AI VIET NAM
@aivietnam.edu.vn

Module 02 – Exercise Class

STATISTIC

Nguyen Quoc Thai

AI VIET NAM
@aivietnam.edu.vn

Objectives

Introduction

- Random Variable
- Discrete Random Variable
- Continuous Random Variable
- Mean: $\mu = \frac{1}{n}\sum_k x_k$
- Variance: $\mathrm{Var}(X) = \frac{1}{n}\sum_k (x_k - \mu)^2$
- Standard Deviation: $\sigma(X) = \sqrt{\mathrm{Var}(X)}$
- Covariance & Correlation:
  $\mathrm{Cov}(X, Y)$, $\mathrm{Corr}(X,Y)$
- Important Probability Distributions
  - Bernoulli
  - Uniform
  - Normal

Application

- Tabular Data Analysis
- Text Retrieval

Outline

SECTION 1
Basic Statistics

SECTION 2
Important Probability Distribution

SECTION 3
Tabular Data Analysis

SECTION 4
Text Retrieval

Basic Statistics

A Random Variable

A random variable X is a function X: $\Omega \rightarrow \mathbb{R}$, maps an outcome $s \in \Omega$ to a number on the real line $X(s) \in \mathbb{R}$

A continuous random variable X is a function X(s): $\Omega \rightarrow \mathbb{R}$, maps an outcome s from **an uncountably infinite** to a number on the real line $X(s) \in \mathbb{R}$

A discrete random variable X is a function X(s): $\Omega \rightarrow \mathbb{R}$, maps an outcome from a finite or countably infinite sample space to a number on the real line $X(s) \in \mathbb{R}$

Basic Statistics

A Continuous Random Variable

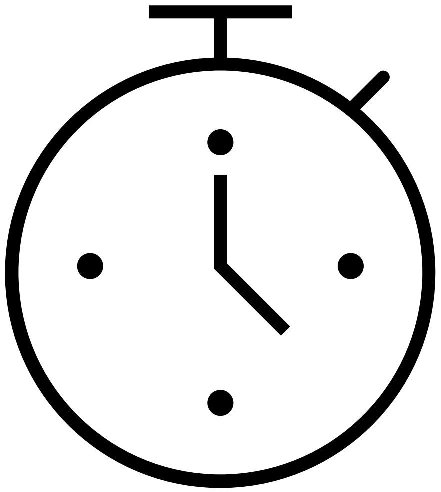
A continuous random variable X is a function X(s): $\Omega \rightarrow \mathbb{R}$, maps an outcome s from **an uncountably infinite** to a number on the real line $X(s) \in \mathbb{R}$

Rotate a pointer about a pivot in a plane (a clock)

Outcome: the angle where stops: $2\pi\theta, \theta \in (0,1]$

$\Omega = (0\ ,\ 1]$

A continuos random variable: $X(\theta) = \theta$



Basic Statistics

A Discrete Random Variable

A discrete random variable X is a function $X(s): \Omega \rightarrow \mathbb{R}$, maps an outcome from a finite or countably infinite sample space to a number on the real line $X(s) \in \mathbb{R}$

Toss a coin 3 times in sequence

Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

- X(s) = the number of Heads in the sequence

  X(HTH) = 2 X(THT) = 1 …

- $Y(s) \begin{cases} \text{The index of the first H} \\ 0 \text{ if the sequence has no H} \end{cases}$

  X(TTH) = 3 X(TTT) = 0 …

Basic Statistics

A Discrete Random Variable

A discrete random variable X is a function $X(s): \Omega \rightarrow \mathbb{R}$, maps an outcome from a finite or countably infinite sample space to a number on the real line $X(s) \in \mathbb{R}$

Toss a coin 3 times in sequence

Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

- $X(s)$ = the number of Heads in the sequence

  $X(s) = 2$

  => corresponds to the event s = {HHT, HTH, THH}

  $1 < X(s) \leq 3$

  => s = {HHH, HHT, HTH, THH}

Basic Statistics

A Discrete Random Variable

Probability measure on discrete random variables

Toss a coin 3 times in sequence

Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

- X(s) = the number of Heads in the sequence

  X(s) = 2 => corresponds to the event s = {HHT, HTH, THH}

  => P(X=2) = P({HHT, HTH, THH}) = 3/8

  $1 < X(s) \leq 3$ => s = {HHH, HHT, HTH, THH}

  => $P(1 < X \leq 3)$ = P({HHH, HHT, HTH, THH}) = 3/8

Question

- P(X=1)
- P(X=2)
- P(X<3)
- $P(X \leq -1)$
- $P(X \leq 3)$
- $P(1 < X \leq 3)$

Discrete Random Variables

Probability Mass Function

Probability mass function
$$p_X(x) = P(X = x)$$

Toss a coin 3 times in sequence

Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

X(s) = the number of Heads in the sequence

- $\mathrm{p}(0) = 1/8$
- $\mathrm{p}(1) = 3/8$
- $\mathrm{p}(2) = 3/8$
- $\mathrm{p}(3) = 1/8$

Discrete Random Variables

Probability Distribution Function

Probability mass function
$p_X(x) = P(X = x)$

(Cumulative) Probability distribution function
$F_X(x) = P(X \leq x)$

Toss a coin 3 times in sequence

Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

X(s) = the number of Heads in the sequence

- $\mathrm{p}(0) = 1/8$
- $\mathrm{p}(1) = 3/8$
- $\mathrm{p}(2) = 3/8$
- $\mathrm{p}(3) = 1/8$

- $\mathrm{F}(-1) = \mathrm{P}(\mathrm{X} \leq -1) = 0/8$
- $\mathrm{F}(0) = \mathrm{P}(\mathrm{X} \leq 0) = 1/8$
- $\mathrm{F}(1) = \mathrm{P}(\mathrm{X} \leq 1) = 4/8$
- $\mathrm{F}(2) = \mathrm{P}(\mathrm{X} \leq 2) = 7/8$

- $\mathrm{F}(3) = \mathrm{P}(\mathrm{X} \leq 3) = 1$
- $\mathrm{F}(4) = \mathrm{P}(\mathrm{X} \leq 4) = 1$

Discrete Random Variables

AI VIET NAM
@aivietnam.edu.vn

Probability Distribution Function

(Cumulative) probability distribution function

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

Toss a coin 3 times in sequence

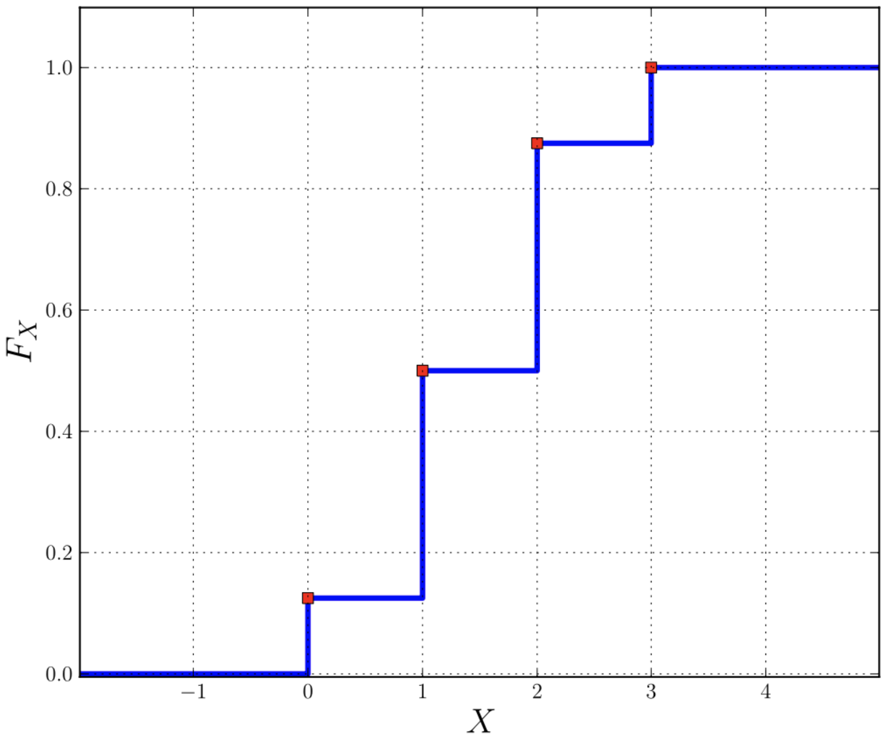
Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

X(s) = the number of Heads in the sequence

- $F(-1) = P(X \leq -1) = 0/8$
- $F(0) = P(X \leq 0) = 1/8$
- $F(1) = P(X \leq 1) = 4/8$
- $F(2) = P(X \leq 2) = 7/8$
- $F(3) = P(X \leq 3) = 1$
- $F(4) = P(X \leq 4) = 1$

$$P(0 < X \leq 2) = P(X = 1) + P(X = 2) = F(2) - F(0)$$

The graph of the probability distribution function



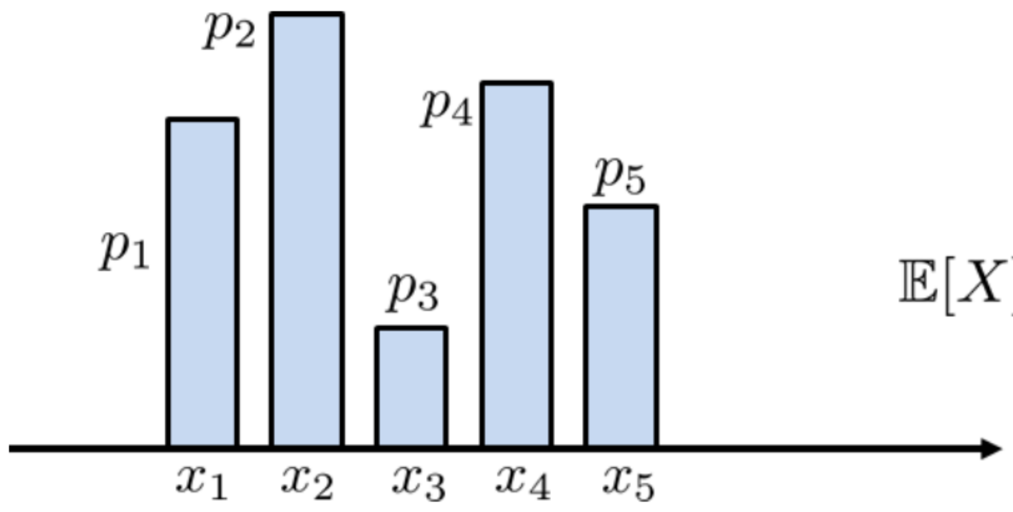
Basic Statistics

Expectation

The expected value of a discrete random variable X is

$$E[X] = \sum_k x_k \,.\, P(X = x_k) = \sum_k x_k \,.\, p_X(x_k)$$

The **E[X]** represents the weighted average value of **X**
E**[X]** is also called the mean of **X**



$$\mathbb{E}[X] = p_1 x_1 + \ldots + p_5 x_5$$

Basic Statistics

Expectation

The expected value of a discrete random variable X is

$$E[X] = \sum_k x_k \,.\, P(X = x_k) = \sum_k x_k \,.\, p_X(x_k)$$

Rolling a die

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

x	1	2	3	4	5	6
p(x)	1/6	1/6	1/6	1/6	1/6	1/6

$$E[X] = 1 * \frac{1}{6} + 2 * \frac{1}{6} + 3 * \frac{1}{6} + 4 * \frac{1}{6} + 5 * \frac{1}{6} + 6 * \frac{1}{6} = \frac{7}{2}$$

