

---

## UNIT 1- MEASURES OF MOMENTS, SKEWNESS AND KURTOSIS

---

### Structure

Objectives  
Introduction  
Concept of Skewness  
Karl Pearson's Measure of Skewness  
Bowley's Measure of Skewness  
Kelly's Measure of Skewness  
Moments  
Concept and Measure of Kurtosis  
Sum Up

---

### OBJECTIVES

---

After going through this Unit, you will be able to :

- distinguish between a symmetrical and a skewed distribution;
- compute various coefficients to measure the extent of skewness in a distribution;
- distinguish between platykurtic, mesokurtic and leptokurtic distributions; and
- compute the coefficient of kurtosis.

---

### INTRODUCTION

---

In this Unit you will learn various techniques to distinguish between various shapes of a frequency distribution. This is the final Unit with regard to the summarisation of univariate data. This Unit will make you familiar with the concept of skewness and kurtosis. The need to study these concepts arises from the fact that the measures of central tendency and dispersion fail to describe a distribution completely. It is possible to have frequency distributions which differ widely in their nature and composition and yet may have same central tendency and dispersion. Thus, there is need to supplement the measures of central tendency and dispersion. Consequently, in this Unit, we shall discuss two such measures, viz, measures of skewness and kurtosis.

---

### CONCEPT OF SKEWNESS

---

The skewness of a distribution is defined as the lack of symmetry. In a symmetrical distribution, the Mean, Median and Mode are equal to each other and the ordinate at mean divides the distribution into two equal parts such that one

part is mirror image of the other (Fig. 6.1). If some observations, of very high (low) magnitude, are added to such a distribution, its right (left) tail gets elongated.

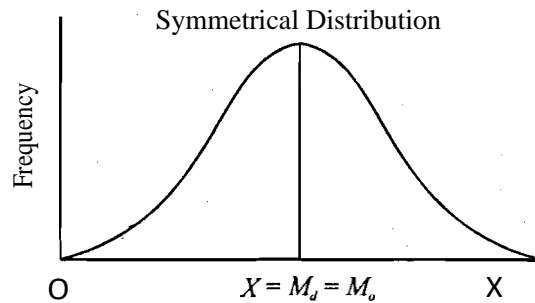


Fig. 1

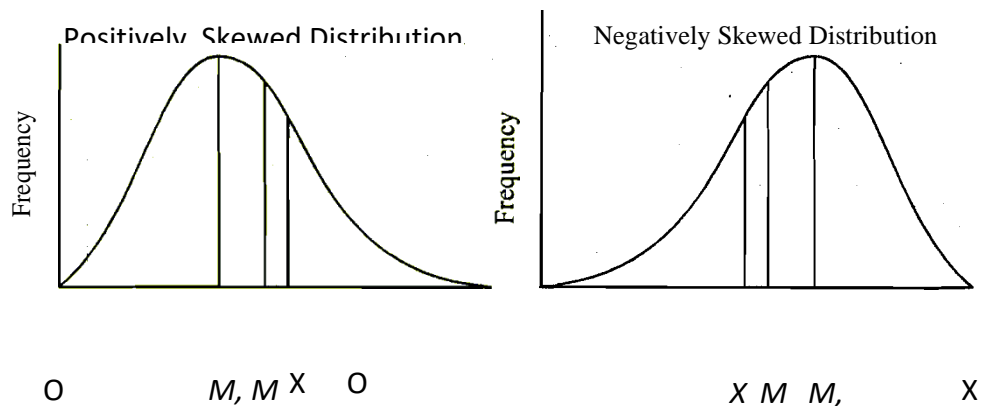


Fig. 2

These observations are also known as extreme observations. The presence of extreme observations on the right hand side of a distribution makes it positively skewed and the three averages, viz., mean, median and mode, will no longer be equal. We shall in fact have  $\text{Mean} > \text{Median} > \text{Mode}$  when a distribution is positively skewed. On the other hand, the presence of extreme observations to the left hand side of a distribution make it negatively skewed and the relationship between mean, median and mode is:  $\text{Mean} < \text{Median} < \text{Mode}$ . In Fig. 2 we depict the shapes of positively skewed and negatively skewed distributions.

The direction and extent of skewness can be measured in various ways. We shall discuss four **measures** of skewness in this Unit.

### Karl Pearson's Measure of Skewness

In Fig. 2 you noticed that the mean, median and mode are not equal in a skewed distribution. The Karl Pearson's measure of skewness is based upon the *divergence of mean from mode* in a skewed distribution.

Since  $\text{Mean} = \text{Mode}$  in a symmetrical distribution,  $(\text{Mean} - \text{Mode})$  can be taken as

an *absolute measure of skewness*. The absolute measure of skewness for a distribution depends upon the unit of measurement. For example, if the mean = 2.45 metre and mode = 2.14 metre, then absolute measure of skewness will be  $2.45 \text{ metre} - 2.14 \text{ metre} = 0.31 \text{ metre}$ . For the same distribution, if we change the unit of measurement to centimetres, the absolute measure of skewness is 245

centimetre — 214centimetre = 31centimetre. In order to avoid such a problem Karl Pearson takes a relative measure of skewness  $S_k$ .

A relative measure, independent of the units of measurement, is defined as the *Karl Pearson's Coefficient of Skewness*  $St$ , given by

$$S_k = \frac{\text{Mean} - \text{Mode}}{\text{s.d.}}$$

The sign of  $St$  gives the direction and its magnitude gives the extent of skewness.

If  $St > 0$ , the distribution is positively skewed, and if  $St < 0$  it is negatively skewed. So

far we have seen that  $St$  is strategically dependent upon mode. If mode is not defined for a distribution we cannot find  $St$ . But empirical relation between mean, median and mode states that, for a moderately symmetrical distribution, we have

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median})$$

Hence Karl Pearson's coefficient of skewness is defined in terms of median as

$$S_k = \frac{3(\text{Mean} - \text{Median})}{\text{s.d.}}$$

Example 6.1: Compute the Karl Pearson's coefficient of skewness from the following data:

Table 6.1

Height (in inches)	Number of Persons
58	10
59	18
60	30
61	42
62	35
63	28
64	16
65	8

Table for the computation of mean and s.d.

Height(	n=X-61	No. of persons		/u°
58	-3	10	-30	90
59	-2	18	-36	72
60	-1	30	-30	30
61	0	42	0	0
62	1	35	35	35
63	2	28	56	112
64	3	16	48	144
65	4	8	32	128



Total		187	75	611
-------	--	-----	----	-----

$$\text{Mean} = 61 + \frac{75}{18} = 61.4$$

$$\text{s.d.} = \sqrt{\frac{611}{187} - \left(\frac{75}{187}\right)^2} = 1.76$$

To find mode, we note that height is a continuous variable. It is assumed that the height has been measured under the approximation that a measurement on height that is, e.g., greater than 58 but less than 58.5 is taken as 58 inches while a measurement greater than or equal to 58.5 but less than 59 is taken as 59 inches. Thus the given data can be written as

Height (in inches)	No. of persons
57.5 - 58.5	10
58.5 - 59.5	18
59.5 - 60.5	30
60.5 - 61.5	42
61.5 - 62.5	35
62.5 - 63.5	28
63.5 - 64.5	16
64.5 - 65.5	8

By inspection, the modal class is 60.5 — 61.5. Thus, we have

$$l = 60.5, f_t = 42 - 30 = 12, A_2 = 42 - 35 = 7 \text{ and } f_i = 1.$$

$$\text{Mode} = 60.5 + \frac{12}{12+7} \times 1 = 61.13$$

$$\text{Hence, the Karl Pearson's coefficient of skewness } S_k = \frac{61.4 - 61.13}{1.76} = 0.153.$$

Thus the distribution is positively skewed.

### Bowley's Measure of Skewness

This measure is based on quartiles. For a symmetrical distribution, it is seen that  $Q_3$  and  $Q_1$  are equidistant from median. Thus  $Q_3 - M_p = M_p - Q_1$  can be taken as an absolute measure of skewness.

A relative measure of skewness, known as Bowley's coefficient ( $S_b$ ), is given by

$$S_b = \frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)}$$

$$= \frac{Q_3 - 2M_d + Q_1}{Q_3 + Q_1 - 2M_d}$$



The Bowley's coefficient for the data on heights given in Table 6.1 is computed below. Skewness and

**Kurtosis**

Height (in inches)	No. of persons (/)	Cumulative Frequency
57.5 - 58.5	10	10
58.5 - 59.5	18	28
59.5 - 60.5	30	58
60.5 - 61.5	42	100
61.5 - 62.5	35	135
62.5 - 63.5	28	163
63.5 - 64.5	16	179
64.5 - 65.5	8	187

*Computation of  $Q_1$  :*

Since  $\frac{N}{4} = 46.75$ , the first quartile class is 59.5 — 60.5. Thus  $q_1$ ,

$$= 59.5, C = 28, /q_1 = 30 \text{ and } h = 1.$$

$$Q_1 = 59.5 + \frac{46.75 - 28}{30} \cdot 1 = 60.125.$$

*Computation of  $M$  ( $Q_2$ ) :*

Since  $\frac{N}{2} = 93.5$ , the median class is 60.5 — 61.5. Thus

$$lq = 60.5, C = 58, fg = 42 \text{ and } h = 1.$$

$$M_q = 60.5 + \frac{93.5 - 58}{42} \cdot 1 = 61.345.$$

*Computation of  $Q_3$  :*

Since  $\frac{3N}{4} = 140.25$ , the third quartile class is 62.5 — 63.5. Thus

$$lq_3 = 62.5, C = 135, /q_3 = 28 \text{ and } h = 1.$$

$$Q_3 = 62.5 + \frac{140.25 - 135}{28} \cdot 1 = 62.688.$$

$$\text{Hence, Bowley's coefficient } S_p = \frac{62.688 - 2 \times 61.345 + 60.125}{62.688 - 60.125} = 0.048.$$

### **Kelly's Measure of Skewness**

Bowley's measure of skewness is based on the middle 50% of the observations because it leaves 25% of the observations on each extreme of the distribution. As an improvement over Bowley's measure, Kelly has suggested a measure based on  $P_{10}$  and  $P_{90}$  so that only 10% of the observations on each extreme are

ignored.

Kelly's coefficient of skewness, denoted by  $S$  is given by

$$P' = \frac{(r_0 - s_0) * (s_0 - i_0)}{(r_0 - s_0) + (s_0 - i_0)}$$

$$\frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - 2P_{50} + P_{10}}$$

Note that  $P_{50} = M_d$  (median).

The value of  $S_q$ , for the data given in Table 6.1, can be computed as given below.

*Computation of  $P_{50}$ :*

Since  $\frac{10N}{100} - \frac{10 \times 187}{100} = 18.7$ , 18.7th percentile lies in the class 58.5 – 59.5. Thus

$$l_{p50} = 58.5, C = 10, p_{50} = 18 \text{ and } h = 1.$$

$$\therefore P_{50} = 58.5 + \frac{18.7 - 10}{18} \times 1 = 58.983$$

*Computation of  $P_{90}$ :*

Since  $\frac{90N}{100} - \frac{90 \times 187}{100} = 168.3$ , 90th percentile lies in the class 63.5 – 64.5. Thus

$$l_{p90} = 63.5, C = 163, p_{90} = 16 \text{ and } h = 1.$$

$$P_{90} = 63.5 + \frac{168.3 - 163}{16} \times 1 = 63.831$$

$$\text{Hence, Kelly's coefficient } S_p = \frac{63.831 - 2 \times 58.983 + 58.983}{63.831 - 58.983} = 0.026.$$

It may be noted here that although the coefficient  $S_p$ ,  $\rho_S$  and  $S_p$  are not comparable, however, in the absence of skewness, each of them will be equal to zero.

**Check Your Progress 1**

1) Compute the Karl Pearson's coefficient of skewness from the following data :

Daily Expenditure (Rs.) :	0-20	20-40	40-60	60-80	80-100
No. of families :	13	25	27	19	16



- 2) The following figures relate to the size of capital of 285 companies :

Capital (in Rs. lacs.)	1-5	6-10	11-15	16-20	21-25	26-30	31-35	Total
No. of companies	R	27	B	38	48	53	70	285

Compute the Bowley's and Kelly's coefficients of skewness and interpret the results.

---



---



---



---



---

- 3) The following measures were computed for a frequency distribution :

Mean = 50, coefficient of Variation = 35% and

Karl Pearson's Coefficient of Skewness = - 0.25.

Compute Standard Deviation, Mode and Median of the distribution.

---



---



---



---

## MOMENTS

The  $r$ th moment about mean of a distribution, denoted by  $\mu_r$ , is given by

$$\mu_r = \frac{1}{N} \sum_{i=1}^n f_i (X_i - \bar{X})^r \quad \text{where } r = 0, 1, 2, 3, 4, \dots$$

Thus,  $r$ th moment about mean is the mean of the  $r$ th power of deviations of observations from their arithmetic mean. In particular,

$$\text{if } r = 0, \text{ we have } \mu_0 = \frac{1}{N} \sum_{i=1}^n f_i (X_i - \bar{X})^0 = 1$$

$$\text{if } r = 1, \text{ we have } \mu_1 = \frac{1}{N} \sum_{i=1}^n f_i (X_i - \bar{X}) = 0$$



if  $r = 2$ , we have  $\mu_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \mu^2$ ,

if  $r = 3$ , we have  $\mu_3 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^3$  and so on.

These moments are also known as *central moments*.

In addition to the above, we can define *raw moments* as moments about any arbitrary mean.

Let  $A$  denote an arbitrary mean, then  $r$ th moment about  $A$  is defined as

$$\mu'_r = \frac{1}{N} \sum_{i=1}^n f_i (X_i - A)^r, r = 0, 1, 2, 3, \dots$$

When  $A = 0$ , we get various moments about origin.

### Moment Measure of Skewness

The moment measure of skewness is based on the property that, for a symmetrical distribution, all odd ordered central moments are equal to zero.

We note that  $\mu'_1 = 0$ , for every distribution, therefore, the lowest order moment that can provide an absolute measure of skewness is  $\mu'_3$ .

Further, a coefficient of skewness, independent of the units of measurement, is given by

$\gamma_3 = \frac{\mu'_3}{\sigma^3} = \frac{\mu'_3}{\sigma^3} = \frac{\mu'_3}{\sigma^3}$ , where  $Q_3$  and  $\gamma_3$  are defined as the *third beta* and *third gamma* coefficients respectively.  $Q_4$  is measure of kurtosis as you will come to know in the next Section.

Very often, the skewness is measured in terms of  $\frac{\mu'_3}{\sigma^3}$ , where the sign of skewness is determined by the sign of  $\mu'_3$ .

**Example 6.2:** Compute the Moment coefficient of skewness ( $Q_3$ ) from the following data.

Marks Obtained :	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency :	6	12	22	24	16	12	8

Table for the computations of mean, s.d. and  $\mu'_3$ .

Class Intervals	Frequency (f)	Mid-values (J)	$\frac{J-35}{10}$		$fu^2$	$fu^3$
0 - 10	6	5	-3	-18	54	-162
10 - 20	12	15	-2	-24	48	-96
20 - 30	22	25	-1	-22	22	-22
30 - 40	24	35	0	0	0	0
40 - 50	16	45	1	16	16	16
50 - 60	12	55	2	24	48	96
60 - 70	8	65	3	24	72	216
<b>Total</b>	100			0	260	48

Since  $\sum fJ = 0$ , the mean of the distribution is 35.

The second moment  $\mu_2$  is equal to the variance (H) and its positive square root is equal to standard deviation (w).

$$\mu_2 = \frac{260^2}{100} = 676, \text{ and}$$

$$\text{s.d. } w = \sqrt{676} = 26.$$

$$\text{Also } \mu_1 = \frac{48}{100} = 0.48.$$

$$\text{Thus, } \beta_1 = \frac{(480)^2}{(260)^3} = 0.01.$$

Since the sign of  $\beta_1$  is positive and  $\beta_1$  is small, the distribution is slightly positively skewed.

If the mean of a distribution is not a convenient figure like 35, as in the above example, the computation of various central moments may become a cumbersome task. Alternatively, we can first compute raw moments and then convert them into central moments by using the equations obtained below.

### Conversion of Raw Moments into Central Moments

We can write

$$\begin{aligned} \mu_r &= \frac{1}{N} \sum_{i=1}^n f_i (X_i - \bar{X})^r = \frac{1}{N} \sum_{i=1}^n f_i [(X_i - A) - (\bar{X} - A)]^r \\ &= \frac{1}{N} \sum_{i=1}^n f_i [(X_i - A) - h]^r \quad (\text{since } \bar{X} - A = h) \end{aligned}$$

Expanding the term within brackets by *binomial theorem*, we get

$$\begin{aligned} &= \frac{1}{N} \sum_{i=1}^n f_i \left[ {}^r C_0 (X_i - A)^r - {}^r C_1 (X_i - A)^{r-1} h + {}^r C_2 (X_i - A)^{r-2} h^2 - \dots \right] \\ &= \frac{1}{N} \sum_{i=1}^n f_i (X_i - A)^r - {}^r C_1 h \frac{1}{N} \sum_{i=1}^n f_i (X_i - A)^{r-1} + {}^r C_2 h^2 \frac{1}{N} \sum_{i=1}^n f_i (X_i - A)^{r-2} - \dots \end{aligned}$$

From the above, we can write

$$\mu_r = \mu_r' - {}^r C_1 \mu_{r-1}' h + {}^r C_2 \mu_{r-2}' h^2 - {}^r C_3 \mu_{r-3}' h^3 + \dots$$

In particular, taking  $r = 2, 3, 4$ , etc., we get

$$\mu_2 = \mu'_2 - {}^2C_1 \mu_1'^2 + {}^2C_2 \mu'_0 \mu_1'^2 = \mu'_2 - \mu_1'^2 \text{ (since } q_1 = 1)$$

$$\mu_3 = \mu'_3 - 3\mu'_2 \mu_1' + 3\mu_1'^3 - \mu_1'^3 = \mu'_3 - 3\mu'_2 \mu_1' + 2\mu_1'^3$$

Example 6.3: Compute the first four moments about mean from the following data.

Class Intervals .	0 - 10	10 - 20	20 - 30	30 - 40
Frequency (/)	1	3	4	2

Table for computations of raw moments (Take A = 25).

Class Intervals	f	Mid-Value	$u = \frac{x - 25}{10}$	fu	fu <sup>2</sup>	fu <sup>3</sup>	fu <sup>4</sup>
<b>0-10</b>	1	5	-2	-2	4	-8	16
<b>10-20</b>	3	15	-1	-3	3	-3	3
<b>20-30</b>	4	25	0	0	0	0	0
<b>30-40</b>	2	35	1	2	2	2	2
<b>Total</b>	<b>10</b>			- L	9	- 9	21

From the above table, we can write

$$\mu'_1 = \frac{-3 \times 10}{10} = -3,$$

$$\mu'_2 = \frac{9 \times 10^2}{10} = 90,$$

$$\mu'_3 = \frac{-9 \times 10^3}{10} = -900 \text{ and}$$

$$\mu'_4 = \frac{21 \times 10^4}{10} = 21000$$

### Moments about Mean

By definition,

$$p_1 = 0,$$

$$p_2 = 90 - 9 = 81,$$

$$p_3 = -900 - 3 \times 90 \times (-3) + 2 \times (-3) = -900 + 810 - 6 = -90 \text{ and}$$

$$p_4 = 21000 - 4 \times (-900) \times (-3) + 6 \times 90 \times (-3)^2 - 3 \times (-3)^3$$

$$= 21000 - 10800 + 4860 - 243 = 14817.$$

### Check Your Progress 2

1) Calculate the first four moments about mean for the following

distribution. Also calculate Q, and comment upon the nature of skewness.

Marks	:	0 - 20	20 - 40	40 - 60	60 - 80	80 - 100
Frequency	:	8	28	35	17	12

- 2) The first three moments of a distribution about the value 3 of a variable are 2, 10 and 30 respectively. Obtain  $\mu_2$ ,  $\mu_3$ , and hence  $\mu_4$ . Comment upon the nature of skewness.

---

### **CONCEPT AND MEASURE OF KURTOSIS**

---

Kurtosis is another measure of the shape of a distribution. Whereas skewness measures the lack of symmetry of the frequency curve of a distribution, kurtosis is a measure of the relative peakedness of its frequency curve. Various frequency curves can be divided into three categories depending upon the shape of their peak. The three shapes are termed as Leptokurtic, Mesokurtic and Platykurtic as shown in Fig.3.

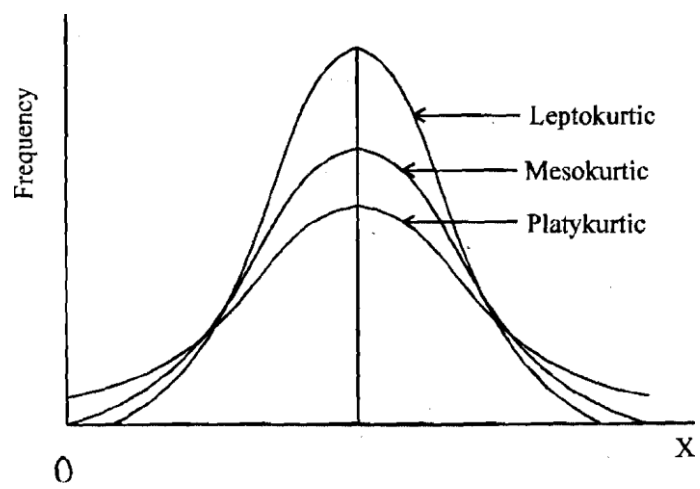


Fig.3



A measure of kurtosis is given by  $\beta_2$ , a coefficient given by Karl Pearson.

