TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

MSc European Economy and Business Law

Statistics Pre-course 2022

Topic 3: Random Variables

Francesca Centofanti

francesca.centofanti@students.uniroma2.eu

Outline

- Recap
- Discrete Random Variables: famous distributions
- Continuous Random Variables: famous distributions
- Introduction to Central Limit Theorem

Recap: What is a Random Variable?

- A *random variable* is any **function** from the sample space to the real numbers
- Some particular probability distributions occur often because they are useful description of certain chance phenomenon under study
- Sometimes is the experiment itself that allows us to determinate which probability distribution we have to take into account
- **Discrete Random Variable**:
  1. *Probability mass distribution*: $p_x = P(X = x) = 0 \ \forall x \in \chi$
  2. *Cumulative distribution function*•:
     $F_X(x) = P(X \leq x) = \sum_{y \leq x} P(X = y) = \sum_{y \leq x} p_y$
- **Continuous Random Variable**:
  1. *Cumulative distribution function*:
     $F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt \ \forall x \in \chi$
  2. *Probability density distribution*: $f_X(x) = \frac{dF_X(x)}{dx} \ \forall x \in \chi$

Discrete Random Variables: famous distributions

- Uniform
- Bernoulli
- Binomial
- Poisson

Discrete Random Variables: Uniform distribution

The *discrete uniform distribution* is a symmetric probability distribution whereby a finite number of values are equally likely to be observed

- Given n possible values for X, distributed uniformly, the probability to observe each value is equal to $1/n$

Example: throwing a fair die, the possible values are 1, 2, 3, 4, 5, 6, and each time the dice is thrown the probability of a given score is 1/6:

$$P(X = x) = \frac{1}{6} \qquad with \quad x = 1, 2, 3, 4, 5, 6$$

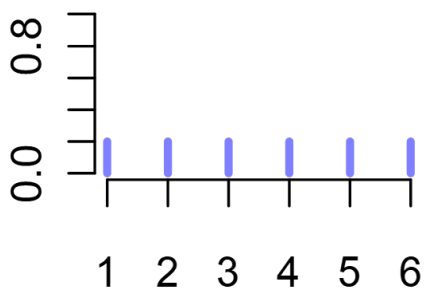
BE CAREFUL! If two dice are thrown and their values added, the resulting distribution is no longer uniform since not all sums have equal probability

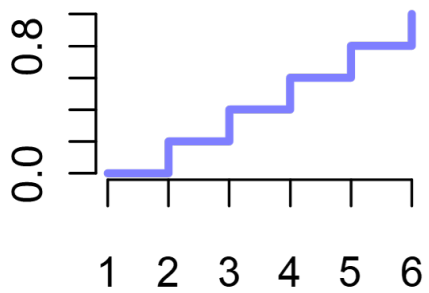
Uniform Distribution

- The discrete uniform distribution itself is inherently non-parametric. It is convenient, however, to represent its values generally by all integers in an *interval* $[a, b]$, so that a and b become the main **parameters** of the distribution
- $X \sim Uniform(a, b)$

$$p_x = P(X = x) = \frac{\lfloor x \rfloor - a + 1}{b - a + 1}$$

$$F_X(x) = P(X \leq x) = \frac{\lfloor x \rfloor - a + 1}{b - a + 1}$$





Uniform Distribution: Expected Value

- $X \sim Uniform(a, b)$
- $X \in [a, b] = [a, a + k, a + 2k, ..., b]$ where $b = a + (n - 1)k$

$$\mathbb{E}[X] = \sum_x x p_x = \sum_{l=0}^{n-1} x \frac{1}{n} = \frac{1}{n} \sum_{l=0}^{n-1} a + lk$$

