BITS Pilani
Pilani Campus

Random Variable

Akanksha Bharadwaj
Asst. Professor, BITS Pilani

innovate achieve lead

BITS Pilani
Pilani Campus

SS ZG536, ADV STAT TECHNIQUES FOR ANALYTICS
Contact Session 3

Random variable

- A random variable can be thought of as a function that associates exactly one of the possible numerical outcomes to each trial of a random experiment.
- However, that number can be the same for many of the trials.
- The set of possible values is called the **Sample Space**.
- A Random Variable is given a capital letter, such as X or Z.
- Any random variable whose only possible values are 0 and 1 is called a **Bernoulli random variable.**

True / False

Example

- Consider an experiment in which 9-volt batteries are tested until one with an acceptable voltage (S) is obtained. The sample space is {S, FS, FFS, FFFS…..}. Define a random variable X by
- X = the number of batteries tested before the experiment terminates

X = ?

1, 2, 3, 4, 5 . . . . . . ∞

or num of batteries available

Example

- Consider the random experiment of flipping a coin twice. The sample space of possible outcomes is S = { HH, HT, TH, TT }.
- Now, let's define the variable X to be the number of heads that the random experiment will produce.

$X = ?$ 0, 1, 2

- If the outcome is HH, we have two heads, so the value for X is 2.
- If the outcome is HT, we got one head, so the value for X is 1.
- If the outcome is TH, we again got one head, so the value for X is 1.
- Lastly, if the outcome is TT, we got zero heads, so the value for X is 0
- As the definition suggests, X is a quantitative variable that takes the possible values of 0, 1, or 2.

What is the probability that X will be 2?

$1/4$

Example

- Assume we choose a 13 year old boy at random and record his exact weight. The average weight for a 13-year-old boy is between 75 and 145 pounds, so the sample space here is S = { All the numbers in the interval 75-145 }.

- We'll define X to be the weight of a 13 year old boy. Here X can take any value between 75 and 145.

- What is the probability that X will be more than 120?

?

Difference between examples

- What is the difference between the random variables in these examples?
- In the first example of coins, X has three distinct possible values: 0, 1, and 2. You can list them.
- In contrast, in the second example, X takes any value in the interval 75-145, and thus the possible values of X cover an infinite range of possibilities, and cannot be listed.

Types of random variable

- A random variable values are a list of distinct values, is called a **discrete random variable**.

- A random variable that can take any value in an interval, is called a **continuous random variable**.

- A good rule of thumb is that **discrete** random variables are things we **count**, while **continuous** random variables are things we **measure**.

Probability distribution

- For a random variable X, the list of possible values and probabilities is called the **probability distribution or probability mass function(pmf)**.

- Now, let's define the variable X to be the number of heads that the random experiment will produce with 2 coins.

  HH, HT, TH, TT

  $x = 0, 1, 2$

- The **probability distribution of the random variable X** is easily summarized in a table:

X	0	1	2
P(X)	1/4	1/2	1/4

Example: Two dice are tossed

- The Random Variable is X = "The sum of the scores on the two dice".

- Let's make a table of all possible values:

		1st Die					
		1	2	3	4	5	6
2nd Die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

- Sample Space is {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}

Solution

innovate achieve lead

$P(X = 2)$ $1/36$

$P(X = 6)$ $5/36$

$P(X = 9)$ $4/36$

$P(X = 11)$ $2/36$

$P(5 \leq X \leq 8)$ $= P(X=5) + P(X=6) + P(X=7) + P(X=8)$

$4/36 + 5/36 + 6/36 + 5/36 = 20/36$

Example

- Consider a group of five potential blood donors—a, b, c, d, and e—of whom only a and b have type O+ blood. Five blood samples, one from each individual, will be typed in random order until an O+ individual is identified. Let the rv Y= the number of typings necessary to identify an O+ individual. Then the pmf of Y is

