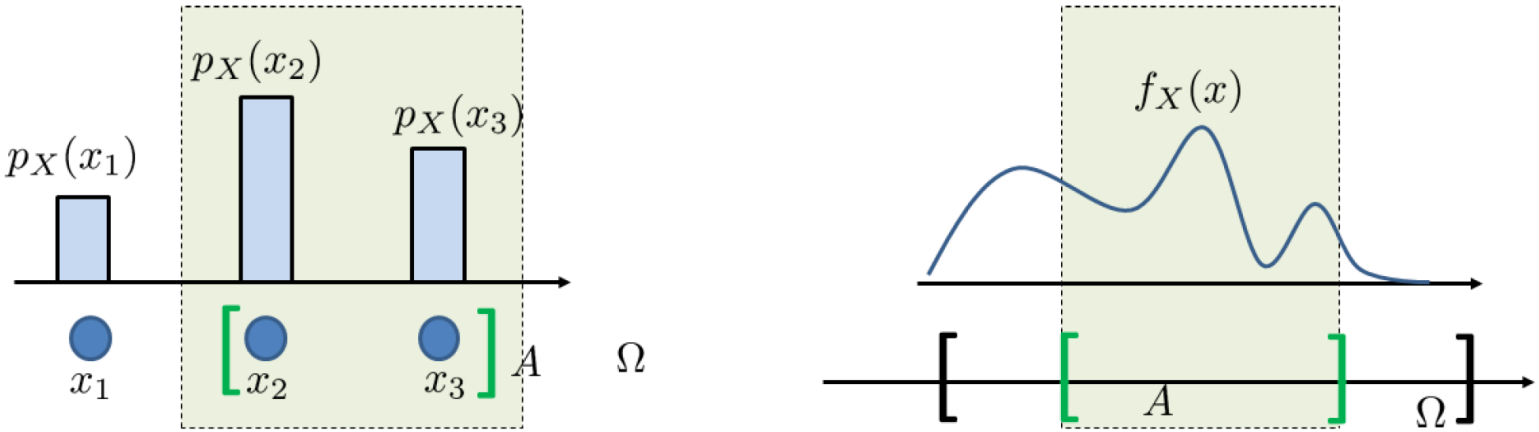
Probability & Statistics for DS & AI

Continuous Random Variables

Michele Guindani

Summer

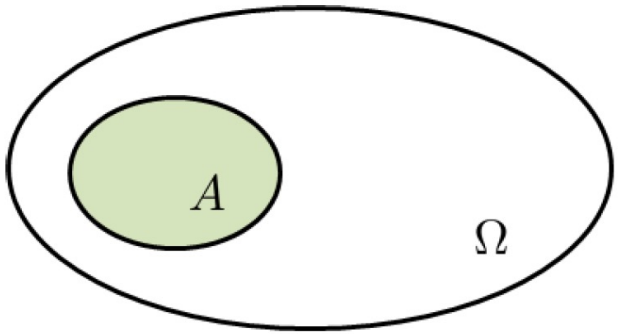
How to define probability for continuous events?

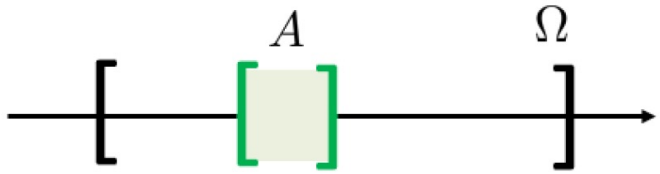


Intuition

How would you define $\mathbb{P}[\{x \in A\}]$? Measure the size of a set:

$$\mathbb{P}[\{x \in A\}] = \frac{\text{"size" of } A}{\text{"size" of } \Omega}$$

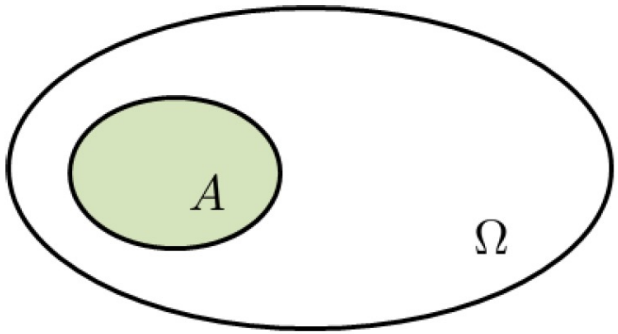


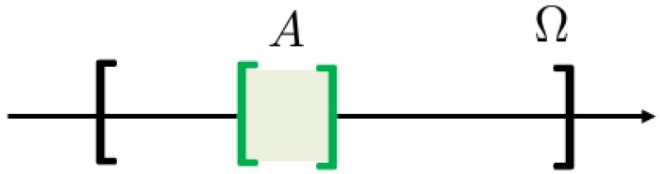


Intuition

How would you define $\mathbb{P}[\{x \in A\}]$? Measure the size of a set:

$$\mathbb{P}[\{x \in A\}] = \frac{\text{"size" of } A}{\text{"size" of } \Omega}$$





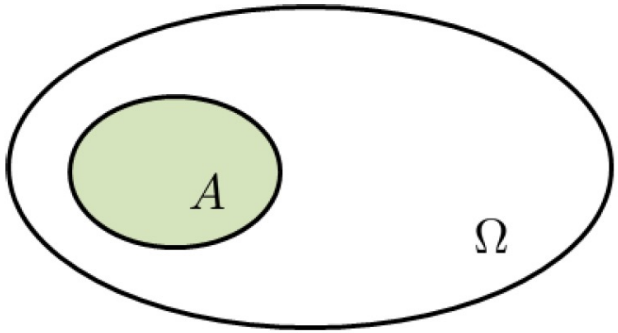
Suppose that the sample space is the interval $\Omega = [0, 5]$ and the event is $A = [2, 3]$. To measure the "size" of A, we can integrate A to determine the length. That is,

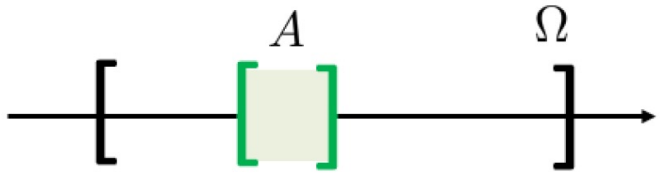
$$\mathbb{P}[\{x \in [2,3]\}] = \frac{\text{"size" of } A}{\text{"size" of } \Omega} = \frac{\int_A dx}{\int_\Omega dx} = \frac{\int_2^3 dx}{\int_0^5 dx} = \frac{1}{5}.$$

Intuition

How would you define $\mathbb{P}[\{x \in A\}]$? Measure the size of a set:

$$\mathbb{P}[\{x \in A\}] = \frac{\text{"size" of } A}{\text{"size" of } \Omega}$$





More formally,

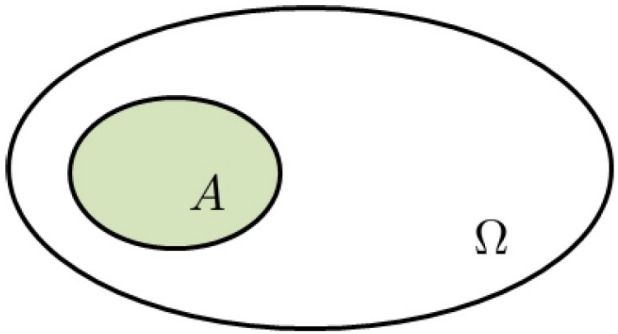
$$\mathbb{P}[\{x \in A\}] = \frac{\int_A dx}{\int_\Omega dx} = \frac{\int_A dx}{|\Omega|}$$

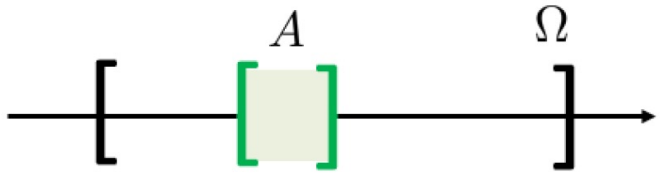
$$= \int_A \underbrace{\frac{1}{|\Omega|}}_{\text{equally important over } \Omega} dx.$$

Relax equiprobable assumption

How would you define $\mathbb{P}[\{x \in A\}]$? Measure the size of a set:

$$\mathbb{P}[\{x \in A\}] = \frac{\text{"size" of } A}{\text{"size" of } \Omega}$$



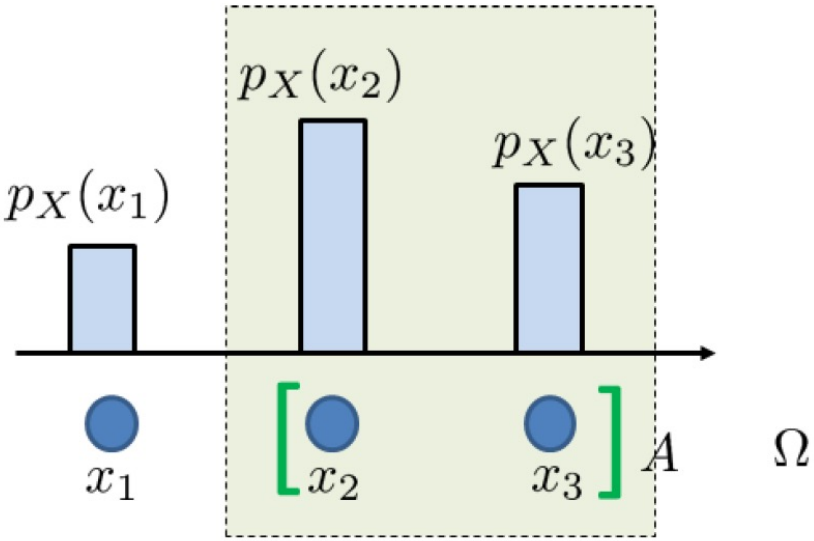


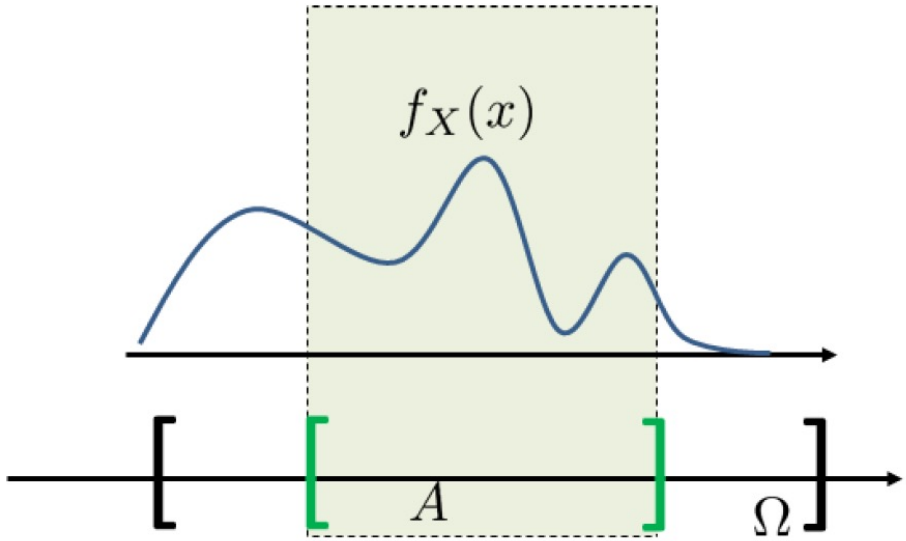
What happens if we want to relax the "equiprobable" assumption?