$$= \frac{1}{n}[na + k \sum_{l=0}^{n-1} l] = a + \frac{k(n-1)n}{2n}$$

$$= a + \frac{k(n-1)}{2} = a + \frac{b-a}{2}$$

$$= \frac{a+b}{2}$$

Uniform Distribution: Expected Value

- **Useful expected value**: $\mathbb{E}[X^2]$
- $X \sim Uniform(a, b)$

$$\mathbb{E}[X^2] = \sum_x x^2 p_x = \frac{1}{b-a+1} \sum_{x=a}^{b} x^2$$

$$= \frac{1}{b-a+1} \left(\frac{(b^2+b)(2b+1) - (a^2-a)(2a-1)}{6} \right)$$

$$= \frac{1}{b-a+1} \left(\frac{(2b^3+3b^2+b) - (2a^3-3a^2+a)}{6} \right)$$

Uniform Distribution: Variance

- $X \sim Uniform(a, b)$
- $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

$$\mathbb{V}[X] = \frac{1}{b-a+1}\left(\frac{(2b^3+3b^2+b)-(2a^3-3a^2+a)}{6}\right) - \left(\frac{a+b}{2}\right)^2$$
$$= \frac{(b-a+1)^2-1}{12}$$

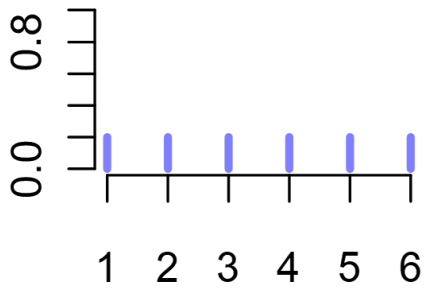
Discrete Random Variables: Bernoulli distribution

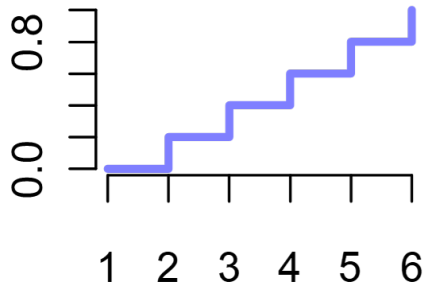
- Only two possible outcomes (mutually exclusive and exhaustive): *success* and *failure*
- Probability of **success** p is the only one parameter
- Probability of **failure** is equal to $1 - p$

Example: let $X = 1$ if the price next month of Microsoft stock goes up and $X = 0$ if the price goes down (assuming it cannot stay the same). The probability of the event "the price next month of Microsoft stock goes up" is $p = 3/5$:

$$X = \begin{cases} 1 & \textit{with probability} \quad 3/5 \\ 0 & \textit{with probability} \quad 1 - \frac{3}{5} = \quad 2/5 \end{cases}$$

Bernoulli distribution: Expected Value





- $X \sim Bernoulli(p)$

$$\begin{aligned}\mathbb{E}[X] = \sum_x x p_x &= 0 \times p_0 + 1 \times p_1 \\ &= 0 \times (1-p) + 1 \times p \\ &= p\end{aligned}$$

Bernoulli Distribution: Expected Value

- **Useful expected value**: $\mathbb{E}[X^2]$
- $X \sim Bernoulli(p)$

$$\mathbb{E}[X^2] = \sum_x x^2 p_x = 0 \times p_0 + 1 \times p_1$$
$$= 0 \times (1-p) + 1 \times p$$
$$= p$$

Bernoulli Distribution: Variance

- $X \sim Bernoulli(p)$
- $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

$$\begin{aligned} \mathbb{V}[X] &= p - p^2 \\ &= p(1-p) \end{aligned}$$

Discrete Random Variables: Binomial distribution

Given n Bernoulli trials with probability of success p, the random variable X representing the **number of successes** is called *Binomial random variable* $Bin(n, p)$ and its probability mass function is given by

$$P(X = x) = \binom{n}{x} p^k (1-p)^{n-x}$$

Example

Suppose 40% of a population supports Obama. A random sample of $n = 5$ voters is selected. Let X be the random variable representing the number of Obama supporters among $n = 5$ voters. Every time we randomly select a voter we are doing a Bernoulli trial $Y_i \sim Be(0.4)$ with $i = 1, 2, 3, 4, 5$.