$$P(Y=1) = 2/5$$

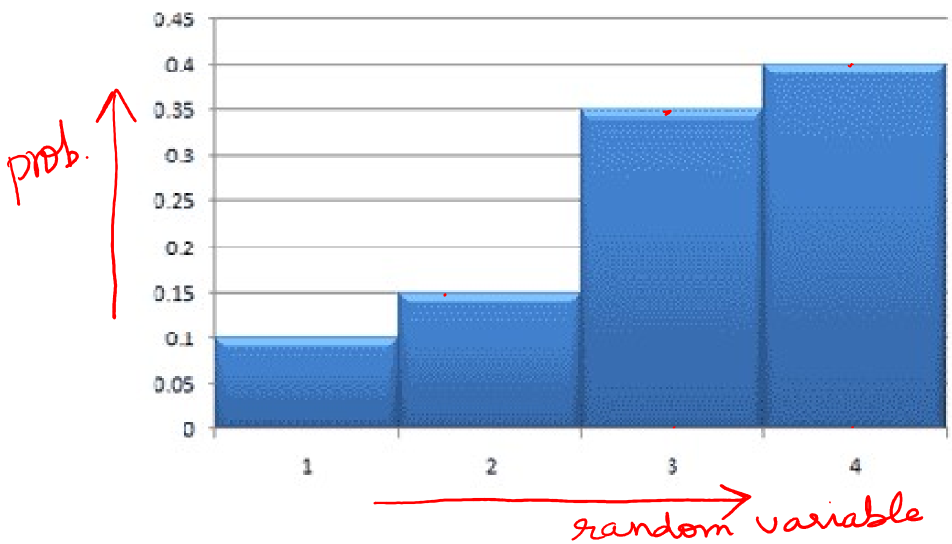
$$P(Y=2) = \frac{3}{5} * \frac{2}{4}$$

$$P(Y=3) = \frac{{}^3C_2}{{}^5C_2} * \frac{{}^2C_1}{{}^3C_1}$$

$$P(Y=4) = \frac{{}^3C_3}{{}^5C_3} * \frac{{}^2C_1}{{}^2C_1}$$

Probability distribution histogram

- The horizontal axis represents the range of all possible values of the random variable, and the vertical axis represents the probabilities of those values.



- The sum of the areas of all of the rectangles is the same as the sum of all of the probabilities.
- Therefore, the total area = 1.

Image: google

Valid Probability Model

- Alex is playing cricket. Various possible scenarios for possibility of catch on the next two balls are given below. Is it a valid model?

?

Scenarios	Probability
Miss both the catch	0.3
Miss one catch	0.4
Miss none	0.2

No

here $\sum_{\text{all}}$ probabilities $\neq 1$

Exercise

- Six lots of components are ready to be shipped by a certain supplier. The number of defective components in each lot is as follows:

Lot	1	2	3	4	5	6
Number of defectives	0	2	0	1	2	0

- One of these lots is to be randomly selected for shipment to a particular customer. Let X be the number of defectives in the selected lot. What are the possible values of X and P(X)

Solution

$$P(X=0) = 3/6$$

$$P(X=1) = 1/6$$

$$P(X=2) = 2/6$$

Cumulative distribution function

- The **cumulative distribution function** (cdf) $F(x)$ of a discrete rv variable X with pmf $p(x)$ is defined for every number x by

$$F(x) = P(X \le x) = \sum_{y: y \le x} p(y)$$

eg:- HH, HT, TH, TH

$X = 0, 1, 2$

$F(1) = P(X \le 1) = P(X=0) + P(X=1)$

Example

- A store carries flash drives with either 1 GB, 2 GB, 4 GB, 8 GB, or 16 GB of memory. The accompanying table gives the distribution of Y= the amount of memory in a purchased drive:

y	1	2	4	8	16
$p(y)$	.05	.10	.35	.40	.10

Solution

y	1	2	4	8	16
$p(y)$	.05	.10	.35	.40	.10

$F(1) = 0.05$

$F(2) = 0.05 + 0.10 = 0.15$

$$F(4) = P(Y \le 4) = P(Y = 1 \text{ or } 2 \text{ or } 4) = p(1) + p(2) + p(4) = \underline{.50}$$

$$F(8) = P(Y \le 8) = p(1) + p(2) + p(4) + p(8) = .90$$

$$F(16) = P(Y \le 16) = 1$$

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 by *E*(*X*) or μ_X or just μ, is

