BITS Pilani

Pilani|Dubai|Goa|Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Session No 3

Bayes Theorem , Random Variables

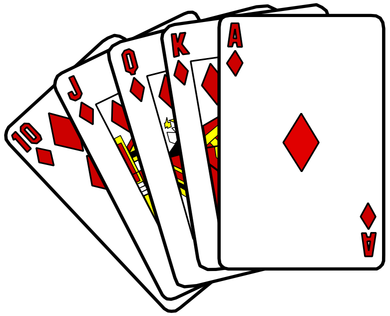
(Session 3: 20th /21st Nov 2021)

Session No 3 Course Handout

Contact Session	List of Topic Title	Reference
CS - 3	Random Variables – Discrete & Continuous (single variable)	T1:Chapter 3 & 4
HW	Problems on Random Variables	T1:Chapter 3 & 4
Lab		

Agenda ➡ Here is what you learn in the entire session

1	Definition of Random variables and Different types
2	Discrete and Continuous Probability Distributions



Variable ➡ Meaning/ Definition

In which of the following data it is possible to apply any Statistical Methods to summarize?

Height (cms) by 10 persons

168, 168, 168, 168, 168, 168, 168, 168, 168, 168

No — ∵ No changes in each of the data (Constant)

Height (cms) by 10 persons

168, 165, 178, 166, 158, 181, 154, 170, 168, 178

Yes — ∵ Changes in each of the data (Variable)

Variable ➡ Meaning/ Definition

Statistical Methods are possible to apply only if the data varies (Variable).

- **Discrete (ex. Countable #s)**
- **Continuous (ex. Real #s)**

Can the data on height (cms) be called random variable?

- **Yes/No (Why?)**
- **Let us explore the reasons after a while**

Like Height, other measurements viz., Weight, Age, BP, Wages etc are also called variables, usually denoted by X, Y, Z etc

Variable ⇨ Meaning/ Definition

Example 1

If a computer chip is selected at random, the sample space will be Ω

$\Omega = \{N, D\}$

# of events = $2^1 = 2$

Example 2

If two computer chip is selected at random, the sample space will be Ω

$\Omega = \{NN, ND, DN, DD\}$

# of events = $2^2 = 4$

Variable ⇨ Meaning/ Definition

Example 3

If three computer chips are selected at random, the sample space will be Ω

Ω = {NNN, NND, NDN, DNN, NDD, DND, DDN, DDD}

# of events = $2^3 = 8$

Example 4

If four computer chips are selected at random, the sample space will be Ω

Ω = {NNNN, NNND,…, NNDN, NDDN,…,DNDN, NDDD,…DNDD, DDDD}

# of events = $2^4 = 16$

Variable ➡ Meaning/ Definition

If the number of Computer chips selected are 5, 6, …, n, the size of the sample space also increases $2^5 = 32$, $2^6 = 64$, …, 2^n.

- Basically, instead of how many events in a Ω are there
- we are interested in counting a # of times a favourable event occurs like No. of defective chips

Variable ⇨ Meaning/ Definition

Example 1

If a computer chip is selected at random, the sample space will be Ω

Ω = {N, D}

Ω = {0, 1}

Example 2

If two computer chip is selected at random, the sample space will be Ω

Ω = {NN, ND, DN, DD}

Ω = {0, 1, 1, 2}

Favourable event (# of defectives)

No. of defective chips

Variable ⇨ Meaning/ Definition

No. of defective chips

Example 3

If three computer chips are selected at random, the sample space will be Ω

Ω = {NNN, NND, NDN, DNN, NDD, DND, DDN, DDD }

Ω = {0, 1, 1, 1, 2, 2, 2, 3 }

Example 4

If four computer chips are selected at random, the sample space will be Ω

Ω = {NNNN, NNND,…., NNDN, NDDN,….,DNDN, NDDD,…DNDD, DDDD}

Ω = {0, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4 }

Pascal triangle

Variable → Meaning/ Definition

Pascal triangle

1	n = 0
1 1	n = 1
1 2 1	n = 2
1 3 3 1	n = 3
1 4 6 4 1	n = 4
1 5 10 10 5 1	n = 5

Variable ⇨ Meaning/ Definition

Let us denote the counting numbers based on the defined favourable event by X, Y, Z etc.

- **X, Y, Z are called Random variable**
- **The examples of Height, Weight, Age, BP, Wages which are denoted as X, Y, Z, are called Variables whereas in case of example on Computer chips X, Y, Z are called Random variables.**
- **What is the difference between the two?**

Random Variable → Explanation – Every value of X based on some chance

Example 1

Outcome	X	Favourable events (m)	Probability
N	0	1	1/2
D	1	1	1/2
Total		2	1

Example 2

Outcome	X	m	Frequency (f)	Probability
NN	0	1	1	1/4
ND	1	1	2	2/4
DN	1	1		
DD	2	1	1	1/4
Total		**4**	**4**	**1**

Random variable

Explanation – Every value of X based on some chance

Example 3

Outcome	X	m	f	Probability
NNN	0	1	1	1/8
NND	1	1	3	3/8
NDN	1	1		
DNN	1	1		
NDD	2	1	3	3/8
DND	2	1		
DDN	2	1		
DDD	3	1	1	1/8
Total		8	8	1

Random variable → Probability Distribution

Example 1

X	0	1	Total
Frequency	1	1	2
P(X=x)	0.50	0.50	1

For each value taken by X, there is a probability associated called probability mass. Why?

