Probability and Statistics for Data Science

Data Science Course

Dr. Ariel Mantzura

2025-02-03

Topics to be Covered in this Lecture

- Continuous Random Variables
- The Density function
- The Uniform Distribution
- The Exponential Distribution
- The Normal Distribution
- Calculating densities with Python.
- Expectation and Variance
- Multivariate Random Variables
- Conditioning on Random Variables

Sources: **Practical statistics for Data Scientist** Peter Bruce, Andrew Bruce & Peter Gedeck **Probability and Statistics for Data Science** Carlos Fernandez-Granda

Continuous Random Variables - Why not assign probabilty to each value?

- We cannot assign nonzero probabilities to specific outcomes of an uncertain continuous quantity.
- This would result in an infinite number of disjoint outcomes with nonzero probability.
- The sum of an infinite number of positive values is infinite, so the probability of their union would be greater than one.
- This obviously does not make sense.

Continuous Random Variables - bad example

- Assume that a random variable can get values on the interval $[0, 1]$ with equal probability.
- For each value we give a very small probability mass.
- The pmf (probability mass function) for each value is

$$P(X = x) = \epsilon$$

  where $\epsilon > 0$.
- Then we get that

$$\sum_{x \in [0,1]} P(X = x) = \infty$$

Continuous Random Variables

- Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space and

$$X : \Omega \to \mathcal{R}$$

  a random variable.
- The cumulative distribution function (cdf) of X is defined as

$$F_X(x) = P(X \le x)$$

- In words, $F_X(x)$ is the probability of X being equal or smaller than x.
- Note that the cumulative distribution function can be defined for both continuous and discrete random variables.

Continuous Random Variables - Properties

For any continuous random variable X:

$$\lim_{x \to -\infty} F_X(x) = 0$$

$$\lim_{x \to \infty} F_X(x) = 1$$

$$F_X(a) \geq F_X(b) \quad \text{if b>a}, \quad i.e \quad F_X(x) \quad \text{is non-decreasing}$$

- Hence, the probability of a random variable X belonging to an interval (a; b] is given by

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)$$

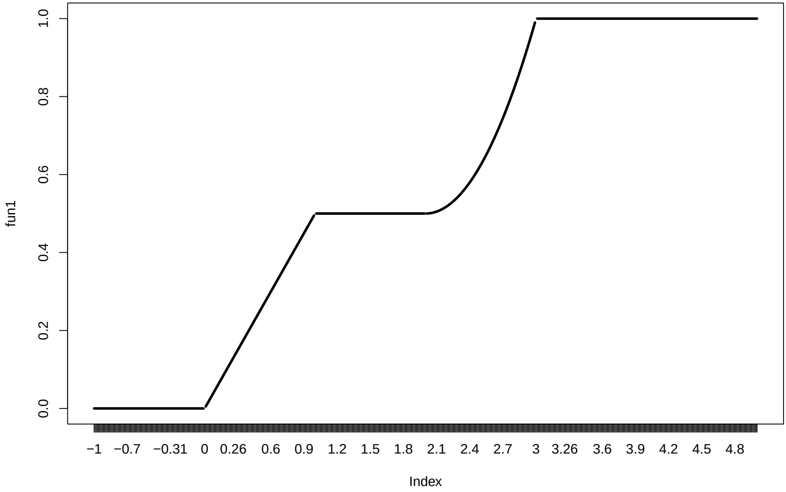
Continuous Random Variables - example

- Consider a continuous random variable X with a cdf given by:

$$F_X(x) = \begin{cases} 0 \quad if \quad x < 0 \\ 0.5x \quad if \quad 0 \le x \le 1 \\ 0.5 \quad if \quad 1 \le x \le 2 \\ 0.5(1 + (x-2)^2) \quad if \quad 2 \le x \le 3 \\ 1 \quad if \quad x > 3 \end{cases}$$

- Check that this function satisfies the properties of a cdf.
- What is the probability that X is between 0.5 and 2.5?

Continuous Random Variables - example



Probability density functions

- If the cdf of a continuous random variable is differentiable, its derivative can be interpreted as a density function.
- This density can then be integrated to obtain the probability of the random variable belonging to an interval or a union of intervals.

