X}_n$ is *asymptotically normally distributed*:[b]

a. $\bar{X}_n \overset{\text{a}}{\sim} \mathcal{N}\left(\lim_{n\to\infty} \mathbb{E}[\bar{X}_n],\ \lim_{n\to\infty} \mathbb{V}[\bar{X}_n]\right)$

b. $\bar{X}_n \overset{\text{a}}{\sim} \mathcal{N}\left(\lim_{n\to\infty} \frac{1}{n}\sum_i \mathbb{E}[X_i],\ \lim_{n\to\infty} \frac{1}{n^2}\sum_i \mathbb{V}[X_i]\right)$

c. $\bar{X}_n \overset{\text{a}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

[a] In the previous example where X_i is the probability that the i-th coin flip is heads, if one flips a coin n times, the probability of getting $\frac{n}{2}$ heads will get increasingly close to a normal distribution centered on 0 as n increases.

[b] The asymptotics here correspond to a n is large enough that it's reasonable to consider the approximation, but not so large that the asymptotic variance goes to zero and makes the distribution degenerate.

Remarks:

- By an LLN, $\bar{X}_n$ has a degenerate distribution as it converges to a constant, $\mathbb{E}[\bar{X}_n]$. To apply the CLT, we first scale $\left(\bar{X}_n - \mathbb{E}[\bar{X}_n]\right)$ by its standard deviation $\sqrt{\mathbb{V}[\bar{X}_n]}$ to construct a random variable with variance 1, i.e., with a nondegenerate distribution.
- LLNs and CLTs are widely used in econometrics because extremum estimators involve averages.
  - LLNs give consistency. Ex: we can rewrite $\hat{\beta}_{\text{OLS}}$ to make two averages appear and apply LLNs:

$$\begin{aligned}\hat{\beta}_{\text{OLS}} &= \beta_0 + (X'X)^{-1}X'e \;=\; \beta_0 + \left(\tfrac{1}{n}X'X\right)^{-1}\tfrac{1}{n}X'e \\ &= \beta_0 + \underbrace{\left(\frac{1}{n}\sum_i x_i x_i'\right)^{-1}}_{\xrightarrow[n\to\infty]{p} \text{ finite, } \neq 0} \underbrace{\frac{1}{n}\sum_i x_i e_i}_{\xrightarrow[n\to\infty]{p} 0} \qquad \text{as } \mathbb{E}[x_i e_i] = 0\end{aligned}$$

– CLTs give limit distributions, after rescaling. Ex: we can center and rescale $\hat{\beta}_{\text{OLS}}$ to apply a CLT:

$$\sqrt{n}\big(\hat{\beta}_{\text{OLS}} - \beta_0\big) = \underbrace{\Big(\frac{1}{n}\sum_i x_i x_i'\Big)^{-1}}_{\xrightarrow[n\to\infty]{p}\ \text{finite},\ \neq 0} \underbrace{\frac{1}{\sqrt{n}}\sum_i x_i e_i}_{\xrightarrow[n\to\infty]{d}\ \mathcal{N}(0,\ldots)} \xrightarrow[n\to\infty]{d} \mathcal{N}\Big(0, M_{\text{X'X}}^{-1} M_{\text{X'}\Sigma\text{X}} M_{\text{X'X}}^{-1}\Big)$$