R V COLLEGE OF ENGINEERING
(An autonomous institution affiliated to VTU, Belgaum)
Department of Mathematics
MA231CT : Linear Algebra and Probability Theory
Unit-IV: Probability Distributions and Sampling Theory

Topic Learning Objectives:

- To apply the knowledge of the statistical analysis and theory of probability in the study of uncertainties.
- To use probability theory to solve random physical phenomena and implement appropriate distribution models.

Introduction:

In this unit, discrete probability distributions and continuous probability distributions are discussed. Discrete probability distribution is used when the sample space is discrete but not countable, whereas continuous probability distribution is used when the sample space is continuous or sample space is defined in a continuous interval.

In discrete distributions, the variables are distributed according to some definite probability law which can be expressed mathematically. The present study will also enable us to fit a mathematical' model or a function of the form $y = p(x)$ to the observed data. In discrete distributions Binomial distributions and Poisson distributions are discussed. In continuous distributions, Exponential distributions and Normal distributions are discussed.

Bernoulli distribution: A random variable X which takes two values 0 and 1, with probabilities q and p respectively, i.e ., P (X = 1) = p, P(X = 0) = q, q = 1 - p is called a Bernoulli variate and is said to have a Bernoulli distribution.

The probability of getting a head or a tail on tossing a coin is 1/2. If a coin is tossed thrice, the sample space S = {HHH, HHT, HTH, THH, TTH, THT, HTT, TTT}.

The probability of getting one head and two tails = 3/8. i.e., {HTT, TTH, THT}.

The probability of each one (one head, one tail, one tail) of these being (1/2) * (1/2) * (1/2) i.e., $(1/2)^3$, their total probability shall be 3 * $(1/2)^3$.

Similarly if a trial is repeated 'n' times and if 'p' is the probability of a success and 'q' that of a failure, then the probability of 'r' successes and 'n – r' failures is given by '$p^r q^{n-r}$'. But these 'r' successes and 'n – r' failures can occur in any of n_{C_r} ways in each of which the probability is same. Thus the probability of 'r' successes is $n_{C_r} p^r q^{n-r}$. The probability of at least 'r' successes in 'n' trials = Sum of probabilities of 'r, r + 1, ..., n' successes.

$$= n_{C_r} p^r q^{n-r} + n_{C_{r+1}} p^{r+1} q^{n-r-1} + \cdots + n_{C_n} p^n .$$

Binomial distribution: Binomial distribution was discovered by James Bernoulli (1654-1705) in the year 1700. It is concerned with trials of a repetitive nature in which only the occurrence or non-occurrence, success or failure, acceptance or rejection, yes or no of a particular event is of interest.

If a series of independent trials are performed such that for each trial 'p' is the probability of success and 'q' that of a failure, then the probability of 'r' successes in a series of 'n' trials is given by $n_{C_r} p^r q^{n-r}$, where 'r' takes any integral value from 0 to n. The probabilities of 0, 1, 2, ..., r, ..., n successes are, therefore, given by

$$q^n, n_{C_1} pq^{n-1}, n_{C_2} p^2 q^{n-2}, \dots, n_{C_r} p^r q^{n-r}, \dots, p^n$$

The probability of the number of successes so obtained is called the **Binomial distribution**.

The sum of the probabilities $= q^n + n_{C_1} pq^{n-1} + n_{C_2} p^2 q^{n-2} + \cdots + p^n = (q+p)^n = 1$.

The most obvious application deals with the testing of items as they come off an assembly line, where each trial may indicate a defective or a non defective item. One may choose to define either outcome as a success. The process is referred to as a **Bernoulli process**. Each trial is called a **Bernoulli trial**. The number *X* of successes in *n* Bernoulli trials is called a **binomial random variable**. The probability distribution of this discrete random variable is called the **binomial distribution**, and its values will be denoted by *b* (*x*; *n*, *p*) since they depend on the number of trials and the probability of a success 'p' on a given trial.