Example 2

X	0	1	2	Total
Frequency	1	2	1	4
P(X=x)	0.25	0.50	0.25	1

Example 3

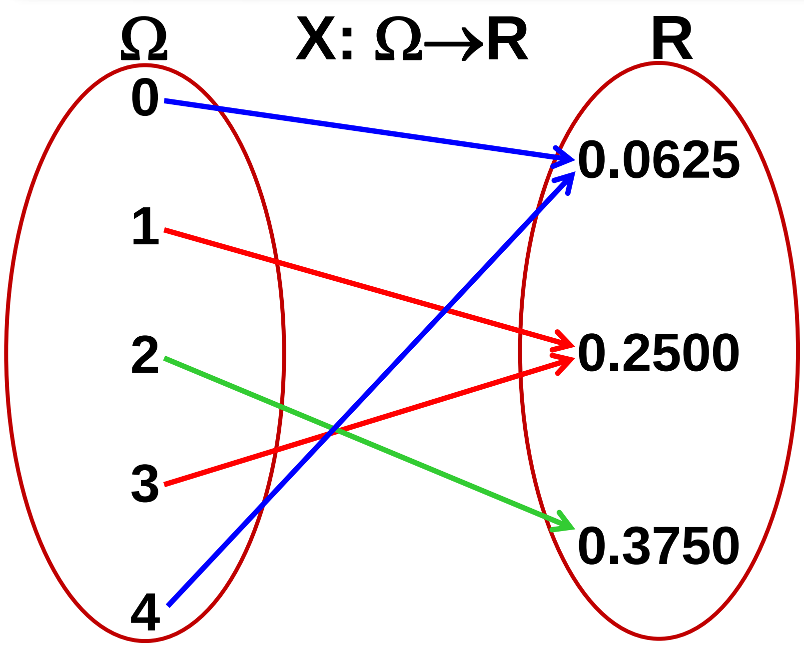
X	0	1	2	3	Total
Frequency	1	3	3	1	8
P(X=x)	0.125	0.375	0.375	0.125	1

Example 4

X	0	1	2	3	4	Total
Frequency	1	4	6	4	1	16
P(X=x)	0.0625	0.2500	0.3750	0.2500	0.0625	1

Random variable ⇨ Probability Distribution

A random variable is a real valued function defined on the sample space Ω



- **With the introduction of X we can write the probabilities as**

- **p(0) = P(X=0) = 0.0625**
- **p(1) = P(X=1) = 0.2500**
- **p(2) = P(X=2) = 0.3750**
- **p(3) = P(X=3) = 0.2500**
- **p(4) = P(x=4) = 0.0625 such that**

p(0) + p(1) + p(2) + p(3) + p(4) = 1

and each p(x) ≥ 0, x = 0, 1, 2, 3, 4

Random variable ➡ Definition

A **random variable** is a real valued function which is a mapping from the sample space Ω to the set of real numbers, ie., $X: \Omega \to R$. There are two types viz.,

- Discrete
- Continuous

Types of Random Variables

Discrete random variables

- Number of sales
- Number of calls
- Shares of stock
- People in line
- Mistakes per page

- **Continuous random variables**
  - Length
  - Depth
  - Volume
  - Time
  - Weight

Random variable ⇨ Types of random variables

Two types of random variables

Discrete random variable:
A random variable which take on countable numbers (may be finite or countable infinite values) i.e., without decimal like Natural #s, Whole #s, Integers etc.

Continuous random variable:
A random variable which take on any values in an interval i.e., in the set of real #s which includes, negative, positive, rational, irrational, decimal etc

Random variable ⇨ Probability distributions based on RVs

Two types of Probability distributions

Discrete Probability Distribution:
A probability distribution based on discrete random variable is called discrete probability distribution.

Continuous Probability Distribution:
A probability distribution based on continuous random variable is called continuous probability distribution

Discrete Probability Distribution

Probability Distribution → Meaning/ Definition

Frequency distribution

Age (yrs)	No. of Persons
≤ 1	141
2-5	187
6-12	206
13-19	353
20-29	365
30-39	386
40-49	269
50-60	63
> 60	30
Total	2000

Probability distribution

No. of defective RAM chips (X)	Probability
0	0.0625
1	0.2500
2	0.3750
3	0.2500
4	0.0625
Total	1

Probability Distribution → Definition and Properties

Definition

A random variable ‘X’ is said to have discrete probability distribution if it satisfy the following conditions

$$\text{(i)}\quad 0 \le p(x) \le 1, \text{ for all } x$$

$$\text{(ii)}\quad \sum_{\text{all } x_i} p(x) = 1$$

p(x) = P(X=x) is called probability mass function (pmf)