The value of  $\beta_2 = 3$  for a mesokurtic curve. When  $\beta_2 > 3$ , the curve is more peaked than the mesokurtic curve and is termed as leptokurtic. Similarly, when  $\beta_2 < 3$ , the curve is less peaked than the mesokurtic curve and is called as platykurtic curve.

**Example 4:** The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Examine the skewness and kurtosis of the

distribution. To examine skewness, we compute  $\beta_1$ .

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.7)^2}{(2.5)^3} = 0.031$$

Since  $\beta_1 > 0$  and  $\beta_1$  is small, the distribution is moderately positively skewed.

Kurtosis is given by the coefficient  $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18.75}{(2.5)^2} = 3.0$ .

Hence the curve is mesokurtic.

### Check Your Progress 3

- 1) Compute the first four central moments from the following data. Also find the two beta coefficients.

Vvalue	:	5	10	15	20	25	30	35
Frequency	:	8	15	20	32	23	17	5

- 2) The first four moments of a distribution are 1, 4, 10 and 46 respectively. Compute the moment coefficients of skewness and kurtosis and comment upon the nature of the distribution.

---

## LET US SUM UP

---

In this Unit you have learned about the measures of skewness and kurtosis. These two concepts are used to get an idea about the shape of the frequency curve of a distribution. Skewness is a measure of the lack of symmetry whereas kurtosis is a measure of the relative peakedness of the top of a frequency curve.

---

**Unit – 1 Binomial distribution-properties**  
**Poisson Distributions - properties, Normal Distributions- properties**

**Theoretical Distributions**

**Theoretical distributions are**

- |                          |                       |                         |
|--------------------------|-----------------------|-------------------------|
| 1. Binomial distribution | Discrete distribution | }                       |
| 2. Poisson distribution  |                       |                         |
| 3. Normal distribution   | →                     | Continuous distribution |

**Discrete Probability distribution**

**Bernoulli distribution**

A random variable  $x$  takes two values 0 and 1, with probabilities  $q$  and  $p$  ie.,  $p(x=1) = p$  and  $p(x=0)=q$ ,  $q=1-p$  is called a Bernoulli variate and is said to be Bernoulli distribution where  $p$  and  $q$  are probability of success and failure. It was given by Swiss mathematician James Bernoulli (1654-1705)

**Example**

- Tossing a coin(head or tail)
- Germination of seed(germinate or not)

**Binomial distribution**

Binomial distribution was discovered by James Bernoulli (1654-1705). Let a random experiment be performed repeatedly and the occurrence of an event in a trial be called as success and its non-occurrence is failure. Consider a set of  $n$  independent trials ( $n$  being finite), in which the probability  $p$  of success in any trial is constant for each trial. Then  $q=1-p$  is the probability of failure in any trial.

The probability of x success and consequently n-x failures in n independent trials. But x successes in n trials can occur in  $nc_x$  ways. Probability for each of these ways is  $p^x q^{n-x}$ .

$$\begin{aligned} P(sss\dots ff\dots fsf\dots f) &= p(s)p(s)\dots p(f)p(f)\dots \\ &= p,p\dots q,q\dots \\ &= (p,p\dots p)(q,q\dots q) \\ &\quad (x \text{ times}) (n-x \text{ times}) \end{aligned}$$

Hence the probability of x success in n trials is given by

$$nc_x p^x q^{n-x}$$

### Definition

A random variable x is said to follow binomial distribution if it assumes non-negative values and its probability mass function is given by

$$P(X=x) = p(x) =$$

$$\begin{aligned} nc_x p^x q^{n-x}, & \quad x=0,1,2,\dots,n \\ & \quad q=1-p \\ 0, & \quad \text{otherwise} \end{aligned}$$

The two independent constants n and p in the distribution are known as the parameters of the distribution.

### Condition for Binomial distribution

We get the binomial distribution under the following experimentation conditions

1. The number of trial n is finite
2. The trials are independent of each other.

3. The probability of success  $p$  is constant for each trial.
4. Each trial must result in a success or failure.
5. The events are discrete events.

### Properties

1. If  $p$  and  $q$  are equal, the given binomial distribution will be symmetrical. If  $p$  and  $q$  are not equal, the distribution will be skewed distribution.
2. Mean =  $E(x) = np$
3. Variance =  $V(x) = npq$  (mean > variance)

### Application

1. Quality control measures and sampling process in industries to classify items as defectives or non-defective.
2. Medical applications such as success or failure, cure or no-cure.

### Example 1

Eight coins are tossed simultaneously. Find the probability of getting atleast six heads.

### Solution

Here number of trials,  $n = 8$ ,  $p$  denotes the probability of getting a head.

$$\therefore p = \frac{1}{2} \text{ and } q = \frac{1}{2}$$

If the random variable  $X$  denotes the number of heads, then the probability of a success in  $n$  trials is given by

$$P(X = x) = {}^n C_x p^x q^{n-x}, x = 0, 1, 2, \dots, n$$

$$= {}^8 C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{8-x} = {}^8 C_x \left(\frac{1}{2}\right)^8$$

$$= \frac{1}{2^8} {}^8 C_x$$

Probability of getting atleast six heads is given by

$$P(x \geq 6) = P(x = 6) + P(x = 7) + P(x = 8)$$

$$= \frac{1}{2^8} 8C_6 + \frac{1}{2^8} 8C_7 + \frac{1}{2^8} 8C_8$$

$$= \frac{1}{2^8} [8C_6 + 8C_7 + 8C_8]$$

$$= \frac{1}{2^8} [28 + 8 + 1] = \frac{37}{256}$$

**Example 2** Ten coins are tossed simultaneously. Find the probability of getting (i) atleast seven heads (ii) exactly seven heads (iii) atleast seven heads

**Solution**

$$p = \text{Probability of getting a head} = \frac{1}{2}$$

$$q = \text{Probability of not getting a head} = \frac{1}{2}$$

The probability of getting x heads throwing 10 coins simultaneously is given by

$$P(X = x) = {}^nC_x p^x q^{n-x}, x = 0, 1, 2, \dots, n$$

$$= {}^{10}C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{10-x} = {}^{10}C_x \left(\frac{1}{2}\right)^{10}$$

$$= \frac{1}{2^{10}} {}^{10}C_x$$

i) Probability of getting atleast seven heads

$$P(x \geq 7) = P(x = 7) + P(x = 8) + P(x = 9) + P(x = 10)$$

$$= \frac{1}{2^{10}} [{}^{10}C_7 + {}^{10}C_8 + {}^{10}C_9 + {}^{10}C_{10}]$$

$$= \frac{1}{1024} [120 + 45 + 10 + 1] = \frac{176}{1024}$$

ii) Probability of getting exactly 7 heads

$$P(x = 7) = \frac{1}{2^{10}} {}^{10}C_7 = \frac{1}{2^{10}} (120) = \frac{120}{1024}$$

iii) Probability of getting atleast 7 heads

$$P(x \leq 7) = 1 - P(x > 7)$$

$$\begin{aligned}
&= 1 - \text{symbol } \{P(x=8) + P(x=9) + P(x=10)\} \\
&= 1 - \frac{1}{2^{10}} [10C_8 + 10C_9 + 10C_{10}] \\
&= 1 - \frac{1}{2^{10}} [45 + 10 + 1] \\
&= 1 - \frac{56}{1024} \\
&= \frac{968}{1024}
\end{aligned}$$

**Example 3:** 20 wrist watches in a box of 100 are defective. If 10 watches are selected at random, find the probability that (i) 10 are defective (ii) 10 are good (iii) at least one watch is defective (iv) at most 3 are defective.

**Solution**

20 out of 100 wrist watches are defective

Probability of defective wrist watch,  $p = \frac{20}{100} = \frac{1}{5}$

$$q = 1 - p = \frac{4}{5}$$

Since 10 watches are selected at random,  $n = 10$

$P(X = x) = nC_x p^x q^{n-x}$ ,  $x = 0, 1, 2, \dots, 10$

$$= 10C_x \left(\frac{1}{5}\right)^x \left(\frac{4}{5}\right)^{10-x}$$

i) Probability of selecting 10 defective watches

$$P(x=10) = 10C_{10} \left(\frac{1}{5}\right)^{10} \left(\frac{4}{5}\right)^0 = 1 \cdot \frac{1}{5^{10}} \cdot 1 = \frac{1}{5^{10}}$$

ii) Probability of selecting 10 good watches (i.e. no defective)

$$\begin{aligned}
P(x=0) &= 10C_0 \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^{10} \\
&= 1 \cdot 1 \left(\frac{4}{5}\right)^{10} = \left(\frac{4}{5}\right)^{10}
\end{aligned}$$

iii) Probability of selecting at least one defective watch

$$P(x \geq 1) = 1 - P(x < 1)$$

$$= 1 - P(x = 0)$$

$$= 1 - {}^{10}C_0 \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^{10}$$

$$= 1 - \left(\frac{4}{5}\right)^{10}$$

iv) Probability of selecting at most 3 defective watches

$$P(x \leq 3) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)$$

$$= {}^{10}C_0 \left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^{10} + {}^{10}C_1 \left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^9 + {}^{10}C_2 \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^8 + {}^{10}C_3 \left(\frac{1}{5}\right)^3 \left(\frac{4}{5}\right)^7$$

$$= 1.1 \left(\frac{4}{5}\right)^{10} + 10 \left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^9 + \frac{10 \cdot 9}{1 \cdot 2} \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^8 + \frac{10 \cdot 9 \cdot 8}{1 \cdot 2 \cdot 3} \left(\frac{1}{5}\right)^3 \left(\frac{4}{5}\right)^7$$

$$= 1. (0.107) + 10 (0.026) + 45 (0.0062) + 120 (0.0016)$$

$$= 0.859 \text{ (approx)}$$

### Poisson distribution

The Poisson distribution, named after Simeon Denis Poisson (1781-1840). Poisson distribution is a discrete distribution. It describes random events that occurs rarely over a unit of time or space.

It differs from the binomial distribution in the sense that we count the number of success and number of failures, while in Poisson distribution, the average number of success in given unit of time or space.

### Definition

The probability that exactly x events will occur in a given time is as follows

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!}, x=0,1,2,\dots$$

called as probability mass function of Poisson distribution.



where  $\lambda$  is the average number of occurrences per unit of time

$$\lambda = np$$

### **Condition for Poisson distribution**

Poisson distribution is the limiting case of binomial distribution under the following assumptions.

1. The number of trials  $n$  should be indefinitely large ie.,  $n \rightarrow \infty$
2. The probability of success  $p$  for each trial is indefinitely small.
3.  $np = \lambda$ , should be finite where  $\lambda$  is constant.

### **Properties**

1. Poisson distribution is defined by single parameter  $\lambda$ .
2. Mean =  $\lambda$
3. Variance =  $\lambda$ . Mean and Variance are equal.

### **Application**

1. It is used in quality control statistics to count the number of defects of an item.
2. In biology, to count the number of bacteria.
3. In determining the number of deaths in a district in a given period, by rare disease.
4. The number of error per page in typed material.
5. The number of plants infected with a particular disease in a plot of field.
6. Number of weeds in particular species in different plots of a field.

**Example 4:** Suppose on an average 1 house in 1000 in a certain district has a fire during a year. If there are 2000 houses in that district, what is the probability that exactly 5 houses will have a fire during the year? [given that  $e^{-2} = 0.13534$ ]

**Solution:**

Mean,  $\bar{x} = np$ ,  $n = 2000$  and  $p = \frac{1}{1000}$

$$= 2000 \times \frac{1}{1000}$$

$$\lambda = 2$$

The Poisson distribution is

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

$$\begin{aligned} P(X = 5) &= \frac{e^{-2} 2^5}{5!} \\ &= \frac{(0.13534) \times 32}{120} \\ &= \mathbf{0.036} \end{aligned}$$

### Example 5

If 2% of electric bulbs manufactured by a certain company are defective. Find the probability that in a sample of 200 bulbs i) less than 2 bulbs ii) more than 3 bulbs are defective. [ $e^{-4} = 0.0183$ ]

**Solution**

The probability of a defective bulb =  $p = \frac{2}{100} = 0.02$

Given that  $n = 200$  since  $p$  is small and  $n$  is large

We use the Poisson distribution

mean,  $m = np = 200 \times 0.02 = 4$

Now, Poisson Probability function,  $P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$

i) Probability of less than 2 bulbs are defective

$$\begin{aligned}
&= P(X < 2) \\
&= P(x = 0) + P(x = 1) \\
&= e^{-4} + e^{-4} (4) \\
&= e^{-4} (1 + 4) = 0.0183 \times 5 \\
&= 0.0915
\end{aligned}$$

ii) Probability of getting more than 3 defective bulbs

$$\begin{aligned}
P(x > 3) &= 1 - P(x \leq 3) \\
&= 1 - \{P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)\} \\
&= 1 - e^{-4} \left\{ 1 + 4 + \frac{4^2}{2!} + \frac{4^3}{3!} \right\} \\
&= 1 - \{0.0183 \times (1 + 4 + 8 + 10.67)\} \\
&= 0.567
\end{aligned}$$

### Normal distribution

Continuous Probability distribution is normal distribution. It is also known as error law or Normal law or Laplacian law or Gaussian distribution. Many of the sampling distribution like student-t, f distribution and  $\chi^2$  distribution.

### Definition

A continuous random variable  $x$  is said to be a normal distribution with parameters  $\mu$  and  $\sigma^2$ , if the density function is given by the probability law

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}; -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$$

### Note

The mean  $m$  and standard deviation  $s$  are called the parameters of Normal distribution.

The normal distribution is expressed by  $X \sim N(m, s^2)$

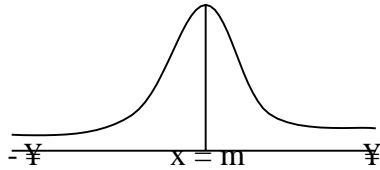
### Condition of Normal Distribution

i) Normal distribution is a limiting form of the binomial distribution under the following conditions.

- a)  $n$ , the number of trials is indefinitely large i.e.,  $n \rightarrow \infty$  and
  - b) Neither  $p$  nor  $q$  is very small.
- ii) Normal distribution can also be obtained as a limiting form of Poisson distribution with parameter  $m \rightarrow \infty$
- iii) Constants of normal distribution are mean =  $m$ , variation =  $s^2$ , Standard deviation =  $s$ .

### Normal probability curve

The curve representing the normal distribution is called the normal probability curve. The curve is symmetrical about the mean ( $m$ ), bell-shaped and the two tails on the right and left sides of the mean extends to the infinity. The shape of the curve is shown in the following figure.



### Properties of normal distribution

1. The normal curve is bell shaped and is symmetric at  $x = m$ .
2. Mean, median, and mode of the distribution are coincide  
i.e., Mean = Median = Mode =  $m$
3. It has only one mode at  $x = m$  (i.e., unimodal)
4. The points of inflection are at  $x = m \pm s$
5. The maximum ordinate occurs at  $x = m$  and its value is  $= \frac{1}{\sigma\sqrt{2\pi}}$
6. Area Property  $P(m - s < x < m + s) = 0.6826$   
 $P(m - 2s < x < m + 2s) = 0.9544$   
 $P(m - 3s < x < m + 3s) = 0.9973$

### Standard Normal distribution

Let  $X$  be random variable which follows normal distribution with mean  $\mu$  and variance  $\sigma^2$ . The standard normal variate is defined as  $Z = \frac{X - \mu}{\sigma}$  which follows

standard normal distribution with mean 0 and standard deviation 1 i.e.,  $Z \sim N(0,1)$ . The

standard normal distribution is given by  $\phi(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$ ;  $-\infty < z < \infty$

The advantage of the above function is that it doesn't contain any parameter. This enables us to compute the area under the normal probability curve.

### Note

#### Property of $\phi(Z)$

1.  $\phi(-Z) = 1 - \phi(Z)$
2.  $P(a \leq Z \leq b) = \phi(b) - \phi(a)$

**Example 6:** In a normal distribution whose mean is 12 and standard deviation is 2. Find the probability for the interval from  $x = 9.6$  to  $x = 13.8$

### Solution

Given that  $Z \sim N(12, 4)$

$$\begin{aligned}
 P(9.6 \leq Z \leq 13.8) &= P\left(\frac{9.6 - 12}{2} \leq Z \leq \frac{13.8 - 12}{2}\right) \\
 &= P(-1.2 \leq Z \leq 0) + P(0 \leq Z \leq 0.9) \\
 &= P(0 \leq Z \leq 1.2) + P(0 \leq Z \leq 0.9) \quad [\text{by using symmetric property}] \\
 &= 0.3849 + 0.3159 \\
 &= 0.7008
 \end{aligned}$$

When it is converted to percentage (ie) 70% of the observations are covered between 9.6 to 13.8.

**Example 7:** For a normal distribution whose mean is 2 and standard deviation 3. Find the value of the variate such that the probability of the variate from the mean to the value is 0.4115

**Solution:**

Given that  $Z \sim N(2, 9)$

To find  $X_1$ :

We have  $P(2 \leq Z \leq X_1) = 0.4115$

$$P\left(\frac{2-2}{3} \leq \frac{X-\mu}{\sigma} \leq \frac{X_1-2}{3}\right) = 0.4115$$

$$P(0 \leq Z \leq Z_1) = 0.4115 \text{ where } Z_1 = \frac{X_1 - 2}{3}$$

[From the normal table where 0.4115 lies is the value of  $Z_1$ ]

From the normal table we have  $Z_1 = 1.35$

$$\therefore 1.35 = \frac{X_1 - 2}{3}$$

$$\Rightarrow 3(1.35) + 2 = X_1$$

$$= X_1 = 6.05$$

(i.e) 41 % of the observation converged between 2 and 6.05

### Questions

1. For a Poisson distribution

- |                     |                     |
|---------------------|---------------------|
| (a) mean > variance | (b) mean = variance |
| (c) mean < variance | (d) mean < variance |

**Ans: mean = variance**

2. In normal distribution, skewness is

- |                      |                   |
|----------------------|-------------------|
| (a) one              | (b) zero          |
| (c) greater than one | (d) less than one |

**Ans: zero**

3. Poisson distribution is a distribution for rare events

**Ans: True**

4. The total area under normal probability curve is one.

**Ans: True**

5. Poisson distribution is for continuous variable.

**Ans: False**

6. In a symmetrical curve mean, median and mode will coincide.

**Ans: True**

7. Give any two examples of Poisson distribution

8. The variance of a Poisson distribution is 0.5. Find  $P(x = 3)$ .

**$[e^{-0.5} = 0.6065]$**

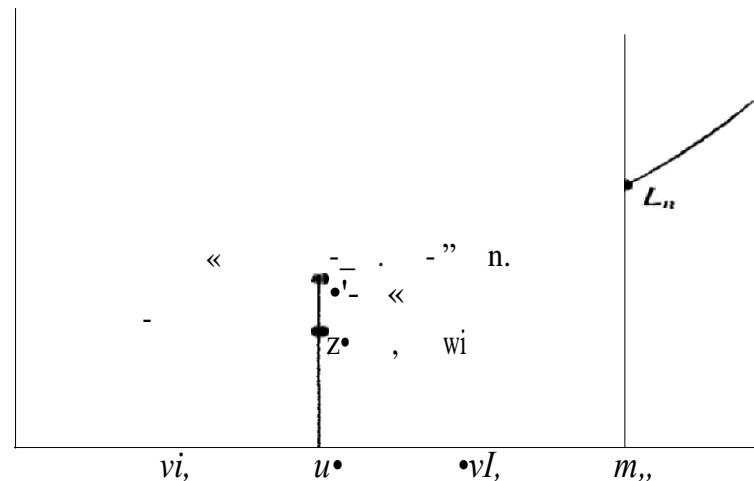
9. The customer accounts of a certain departmental store have an average balance of Rs.1200 and a standard deviation of Rs.400. Assuming that the account balances are normally distributed. (i) what percentage of the accounts is over Rs.1500? (ii) What percentage of the accounts is between Rs.1000 and Rs.1500? (iii) What percentage of the accounts is below Rs.1500?

10. State the Properties of normal distribution

# Curve fitting

## PRINCIPLE OF LEAST SQUARES

The graphical method has the drawback in that the straight line drawn may not be unique but principle of least squares provides a unique set of values to the constants and hence suggests a curve of best fit to the given data. The method of least square is probably the most systematic procedure to fit a unique curve through the given data points.



We will consider some of the best fitting curves of the type:

1. A straight line.
2. A second degree curve.
3. The exponential curve  $y = ae^{bx}$ .
4. The curve  $y = ax^{Tt}$ .

### 1. Fitting a straight line by the method of least squares:

Let  $(x_i, y_i)$ ,  $i = 0, 1, 2, \dots, n$  be the  $n$  sets of observations and let the related relation by  $y = ax + b$ . Now we have to select  $a$  and  $b$  so that the straight line is the best fit to the data.

As explained earlier, the residual at  $z = z_i$  is

$$d_i = y_i - (ax_i + b) \quad i = 1, 2, \dots, n$$

$$E = \sum_{i=1}^n d_i^2 = \sum_{i=1}^n [y_i - (ax_i + b)]^2$$

By the principle of least squares,  $E$  is minimum.



$$\frac{dE}{dn} = 0 \quad \text{and} \quad \frac{dE}{db} = 0$$

$$\text{i.e., } 2[y; -(n_0 + b)](-x_i) = 0 \quad \& \quad 2[y; -(a x_i + b)](-1) = 0$$

$$\text{i.e., } \sum_{i=1}^n (x_i y_i - a x_i^2 - b x_i) = 0 \quad \& \quad \sum_{i=1}^n (y_i - a x_i - b) = 0$$

$$\text{i.e., } \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i \quad \dots \text{ (eq.1)}$$

$$\text{And } \sum_{i=1}^n x_i + n b = \sum_{i=1}^n y_i \quad \dots \text{ (eq.2)}$$

Since,  $x_i, y_i$  are known, equations (1) & (2) give two equations in  $a$  &  $b$ . Solve for  $a$  &  $b$  from (1) & (2) & obtain the best fit  $y = ax + b$ .