Replace the constant function $1/|\Omega|$ with $f_X(x)$. This will give us

$$\mathbb{P}[\{x \in A\}] = \int_A \underbrace{f_X(x)}_{\text{replace } 1/|\Omega|} dx.$$

Probability density function





A probability mass function (PMF) tells us the relative frequency of a state when computing the probability.

A probability density function (PDF) is the infinitesimal version of the PMF. Thus, the "size" of A is the integration over the PDF.

More formally...

Probability Density function (Pdf)

A probability density function f_X of a random variable X is a mapping $f_X : \Omega \to \mathbb{R}$, with the property that

- Non-negativity: $f_X(x) \geq 0$ for all $x \in \Omega$
- Unity: $\int_\Omega f_X(x)\,dx = 1$
- Measure of a set: $\mathbb{P}[\{x \in A\}] = \int_A f_X(x)\,dx$

More formally...

Probability Density function (Pdf)

A probability density function f_X of a random variable X is a mapping $f_X : \Omega \to \mathbb{R}$, with the property that

- Non-negativity: $f_X(x) \geq 0$ for all $x \in \Omega$
- Unity: $\int_\Omega f_X(x)\,dx = 1$
- Measure of a set: $\mathbb{P}[\{x \in A\}] = \int_A f_X(x)\,dx$

Probability

Let X be a continuous random variable. The probability density function (PDF) of X is a function $f_X : \Omega \to \mathbb{R}$, when integrated over an interval $[a, b]$, yields the probability of obtaining $a \leq X \leq b$:

$$\mathbb{P}[a \leq X \leq b] = \int_a^b f_X(x)\,dx$$

Example

Let $f_X(x) = 3x^2$ with $\Omega = [0, 1]$. Let $A = [0, 0.5]$. Then, the probability $\mathbb{P}[\{X \in A\}]$ is

$$\mathbb{P}[0 \leq X \leq 0.5] = \int_0^{0.5} 3x^2 \, dx = \frac{1}{8}$$

Example

Let $f_X(x) = 1/|\Omega|$ with $\Omega = [0, 5]$. Let $A = [3, 5]$. Then, the probability $\mathbb{P}[\{X \in A\}]$ is

$$\mathbb{P}[3 \leq X \leq 5] = \int_3^5 \frac{1}{|\Omega|} \, dx = \int_3^5 \frac{1}{5} \, dx = \frac{2}{5}$$

Remark

Since isolated points have zero measure in the continuous space, the probability of an open interval (a, b) is exactly the same as the probability of a closed interval:

$$\mathbb{P}[[a, b]] = \mathbb{P}[(a, b)] = \mathbb{P}[(a, b]] = \mathbb{P}[[a, b)]$$

Expectation

The expectation of a continuous random variable X is

$$\mathbb{E}[X] = \int_{\Omega} x\, f_X(x)\, dx$$

Example (Uniform random variable)

Let X be a continuous random variable with PDF $f_X(x) = \frac{1}{b-a}$ for $a \leq x \leq b$, and 0 otherwise.

Expectation

The expectation of a continuous random variable X is

$$\mathbb{E}[X] = \int_{\Omega} x\, f_X(x)\, dx$$

Example (Uniform random variable)

Let X be a continuous random variable with PDF $f_X(x) = \frac{1}{b-a}$ for $a \leq x \leq b$, and 0 otherwise.

The expectation is

$$\mathbb{E}[X] = \int_{\Omega} x f_X(x)\, dx = \int_a^b x \cdot \frac{1}{b-a}\, dx = \frac{1}{b-a} \underbrace{\int_a^b x dx}_{= \frac{x^2}{2}\Big|_a^b}$$

$$= \frac{1}{b-a} \cdot \frac{b^2 - a^2}{2} = \frac{a+b}{2}$$

Example (Exponential random variable)

Let X be a continuous random variable with PDF $f_X(x) = \lambda e^{-\lambda x}$, for $x \geq 0$.

The expectation is

$$\mathbb{E}[X] = \int_0^\infty x\lambda e^{-\lambda x} dx$$

$$= -\int_0^\infty x \, de^{-\lambda x} \quad = \underbrace{-\, xe^{-\lambda x}\big|_0^\infty}_{=0} + \int_0^\infty e^{-\lambda x} dx = -\frac{1}{\lambda} \underbrace{e^{-\lambda x}\big|_0^\infty}_{-=1} = \frac{1}{\lambda}$$

where the red step is due to integration by parts.

Function of random variables

Let $g : \Omega \to \mathbb{R}$ be a function and X be a continuous random variable, then

$$\mathbb{E}[g(X)] = \int_{\Omega} g(x)\, f_X(x)\, dx$$

Example (Uniform random variable)

Let X be a continuous random variable with $f_X(x) = \frac{1}{b-a}$ for $a \leq x \leq b$, and 0 otherwise. If $g(t) = t^2$, then

$\mathbb{E}[g(X)]$

Function of random variables

Let $g : \Omega \to \mathbb{R}$ be a function and X be a continuous random variable, then

$$\mathbb{E}[g(X)] = \int_{\Omega} g(x)\, f_X(x)\, dx$$

Example (Uniform random variable)

Let X be a continuous random variable with $f_X(x) = \frac{1}{b-a}$ for $a \leq x \leq b$, and 0 otherwise. If $g(t) = t^2$, then

$\mathbb{E}[g(X)] = \mathbb{E}\left[X^2\right]$

Function of random variables

Let $g : \Omega \to \mathbb{R}$ be a function and X be a continuous random variable, then

$$\mathbb{E}[g(X)] = \int_{\Omega} g(x)\, f_X(x)\, dx$$

Example (Uniform random variable)

Let X be a continuous random variable with $f_X(x) = \frac{1}{b-a}$ for $a \leq x \leq b$, and 0 otherwise. If $g(t) = t^2$, then

$$\mathbb{E}[g(X)] = \mathbb{E}\left[X^2\right] = \int_{\Omega} x^2 f_X(x)\, dx$$

$$= \frac{1}{b-a} \cdot \underbrace{\int_a^b x^2\, dx}_{= \frac{b^3 - a^3}{3}} = \frac{a^2 + ab + b^2}{3}$$

Interesting case

Interesting case

Example

Let $A \subseteq \Omega$. Let $\mathbb{I}_A(X)$ be an indicator function such that

$$\mathbb{I}_A(X) = \begin{cases} 1, & \text{if } X \in A \\ 0, & \text{if } X \notin A \end{cases}$$

Find $\mathbb{E}[\mathbb{I}_A(X)]$

Interesting case

Example

Let $A \subseteq \Omega$. Let $\mathbb{I}_A(X)$ be an indicator function such that

$$\mathbb{I}_A(X) = \begin{cases} 1, & \text{if } X \in A \\ 0, & \text{if } X \notin A \end{cases}$$

Find $\mathbb{E}\left[\mathbb{I}_A(X)\right]$

$$\mathbb{E}\left[\mathbb{I}_A(X)\right] = \int_{\Omega} \mathbb{I}_A(x) f_X(x)\, dx = \int_{x \in A} f_X(x)\, dx = \mathbb{P}[X \in A]$$

So the probability of $\{X \in A\}$ can be equivalently represented in terms of expectation.

Properties of Expectation

Linear Operator

- -$\mathbb{E}[aX] = a\mathbb{E}[X]$: Scalar multiple of random variable will scale the expectation.
- $\mathbb{E}[X + a] = \mathbb{E}[X] + a$: Constant addition of a random variable will offset the expectation.
- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$: Linear transformation of a random variable will translate to the expectation.

Variance of a continuous random variable

The variance of a continuous random variables X is

$$\text{Var}[X] = \mathbb{E}\left[(X - \mu)^2\right] = \int_{\Omega} (x - \mu)^2 f_X(x)\, dx$$

where $\mu \overset{\text{def}}{=} \mathbb{E}[X]$

Variance of a continuous random variable

The variance of a continuous random variables X is

$$\text{Var}[X] = \mathbb{E}\left[(X-\mu)^2\right] = \int_{\Omega} (x-\mu)^2 f_X(x)\, dx$$

where $\mu \stackrel{\text{def}}{=} \mathbb{E}[X]$

It can be shown that

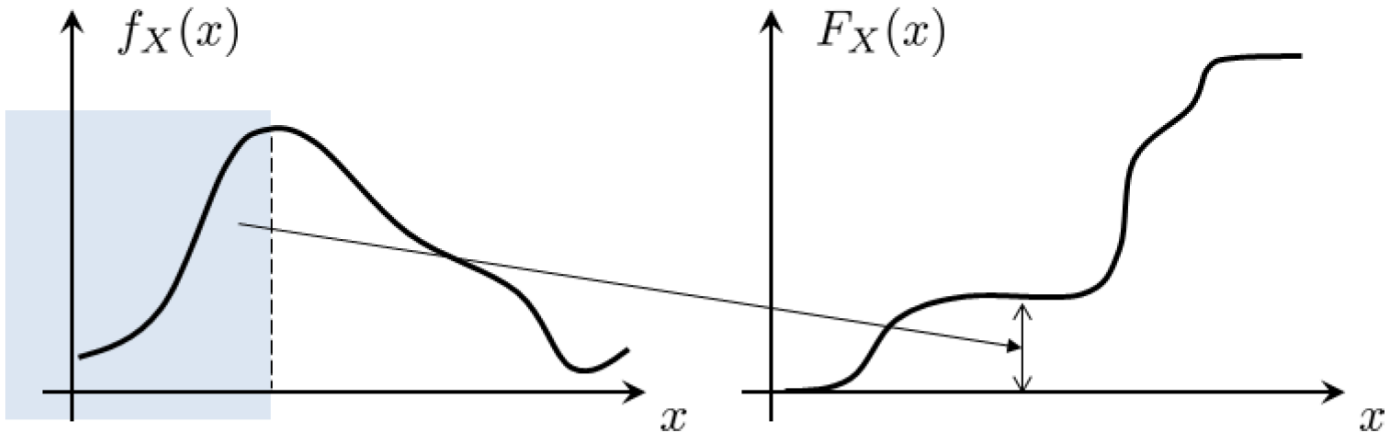
$$\text{Var}[X] = \mathbb{E}\left[X^2\right] - \mu^2$$

Cumulative Distribution Function (CDF)

- The definition of CDF is the same as before: $F_X(x) \stackrel{\text{def}}{=} \mathbb{P}[X \leq x]$.

Cumulative Distribution Function (CDF)

- The definition of CDF is the same as before: $F_X(x) \stackrel{\text{def}}{=} \mathbb{P}[X \leq x]$.

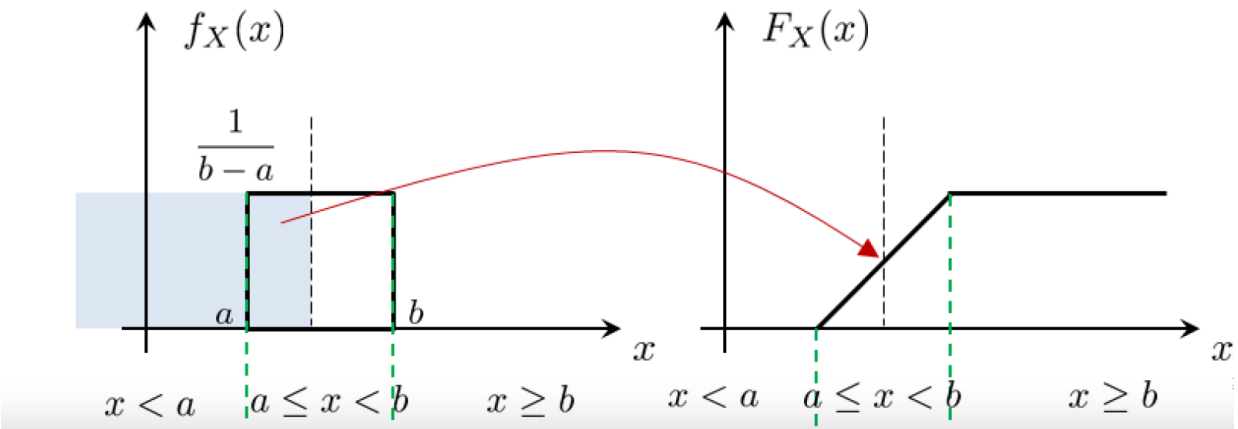


Example (Uniform random variable)

Let X be a continuous random variable with PDF $f_X(x) = \frac{1}{b-a}$ for $a \leq x \leq b$, and is 0 otherwise. Find the CDF of X.