- Compute $P(X = 0)$:

$$P(Y_1 = 0 \cap Y_2 = 0 \cap Y_3 = 0 Y_4 = 0 \cap Y_5 = 0) = (1 - 0.4)^5 = 0.6^5$$

- Compute $P(X = 1)$:

$$\begin{aligned} P(X = 1) = &\mathbf{(0.4)}(0.6)(0.6)(0.6)(0.6) + (0.6)\mathbf{(0.4)}(0.6)(0.6)(0.6) \\ &+ (0.6)(0.6)\mathbf{(0.4)}(0.6)(0.6) + (0.6)(0.6)(0.6)\mathbf{(0.4)}(0.6) \\ &+ (0.6)(0.6)(0.6)(0.6)\mathbf{(0.4)} = 5(0.6)^4(0.4) \end{aligned}$$

- Compute $P(X = 2)$: we have to compute the umber of possible configurations of 2 "Yes" and 3 "No"

$$\binom{5}{2} = \frac{5!}{(5-2)!2!}$$

Binomial distribution: Expected Value and Variance

- If $X \sim Bin(n, p)$
  1. $\mathbb{E}[X] = np$
  2. $\mathbb{V}[V] = np(1-p)$

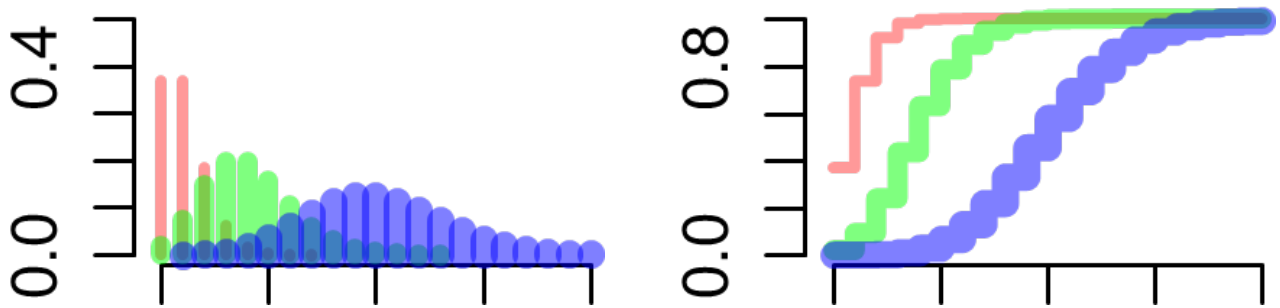
Discrete Random Variables: Poisson distribution

The *Poisson distribution* describes the number of events in a given interval of time or in a given space

Examples: # of clients calling a call-center, # of defects in a square meter of a manufactured good, # of patients arriving to the emergency hospital in the last hour

- $X \sim Poisson(\lambda)$

$$p_x = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$



Poisson distribution: Expected Value

- If $X \sim Poisson(\lambda)$

$$\begin{aligned}
\mathbb{E}[X] = \sum_x x p_x &= \sum_{x=0}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!} \\
&= \sum_{x=1}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x(x-1)!} \\
&= \sum_{x=1}^{\infty} \frac{\lambda^x e^{-\lambda}}{(x-1)!} \\
&\sum_{z=0}^{\infty} \frac{\lambda^{z+1} e^{-\lambda}}{z!} \\
&= \lambda \sum_{z=0}^{\infty} \frac{\lambda^z e^{-\lambda}}{z!}
\end{aligned}$$