Note:

- Equations (1) & (2) are called normal equations.
- Dropping suffix  $i$  from (1) & (2), the normal equations are

$$a \sum x_i^2 + b \sum x_i = \sum x_i y_i$$

Which are get taking  $\sum$  on both sides of  $y = ax + b$  & also taking  $\sum$  on both sides after multiplying by  $x$  both sides of  $y = ax + b$ .

- Transformation like  $I = \frac{1}{h} Y - \frac{1}{h}$  reduce the linear equation  $y = ax + b$  to the form  $Y = AX + B$ . Hence, a linear fit is another linear fit in both systems of coordinates.

Example 1:

By the method of least squares find the straight line to the data given below:

x	5	10	15	20	25
y	16	19	23	26	30

Solution:

Let the straight line be  $y = ax + b$

The normal equations are  $\sum x_i^2 + b \sum x_i = \sum x_i y_i$  .... (eq.1)

$$\sum x_i^2 + b \sum x_i = \sum x_i y_i \quad \dots \text{ (eq.2)}$$

To calculate  $\sum x_i^2, \sum x_i, \sum y_i, \sum x_i y_i$  we form below the table.

			$x^2$	$xy$
	5	16		80
	10	19	100	190
	15	23	225	345
	20	26	400	520
	25	30	625	750
Total	75	114	1375	1885

The normal equations are  $75a+5b=114$  ..... (eq.1)

$$1375a+75b=1885 \quad \text{.....(eq.2)}$$

Eliminate b, multiply (1) by 15

$$1125a+75b=1710 \quad \text{.. ... (eq.3)}$$

(eq.2) — (eq.3) gives,  $250a=175$  or  $a=0.7$ , hence  $b=19.2$

Hence, the best fitting line is  $y=0.7x+12.3$

$$\text{Let } X = \frac{x - x_{mid}}{h} = \frac{x - 15}{5}, \quad Y = \frac{y - y_{mid}}{h} = \frac{y - 23}{5}$$

Let the line in the new variable by  $Y=AX+B$

	$x$	$y$	$X$	$X^2$	$Y$	$XY$
	5	16	-2	4	-1.4	2.8
	10	19	-1	1	-0.8	0.8
	15	23	0	0	0	0
	20	26	1	1	0.6	0.6
	25	30	2	4	1.4	2.8
Total			0	10	-0.2	7

The normal equations are  $A \sum X + 5B = \sum Y$  .....(eq.4)

$$\text{fi } \sum X^2 = \sum XY \quad \text{.....(eq.5)}$$

Therefore,  $-5B = -0.2 \quad B = 0.04$

$$10A = 7 \rightarrow A = 0.7$$

The equations  $Y = 0.7X - 0.04$

$$\text{i.e.} \rightarrow 0.7 \left( \frac{x-15}{5} \right) - 0.04 \rightarrow y - 23 \quad 0.7x - 10.5 - 0.2$$

$$\text{i.e.} \quad y = 0.7x + 33.3$$

Which is the same equation as seen before.

### Example 2:

Fit a straight line to the data given below. Also estimate the value of y at x=2.5

x	0	1	2	3	4
y	1	1.8	3.3	4.5	6.3

Solution:

Let the best fit be  $y = ax + b$  . . . (eq.1)

The normal equations are  $n \quad x + 5b =$  .. ... (eq.2)

$$a \sum x^2 + b \sum x = \sum xy \quad \dots \text{ (eq.3)}$$

We prepare the table for easy use.

	<b>x</b>	<b>y</b>	<b>x<sup>2</sup></b>	<b>xy</b>
	0	1	0	0
	1	1.8	1	1.8
	2	3.3	4	6.6
	3	4.5	9	13.5
	4	6.3	16	25.2
<b>Total</b>	<b>10</b>	<b>16.9</b>	<b>30</b>	<b>47.1</b>

Substituting in (eq.2) and (eq.3), we get,

$$10a + 5b = 16.9$$

$$30a + 10b = 47.1$$

Solving eq.(2)-eq.(1), we get,  $a = 1.33$ ,  $b = 0.72$

Hence, the equation is  $y = 1.33x + 0.72$

$$y \text{ (at } x=2.5) = 1.33 (2.5) + 0.72 = 4.045$$

### Example 3:

By proper transformation, convert the relation  $y = a + bxy$  to a linear form & find the equation to fit the data.

x	-4	1	2	3
y	4	6	10	8

Solution:

Let  $X = xy$ ,

—

The equation becomes  $y = a + bX$ .

The normal equations are  $a \sum A + b \sum J^2$  by

$$4a + b \sum X = \sum y \quad \dots (eq.1)$$

$$\dots (eq.2)$$

And

	x	y	X	A <sup>2</sup>	Xy
	-4	4	-16	256	-64
	1	6	6	36	36
	2	10	20	400	200
	3	8	24	576	192
Total		28	34	1268	364

Substituting in (eq.2) and (eq.3), we get,

$$34a + 1268b = 364$$

$$4a + 34b = 28$$

Solving, we get,  $a = 5.90605$ ,  $b = 0.12870$

Therefore, the equation is  $y = 5.90605 + 0.12870X$

i.e.  $y = 5.90605 + 0.12870X$

Using this equation we get  $y(1 - 0.12870x) = 5.90605$

i.e.,  $y = \frac{5.90605}{1 - 0.12870x}$

We tabulate the value to verify:

x	-4	1	2	3
y	3.89890	6.77843	7.95320	9.61054

Note: if we take  $u = \frac{1}{xy}$ ,  $v = \frac{1}{x}$  we get

$v = ax + b$ . Taking this as linear, we get

$$a = 10.5, b = -0.13$$

That is  $y = 10.5 - 0.13xy$

i.e.  $y = \frac{10.5}{1 - 0.13x}$

---

x	-4	1	2	3
y	21.875	92.9203	8.33333	7.55396

The values of y are far away from the given values. Perhaps, the selection of the form is not correct.

## **Hypothesis Testing**

In this chapter we will learn ....

- ❑ To use an inferential method called a hypothesis test
- ❑ To analyze evidence that data provide
- ❑ To make decisions based on data

### **Major Methods for Making Statistical Inferences about a Population**

- ❑ **The traditional Method**
- ❑ **The p-value Method**
- ❑ **Confidence Interval**

## **Section 8-1: Steps in Hypothesis Testing – Traditional Method**

The main goal in many research studies is to check whether the data collected support certain statements or predictions.

**Statistical Hypothesis – a conjecture about a population parameter. This conjecture may or may not be true.**

**Example:** The mean income for a resident of Denver is equal to the mean income for a resident of Seattle.

- ❑ Population parameter is mean income
- ❑ One population consists of residents of Denver while the other consists of residents of Seattle.

There are two types of statistical hypotheses:

**Null Hypothesis ( $H_0$ )** – a statistical hypothesis that states that there is **no** difference between a parameter and a specific value, or that there is **no** difference between two parameters.

**Alternative Hypothesis ( $H_1$ )** – a statistical hypothesis that states the existence of a **difference** between a parameter and a specific value, or states that there is a **difference** between two parameters.

*Can you formulate a null and alternative hypothesis for the income example?*

---



We tend to want to reject the null hypothesis so we assume it is true and look for enough evidence to conclude it is incorrect.

We tend to want to accept the alternative hypothesis. If the null hypothesis is rejected then we must accept that the alternative hypothesis is true.

Note:  $H_0$  **will ALWAYS have an equal sign** (and possibly a less than or greater than symbol, depending on the alternative hypothesis). The alternative hypothesis has a range of values that are alternatives to the one in  $H_0$ .

The null and alternative hypotheses are stated together.  
are typical hypothesis for means, where  $k$  is a specified number.

The following

**Two-tailed test**

$$H_0: \mu = k$$

$$H_1: \mu \neq k$$

**Right-tailed test**

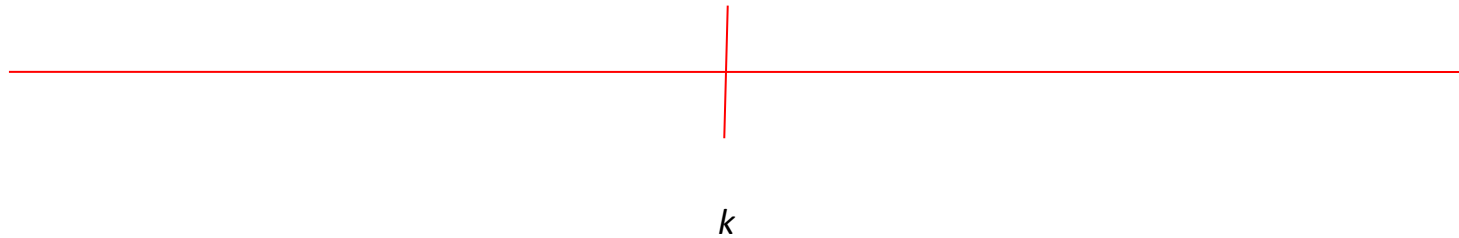
$$H_0: \mu = k$$

$$H_1: \mu > k$$

**Left-tailed test**

$$H_0: \mu = k$$

$$H_1: \mu < k$$



*Note:* Right-tailed and left-tailed tests are distinguished by the way the greater than or less than points. It is the direction where the alternative places the true mean.

**Examples:** State the  $H_0$  and  $H_1$  for each case.

A researcher thinks that if expectant mothers use vitamins, the birth weight of the babies will increase. The average birth weight of the population is 8.6 pounds.

$$H_0: \mu = 8.6$$

$$H_1: \mu > 8.6$$

An engineer hypothesizes that the mean number of defects can be decreased in a manufacturing process of compact disks by using robots instead of humans for certain tasks. The mean number of defective disks per 1000 is 18.

$$H_0 :$$

$$H_1 :$$

A psychologist feels that playing soft music during a test will change the results of the test. The psychologist is not sure whether the grades will be higher or lower. In the past, the mean of the scores was 73.

$H_0$  :

$H_1$  :

When a researcher conducts a study, he or she is generally looking for evidence to support a claim of some type of difference. In this case, the claim should be stated as the alternative hypothesis.

Because of this, the alternative hypothesis is sometimes called the **research hypothesis**.

Keywords help to indicate what the null and/or alternative hypotheses should be.

**Table 8–1** Hypothesis-Testing Common Phrases

$>$	$<$
Is greater than	Is less than
Is above	Is below
Is higher than	Is lower than
Is longer than	Is shorter than
Is bigger than	Is smaller than
Is increased	Is decreased or reduced from
$=$	$\neq$
Is equal to	Is not equal to
Is the same as	Is different from
Has not changed from	Has changed from
Is the same as	Is not the same as

After stating the hypotheses, the researcher designs the study.

- ☐ Select the correct statistical test
- ☐ Choose an appropriate level of significance
- ☐ Formulate a plan for conducting the study

**Statistical Test** – uses the data obtained from a sample to make a decision about whether the null hypothesis should be rejected.




**Test Value** (test statistic) – the numerical value obtained from a statistical test.

When we make a conclusion from a statistical test there are two types of errors that we could make. They are called: Type I and Type II Errors

**Type I error** – reject  $H_0$  when  $H_0$  is true.

**Type II error** – do not reject  $H_0$  when  $H_0$  is false.

Results of a statistical test:

	$H_0$ is True	$H_0$ is False
Reject $H_0$	<b>Type I Error</b> 	Correct Decision 
Do not Reject $H_0$	Correct Decision 	<b>Type II Error</b>



**Example:** Decision Errors in a Legal Trial .

What are  $H_0$  and  $H_1$ ?

	$H_0$ true	$H_0$ false
Reject $H_0$	<b>Error</b> Type I	Correct decision
Do not reject $H_0$	Correct decision	<b>Error</b> Type II

$H_0$ : Defendant is innocent.

$H_1$ : Defendant is not innocent, i.e., guilty



If you are the defendant, which is the worse error?

Why?

- ❑ The decision of the jury does not prove that the defendant did or did not commit the crime.
  - ❑ The decision is based on the evidence presented.
  - ❑ If the evidence is strong enough the defendant will be convicted in most cases, if it is weak the defendant will be acquitted.
  - ❑ So the decision to reject the null hypothesis does not prove anything.
  - ❑ The question is how large of a difference is enough to say we have enough evidence to reject the null hypothesis?
-

**Significance level** - is the maximum probability of committing a Type I error.  
symbolized by  $\alpha$ .

This probability is

$$P(\text{Type I error} | H_0 \text{ is true}) = \alpha$$

**Critical or Rejection Region** – the range of values for the test value that indicate a significant difference and that the null hypothesis should be rejected.

**Non-critical or Non-rejection Region** – the range of values for the test value that indicates that the difference was probably due to chance and that the null hypothesis should not be rejected.

---

**Critical Value (CV)** – separates the critical region from the non-critical region, i.e., when we should reject  $H_0$  from when we should not reject  $H_0$ .

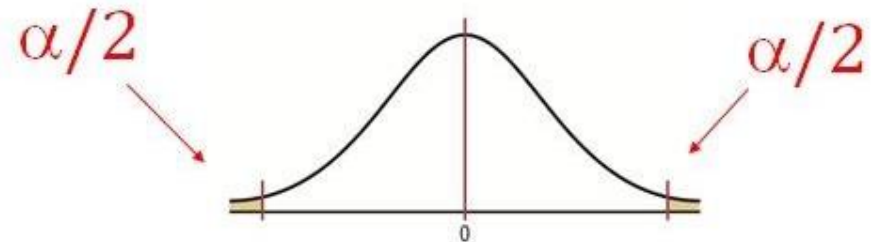
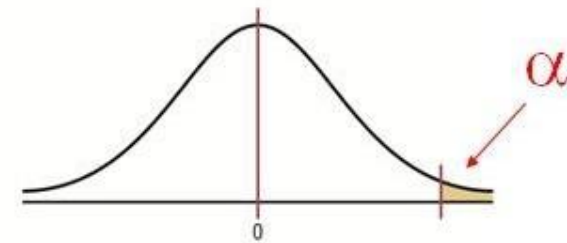
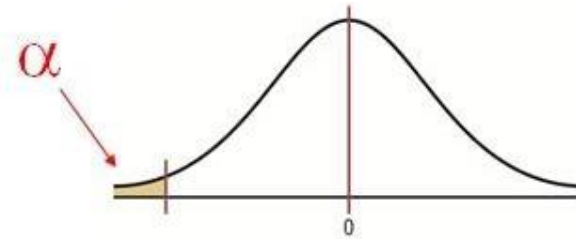
- ❑ The location of the critical value depends on the inequality sign of the alternative hypothesis.
- ❑ Depending on the distribution of the test value, you will use different tables to find the critical value.

**One-tailed test** – indicates that the null hypothesis should be rejected when the test value is in the critical region on one side.

❓ **Left-tailed test** – when the critical region is on the left side of the distribution of the test value.

❓ **Right-tailed test** – when the critical region is on the right side of the distribution of the test value.

**Two-tailed test** – the null hypothesis should be rejected when the test value is in either of two critical regions on either side of the distribution of the test value.



To obtain the critical value, the researcher must choose the significance level,  $\alpha$ , and know the distribution of the test value.

- The distribution of the test value indicates the shape of the distribution curve for the test value. This will have a shape that we know (like the standard normal or  $t$  distribution).
  - Let's assume that the test value has a standard normal distribution.
  - We should use Table E (the standard normal table) or Table F (using the bottom row of the  $t$  distribution, which is equivalent to a standard normal distribution) to find the critical value.
-

## Finding the Critical Values for Specific $\alpha$ Values, Using Table E

**Step 1:** Draw a figure for the distribution of the test values and indicate the appropriate area for the rejection region.

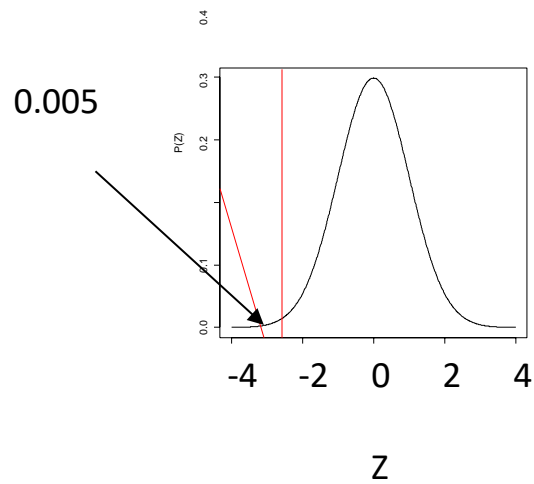
- ❑ If the test is left-tailed, the critical region, with area equal to  $\alpha$ , will be on the left side of the distribution curve.
  - ❑ If the test is right-tailed, the critical region, with area equal to  $\alpha$ , will be on the right side of the distribution curve.
  - ❑ If the test is two-tailed,  $\alpha$  must be divided by 2; the critical regions will be in each end of the distribution curve - half the area in the left part of the distribution and half of the area in the right part of the distribution.
-

## Step 2:

- ☐ For a left-tailed test, use the z value that corresponds to the area equivalent to in Table E, i.e.,  $z_{1-\alpha}$ , the percentile of the distribution.
  - ☐ For a right-tailed test, use the z value that corresponds to the area equivalent to  $1-\alpha$  in Table E, i.e.,  $z_{\alpha}$ , the 1  $\alpha$  percentile of the distribution.
  - ☐ For a two-tailed test, use the z value that corresponds to  $\alpha/2$  for the left lower CV. It will be negative. Change the sign to positive and you will get the critical value for the right side.
-

**Example:** Find the critical value(s) for each situation and draw the appropriate figure, showing the critical region.

Left-tailed test with  $\alpha = 0.005$

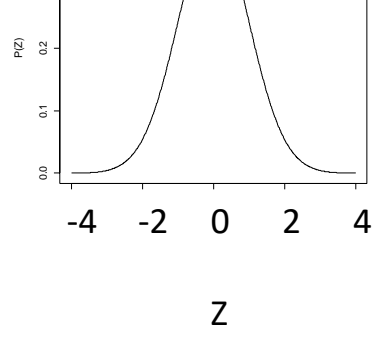


Looking up 0.005 in the Z table We have  $Z = -2.575$ .

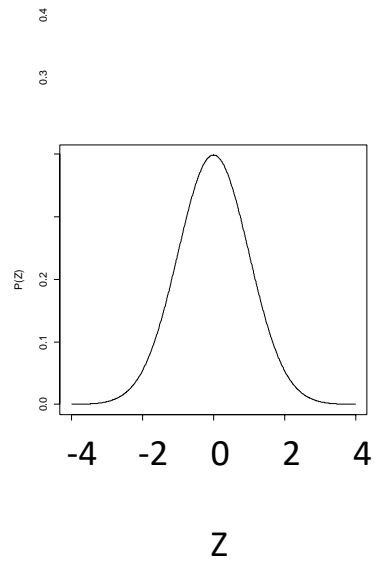
Right-tailed test with  $\alpha = 0.01$

0.4  
0.3





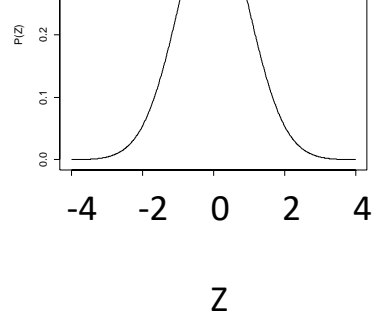
Two-tailed test with  $\alpha = 0.1$



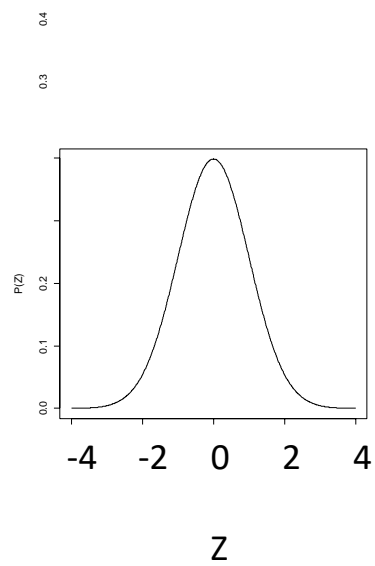
Left-tailed test with  $\alpha = 0.2$

0.4  
0.3

---



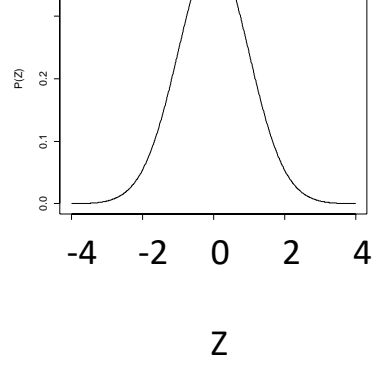
Right-tailed test with  $\alpha = 0.15$



Two-tailed test with  $\alpha = 0.09$

0.4  
0.3

---



### **Hypothesis Test Procedure (Traditional Method)**

Step 1 State the hypotheses and identify the claim.

Step 2 Find the critical value(s) from the appropriate table. Step 3 Compute the test value.

Step 4 Make the decision to reject or not reject the null hypothesis.