Solution.

$$F_X(x) = \begin{cases} 0, & x \leq a, \\ \frac{x-a}{b-a}, & a < x \leq b, \\ 1, & x > b. \end{cases}$$

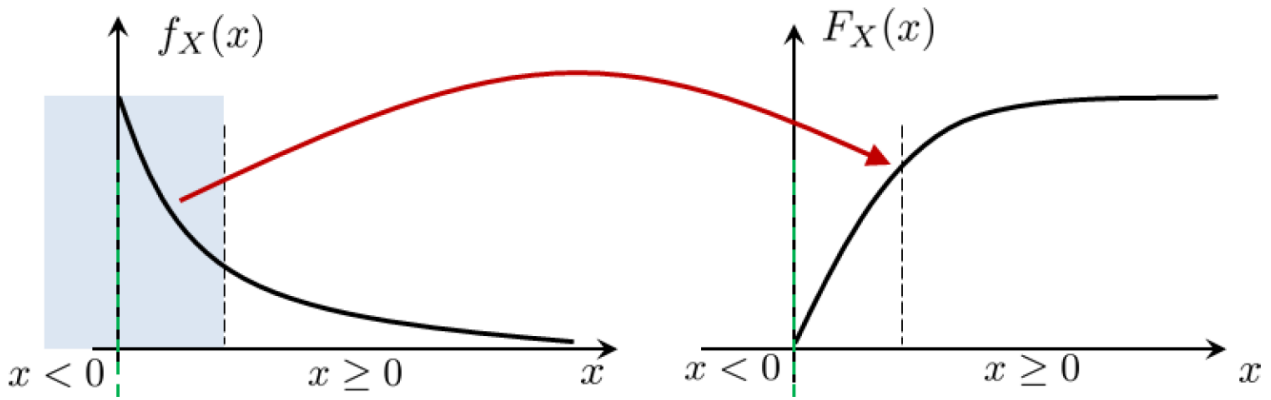


Example (Exponential random variable)

Let X be a continuous random variable with PDF $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, and is 0 otherwise. Find the CDF of X.

Solution.

$$F_X(x) = \qquad = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$$



Properties of the CDF

Property 1

Let X be a random variable (either continuous or discrete), then the CDF of X has the following properties:

(i) The CDF is a non-decreasing.

(ii) The maximum of the *CDF* is when $x = \infty : F_X(+\infty) = 1$.

(iii) The minimum of the *CDF* is when $x = -\infty : F_X(-\infty) = 0$.

Properties of the CDF

Property 1

Let X be a random variable (either continuous or discrete), then the CDF of X has the following properties:

(i) The CDF is a non-decreasing.

(ii) The maximum of the *CDF* is when $x = \infty : F_X(+\infty) = 1$.

(iii) The minimum of the *CDF* is when $x = -\infty : F_X(-\infty) = 0$.

Property 2

Let X be a continuous random variable. If the CDF F_X is continuous at any $a \leq x \leq b$, then

$$\mathbb{P}[a \leq X \leq b] = F_X(b) - F_X(a)$$

The CDF is right continuous

- For any random variable X (discrete or continuous), $F_X(x)$ is always right-continuous.

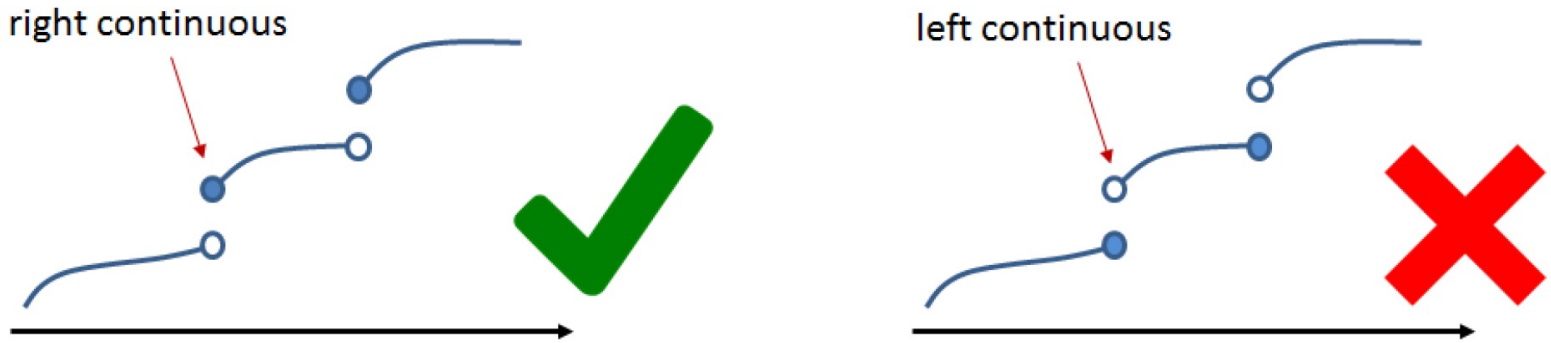


Figure: A CDF must be right continuous

The CDF is right continuous

- For any random variable X (discrete or continuous), $F_X(x)$ is always right-continuous.

Figure: A CDF must be right continuous

Retrieving PDF from CDF

The probability density function (PDF) is the derivative of the cumulative distribution function (CDF):

$$f_X(x) = \frac{dF_X(x)}{dx} = \frac{d}{dx}\int_{-\infty}^{x} f_X(x')\,dx'$$

provided F_X is differentiable at x.

If F_X is not differentiable at x, then,

$$f_X(x) = \mathbb{P}[X = x] = F_X(x) - \lim_{h \to 0^+} F_X(x - h)$$

Summary

The **cumulative distribution function (CDF)** of X is

$$F_X(x) \stackrel{\text{def}}{=} \mathbb{P}[X \leq x]$$

CDF must satisfy these **properties**:

- Non-decreasing, $F_X(-\infty) = 0$, and $F_X(\infty) = 1$.
- $\mathbb{P}[a \leq X \leq b] = F_X(b) - F_X(a)$.
- Right continuous: Solid dot on at the start.
- If discontinuous at b, then $\mathbb{P}[X = b] =$ Gap.

Relationship between CDF and PDF:

- PDF $\rightarrow$ CDF: Integration
- CDF $\rightarrow$ PDF: Differentiation

Probability & Statistics for DS & AI

Median, Mode, and Mean

Michele Guindani

Summer

Median

Given a sequence of numbers

n	1	2	3	4	5	6	7	8	9	...	100
x_n	1.5	2.5	3.1	1.1	-0.4	-4.1	0.5	2.2	-3.4	...	-1.4

🤔 How do you compute the median?

Median

Given a sequence of numbers

n	1	2	3	4	5	6	7	8	9	...	100
x_n	1.5	2.5	3.1	1.1	-0.4	-4.1	0.5	2.2	-3.4	...	-1.4

- How do you compute the median?

  Step 1: You sort the sequence (either in ascending order or descending order)

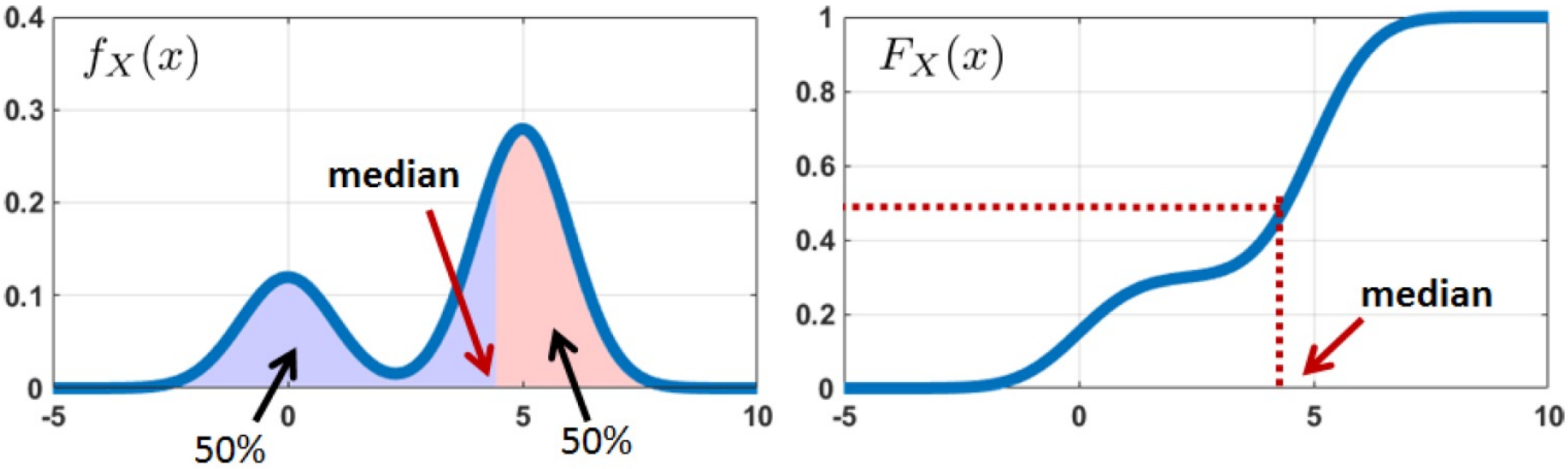
  Step 2 Pick the middle one.

If we have a random variable, what is the ideal median?

Median from a PDF

Let X be a continuous random variable with PDF f_X. The median of X is a point $c \in \mathbb{R}$ such that

$$\int_{-\infty}^{c} f_X(x)\,dx = \int_{c}^{\infty} f_X(x)\,dx$$



[Left] The median is computed as the point such that the two areas under the curve are equal. [Right] The median is computed as the point such that F_X hits 0.5.

