Probability distributions in regression modeling

immediate

Contents

Disclaimer: sections and lines in brown correspond to content which is very much '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.**

Note: is X a random or fixed variable?

In a regression model $y \sim f(X, e|\theta)$, e is a random variable and therefore y, a transformation of e, is itself a random variable. However, the explanatory variables X, 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's marginal or conditional distribution. The details go beyond the point of this document; we will simply make the following summary:

- In experimental studies, the researcher "sets" the values of X, therefore X are typically considered fixed, i.e., real vectors $X \in \mathbb{R}^p$.
  - model: a function f s.t. $y \sim f(X, e|\theta)$
  - estimation: in OLS, we assume $\mathbb{E}[e] = 0$ and look for $\mathbb{E}[y]$.
- In observational studies, we draw a sample $\{y, X\}$ from the population, therefore X are typically considered random.
  - model: f now gives the conditional probability: $y|X \sim f(X, e|\theta)$
  - estimation: in OLS, we assume in addition $\mathbb{E}[e|X] = 0$ and look for $\mathbb{E}[y|X]$.

In general, we will assume the context of an observational study, and therefore consider X a random variable.

2 Definitions

2.1 Random variables & probability distributions

A **random variable (r.v.)** $\boldsymbol{X : S \mapsto} \mathbb{R}$ is a function that maps the outcome of a random process in the sample space S to a number (i.e., it assigns to each point in S a number):[1] It can be:

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

Its **probability distribution** gives the probability of occurrence of each value. It is described using:

- the density or **probability function** $\boldsymbol{f_X(x) \equiv \mathrm{P}(X = x)}$; or
- the **cumulative distribution function** $\boldsymbol{F_X(x) \equiv \mathrm{P}(X{<}x)}$: the area under $f_X()$ over $]{-\infty}, x[$.

As a function on S, X should be written $X()$, defined as $X(s)$, $\forall s \in S$. In a statistical model with a sample of size N, we would write $X(i)$ or X_i, $\forall i \in N$. Typically, we will drop the notations $X(s)$ or X_i and just write X, leaving it implicit that this 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 $\mathbb{E}[X] \equiv \displaystyle\int x f_X(x)\, dx = \int x\, dF_X(x)$ is the distribution's center of mass or **mean**.

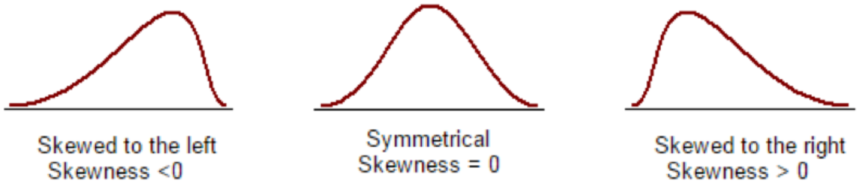
Variance $\mathbb{V}[X] \equiv \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - \mathbb{E}\,[X]^2$

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)



[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 $\mathit{Skew} = 0$, but the reverse isn't true. E.g., a distribution with one long but thin tail, and one short but fat tail, will have $\mathit{Skew} = 0$.

For X, Y:

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

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

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 indicate only 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 $\{x_i, y_i\}_i$ of n realizations of the random variables X and Y, 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	$\boldsymbol{\sigma_X^2} = \mathbb{E}\left[(X-\mathbb{E}[X])^2\right]$	$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$
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:

- central tendency: the **median** $\boldsymbol{\tilde{X}}$. 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

[5] For example: by definition, an *unbiased* estimate of σ^2 is an s^2 s.t. $\mathbb{E}[s^2] = \sigma^2$. We can show that $\mathbb{E}[\sum_{i=1}^{n}(x_i - \bar{x})^2] = ... = (n-1)\sigma^2$. So we must divide the sum by $n-1$, for it to be an *unbiased* estimate of σ^2.
This bias has to do with *degrees of freedom*. Informally: what causes this bias is using the mean calculated from the sample $\bar{x}$ instead of the mean of the population $\mathbb{E}[X]$. (Of course, we don't know the mean of the population, so we are forced to use the sample mean. But we understand that and adjust for it.) Now, why does that bias manifest itself as $n-1$ and not n? Because we lost one degree of freedom when we calculated (then used) the sample mean. Remember that every time you calculate a statistic, you 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.

[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.

