IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

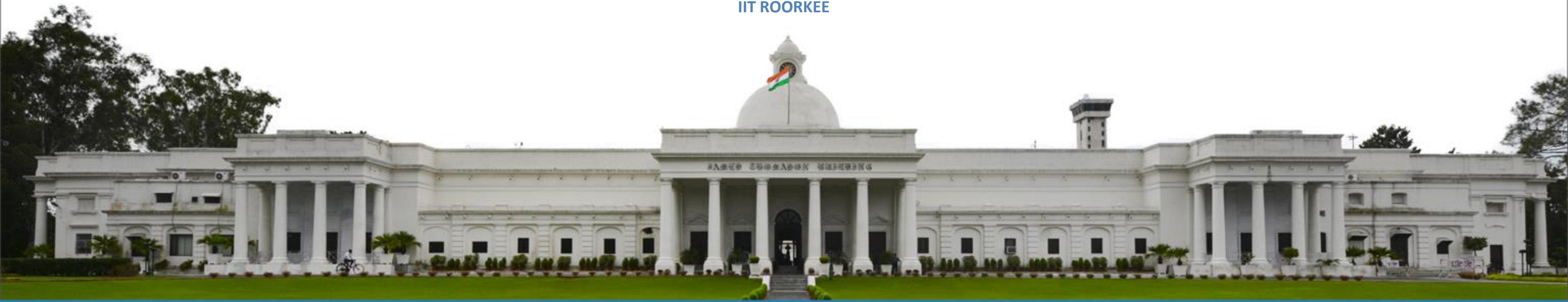
Data Analytics with Python

Lecture 8: Probability Distributions

Dr. A. Ramesh

DEPARTMENT OF MANAGEMENT

IIT ROORKEE



Lecture Objectives

- Empirical Distribution
- Discrete Distributions
- Continuous Distributions

What is a distribution?

- Describes the ‘shape’ of a batch of numbers
- The characteristics of a distribution can sometimes be defined using a small number of numeric descriptors called ‘parameters’

Why distribution?

- Can serve as a basis for standardized comparison of empirical distributions
- Can help us estimate confidence intervals for inferential statistics
- Form a basis for more advanced statistical methods
  - ‘fit’ between observed distributions and certain theoretical distributions is an assumption of many statistical procedures

Random variable

- A variable which contains the outcomes of a chance experiment
- “Quantifying the outcomes”
- Example X= (1 = Head, 0 = Tails)
- A variable that can take on different values in the population according to some “random” mechanism
- Discrete
  - Distinct values, countable
  - Year
- Continuous
  - Mass

Probability Distributions

- The probability distribution function or probability density function (PDF) of a random variable X means the values taken by that random variable and their associated probabilities.

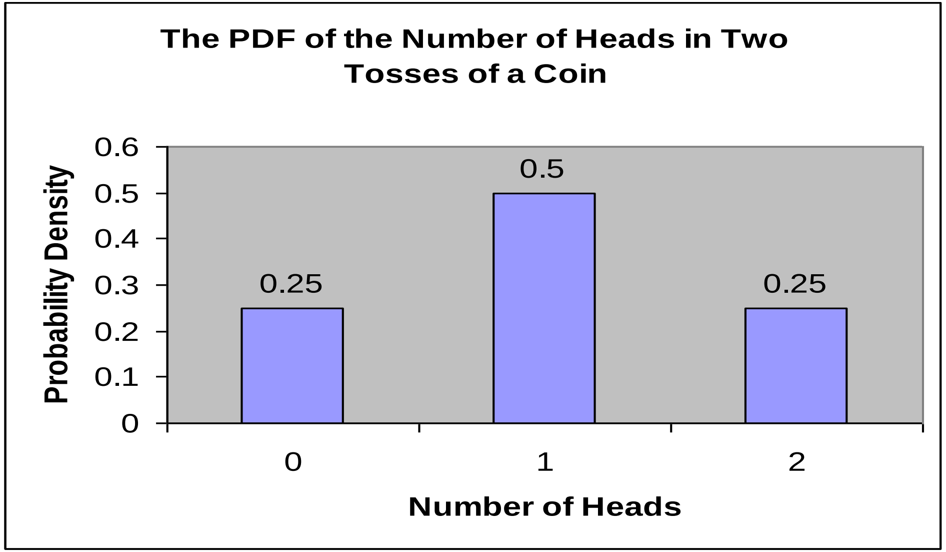
- PDF of a discrete r.v. (also known as PMF):

  Example 1: Let the r.v. X be the number of heads obtained in two tosses of a coin.