Step 5 Summarize the results.

## **Section 8-2: z Test for a Mean**

MOTIVATING SCENARIO: It has been reported that the average credit card debt for college seniors is \$3262.

The student senate at a large university feels that their seniors have a debt much less than this, so it conducts a study of 50 randomly selected seniors and finds that the average debt is \$2995, and the population standard deviation is \$1100.

**Can we support the student senate's claim using the data collected?**

## How....the z Test for a Mean

A **statistical test** uses the data obtained from a sample to make a decision about whether the null hypothesis should be rejected.

The numerical value obtained from a statistical test is called the **test value**.

You will notice that our statistical tests will resemble the general formula for a z-score:

$$\text{Test Value} = \frac{\text{observed value} - \text{expected value}}{\text{standard error}}$$



## The z test for Means

The z test is a statistical test for the mean of a population. It can be used when  $n \geq 30$ , or when the population is normally distributed and  $\sigma$  is known.

The formula for the z-test is:

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}, \text{ where}$$

$$\bar{X} =$$

$$\frac{\sum X}{n}$$

We use our standard normal distribution...our z table!

**Hypothesis Test Procedure (Traditional Method) Step 1** State the hypotheses and identify the claim.

**Step 2** Find the critical value(s) from the appropriate table.

**Step 3** Compute the test value.

**Step 4** Make the decision to reject or not reject the null hypothesis.

**Step 5** Summarize the results.

**Example:** It has been reported that the average credit card debt for college seniors is \$3262. The student senate at a large university feels that their seniors have a debt much less than this, so it conducts a study of 50 randomly selected seniors and finds that the average debt is \$2995, and the population standard deviation is \$1100. Let's conduct the test based on a Type I error of  $\alpha=0.05$ .

**Step 1** State the hypotheses and identify the claim.  $H_0: \mu = \$3262$   $H_1: \mu < \$3262$

CLAIM



**Step 2** Find the critical value(s) from the appropriate table.

Left-tailed test,  $\alpha=0.05$   $\Rightarrow$  Z will be negative and have probability 0.05 underneath it

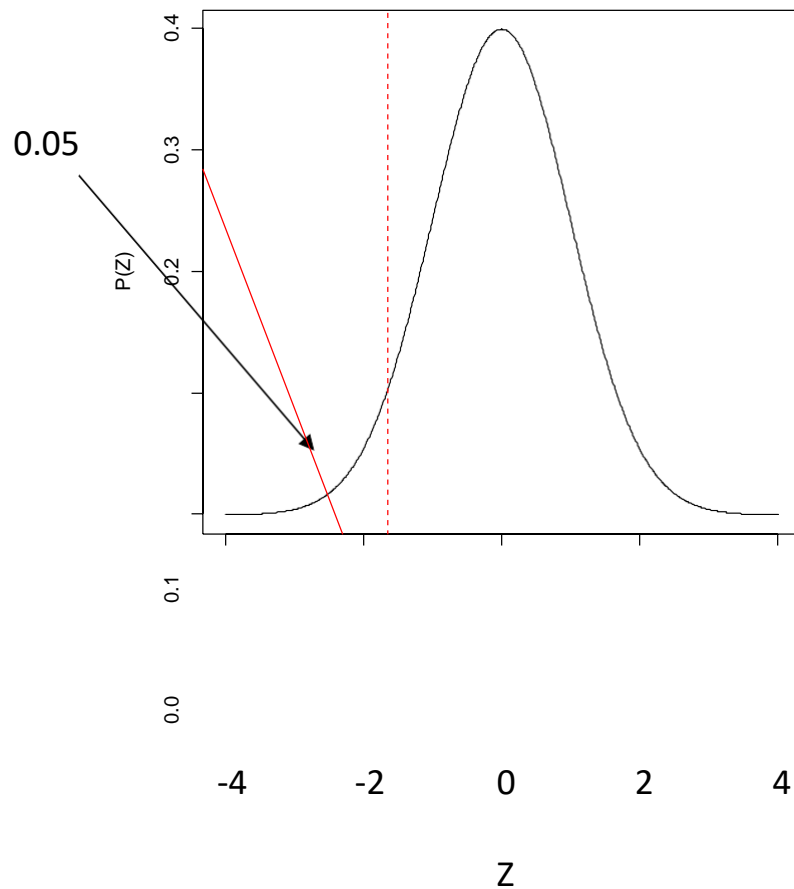


Table E The Standard Normal Distribution						
Cumulative Standard Normal Distribution						
z	.00	.01	.02	.03	.04	.05
-3.4	.0003	.0003	.0003	.0003	.0003	.0003
-3.3	.0005	.0005	.0005	.0004	.0004	.0004
-3.2	.0007	.0007	.0006	.0006	.0006	.0006
-3.1	.0010	.0009	.0009	.0009	.0008	.0008
-3.0	.0013	.0013	.0013	.0012	.0012	.0011
-2.9	.0019	.0018	.0018	.0017	.0016	.0016
-2.8	.0026	.0025	.0024	.0023	.0023	.0022
-2.7	.0035	.0034	.0033	.0032	.0031	.0030
-2.6	.0047	.0045	.0044	.0043	.0041	.0040
-2.5	.0062	.0060	.0059	.0057	.0055	.0054
-2.4	.0082	.0080	.0078	.0075	.0073	.0071
-2.3	.0107	.0104	.0102	.0099	.0096	.0094
-2.2	.0139	.0136	.0132	.0129	.0125	.0122
-2.1	.0179	.0174	.0170	.0166	.0162	.0158
-2.0	.0228	.0222	.0217	.0212	.0207	.0202
-1.9	.0287	.0281	.0274	.0268	.0262	.0256
-1.8	.0359	.0351	.0344	.0336	.0329	.0322
-1.7	.0446	.0436	.0427	.0418	.0409	.0401
-1.6	.0548	.0537	.0526	.0516	.0505	.0495

$$Z = -1.645 \text{ or } Z = -1.65$$

**Step 3** Compute the test value.

$$z = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{2995 - 3262}{\frac{1100}{\sqrt{50}}} = -1.716341$$

**Step 4** Make the decision to reject or not reject the null hypothesis.

Since this is a left-tailed test, our rejection region consists of values of Z that are smaller than our critical value of Z = -1.645.

Since our test value (-1.716341) is less than our critical value (-1.645), we reject the null hypothesis.

**Step 5** Summarize the results.

We have evidence to support the student senate claim that the university's seniors have credit card debt that is less than the reported average debt.

This is based on a Type I error rate of 0.05. This means we falsely make the claim above 5% of the time.

**Example:** The medical Rehabilitation Education Foundation reports that the average cost of rehabilitation for stroke victims is \$24,672.

To see if the average cost of rehab is different at a particular hospital, a researcher selects a random sample of 35 stroke victims at the hospital and finds the average cost of their rehab is

\$25,250. The standard deviation of the population is \$3251.

At  $\alpha = 0.01$ , can it be concluded that the average cost of stroke rehabilitation at a particular hospital is different from \$24,672?



**Step 1** State the hypotheses and identify the claim.

**Step 2** Find the critical value(s) from the appropriate table.

**Step 3** Compute the test value.

**Step 4** Make the decision to reject or not reject the null hypothesis.

**Step 5** Summarize the results.

**IMPORTANT NOTE:**

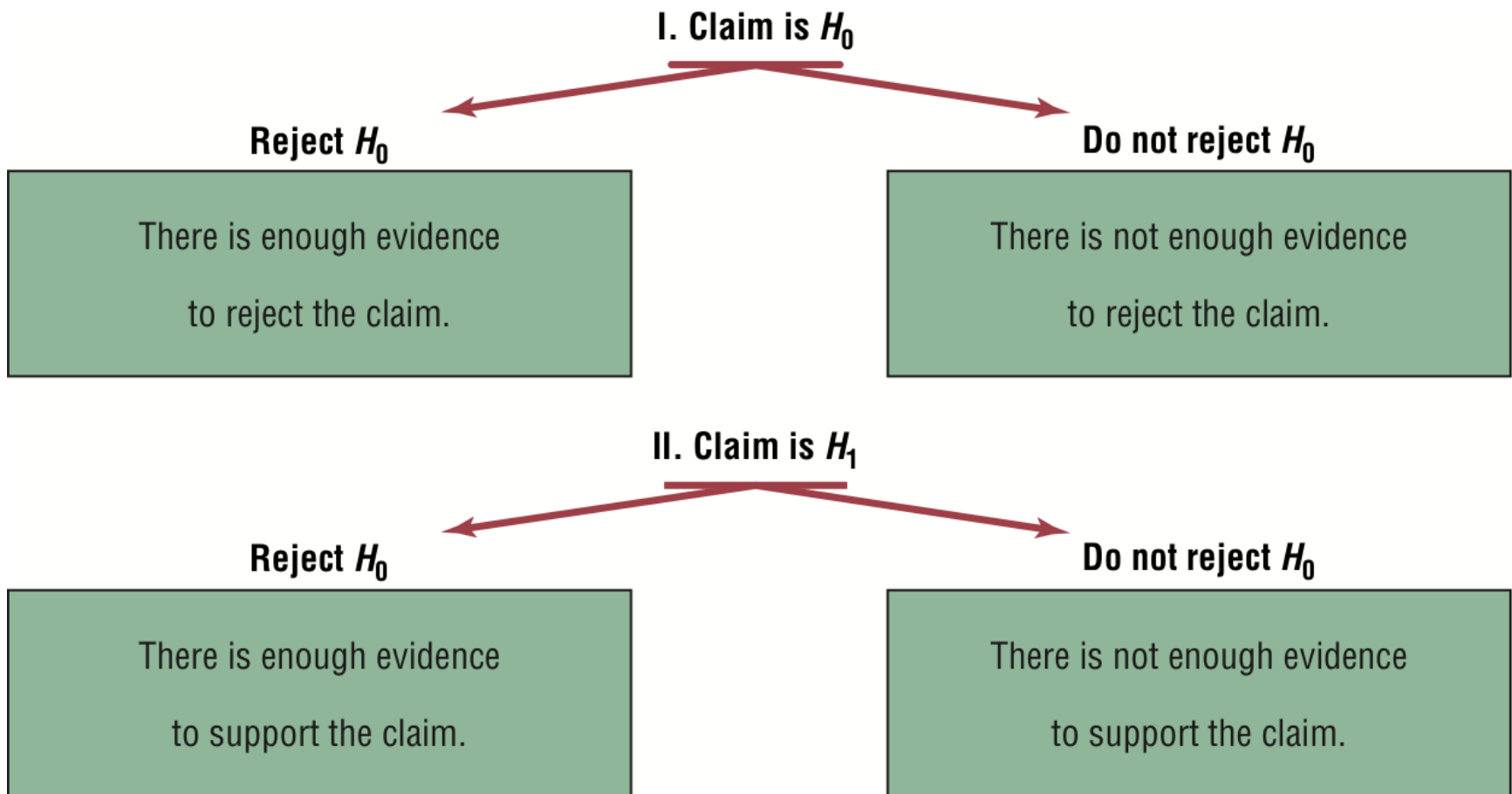
**When the null hypothesis is not rejected, we do not accept it as true. There is merely not enough evidence to say that it is false.**

Consider the jury trial analogy.

We don't find people innocent, only guilty or not guilty. If someone is found not guilty, it does not mean that they were proved innocent; it only means that there is not enough evidence to reach a guilty verdict.

**We conclude the alternative hypothesis (when we reject the null) because the data clearly support that conclusion.**

See the following slide for ways of describing your results.



See page 415 to help you remember what you can say in summarizing the results based on where the claim is ( $H_0$  or  $H_1$ ) and whether you rejected  $H_0$  or not.

## P-Value Method for Hypothesis Testing

We often test hypotheses at common levels of significance ( $\alpha = 0.05$ , or  $0.01$ ). Recall that the choice of alpha depends on the seriousness of the Type I error. There is another approach that utilizes a P-value.

The P-Value (or probability value) is the probability of getting a sample statistic (such as the mean) or a more extreme sample statistic in the direction of the alternative hypothesis when the null hypothesis is true.

The P-value is the actual area under the standard normal distribution curve of the test value or a more extreme value (further in the tail).

## A General Rule for Finding P-values using the Z

**distribution or the  $t$  distribution:**

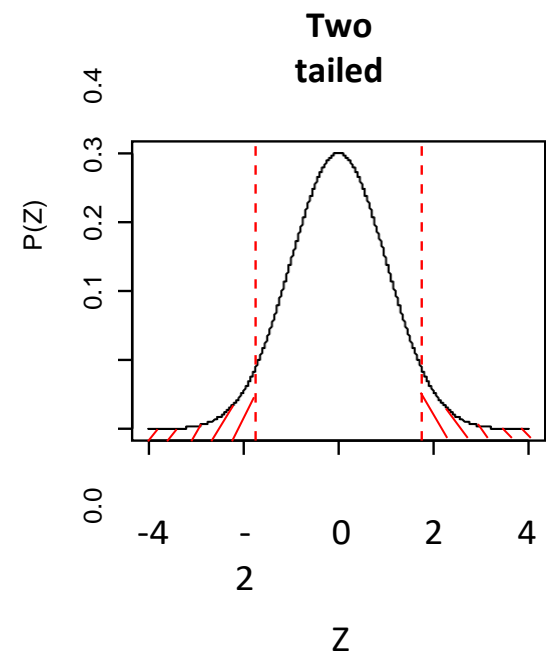
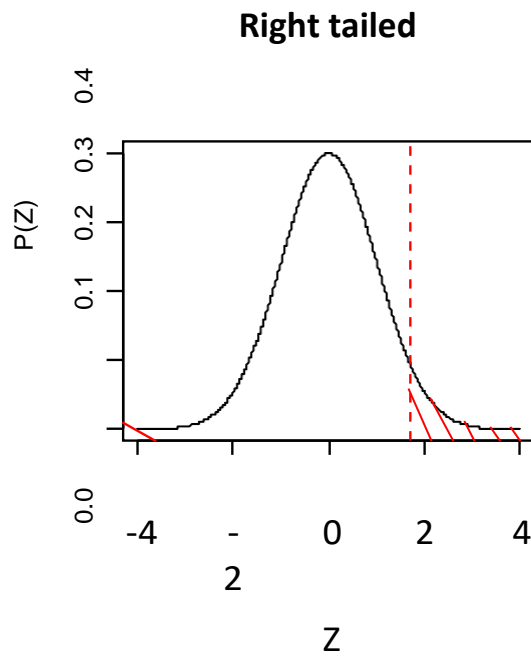
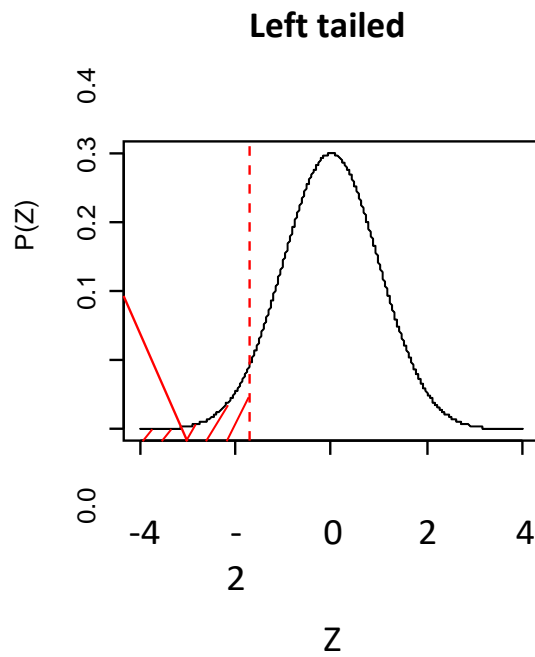
Suppose that  $z^*$  is the test statistic of a  $z$  test and  $t^*$  is the test statistic of a  $t$  test

**Left-tailed test:** p-value  $= P(Z \leq z^*)$  or p-value  $= P(T \leq t^*)$   
(depending on whether we are doing a  $z$  test or a  $t$  test).

**Right-tailed test:** p-value  $= P(Z \geq z^*)$  or p-value  $= P(T \geq t^*)$   
(depending on whether we are doing a  $z$  test or a  $t$  test).

**Two-tailed test:** p-value  $= 2P(Z \geq z^*)$  or p-value  $= 2P(T \geq t^*)$   
(depending on whether we are doing a  $z$  test or a  $t$  test).

The smaller the P-value, the stronger the evidence is against  $H_0$ . We use Table E to find P-Values that use a z Test.



***Examples:***

Suppose you have a left-tailed test and find the area in the tail to be 0.0489. What is the P-value? Would you reject this at  $\alpha =$   
0.05?       $\alpha = 0.01$ ?

Suppose you have a two-tailed test and find the area in one tail to be 0.0084. What is the P-value? Would you reject this at  $\alpha = 0.05$ ?       $\alpha = 0.01$ ?



## P-Value

If p-value is  $\leq \alpha$  then we \_\_\_\_\_.

If p-value is  $> \alpha$  then we \_\_\_\_\_.

Let's reconsider the hypothesis tests we did earlier and find the corresponding P-values.

**Example:** It has been reported that the average credit card debt for college seniors is \$3262. The student senate at a large university feels that their seniors have a debt much less than this, so it conducts a study of 50 randomly selected seniors and finds that the average debt is \$2995, and the population standard deviation is \$1100. Let's conduct the test based on a Type I error of  $\alpha=0.05$ .

**Step 3** Compute the test value and find the P-value.  $Z = -1.72$

**Step 4 and Step 5**

**Example:** The medical Rehabilitation Education Foundation reports that the average cost of rehabilitation for stroke victims is \$24,672. To see if the average cost of rehab is different at a particular hospital, a researcher selects a random sample of 35 stroke victims at the hospital and find the average cost of their rehab is \$25,250. The standard deviation of the population is \$3,251. At  $\alpha = 0.01$ , can it be concluded that the average cost of stroke rehabilitation at a particular hospital is different from \$24,672?

**Step 3**      Compute the test value and find the P-value.  $Z = 1.05$

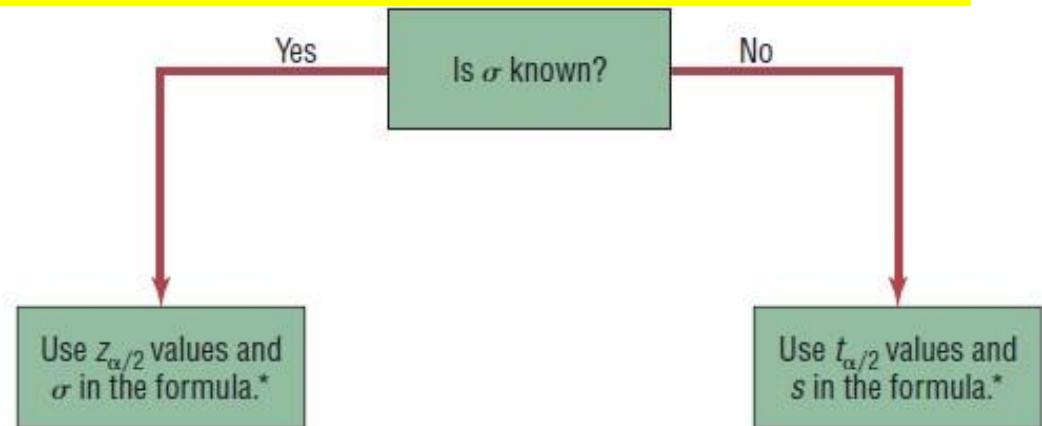
**Step 4 and Step 5**

### ***t* Test for a Mean**

When a population is normally or approximately normally distributed, but the population standard deviation is unknown, the *z* test is inappropriate for testing hypotheses involving means. Instead we will use the *t* test when  $\sigma$  is unknown and the distribution of the variable is approximately normal.

**Figure 8–25**

Using the *z* or *t* Test

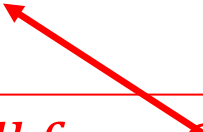


\*If  $n < 30$ , the variable must be normally distributed.

The **one-sample  $t$  test** is a statistical test for the mean of a population and is used when the population is normally or approximately **normally distributed** and  **$\sigma$  is unknown**.

The formula for the test value of the one-sample  $t$  test is:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \quad d.f. = n-1$$



*Recall from Chapter 7 that the  $t$  distribution is a family of curves and requires us to know the degrees of freedom in order to find the appropriate  $t$  value.*

## Finding Critical Values Using Table F

- Determine if the test is one-tailed or two tailed.
- Determine the degrees of freedom
- Find the significance level by looking in the appropriate row for the number of tails and down the column with the degrees of freedom.
- If the degrees of freedom is not in the table, always round DOWN to the nearest table value (this is a conservative approach).
- If the test is a left-tailed test, then the critical value is the NEGATIVE of the value given in the table.