Poisson distribution: Expected Value

- **Useful expected value**: $\mathbb{E}[X^2]$
- If $X \sim Poisson(\lambda)$

$$\begin{aligned}
\mathbb{E}[X^2] &= \sum_x x^2 p_x = \sum_{x=0}^{\infty} x^2 \frac{\lambda^x e^{-\lambda}}{x!} \\
&= \sum_{x=1}^{\infty} x^2 \frac{\lambda^x e^{-\lambda}}{x(x-1)!} = \sum_{x=1}^{\infty} x \frac{\lambda^x e^{-\lambda}}{(x-1)!} \\
&= \sum_{z=0}^{\infty} (z+1) \frac{\lambda^{z+1} e^{-\lambda}}{z!} = \lambda \left(\sum_{z=0}^{\infty} z \frac{\lambda^z e^{-\lambda}}{z!} + \sum_{z=0}^{\infty} \frac{\lambda^z e^{-\lambda}}{z!} \right) \\
&= \lambda(\lambda+1) = \lambda^2 + \lambda
\end{aligned}$$

Poisson distribution: Variance

- If $X \sim Poisson(\lambda)$

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= \lambda^2 + \lambda - \lambda^2 \\ &= \lambda\end{aligned}$$

Continuous Random Variables: famous distributions

- Uniform
- Exponential
- Normal (or Gaussian)
- Standard Normal

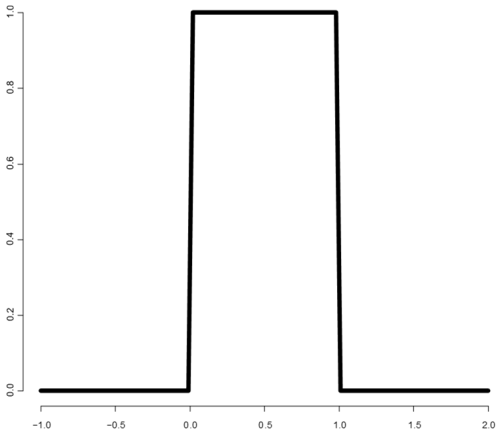
Continuous Random Variables: Uniform Distributions

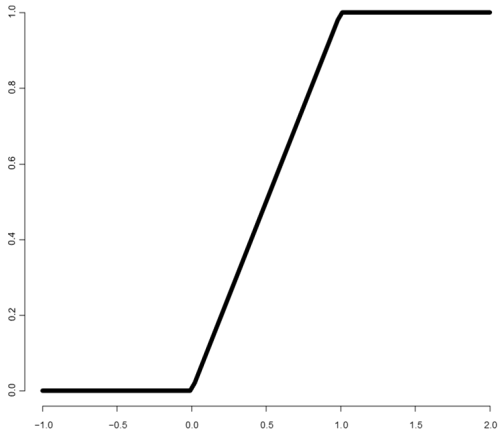
A random variable X is *uniformly distributed* between a and b, if X takes value in any interval of a given size with equal probability

- The probability of X being in an interval is proportional to the length of the interval
- If $X \sim Unif(a, b)$:
  1. $f_X(x) = \frac{1}{b-a}$
  2. $F_X(x) = \frac{x-a}{b-a}$
- Closed form expression of the c.d.f. for continuous r.v.:
  $F_X(x) = \int_a^x \frac{1}{b-a} dt = \frac{1}{b-a}(t|_a^x) = \frac{x-a}{b-a}$ with $a \leq x \leq b$
- The probability of a set only depends on its size

$$P(X \in [a_1, b_1]) = F_X(b_1) - F_X(a_1) = \frac{b_1 - a}{b - a} - \frac{a_1 - a}{b - a} = \frac{b_1 - a_1}{b - a}$$