Probability density functions

Let $X : \Omega \to R$ be a random variable with cdf F_X. If F_X is differentiable at point x then the probability density function or pdf (probability function density) of X is defined as:

$$f_X(x) = \frac{\partial F(x)}{\partial x}$$

- The probability of a random variable X belonging to an interval is given by:

$$P(a < X \leq b) = F_X(b) - F_X(a) =$$

$$= \int_a^b f_X(x)dx$$

הטכניון
ביה"ס ללימודי המשך
של הטכניון ע"ש עזריאלי

Probability density functions - properties

- The probability density functions has 2 main properties:

$$\int_{-\infty}^{\infty} f_X(x) = 1$$

$$f_X(x) \geq 0$$

  for all $x \in \mathcal{R}$.
- The pdf is a function which must be integrated to yield a probability.
- In particular, $f_X(x)$ is not necessarily smaller than one for some point x.

Probability density functions - example

- To compute the pdf of the preivous random variable we differentiate its cdf:

$$f_X(x) = \begin{cases} 0 \quad if \quad x < 0 \\ 0.5 \quad if \quad 0 \leq x \leq 1 \\ 0 \quad if \quad 1 \leq x \leq 2 \\ x - 2 \quad if \quad 2 \leq x \leq 3 \\ 0 \quad if \quad x > 3 \end{cases}$$

- Calculate the probability that $P(0.5 < X \leq 2.5)$

Probability density functions - Uniform

- A uniform random variable models an experiment in which every outcome within a continuous interval is equally likely.
- As a result the pdf is constant over the interval [a,b].
- The pdf of a uniform random variable with domain [a, b], where b > a are real numbers, is given by:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & if \quad a \leq x \leq b \\ 0 & otherwise \end{cases}$$

Probability density functions - Uniform

- The cdf of a uniform random variable with domain [a, b] is:

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

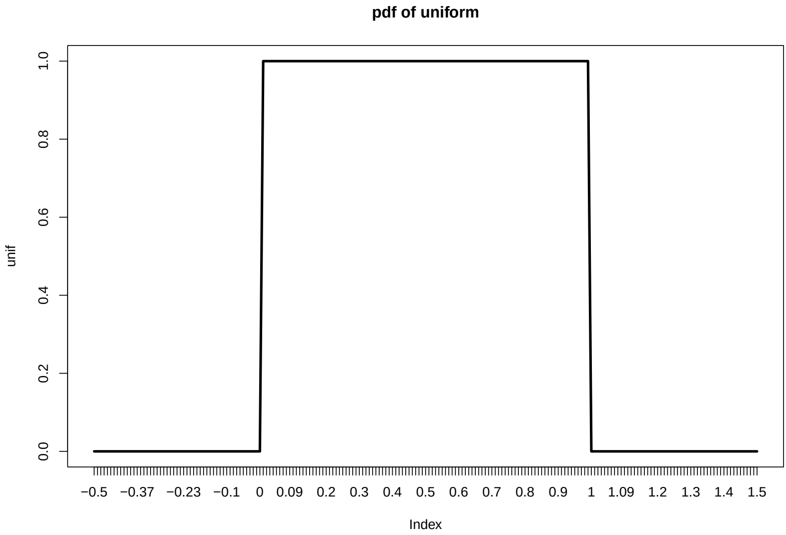
- A special case is when a=0 and b=1.
- We then get that the pdf is :

$$f_X(x) = \begin{cases} 1 \quad if \quad 0 \leq x \leq 1 \\ 0 \quad otherwise \end{cases}$$

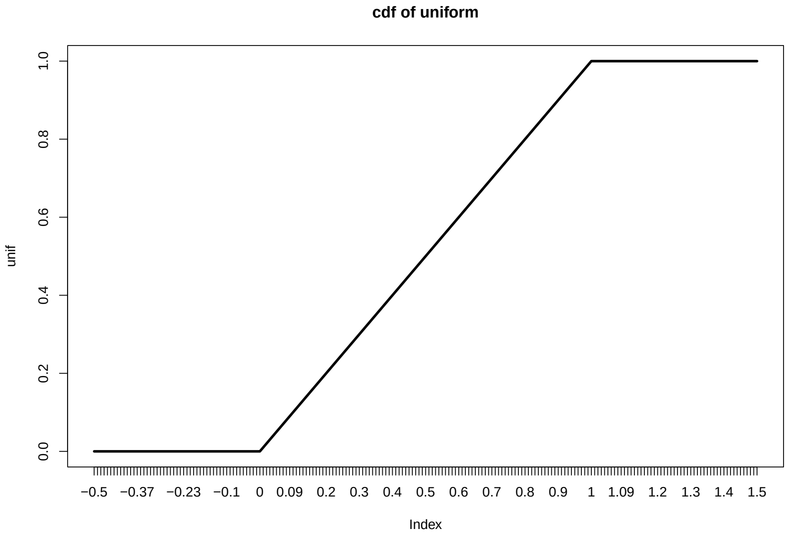
- The cdf is:

$$F_X(x) = \begin{cases} 0 \quad if \quad x < 0 \\ x \quad if \quad 0 \leq x \leq 1 \\ 1 \quad if \quad x > 1 \end{cases}$$