Median from CDF

Theorem 4

The median of a random variable X is the point c such that

$$F_X(c) = \frac{1}{2}$$

Median from CDF

Theorem 4

The median of a random variable X is the point c such that

$$F_X(c) = \frac{1}{2}$$

Example (Uniform random variable)

(Uniform random variable) Let X be a continuous random variable with

- PDF: $f_X(x) = \frac{1}{b-a}$ for $a \leq x \leq b$, and is 0 otherwise.
- CDF: $F_X(x) = \frac{x-a}{b-a}$ for $a \leq x \leq b$

⇨ Find the median

Median from CDF

Theorem 4

The median of a random variable X is the point c such that

$$F_X(c) = \frac{1}{2}$$

Example (Uniform random variable)

(Uniform random variable) Let X be a continuous random variable with

- PDF: $f_X(x) = \frac{1}{b-a}$ for $a \leq x \leq b$, and is 0 otherwise.
- CDF: $F_X(x) = \frac{x-a}{b-a}$ for $a \leq x \leq b$

⇨ Find the median

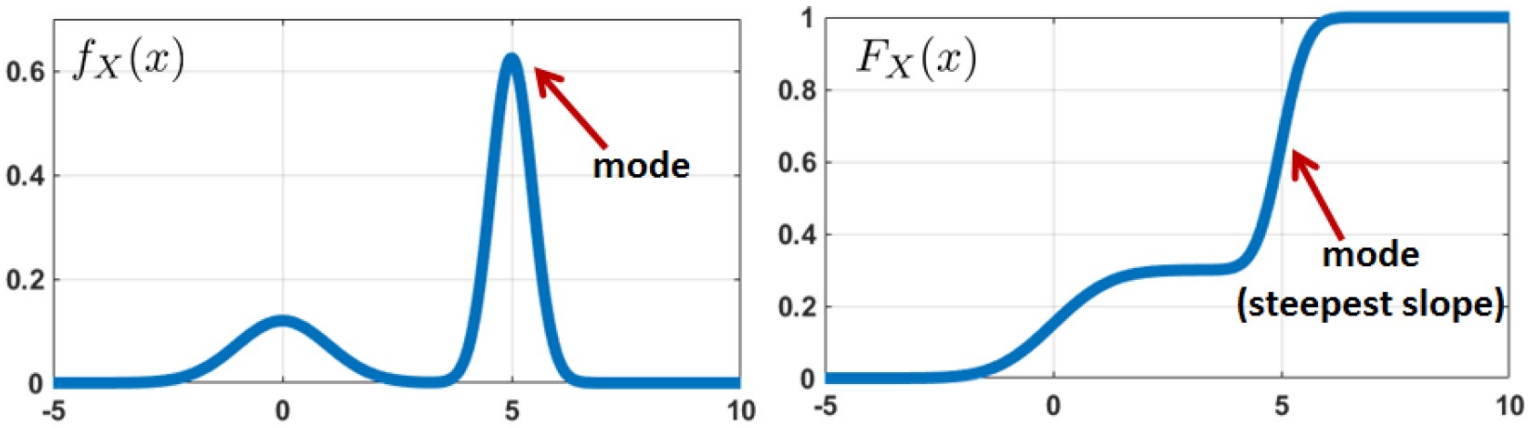
Since

$$F_X(c) = \frac{c-a}{b-a} = \frac{1}{2}c = \frac{a+b}{2}$$

Mode

Mode

The mode is the peak of the PDF, i.e. the point where the pdf $f_X(x)$ attains its maximum.



[Left] The mode appears at the peak of the PDF. [Right] The mode appears at the steepest slope of the CDF.

Probability & Statistics for DS & AI

Common models used to describe Continuous Random Variables

Michele Guindani

Summer

Uniform random variable

Uniform random variable

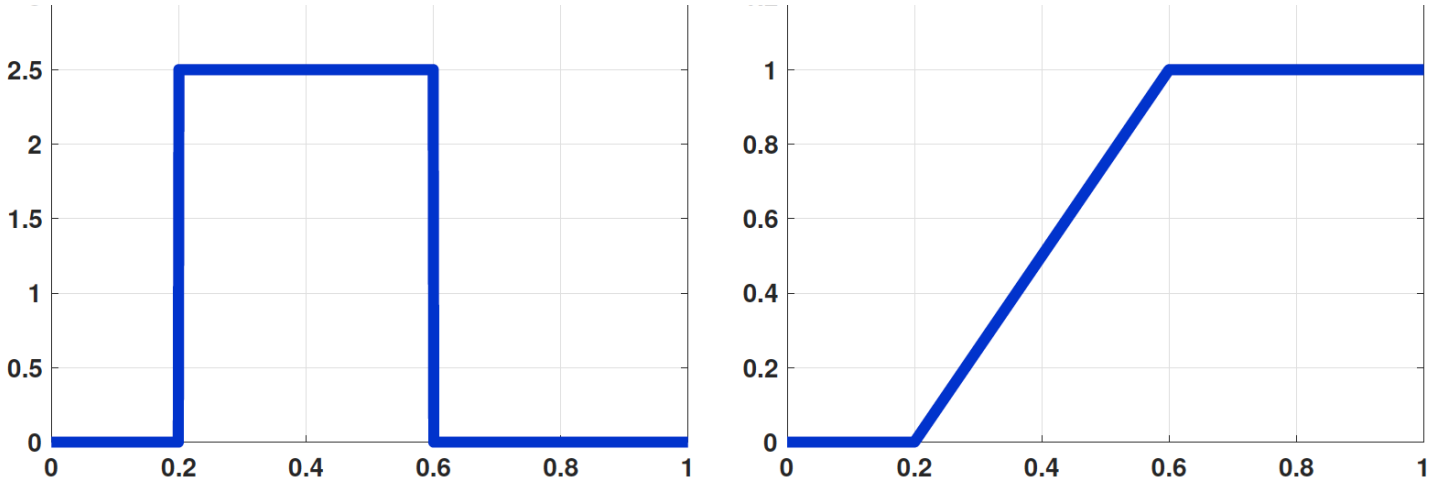
- We have already seen the pdf and CDF of a Uniform random variable

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{otherwise} \end{cases} \qquad F_X(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \leq x \leq b, \\ 1, & x > b. \end{cases}$$

Uniform random variable

- We have already seen the pdf and CDF of a Uniform random variable

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{otherwise} \end{cases} \qquad F_X(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \leq x \leq b, \\ 1, & x > b. \end{cases}$$



(a) PDF (b) CDF

The PDF and CDF of $X \sim \text{Uniform}(0.2, 0.6)$.

$$\mathbb{E}[X] = \frac{a+b}{2}, \quad \text{and } \text{Var}[X] = \frac{(b-a)^2}{12}.$$

Ideal vs Empirical pdf

- Similar ideas as the ones we discussed for discrete r.v.'s apply also for continuous r.v.'s and their pdf's

Ideal vs Empirical pdf

- Similar ideas as the ones we discussed for discrete r.v.'s apply also for continuous r.v.'s and their pdf's
- In real life, we do not observe the pdf, but we observe data $x_1, \ldots, x_n$ that we can use to estimate quantities of interest (statistics) of the underlying distribution.

Ideal vs Empirical pdf

- Similar ideas as the ones we discussed for discrete r.v.'s apply also for continuous r.v.'s and their pdf's
- In real life, we do not observe the pdf, but we observe data $x_1, \ldots, x_n$ that we can use to estimate quantities of interest (statistics) of the underlying distribution.

Example

- Consider $X \sim \text{Uniform}(0.2, 0.6)$. Then $E(X) = 0.4$.

Ideal vs Empirical pdf

- Similar ideas as the ones we discussed for discrete r.v.'s apply also for continuous r.v.'s and their pdf's
- In real life, we do not observe the pdf, but we observe data $x_1, \ldots, x_n$ that we can use to estimate quantities of interest (statistics) of the underlying distribution.

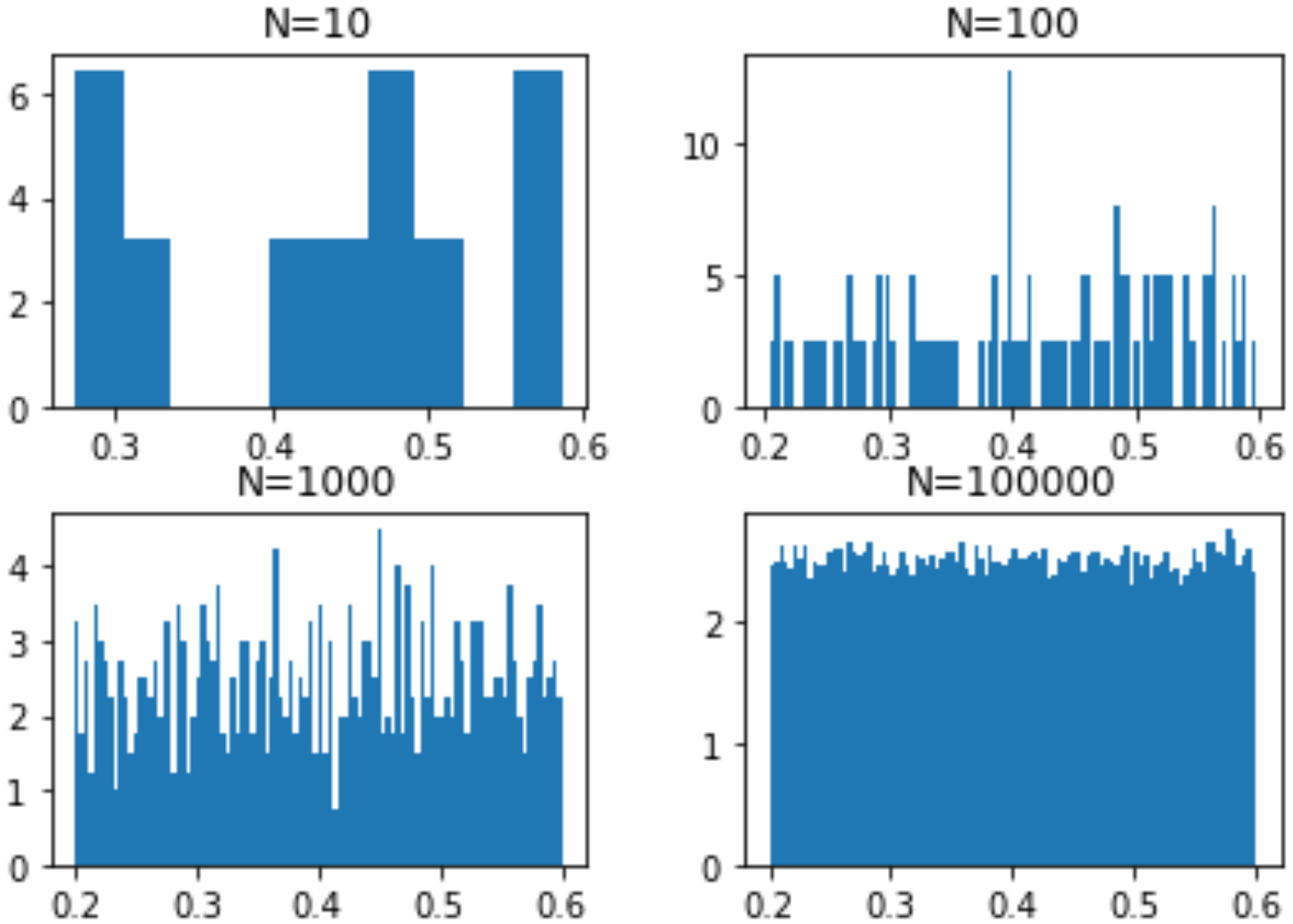
Example

- Consider $X \sim \text{Uniform}(0.2, 0.6)$. Then $E(X) = 0.4$.
- However, what we observe in reality is a histogram.

  For example, suppose we observe the following 10 data values

$$0.57184644, 0.32655022, 0.27356752, 0.28182411,$$
$$0.42709001, 0.43821788, 0.58580581, 0.46127084,$$
$$0.49956266, 0.46142795$$

The sample mean $\bar{X}$ is ≈ 0.4327. This is an estimate of the true mean, but of course it does not equal the true (unknown) mean.