**Example:** Find the critical  $t$  value for  $\alpha = 0.01$  with sample size of 13 for a left-tailed test.

- ◆ Left tailed means the critical  $t$  value will be negative
- ◆  $n=13$  means the degrees of freedom are  $n-1 = 12$
- ◆ The critical value is -2.681

Table F The $t$ Distribution						
	Confidence intervals	80%	90%	95%	98%	99%
	One tail, $\alpha$	0.10	0.05	0.025	0.01	0.005
d.f.	Two tails, $\alpha$	0.20	0.10	0.05	0.02	0.01
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.132	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032
6		1.440	1.943	2.447	3.143	3.707
7		1.415	1.895	2.365	2.998	3.499
8		1.397	1.860	2.306	2.896	3.355
9		1.383	1.833	2.262	2.821	3.250
10		1.372	1.812	2.228	2.764	3.169
11		1.363	1.796	2.201	2.718	3.106
12		1.356	1.782	2.179	2.681	3.055
13		1.350	1.771	2.160	2.650	3.012
14		1.345	1.761	2.145	2.624	2.977
15		1.341	1.753	2.131	2.602	2.947

**Example:** Find the critical  $t$  value for  $\alpha = 0.05$  with sample size of 21 for a right-tailed test.

**Example:** Find the critical  $t$  value for  $\alpha = 0.10$  with sample size of 59 for a two-tailed test.



## Finding P-Values Using Table F

To find a p-value:

- Draw a picture of the area we're trying to find.
- Find row corresponding to  $df = n - 1$ .
- Find the 2 values in the row that the value of the test statistic falls between.
- Find the  $\alpha$ 's that corresponds to these values
- The  $\alpha$ 's are the values that the p-value falls between

You can calculate the exact p-value using your calculator:

- 2<sup>nd</sup> -> VARS(DISTR)
- tcdf(lowerbound, upperbound, df)  
lowerbound corresponds to the value on the left that you are computing probability in between  
upperbound corresponds to the value on the right that you are computing probability in between  
use EE99 for  $\infty$   
use -EE99 for  $-\infty$
- You will need to multiply by 2 if you are performing a two- tailed test.

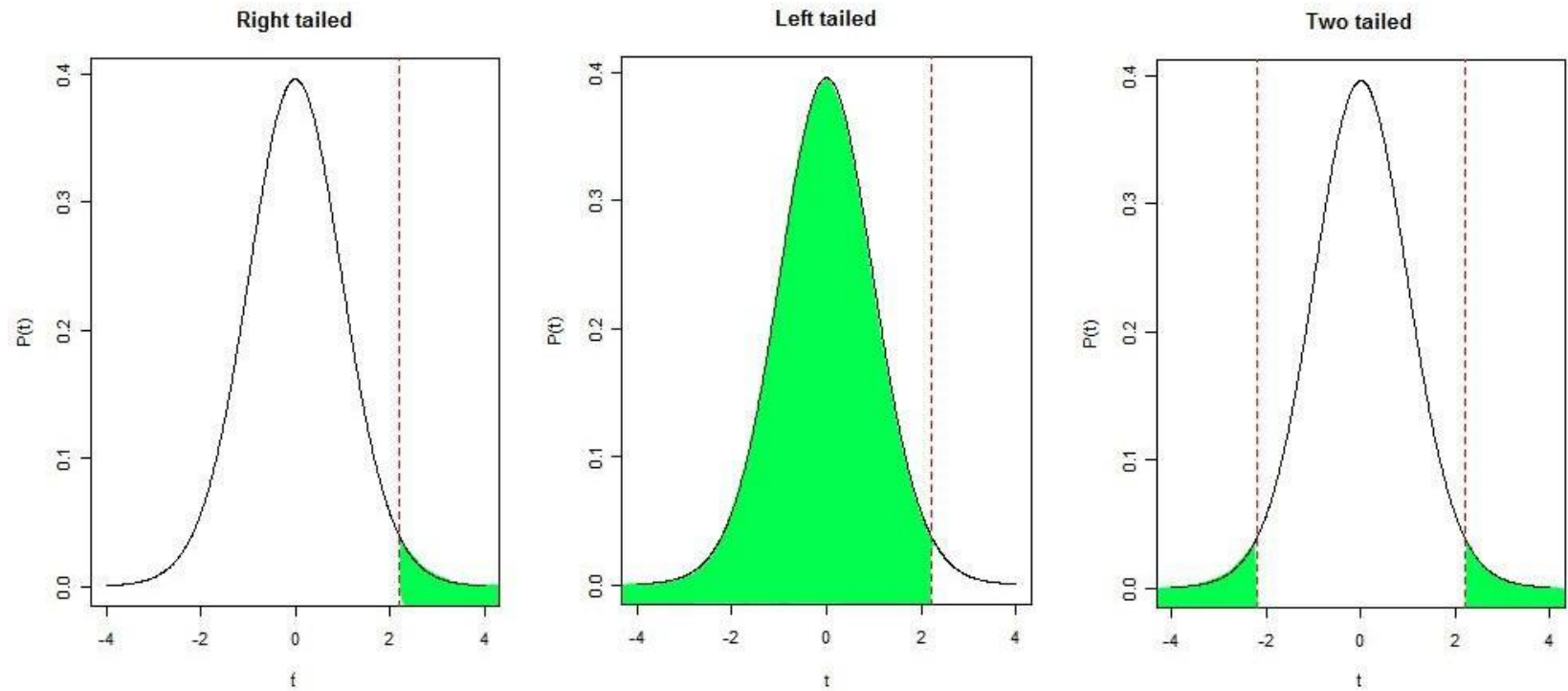
**Example:** Let  $n = 29$  and your test value be  $t = 2.21$ .

Find the p-value for a right-tailed hypothesis test.

Find the p-value for a left-tailed test.

Find the p-value for a right-tailed

Find the p-value for a two-tailed hypothesis



Degrees of freedom =  $29 - 1 = 28$

Table F The $t$ Distribution						
	Confidence intervals	80%	90%	95%	98%	99%
	One tail, $\alpha$	0.10	0.05	0.025	0.01	0.005
d.f.	Two tails, $\alpha$	0.20	0.10	0.05	0.02	0.01
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.132	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032
6		1.440	1.943	2.447	3.143	3.707
7		1.415	1.895	2.365	2.998	3.499
8		1.397	1.860	2.306	2.896	3.355
9		1.383	1.833	2.262	2.821	3.250
10		1.372	1.812	2.228	2.764	3.169
11		1.363	1.796	2.201	2.718	3.106
12		1.356	1.782	2.179	2.681	3.055
13		1.350	1.771	2.160	2.650	3.012
14		1.345	1.761	2.145	2.624	2.977
15		1.341	1.753	2.131	2.602	2.947
16		1.337	1.746	2.120	2.583	2.921
17		1.333	1.740	2.110	2.567	2.898
18		1.330	1.734	2.101	2.552	2.878
19		1.328	1.729	2.093	2.539	2.861
20		1.325	1.725	2.086	2.528	2.845
21		1.323	1.721	2.080	2.518	2.831
22		1.321	1.717	2.074	2.508	2.819
23		1.319	1.714	2.069	2.500	2.807
24		1.318	1.711	2.064	2.492	2.797
25		1.316	1.708	2.060	2.485	2.787
26		1.315	1.706	2.056	2.479	2.779
27		1.314	1.703	2.052	2.473	2.771
28		1.313	1.701	2.048	2.467	2.763
29		1.311	1.699	2.045	2.462	2.756

### Right-tailed p-value

☐ is between 0.025 and 0.01 based on the table

☐ is 0.018 based on the calculator

### Left-tailed p-value

☐ is between 0.975 and 0.99 based on the table

☐ is 0.982 based on the calculator

### Two-tailed p-value

☐ is between 0.05 and 0.02 based on the table

☐ is 0.035 based on the calculator

**Right tailed**

**Left tailed**

**Two-tailed**

**Example:** Let  $n = 11$  and your test value be  $t = 1.35$ .

a left-tailed hypothesis test. Find the p-value for a right-tailed hypothesis test.

a two-tailed hypothesis test.

Find the p-value for

Find the p-value for

**Example:** We wish to check that normal body temperature may be less than 98.6 degrees. In a random sample of  $n = 18$  individuals, the sample mean was found to be 98.217 and the standard deviation was .684. Assume the population is normally distributed. Use  $\alpha = 0.05$ .

**Step 1 State the hypotheses and identify the claim.**

$H_0: \mu = 98.6$

$H_1: \mu < 98.6$                        $\mu$                       CLAIM

**Step 2 Find the critical value(s) from the appropriate table.** Left tailed,  $\alpha = 0.05$ ,  $df = 18 - 1 = 17$                        $t$  critical

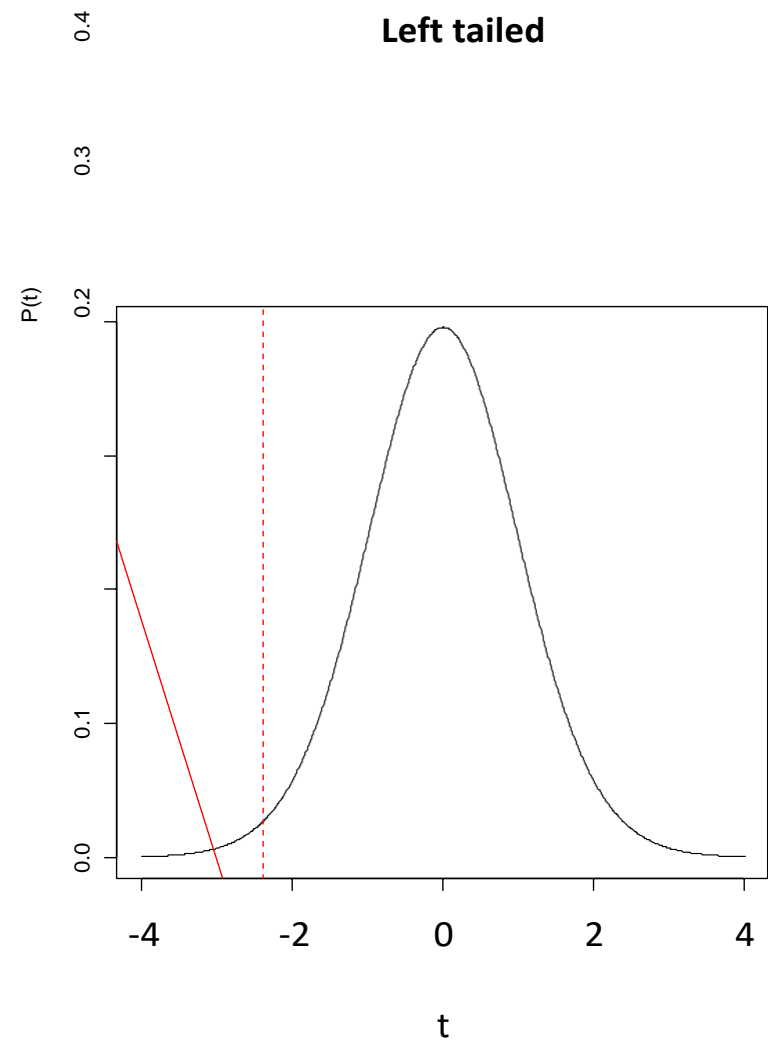
value = 1.740 **Step 3 Compute the test value and determine the P-value.**



$$t = \frac{98.217 - 98.6}{0.684 \sqrt{18}} = -2.375631 \approx -2.38$$

p-value is between 0.01 and 0.025

p-value = 0.0146



**Step 4 Make the decision to reject or not reject the null hypothesis.**

Since our p-value is less than our  $\alpha = 0.05$ , we reject the null hypothesis.

The same conclusion is reached by looking at the critical value. Our test value is smaller than the critical value of -1.74.

You only need to do it one way. The decision will always match.

**Step 5 Summarize the results.**

We have enough evidence to support the claim that average body temperature is less than 98.6 degrees.

**Example:** A certain company would like to determine the amount of time employees waste at work each day. A random sample of 10 of its employees shows a mean time of 121.80 minutes wasted per day with a standard deviation of 9.45 minutes per day. Does the data provide evidence that the mean amount of time wasted by employees each day is more than 120 minutes? Test at  $\alpha=.05$ . Assume the population is at least approximately normally distributed.

Let  $\mu$  = mean daily wasted time for employees of this company.

**Step 1 State the hypotheses and identify the claim.**

**Step 2 Find the critical value(s) from the appropriate table.**

**Step 3 Compute the test value and determine the P-value.**

**Step 4 Make the decision to reject or not reject the null hypothesis.**

**Step 5 Summarize the results.**

Suppose the previous test were two-tailed instead of one-tailed. What would the critical values/regions be?

What would the p-value for the test be?

What would our decision be?

## z-Test for a Proportion

A hypothesis test involving proportions can be considered as a binomial experiment when there are only two outcomes and the probability of success does not change from trial to trial.

However, when both  $n\hat{p}$  and  $n\hat{q}$  are each greater than or equal to 5, the Central Limit Theorem kicks in and a normal distribution can be used to describe the distribution of the sample proportions.

**Recall:** When the central limit theorem applies to data from a binomial distribution, then  $\hat{p}$  can be well approximated by a normal distribution with mean  $p$  and standard deviation

$$\sqrt{pq/n}, \text{ i.e.,}$$

$\hat{p}$  is approximately Normal with mean  $=p$

and standard error =

$$\sqrt{\frac{pq}{n}}.$$

When performing our hypothesis test, we will make an assumption about the value of  $p$  in our null hypothesis, e.g.,

$$H_0 : p \leq k$$

where  $k$  is some fixed value.  
test for  $p$ .

We can use this to determine to perform a hypothesis



## Formula for the z-Test for Proportions

Test value:  $z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$

$$\sqrt{\frac{pq}{n}}$$

$$\hat{p} =$$

$$p = q$$

$$= n =$$

Since this is a **z test** we can use Table E to find critical values and p-values.  
 values using Table F using the row with degrees of freedom row  $\infty$ .

However, it is easier to find critical

Pick appropriate tail and

?

Table F	The <i>t</i> Distribution					
	Confidence intervals	80%	90%	95%	98%	99%
	One tail, $\alpha$	0.10	0.05	0.025	0.01	0.005
d.f.	Two tails, $\alpha$	0.20	0.10	0.05	0.02	0.01
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
100		1.290	1.660	1.984	2.364	2.626
500		1.283	1.648	1.965	2.334	2.586
1000		1.282	1.646	1.962	2.330	2.581
(z) $\infty$		1.282 <sup>a</sup>	1.645 <sup>b</sup>	1.960	2.326 <sup>c</sup>	2.576 <sup>d</sup>



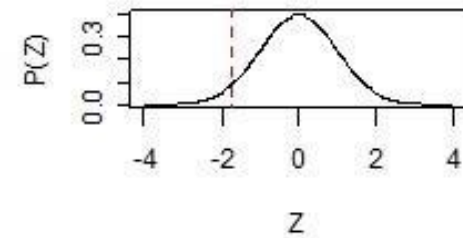
The p-values can be found according to the following procedure. If  $z^*$  is the test value of our test then:

**Left-tailed test:** p-value  $= P(Z \leq z^*)$

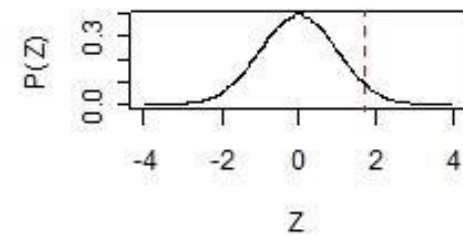
**Right-tailed test:** p-value  $= P(Z \geq z^*)$

**Two-tailed test:** p-value  $= 2P(Z \geq z^*)$

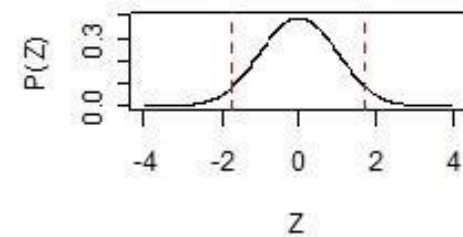
**Left tailed**



**Right tailed**



**Two tailed**



**Example:** An educator estimates that the dropout rate for seniors at high schools in Colorado is 12%. Last year in a random sample of 300 Colorado seniors, 27 withdrew from school. At  $\alpha = 0.05$ , is there enough evidence to reject the educator's claim?

**Step 1** State the hypotheses and identify the claim.  $H_0: p = 0.12$     $H_1: p \neq 0.12$

CLAIM

**Step 2** Find the critical value(s) from the appropriate table.



Two-tailed,  $\alpha = 0.05$

.....

critical value is  $Z = \pm 1.96$

*(peak back at slide 333)*

**Step 3** Compute the test value and determine the P-value.

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.12 - 0.12}{\sqrt{\frac{0.12(1-0.12)}{300}}} = -1.60$$

0.01876166

$$\begin{aligned} \text{p-value} &= 2*(0.0548) \\ &= 0.1096 \end{aligned}$$

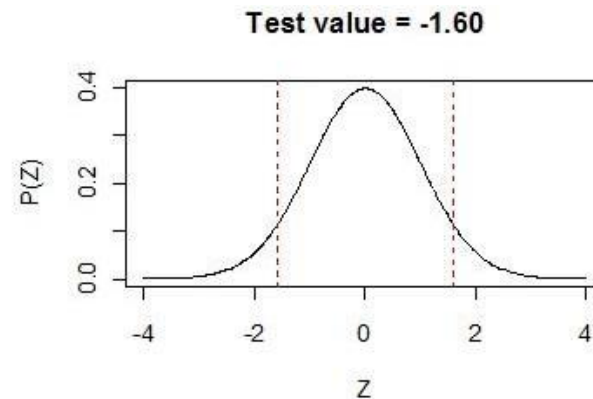


Table E The Standard Normal Distribution			
Cumulative Standard Normal Distribution			
z	.00	.01	.02
-3.4	.0003	.0003	.0003
-3.3	.0005	.0005	.0005
-3.2	.0007	.0007	.0006
-3.1	.0010	.0009	.0009
-3.0	.0013	.0013	.0013
-2.9	.0019	.0018	.0018
-2.8	.0026	.0025	.0024
-2.7	.0035	.0034	.0033
-2.6	.0047	.0045	.0044
-2.5	.0062	.0060	.0059
-2.4	.0082	.0080	.0078
-2.3	.0107	.0104	.0102
-2.2	.0139	.0136	.0132
-2.1	.0179	.0174	.0170
-2.0	.0228	.0222	.0217
-1.9	.0287	.0281	.0274
-1.8	.0359	.0351	.0344
-1.7	.0446	.0436	.0427
-1.6	.0548	.0537	.0526

**Step 4** Make the decision to reject or not reject the null hypothesis.

Noting that our p-value is larger than our Type I error rate of 0.05, we fail to reject the null hypothesis.

You come to the same decision by noting that our test value of

1.60 falls within the non-critical region of -1.96 to 1.96.

**Step 5** Summarize the results.

We do not have sufficient evidence to reject the claim that the dropout rate for seniors at high schools in Colorado is 12%.

*Check check! Were  $n\hat{p}$  and  $n\hat{q}$  both greater than or equal to 5?*

Let's generate some questions about Math 2830 students that would allow us to test a hypothesis concerning proportions using our class as the sample.

1.

2.

3.

4.

Question # \_\_\_\_\_

What level of significance would we like to use? \_\_\_\_\_

Our class data:

**Step 1** State the hypotheses and identify the claim.

**Step 2** Find the critical value(s) from the appropriate table.



**Step 3** Compute the test value and determine the P-value.

**Step 4** Make the decision to reject or not reject the null hypothesis.

**Step 5** Summarize the results.

Question # \_\_\_\_\_

What level of significance would we like to use? \_\_\_\_\_

Our class data:

**Step 1** State the hypotheses and identify the claim.

**Step 2** Find the critical value(s) from the appropriate table.

**Step 3** Compute the test value and determine the P-value.

**Step 4** Make the decision to reject or not reject the null hypothesis.

**Step 5** Summarize the results.

## Unit – 3

### Probability space

The notion of probability space is at the heart of the mathematical treatment of probability. From the point of view of this course, you can think of it as a unifying idea: it unifies discrete and continuous probability.

A **probability space** is a triple  $(\Omega, \mathcal{A}, P)$  where  $\Omega$  is the sample space,  $\mathcal{A}$  is a  $\sigma$ -field of subsets of  $\Omega$  and  $P$  is a probability measure on  $\Omega$ .

Understanding this requires two definitions.

**Definition** A nonempty collection of subsets  $\mathcal{A}$  of a set  $\Omega$  is a  **$\sigma$ -field** if it has the following two properties.

(i) If  $A \in \mathcal{A}$  then  $\bar{A} = \Omega \setminus A \in \mathcal{A}$ .

(ii) If  $A_n \in \mathcal{A}$ ,  $n = 1, 2, \dots$ , then  $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$ .

Note that (i) and (ii) imply that  $\bigcap_{n=1}^{\infty} A_n \in \mathcal{A}$  since  $\bigcap_{n=1}^{\infty} A_n = \Omega \setminus \left( \bigcup_{n=1}^{\infty} (\Omega \setminus A_n) \right)$ .

**$\sigma$ -algebra** is another name for  $\sigma$ -field.

**Definition** A **probability measure**  $P$  on a  $\sigma$ -field  $\mathcal{A}$  is a real valued function having domain  $\mathcal{A}$  satisfying the following properties:

(i)  $P(\Omega) = 1$ .

(ii)  $P(A) \geq 0$  for all  $A \in \mathcal{A}$ .

(iii) If  $A_n$ ,  $n = 1, 2, 3, \dots$ , are pairwise disjoint sets in  $\mathcal{A}$  (so  $A_i \cap A_j = \emptyset$  if  $i \neq j$ ) then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

**Example 1** If  $\Omega$  is a discrete set sample set with a random variable  $X$  and a distribution function  $m$  for  $X$  then we can take  $\mathcal{A}$  to be the collection of all subsets of  $\Omega$  and the probability measure  $P$  to be given by

$$P(E) = \sum_{\omega \in E} m(\omega).$$

**Example 2** If  $\Omega \subseteq \mathbb{R}^n$  (e.g.,  $\Omega$  is an interval or a nice region in the plane) and  $X$  is a continuous real-valued random variable on  $\Omega$  with a density function  $f$  then

$$P(E) = P(X \in E) = \int_E f(x) dx$$

defines a probability measure on any  $\sigma$ -field of nice subsets of  $\Omega$ .