Probability Distribution → Discrete distribution

Example

Check whether the following can serve as Probability distributions

$$p(x) = \frac{x - 2}{2}, \text{for } x = 1, 2, 3, 4$$

No

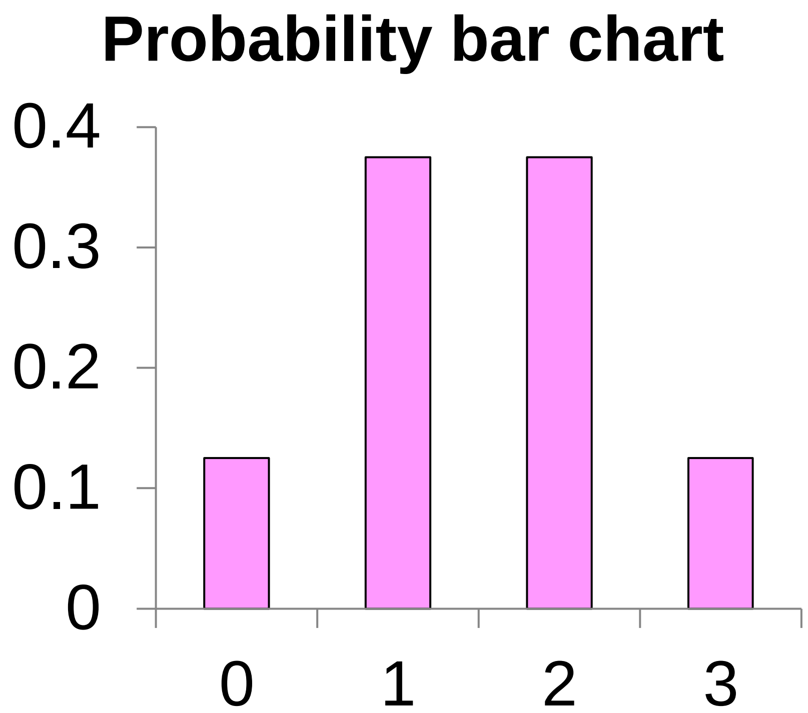
$$p(x) = cx^2, \text{for } x = 0, 1, 2, 3, 4$$

Yes with c =1/30

For what value of c, this can be a probability mass function (pmf)?

Probability Distribution → Discrete distribution

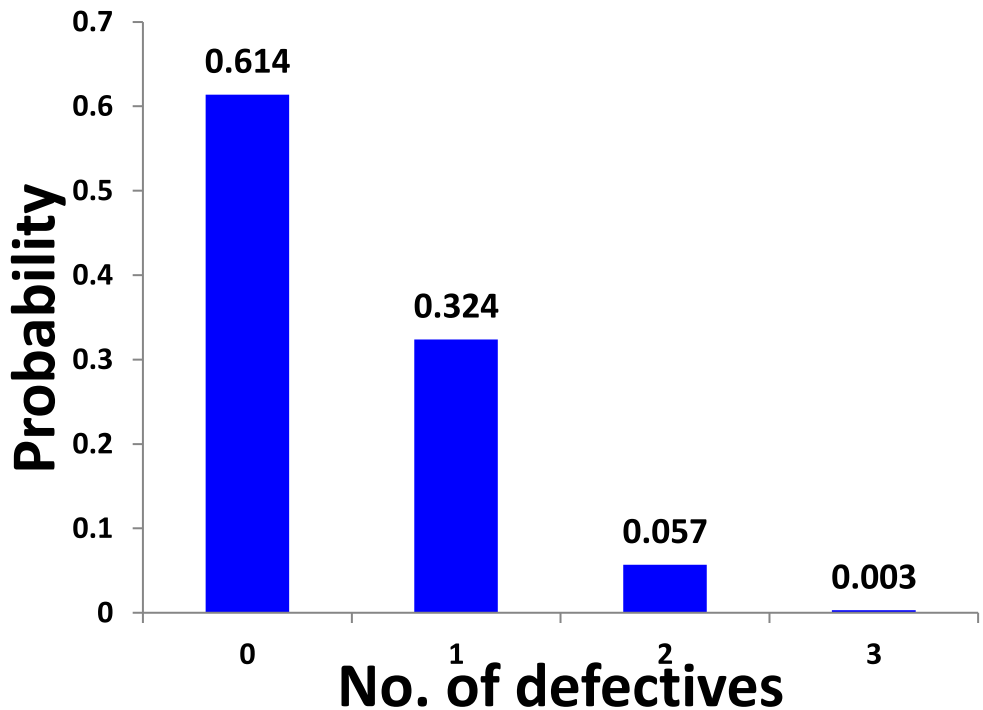
No. of defective RAM chips	Probability
0	⅛
1	⅜
2	⅜
3	⅛
Total	1