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

eg:-

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

$$E(x) = 0 \times 1/4 + 1 \times 1/2 + 2 \times 1/4$$
$$= 1$$

Exercise

- Just after birth, each newborn child is rated on a scale called the Apgar scale. The possible ratings are 0, 1, . . . , 10, with the child's rating determined by color, muscle tone, respiratory effort, heartbeat, and reflex irritability (the best possible score is 10). Let X be the Apgar score of a randomly selected child born at a certain hospital during the next year, and suppose that the pmf of X is

x	0	1	2	3	4	5	6	7	8	9	10
$p(x)$	.002	.001	.002	.005	.02	.04	.18	.37	.25	.12	.01

- What is the expected value of X?

$\Sigma x \cdot p(x)$

$= 7.15$

Exercise

- Pizza point delivers only one kind of pizza, which is sold for Rs150, and costs the pizza point Rs50 to make. The pizza point has the following policy regarding delivery: if the pizza takes longer than half an hour to arrive, there is no charge. Let the random variable X be the pizza point's gain for any one pizza.

- Experience has shown that delivery takes longer than half an hour only 10 percent of the time. Find the mean gain per pizza, μ_X.

Solution

- We first need to establish its probability distribution—the possible values and their probabilities.
- The random variable X has two possible values: either the pizza costs them Rs50 to make and they sell it for Rs150, in which case X takes the value 150 - 50 = Rs100, or it costs them Rs50 to make and they give it away, in which case X takes the value 0-50 = -Rs50.
- The probability of the latter case is given to be 10 percent, or .1, so using complements, the former has probability .9. Here, then is the probability distribution of X:

X	+100	-50
P(X=x)	.9	.1

- So, μ_X = (100)(.9) + (-50)*(.1)= +85
- In the long run, the pizza point gains an average of Rs85 per pizza delivered.

Expected Value of a Function

- If the rv X has a set of possible values D and pmf $p(x)$, then the expected value of any function $h(X)$,

$$E[h(X)] = \sum_{D} h(x) \cdot p(x)$$

Exercise

- The cost of a certain vehicle diagnostic test depends on the number of cylinders X in the vehicle's engine. Suppose the cost function is given by $h(X)=20+3X+.5X^2$. Since, X is a random variable, so is Y=h(X). The pmf of X is as follows:

x	4	6	8
$p(x)$	.5	.3	.2

$h(4) = 20 + 3\times4 + 0.5(4)^2 = 40$

$h(6) = 20 + 3\times6 + 0.5(6)^2 = 56$

$h(8) = 20 + 3\times8 + 0.5(8)^2 = 76$

- Calculate E(h(X))

$E(h(x)) = 40\times0.5 + 56\times0.3 + 76\times0.2$

$= 20 + 16.8 + 15.2$

$= 52$

Rules of Expected Value

Two special cases of the proposition yield two important rules of expected value.

1. For any constant *a*, $E(aX) = a*E(X)$

2. For any constant *b*, $E(X+b) = E(X)+b$

Variance and standard deviation of X

Let X have pmf $p(x)$ and expected value μ. Then the **variance** of X, denoted by $V(X)$ or σ_X^2, or just σ^2, is

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

The **standard deviation** (SD) of X is

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

Exercise

- A library has an upper limit of 6 on the number of videos that can be checked out to an individual at one time. Consider only those who check out videos, and let *X* denote the number of videos checked out to a randomly selected individual. The pmf of *X* is as follows:

x	1	2	3	4	5	6
$p(x)$	.30	.25	.15	.05	.10	.15

$E(X) = 1\times0.3 + 2\times0.25 + 3\times0.15 + 4\times0.05 + 5\times0.10 + 6\times0.15$
$= 2.85$