I don't want to get into technicalities about what subsets are nice, but I will note that if  $\Omega$  is an interval in  $\mathbb{R}$  we can take the  $\sigma$ -field to be the **Borel sets**, the smallest  $\sigma$ -field containing all the open subintervals of  $\Omega$ . It is clear what  $P(E)$  means if  $E$  is an interval. The hard work is to show that  $P(E)$  can be defined for any Borel set  $E$ .

In this example it is not possible to take  $\mathcal{A}$  to be the  $\sigma$ -field of all subsets of  $\Omega$ .

# Conditional Probability

## 1 Conditional Probability

In English, a conditional probability answers the question: “What is the chance of an event  $E$  happening, given that I have already observed some other event  $F$ ?” Conditional probability quantifies the notion of updating one’s beliefs in the face of new evidence.

When you condition on an event happening you are entering the universe where that event has taken place. Mathematically, if you condition on  $F$ , then  $F$  becomes your new sample space. In the universe where  $F$  has taken place, all rules of probability still hold!

The definition for calculating conditional probability is:

### Definition of Conditional Probability

The probability of  $E$  given that (aka conditioned on) event  $F$  already happened:

$$P(E | F) = \frac{P(EF)}{P(F)} = \frac{P(E \cap F)}{P(F)}$$

(As a reminder,  $EF$  means the same thing as  $E \cap F$ —that is,  $E$  “and”  $F$ .)

A visualization might help you understand this definition. Consider events  $E$  and  $F$  which have outcomes that are subsets of a sample space with 50 equally likely outcomes, each one drawn as a hexagon:

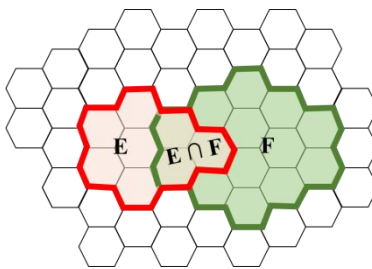


Figure 1: Conditional Probability Intuition

Conditioning on  $F$  means that we have entered the world where  $F$  has happened (and  $F$ , which has 14 equally likely outcomes, has become our new sample space). Given that event  $F$  has occurred, the conditional probability that event  $E$  occurs is the subset of the outcomes of  $E$  that are consistent

with  $F$ . In this case we can visually see that those are the three outcomes in  $E \cap F$ . Thus we have the:

$$P(E | F) = \frac{P(EF)}{P(F)} = \frac{3/50}{14/50} = \frac{3}{14} \approx 0.21$$

Even though the visual example (with equally likely outcome spaces) is useful for gaining intuition, the above definition of conditional probability applies regardless of whether the sample space has equally likely outcomes.

### The Chain Rule

The definition of conditional probability can be rewritten as:

$$P(EF) = P(E | F)P(F)$$

which we call the Chain Rule. Intuitively it states that the probability of observing events  $E$  and  $F$  is the probability of observing  $F$ , multiplied by the probability of observing  $E$ , given that you have observed  $F$ . Here is the general form of the Chain Rule:

$$P(E_1 E_2 \dots E_n) = P(E_1)P(E_2 | E_1) \dots P(E_n | E_1 E_2 \dots E_{n-1})$$

## 2 Law of Total Probability

An astute person once observed that in a picture like the one in Figure 1, event  $F$  can be thought of as having two parts, the part that is in  $E$  (that is,  $E \cap F = EF$ ), and the part that isn't ( $E^C \cap F = E^C F$ ). This is true because  $E$  and  $E^C$  are mutually exclusive sets of outcomes which together cover the entire sample space. After further investigation this was proved to be a general mathematical truth, and there was much rejoicing:

$$P(F) = P(EF) + P(E^C F)$$

This observation is called the **law of total probability**; however, it is most commonly seen in combination with the chain rule:

### The Law of Total Probability

For events  $E$  and  $F$ ,

$$P(F) = P(F | E)P(E) + P(F | E^C)P(E^C)$$

There is a more general version of the rule. If you can divide your sample space into any number of events  $E_1, E_2, \dots, E_n$  that are *mutually exclusive* and *exhaustive*—that is, *every* outcome in sample space falls into *exactly one* of those events—then:

$$P(F) = \sum_{i=1}^n P(F | E_i)P(E_i)$$

The word “total” refers to the fact that the events in  $E_i$  must combine to form the totality of the sample space.

### 3 Bayes' Theorem

Bayes' theorem (or **Bayes' rule**) is one of the most ubiquitous results in probability for computer scientists. Very often we know a conditional probability in one direction, say  $P(E | F)$ , but we would like to know the conditional probability in the other direction. Bayes' theorem provides a way to convert from one to the other. We can derive Bayes' theorem by starting with the definition of conditional probability:

$$P(E | F) = \frac{P(F \cap E)}{P(F)}$$

Now we can expand  $P(F \cap E)$  using the chain rule, which results in Bayes' theorem.

#### Bayes' theorem

The most common form of Bayes' theorem is:

$$P(E | F) = \frac{P(F | E)P(E)}{P(F)}$$

Each term in the Bayes' rule formula has its own name. The  $P(E | F)$  term is often called the **posterior**; the  $P(E)$  term is often called the **prior**; the  $P(F | E)$  term is called the **likelihood** (or the “update”); and  $P(F)$  is often called the **normalization constant**.

If the normalization constant (the probability of the event you were initially conditioning on) is not known, you can expand it using the law of Total Probability:

$$P(E | F) = \frac{P(F | E)P(E)}{P(F | E)P(E) + P(F | E^c)P(E^c)} = \frac{P(F | E)P(E)}{\sum_i P(F | E_i)P(E_i)}$$

Again, for the last version, all the events  $E_i$  must be *mutually exclusive* and *exhaustive*.

A common scenario for applying the Bayes Rule formula is when you want to know the probability of something “unobservable” given an “observed” event. For example, you want to know the probability that a student understands a concept, given that you observed them solving a particular problem. It turns out it is much easier to first estimate the probability that a student can solve a problem given that they understand the concept and then to apply Bayes' theorem.

The “expanded” version of Bayes' rule (at the bottom of the Bayes' theorem box) allows you to work around not immediately knowing the denominator  $P(F)$ . It is worth exploring this in more depth, because this “trick” comes up often, and in slightly different forms. Another way to get to the exact same result is to reason that because the posterior of Bayes' Theorem,  $P(E | F)$ , is a probability, we know that  $P(E | F) + P(E^c | F) = 1$ . If you expand out  $P(E^c | F)$  using Bayes, you get:

$$P(E^c | F) = \frac{P(F | E^c)P(E^c)}{P(F)}$$

Now we have:

$$\begin{aligned}
 1 &= P(E | F) + P(E^C | F) && \text{since } P(E | F) \text{ is a probability} \\
 1 &= \frac{P(F | E)P(E)}{P(F)} + \frac{P(F | E^C)P(E^C)}{P(F)} && \text{by Bayes' rule (twice)} \\
 1 &= \frac{1}{P(F)} [P(F | E)P(E) + P(F | E^C)P(E^C)] \\
 P(F) &= P(F | E)P(E) + P(F | E^C)P(E^C)
 \end{aligned}$$

We call  $P(F)$  the normalization constant because it is the term whose value can be calculated by making sure that the probabilities of all outcomes sum to 1 (they are “normalized”).

## 4 Conditional Paradigm

As we mentioned above, when you condition on an event you enter the universe where that event has taken place, all the laws of probability still hold. Thus, as long as you condition consistently on the same event, every one of the tools we have learned still apply. Let's look at a few of our old friends when we condition consistently on an event (in this case  $G$ ):

Name of Rule	Original Rule	Conditional Rule
First axiom of probability	$0 \leq P(E) \leq 1$	$0 \leq P(E   G) \leq 1$
Corollary 1 (complement)	$P(E) = 1 - P(E^C)$	$P(E   G) = 1 - P(E^C   G)$
Chain Rule	$P(EF) = P(E   F)P(F)$	$P(EF   G) = P(E   FG)P(F   G)$
Bayes Theorem	$P(E   F) = \frac{P(F   E)P(E)}{P(F)}$	$P(E   FG) = \frac{P(F   EG)P(E   G)}{P(F   G)}$



## UNIT-4

①

### Continuous random 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.

#### Example:

If in the study of the ecology of a lake,  $X$ , the random variables may be depth measurements at randomly chosen locations.

Then  $X$  is a continuous random variable. The range for  $X$  is the minimum depth possible to the maximum depth possible.

- Discrete random variables: where the possible events are countable.

#### Example:

The roll of a dice, or the outcome of a horse race, or whether the firm will default or not.



(2)

- Continuous : where the possible outcomes / events are not countable.

Example :

The number of white hairs on my head, or how much dividend INFOSYSTCH will announce next year, or the price of citibank stock



## # Probability Density Function and Cumulative Distribution Function: ③

A random variable is said to be continuous if there exists a real-valued function  $f_x$  such that, for any subset  $B \subseteq \mathbb{R}$ :

$$P(X \in B) = \int_B f_x(x) dx \quad - (1)$$

Then  $f_x$  is called probability density function (pdf) of the random variable  $X$ .

In particular, for any real numbers  $a$  and  $b$ , with  $a < b$ , letting  $B = [a, b]$ , we obtain from eq (1) that:

$$P(a \leq X \leq b) = \int_a^b f_x(x) dx \quad - (2)$$

Prop 1.1 If  $X$  is a continuous random variable, then

- for all  $a \in \mathbb{R}$

$$P(X = a) = 0 \quad - (3)$$



④ In other words, the probability that a continuous random variable takes on any fixed value is zero.

- for any real numbers  $a$  and  $b$ , with  $a < b$

$$P(a \leq x \leq b) = P(a \leq x < b) = P(a < x \leq b) = P(a < x < b) \quad \hookrightarrow (u)$$

The above eq<sup>n</sup> states that including or not the bounds of an interval does not modify the probability of a continuous rv.

Proof: • let us first prove eq<sup>n</sup> (3)

$$P(x = a) = P(x \in [a, a]) = \int_a^a f_X(x) dx = 0$$

- To prove eq<sup>n</sup> (4)

$$\begin{aligned} P(a \leq x \leq b) &= P(a \leq x < b) = P(a < x \leq b) = P(a < x < b) \\ &= \int_a^b f_X(x) dx. \end{aligned}$$



⑤

Prop 1.2 Let  $F_X$  be the cdf of a random variable  $X$ . Following are some prop of  $F_X$ :

- $F_X$  is increasing:  $x \leq y \Rightarrow F_X(x) \leq F_X(y)$
- $\lim_{x \rightarrow \infty} F_X(x) = 1$  and  $\lim_{x \rightarrow -\infty} F_X(x) = 0$
- $F_X$  is càdlàg:
  - $F_X$  is right continuous:  $\lim_{x \downarrow x_0} F_X(x) = F_X(x_0)$ ,  
for  $x_0 \in \mathbb{R}$
  - $F_X$  has left limits:  $\lim_{x \uparrow x_0} F_X(x)$  exists,  
 $x_0 \in \mathbb{R}$

Prop 1.3 Let  $X$  be a continuous rv with pdf  $f_X$ . Then the cumulative distribution function  $F_X$  of  $X$  is given by:

$$F_X(x) = \int_{-\infty}^x f_X(t) dt \quad - (5)$$

Proof: we have,

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(X \in (-\infty, x]) \\ &= \int_{-\infty}^x f_X(t) dt \end{aligned}$$



⑥

Theorem 1.1 . A probability density function completely determine the distribution of a continuous real-valued random variable.

Theorem 1.2 A cumulative distribution completely determine the distribution of a continuous real-valued random variable.

Probability distribution :

- For a discrete RV, the probability distribution (PD) is a table of all the events and their related probabilities.
- For example, in a roll of a die:

Value	Probability	Value	Probability
1	$1/6$	4	$1/6$
2	$1/6$	5	$1/6$
3	$1/6$	6	$1/6$



(7)

## Cumulative Probability Distribution (CD):

- From the distribution we can find:

$$X = 3$$

$$\Pr(X = 3) = 1/6$$

$X =$  even numbers

$$\Pr(X = 2 \text{ or } X = 4 \text{ or } X = 6) = 3/6 = 1/2$$

- we can also find:

$$\Pr(X > 3) = 3/6 = 1/2$$

- Now a table of probabilities cumulated over the events.

Value	Probability	Value	Probability
$X \leq 1$	$1/6$	$X \leq 4$	$4/6$
$X \leq 2$	$2/6$	$X \leq 5$	$5/6$
$X \leq 3$	$3/6$	$X \leq 6$	$1$

- The CD is monotonically increasing set of numbers.
- The CD always ends with at the highest value of 1.



⑤

Example: Calculate the cumulative distribution function of a random variable uniformly distributed over  $(\alpha, \beta)$

Sol<sup>n</sup> Since  $F(a) = \int_{-\infty}^a f(x) dx$ , we obtain from

$$e_s^m \left( f(x) = \begin{cases} \frac{1}{\beta - \alpha}, & \text{if } \alpha < x < \beta \\ 0, & \text{otherwise} \end{cases} \right)$$

$$F(a) = \begin{cases} 0, & a \leq \alpha \\ \frac{a - \alpha}{\beta - \alpha}, & \alpha < a < \beta \\ 1, & a \geq \beta \end{cases}$$

Example: If  $X$  is uniformly distributed over  $(0, 10)$ , calculate the probability that ~~at~~

(a)  $X < 3$

(b)  $X > 7$

(c)  $1 < X < 6$

(a)  $P(X < 3) = \frac{\int_0^3 dx}{10} = \frac{3}{10}$

(b)  $P(X > 7) = \frac{\int_7^{10} dx}{10} = \frac{3}{10}$

(c)  $P(1 < X < 6) = \frac{\int_1^6 dx}{10} = \frac{1}{2}$



(9)

## # Probability density functions:

- The probability density function (pdf) is the PD of a continuous random variable.

- Since continuous random variables are uncountable, it is difficult to write down the probabilities of all possible events.

Therefore, the PDF is always a function which gives the probability of one event,  $x$ .

- If we denote the PDF as function  $f$ , then

$$\Pr(X=x) = f(x)$$

### Problem

- In a set of continuous r.v., the probability of picking out a value of exactly  $x$  is zero.

- we define the  $\Pr(X=x)$  as follows:

$$\Pr(X \leq (x + \Delta)) - \Pr(X \leq x)$$

as  $\Delta$  becomes an infinitesimally small.

- Here  $\Pr(X \leq x)$  is the cumulative density func<sup>n</sup> of  $X$ .



(10)

Ques (1) Determine  $c$  such that  $f_x$  satisfies the properties of a pdf.

Sol<sup>n</sup> Since  $f_x$  is a pdf,  $f_x(n)$  should be nonnegative for all  $n \in \mathbb{R}$ . This is the case for  $n \in (-\infty, 0)$  and  $n \in (1/2, \infty)$  where  $f_x(n)$  equals zero. On the interval  $[0, 1/2]$ ,  $f_x(n) = c$ . This implies that  $c$  should be nonnegative as well.

Let us now focus on second cond<sup>n</sup>.

$$\int_{-\infty}^{\infty} f_x(n) dn = 1 \Leftrightarrow \int_{-\infty}^0 f_x(n) dn + \int_0^{1/2} f_x(n) dn + \int_{1/2}^{\infty} f_x(n) dn = 1$$

$$\Rightarrow \int_{-\infty}^0 0 dn + \int_0^{1/2} c dn + \int_{1/2}^{\infty} 0 dn = 1$$

$$\Rightarrow c \int_0^{1/2} 1 dn = 1$$

$$\Rightarrow c \cdot \frac{1}{2} = 1$$

$$\Rightarrow c = 2$$

And we check that indeed  $c = 2$  is nonnegative.



(d) Give the cdf of  $X$ .

(11)

Sol<sup>n</sup> The cumulative distribution function  $F_X$  of  $X$  is piecewise like its pdf:

- If  $x < 0$ , then  $f_X(t) = 0$  for all  $t \in (-\infty, x]$ .

$$F_X(x) = \int_{-\infty}^x f_X(t) dt = \int_{-\infty}^x 0 dt = 0$$

- If  $0 \leq x \leq 1/2$ , then  $f_X(t) = 2$  for all  $t \in [0, x]$ .

$$F_X(x) = \int_{-\infty}^x f_X(t) dt$$

$$= \int_{-\infty}^0 f_X(t) dt + \int_0^x f_X(t) dt$$

$$= F_X(0) + \int_0^x 2 dt$$

$$= 0 + 2t \Big|_0^x$$

$$= 2x$$

- If  $x > 1/2$ , then  $f_X(t) = 0$ , for all  $t \in [1/2, x]$ .



(12)

$$F_x(n) = \int_{-\infty}^n f_n(t) dt$$

$$= \int_{-\infty}^{1/2} f_n(t) dt + \int_{1/2}^n f_n(t) dt$$

$$= F_x(1/2) + 0$$

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

$$= 1$$

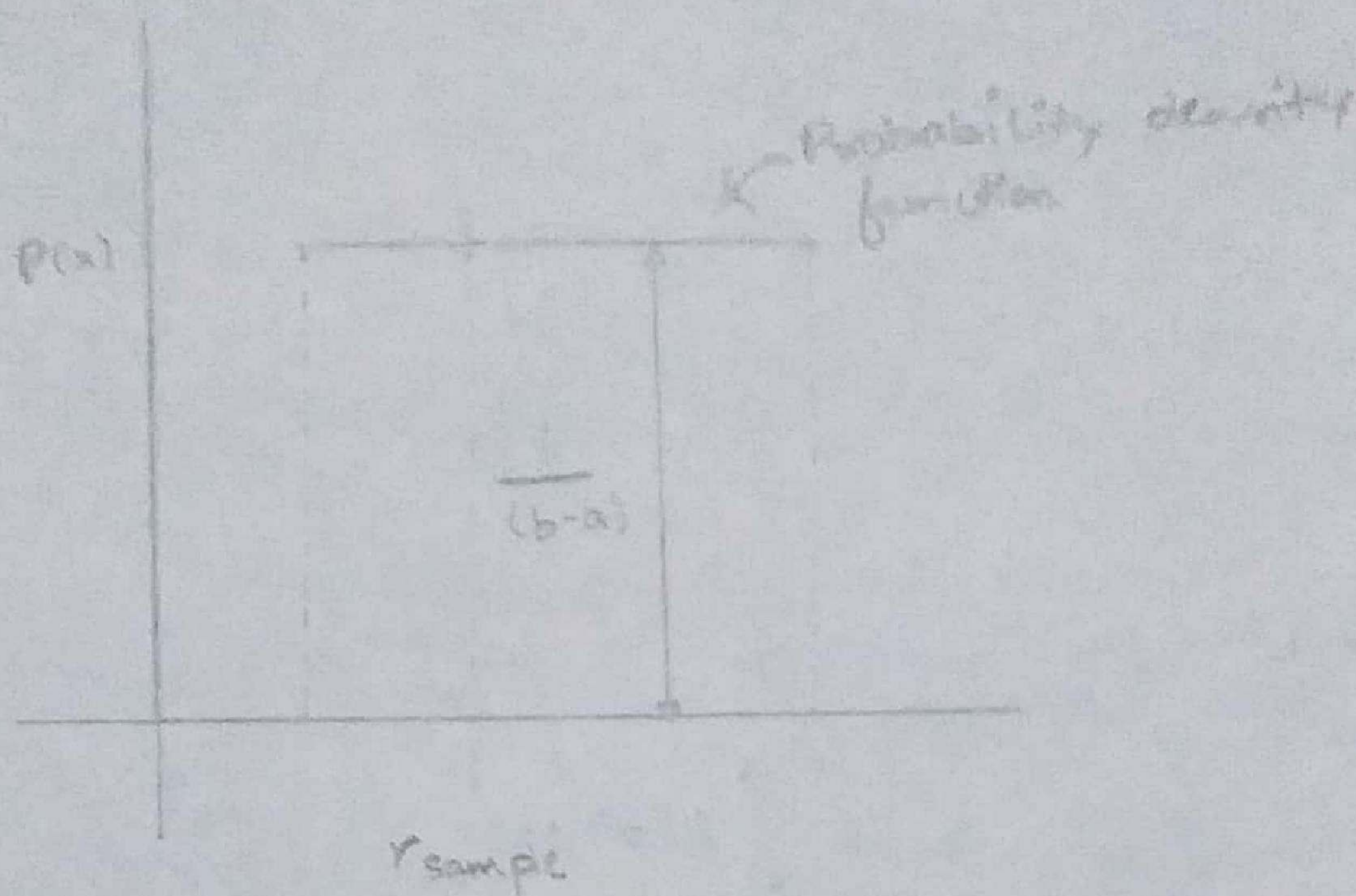
In a nutshell,  $F_x$  is given by:

$$F_x(n) = \begin{cases} 0 & \text{if } n < 0 \\ 2n & \text{if } 0 \leq n \leq 1/2 \\ 1 & \text{if } n > 1/2. \end{cases}$$



## # Uniform distribution:

(13)