With increasing sample size, we get a better idea of the true underlying pdf (and related quantities)

```
import numpy as np
from scipy.stats import uniform
import matplotlib.pyplot as plt

np.random.seed(12345)
a = 0.2; b = 0.4; #this means: (0.2, 0.2+0.4)=(0.2, 0.6)

# Initialise the subplot function using number of rows and columns
figure, axis = plt.subplots(2, 2)
plt.tight_layout()

X_10 = uniform.rvs(loc=a,scale=b,size=10)
print(np.mean(X_10),'\t ', np.var(X_10))
print(X_10)
axis[0, 0].hist(X_10, bins=10, density=True);
axis[0, 0].set_title("N=10")
# 0.4327163437873138 3          0.010833169202909975

X_100 = uniform.rvs(loc=a,scale=b,size=100)
print(np.mean(X_100),'\t ', np.var(X_100))
axis[0, 1].hist(X_100, bins=100, density=True);
axis[0, 1].set_title("N=100")
#0.4171989514782543          0.012729194655870637
```

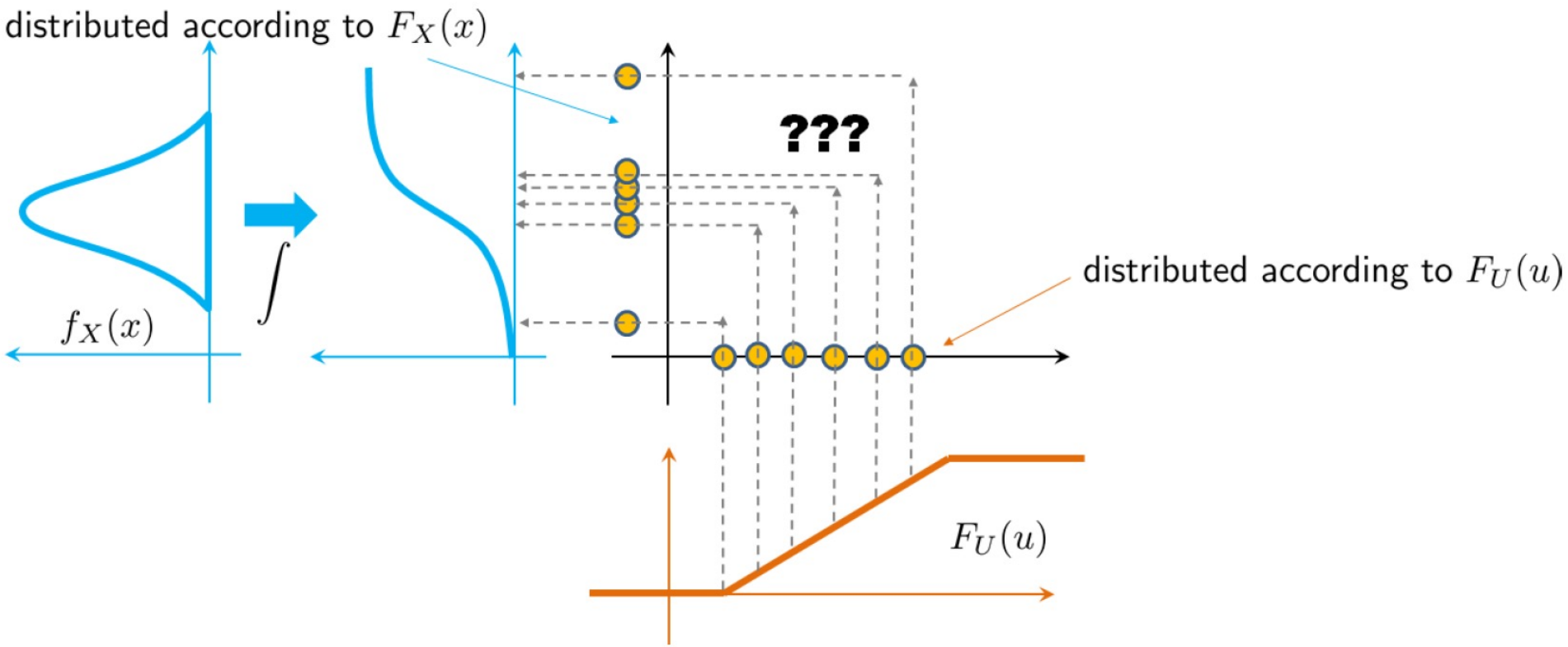
```
X_1000 = uniform.rvs(loc=a,scale=b,size=1000)
print(np.mean(X_1000),'\t ', np.var(X_1000))
axis[1, 0].hist(X_1000, bins=100, density=True);
axis[1, 0].set_title("N=1000")
# 0.402741435150772            0.013278004212539888

X_100000 = uniform.rvs(loc=a,scale=b,size=100000)
print(np.mean(X_100000),'\t ', np.var(X_100000))
axis[1, 1].hist(X_100000, bins=100, density=True);
axis[1, 1].set_title("N=100000")
#0.40005394638188996            0.013369295679455 38

# Combine all the operations and display
plt.show()
```

Applications of Uniform distribution

- Generation of Random numbers



Generating random number according to a known CDF. The idea is to first generate a uniform(0,1) random variable, and then do an inverse mapping F_X^{-1}.

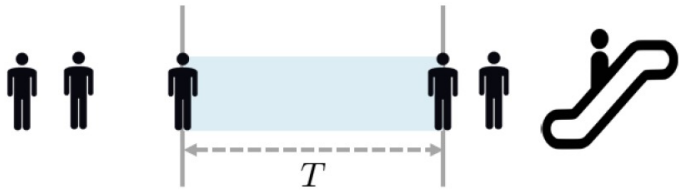
Exponential random variable

Exponential random variable

Example (Energy Efficient Escalator)

Many airports today have installed variable speed escalators.
These escalators change their speeds according to the traffic. If there are no passengers for more than a certain period (say, 60 seconds), the escalator will switch from the full-speed mode to the low-speed mode.
For moderately busy escalators, the variable-speed configuration can save energy.

Can we predict the amount of energy savings?



The variable-speed escalator problem. [Left] We model the passengers as independent Poisson arrivals. Thus, the inter-arrival time is exponential. [Right] A hypothetical passenger arrival rate (number of people per minute), from 06:00 to 23:00.

Exponential Random Variable

Definition

Let X be an exponential random variable. The PDF of X is

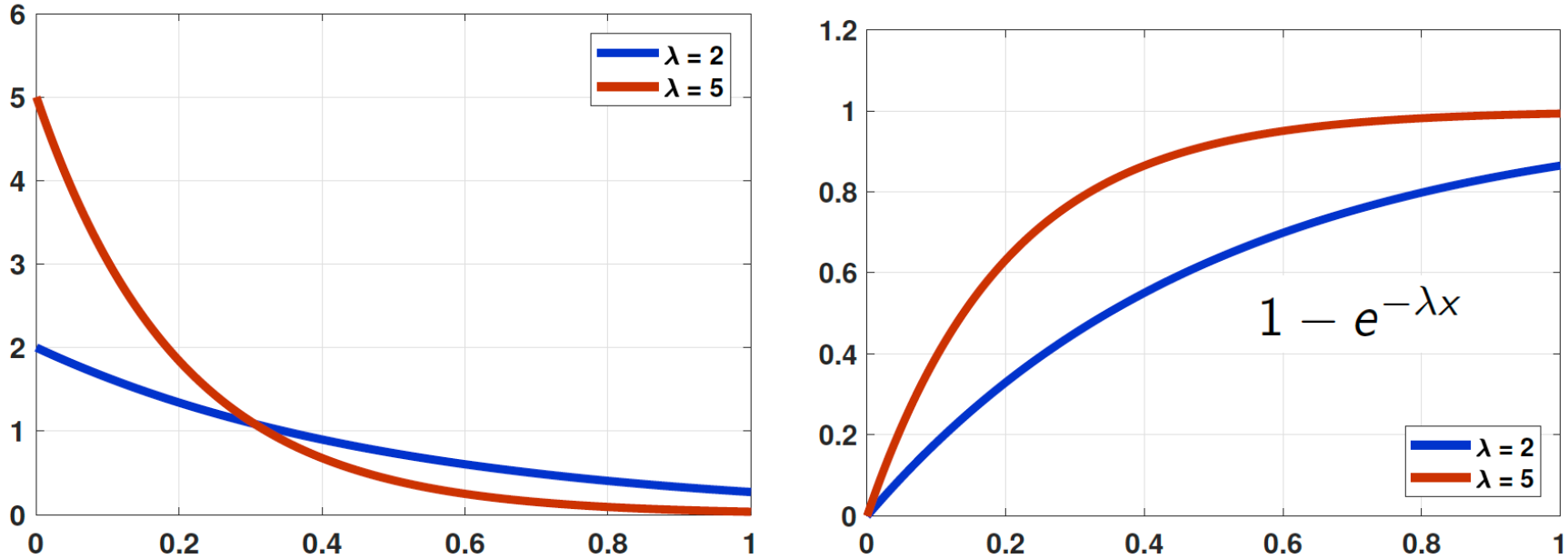
$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

where $\lambda > 0$ is a parameter. We write

$$X \sim \text{Exponential}(\lambda)$$

to say that X is drawn from an exponential distribution of parameter λ.

PDF



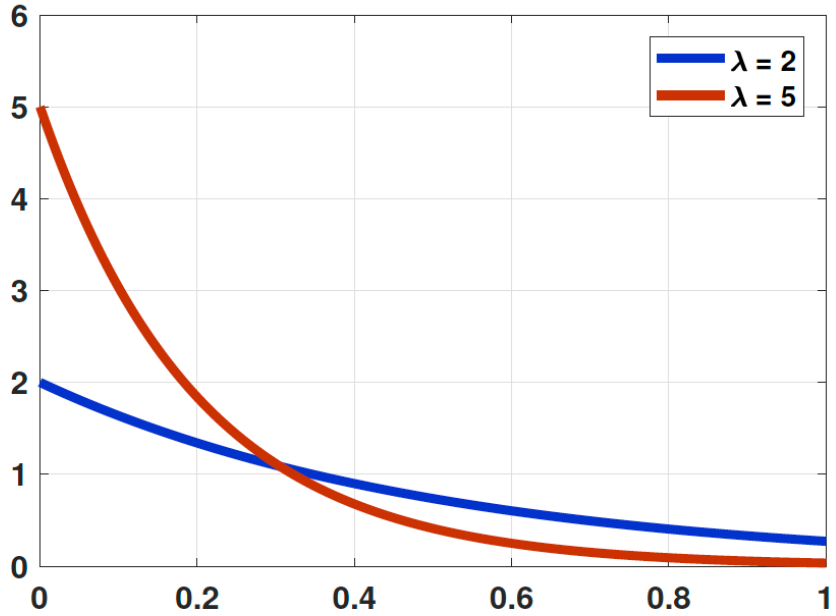
(a) PDF (b) CDF

Figure: The PDF and CDF of $X \sim \text{Exponential}(\lambda)$.