To prove that for a binomial distribution $\sum_{x=0}^{n} p(x) = 1$

Proof:

$$\begin{aligned}\sum_{x=0}^{n} p(x) &= \sum_{x=0}^{n} n_{C_x} p^x q^{n-x} = n_{C_0} p^0 q^{n-0} + n_{C_1} p^1 q^{n-1} + \cdots + n_{C_n} p^n q^{n-n} \\ &= q^n + n_{C_1} p^1 q^{n-1} + n_{C_2} p^2 q^{n-2} + \cdots + p^n \\ &= (p+q)^n = 1\end{aligned}$$

Note:

In a binomial distribution

1. n, the number of trials is finite.
2. each trial has two possible outcomes called success & failure.
3. all the trials are independent
4. p & q are constants for all the trials.

Mean and Variance of a Binomial distribution:

Mean $= \boldsymbol{\mu = np}$

Variance $= \quad \boldsymbol{\sigma^2 = npq}$

Standard deviation = $\sigma = \sqrt{npq}$

Problems:

1. The mean and variance of a binomial variate are respectively 16 & 8. Find (i) $P(X = 0)$ (ii) $P(X \geq 2)$

   Solution: Given, Mean = $\mu = np = 16$ and Variance = $\sigma^2 = npq = 8$

   $\frac{npq}{np} = \frac{8}{16} = \frac{1}{2}$

   i.e., $q = \frac{1}{2}$. Therefore, $p = 1 - q = \frac{1}{2}$. Also, $np = 16$ $i.e., n = 32$.

   (i) $P(X = 0) = n_{C_0} p^0 q^{n-0} = \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{32} = \left(\frac{1}{2}\right)^{32}$

   (ii) $P(X \geq 2) = 1 - P(X < 2) = 1 - P(X = 0) - P(X = 1)$

   $= 1 - \left(\frac{1}{2}\right)^{32} - 32\left(\frac{1}{2}\right)^{32} = 1 - 33\left(\frac{1}{2}\right)^{32}$.

2. Six dice are thrown 729 times. How many times do you expect at least 3 dice to shown a 5 or 6?

   Solution: Here n = 6, N = 729

   $P(x \geq 3) = 6_{C_x} p^x q^{6-x}$

   Let p be the probability of getting 5 or 6 with 1 dice

   i.e., p = 2/6 = 1/3. Thus, q = 1 – 1/3 = 2/3

   $p(x \geq 3) = p(x = 3, 4, 5, 6)$

   $= p(x = 3) + p(x = 4) + p(x = 5) + p(x = 6)$

   = 0 .3196

   Therefore, number of times = 729*0.3196 = 233

3. A basket contains 20 good oranges and 80 bad oranges. 3 oranges are drawn at random from this basket. Find the probability that out of 3 (i) exactly 2 (ii) at least 2 (iii) at most 2 are good oranges.

   Solution: Let p be the probability of getting a good orange i.e., $p = \frac{80C_1}{100C_1}$

   p = 0.8 and q = 1- 0.8 = 0.2

   (i) p (x = 2) = $3C_2(0.8)^2(0.2)^1 = 0.384$

   (ii) p (x ≥ 2) = p(2) + p(3) = 0.896

   (iii) p (x ≤ 2) = p(0) + p(1) + p(2) = 0.488

4. In a sampling a large number of parts manufactured by a machine, the mean number of defective in a sample of 20 is 2. Out of 1000 such samples how many would expected to contain at least 3 defective parts.

   Solution: Given; n = 20, n p = 2

i.e., p = 1/10 and q = 1-p = 9/10

p ($x \geq 3$) = 1 – p ($x < 3$)

= 1 – p (x = 0,1,2) = 0.323

Number of samples having at least 3 defective parts = 0.323 * 1000 = 323

5. The probability that a patient recovers from a rare blood disease is 0.4. If 15 people are known to have contracted this disease, what is the probability that
   (i) at least 10 survive, (ii) from 3 to 8 survive.