Prove the following:

- $E[\alpha X] = \alpha E[X]$
- $E[\alpha X + b] = \alpha E[X] + b$

Basic Statistics

Variance

Mean of X:
$\mu = E[X]$

Variance (The average weighted **square distance** from the mean) of X:
$$\mathrm{Var}(X) = E[(X-\mu)^2] = \sum_k (x_k - \mu)^2 p(x_k)$$

Toss a coin 2 times in sequence

Ω = {HH, TH, HT, TT}

X(s) = the number of Heads in the sequence

- $p(0) = 1/4$
- $p(1) = 2/4$
- $p(2) = 1/4$

x	0	1	2
p(x)	1/4	2/4	1/4

$$E[X] = 0 * \frac{1}{4} + 1 * \frac{2}{4} + 2 * \frac{1}{4} = 1$$

$$Var(X) = (0-1)^2 * \frac{1}{4} + (1-1)^2 * \frac{2}{4} + (2-1)^2 * \frac{1}{4} = 0.5$$

Basic Statistics

Variance

Mean of X:
$\mu = E[X]$

Variance (The average weighted **square distance** from the mean) of X:
$$\mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2] = \sum_k (x_k-\mu)^2 p(x_k) = E[X^2] - \mu^2$$

Toss a coin 2 times in sequence

$\Omega = \{HH, TH, HT, TT\}$

$X(s)$ = the number of Heads in the sequence

- $p(0) = 1/4$
- $p(1) = 2/4$
- $p(2) = 1/4$

x	0	1	2
p(x)	1/4	2/4	1/4

$$E[X] = 0 * \frac{1}{4} + 1 * \frac{2}{4} + 2 * \frac{1}{4} = 1$$

$$E[X^2] = 0 * \frac{1}{4} + 1 * \frac{2}{4} + 4 * \frac{1}{4} = 1.5$$

$$\boldsymbol{Var(X) = 1.5 - 1 = 0.5}$$

Basic Statistics

Standard Deviation

Standard Deviation (The average weighted **distance** from the mean) of X:

$$\sigma(X) = \sqrt{Var(X)} = \sqrt{E[(X-\mu)^2]} = \sqrt{E[X^2] - \mu^2}$$

Toss a coin 2 times in sequence

$\Omega = \{HH, TH, HT, TT\}$

$X(s)$ = the number of Heads in the sequence

- $p(0) = 1/4$
- $p(1) = 2/4$
- $p(2) = 1/4$

x	0	1	2
p(x)	1/4	2/4	1/4

$$E[X] = 0 * \frac{1}{4} + 1 * \frac{2}{4} + 2 * \frac{1}{4} = 1$$

$$Var(X) = 1.5 - 1 = 0.5$$

$$\sigma(X) = \sqrt{1.5 - 1} = \sqrt{0.5} = 0.707$$

Basic Statistics

Standard Deviation

X = {1, 3, 4, 4}

$$E[X] = \mu = \sum_{k} x_k \, . \, p_X(x_k) \qquad Var(X) = \sum_{k} (x_k - \mu)^2 p(x_k) \qquad \sigma(X) = \sqrt{Var(X)}$$

x	1	3	4
times	1	1	2
p(x)	1/4	1/4	1/2

$$E[X] = \mu = 1 * \frac{1}{4} + 3 * \frac{1}{4} + 4 * \frac{1}{2} = 3$$

$$Var(X) = \frac{1}{4} * (1-3)^2 + \frac{1}{4} * (3-3)^2 + \frac{1}{2} * (4-3)^2 = 1.5$$

$$\sigma(X) = \sqrt{Var(X)} = \sqrt{1.5} \approx 1.22$$

AI VIET NAM
@aivietnam.edu.vn

Basic Statistics

Standard Deviation

X = {1, 3, 4, 4}

$$E[X] = \mu = \sum_k x_k \, . \, p_X(x_k) \qquad Var(X) = \sum_k (x_k - \mu)^2 p(x_k) \qquad \sigma(X) = \sqrt{Var(X)}$$

Mean: $\mu = \frac{1}{n}\sum_k x_k$

Variance: $Var(X) = \frac{1}{n}\sum_k (x_k - \mu)^2$

Standard Deviation: $\sigma(X) = \sqrt{Var(X)}$

$$\mu = \frac{1}{4}(1 + 3 + 4 + 4) = 3$$

$$Var(X) = \frac{1}{4}[(1-3)^2 + (3-3)^2 + (4-3)^2 + (4-3)^2] = 1.5$$

$$\sigma(X) = \sqrt{Var(X)} = \sqrt{1.5} \approx 1.22$$

Basic Statistics

Standard Deviation

X = {1, 3, 4, 4}

$$E[X] = \mu = \sum_k x_k \, . \, p_X(x_k) \qquad Var(X) = \sum_k (x_k - \mu)^2 p(x_k) \qquad \sigma(X) = \sqrt{Var(X)}$$

```python
data = np.array([1, 3, 4, 4])
print(data)

print("Mean: ", np.mean(data))
print("Std: ", np.std(data))
print("Variance: ", np.var(data) )
```

```
[1 3 4 4]
Mean:  3.0
Std:  1.224744871391589
Variance:  1.5
```

$$\mu = \frac{1}{4}(1 + 3 + 4 + 4) = 3$$

$$Var(X) = \frac{1}{4}[(1-3)^2 + (3-3)^2 + (4-3)^2 + (4-3)^2] = 1.5$$

$$\sigma(X) = \sqrt{Var(X)} = \sqrt{1.5} \approx 1.22$$

Basic Statistics

Practice

Một xạ thủ có 3 viên đạn được yêu cầu bắn lần lượt từng viên cho đến khi trúng mục tiêu hoặc hết cả 3 viên thì thôi. Tính kỳ vọng, phương sai của số đạn đã bắn, biết rằng xác suất bắn trúng đích của mỗi lần bắn là 0.8

Basic Statistics

Covariance

X, Y: random variables
$E[X] = \mu_X;\ E[Y] = \mu_Y$

Covariance of X and Y:

Sample

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = \frac{\sum(x-\mu_X)(y-\mu_Y)}{n-1}$$

X = {1, 3, 4, 4}

$$E[X] = \mu = \frac{1}{4}(1+3+4+4) = 3$$

$$Var(X) = \frac{1}{4}[(1-3)^2 + (3-3)^2 + (4-3)^2 + (4-3)^2] = 1.5$$

$$\sigma(X) = \sqrt{Var(X)} = \sqrt{1.5} \approx 1.22$$

Y = {1, 2, 3, 2}

$$E[Y] = \mu = \frac{1}{4}(1+2+3+2) = 2$$

$$Var(Y) = \frac{1}{4}[(1-2)^2 + (2-2)^2 + (3-2)^2 + (2-2)^2] = 0.5$$

$$\sigma(Y) = \sqrt{Var(Y)} = \sqrt{0.5} \approx 0.707$$

Basic Statistics

Covariance

X, Y: random variables

$E[X] = \mu_X;\ E[Y] = \mu_Y$

Sample

Covariance of X and Y:

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = \frac{\sum(x-\overset{\bar{x}}{\mu_X})(y-\overset{\bar{y}}{\mu_Y})}{n-1}$$

Population

Covariance of X and Y:

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = \frac{\sum(x-\mu_X)(y-\mu_Y)}{n}$$

Basic Statistics

Covariance

X, Y: random variables
$E[X] = \mu_X;\ E[Y] = \mu_Y$

Covariance of X and Y:

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = \frac{\sum(x-\mu_X)(y-\mu_Y)}{n-1}$$

X = {1, 3, 4, 4}

$E[X] = \mu_X = 3$

Y = {1, 2, 3, 2}

$E[Y] = \mu_Y = 2$

$$\mathrm{Cov}(X,Y) = \frac{\sum(x-\mu_X)(y-\mu_Y)}{n-1}$$

$$= \frac{\begin{matrix}(1-3)(1-2)+(3-3)(2-2)\\ +(4-3)(3-2)+(4-3)(2-2)\end{matrix}}{4-1}$$

$$= \frac{2+1}{3} = 1$$

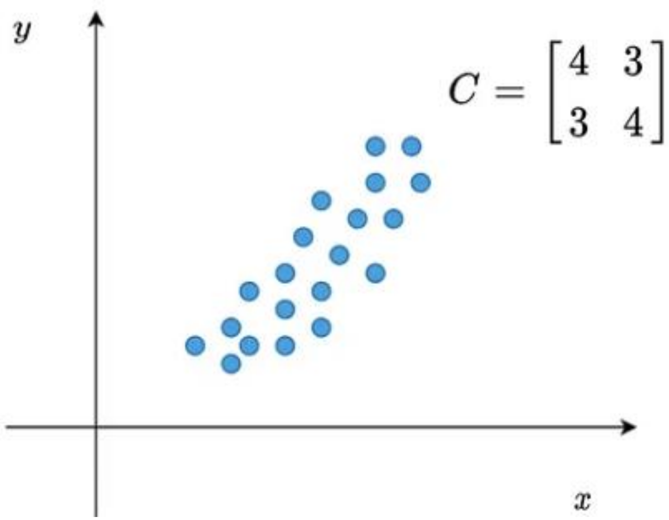
Basic Statistics

Covariance

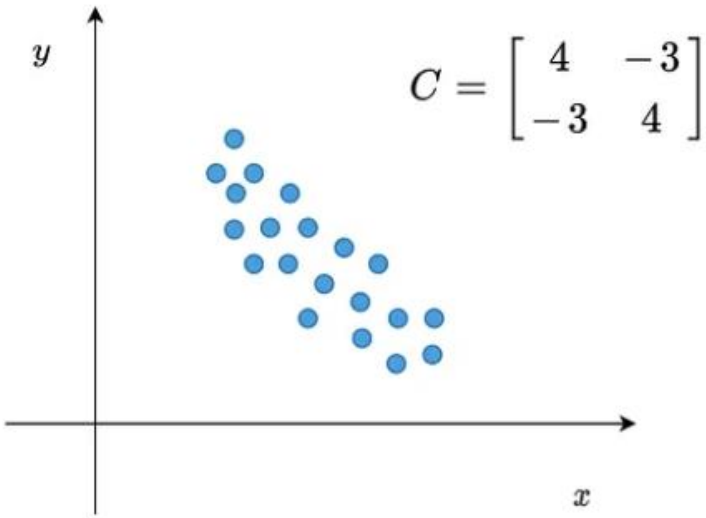
Cov(X,Y)
measures "concordance" or "coherence" of X and Y