Probability Distribution → Discrete distribution

X = No. of defectives

X	0	1	2	3
P(X=x)	0.614	0.324	0.057	0.003

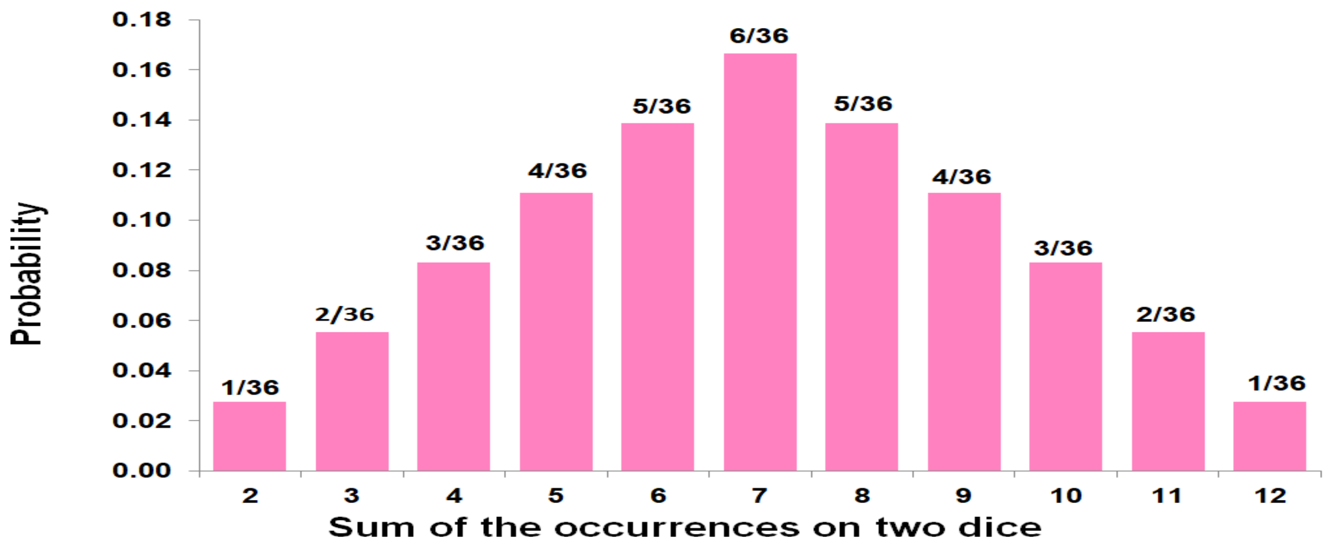


Probability Distribution → Discrete distribution

Example 2:

Let X denote the random variable that is defined as the sum of two fair dice, then,

X=x	2	3	4	5	6	7	8	9	10	11	12
P(X=x)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36



Probability Distribution ⇨ Cumulative Distribution Function

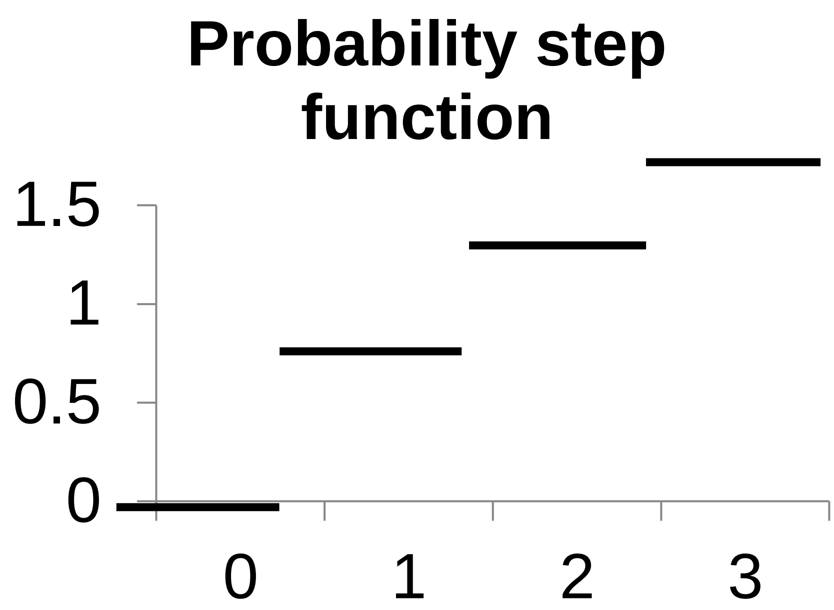
- Let p(x) =P(X=x) is called a (discrete) probability distribution.
- Let F(x) = P(X ≤ x). F(x) is called the Cumulative Distribution Function or Distribution Function (DF) of the discrete random variable X. F(x) has the following properties

$$1.\ 0 \leq F(x) \leq 1, \text{for all } x$$

$$2.\ \sum_{\text{upto } x} P(X \leq x)$$

Probability Distribution → Cumulative Distribution Function

No. of defective RAM chips	Probability	Cumulative probability
0	0.125	0.125
1	0.375	0.500
2	0.375	0.875
3	0.125	1.000
Total	1	



Probability Distribution → Discrete distribution function

Probability distributions can be estimated from relative frequencies. Consider the discrete (countable) number of televisions per household (X) from India survey data …

No. of televisions	No. of households	X	P(x)
0	1,218	0	0.012
1	32,379	1	0.319
2	37,961	2	0.374
3	19,387	3	0.191
4	7,714	4	0.076
5	2,842	5	0.028
Total	101,501		1.000

1,218 ÷ 101,501 = 0.012

e.g. $P(X=4) = P(4) = 0.076 = 7.6\%$

Probability Distribution → Discrete distribution function

What is the probability there is at least one television but no more than three in any given household?