   Solution: Let X be the number of people who survive.

   (i) $p(X \geq 10) = 1 - p(X < 10) = 1 - \sum_{x=0}^{9} b(x; 15,0.4) = 1 - 0.9662 = 0.0338$

   (ii) $p(3 \leq X \leq 8) = \sum_{x=3}^{8} b(x; 15,0.4) = \sum_{x=0}^{8} b(x; 15,0.4) - \sum_{x=0}^{2} b(x; 15,0.4)$

   = 0.9050 - 0.0271 = 0.8779

Exercise:

1. Ten coins are thrown simultaneously. Find the probability of getting at least seven heads.
2. In 256 sets of 12 tosses of a coin, in how many cases one can expect 8 heads and 4 tails.
3. Let ‘x’ be a binomial variate with mean 6 and variance 4. Find the distribution of ‘x’,
4. In a binomial distribution consisting of 5 independent trials, probability of 1 & 2 successes are 0.4096 & 0.2048 respectively. Find the parameter ‘p’ of the distributive function.
5. A and B play a game in which their chances of winning are in the ratio 3:2. Find A’s chance of winning at least three games out of five games played.

Answers: 1. $1.\frac{176}{1024}$ $2.31(\text{approx})$ $3.18_{C_x}\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{18-x}$ $4.p = 0.2$ $5.0.68$

Poisson distribution:

Poisson distribution was discovered by the French mathematician and physicist Simeon Denis Poisson (1781-1840) who published it in 1837. Poisson distribution is a limiting case of the binomial distribution under the following conditions:

i) n, the number of trials is indefinitely large, *i.e.*, $n \to \infty$.

ii) p, the constant probability of success for each trial is indefinitely small, *i.e.*, $p \to 0$.

iii) $np = \lambda$, (say), is finite. Thus $p = \lambda$ /n, $q = 1 - \lambda /n$, where λ is a positive real number.

The probability function of the Poisson distribution is given by

$\boldsymbol{p(x,\lambda) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}}$ where $\boldsymbol{\lambda}$ is known as the parameter of poisson distribution.

Definition: A random variable X is said to follow a Poisson distribution if it assumes only non-negative values and its probability mass function is given by

$$p(\mathrm{x},\lambda) = P(X = \mathrm{x}) = \frac{e^{-\lambda}\lambda^{\mathrm{x}}}{\mathrm{x}!}\ ;\ \ x = 0, 1, 2, \ldots;\ \lambda > 0$$

$= 0$, otherwise

Remarks:

1. It should be noted that $\sum_{x=0}^{\infty} P(X = \mathrm{x}) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^{\mathrm{x}}}{\mathrm{x}!} = e^{-\lambda} e^{\lambda} = 1$
2. The corresponding distribution function is:

   $F(x) = P(X \leq \mathrm{x}) = \sum_{r=0}^{x} P(r) = e^{-\lambda} \sum_{r=0}^{\mathrm{x}} \frac{\lambda^{\mathrm{r}}}{\mathrm{r}!}; x = 0, 1, 2, \ldots$
3. Poisson distribution occurs when there are events which do not occur as outcomes of a definite number of trials (unlike that in binomial) of an experiment but which occur at random points of time and space wherein our interest lies only in the number of occurrences of the event, not in its non-occurrences.
4. Following are some instances where Poisson distribution may be successfully employed.
   i) Number of deaths from a disease (not in the form of an epidemic) such as heart attack or cancer or due to snake bite.
   ii) Number of suicides reported in a particular city.
   iii) The number of defective material in a packing manufactured by a good concern.
   iv) Number of faulty blades in a packet of 100.
   v) Number of air accidents in some unit of time.
   vi) Number of printing mistakes at each page of the book.
   vii) Number of telephone calls received at a particular telephone exchange in some unit of time or connections to wrong numbers in a telephone exchange.
   viii) Number of cars passing a crossing per minute during the busy hours of a day.
   ix) The number of fragments received by a surface area *'t'* from a fragment atom bomb.
   x) The emission of radioactive (alpha) particles.