Covariance of X and Y:

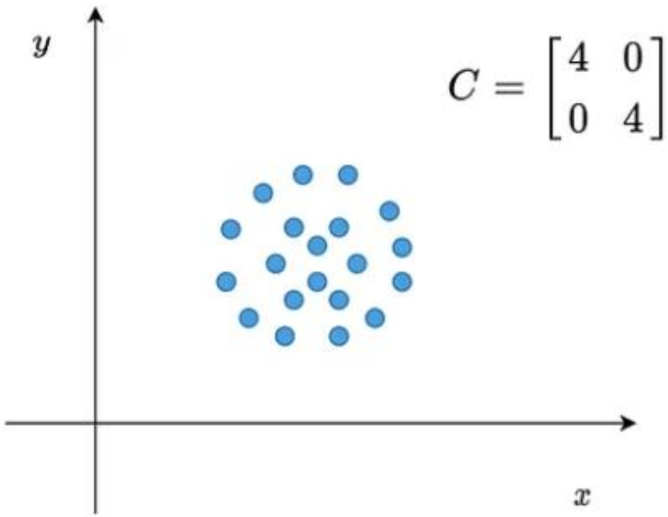
$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = \frac{\sum(x-\mu_X)(y-\mu_Y)}{n-1}$$



Positive Covariance – Cov(X, Y) > 0



Negative Covariance – Cov(X, Y) < 0



Zero Covariance – Cov(X, Y) = 0

Basic Statistics

Correlation

A statistical measure the quantifies the strength and direction of a linear relationshop between two random variables

X = {1, 3, 4, 4}

$$E[X] = \mu_X = 3$$

Y = {1, 2, 3, 2}

$$E[Y] = \mu_Y = 2$$

Correlation of X and Y:

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}$$

$$= \frac{n(\sum_i x_i y_i) - (\sum_i x_i)(\sum_i y_i)}{\sqrt{n\sum_i x_i^2 - (\sum_i x_i)^2}\sqrt{n\sum_i y_i^2 - (\sum_i y_i)^2}}$$

$$\mathrm{Corr}(X, Y) = \frac{27n - 96}{\sqrt{42n - 144}\sqrt{18n - 64}} = \frac{12}{\sqrt{24}\sqrt{8}} \approx 0.866$$

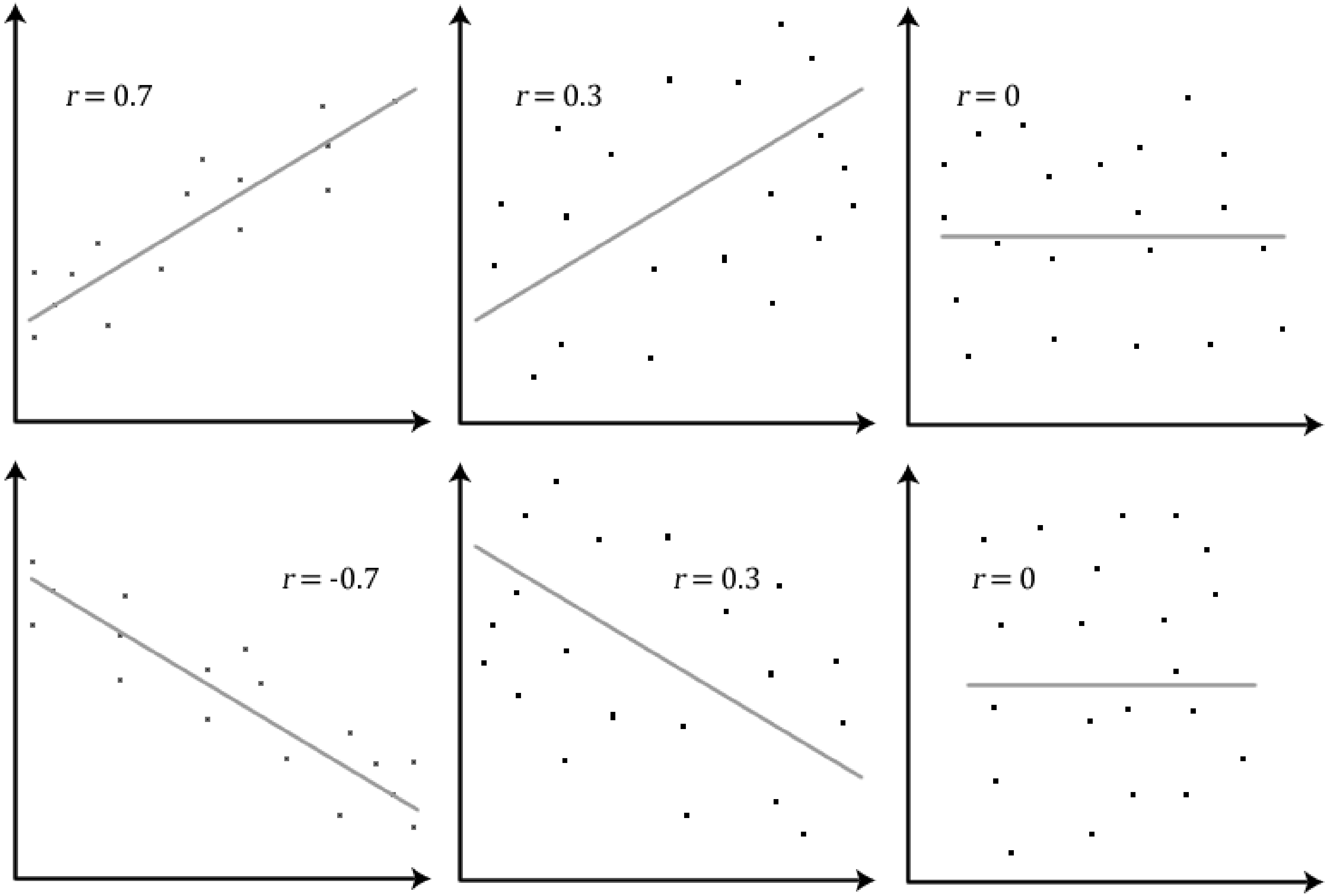
Basic Statistics

Correlation

Correlation of X and Y:

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}$$

$$Corr(X, Y) \in [-1, 1]$$



Basic Statistics

Median

Median: a measure of central tendency which gives the value of the middle – most observation in the data

$$X = \{x_1, x_2, \dots, x_N\}$$

$$\text{If N is odd: } m = S_{\frac{N+1}{2}}$$

$$\text{If N is even: } m = \frac{1}{2}\left(S_{\frac{N}{2}} + S_{\frac{N}{2}+1}\right)$$

X = {1, 3, 4, 2, 5}

$$\text{Sorted}_X = \{1, 2, 3, 4, 5\}$$

$$N = 5 \text{ is odd: } m = S_{\frac{5+1}{2}} = S_3 = 3$$

X = {1, 3, 4, 2}

$$\text{Sorted}_X = \{1, 2, 3, 4\}$$

$$N = 4 \text{ is even: } m = \frac{1}{2}(S_2 + S_3) = 2.5$$

Basic Statistics

Practice

Theo thống kê ở một cửa hàng đậu tương, người ta thấy số lượng đậu tương bán ra X là một biến ngẫu nhiên rời rạc có bảng phân phối dưới đây. Nếu giá nhập là 10000 VNĐ/kg thì cửa hàng sẽ lãi 5000 VNĐ/kg, nếu cuối ngày không bán được thì lỗ 8000 VNĐ/kg. Mỗi ngày cửa hàng nên nhập bao nhiêu để thu được lãi nhiều nhất?