No. of televisions	No. of households	X	P(x)
0	1,218	0	0.012
1	32,379	1	0.319
2	37,961	2	0.374
3	19,387	3	0.191
4	7,714	4	0.076
5	2,842	5	0.028
Total	101,501		1.000

"at least one television but no more than three"

$P(1 \leq X \leq 3) = P(1) + P(2) + P(3) = 0.319 + 0.374 + 0.191 = 0.884$

Probability Distribution → Cumulative Distribution Function

Let X denote the number of tires on a randomly selected automobile that are underinflated.

(a) Which of the following three probability mass functions is legitimate for X and why are the other or not allowed?

X = x	0	1	2	3	4
$p_1(x) = P(X=x)$	0.30	0.20	0.10	0.05	0.05
$p_2(x) = P(X=x)$	0.40	0.10	0.10	0.10	0.30
$p_3(x) = P(X=x)$	0.40	- 0.10	0.20	0.10	0.30

Probability Distribution ⇒ Cumulative Distribution Function

X = x	0	1	2	3	4
p(x) = P(X=x)	0.40	0.10	0.10	0.10	0.30

(b) For legitimate pmf of part (a), compute (i) $P(X \leq 2)$, (ii) $P(2 \leq X \leq 4)$ and (iii) $P(X \neq 0)$

(c) If $p(x) = c(5-x)$, $x = 0, 1, 2, 3, 4$ what is the value of c?

Probability Distribution ⇨ Cumulative Distribution Function

A mail-order computer business has six telephone lines. Let X denote the number of lines in use at a specified time. Suppose that the pmf is as follows:

X = x	0	1	2	3	4	5	6
p(x) = P(X=x)	0.10	0.15	0.20	0.25	0.20	0.06	0.04

Calculate the probability that

(a) At most 3 lines are in use
(b) Fewer than 3 line in use
(c) At least 3 lines are in use
(d) Between 2 and 5 lines are in use
(e) At least four lines are not in use

Probability Distribution ⇨ Cumulative Distribution Function

Find the probability mass function from a given distribution function of X

$$F(x) = \begin{cases} 0, & x < 0 \\ 0.06, & x \le 0 \\ 0.19, & x \le 1 \\ 0.39, & x \le 2 \\ 0.67, & x \le 3 \\ 0.92, & x \le 4 \\ 0.97, & x \le 5 \\ 1.00, & x \le 6 \end{cases}$$

P(X) = F(X) – F(X-1)

$$p(x) = \begin{cases} 0, & x < 0 \\ 0.06, & x = 0 \\ 0.13, & x = 1 \\ 0.20, & x = 2 \\ 0.28, & x = 3 \\ 0.25, & x = 4 \\ 0.05, & x = 5 \\ 0.03, & x = 6 \end{cases}$$

Probability Distribution → Discrete Distribution

How do you describe this frequency distribution?

Wages of employees (Rs)	4001-4500	4501-5000	5001-5500	5501-6000	6001-6500	6501-7000	7001-7500	7501-8000	8001-8500	Total
No. of persons	25	36	45	62	39	55	44	29	15	350

- **Measures of central tendency – Mean, Median**
- **Measures of dispersion – Range, SD, IQR, Skewness**

Probability Distribution → Expected value and Variance

How do you describe this probability distribution?

$$p(x) = \begin{cases} 0, & x < 0 \\ 0.06, & x = 0 \\ 0.13, & x = 1 \\ 0.20, & x = 2 \\ 0.28, & x = 3 \\ 0.25, & x = 4 \\ 0.05, & x = 5 \\ 0.03, & x = 6 \end{cases}$$

Measures of central tendency – Expectation (Expected value)

Measures of dispersion – Variance

Probability Distribution ⇨ Expected value and Variance

Like mean and standard deviation are computed to describe data measured by quantitative variable, a similar measures viz., expected value (mean) and variance for random variable X are computed for describing the probability distribution using the formula

Central measure and Measure of Spread

$E(X)$ | $E(X^2)$ | $V(X)$

Probability Distribution → Expected value and Variance

For a discrete random variable X with probability mass function p(x),

$$\text{Mean or Expected value} = \mu = E(X) = \sum_{\text{over all } x} x\,p(x)$$

$$E(X^2) = \sum_{\text{over all } x} x^2 p(x)$$

$$Variance = V(x) = \sigma^2 = E[(x-\mu)^2] = \sum (x-\mu)^2 p(x)$$
$$= E(x^2) - (E(x))^2$$

$$Standard\ deviation\ \sigma = \sqrt{V(x)}$$

Probability Distribution ➡ Expected value and Variance

How do you describe this probability distribution?