- Calculate variance and standard deviation

$V(x) = (1-2.85)^2\times0.3 + (2-2.85)^2\times0.25 + (3-2.85)^2\times0.15 + (4-2.85)^2\times0.05$
$+ (5-2.85)^2\times0.10 + (6-2.85)^2\times0.15 = 3.2275$

$\sigma_x = \sqrt{3.2275} = 1.8$

Continuous Variables

A random variable *X* is continuous if

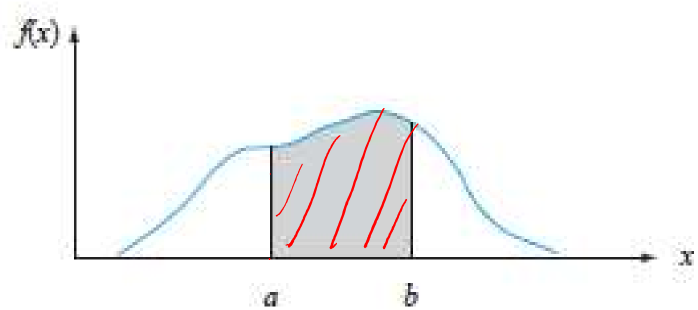
- Possible values comprise either a single interval on the number line or a union of disjoint intervals, and
- $P(X=c) = 0$ for any number *c* that is a possible value of *X*.

PDF of continuous variable

- Let *X* be a continuous rv. Then a **probability distribution** or **probability density function** (pdf) of *X* is a function *f*(*x*) such that for any two numbers *a* and *b* with a<=b ,

$$P(a \le X \le b) = \int_a^b f(x)\,dx$$

- That is, the probability that *X* takes on a value in the interval [*a*, *b*] is the area above this interval and under the graph of the density function, as illustrated



$P(a \le X \le b)$ = the area under the density curve between a and b

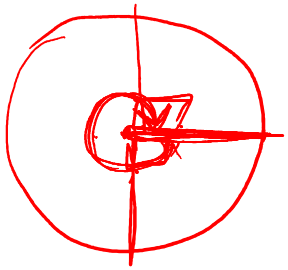
Legitimate pdf

For $f(x)$ to be a legitimate pdf, it must satisfy the following two conditions:

1. $f(x) \geq 0$ for all x
2. $\int_{-\infty}^{\infty} f(x)\,dx =$ area under the entire graph of $f(x)$
   $= 1$

Example

- Consider the reference line connecting the valve stem on a tire to the center point, and let *X* be the angle measured clockwise to the location of an imperfection. One possible pdf for *X* is



$$f(x) = \begin{cases} \frac{1}{360} & 0 \le x < 360 \\ 0 & \text{otherwise} \end{cases}$$

- The probability that the angle is between 90 degree and 180 degree is ?

$$P(90 \le X \le 180) = \int_{90}^{180} \frac{1}{360}\,dx = \frac{x}{360}\Big|_{x=90}^{x=180} = \frac{1}{4} = .25$$

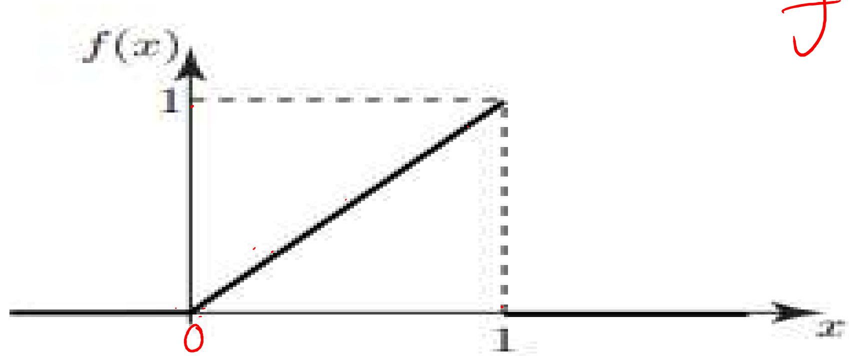
- The probability that the angle of occurrence is within 90 degree of the reference line is ? $0 \le X \le 90$ $\quad 270 \le X \le 360$

$$\int_{270}^{360} \frac{1}{360}dx + \int_{0}^{90} \frac{1}{360}dx = \left[\frac{x}{360}\right]_{0}^{90} + \left[\frac{x}{360}\right]_{270}^{360} = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \text{ or } 0.5$$