X (kg)	**10**	**13**	**16**	**19**
$P(X=x_i)$	0.15	0.2	0.35	0.3

Basic Statistics

Exercise 1

$$\text{Mean: } \mu = \frac{1}{n}\sum_{k} x_k$$

```python
### Question 1
import numpy as np

def compute_mean(X):
    return np.sum(X) / len(X)

X = [2, 0, 2, 2, 7, 4, -2, 5, -1, -1]

print("Mean : ", compute_mean(X))
```

✓ 3.3s

```
Mean :  1.8
```

Basic Statistics

Exercise 2

$$X = \{x_1, x_2, \dots, x_N\}$$

$$\text{If N is odd: } m = S_{\frac{N+1}{2}}$$

$$\text{If N is even: } m = \frac{1}{2}\left(S_{\frac{N}{2}} + S_{\frac{N}{2}+1}\right)$$

```
### Question 2

def compute_median(X):
  size = len(X)
  X = np.sort(X)
  print(X)
  if (size % 2 == 0):
    return (1/2*(X[int(size/2)-1] \
                + (X[int(size/2) + 1 - 1])))
  else:
    return X[int((size+1)/2)-1]

X = [1, 5, 4, 4, 9, 13]
print("Median: ", compute_median(X))
```

✓ 0.0s

```
[ 1  4  4  5  9 13]
Median:  4.5
```

AI VIET NAM
@aivietnam.edu.vn

Basic Statistics

Exercise 3

$$\text{Mean: } \mu = \frac{1}{n}\sum_k x_k$$

$$\text{Variance: } \mathrm{Var}(X) = \frac{1}{n}\sum_k (x_k - \mu)^2$$

$$\text{Standard Deviation: } \sigma(X) = \sqrt{\mathrm{Var}(X)}$$

```python
### Question 3

def compute_std(X):
    mean = compute_mean(X)
    variance = 0
    for x in X:
        variance = variance + (x - mean)**2
    variance = variance / len(X)
    return np.sqrt(variance)

X = [ 171, 176, 155, 167, 169, 182]
print(np.round(compute_std(X),2))
```

✓ 0.0s

```
8.33
```

AI VIET NAM
@aivietnam.edu.vn

Basic Statistics

Exercise 3

$$\text{Mean: } \mu = \frac{1}{n}\sum_{k} x_k$$

$$\text{Variance: Var}(X) = \frac{1}{n}\sum_{k} (x_k - \mu)^2$$

$$\text{Standard Deviation: } \sigma(X) = \sqrt{\text{Var}(X)}$$

```python
data = np.array([ 171, 176, 155, 167, 169, 182])

print("Mean: ", np.mean(data))
print("edian: ", np.median(data))
print("Std: ", np.std(data))
print("Variance: ", np.var(data))
```

✓ 0.0s

```
Mean:  170.0
edian:  170.0
Std:  8.32666399786453
Variance:  69.33333333333333
```

Basic Statistics

Exercise 4

```python
### Question 4

def compute_correlation_cofficient(X, Y):
    N = len(X)
    numerator = N * X.dot(Y) - np.sum(X)*np.sum(Y)
    denominator = np.sqrt(N*np.sum(np.square(X))-np.sum(X)**2) \
        * np.sqrt(N*np.sum(np.square(Y))-np.sum(Y)**2)

    return np.round(numerator / denominator,2)

X = np.asarray([-2, -5, -11, 6, 4, 15, 9])
Y = np.asarray([4, 25, 121, 36, 16, 225, 81])
print("Correlation: ", compute_correlation_cofficient(X,Y))
```

✓ 0.0s

```
Correlation:  0.42
```

Outline

SECTION 1

Basic Statistics

SECTION 2

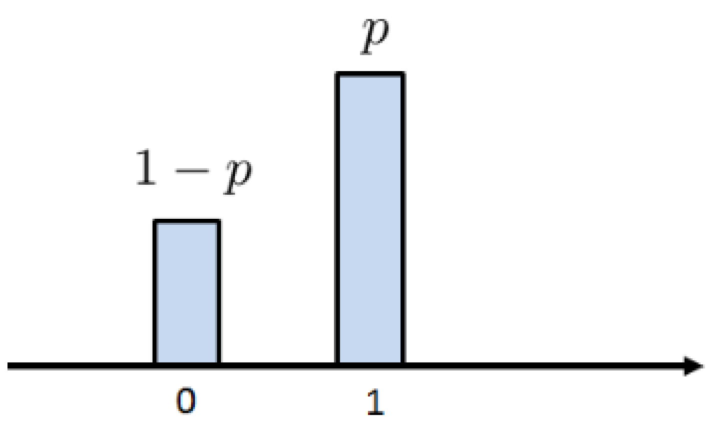
Important Probability Distribution

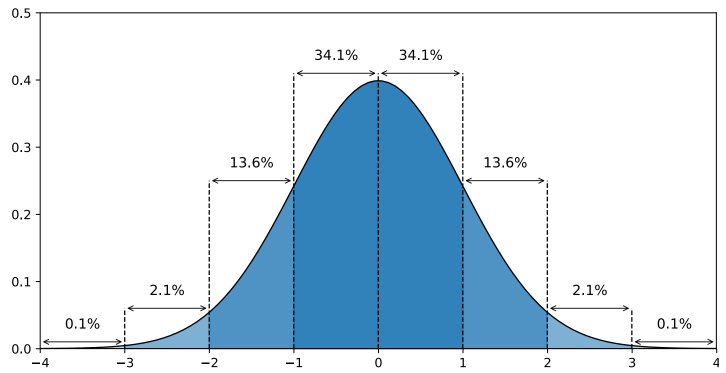
SECTION 3

Tabular Data Analysis

SECTION 4

Text Retrieval





Important Probability Distribution

Bernoulli Random Variable

Bernoulli: two outcomes

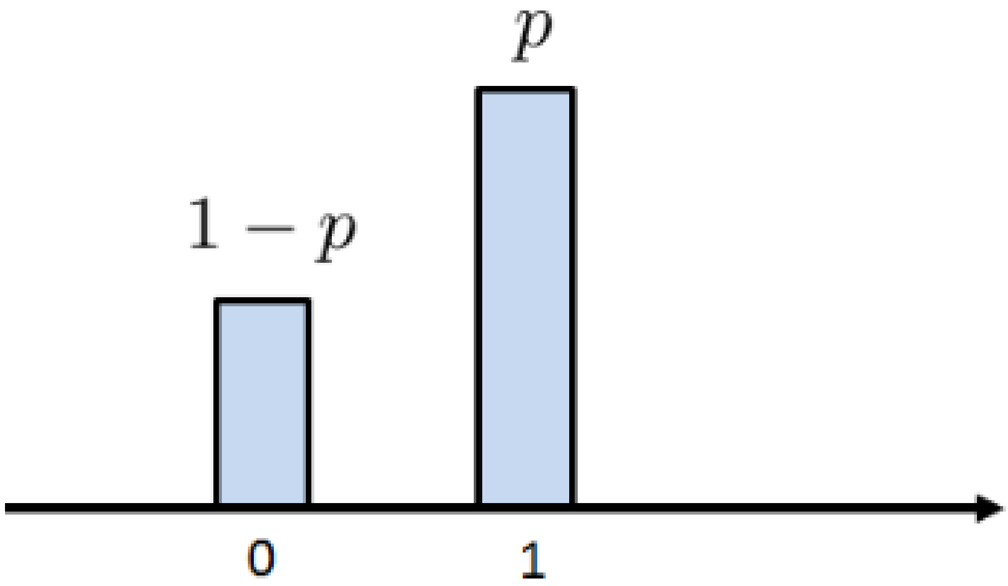
$$P(X) = \begin{cases} P(X=1) = p \\ P(X=0) = 1-p \end{cases}$$

Tossing a coin
Winning or losing a game

$$E[X] = 1 * p + 0 * (1-p) = p$$

$$E[X^2] = 1^2 * p + 0^2 * (1-p) = p$$

$$Var(X) = E[X^2] - E[X]^2 = p - p^2$$
$$= p * (1-p)$$



Important Probability Distribution

Bernoulli Random Variable

Bernoulli: two outcomes

$$P(X) = \begin{cases} P(X = 1) = p \\ P(X = 0) = 1 - p \end{cases}$$

Tossing a coin
Winning or losing a game

Tossing a coin

$$p = \frac{1}{2}$$

$$E[X] = \frac{1}{2} \qquad E[X^2] = \frac{1}{2}$$

$$Var(X) = E[X^2] - E[X]^2 = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}$$

Rolling a die

$$p = \frac{1}{6}$$

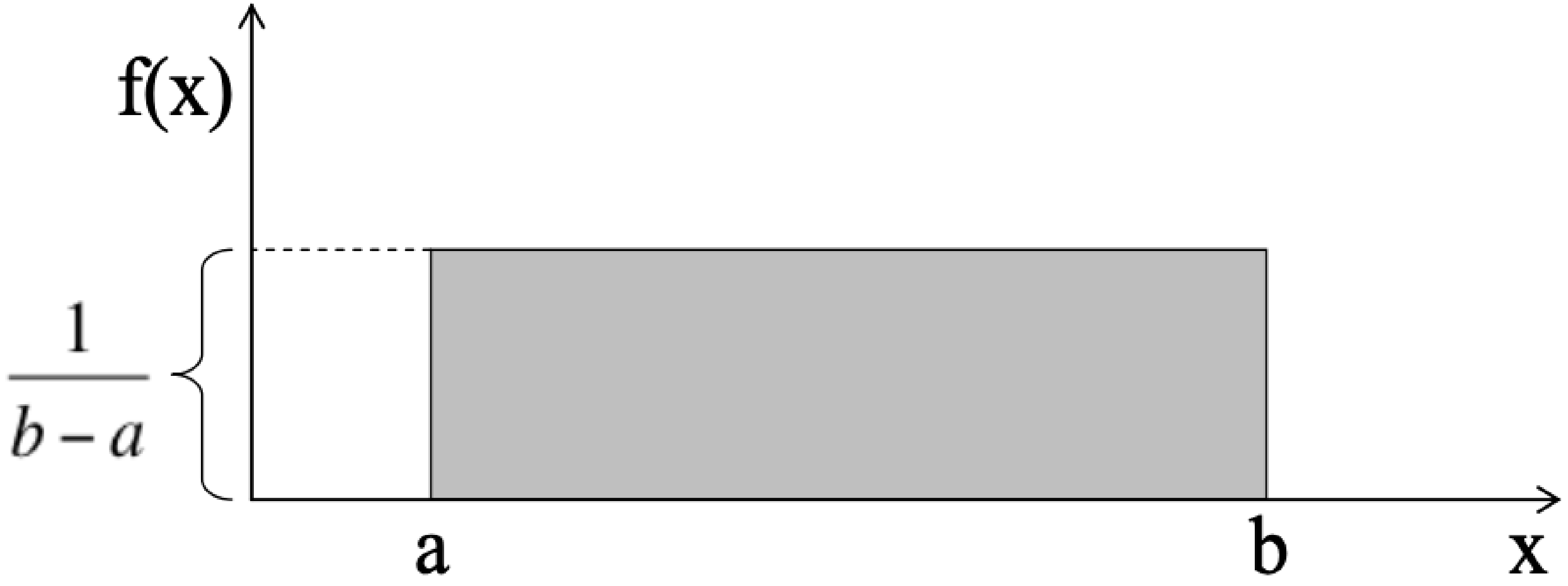
$$E[X] = \frac{1}{6} \qquad E[X^2] = \frac{1}{6}$$

$$Var(X) = E[X^2] - E[X]^2 = \frac{1}{6} * \frac{5}{6} = \frac{5}{36}$$

Important Probability Distribution

Uniform Distribution (Continuous)

$$f(x) = \frac{1}{b - a}$$
$$-\infty < a \leq x \leq b < \infty$$



Important Probability Distribution

Uniform Distribution (Continuous)

$$f(x) = \frac{1}{b-a}$$
$$-\infty < a \leq x \leq b < \infty$$

$$E[X] = \frac{a+b}{2} \qquad Var(X) = \frac{(b-a)^2}{12}$$

Standard Uniform

$a = 0, b = 1$

$$\begin{cases} f(x) = 1; 0 \leq x \leq 1 \\ f(x) = 0; otherwise \end{cases}$$

```python
import numpy as np

data = np.random.uniform(0, 1, (2, 3))
data
```