$$p(x) = \begin{cases} 0, & x < 0 \\ 0.06, & x = 0 \\ 0.13, & x = 1 \\ 0.20, & x = 2 \\ 0.28, & x = 3 \\ 0.25, & x = 4 \\ 0.05, & x = 5 \\ 0.03, & x = 6 \end{cases}$$

$$E(X) = 0*0.6 + 1*0.13 + \ldots + 6*0.03 =$$

$$\mathbf{E(X) = 0 + 0.13 + 0.40 + 0.84 + 1.00 + 0.25 + 0.18 = 2.8}$$

$$E(X^2) = 0^2*0.06 + 1^2*0.13 + \ldots + 6^2*0.03 =$$

$$\mathbf{E(X^2) = 0 + 0.13 + 0.80 + 2.52 + 4.00 + 1.25 + 1.08 = 9.18}$$

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

$$\mathbf{V(X) = 9.18 - (2.8^2) = 1.94}$$

Properties of Expectation

- If 'k' is a constant, then E(k) = k and E(kX) = kE(X), X is an rv
- If X and Y are two random variables $E(X \pm Y) = E(X) \pm E(Y)$
- If $X_1, X_2, \ldots, X_n$ are n RVs, then $E(\sum X) = \sum E(X)$
- If X and Y are two independent random variables (irvs), then $E(XY) = E(X)E(Y)$ and for n irvs, $E(\prod X) = \prod E(X)$

Properties of Variance

- If ‘k’ is a constant, then $V(k) = 0$ and $V(kX) = k^2V(X)$, X is an
- Given $V(X)$, for $Y = a + bX$, then $V(Y) = b^2V(X)$
- If X and Y are two random variables then

  $V(X \pm Y) = V(X) + V(Y) \pm Cov(X, Y)$
- If X and Y are independent random variables (irvs), then

  $V(X \pm Y) = V(X) + V(Y)$

Probability Distribution ⇨ Expected value and Variance

- Let X be a discrete random variable having the probability mass function

X = x	0	1	2	3	4	5	6	7
# Registered	150	450	1950	3750	k	2550	1500	300

- Find the value of k
- Find the probability distribution function of X
- Calculate E(X) and V(X)
- Assume N = 15000

Continuous Probability Distribution

Probability Distribution → Continuous probability distribution

Definition

- A continuous random variable can assume any value in an interval on the real line or in a collection of intervals.
- It is not possible to talk about the probability of the random variable assuming a particular value.
- Instead, we talk about the probability of the random variable assuming a value within a given interval.

Probability Distribution ⇨ Continuous probability distribution

Definition

- **The probability of the continuous random variable assuming a specific value is 0.**

- **The probability of the random variable assuming a value within some given interval from x_1 to x_2 is defined to be the <u>area under the graph</u> of the <u>probability density function</u> between x_1 and x_2.**

Probability Distribution → Continuous probability distribution

Definition

- A random variable is called continuous when it assumes values in a given interval.

- The probability that a random variable X assumes different values x in a given interval, say (a, b), is denoted by $f(x) = P(a \le X \le b)$, called probability density function (pdf).

Introduction

- A continuous random variable can assume any value in an interval on the real line or in a collection of intervals.
- It is not possible to talk about the probability of the random variable assuming a particular value.
- Instead, we talk about the probability of the random variable assuming a value within a given interval.

Probability Distribution → Continuous probability distribution

Introduction

- The probability of the continuous random variable assuming a specific value is 0.

- The probability of the random variable assuming a value within some given interval from x_1 to x_2 is defined to be the <u>area under the graph</u> of the <u>probability density function</u> between x_1 and x_2.