Probability density functions - Uniform



Probability density functions - Uniform



Probability density functions - Exponential

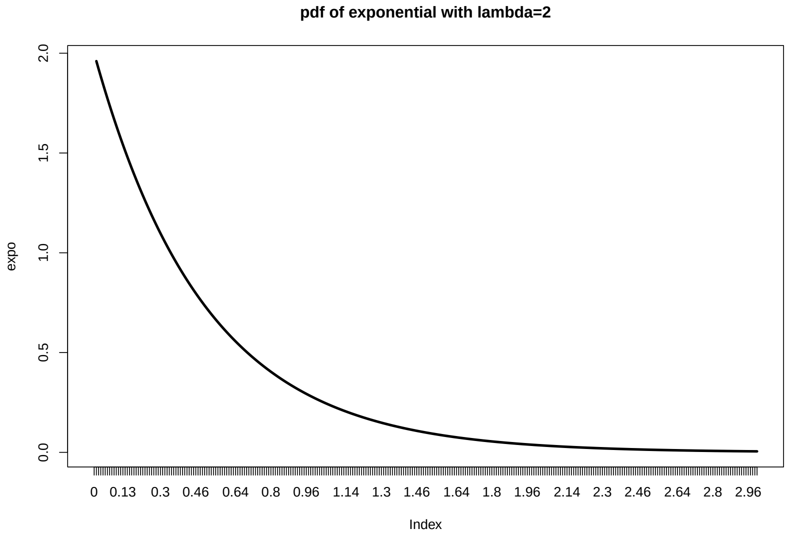
- Exponential random variables are often used to model the time that passes until a certain event.
- The pdf of an exponential random variable with parameter λ is given by:

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

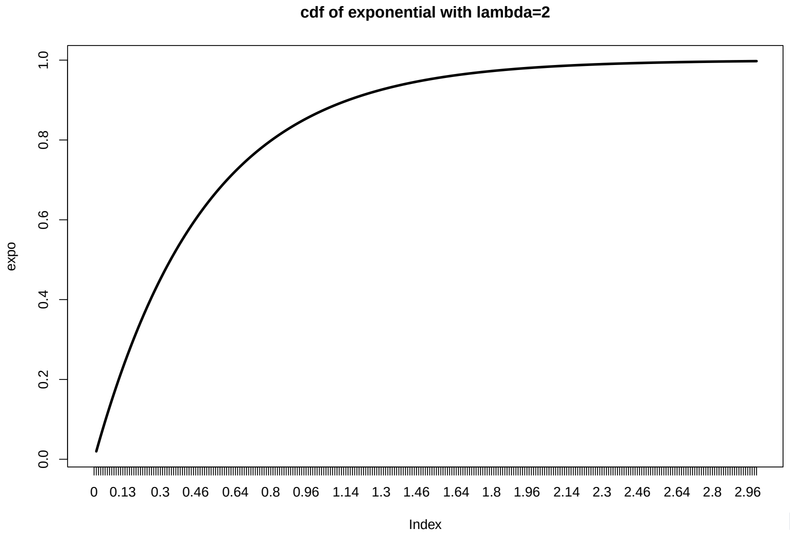
- The cdf of an exponential random variable with parameter λ is given by:

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

Probability density functions - Exponential



Probability density functions - Exponential



Probability density functions - Normal

- The Gaussian or normal random variable is the most popular random variable in all of probability and statistics.
- The pdf of the Normal random variable with $mean = \mu$ and $sd = \sigma$ is

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

- A special case is when $\mu = 0$ and $\sigma = 1$ and is called the standard normal distribution.
- Its pdf is given by:

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

Probability density functions - Normal

- The cdf of the random normal variable does not have a closed form solution.
- This complicates the task of determining the probability that a Gaussian random variable is in a certain interval.
- To mitigate this problem we use the fact that if X is a Gaussian random variable with mean μ and standard deviation σ, then