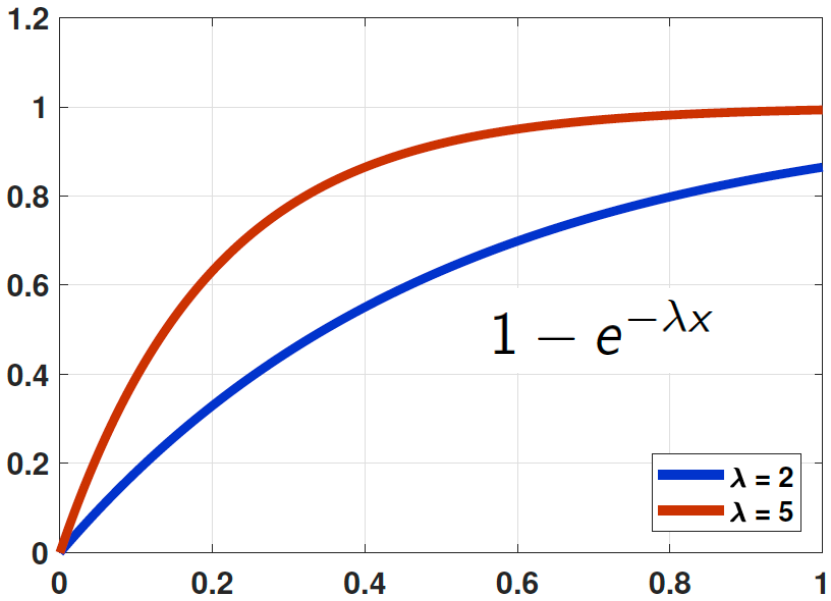
Note that the initial value $f_X(0)$ is

$$f_X(0) = \lambda e^{-\lambda 0} = \lambda.$$

Mean and Variance



(a) PDF



(b) CDF

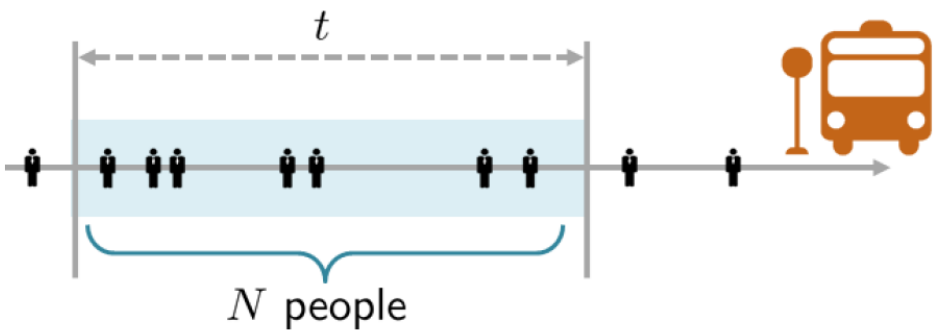
If $X \sim \text{Exponential}(\lambda)$, *then*

$$\mathbb{E}[X] = \frac{1}{\lambda}, \quad \textit{and} \quad \text{Var}[X] = \frac{1}{\lambda^2}.$$

Origin and interpretation of Exponential Random Variables

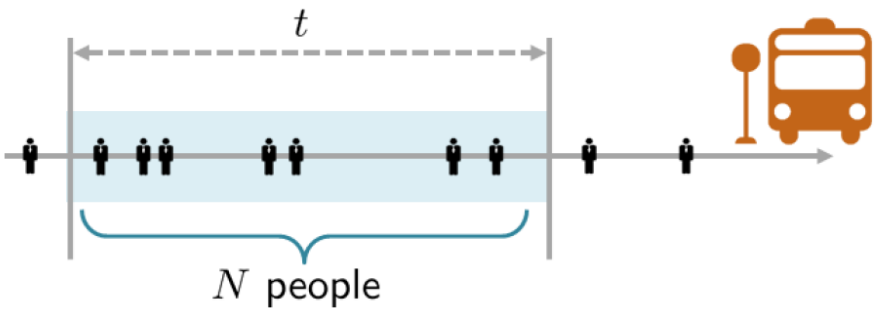
What is the origin of exponential random variables?

- An exponential random variable is the **inter-arrival time** between two consecutive Poisson events.
- That is, how much time it takes to go from N Poisson counts to $N + 1$ Poisson counts.



Question: Find the inter-arrival time between two people.

Deriving Exponential from Scratch



- Imagine that you are waiting a bus.
- The people come with an arrival rate λ per unit time.
- Thus, for a time period of t, the average number of people that arrive is λt.
- Let N be a random variable denoting the number of people. We assume that N is Poisson with a parameter λt.
- That is, for any duration t, the probability of observing n people follows the PMF

$$\mathbb{P}[N = n] = \frac{(\lambda t)^n}{n!} e^{-\lambda t}.$$

Inter-arrival Time T

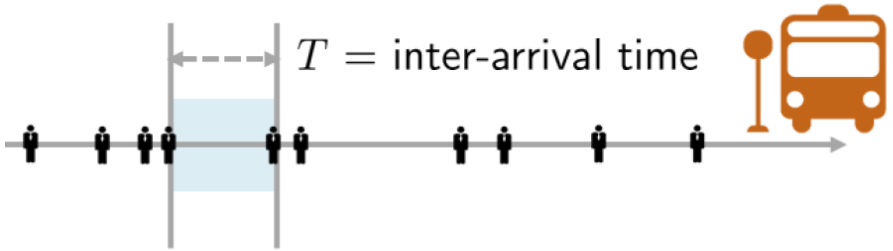


Figure: The inter-arrival time T between two consecutive Poisson events is an exponential random variable.

$$\begin{aligned}\mathbb{P}[T > t] &\overset{(a)}{=} \mathbb{P}[\text{inter-arrival time } > t] \\ &\overset{(b)}{=} \mathbb{P}[\text{no arrival in } t] \\ &\overset{(c)}{=} \mathbb{P}[N = 0] \\ &= \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t}.\end{aligned}$$

CDF and PDF of Inter-arrival Time

Since $\mathbb{P}[T > t] = 1 - F_T(t)$, where $F_T(t)$ is the CDF of T, we can show that

$$F_T(t) = 1 - e^{-\lambda t}$$

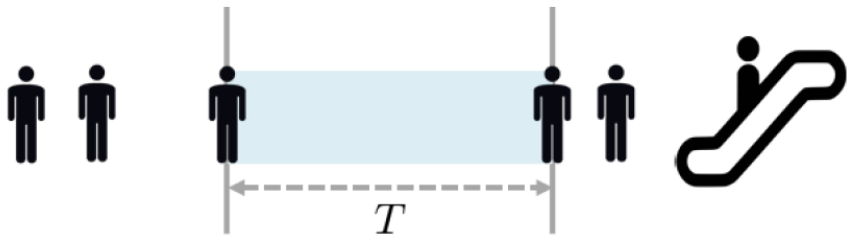
$$f_T(t) = \frac{d}{dt} F_T(t) = \lambda e^{-\lambda t}.$$

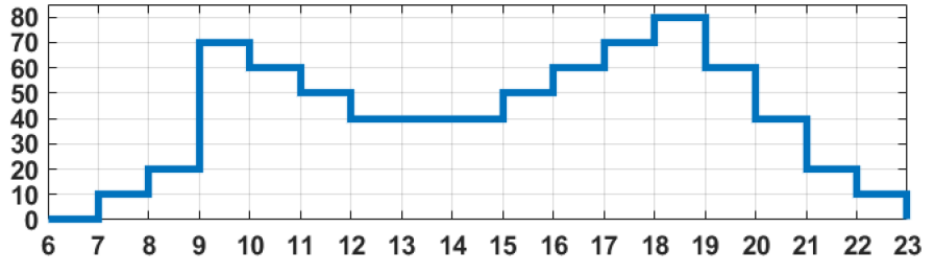
Therefore, the inter-arrival time T follows an **exponential distribution**.

Question: When to use exponential?

- It is the random variable for *time* — inter-arrival time.
- It is derived from Poisson.
- We use it to model photon arrival time, passenger arrival time, etc.
- Widely used in internet traffic, air traffic, congestion analysis.
- Also used in time-of-flight (depth sensing) cameras, LiDAR, etc.

Energy Efficient Escalator





The variable-speed escalator problem. [Left] We model the passengers as independent Poisson arrivals. Thus, the inter-arrival time is exponential. [Right] A hypothetical passenger arrival rate (number of people per minute), from 06:00 to 23:00.

Question:

- Two modes: High-speed mode, and low-speed mode.
- If no passenger arrives in more than τ seconds, then switch to low-speed mode.
- On average, how much saving will you get?

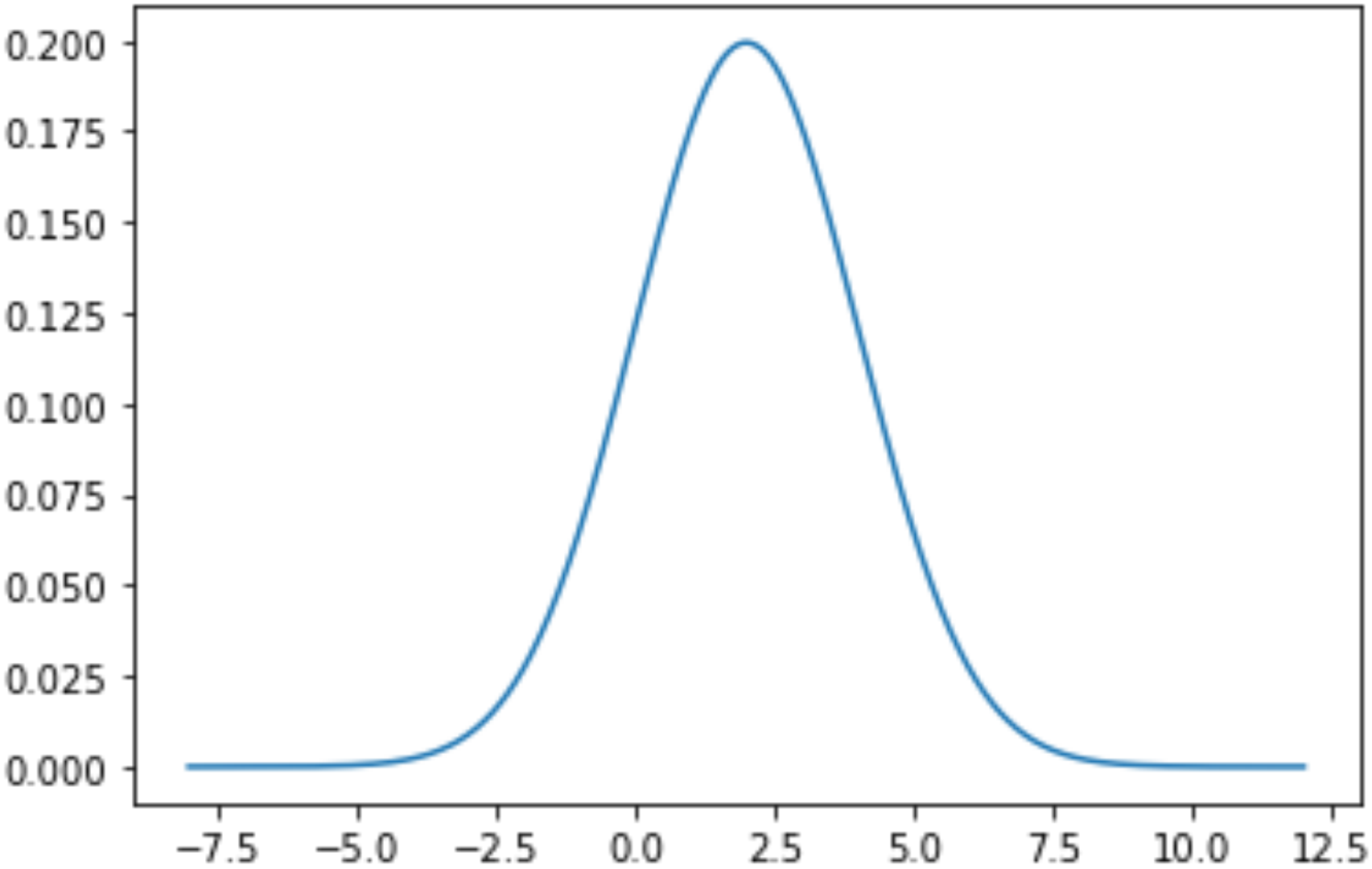
Probability & Statistics for DS & AI

Gaussian (Normal) Random variable

Michele Guindani

Summer

- One of the most used random variables, because of its applications and theorems that make it a good model for inference. Also called “Bell” curve.



- One of the most used random variables, because of its applications and theorems that make it a good model for inference. Also called "Bell" curve.