Mean and Variance of a Poisson distribution

Mean $= \boldsymbol{\mu} = \boldsymbol{\lambda}$

Variance $= \boldsymbol{\sigma^2} = \boldsymbol{\lambda}$

Standard deviation $= \boldsymbol{\sigma} = \sqrt{\boldsymbol{\lambda}}$

Problems

1. It is known that the chance of an error in the transmission of a message through a communication channel is 0.002. 1000 messages are sent through the channel; find the probability that at least 3 messages will be received incorrectly.

   Solution: Here, the random experiment consists of finding an error in the transmission of a message. It is given that n = 1000 messages are sent, a very large number, if p denote the probability of error in the transmission, p = 0.002, relatively a small number, therefore, this problem may be viewed as Poisson oriented. Thus, average number of messages with an error is $\lambda = np = 2$.

Therefore, required probability function is

$$\mathrm{p}(\lambda, x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0,1,2,3,\ldots,\infty.$$

$$p(2, x) = \frac{e^{-2}2^{x}}{x!}, x = 0,1,2,3,\ldots,\infty.$$

Here, the problem is about finding the probability of the event, namely,

$$P(X \geq 3) = 1 - P(X < 3) = 1 - \{P[X = 0] + P[X = 1] + P[X = 2]\}$$

$$= 1 - \left[\sum_{x=0}^{2} \frac{e^{-2}2^{x}}{x!}\right]$$

$$= 1 - e^{-2}(1 + 2 + 2) = 1 - 5e^{-2}$$

2. A Car-hire firm has two cars it hires out daily. The number of demands for a car on each day is distributed as Poisson variate with mean 1.5. Obtain the proportion of days on which i) there was no demand ii) demand is refused.

Solution: Here λ = 1.5

i) $\mathrm{p}(x, 0) = \frac{e^{-1.5}(1.5)^{0}}{0!} = 0.2231$

ii) $\mathrm{p}(x > 2) = 1 - p(x \leq 2) = 1 - p(x = 0, 1, 2)$

$$= 1 - \mathrm{p}(x = 0) - p(x = 1) - p(x = 2)$$

$$= 1 - \frac{e^{-1.5}(1.5)^{0}}{0!} - \frac{e^{-1.5}(1.5)^{1}}{1!} - \frac{e^{-1.5}(1.5)^{2}}{2!}$$

$$= 0.1913$$

3. Assuming that the probability of an individual being killed in a mine accident during a year is 1/2400. Use Poisson distribution to calculate the probability that in a mine employing 200 miners there will be at least one fatal accident in a year?

Solution: Here p = 1/2400, n = 200, λ= np = 0.083

$$\mathrm{p}(x \geq 1) = 1 - p(x < 1) = 1 - p(x = 0)$$

$$= 1 - \mathrm{e}^{-0.083} = 0.0796$$

4. In a Poisson distribution if $\mathrm{P}(2) = \frac{2}{3}\mathrm{P}(1)$, find $\mathrm{P}(0)$. Find also its mean and standard deviation.

Solution: $p(\lambda, x) = \frac{e^{-\lambda}\lambda^x}{x!}$

Given $P(2) = \frac{2}{3}P(1)$

i.e.,$p(\lambda, 2) = \frac{2}{3}p(\lambda, 1)$

$$\frac{e^{-\lambda}\lambda^2}{2!} = \frac{2}{3}\frac{e^{-\lambda}\lambda}{1!}$$

$$\lambda = \frac{4}{3}$$

Thus, $p(\lambda, 0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda} = e^{-4/3} = 0.2636$

Mean $= \mu = \lambda = \frac{4}{3}$ and standard deviation $= \sigma = \sqrt{\lambda} = \sqrt{\frac{4}{3}}$

5. The incidence of occupational disease in an industry is such that the workmen have a 10% chance of suffering from it. What is the probability that in a group of seven, 5 or more will suffer from it.