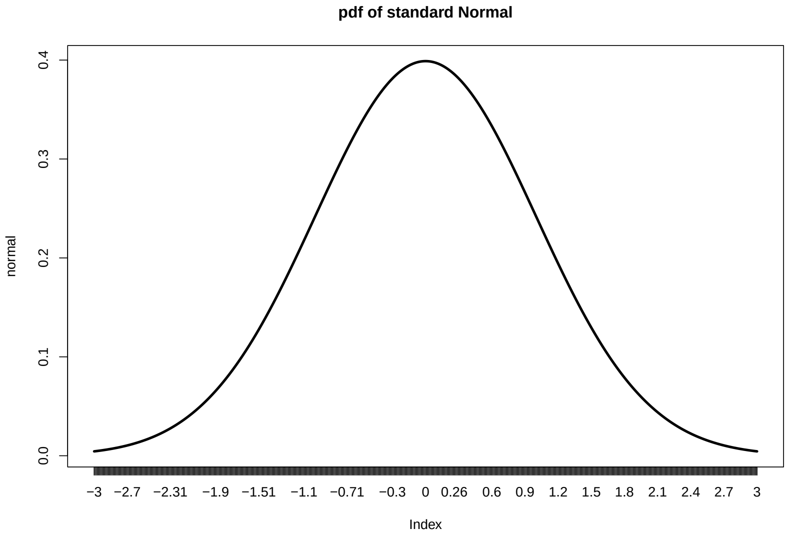
$$\frac{X - \mu}{\sigma}$$

has a standard normal distribution.

Probability density functions - Normal

- This allows us to express the probability of X being in an interval $[a, b]$ in terms of the cdf of a standard Gaussian, which we denote by Φ.
- $$P(X \in [a,b]) = P\left(\frac{x-\mu}{\sigma} \in \left[\frac{a-\mu}{\sigma}, \frac{b-\mu}{\sigma}\right]\right) =$$
$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Probability density functions - Normal



Calculating densities with Pthon - Uniform

```
from scipy.stats import uniform
pdf_value = uniform.pdf(0.75, loc=0, scale=3)
print("PDF at 0.75:", round(pdf_value,3))
```

```
## PDF at 0.75: 0.333
```

```
cdf_value = uniform.cdf(1.25, loc=0, scale=3)
print("CDF at 1.25:", round(cdf_value,3))
```

```
## CDF at 1.25: 0.417
```

```
quantile_value = uniform.ppf(0.23, loc=0, scale=3)
print("Quantile at 0.23:", round(quantile_value,3))
```

```
## Quantile at 0.23: 0.69
```

Calculating densities with Python - Exponential

```
from scipy.stats import expon
pdf_value = expon.pdf(0.5, scale=1/2)
print("PDF at 0.5:", round(pdf_value,3))
```

```
## PDF at 0.5: 0.736
```

```
cdf_value = expon.cdf(1.25, scale=1/2)
print("CDF at 1.25:", round(cdf_value,3))
```

```
## CDF at 1.25: 0.918
```

```
quantile_value = expon.ppf(0.23, scale=1/2)
print("Quantile at 0.23:", round(quantile_value,3))
```

```
## Quantile at 0.23: 0.131
```

Calculating densities with Python - Normal

```
from scipy.stats import norm
pdf_value = norm.pdf(0.5, loc=0, scale=1)
print("PDF at 0.5:", pdf_value)
```

```
## PDF at 0.5: 0.3520653267642995
```

```
cdf_value = norm.cdf(0.05, loc=0, scale=1)
print("CDF at 0.05:", cdf_value)
```

```
## CDF at 0.05: 0.5199388058383725
```

```
quantile_value = norm.ppf(0.23, loc=0, scale=1)
print("Quantile at 0.23:", quantile_value)
```

```
## Quantile at 0.23: -0.7388468491852137
```

Exercise with Python

- Assume that $X \sim uniform(0, 4)$. Calculate $P(X > 2.6)$
- Assume tht $X \sim exp(3)$. Calculate $P(0.2 < X < 0.5)$
- Assume that $X \sim N(3, 2)$. Find the value such that the probability to be larger than it is 0.05.