Definition

Let X be an Gaussian random variable. The PDF of X is

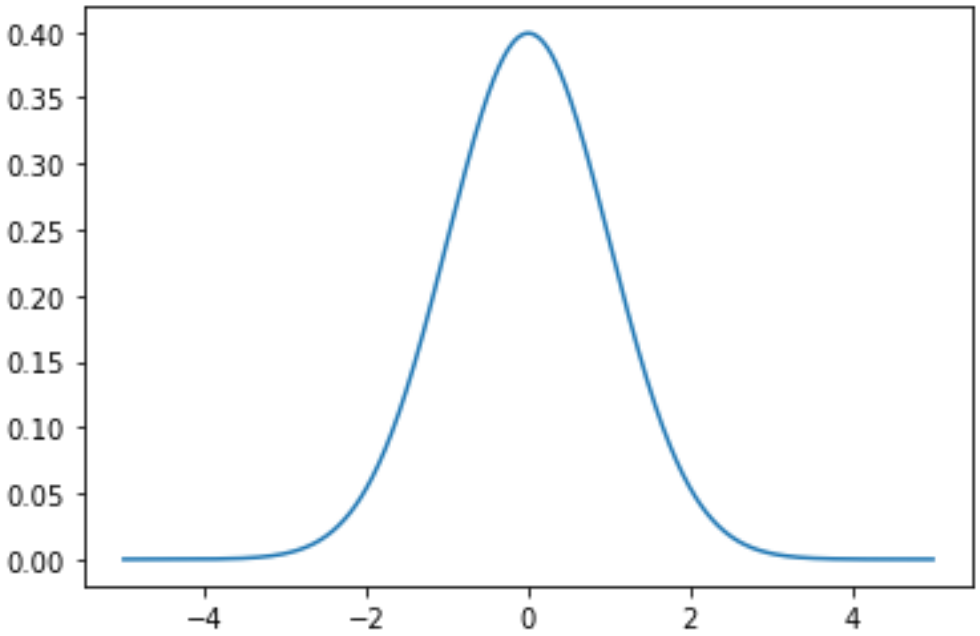
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where (μ, σ^2) are parameters of the distribution. We write

$$X \sim \text{ Gaussian } (\mu, \sigma^2) \quad \text{ or } \quad X \sim \mathcal{N}(\mu, \sigma^2)$$

to say that X is drawn from a Gaussian distribution of parameter (μ, σ^2).

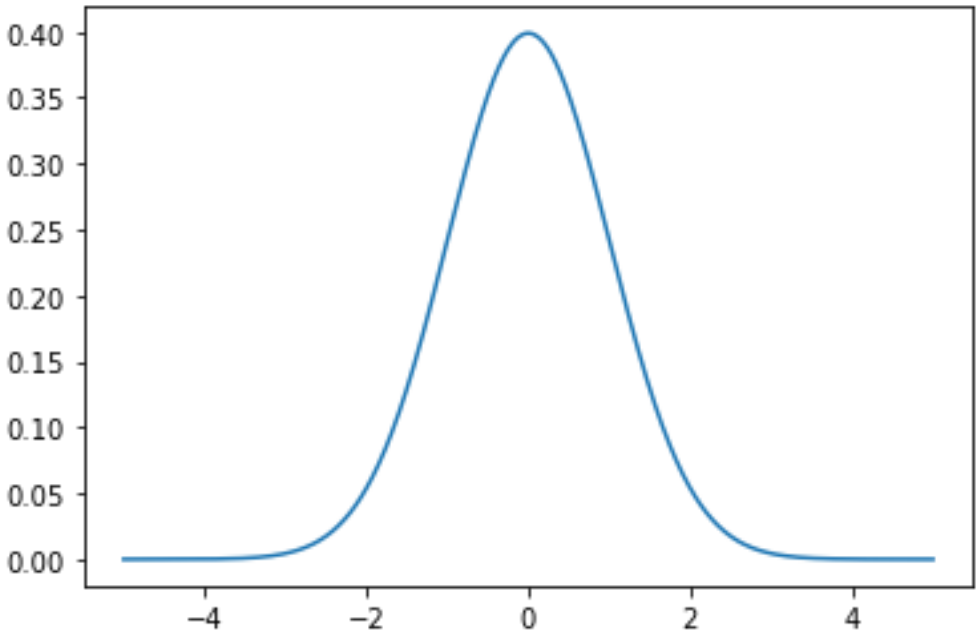
Standard Normal Distribution



Standard Normal Distribution (mean 0, variance 1)

The CDF of a Standard normal is often indicated as $\Phi(x) = Prob(X \leq x)$ and the pdf as $\phi(x) \equiv N(0, 1)$.

Standard Normal Distribution

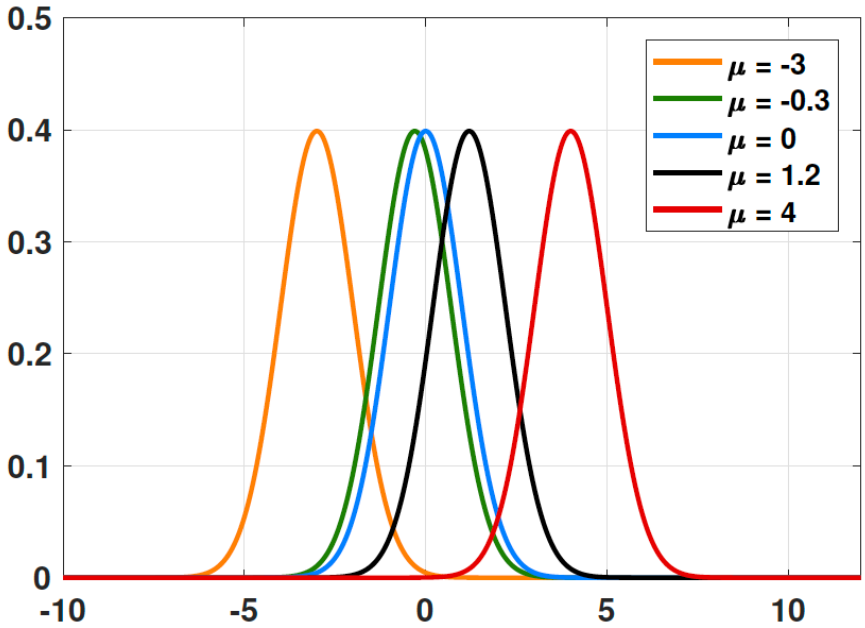


Standard Normal Distribution (mean 0, variance 1)

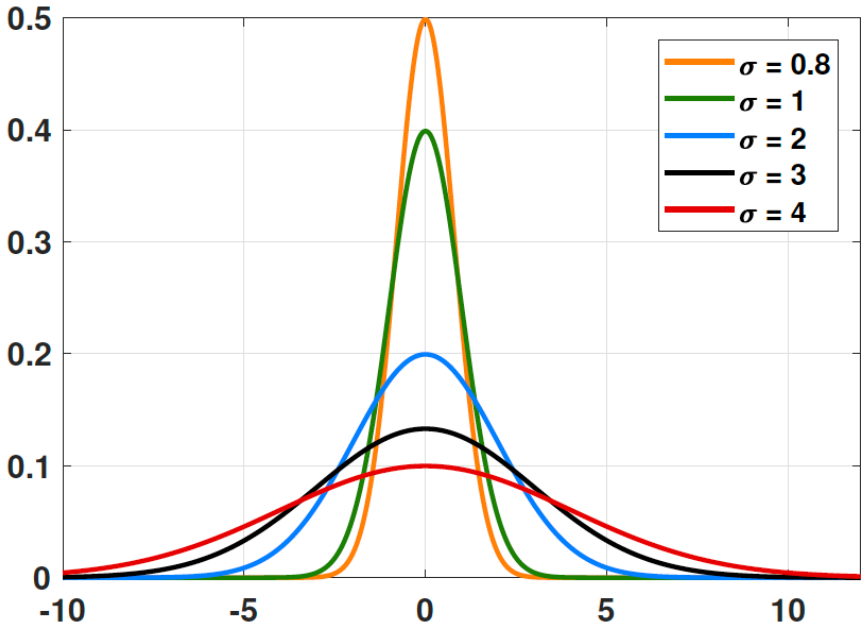
The CDF of a Standard normal is often indicated as $\Phi(x) = Prob(X \leq x)$ and the pdf as $\phi(x) \equiv N(0, 1)$.

```
# Python to generate a Gaussian PDF
import numpy as np
import matplotlib.pyplot as pltimport scipy.stats as stats
x = np.linspace(-5,5,1000)
mu = 0; sigma = 1;
f = stats.norm.pdf(x,mu,sigma)
plt.plot(x,f)
```

Interpreting the mean μ and the variance σ^2



μ changes, $\sigma = 1$



$\mu = 0$, σ changes

A Gaussian random variable with different μ and σ

Some important results

CDF of an arbitrary Gaussian from standard Gaussian CDF

Let $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$. Then

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Some important results

CDF of an arbitrary Gaussian from standard Gaussian CDF

Let $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$. Then

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

As a result,

$$\mathbb{P}[a < X \leq b] = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Some important results

CDF of an arbitrary Gaussian from standard Gaussian CDF

Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

As a result,

$$\mathbb{P}[a < X \leq b] = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

and

Additional results

$\Phi(y) = 1 - \Phi(-y)$

$\mathbb{P}[X \geq b] = 1 - \Phi\left(\frac{b-\mu}{\sigma}\right)$

$\mathbb{P}[|X| \geq b] = 1 - \Phi\left(\frac{b-\mu}{\sigma}\right) + \Phi\left(\frac{-b-\mu}{\sigma}\right)$

Uses of the Gaussian distribution in practice

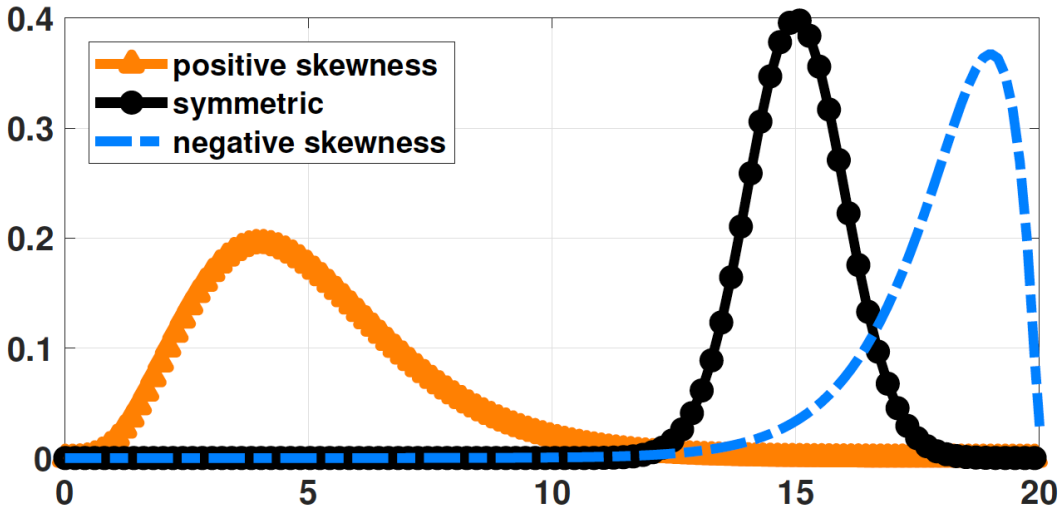
- The origins of the Gaussian distribution are to model measurement errors, or more in general the deviation between a theoretical model and actual measurements
- We expect that errors can be positive or negative, so we expect a distribution of errors to be symmetric.

Uses of the Gaussian distribution in practice

- The origins of the Gaussian distribution are to model measurement errors, or more in general the deviation between a theoretical model and actual measurements
- We expect that errors can be positive or negative, so we expect a distribution of errors to be symmetric.
- Also, many dataset with continuous measurements can be adequately represented by a Gaussian distribution [although more often than not, people force the data to be represented by a Gaussian for the lack of better alternatives]
- In practice, the Gaussian distribution is a good reference distribution for many data and purposes.