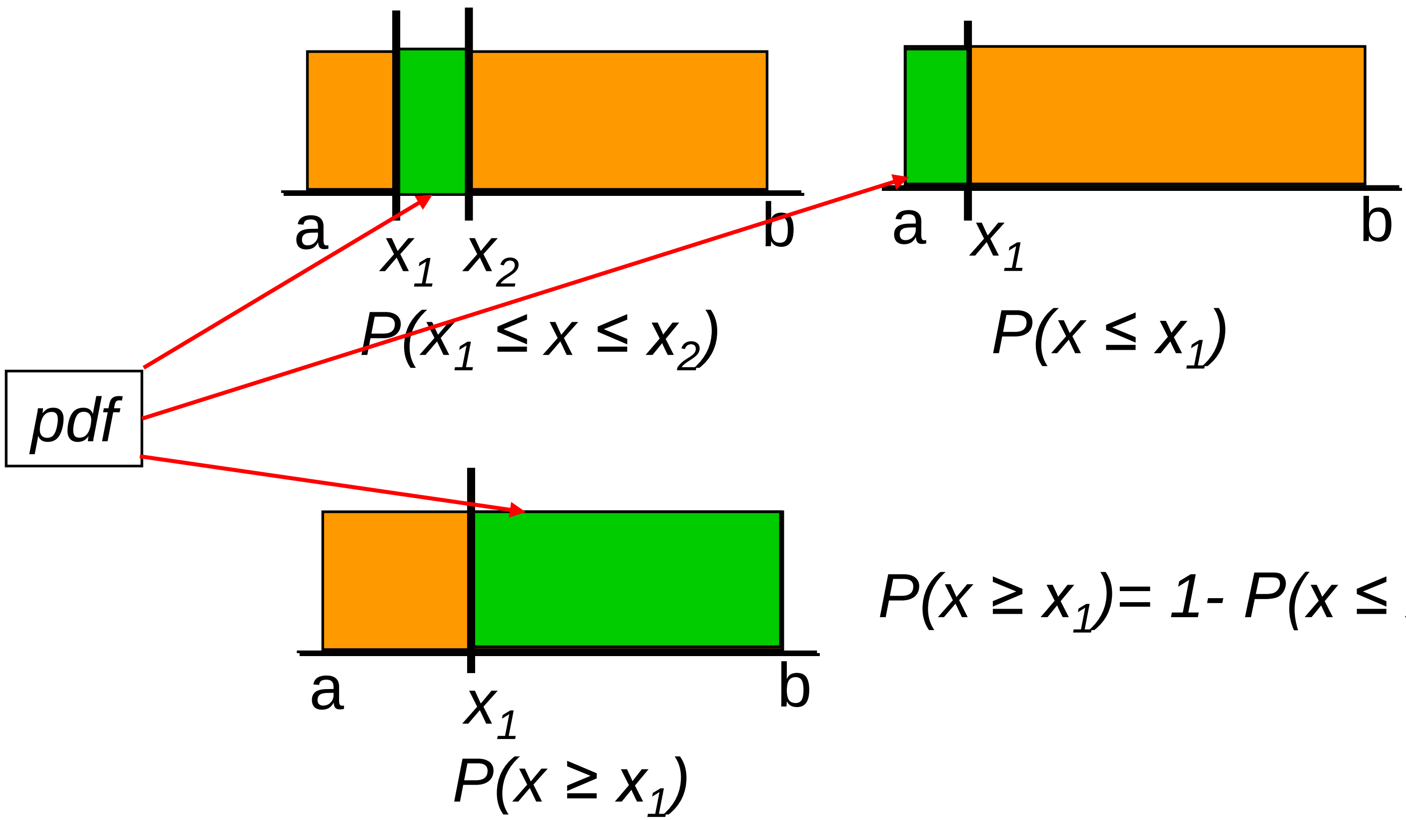
Probability Distribution → Continuous probability distribution

Introduction

- A random variable is called continuous when it assumes values in a given interval.

- The probability that a random variable X assumes different values x in a given interval, say (a, b), is denoted by $f(x) = P(a \le X \le b)$, called probability density function (pdf).

Probability Distribution → Continuous probability distribution



$$P(x \ge x_1) = 1 - P(x \le x_1)$$

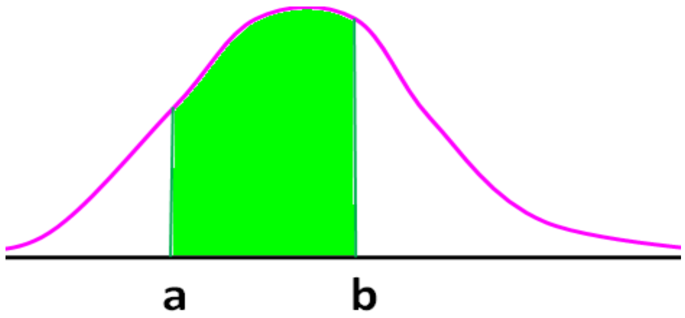
Probability Distribution ➡ Continuous probability distribution

Indeed, the probabilities are the area under the curve in a given interval. Thus, a function with values f(x) defined over the set of all real numbers (a, b) is given by

$$\mathbf{P(a \le X \le b) = \int_{a}^{b} f(x)dx}$$

Probability Distribution → Continuous probability distribution

Introduction



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

Probability Distribution ⇨ Continuous probability distribution

It is important to note that f(c), the value of the pdf of X at a constant c does not give P(X=c) as in the discrete case and in continuous case probabilities are always associated with intervals and **P(X=c) = 0**. That is

$$\mathbf{P(X = c) = P(c \le X \le c) = \int_{c}^{c} f(x)dx = 0}$$

Probability Distribution → Continuous probability distribution

An f(x) is called a **probability density function** of a continuous random variable if it satisfy the following properties