  Sample Space: {HH, HT, TH, TT}

PDF of Discrete r.v.

Number of Heads (X):	0	1	2	sum
PDF ($P(X)$):	¼	½	¼	1



Probability Distribution for the Random Variable *X*

A probability distribution for a discrete random variable X:

x	–8	–3	–1	0	1	4	6
$P(X = x)$	0.13	0.15	0.17	0.20	0.15	0.11	0.09

Find

a. $P(X \le 0)$ $\boxed{0.65}$

b. $P(-3 \le X \le 1)$ $\boxed{0.67}$

Discrete Distribution -- Example

Distribution of Daily Crises	
Number of Crises	Probability
0	0.37
1	0.31
2	0.18
3	0.09
4	0.04
5	0.01

Requirements for a Discrete Probability Function

- Probabilities are between 0 and 1, inclusively

- Total of all probabilities equals 1

$$\boxed{0 \le P(X) \le 1 \quad \text{for all X}}$$

$$\boxed{\sum_{\text{over all x}} P(X) = 1}$$

Cumulative Distribution Function

- The CDF of a random variable X (defined as $F(X)$) is a graph associating all possible values, or the range of possible values with $P(X \le x)$.
- CDFs always lie between 0 and 1 i.e., $0 \le F(X_i) \le 1$, Where $F(X_i)$ is the CDF.

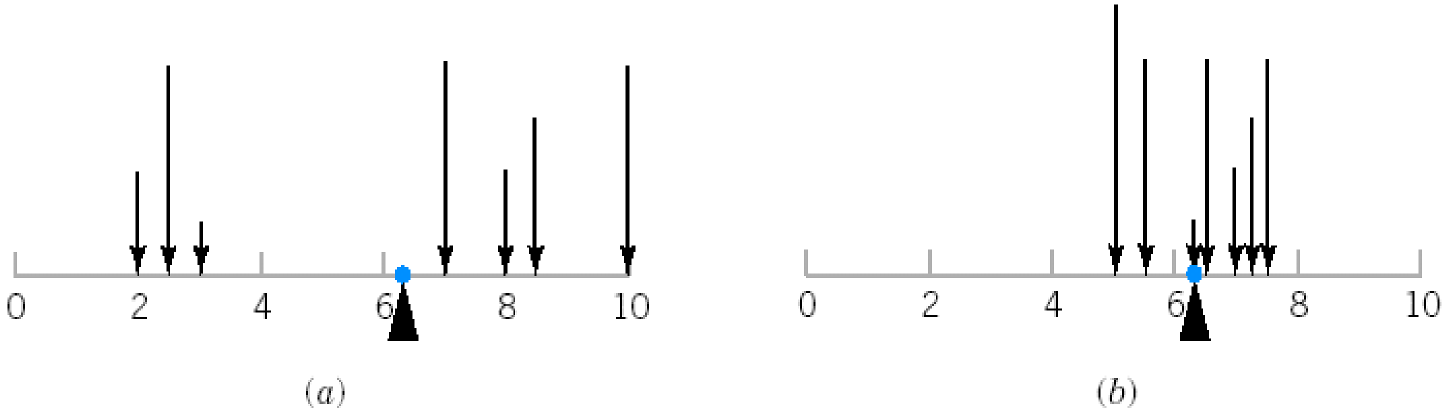
The Expected Value of X

Let X be a discrete rv with set of possible values D and pmf $p(x)$. The *expected value* or *mean value* of X, denoted