Expectation of continuous random variable

- Let X be a continuous random variable with a pdf f_X.
- The expectation of X is defined as:

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

- Moreover, for any function g(X) such that $g : \mathcal{R} \rightarrow \mathcal{R}$ the expectation of $g(x)$ is:

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Expectation of continuous random variable

- The expected value of x such that $X \sim unif(a,b)$ is

$$E(x) = \frac{(b+a)}{2}$$

- The expected value of x such that $X \sim exp(\lambda)$ is

$$E(x) = \frac{1}{\lambda}$$

- The expected value of x such that $X \sim N(\mu, \sigma)$ is

$$E(x) = \mu$$

Variance of continuous random variable

- Let X be a continuous random variable with a pdf f_X.
- The Variance of $X$is defined as

$$Var(X) = E([x - E(x)]^2) = \int_{-\infty}^{\infty} (x - EX)^2 fX(x)dx$$

- The variance can also be calculated as follows:

$$Var(X) = E(X^2) - (EX)^2$$

Variance of continuous random variable

- The variance of x such that $X \sim unif(a,b)$ is

$$Var(x) = \frac{(b-a)^2}{12}$$

- The variance of x such that $X \sim exp(\lambda)$ is

$$Var(x) = \frac{1}{\lambda^2}$$

- The variance of x such that $X \sim N(\mu,\sigma)$ is

$$Var(x) = \sigma^2$$

Exercise

- Create a Data frame with 4 columns: uniform,poisson, Normal,exp. and 4 rows.
- Name the rows: Average, Sample Variance, Mean, Variance.
- Simulate 100 observations from a uniform variable where $a = 5, b = 10$.
- Plot the density.
- Calculate its average and sample variance.
- Compare to the theoretical mean and variance.
- Repeat this process for poisson(5), Normal(2,4),exp(2).

Multivariate Discrete Random Variables - 2 random variables

- Probabilistic models usually include multiple random variables.
- We describe how to specify random variables to represent such quantities and their interactions.
- We will group these random variables as random vectors.
- Let X and let Y be discrete random variables on the same probability space.
- The joint probability mass function (pmf) is defined as

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

- The probability of an event $\{X, Y\} \in A$ is:

$$P(\{X, Y\} \in A) = \sum_{x,y \in A} P_{X,Y}(x, y)$$

Multivariate Discrete Random Variables - properties

- $$P_{X,Y}(x, y) \geq 0$$
- $$\sum_{k=1}^{\infty} \sum_{j=1}^{\infty} P_{X,Y}(x_k, y_j) = 1$$
- Marginalizing out Y -
$$P(X = x_k) = \sum_{j=1}^{\infty} P_{X,Y}(x_k, y_j)$$
- Marginalizing out X -
$$P(Y = y_j) = \sum_{k=1}^{\infty} P_{X,Y}(x_k, y_j)$$

Example

- Consider the joint distribution of the number of products sold X and customer satisfaction rating Y in a small retail store.
- X represents the number of products sold in a transaction: 1, 2, or 3.
- Y represents the customer satisfaction rating for the transaction: 1 (unsatisfied), 2 (neutral), or 3 (satisfied).
- The joint probabilities $P(X = x, Y = y)$ are presented in the following table:

```
##       X=1  X=2  X=3
## Y=1 0.05 0.10 0.05
## Y=2 0.10 0.20 0.05
## Y=3 0.05 0.25 0.15
```

Example

- $P(X = 1, Y = 1) = 0.05$ means the probability of selling 1 product with a satisfaction rating of 1 (unsatisfied) is 0.05.
- $P(X = 2, Y = 2) = 0.20$ means the probability of selling 2 products with a satisfaction rating of 2 (neutral) is 0.20.
- $P(X = 3, Y = 3) = 0.15$ means the probability of selling 3 products with a satisfaction rating of 3 (satisfied) is 0.15.
- And so on for the other combinations.
- To ensure this is a valid probability distribution, the sum of all the probabilities should equal 1:

$$0.05 + 0.10 + 0.05 + 0.10 + 0.20 + 0.05 + 0.05 + 0.25 + 0.15 = 1.0$$

Example- Marginal porbability of Y.