- The pdf for values uniformly distributed across  $[a, b]$  is given by

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

### Sampling

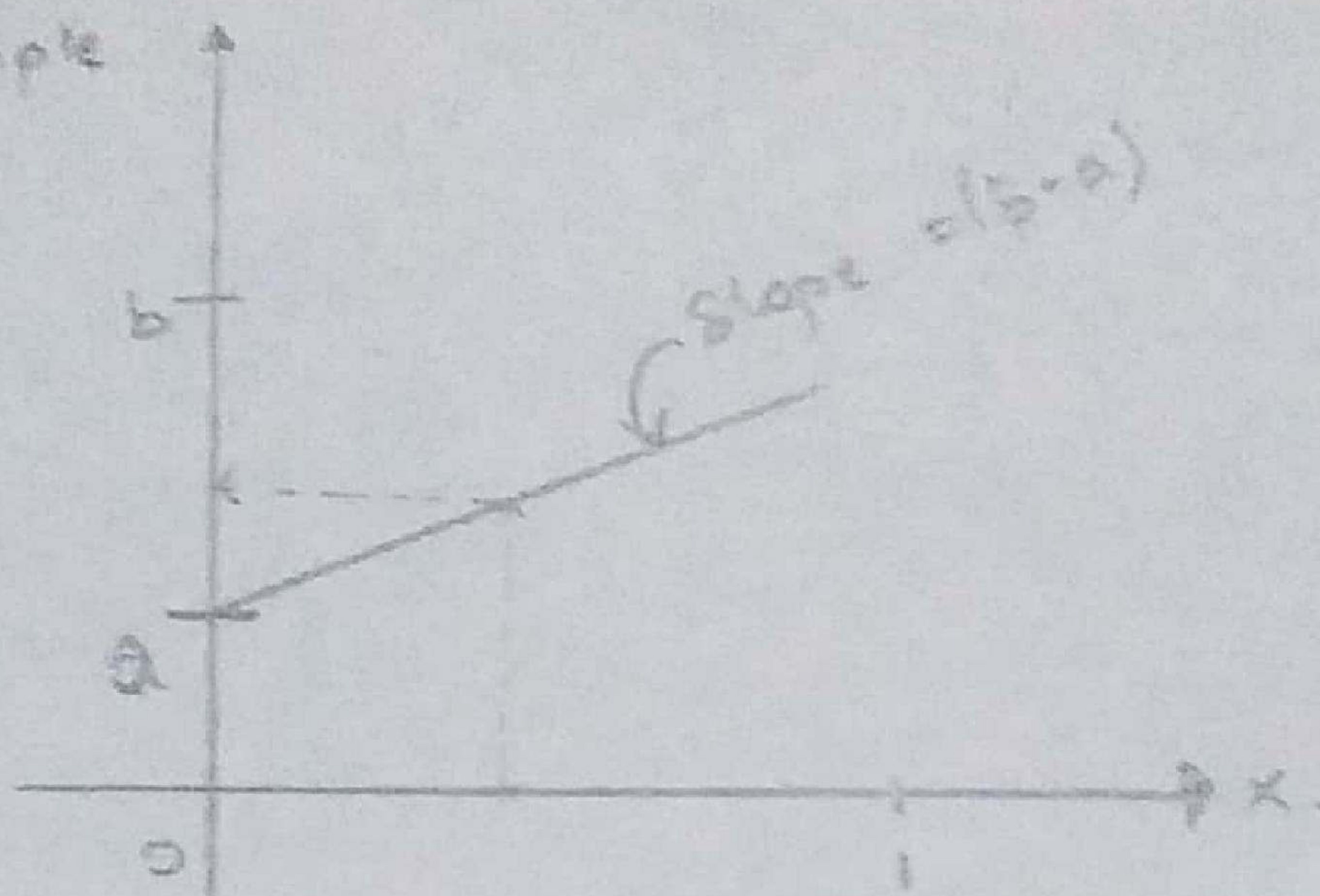
- Random numbers  $x$  drawn from  $[0, 1]$  distribute uniformly across the unit interval, so it is evident that the corresponding values

$$r_{\text{sample}} = a + x(b-a)$$

will distribute uniformly across  $[a, b]$



(14)

 $v_{\text{sample}}$ 

• directly solving

$$x = \int_{-\infty}^{v_{\text{sample}}} f(z) dz$$

for  $v_{\text{sample}}$  as per

$$x = \int_a^{v_{\text{sample}}} \frac{1}{b-a} dz = \left[ \frac{z}{b-a} \right]_a^{v_{\text{sample}}} = \frac{v_{\text{sample}}}{b-a} - \frac{a}{b-a}$$

also yields  $v_{\text{sample}} = a + x(b-a)$



## # Mean and Variance for the uniform <sup>(13)</sup> distribution:

- The mean  $\mu$  of the uniform distribution is given by

$$\mu = E(x) = \int_a^b z \left( \frac{1}{b-a} \right) dz = \frac{b^2 - a^2}{2} \frac{1}{b-a} = \frac{b+a}{2}$$

- The standard deviation ( $\sigma$ ) of the uniform distribution is obtained from the variance  $\sigma^2$  where

$$\sigma^2 = E((x-\mu)^2) = \int_a^b \left( z - \frac{b+a}{2} \right)^2 \left( \frac{1}{b-a} \right) dz = \frac{(b-a)^2}{12}$$



(16)

## # Normal Distribution :

- For a finite population the mean ( $m$ ) and standard deviation ( $s$ ) provide a measure of average value and degree of variation from the average value.
- If random samples of size  $n$  are drawn from the population, then it can be shown that the distribution of the sample means approximates that of a distribution with

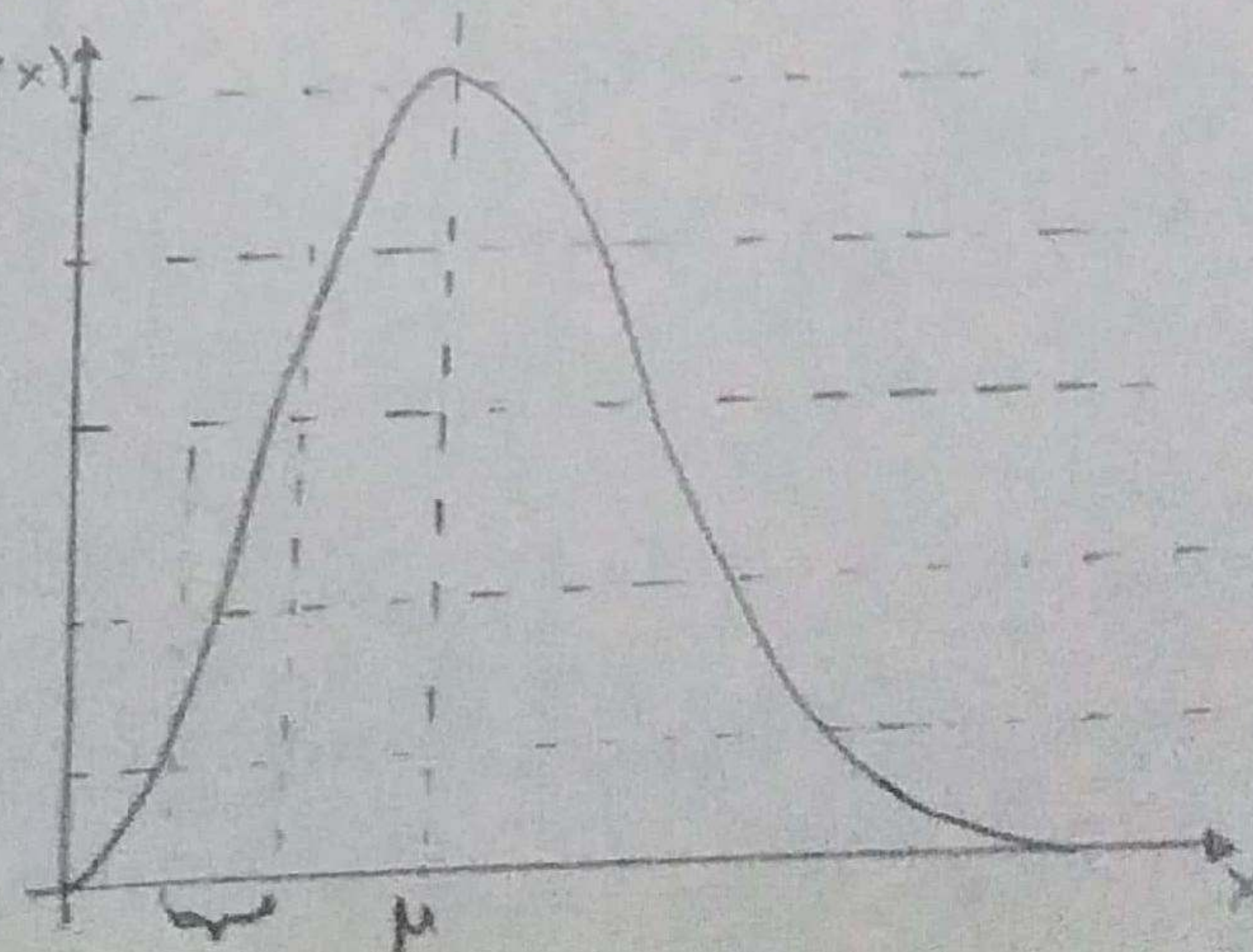
mean :  $\mu = m$  and  $\sigma$

standard deviation :  $\sigma = \frac{s}{\sqrt{n}}$

Pdf : 
$$f(m) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(m-\mu)^2}{2\sigma^2}}$$

Example plot :  $f(x)$

- plot with  $\mu = 30$   
and  $\sigma = 10$



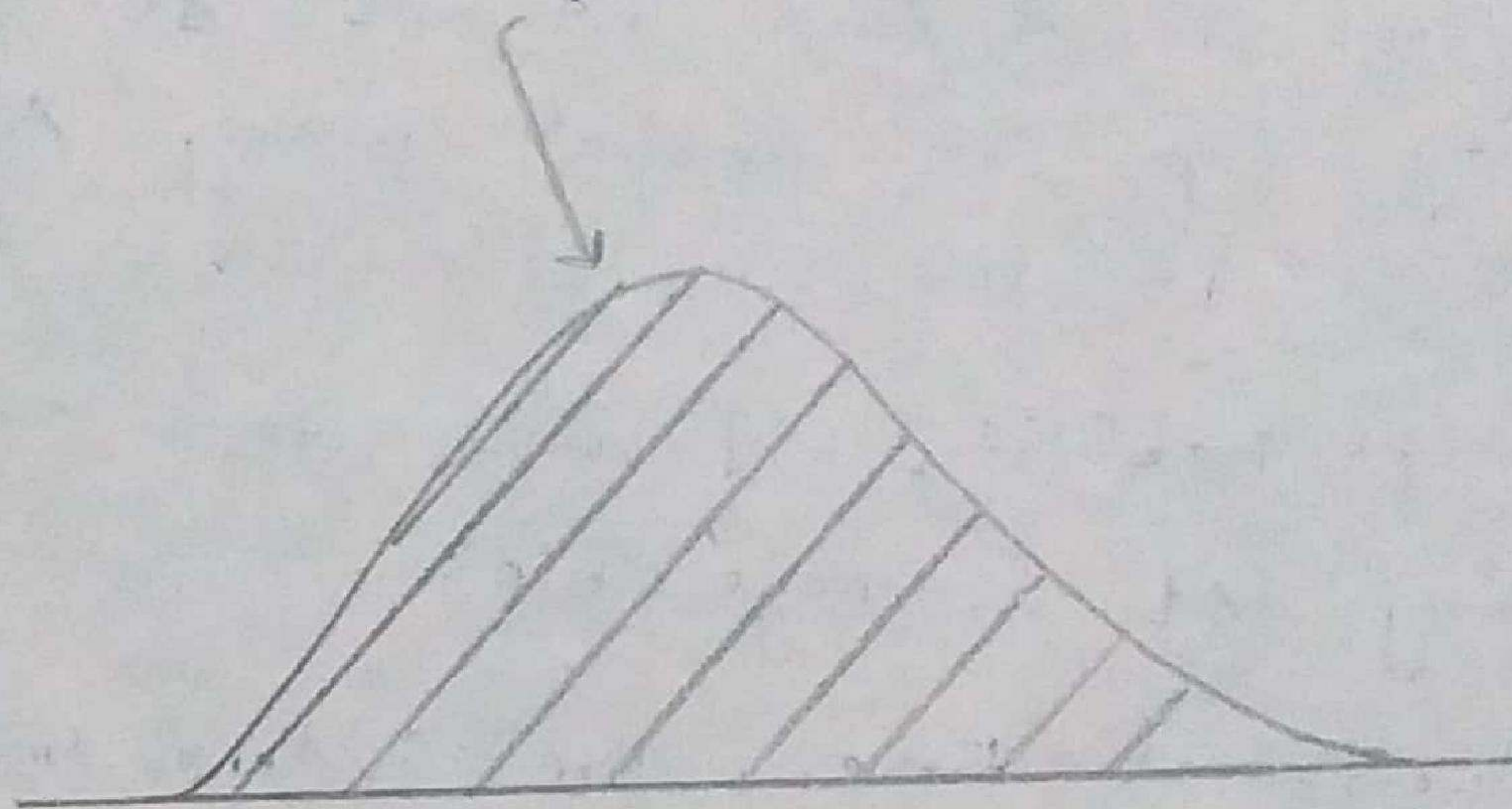


## # Standard normal distribution

(14)

- The standard normal distribution is given by  $\mu = 0$  and  $\sigma = 1$ , in which case the pdf becomes

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



$$X = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$



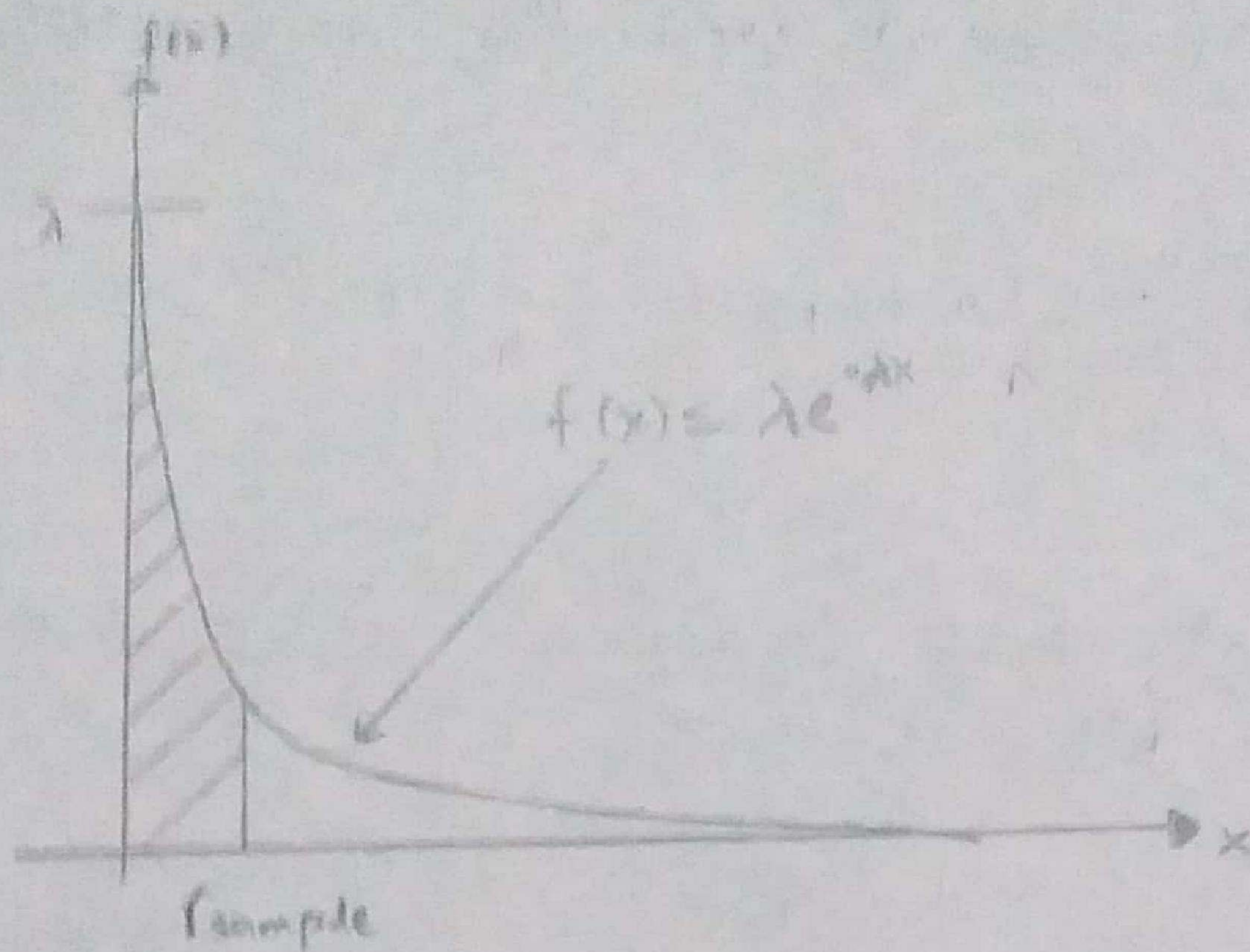
(18)

## # Exponential Distribution:

- The exponential distribution arises in connection with Poisson processes.
  - A Poisson process is one exhibiting a random arrival pattern in the following:
    - For a small time interval  $\Delta t$ , the probability of an arrival during  $\Delta t$  is  $\lambda \Delta t$ , where  $\lambda =$  the mean arrival rate.
    - The probability of more than one arrival during  $\Delta t$  is negligible.
    - Interarrival times are independent of each other.
  - This is a kind of "stochastic" process, one for which events occur in a random fashion.
- Under these assumptions, it can be shown that the pdf for the distribution of interarrival times is given by
$$f(x) = \lambda e^{-\lambda x}$$
which is the exponential distribution.



## ## Graphical Appearance (Exponential Distribution):



## ## Standard Exponential Distribution:

• The case  $\lambda = 1$  gives the standard exponential distribution

- Investing is straightforward since

$$\int_0^{n_{\text{sample}}} e^{-z} dz = -e^{-z} \Big|_0^{n_{\text{sample}}} = 1 - e^{-n_{\text{sample}}} = x \text{ so } n_{\text{sample}} = -\log_e(1-x)$$

- closed form formula for obtaining a normalized sample value ( $n_{\text{sample}}$ ) using a random probability  $x$

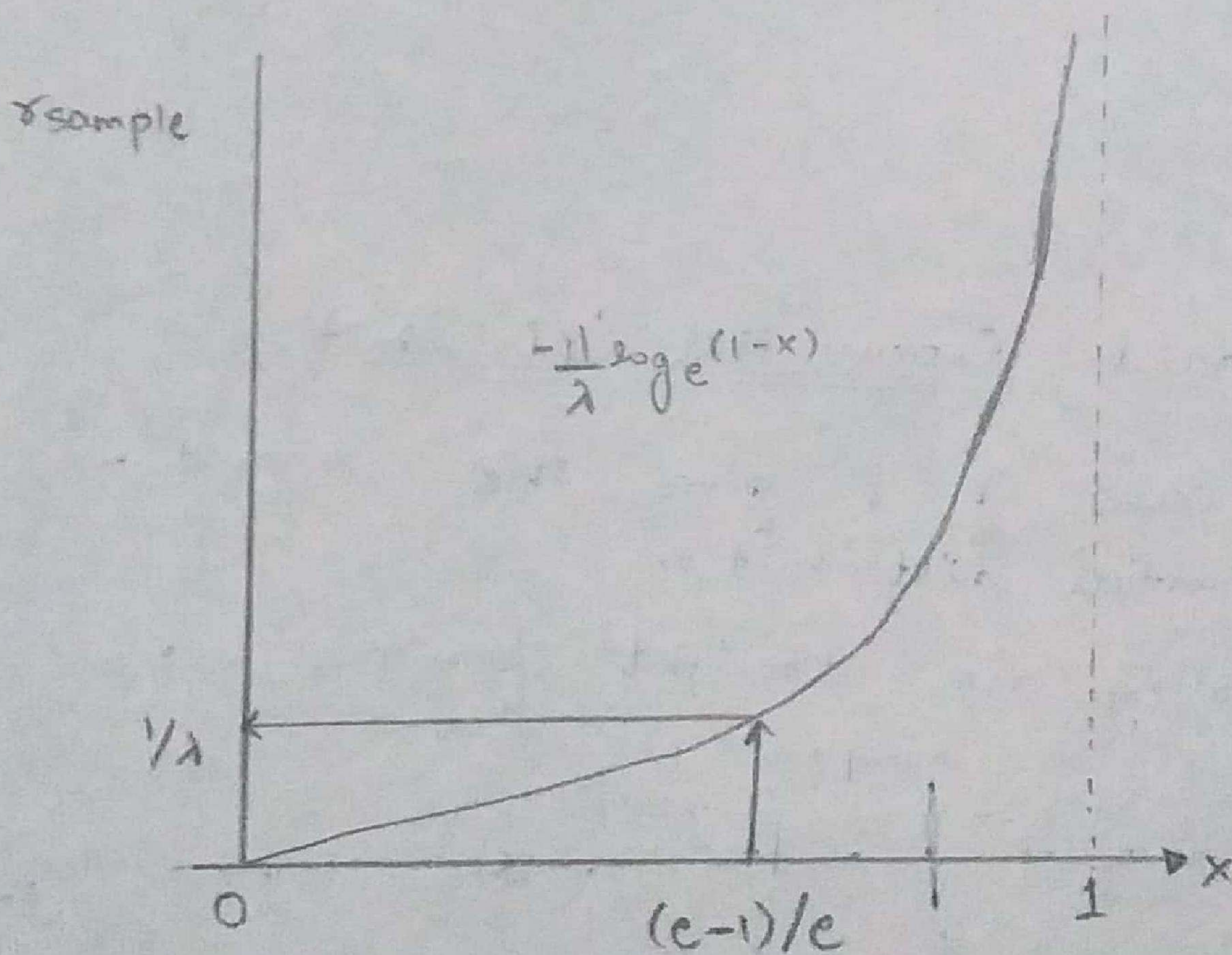


(20)

• General sample values ( $r_{\text{sample}}$ ) are obtained from the standard exponential distribution by

$$r_{\text{sample}} = \frac{1}{\lambda} r_{\text{sample}} = -\frac{1}{\lambda} \log e^{(1-x)}$$

# Sampling function for the standard exponential distribution:





## # Bivariate distributions and their prop. <sup>(2)</sup>

### 1. Distributions of two random variables:

Let  $X$  and  $Y$  be two random variables on probability space  $(\Omega, \mathcal{A}, \mathbb{P})$ . A two-dimensional random variable  $(X, Y)$  is a function mapping  $(X, Y):$

$\Omega \rightarrow \mathbb{R}^2$ , such that for any numbers  $x, y \in \mathbb{R}$

$$\{\omega \in \Omega \mid X(\omega) \leq x \text{ and } Y(\omega) \leq y\} \in \mathcal{A} \quad - (1)$$

### 1.1 Distributions of two discrete R.V

Ex: According to a new report from the research firm Technomic, 42% of millennials say they visit "upscale casual-dining restaurants" at least once a month. That's a higher percentage than Gen X (33%) and Baby Boomers (24%) who go to such restaurants once or more monthly."