- 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.

$$\begin{aligned} &\text{Population:} && z = \frac{y - \mu}{\sigma_y} \\ &\text{Sample:} && z = \frac{y - \bar{y}}{s_y}, \quad \text{where } s_y = \text{the sample standard deviation} \end{aligned}$$

2.6 Independence $\implies$ uncorrelation = orthogonality

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

- 2 random variables X and Y are **independent ($X \perp\!\!\!\perp Y$)** 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 $<u, v> = 0$
  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 $\text{cov}[X, Y] = 0$.

$$\begin{tikzcd} \text{independent} \arrow[r, Rightarrow] \arrow[d, equal] & \text{uncorrelated, orthogonal} \arrow[d, equal] \\ X \perp\!\!\!\perp Y & \text{cov}[X, Y] = 0 \end{tikzcd}$$

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** $\mathbf{P}(\boldsymbol{X}=\boldsymbol{x}\vert\boldsymbol{\theta}) = ...$	**Moments** $\mathbb{E}[\boldsymbol{X}\vert\boldsymbol{\theta}], \mathbb{V}[\boldsymbol{X}\vert\boldsymbol{\theta}]$
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. $X = 0, 1, ...$	$= \dfrac{\lambda^x e^{-\lambda}}{x!}$	$\mathbb{E} = \lambda$ $\mathbb{V} = \lambda$
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 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: $\mathbb{V}[X] > \mathbb{E}[X]$. With overdispersed count data, we will 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** $\mathbf{P(X{=}x\|\theta) = ...}$	**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 random variable 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 $X \sim \Gamma(\alpha, \beta)$	$\alpha > 0$ is a shape parameter and $\beta > 0$ a rate or 'inverse scale' parameter. X is $X \in (0, \infty)$	$= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\mathbb{E} = \frac{\alpha}{\beta}$ $\mathbb{V} = \frac{\alpha}{\beta^2}$
Chi-squared $X \sim \chi^2_k$	X is a sum of the squares of k independent standard normal random variables. The chi-square distribution is used primarily in hypothesis testing. The χ^2 distribution with k degrees of freedom is a gamma distribution $\Gamma\left(\frac{k}{2}, \frac{1}{2}\right)$ $X \in (0, \infty)$	$= \frac{2^{-\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$	$\mathbb{E} = k$ $\mathbb{V} = 2k$
Weibull			
Logistic $X \sim Logistic(\mu,s)$	Its 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 $X \sim \mathcal{N}(\mu, \sigma^2)$	$X \in (-\infty, \infty)$	$= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$\mathbb{E} = \mu$ $\mathbb{V} = \sigma^2$
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 $k > 2$

As conjugate prior distributions in Bayesian inference:

- Beta distribution, for a single probability (real number between 0 and 1); conjugate to the Bernoulli distribution and binomial distribution
- Gamma distribution, for a non-negative scaling parameter; conjugate to the rate parameter of a Poisson distribution or exponential distribution, the precision (inverse variance) of a normal distribution, etc.
- Dirichlet distribution, for a vector of probabilities that must sum to 1; conjugate to the categorical distribution and multinomial distribution; generalization of the beta distribution
- Wishart distribution, for a symmetric non-negative definite matrix; conjugate to the inverse of the covariance matrix of a multivariate normal distribution; generalization of the gamma distribution

[8] Rmk: If want a bell-shape distribution, defined by mean and sd, but need it to be bounded (i.e., have all values within a given interval, so normal distrib doesn't work). e.g. simulate test scores data, which can only be between 0-100, and want to impose their mean and sd. Do I use a Beta distrib or a truncated normal? Gelman recommends using a truncated normal. I.e. use a normal, and if get some simulated data at 104, transform them to 100. This is kind of representing the underlying process: don't want to throw out these ¿100 points but instead set them to the max. Represents 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. Doesn't recommend using a Beta, this distribution comes out of nowhere, does not represent at all any underlying process.

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: $a_n \xrightarrow[n\to\infty]{} a_\infty = 4, \quad \lim\limits_{n\to\infty} a_n = a_\infty$
- 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]

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

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

[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 $\bar{X}_n$.[9] We distinguish three situations:[10]

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:

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

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

c. If X_i iid: $\operatorname*{plim}\limits_{n\to\infty} \bar{X}_n = \lim\limits_{n\to\infty} \frac{1}{n} n\mu = \mu \quad \Longleftrightarrow \quad \bar{X}_n \xrightarrow[n\to\infty]{p} \mu$

[9] 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 $\frac{1}{2}$.

[10] 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 X_i independent:

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

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

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

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 $\big(\bar{X}_n - \mathbb{E}[\bar{X}_n]\big)$ 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:

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