✓ 0.2s

```
array([[0.93003596, 0.18329112, 0.44956657],
       [0.51529433, 0.17943308, 0.80715331]])
```

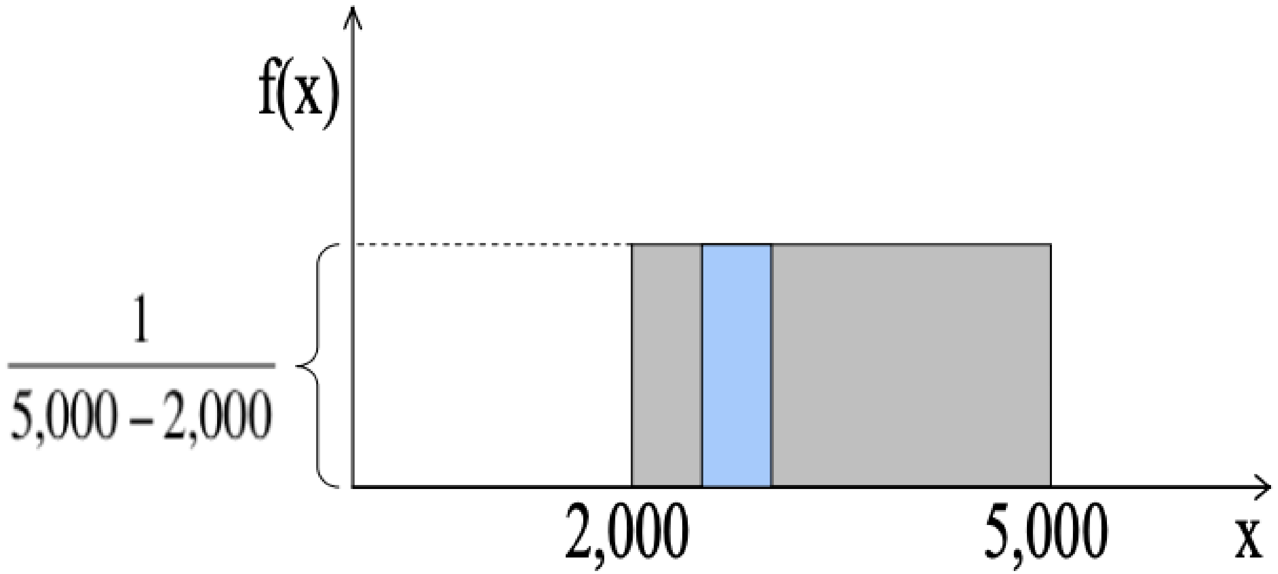
Important Probability Distribution

Uniform Distribution (Continuous)

$$f(x) = \frac{1}{b - a}$$
$$-\infty < a \leq x \leq b < \infty$$

Suppose the amount of gasoline sold daily at a service station is uniformly distributed with a minimum of 2,000 gallons and a maximum of 5,000 gallons.

$$P(2500 < X \leq 3000)$$
$$= \frac{1}{5000 - 2000}(3000 - 2500)$$
$$= 0.1667$$



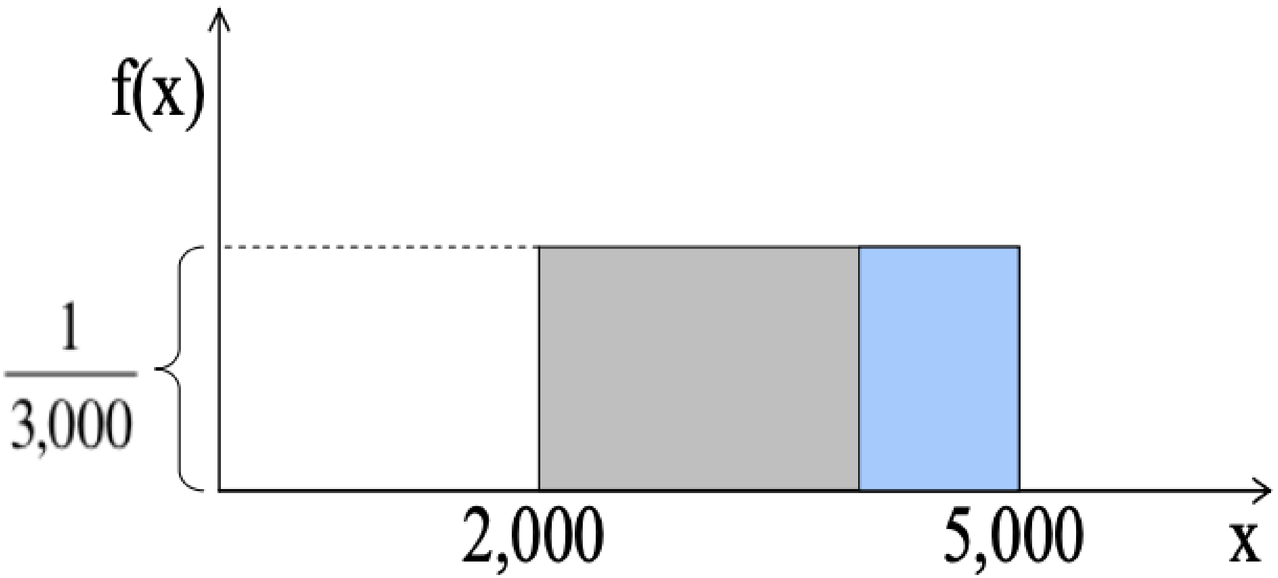
Important Probability Distribution

Uniform Distribution (Continuous)

$$f(x) = \frac{1}{b-a}$$
$$-\infty < a \leq x \leq b < \infty$$

Suppose the amount of gasoline sold daily at a service station is uniformly distributed with a minimum of 2,000 gallons and a maximum of 5,000 gallons.

$$P(X > 4000)$$
$$= \frac{1}{5000 - 2000}(5000 - 4000)$$
$$= 0.3333$$



Important Probability Distribution

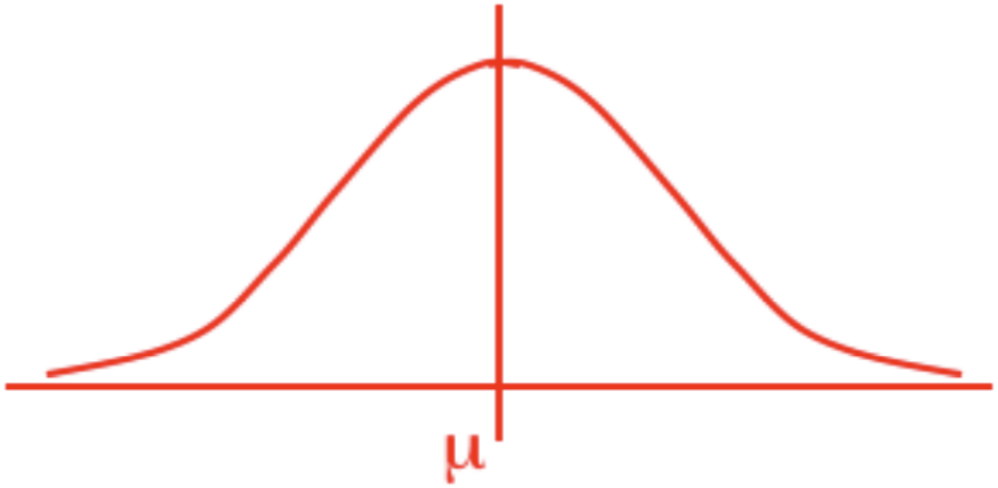
Normal Distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

$$-\infty < x < \infty$$

The curve is bell shaped and is symmetric around the mean μ

Standard deviation σ controls the "flatness" of the curve



Important Probability Distribution

Normal Distribution

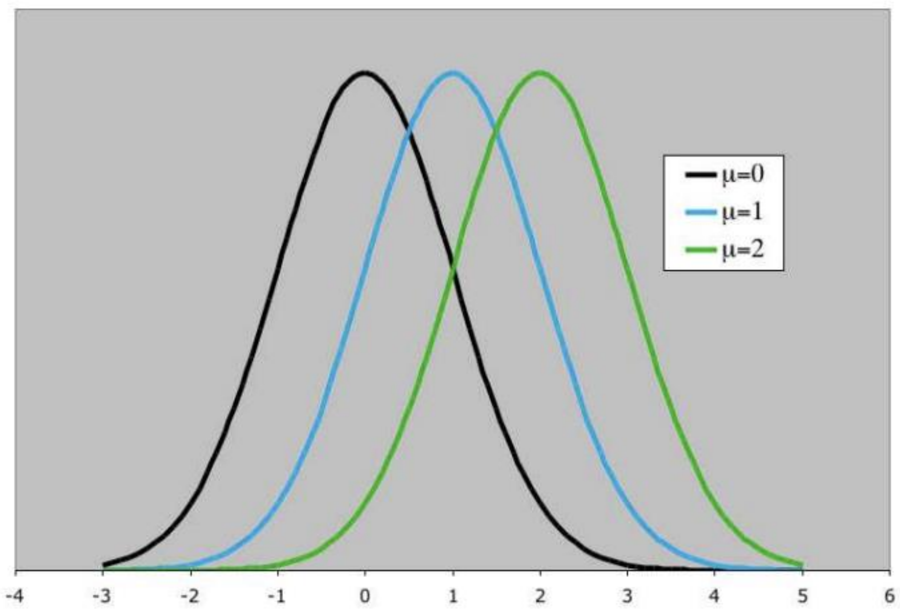
$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

$$-\infty < x < \infty$$

The curve is bell shaped and is symmetric around the mean μ
Increasing the mean shifts the density curve to the right