$$E(X) \text{ or } \mu_X, \text{ is}$$

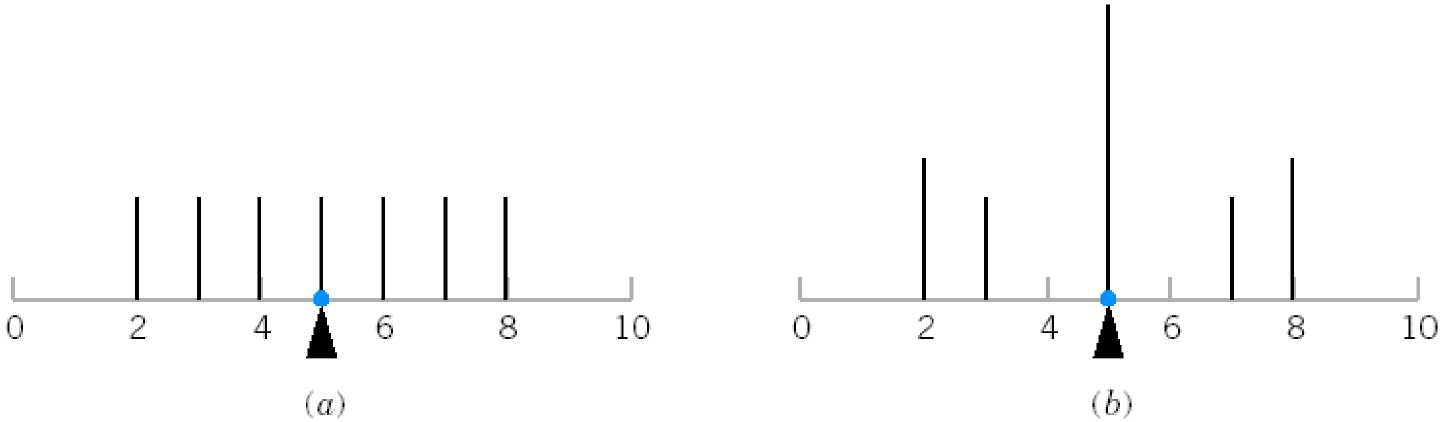
$$E(X) = \mu_X = \sum_{x \in D} x \cdot p(x)$$

Mean and Variance of a Discrete Random Variable



A probability distribution can be viewed as a loading with the mean equal to the balance point. Parts (a) and (b) illustrate equal means, but Part (a) illustrates a larger variance.

Mean and Variance of a Discrete Random Variable



The probability distribution illustrated in Parts (a) and (b) differ even though they have equal means and equal variances.

Example – Expected Value

- Use the data below to find out the expected number of credit cards that a customer to a retail outlet will possess.

x = # credit cards

x	$P(x = X)$
0	0.08
1	0.28
2	0.38
3	0.16
4	0.06
5	0.03
6	0.01

$$E(X) = x_1 p_1 + x_2 p_2 + \ldots + x_n p_n$$

$$= 0(.08) + 1(.28) + 2(.38) + 3(.16)$$

$$+ 4(.06) + 5(.03) + 6(.01)$$

$$= 1.97$$

About 2 credit cards

The Variance and Standard Deviation

Let X have pmf p(x), and expected value Then the variance of X, denoted V(X) μ

$$(\text{or } \sigma_X^2 \text{ or } \sigma^2), \text{ is}$$

$$V(X) = \sum_D (x-\mu)^2 \cdot p(x) = E[(X-\mu)^2]$$

The standard deviation (SD) of X is

$$\sigma_X = \sqrt{\sigma_X^2}$$

The quiz scores for a particular student are given below:

22, 25, 20, 18, 12, 20, 24, 20, 20, 25, 24, 25, 18

Find the variance and standard deviation.

Value	12	18	20	22	24	25
Frequency	1	2	4	1	2	3
Probability	.08	.15	.31	.08	.15	.23