Skewness

- Skewness measures the asymmetry of a distribution with respect to the Gaussian
- Gaussian has skewness =0

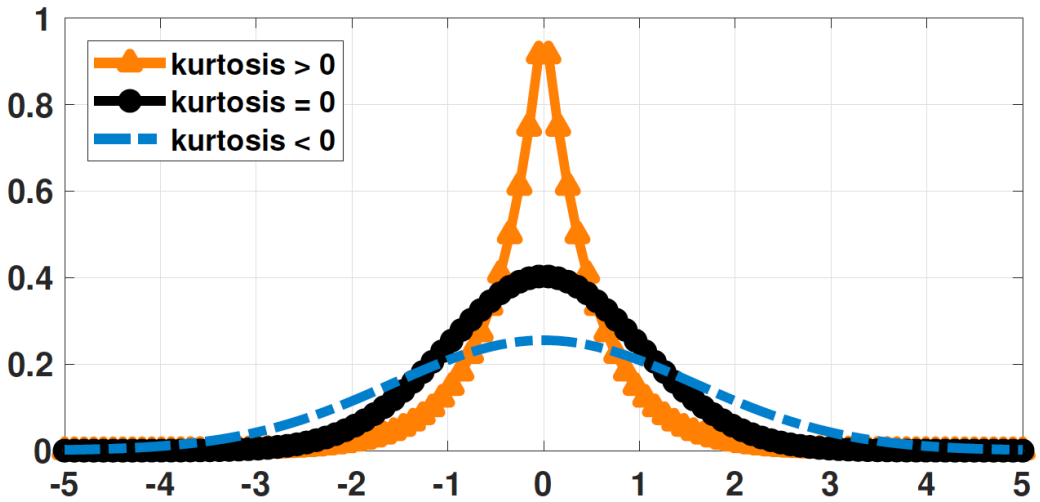


Skewness of a distribution measures the asymmetry of the distribution. In this example the skewnesses are: orange = 0.8943, black = 0, blue = -1.414.

- Skewed towards left: positive
- Skewed towards right: negative
- Symmetric: zero

Kurtosis

- Kurtosis measures how heavy-tailed the distribution is.
- There are two forms of kurtosis: one is the standard kurtosis, which is the fourth central moment, and the other is the excess kurtosis, which is $\kappa_{\text{excess}} = \kappa - 3$.
- The constant 3 comes from the kurtosis of a standard Gaussian.
- Excess kurtosis is more widely used in data analysis.

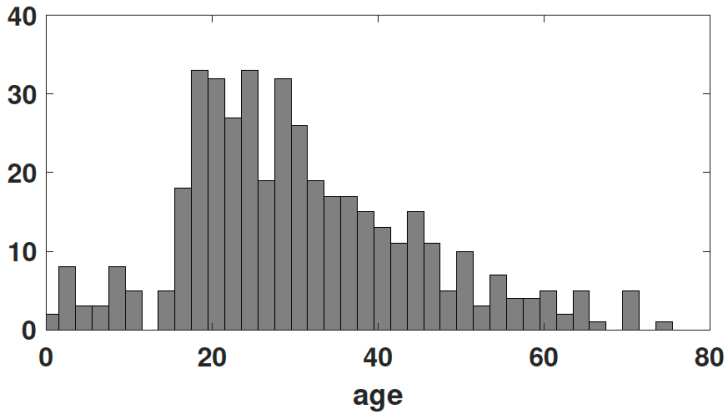


Kurtosis of a distribution measures how heavy-tailed the distribution is. In this example, the (excess) kurtoses are: orange = 2.8567, black = 0, blue = −0.1242.

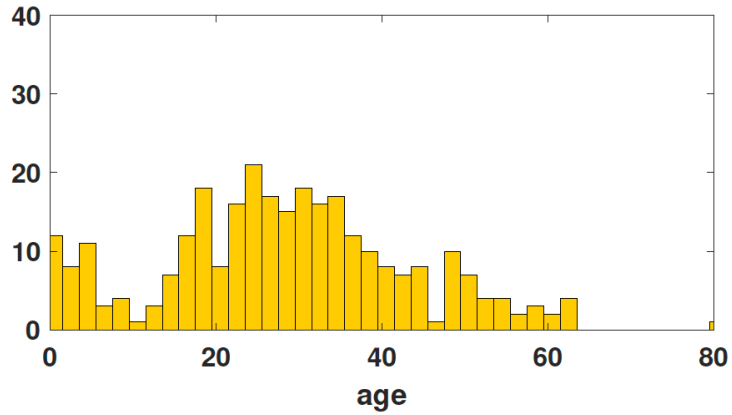
The interpretation of kurtosis is the comparison to a Gaussian.
If the kurtosis is positive, the distribution has a tail that decays faster than a Gaussian.

Beware: Often data are not Gaussian!!

- On April 15, 1912, RMS Titanic sank after hitting an iceberg. The disaster killed 1502 out of 2224 passengers and crew. A hundred years later, we want to analyze the data.
- At https://www.kaggle.com/c/titanic/ there is a dataset collecting the identities, age, gender, etc., of the passengers.
- We partition the dataset into two: one for those who died and the other one for those who survived.
- We plot the histograms of the ages of the two groups and compute several statistics of the dataset.



Group 1 (died)



Group 2 (survived)

Beware: Often data are not Gaussian!!

- Mean and standard deviation cannot tell the difference.
- Skewness and kurtosis can tell the difference.

Statistics	Group 1 (Died)	Group 2 (Survived)
Mean	30.6262	28.3437
Standard Deviation	14.1721	14.9510
Skewness	0.5835	0.1795
Excess Kurtosis	0.2652	-0.0772

So, what's the fuss about the Gaussian distribution?

- So...not all data are Gaussian (actually, quite the opposite!)

So, what's the fuss about the Gaussian distribution?

- So...not all data are Gaussian (actually, quite the opposite!)
- Then, why do we care about the Gaussian distribution so much?
- It turns out (we will see more in detail very soon) that when we see n observations, say $x_1, \ldots, x_n$, for all the probability models we have considered (and a few more) a good estimate of the population mean (and other parameters of interest) is often the sample average

$$\bar{x}_N = \frac{1}{N} \sum_{n=1}^{N} x_n.$$

So, what's the fuss about the Gaussian distribution?

- So...not all data are Gaussian (actually, quite the opposite!)
- Then, why do we care about the Gaussian distribution so much?
- It turns out (we will see more in detail very soon) that when we see n observations, say $x_1, \ldots, x_n$, for all the probability models we have considered (and a few more) a good estimate of the population mean (and other parameters of interest) is often the sample average

$$\bar{x}_N = \frac{1}{N} \sum_{n=1}^{N} x_n.$$

- In more statistical terms, we can say that the observations $x_1, \ldots, x_n$ are the result of a random draw from a common underlying population, represented by a probability model $f_X(x; \boldsymbol{\theta})$ for some vector of parameters $\boldsymbol{\theta}$.

So, what's the fuss about the Gaussian distribution?

- We often write

$$X_i \overset{i.i.d.}{\sim} f_X(x_i; \boldsymbol{\theta})$$

- It turns out that in many cases the estimates of the elements of $\boldsymbol{\theta}$ are a function of some sample statistics, e.g.

$$\bar{X}_N = \frac{1}{N} \sum_{n=1}^{N} X_n \quad \text{for the mean}$$

or

$$\sum_{n}^{N} X_n^2 \quad \text{for the variance}$$

- Note the use of capital letters (to highlight r.v.'s) vs small letters (observations) in this slide.

So, Michele, can you tell us now what the fuss is about the Gaussian distribution?

So, Michele, can you tell us now what the fuss is about the Gaussian distribution?

Central Limit Theorem

So, Michele, can you tell us now what the fuss is about the Gaussian distribution?

Central Limit Theorem

- See a proper write-up of the Theorem in Section 6.4.2 of the textbook. Here I am gong to go with a layman description

So, Michele, can you tell us now what the fuss is about the Gaussian distribution?

Central Limit Theorem

- See a proper write-up of the Theorem in Section 6.4.2 of the textbook. Here I am gong to go with a layman description
- Suppose $X_1, \ldots, X_N$ are i.i.d. random variables, with mean μ and variance σ^2. They are not necessarily Gaussians.

So, Michele, can you tell us now what the fuss is about the Gaussian distribution?

Central Limit Theorem

- See a proper write-up of the Theorem in Section 6.4.2 of the textbook. Here I am gong to go with a layman description
- Suppose $X_1, \ldots, X_N$ are i.i.d. random variables, with mean μ and variance σ^2. They are not necessarily Gaussians.
- Define the sample average as $\bar{X}_N = (1/N) \sum_{n=1}^{N} X_n$, and let $Z_N = \sqrt{N} \left(\frac{\bar{X} - \mu}{\sigma} \right)$.

So, Michele, can you tell us now what the fuss is about the Gaussian distribution?

Central Limit Theorem

- See a proper write-up of the Theorem in Section 6.4.2 of the textbook. Here I am gong to go with a layman description
- Suppose $X_1, \ldots, X_N$ are i.i.d. random variables, with mean μ and variance σ^2. They are not necessarily Gaussians.
- Define the sample average as $\bar{X}_N = (1/N) \sum_{n=1}^{N} X_n$, and let $Z_N = \sqrt{N} \left(\frac{\bar{X} - \mu}{\sigma} \right)$.
- The Central Limit Theorem says

$$Z_N \xrightarrow{d} \text{Gaussian}(0, 1)$$

Equivalently, the theorem says that $N\bar{X}_N \xrightarrow{d} \text{Gaussian}\left(N\mu, N\sigma^2\right)$. (check, error in the textbook)

- It is common to consider $\bar{X}_N \xrightarrow{d} N(\mu, \sigma^2/N)$.

And soooo?

- Suppose that we are interested in a parameter θ and the best sample statistic to estimate θ is

$$\hat{\theta} = \bar{X}_N = \frac{1}{N} \sum_{n=1}^{N} X_n$$

And soooo?

- Suppose that we are interested in a parameter θ and the best sample statistic to estimate θ is

$$\hat{\theta} = \bar{X}_N = \frac{1}{N} \sum_{n=1}^{N} X_n$$

- Then, we can easily build assessments of the uncertainty of the estimates based on the CLT and the fact that:

$$\mathbb{P}\left[a \leq \bar{X}_N \leq b\right] \approx \Phi\left(\sqrt{N}\frac{b-\mu}{\sigma}\right) - \Phi\left(\sqrt{N}\frac{a-\mu}{\sigma}\right)$$

where $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx$ is the CDF of the standard Gaussian (more later)

⚠ Note that we can do this whatever the distribution of $X_1, \ldots, X_n$ 💪

And soooo?

- Suppose that we are interested in a parameter θ and the best sample statistic to estimate θ is

$$\hat{\theta} = \bar{X}_N = \frac{1}{N} \sum_{n=1}^{N} X_n$$

- Then, we can easily build assessments of the uncertainty of the estimates based on the CLT and the fact that:

$$\mathbb{P}\left[a \leq \bar{X}_N \leq b\right] \approx \Phi\left(\sqrt{N}\frac{b-\mu}{\sigma}\right) - \Phi\left(\sqrt{N}\frac{a-\mu}{\sigma}\right)$$

where $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx$ is the CDF of the standard Gaussian (more later)

⚠ Note that we can do this whatever the distribution of $X_1, \ldots, X_n$ 💪

⇨ The CLT enables statistical inference in the frequentist paradigm 🥳 🎉

Some illustrations of the CLT

- See Sections 4.6.4 and 6.4 in your textbook for examples and python code. Also, see the lab.
- Have fun trying different populations and running simulations and these applets:
  - Central Limit Theorem for Means
  - Sampling Distributions and the Central Limit Theorem
  - Population and sample mean