Standard deviation σ controls the "flatness" of the curve

Same variance, different means



AI VIET NAM
@aivietnam.edu.vn

Important Probability Distribution

Normal Distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

$$-\infty < x < \infty$$

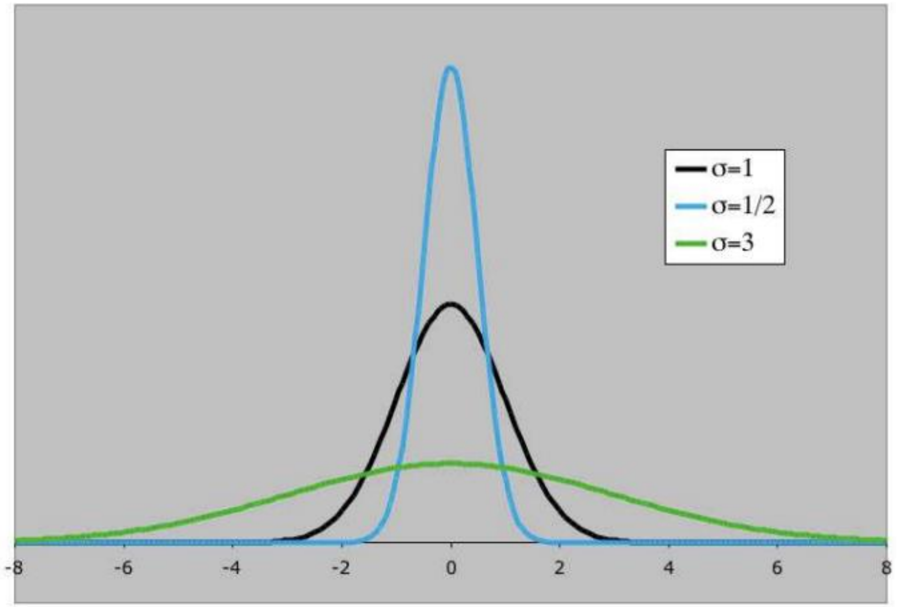
The curve is bell shaped and is symmetric around the mean μ

Increasing the mean shifts the density curve to the right

Standard deviation σ controls the “flatness” of the curve

Increasing the standard deviation flattens the density curve

Same mean, different standard deviations



AI VIET NAM
@aivietnam.edu.vn

Important Probability Distribution

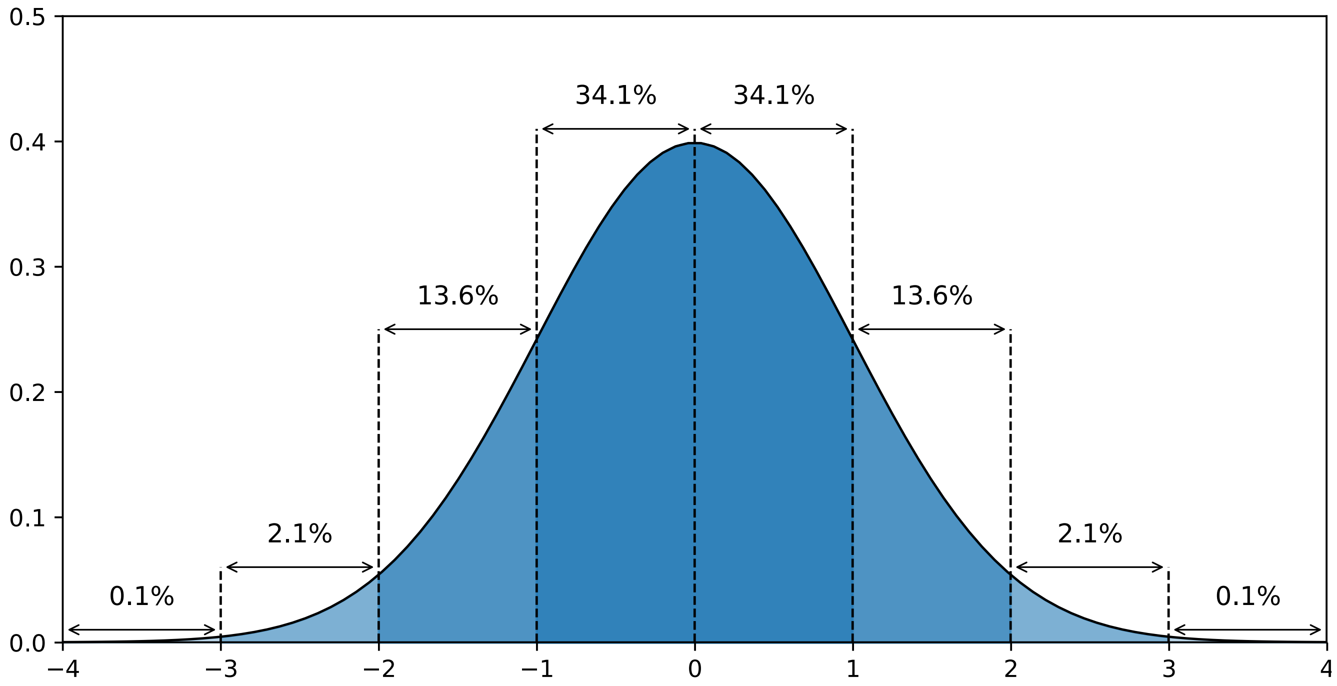
Normal Distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$
$$-\infty < x < \infty$$

$$\mu = 0, \sigma = 1$$

Standard Normal Distribution

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{x^2}{2}}$$



Important Probability Distribution

Normal Distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$
$$-\infty < x < \infty$$

$$\mu = 0, \sigma = 1$$

Standard Normal Distribution

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{x^2}{2}}$$

```python
import numpy as np

data = np.random.normal(0, 1, (2, 3))
data
```

✓ 0.0s

```
array([[-0.31501619,  1.06958159, -0.0243189 ],
       [ 0.2949796 , -0.15386693,  0.00876236]])
```

QUIZ TIME

AI VIET NAM
@aivietnam.edu.vn

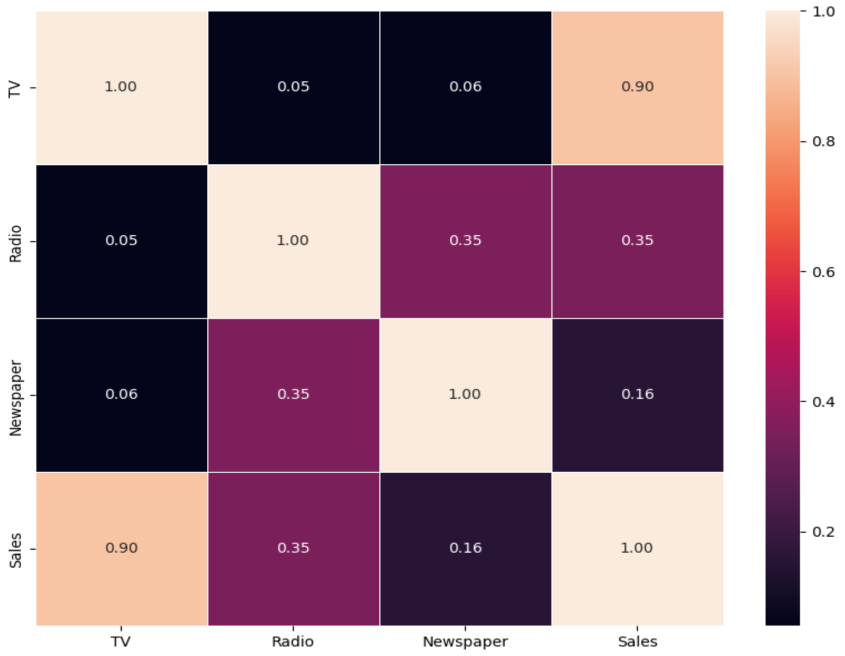
Outline

SECTION 1
Basic Statistics

SECTION 2
Important Probability Distribution

SECTION 3
Tabular Data Analysis

SECTION 4
Text Retrieval



Tabular Data Analysis

Advertising Dataset

```python
import pandas as pd
data = pd.read_csv("advertising.csv")
data
```

```python
data.head(5)
data.tail(5)
```

	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	12.0
3	151.5	41.3	58.5	16.5
4	180.8	10.8	58.4	17.9
...	...	...	...	...
195	38.2	3.7	13.8	7.6
196	94.2	4.9	8.1	14.0
197	177.0	9.3	6.4	14.8
198	283.6	42.0	66.2	25.5
199	232.1	8.6	8.7	18.4

Tabular Data Analysis

Basic EDA

```
1 data.describe()
```

	TV	Radio	Newspaper	Sales
count	200.000000	200.000000	200.000000	200.000000
mean	147.042500	23.264000	30.554000	15.130500
std	85.854236	14.846809	21.778621	5.283892
min	0.700000	0.000000	0.300000	1.600000
25%	74.375000	9.975000	12.750000	11.000000
50%	149.750000	22.900000	25.750000	16.000000
75%	218.825000	36.525000	45.100000	19.050000
max	296.400000	49.600000	114.000000	27.000000

```
1 data.info()
```