$$f(x) \geq 0$$

$$\int_a^b f(x)\,dx = 1$$

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

(i) $0 \leq p(x) \leq 1$, for all x

(ii) $\sum_{\text{all } x_i} p(x) = 1$

Discrete probability distribution

Probability Distribution → Continuous probability distribution

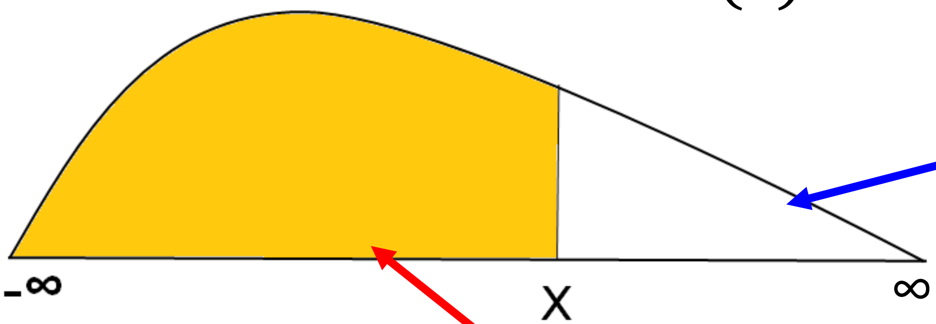
- Let $F(x) = P(X \leq x)$. F(x) is called the Distribution Function (DF) of the continuous random variable X. F(x) has the following properties.

$$1.\ F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt \qquad 2.\ 0 \leq F(x) \leq 1$$

- where f(x) is called probability density function.

Probability Distribution ⇨ Continuous probability distribution

$$1 - F(x) = P(X > x) = \int_{x}^{\infty} f(t)\,dt$$



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

Probability Distribution → Continuous probability distribution

- As in case of discrete probability distribution, the expected value E(X) and variance V(X) can be computed for the continuous probability distribution

$$\mathbf{E(X) = \int_{-\infty}^{\infty} xf(x)dx} \qquad \mathbf{E(X^2) = \int_{-\infty}^{\infty} x^2f(x)dx}$$

$$\mathbf{V(X) = E(X^2) - (E(X))^2}$$

Probability Distribution → Continuous probability distribution

- If $f(x) = 3e^{-3x}$, $x>0$
- Find E(X) and V(X)

$$\mathbf{E(X) = \int_{0}^{\infty} 3xe^{-3x}dx} \qquad \mathbf{E(X^2) = \int_{0}^{\infty} 3x^2e^{-3x}dx}$$

$$\mathbf{V(X) = E(X^2) - (E(X))^2}$$

$$\mathrm{E(X)} = \frac{1}{3} \qquad \mathrm{V(X)} = \frac{1}{9}$$

Homework Problems → Book T1: Section 3.2, Ex 13

A mail-order computer business has six telephone lines. Let X denote the number of lines in use at a specified time. Suppose the pmf of X is as given in the accompanying table.

x	0	1	2	3	4	5	6
$p(x)$	.10	.15	.20	.25	.20	.06	.04

Calculate the probability of each of the following events.

a. {at most three lines are in use}
b. {fewer than three lines are in use}
c. {at least three lines are in use}
d. {between two and five lines, inclusive, are in use}
e. {between two and four lines, inclusive, are not in use}
f. {at least four lines are not in use}

Homework Problems → Book T1: Section 3.2, Ex 23

- A consumer organization that evaluates new automobiles customarily reports the number of major defects in each car is examined. Let x denote the number of the major defects in the randomly selected car of certain type. The cdf of x is given as follows
- Calculate the following probabilities

$$F(x) = \begin{cases} 0 & x < 0 \\ 0.06 & 0 \le x < 1 \\ 0.19 & 1 \le x < 2 \\ 0.39 & 2 \le x < 3 \\ 0.67 & 3 \le x < 4 \\ 0.92 & 4 \le x < 5 \\ 0.97 & 5 \le x < 6 \\ 1 & 6 \le x \end{cases}$$

a. $p(2)$, that is, $P(X = 2)$
b. $P(X > 3)$
c. $P(2 \le X \le 5)$
d. $P(2 < X < 5)$

Homework Problems → Book T1: Section 3.3, Ex 29

- The pdf is given by

x	1	2	4	8	16
P(x)	0.05	0.10	0.35	0.40	0.10

1. Find E(x)
2. Find E(x^2)
3. Find V(x) directly from definition
4. Find V(x) using shortcut formula
5. Find $P(x \geqslant 2)$

Homework Problems → Book T1: Section 4.2, Ex 11

- **Let x denote the amount of time a book on two hour reserve is actually checked out, and suppose cdf is**

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{4} & 0 \le x < 2 \\ 1 & 2 \le x \end{cases}$$

a. $P(X \le 1)$

b. $P(0.5 \le X \le 1)$

c. $P(X > 1.5)$

d. $E(X)$

e. V(X)

Homework Problems → Probability distribution

- **The pdf of weekly gravel sales is given by**

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

1. **Find E(x)**
2. **Find E(x^2)**
3. **Find V(x)**

Homework Problems → Probability distribution

- **Let x be a random variable with pdf given by**

$$f(x) = \begin{cases} cx^2, & -1 \leqslant x \leqslant 1 \\ 0, & otherwise \end{cases}$$

1. **Find constant 'c'**
2. **Find E(x)**
3. **Find V(x)**
4. **Find** $P(x \geqslant 1/2)$

Thanks