$$P(Y=1)=\sum_{j=1}^{3}P(Y=1,X=j)=0.05+0.10+0.05=0.2$$

$$P(Y=2)=\sum_{j=1}^{3}P(Y=2,X=j)=0.1+0.2+0.05=0.35$$

$$P(Y=3)=\sum_{j=1}^{3}P(Y=3,X=j)=0.05+0.25+0.15=0.45$$

Multivariate Disdrete Random Variables - Vector of random variables

- The joint pmf of a discrete random vector of dimension n:

$$\tilde{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

is defined as:

$$p_{\tilde{X}}(\tilde{x}) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

Multivariate Random Variables - multinomial example

- Suppose that there are r types of successes from n trials with X_1 denoting number of the first type, X_2 the second type etc.
- $$P(X_1 = n_1, X_2 = n_2, \ldots, X_r = n_r) = \frac{n!}{n_1!n_2!\ldots n_r!} p_1^{n_1} p_2^{n_2} \ldots p_r^{n_r}$$
- where $n_1 + n_2 + \cdots + n_r = n$ and $p_1 + p_2 + \cdots + p_r = 1$

Multivariate continuous Random Variables- 2 random variables

- Two random variables X Y are jointly continuous if there exists a non negative function $f_{XY} : \mathcal{R} \times \mathcal{R} \to \mathcal{R}$, such that for any set $A \subseteq \mathcal{R} \times \mathcal{R}$, we have

$$P((X,Y) \in A) = \int\int_{x,y \in A} f_{XY}(x,y)dxdy$$

- The function $f_{XY}(x,y)$ is called the joint density function of X and Y.

Multivariate continuous Random Variables- 2 random variables

- The multivariate is defined as

$$F_{X,Y(x',y')} = P(X \leq x', Y \leq y') = \int_{-\infty}^{x'} \int_{-\infty}^{y'} f_{XY}(x,y)dxdy$$

- So we have that:

$$f(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$$

Multivariate continuous Random Variables - properties

- Suppose $f_{X,Y}(x,y)$ is a joint pdf of X and Y, then the marginal densities of x and Y are given by:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

and,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

Covarianve of 2 random variables

- The covariance of two random variables describes their joint behavior.
- It is the expected value of the product between the diverence of the random variables and their respective means.
- Intuitively, it measures to what extent the random variables fluctuate together.

$$Cov(X, Y) = E([X - E(X)][Y - E(Y)]) =$$

$$= E(XY) - E(X)E(Y)$$

- This can be stacked in a 2×2 covariance matrix:

$$\Sigma_{XY} = \begin{bmatrix} Var[X] & Cov(X, Y) \\ Cov(X, Y) & Var(Y) \end{bmatrix}$$

Matrix Covarianve of n random variables

- This notion can be generalized to a vector of random variables
- The covariance can be calculatd between all the pairs of random variables and stacked into a matrix.

$$\Sigma = \begin{bmatrix} Cov[X_1, X_1] & Cov(X_1, X_2) \ldots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Cov(X_1, X_2) & Cov(X_2, X_n) \\ & \vdots & \\ Cov(X_n, X_1) & Cov(X_n, X_2) & Cov(X_n, X_n) \end{bmatrix}$$

- Note that $Var(X_i) = Cov(X_i, X_i)$. Prove.

Multivariate continuous Normal Random Variables - vector of Random Normals

- Gaussian random vectors are a multidimensional generalization of Gaussian random variables.
- They are parametrized by a mean vector and a covariance matrix that correspond to their mean and covariance.
- A Gaussian random vector X is a random vector with joint pdf:

$$f_{\tilde{X}}(\tilde{x}) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{\left(-\frac{1}{2}(\tilde{x}-\mu)^T\Sigma^{-1}(\tilde{x}-\mu)\right)}$$

Example

```
import numpy as np
from numpy.random import multivariate_normal
import pandas as pd
mu1 = np.array([1, 2])
Sigma1 = np.array([[1, 0.5],
                   [0.5, 1]])
print("Mean Vector (mu1):", mu1)
```

```
## Mean Vector (mu1): [1 2]
```

```
print("Covariance Matrix (Sigma1):\n", Sigma1)
```

```
## Covariance Matrix (Sigma1):
##  [[1.  0.5]
##  [0.5 1. ]]
```

Example

```
mv1=multivariate_normal(mean=mu1,cov=Sigma1,size=100)
cov = np.cov(mv1[:, 0], mv1[:, 1])[0, 1]
mean_mv1_0 = np.mean(mv1[:, 0])
std_mv1_0 = np.std(mv1[:, 0], ddof=1)
print("Cov of mv1[:, 0] and mv1[:, 1]:", round(cov,3))
```

```
## Cov of mv1[:, 0] and mv1[:, 1]: 0.44
```