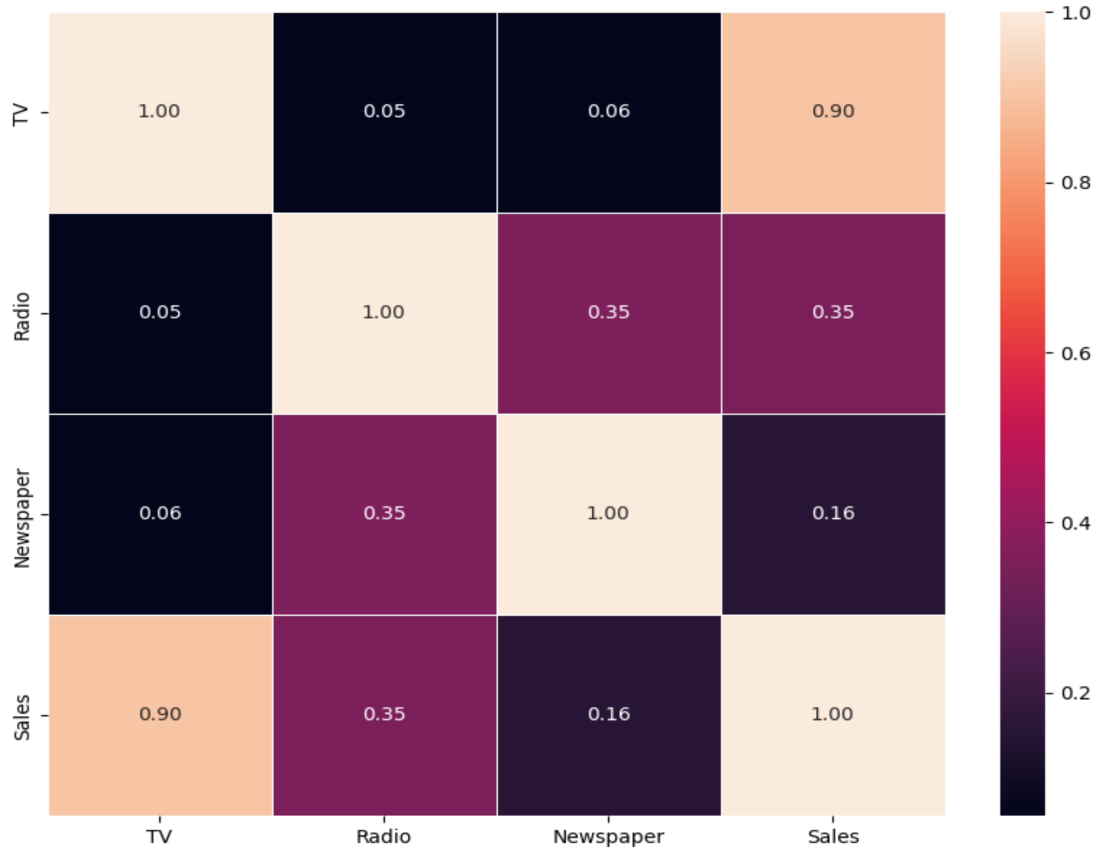
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   TV         200 non-null    float64
 1   Radio      200 non-null    float64
 2   Newspaper  200 non-null    float64
 3   Sales      200 non-null    float64
dtypes: float64(4)
memory usage: 6.4 KB
```

Tabular Data Analysis

Correlation

```
1 data_corr_coef = data.corr()
2 data_corr_coef
```

	TV	Radio	Newspaper	Sales
TV	1.000000	0.054809	0.056648	0.901208
Radio	0.054809	1.000000	0.354104	0.349631
Newspaper	0.056648	0.354104	1.000000	0.157960
Sales	0.901208	0.349631	0.157960	1.000000



Outline

SECTION 1

Basic Statistics

SECTION 2

Important Probability Distribution

SECTION 3

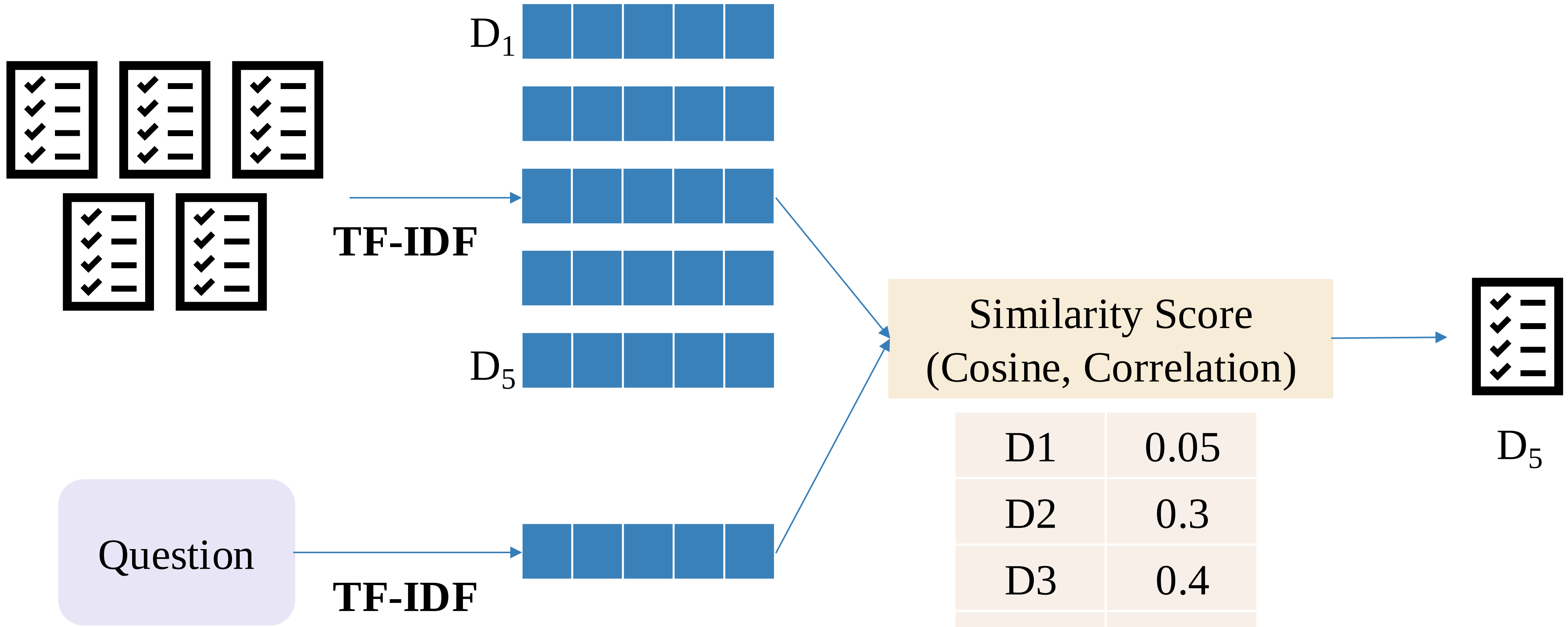
Tabular Data Analysis

SECTION 4

Text Retrieval

Text Retrieval

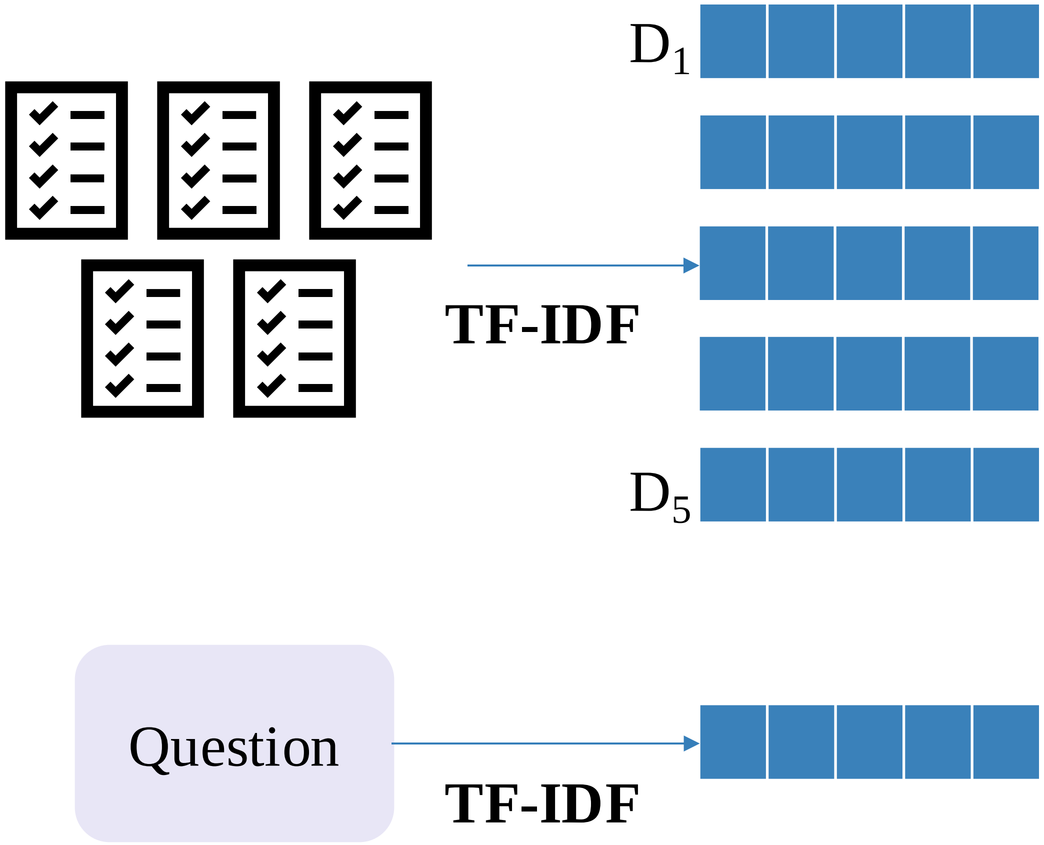
Text Retrieval



D1	0.05
D2	0.3
D3	0.4
D4	0.2
D5	0.9

Text Retrieval

Text Embedding (TF-IDF)



Text Retrieval

Term Frequency (TF)

$$tf_{t,d} = count(t, d)$$

- Some ways to reduce the raw frequency:
  - Using log space + add 1:

$$tf_{t,d} = log(count(t, d) + 1)$$

  - Divide the number of occurrences by the length of document:

$$tf_{t,d} = \frac{count(t, d)}{len(d)}$$

Text Retrieval

Term Frequency (TF)

$$tf_{t,d} = count(t, d)$$

Example

[dog, bites, man]
[man, bites, dog]
[dog, eats, meat]
[man, eats, food]

	bites	dog	eats	food	man	meat
D1						
D2						
D3						
D4						

Text Retrieval

Term Frequency (TF)

$$tf_{t,d} = count(t, d)$$

Example

[dog, bites, man]
[man, bites, dog]
[dog, eats, meat]
[man, eats, food]

	bites	dog	eats	food	man	meat
D1	1/3	1/3	0	0	1/3	0
D2	1/3	1/3	0	0	1/3	0
D3	0	1/3	1/3	0	0	1/3
D4	0	0	1/3	1/3	1/3	0

Text Retrieval

Inverse Document Frequency (IDF)

$$idf_t = \frac{N}{df_t}$$

- Measures the importance of the word across a corpus

  N: The total number of documents in the corpus

  df_t: The number of documents with term t in them

- Using log space:

$$idf_t = log\frac{N}{df_t} \qquad idf_t = log\frac{N}{df_t} + 1 \qquad idf_t = log\frac{N+1}{df_t+1} + 1$$

Text Retrieval

Inverse Document Frequency (IDF)

$$idf_t = ln\frac{N+1}{df_t+1}+1$$

Example

[dog, bites, man]
[man, bites, dog]
[dog, eats, meat]
[man, eats, food]

bites	dog	eats	food	man	meat

Text Retrieval

Inverse Document Frequency (IDF)

$$idf_t = ln\frac{N+1}{df_t+1} + 1$$

Example

[dog, bites, man]
[man, bites, dog]
[dog, eats, meat]
[man, eats, food]

bites	dog	eats	food	man	meat
1.511	1.223	1.511	1.916	1.223	1.916

Text Retrieval

TF-IDF

$$w_{t,d} = tf_{t,d} \times idf_t$$

- The weighted value $w_{t,d}$ for word t in document d
- IDF weighs down the terms: very common across a corpus and rare terms
- The TF-IDF vector representation for a document is then simply TF-IDF score for each term in that document.

Text Retrieval

TF-IDF

bites	1.511
dog	1.223
Eats	1.511
Food	1.916
Man	1.223
meat	1.916

	bites	dog	eats	food	man	meat
D1	1/3	1/3	0	0	1/3	0
D2	1/3	1/3	0	0	1/3	0
D3	0	1/3	1/3	0	0	1/3
D4	0	0	1/3	1/3	1/3	0

Example

[dog, bites, man]
[man, bites, dog]
[dog, eats, meat]
[man, eats, food]

	bites	dog	eats	food	man	meat
D1						
D2						
D3						
D4						

Text Retrieval

TF-IDF

Example

[dog, bites, man]
[man, bites, dog]
[dog, eats, meat]
[man, eats, food]

bites	1.511
dog	1.223
Eats	1.511
Food	1.916
Man	1.223
meat	1.916

	bites	dog	eats	food	man	meat
D1	1/3	1/3	0	0	1/3	0
D2	1/3	1/3	0	0	1/3	0
D3	0	1/3	1/3	0	0	1/3
D4	0	0	1/3	1/3	1/3	0

	bites	dog	eats	food	man	meat
D1	0.504	0.408	0	0	0.400	0
D2	0.504	0.408	0	0	0.408	0
D3	0	0.408	0.504	0	0	0.639
D4	0	0	0.504	0.639	0.408	0

AI VIET NAM
@aivietnam.edu.vn

Text Retrieval

TF-IDF

- Information Retrieval

Example

[dog, bites, man]
[man, bites, dog]
[dog, eats, meat]
[man, eats, food]

	bites	dog	eats	food	man	meat
D1	0.504	0.408	0	0	0.400	0
D2	0.504	0.408	0	0	0.408	0
D3	0	0.408	0.504	0	0	0.639
D4	0	0	0.504	0.639	0.408	0

Search: “dog, meat”

Text Retrieval

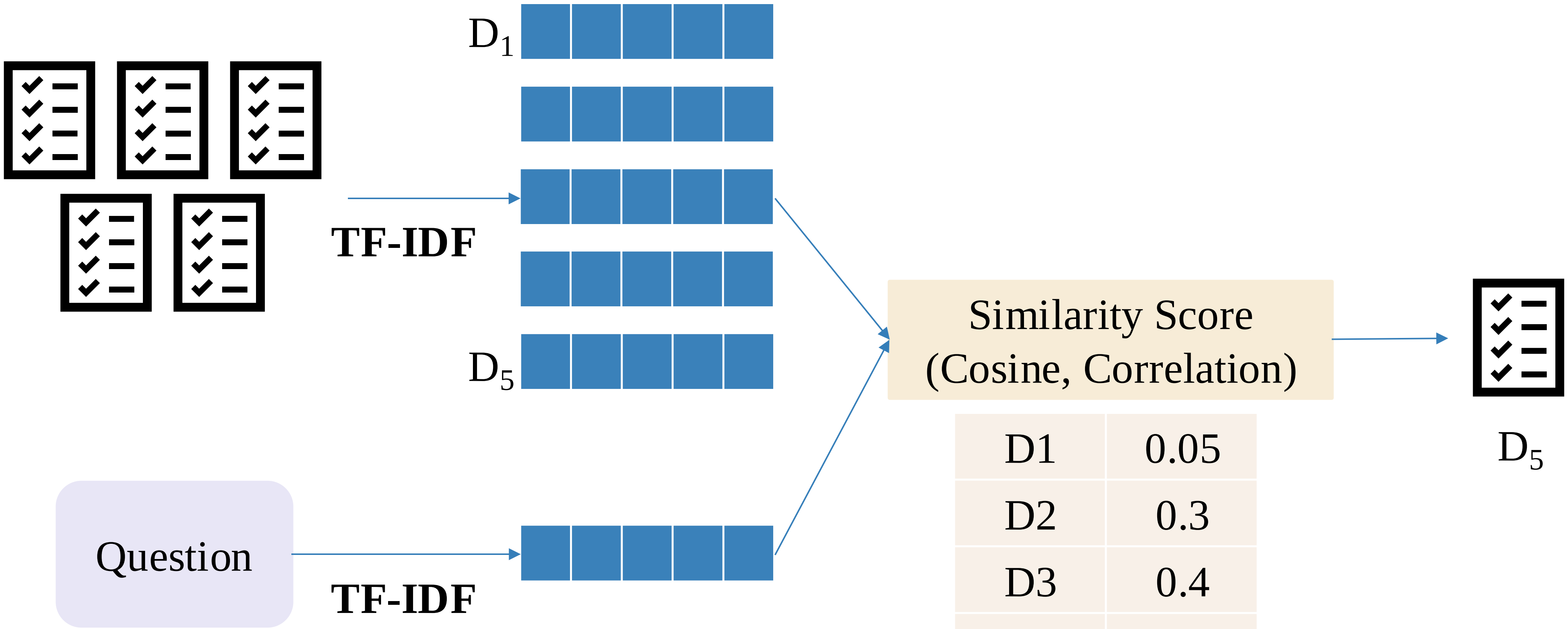
TF-IDF