Uniform Distributions: Expected Value and Variance





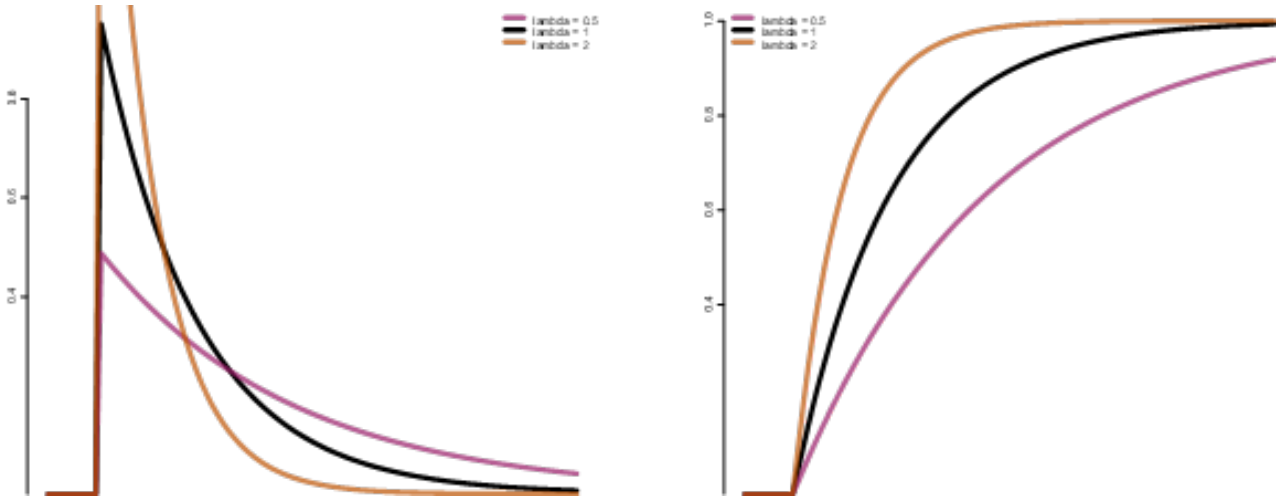
- $X \sim Unif(a, b)$
  1. $\mathbb{E}[X] = \frac{a+b}{2}$
  2. $\mathbb{V}[X] = \frac{(b-a)^2}{12}$
- **Remarks**:
  1. Since the expected value is a location/center summary, it depends on the specific values the random variable assumes
  2. Since the variance is a scale/dispersion summary, it depends only on the size of the support

Continuous Random Variables: Exponential Distributions

The *Exponential* is typically used to model time until some specific event occurs, and its *parameter* λ affects the mean time between events

- $X \sim Exp(\lambda)$, $\lambda > 0$, $x \geq 0$
  1. $f_X(x) = \lambda e^{-\lambda x}$
  2. $F_X(x) = 1 - e^{-\lambda x}$
- The larger is a value of an Exponential random variable, the less likely it is
- **Properties**:
  1. The exponential is memoryless: $P(T \geq t) = P(T \geq t + s | T \geq s)$
  2. The exponential represent the waiting time between two Poisson events

Exponential Distributions: Expected Value and Variance

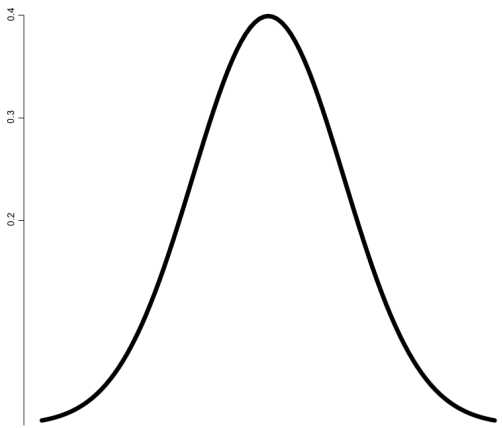


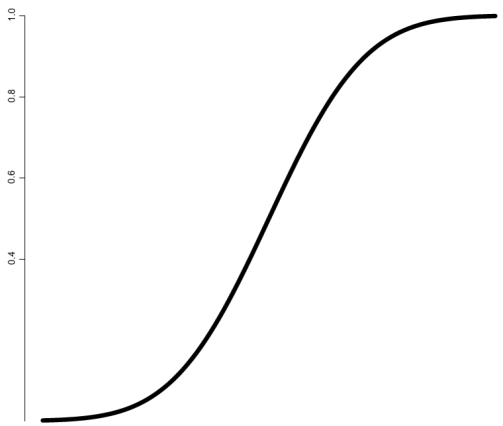
- If $X \sim Exp(\lambda)$, $\lambda > 0$
  1. Expected Value: $\mathbb{E}[X] = \frac{1}{\lambda}$
  2. Variance: $\mathbb{V}[X] = \frac{1}{\lambda^2}$