Exercise

Ques. Is this a valid pdf?

$$f(x) = x \quad 0 \le x \le 1$$
$$= 0 \quad \text{otherwise}$$



Solution

f(x)= x for 0<= x <=1 and 0 elsewhere

f(x) >= 0 for all x

But,

$$\int_0^1 f(x)\,dx = \int_0^1 x\,dx = \left[\frac{x^2}{2}\right]_0^1$$

$$= 1/2$$

Not a valid pdf

Exercise

$$\int x^n dx = \frac{x^{n+1}}{n+1}$$

$$\int c\,dx = cx$$

Is this a valid pdf?

$$f(x) = \begin{cases} x^2 - 4x + \frac{10}{3}, & 0 \leq x \leq 3 \\ 0, & \text{elsewhere} \end{cases}$$

$$f(0) = 10/3$$

$$f(1) = 1 - 4 + \frac{10}{3} = \frac{-9+10}{3} = 1/3$$

$$f(2) = 4 - 8 + \frac{10}{3} = \frac{-12+10}{3} = -2/3$$

$$f(3) = 9 - 12 + \frac{10}{3} = 1/3$$

Solution

f(x) is not >= 0 for all values of x

It is < 0 for x=2 ~~and 3~~

$$\int_0^3 \left(x^2 - 4x + \frac{10}{3}\right) dx$$

$$= \left[\frac{x^3}{3} - \frac{4x^2}{2} + \frac{10}{3}x\right]_0^3$$

$$= \frac{9\times\cancel{3}}{\cancel{3}} - \frac{\overset{2}{\cancel{4}}\times 9}{\cancel{2}} + \frac{10}{\cancel{3}}\times\cancel{3} = 1$$

Cumulative distribution function

The **cumulative distribution function** $F(x)$ for a continuous rv X is defined for every number x by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)\, dy$$

For each x, $F(x)$ is the area under the density curve to the left of x. This is illustrated in Figure, where $F(x)$ increases smoothly as x increases.