Sol: We can find the populations of each category provided by US Census Bureau:



Age	Population
0-19	83,267,556
20-34 (Millennials)	62,649,947
35-49 (Gen X)	63,779,197
50-69 (Baby Boomers)	71,216,117
70+	27,832,721

Now,

Let  $X$  and  $Y$  be two random variables

- $X=1$  : a person between 20 and 69 visits upscale restaurants at least once a month and  $X=0$  otherwise.
- $Y=1$  : a person between 20 and 69 is millennial
- $Y=2$  : a person between 20 and 69 is a Gen X
- $Y=3$  : a person between 20 and 69 is a Baby Boomer



(23)

we can translate the statement of the article in the form of a contingency table shown below:

X/Y	1	2	3
0	36,337	42,732	47,003
1	26,313	21,049	24,213

Table: Count data ( $\times 1000$ )

The probability that the couple  $(X, Y)$  takes on a particular value can be found by dividing each cell by the total population of people between 20 and 69.

X/Y	1	2	3
0	0.184	0.216	0.238
1	0.133	0.106	0.123

Now, let us define formally the joint probability mass function of two discrete random variables  $X$  and  $Y$ .



(24)  
Prop 1.1 let  $X$  and  $Y$  be two discrete r.v.s on probability space  $(\Omega, \mathcal{A}, P)$  with joint pmf  $P_{XY}$ , then the following holds:

- $P_{XY}(n, y) \geq 0$ , for  $n \in X(\Omega)$  and  $y \in Y(\Omega)$

- $\sum_{n \in X(\Omega)} \sum_{y \in Y(\Omega)} P_{XY}(n, y) = 1$

Prop 1.2 let  $X$  and  $Y$  be two discrete r.v.s on probability space  $(\Omega, \mathcal{A}, P)$  with joint pmf  $P_{XY}$ . Then, for any subset  $S \subset X(\Omega) \times Y(\Omega)$

$$P((X, Y) \in S) = \sum_{(n, y) \in S} P_{XY}(n, y)$$

The above property tells us that in order to determine the probability of event  $\{(X, Y) \in S\}$ , you simply sum up the probabilities of the events  $\{X = n, Y = y\}$  with values  $(n, y)$  in  $S$ .



## Examples

(2)

Q1 Consider the following joint probability mass function:

$$P_{XY}(n, y) = \frac{xy^2}{13} \mathbb{1}_S(n, y)$$

$$\text{with } S = \{(1, 1), (1, 2), (2, 2)\}$$

(a) Show that  $P_{XY}$  is a valid joint probability mass function.

Sol<sup>n</sup> The joint probability mass function of  $X$  and  $Y$  is given by the following table:

$X/Y$	1	2
1	$1/13$	$4/13$
2	0	$8/13$

We first note that  $P_{XY}(n, y) \geq 0$  for all  $n, y = 1, 2$ . Second,

$$\sum_{n=1}^2 \sum_{y=1}^2 P_{XY}(n, y) = \frac{1}{13} + \frac{4}{13} + 0 + \frac{8}{13} = 1$$

Hence  $P_{XY}$  is indeed a valid joint probability function.



(b) What is  $P(X+Y \leq 3)$ ?

Sol<sup>n</sup> Let us denote  $B$  the set of values such that  $X+Y \leq 3$ :

$$B = \{(1,1), (1,2), (2,1)\}$$

$\therefore$

$$P(X+Y \leq 3) = \sum_{(x,y) \in B} P_{XY}(x,y)$$

$$= P_{XY}(1,1) + P_{XY}(1,2) + P_{XY}(2,1)$$

$$= \frac{1}{13} + \frac{4}{13} + 0$$

$$= \frac{5}{13}$$

(c) Give the marginal probability mass functions of  $X$  and  $Y$ .

Sol<sup>n</sup> The marginal probability mass func<sup>n</sup> of  $X$  is given by:

$$P_X(1) = \sum_{y=1}^2 P_{XY}(1,y) = P_{XY}(1,1) + P_{XY}(1,2) = \frac{1}{13} + \frac{4}{13} = \frac{5}{13}$$

$$P_X(2) = \sum_{y=1}^2 P_{XY}(2,y) = P_{XY}(2,1) + P_{XY}(2,2) = 0 + \frac{8}{13} = \frac{8}{13}$$



(29)

Similarly, the marginal probability mass func<sup>n</sup> of  $Y$  is given by:

$$P_Y(1) = \sum_{x=1}^2 P_{XY}(x,1) = P_{XY}(1,1) + P_{XY}(2,1) = \frac{1}{13} + 0 = \frac{1}{13}$$

$$P_Y(2) = \sum_{x=1}^2 P_{XY}(x,2) = P_{XY}(1,2) + P_{XY}(2,2) = \frac{4}{13} + \frac{8}{13} = \frac{12}{13}$$

(d) What are the expected values of  $X$  and  $Y$ ?

sol<sup>n</sup> The expected value of  $X$  is given by:

$$E[X] = \sum_{x=1}^2 \sum_{y=1}^2 x P_{XY}(x,y)$$

$$= \sum_{x=1}^2 x \left\{ \sum_{y=1}^2 P_{XY}(x,y) \right\}$$

$$= \sum_{x=1}^2 x P_X(x)$$

$$= 1 \cdot \frac{5}{13} + 2 \cdot \frac{8}{13}$$

$$= \frac{21}{13}$$

Similarly, the expected value of  $Y$  is:

$$E[Y] = \sum_{x=1}^2 \sum_{y=1}^2 y P_{XY}(x,y)$$



(28)

$$= \sum_{y=1}^2 y \left\{ \sum_{x=1}^2 p_{xy}(x, y) \right\}$$

$$= \sum_{x=1}^2 y p_X(y)$$

$$= 1 \cdot \frac{1}{13} + 2 \cdot \frac{12}{13}$$

$$= 25/13$$

(c) Are  $X$  and  $Y$  independent?

sol<sup>n</sup> It is easy to see that  $X$  and  $Y$  are not independent, since, for example,

$$P_X(2) P_Y(1) = \frac{8}{13} \cdot \frac{1}{13}$$

$$= \frac{8}{169} \neq 0$$

$$= p_{XY}(2, 1)$$



## # Baye's Theorem

In probability theory and statistics, Bayes's theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

Bayes's theorem is stated mathematically as the following equation:

$$P(A/B) = \frac{P(B/A) P(A)}{P(B)}$$

where A and B are events  
and

$$P(B) \neq 0$$

- $P(A/B)$  is a conditional probability: the likelihood of event A occurring given that B is true.
- $P(B/A)$  is also a conditional probability: the likelihood of event B occurring given that A is true.
- $P(A)$  and  $P(B)$  are the probabilities of observing A and B respectively, they are known as the



②  
marginal probability

Example

Q Suppose that a test for using a particular drug is 99% sensitive and 99% specific. That is the test will produce 99% true positive results for drugs users and 99% true negative result for non-drug users. Suppose that 0.5% of people are users of the drug. What is the probability that a randomly selected individual with a positive test is drug user?

Sol<sup>n</sup>  $P(\text{User} | +) = \frac{P(+ | \text{User}) P(\text{User})}{P(+)}$

$$= \frac{P(+ | \text{User}) P(\text{User})}{P(+ | \text{User}) P(\text{User}) + P(+ | \text{Non-user}) P(\text{Non-user})}$$

$$= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995}$$

$$\approx 33.2\%$$



(3)

Q Even if 100% of patients with pancreatic cancer have a certain symptom, when someone has the same symptom, it does not mean that this person has 100% chance of getting pancreatic cancer. Assume the incidence rate is 1/100000, while 1/10000 healthy individuals have the same symptoms worldwide, the probability of having pancreatic cancer is only 9% and the other 91% could be positives.

sol<sup>n</sup> Based on incidence rate, the following table represents the corresponding no. per 100,000 people.

Symptom	Cancer		Total
	NO	Yes	
No	99989	0	99989
Yes	10	1	11
Total	99999	1	100000

$$P(\text{Cancer} | +) = \frac{P(+ | \text{Cancer}) P(\text{Cancer})}{P(+)}$$

$$= \frac{P(+ | \text{Cancer}) P(\text{Cancer})}{P(+ | \text{Cancer}) P(\text{Cancer}) + P(+ | \text{Non-cancer}) P(\text{Non-cancer})}$$



(32)

$$= \frac{1 \times 0.00001}{1 \times 0.00001 + (10/9999) \times 0.99999}$$

$$= \frac{1}{11}$$

$$\approx 9.1\%$$

Q A company manufactures TVs at two diff. plants A and B. Plant 'A' produce 80% and B produces 20% of the total production. 85 out of 100 TVs produced at plant A meet the quality standards while 65 out of 100 TVs produced at plant B meet the quality standard. A T.V produced by company is selected at random and is not found to meeting the quality standard. Find the probability that selected T.V was manufactured by plant B?

sol<sup>n</sup> Let  $B_1$  ( $B_2$ ) be the event that plant A (B) produce a T.V that does not meet the quality standard.



$$P(B_1) = 1 - \frac{85}{100} = \frac{15}{100} = \frac{3}{20}$$

$$P(B_2) = 1 - \frac{65}{100} = \frac{35}{100} = \frac{7}{20}$$

Let  $A_1(A_2)$  be the event that selected T.V is produced by plant  $A(B)$

$$\Rightarrow P(A_1) = 0.8, P(A_2) = 0.2$$

Let 'A' be the event that selected T.V does meet the quality standards.

$$\Rightarrow P(A/A_1) = P(B_1) = \frac{3}{20}$$

and,  $P(A/A_2) = P(B_2) = \frac{7}{20}$

$$= P(A_1) \cdot P(A/A_1) + P(A_2) \cdot P(A/A_2)$$

$$= 0.8 \times \frac{3}{20} + (0.2) \frac{7}{20} = \frac{3.8}{20}$$

$$\Rightarrow P(A/A_2) = \frac{P(A_2) \cdot P(A/A_2)}{P(A)} = \frac{(0.2)(7/20)}{(3.8/20)}$$

$$= \frac{14}{38} = \frac{7}{19}$$



## # Conditional Density:

$$f_{X/A}(x|A) = \frac{d}{dx} F_X(x|A)$$

## Conditional Distribution:

$$\begin{aligned} F_{X/A}(x|A) &= \Pr\{X \leq x|A\} \\ &= \frac{\Pr\{(X \leq x) \cap A\}}{\Pr\{A\}} \end{aligned}$$

## \* Probability conditioned on continuous RV

$$\Pr\{A|\bar{X} = x\} = ?$$

$$\therefore \Pr(A|\bar{X} = x) \approx \lim_{\Delta x \rightarrow 0} \frac{\Pr\{A \cap (x \leq \bar{X} \leq x + \Delta x)\}}{\Pr\{x \leq \bar{X} \leq x + \Delta x\}}$$

$$= \lim_{\Delta x \rightarrow 0} \frac{\Pr\{(x \leq \bar{X} \leq x + \Delta x)|A\} \Pr\{A\}}{\Pr\{x \leq \bar{X} \leq x + \Delta x\}}$$

$$\text{Now, } \Pr\{(x \leq \bar{X} \leq x + \Delta x)|A\} = f(x|A)\Delta x$$

$$\therefore = \lim_{\Delta x \rightarrow 0} \frac{f_X(x|A)\Delta x \Pr\{A\}}{f_X(x)\Delta x}$$

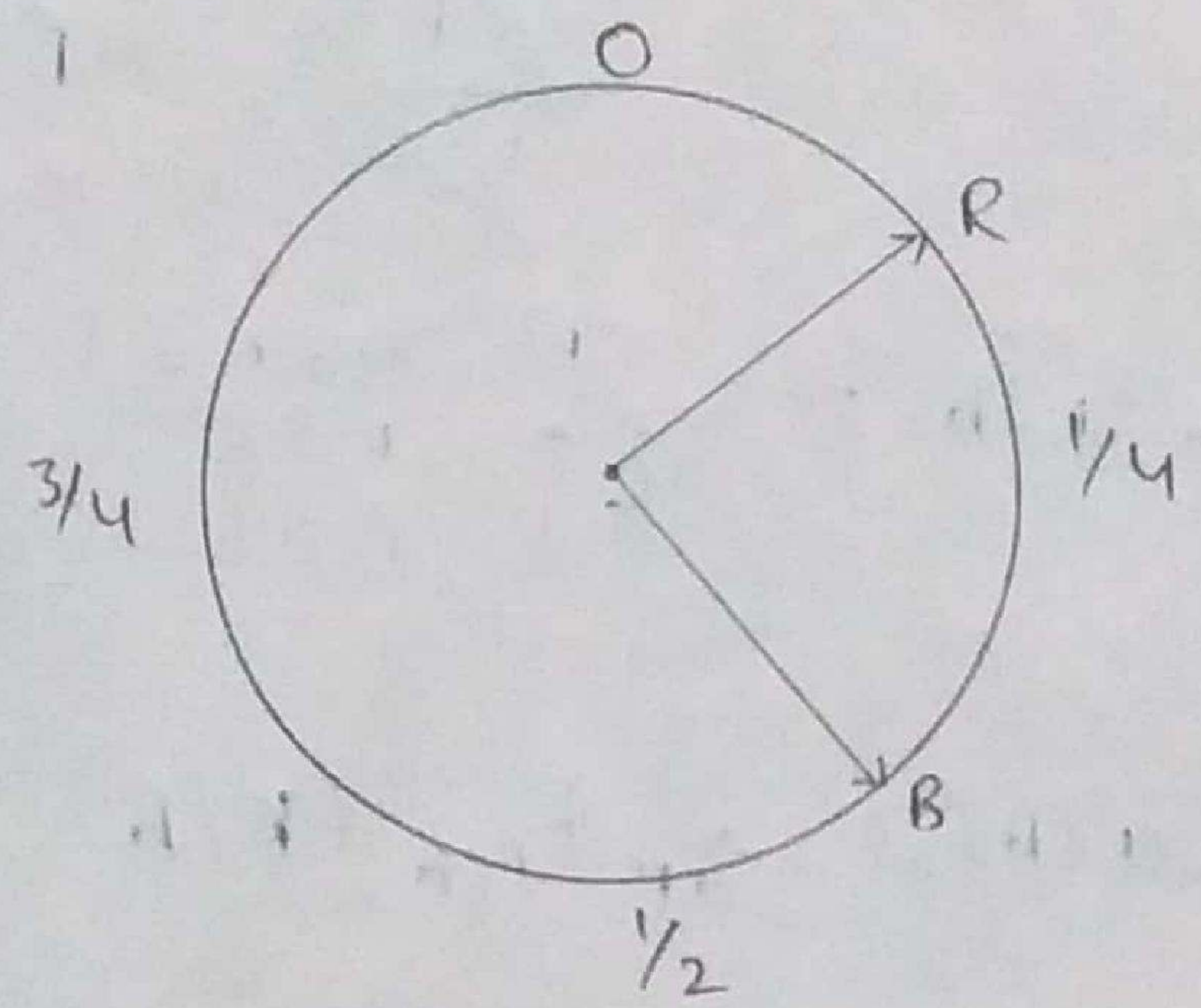


$$\Pr\{A | \bar{X} = x\} = \frac{f_{\bar{X}|A}(x|A) \Pr\{A\}}{f_{\bar{X}}(x)}$$

where,

$$\frac{f_{\bar{X}|A}(x|A)}{f_{\bar{X}}(x)} \rightarrow \text{Density Ratio}$$

Example



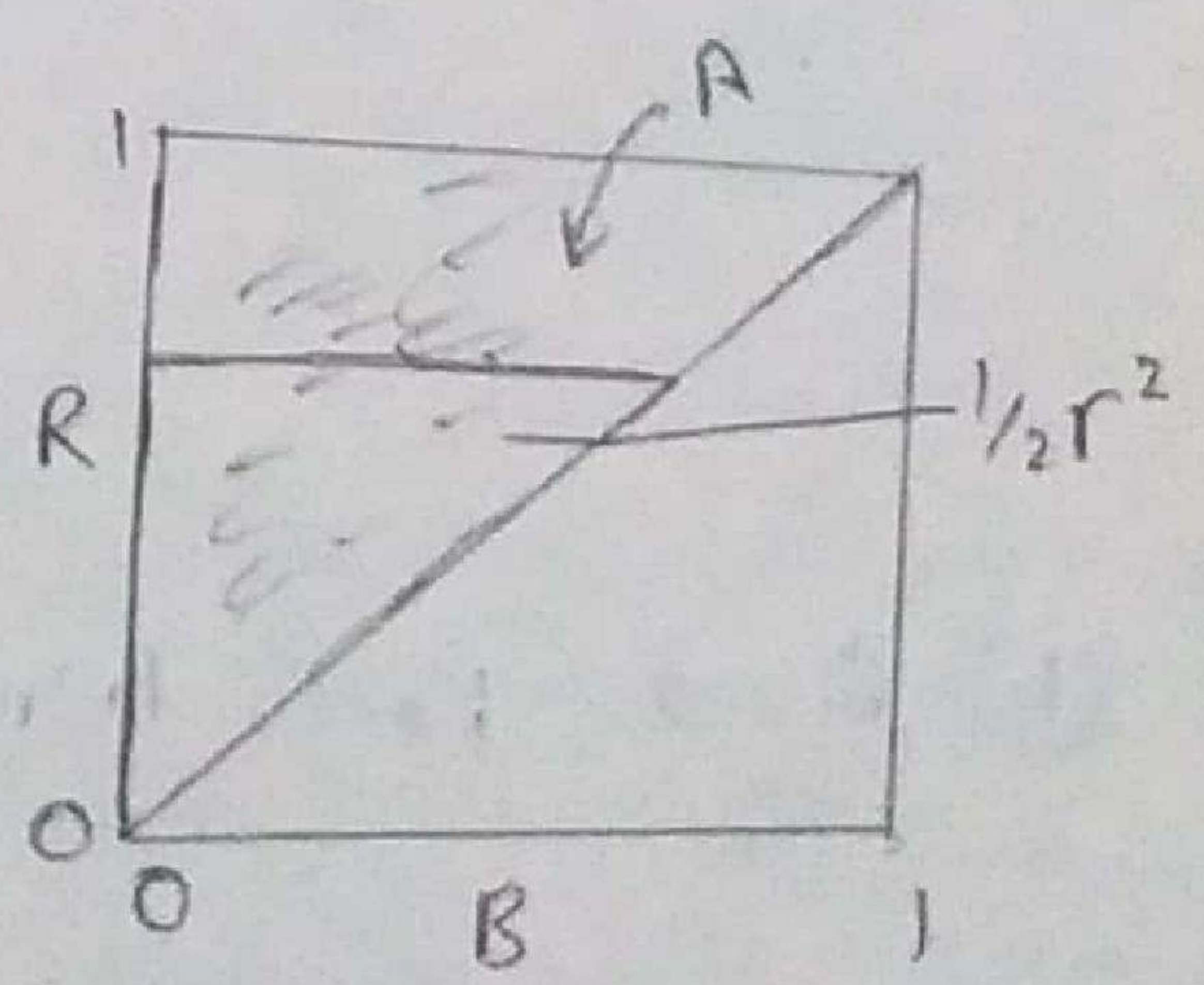
A = event  $R \geq B$

(a) Find  $F_{R|A}(r|A)$

A =  $R \geq B$

① Find  $F_{R|A}(r|A)$

②  $d/dx$ .





4/17

$$\textcircled{1} F_{R|A}(r/A) = \frac{\Pr\{(R \leq r) \cap A\}}{\Pr\{A\}}$$

$$= \frac{\Pr\{R \leq r \cap (R \geq B)\}}{\Pr\{R \geq B\}}$$

$$= \frac{\frac{1}{2}r^2}{\frac{1}{2}} = r^2$$

$$F_{R|A}(r/A) = \begin{cases} r^2 & 0 \leq r \leq 1 \\ 0 & r < 0 \\ 1 & r \geq 1 \end{cases}$$

$$\textcircled{2} f_{R|A}(r/A) = \frac{d}{dr} F_{R|A}(r/A)$$

$$= \begin{cases} 0 & r < 0 \\ 2r & 0 \leq r \leq 1 \\ 0 & r > 1 \end{cases}$$

(b) Find  $\Pr\{A/r\}$

$$\Pr\{A/r\} = \frac{F_{R|A}(r)}{F_R(r)} \cdot \Pr\{A\}$$



$$= \frac{2r}{1} \times \frac{1}{2} = r$$