Continuous Random Variables: Normal (Gaussian) Distributions

The *Normal* or *Gaussian Distribution* is the queen of the random variables

- It represents many natural and economic phenomena
- It approximates other distributions
- It is key to inference in sampling
- $X \sim N(\mu, \sigma^2)$ with $\sigma^2 > 0$, $\mu \in \mathbb{R}$
  1. $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$
  2. $F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t-\mu)^2} dt$





Normal Distributions: Expected Value and Variance

A random variable $X \sim N(\mu, \sigma^2)$ has an interpretable parametrization:

- $\mu = \mathbb{E}[X]$
- $\sigma^2 = \mathbb{V}[X]$

Properties:

1. A linear transformation of a Normal random variable is still a Normal random variable

$$if \quad Y = aX + b, \quad where \quad a, b \in R \quad \Rightarrow Y \sim N(a\mu + b, a^2\sigma^2)$$

2. A linear combination of Normal random variables is still a Normal random variable

$$X_1, ..., X_n \ \ i.r.v. \ \ such \ \ that \ \ X_i \sim N(\mu_i, \sigma_i^2)$$

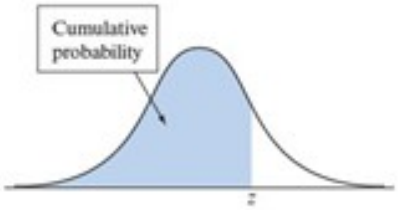
$$\Rightarrow Y = \sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right)$$

Continuous Random Variables: Standard Normal Distributions

Every Normal distribution can be turn into a *Standard Normal* by means of **standardization**

- When $\mu = 0$ and $\sigma^2 = 1$, the random variable $X \sim N(0, 1)$ is called a **Standard Normal** and it is denoted by Z:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$



Cumulative probability for z is the area under the standard normal curve to the left of z

Table A Standard Normal Cumulative Probabilities (*continued*)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549

Standard Normal Distributions: Expected Value and Variance

REMEMBER! A Standard Normal is just a *linear transformation* of a Normal, so we have:

- **Expected Value**

$$\mathbb{E}[Z] = \mathbb{E}\left[\frac{X-\mu}{\sigma}\right] = \frac{\mathbb{E}[X]-\mu}{\sigma} = \frac{\mu-\mu}{\sigma} = 0$$

- **Variance**

$$\mathbb{V}[Z] = \mathbb{V}\left[\frac{X-\mu}{\sigma}\right] = \frac{\mathbb{V}[X]}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1$$

- **Properties**:
  1. Standard Normal Table: where all the (positive) values of the cumulative distribution function of a Standard Normal are stored
  2. Relation between positive and negative values for a z standardized random variable: $P(Z < -a) = 1 - P(Z < a)$

Introduction to Central Limit Theorem

- Suppose you have $X_1, ..., X_n$ random variables independent and with the same distribution
- Identical distribution implies that all the variables have the same expected value $\mu = \mathbb{E}[X_i]$ and variance $\sigma^2 = \mathbb{V}[X_i]$
- The average of this collection is also a random variable: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Even if we don't know the distribution of $\overline{X}$, the *Central Limit Theorem* (CLT) tell us that, as $n \to \infty$,

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \to Z$$

Central Limit Theorem: Properties

- If $X_1, ..., X_n$ are already Normals, then the result of the CLT is exact, that is, it works for any n
- Even if we have no idea of what distribution generated the collection $X_1, ..., X_n$, we can always (albeit asymptotically) derive a distribution for its mean
- The CLT is very useful in **statistical inference**. We typically consider our data as realization of a collection of random variables $X_1, ..., X_n$, whose distribution we do not know; it is crucial to have a summary whose distribution we know in order to draw inferential conclusions