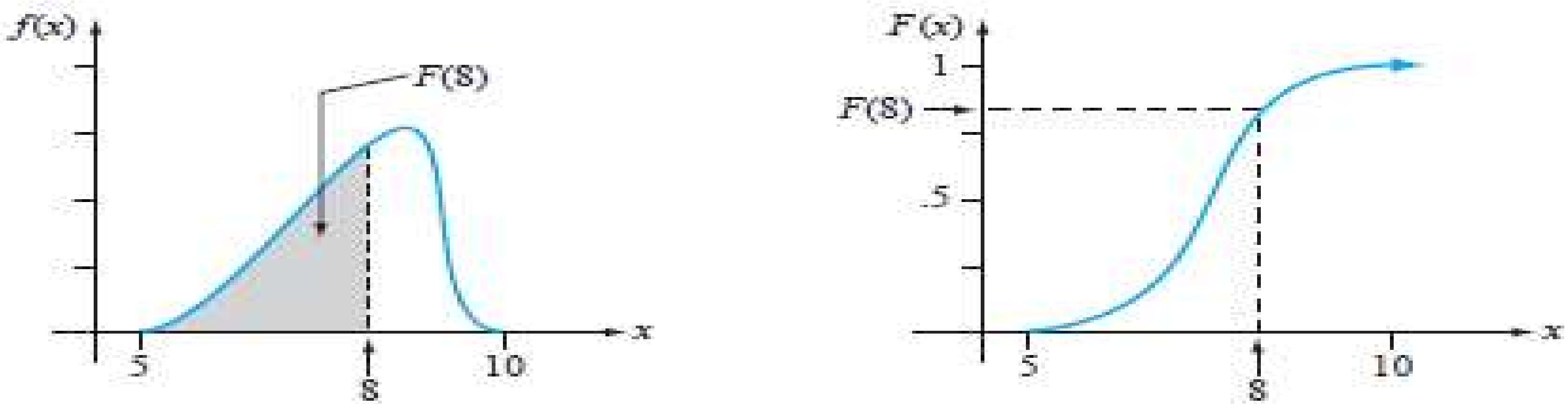


Figure 4.5 A pdf and associated cdf

Using F (x) to Compute Probabilities

Let X be a continuous rv with pdf $f(x)$ and cdf $F(x)$. Then for any number a,

$$P(X > a) = 1 - F(a)$$

and for any two numbers a and b with $a < b$,

$$P(a \leq X \leq b) = F(b) - F(a)$$

Example

- Suppose the pdf of the magnitude *X* of a dynamic load on a bridge (in newtons) is given by

$$f(x) = \begin{cases} \frac{1}{8} + \frac{3}{8}x & 0 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

- For any number *x* between 0 and 2, F(X) is

$$F(X) = \int_{-\infty}^{x} f(x)\,dx$$

$$= \int_0^2 \left(\frac{1}{8} + \frac{3}{8}x\right) dx = \left[\frac{x}{8} + \frac{3}{8} \times \frac{x^2}{2}\right]_0^2$$

$$= \left(\frac{2}{8} + \frac{3}{8} \times \frac{\cancel{4}\,2}{2}\right) = 1$$

Example

The probability that the load is between 1 and 1.5 is

$$P(1 \le X \le 1.5) = \int_1^{1.5} f(x)\,dx.$$

$$= F(1.5) - F(1)$$

$$= \left[\frac{1}{8}\times(1.5) + \frac{3}{16}(1.5)^2\right] - \left[\frac{1}{8}\times(1) + \frac{3}{16}\times(1)^2\right]$$

$$= 19/64$$

The probability that the load exceeds 1 is

$$P(X > 1) = 1 - F(1)$$

$$= 1 - \left[\frac{1}{8} + \frac{3}{16}(1)^2\right]$$

$$= 1 - \frac{5}{16} = 11/16$$

Expected Values

$p(x) = pmf$
$f(x) = pdf$

$F(X) =$ cumulative prob. dis.

The **expected** or **mean value** of a continuous rv *X* with pdf *f* (*x*) is

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

Example

- The pdf of weekly gravel sales *X* was

$$f(x) = \begin{cases} \frac{3}{2}(1 - x^2) & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

- Calculate mean value of X

$$\int_{-\infty}^{\infty} x f(x)\, dx$$

$$= \frac{3}{2} \int_0^1 x(1-x^2)\, dx = \frac{3}{2}\left[\frac{x^2}{2} - \frac{x^4}{4}\right]_0^1$$

$$= \frac{3}{2}\left[\frac{1}{2} - \frac{1}{4}\right] = 3/8$$

Variance and standard deviation

The **variance** of a continuous random variable *X* with pdf *f*(*x*) and mean value μ is

$$\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\,dx = E[(X-\mu)^2]$$

$$V(X) = E(X^2) - [E(X)]^2$$

The **standard deviation** (SD) of *X* is

$$\sigma_X = \sqrt{V(X)}.$$

Example

① A random variable X has the following probability function

(i) find the value of k

(ii) Mean

(iii) variance

(iv) $P(X \geq 3)$ (v) $P(1 < X \leq 5)$

x	1	2	3	4	5	6
P(x)	k	3k	5k	7k	9k	11k

Solution

(i) w.k.t by def of discrete random variable

$$\Rightarrow \sum_{i=1}^{n} P(x_i) = 1$$

$$\Rightarrow P(X=1) + P(X=2) + \cdots + P(X=6) = 1,$$

$$\Rightarrow \therefore K + 3K + 5K + 7K + 9K + 11K = 36K = 1$$

$$\boxed{K = \frac{1}{36}}$$

$$E(x) = \sum x p(x) = 1\cdot\frac{1}{36} + 2\cdot\frac{3}{36} + 3\cdot\frac{5}{36} + 4\cdot\frac{7}{36}$$