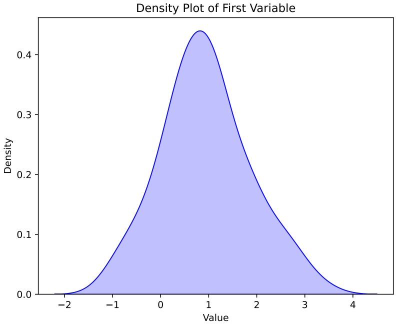
```
print("Mean of mv1[:, 0]:", round(mean_mv1_0,3))
```

```
## Mean of mv1[:, 0]: 0.941
```

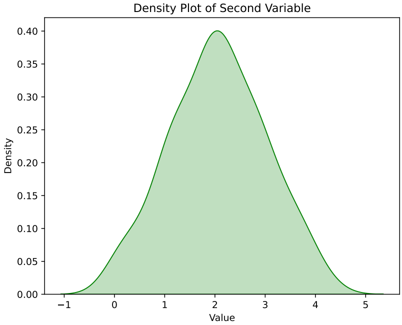
```
print("Std of mv1[:, 0]:", round(std_mv1_0,3))
```

```
## Std of mv1[:, 0]: 0.923
```

Example

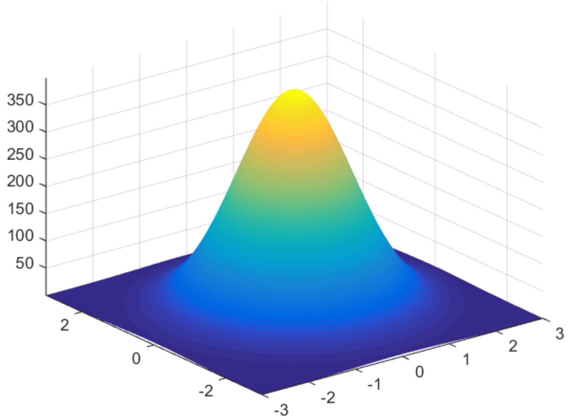


Density Plot of First Variable
Density
Value
0.0
0.1
0.2
0.3
0.4
−2
−1
0
1
2
3
4



Density Plot of Second Variable
Density
Value
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
−1
0
1
2
3
4
5

Multivariate continuous Normal Random Variables - vector of Random Normals



Example

```
# Define the mean vector
mu2 = np.array([1, 2, 3, 4])

# Define the 4x4 covariance matrix
cov_matrix = np.array([[1, 0.5, 0.2, 0.1],
                       [0.5, 2, 0.3, 0.4],
                       [0.2, 0.3, 1.5, 0.6],
                       [0.1, 0.4, 0.6, 3]])
```

Example

```
mv2=multivariate_normal(mean=mu2,cov=cov_matrix,size=1000)
mv2_df=pd.DataFrame(mv2, columns=["Var1","Var2","Var3","Va
covariance_matrix = mv2_df.cov()
print(covariance_matrix)
```

```
##           Var1      Var2      Var3      Var4
## Var1  1.008258  0.499977  0.237448  0.157531
## Var2  0.499977  1.979687  0.335122  0.351404
## Var3  0.237448  0.335122  1.442853  0.514906
## Var4  0.157531  0.351404  0.514906  2.921139
```

Independence and conditional independence

- The n entries $X_1, X_2, ..., X_n$ in a random vector $\tilde{X}$ are independent if and only if:

$$F_{\tilde{X}}(\tilde{x}) = \Pi_{i=1}^{n} F_{X_i}(X_i)$$

- Which is equivalent to

$$P_{\tilde{X}}(\tilde{x}) = \Pi_{i=1}^{n} p_{X_i}(X_i)$$

  for discrete vectors.

- For continuous vectors when the joint pdf exists:

$$f_{\tilde{X}}(\tilde{x}) = \Pi_{i=1}^{n} f_{X_i}(X_i)$$

- There are cases when conditional on some variable Y $X_1, X_2, \ldots, X_n$ are independent.

$$F_{\tilde{X}|Y}(\tilde{x}) = \Pi_{i=1}^{n} F_{X_i|Y}(X_i)$$

Exercise

- Write the pmf of $X_1, \ldots, X_n$ given Y. Write the pmf if given Y, $X_1, \ldots, X_n$ are independent.
- Write the density of two independent normal variables.
- Assume that X_1, X_2 have a multivariate Normal distribution:

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$

  and that

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- Prove that X_1 and X_2 independent?