   Solution: p = 10% = 0.1, n = 7

   $\mu = np = 0.1 * 7 = 0.7$

   $P(x \geq 5) = P(5) + P(6) + P(7)$

   $$= \frac{e^{-\lambda}\lambda^5}{5!} + \frac{e^{-\lambda}\lambda^6}{6!} + \frac{e^{-\lambda}\lambda^7}{7!}$$

   $$= \frac{e^{-0.7}(0.7)^5}{5!} + \frac{e^{-0.7}(0.7)^6}{6!} + \frac{e^{-0.7}(0.7)^7}{7!} = 0.0008$$

Exercise:

1. For a Poisson variable $3P(2) = P(4)$, find standard deviation.
2. If the probability of a bad reaction from a certain injection is 0.001, determine the chance that out of 2000 individuals more than two will get a bad reaction.
3. Fir a Poisson distribution to the set of observations given below.

x	0	1	2	3	4
f(x)	122	60	15	2	1

4. In a certain factory turning out razor blades there is a small chance of 0.002 for any blade to be defective. The blades are supplied in packets of 10. Use Poisson distribution to calculate the approximate number of packets containing no defective, one defective two blades defective respectively ina consignment of 10,000 packets.
5. A manufacturer of cotter pins knows that 5% of his product is defective. If he sells cotter pins in boxes of 100 and guarantees that not more than 10 pins will be defective, what is the approximate probability that a box will fail to meet the guaranteed quality?

Answers: 1. 2.45 2. 0.32 3. $f(x) = \frac{e^{-0.5}(0.5)^x}{x!}$, for N = 200, it is $N * f(x)$.

4. 9802, 196, 2 5. $1 - e^{-5} \sum_{x=0}^{10} \frac{(5)^x}{x!}$

Exponential distribution

Many experiments involve the measurement of time X between an initial point of time and the occurrence of some phenomenon of interest. Exponential distribution deals with such type of continuous random variable X.

A continuous random variable X assuming non-negative values is said to have an exponential distribution with parameter $\lambda > 0$, if its probability density function is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Examples such as time between two successive job arrivals, duration of telephone calls, life time of a component or a product, server time at a server in a queue can be taken under Exponential distribution.

Mean and variance of Exponential distribution

$$\textbf{Mean} = \boldsymbol{\mu} = \int_0^\infty x\, f(x)dx$$

$$= \int_0^\infty x\, \lambda e^{-\lambda x} dx$$

$$= \lambda \int_0^\infty x\, e^{-\lambda x} dx$$

$$= \lambda \left[x * \frac{e^{-\lambda x}}{(-\lambda)} - 1 * \frac{e^{-\lambda x}}{(-\lambda)^2} \right]_{x=0 \text{ to } \infty}$$

$$\boldsymbol{\mu} = \frac{\mathbf{1}}{\boldsymbol{\lambda}}$$

$$\textbf{Variance} = \boldsymbol{\sigma}^2 = \int_0^\infty x^2\, f(x)dx - \mu^2$$

$$= \int_0^\infty x^2\, \lambda e^{-\lambda x} dx - \left(\frac{1}{\lambda}\right)^2$$

$$= \lambda \left[x^2 * \frac{e^{-\lambda x}}{(-\lambda)} - 2x * \frac{e^{-\lambda x}}{(-\lambda)^2} + 2 * \frac{e^{-\lambda x}}{(-\lambda)^3} \right]_{x=0 \text{ to } \infty} - \left(\frac{1}{\lambda}\right)^2$$

$$\boldsymbol{\sigma}^2 = \left(\frac{\mathbf{1}}{\boldsymbol{\lambda}}\right)^{\mathbf{2}}$$

Standard deviation = $\boldsymbol{\sigma} = \frac{\mathbf{1}}{\boldsymbol{\lambda}}$

Problems:

1. Let the mileage (in thousands of miles) of a particular tyre be a random variable X having the probability density $(x) = \begin{cases} \frac{1}{20} e^{\frac{-x}{20}} & , x > 0 \\ 0, x \leq 0 \end{cases}$.