$$+ 5\cdot\frac{9}{36} + 6\cdot\frac{11}{36} = 4.47$$

(iii) Variance $\sigma^2 = \sum (x-\mu)^2 p(x)$

$$\Rightarrow \quad \sigma^2 = E(x^2) - [E(x)]^2$$

$$\Rightarrow \sigma^2 = \sum x^2 p(x) - [E(x)]^2$$

$$\Rightarrow \sigma^2 = \frac{1}{36} + 4\cdot\frac{(3)}{36} + 9\cdot\frac{(5)}{36} + 16\frac{(7)}{36} + 25\left(\frac{9}{36}\right)$$

$$+ 36\left(\frac{11}{36}\right) - (4.47)^2$$

$$\sigma^2 = \frac{791}{36} - (4.47)^2 = 1.99$$

$$\therefore \sigma^2 = 1.99$$

(iv) $P(X \geq 3)$

$$P(X \geq 3) = P(X=3) + P(X=4) + P(X=5) + P(X=6)$$

$$= \frac{5}{36} + \frac{7}{36} + \frac{9}{36} + \frac{11}{36} = \frac{32}{36} = \frac{8}{9}$$

$$P(X \geq 3) = \frac{8}{9}$$

(v) $P(1 < X \leq 5) = P(X=2) + P(X=3) + P(X=4) + P(X=5)$

$$= \frac{3}{36} + \frac{5}{36} + \frac{7}{36} + \frac{9}{36} = \frac{24}{36} = \frac{2}{3}$$

③. If p.d.f $f(x) = kx^3$ in $1 \le x \le 3$

0 else where.

Find the value of k and find the probability between $x = \frac{1}{2}$ and $x = \frac{3}{2}$

Sol:- $\int_{-\infty}^{\infty} f(x)\,dx = 1$

$$\Rightarrow \int_{-\infty}^{1} f(x)\,dx + \int_{1}^{3} f(x)\,dx + \int_{3}^{\infty} f(x)\,dx = 1$$

$$\Rightarrow 0 + \int_{1}^{3} kx^3\,dx + 0 = 1$$

$$\Rightarrow k\left[\frac{x^4}{4}\right]_1^3 = 1 \Rightarrow \frac{k}{4}\left[3^4 - 1\right] = 1$$

$$\Rightarrow \frac{k}{4}[81-1] = 1 \Rightarrow \frac{k}{4}(80) = 1$$

$$\Rightarrow k = \frac{4}{80} \Rightarrow \boxed{k = \frac{1}{20}}$$

$$P\left(\frac{1}{2} \leq x \leq \frac{3}{2}\right) = \int_{1/2}^{3/2} f(x)\,dx = \int_{1/2}^{3/2} kx^3\,dx$$

Homework

For the variable X with pdf, find E(X) and V(X)

$$f(x) = \begin{cases} \frac{1}{2}x, & 0 \le x \le 2 \\ 0, & \text{elsewhere} \end{cases}$$

Solution

$$E(X) = \int_0^2 \frac{1}{2} x.x\, dx = \left[\frac{1}{6}x^3\right]_0^2 = \frac{8}{6} = \frac{4}{3}.$$

$$E(X^2) = \int_0^2 \frac{1}{2} x.x^2\, dx = \left[\frac{1}{8}x^4\right]_0^2 = 2.$$

$$\begin{aligned} V(X) &= E(X^2) - \{E(X)\}^2 \\ &= 2 - \frac{16}{9} = \frac{2}{9}. \end{aligned}$$

References

- Probability and Statistics for Engineering and Sciences,8th Edition, Jay L Devore, Cengage Learning
- Applied Business Statistics, Ken Black