```python
from sklearn.feature_extraction.text import TfidfVectorizer

tfidf_vectorizer = TfidfVectorizer()
context_embedded = tfidf_vectorizer.fit_transform(context)

question = vi_data_df.iloc[0]['question']
query_embedded = tfidf_vectorizer.transform([question.lower()])
query_embedded.shape
```

	id	question	text
0	1570446247	Quang Hải giành được chức vô địch U21 quốc gia...	Năm 2013 , Nguyễn Quang Hải giành chức vô địch...
1	1570445661	Mỗi hiệp bóng đá kéo dài bao lâu	Một trận đấu bóng đá thông thường có hai hiệp ...
2	1570382095	Quân đội Hoa Kỳ gồm những lực lượng nào	Quân đội Hoa Kỳ hay Các lực lượng vũ trang Hoa...
3	1570382072	Ngọc Lan là ai	Ngọc Lan (28 tháng 12 năm 1956 - 6 tháng 3 20...
4	1570382037	Thu Phương từng được những giải thưởng nào	Cô được coi là một trong những ca sĩ thuộc thế...

Text Retrieval

Similarity Scoring



D1	0.05
D2	0.3
D3	0.4
D4	0.2
D5	0.9

AI VIET NAM
@aivietnam.edu.vn

Text Retrieval

Cosine Similarity

$$\mathrm{cs}(\vec{x},\vec{y}) = \frac{\vec{x}.\vec{y}}{\|\vec{x}\|\,\|\vec{y}\|} = \frac{\sum_1^n x_i y_i}{\sqrt{\sum_1^n x_i^2}\sqrt{\sum_1^n y_i^2}}$$

```python
def tfidf_search(question, tfidf_vectorizer, top_d=5):
    query_embedded = tfidf_vectorizer.transform([question.lower()])
    cosine_scores = cosine_similarity(context_embedded, query_embedded).reshape((-1,))
    results = []
    for idx in cosine_scores.argsort()[-top_d:][::-1]:
        doc = {
            'id': idx,
            'cosine_score':cosine_scores[idx]
        }
        results.append(doc)
    return results
```

Text Retrieval

Correlation Similarity

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}} = \frac{n(\sum_i x_i y_i) - (\sum_i x_i)(\sum_i y_i)}{\sqrt{n\sum_i x_i^2 - (\sum_i x_i)^2}\sqrt{n\sum_i y_i^2 - (\sum_i y_i)^2}}$$

```python
def corr_search(question, tfidf_vectorizer, top_d=5):
    query_embedded = tfidf_vectorizer.transform([question.lower()])
    corr_scores = np.corrcoef(
        query_embedded.toarray()[0],
        context_embedded.toarray()
    )
    corr_scores = corr_scores[0][1:]
    results = []
    for idx in corr_scores.argsort()[-top_d:][::-1]:
        doc = {
            'id': idx,
            'corr_score':corr_scores[idx]
        }
        results.append(doc)
    return results
```

Summary

Introduction

- Random Variable
- Discrete Random Variable
- Continuous Random Variable
- Mean: $\mu = \frac{1}{n}\sum_k x_k$
- Variance: $\mathrm{Var}(X) = \frac{1}{n}\sum_k (x_k - \mu)^2$
- Standard Deviation: $\sigma(X) = \sqrt{\mathrm{Var}(X)}$
- Covariance & Correlation:
  $\mathrm{Cov}(X, Y)$, $\mathrm{Corr}(X,Y)$
- Important Probability Distributions
  - Bernoulli
  - Uniform
  - Normal

Application

- Tabular Data Analysis
  Advertising Dataset
- Text Retrieval
  TF-IDF

$$tf_{t,d} = count(t, d) \quad idf_t = \frac{N}{df_t}$$

$$w_{t,d} = tf_{t,d} \times idf_t$$

Cosine Similarity

$$\mathbf{cs}(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = \frac{\vec{\boldsymbol{x}}.\vec{\boldsymbol{y}}}{\|\vec{\boldsymbol{x}}\|\|\vec{\boldsymbol{y}}\|} = \frac{\sum_1^n x_i y_i}{\sqrt{\sum_1^n x_i^2}\sqrt{\sum_1^n y_i^2}}$$

Correlation Similarity

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Thanks!

Any questions?