$$\mu = 21$$

$$V(X) = p_1(x_1-\mu)^2 + p_2(x_2-\mu)^2 + \ldots + p_n(x_n-\mu)^2$$

$$\sigma = \sqrt{V(X)}$$

$$V(X) = .08(12-21)^2 + .15(18-21)^2 + .31(20-21)^2$$

$$+ .08(22-21)^2 + .15(24-21)^2 + .23(25-21)^2$$

$$\boxed{V(X) = 13.25}$$

$$\sigma = \sqrt{V(X)} \quad \boxed{= \sqrt{13.25} \approx 3.64}$$

Shortcut Formula for Variance

$$V(X)=\sigma^2=\left[\sum_D x^2\cdot p(x)\right]-\mu^2$$

$$=E\left(X^2\right)-\left[E(X)\right]^2$$

Mean of a Discrete Distribution

$$\boxed{\mu = E(X) = \sum X \cdot P(X)}$$

X	P(X)	X.P(X)
-1	.1	-.1
0	.2	.0
1	.4	.4
2	.2	.4
3	.1	$\underline{.3}$
		1.0

Variance and Standard Deviation of a Discrete Distribution

$$\sigma^2 = \sum (X-\mu)^2 \cdot P(X) = 1.2$$

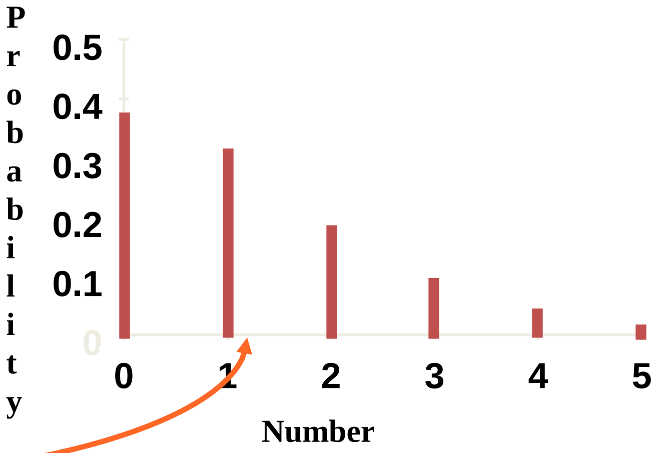
$$\sigma = \sqrt{\sigma^2} = \sqrt{1.2} \cong 1.10$$

X	P(X)	$X - \mu$	$(X-\mu)^2$	$(X-\mu)^2 \cdot P(X)$
-1	.1	-2	4	.4
0	.2	-1	1	.2
1	.4	0	0	.0
2	.2	1	1	.2
3	.1	2	4	.4
				1.2

Mean of the Data Example

$$\mu = E(X) = \sum X \cdot P(X) = 1.15$$

X	P(X)	X•P(X)
0	.37	.00
1	.31	.31
2	.18	.36
3	.09	.27
4	.04	.16
5	.01	.05
		1.15



Properties of Expected Value

1. $E(b) = b, b$ is a constant.

2. $E(X+Y) = E(X) + E(Y).$

3. $E\left(X/Y\right) \neq \dfrac{E(X)}{E(Y)}.$

4. $E(XY) \neq E(X)E(Y)$ unless they are indpendendent.

5. $E(aX) = aE(X),\ a$ constant.

6. $E(aX+b) = aE(X)+b,\ a$ and b are constants.

Properties of Variance

1. Var(constant) = 0
2. If X and Y are two independent random variables, then

   $Var(X + Y) = Var(X) + Var(Y)$ and

   $Var(X - Y) = Var(X) + Var(Y)$
3. If b is a constant then $Var(b+X) = Var(X)$
4. If a is a constant then $Var(aX) = a^2Var(X)$
5. If a and b are constants then $Var(aX+b) = a^2Var(X)$
6. If X and Y are two independent random variables and a and b are constants then $Var(aX+bY) = a^2Var(X) + b^2Var(Y)$

Covariance

Covariance: For two discrete random variables X and Y with $E(X) = \mu_x$ and $E(Y) = \mu_y$, the covariance between X and Y is defined as Cov(XY$) = \sigma_{xy} = E(X - \mu_x)\, E(Y - \mu_y) = E(XY) - \mu_x \mu_y$.

Covariance

- In general, the covariance between two random variables can be positive or negative.
- If two random variables move in the same direction, then the covariance will be positive, if they move in the opposite direction the covariance will be negative.

Properties:

1. If X and Y are independent random variables, their covariance is zero. Since $E(XY) = E(X)E(Y)$
2. Cov(XX) = Var(X)
3. Cov(YY) = Var(Y)

Correlation Coefficient

- The covariance tells the sign but not the magnitude about how strongly the variables are positively or negatively related. The correlation coefficient provides such measure of how strongly the variables are related to each other.
- For two random variables X and Y with $E(X) = \mu_x$ and $E(Y) = \mu_y$, the correlation coefficient is defined as

$$\rho_{xy} = \frac{Cov(XY)}{\sigma_x \sigma_y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Thank You