Find the probability that one of these tyres will last (i) at most 10,000 miles (ii) anywhere between 16,000 to 24,000 miles (iii) at least 30,000 miles. Also, find The mean and variance of the given probability density function.

Solution: (i) $\mathrm{P}(x \leq 10) = \int_0^{10} f(x)dx = \int_0^{10} \frac{1}{20} \mathrm{e}^{\frac{-x}{20}} dx = \frac{1}{20}\int_0^{10} \mathrm{e}^{\frac{-x}{20}} dx = 0.3934$

(ii) $\mathrm{P}(16 \leq x \leq 24) = \int_{16}^{24} f(x)dx = \int_{16}^{24} \frac{1}{20} \mathrm{e}^{\frac{-x}{20}} dx = \frac{1}{20}\int_{16}^{24} \mathrm{e}^{\frac{-x}{20}} dx = \ 0.148$

(iii) $\mathrm{P}(x \geq 30) = \int_{30}^{\infty} f(x)dx = \int_{30}^{\infty} \frac{1}{20} \mathrm{e}^{\frac{-x}{20}} dx = \frac{1}{20}\int_{30}^{\infty} \mathrm{e}^{\frac{-x}{20}} dx = \ 0.223$

Mean $= \mu = \frac{1}{\lambda} = \frac{1}{\frac{1}{20}} = 20$ and Variance $= \sigma^2 = \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\left(\frac{1}{20}\right)^2} = 400.$

2. The length of time for one person to be served at a cafeteria is a random variable X having an exponential distribution with a mean of 4 minutes. Find the probability that a person is served in less than 3 minutes on at least 4 of the next 6 days.

Solution: Given, Mean = 4. i.e., Mean $= 4 = \frac{1}{\lambda} \to \lambda = \frac{1}{4}$

The probability density function is $f(x) = \lambda \mathrm{e}^{-\lambda x} = \frac{1}{4}\mathrm{e}^{\frac{-x}{4}}$

$$\mathrm{P}(x < 3) = 1 - \mathrm{P}(x \geq 3) = 1 - \int_3^{\infty} f(x)dx$$

$$= 1 - \int_3^{\infty} \frac{1}{4} \mathrm{e}^{\frac{-x}{4}} dx = 1 - \mathrm{e}^{\frac{-3}{4}} = 0.9875$$

Let D represents the number of days on which a person is served in less than 3 minutes. Then using the binomial distribution, the probability that a person is served in less than 3 minutes on at least 4 of next 6 days is;

$\mathrm{P}(\mathrm{D} \geq 4) = \mathrm{P}(\mathrm{D} = 4) + \mathrm{P}(\mathrm{D} = 5) + \mathrm{P}(\mathrm{D} = 6)$

$= 6_{C_4}\left(1 - e^{-\frac{3}{4}}\right)^4 (e^{-\frac{3}{4}})^2 + 6_{C_5}(1 - e^{-3/4})^5 (e^{-\frac{3}{4}})^1 + 6_{C_6}(1 - e^{-3/4})^6 (e^{-\frac{3}{4}})^0$

$= 0.3968$

3. The increase in sales per day in a shop is exponentially distributed with Rs 800 as the average. If sales tax is paid at the rate of 6%, find the probability that increase in sales tax return from that shop will exceed Rs 30 per day.

Solution: Given, Mean = 800

i.e., Mean $= 800 = \frac{1}{\lambda} \to \lambda = \frac{1}{800}$

The probability density function is $f(x) = \lambda \mathrm{e}^{-\lambda x} = \frac{1}{800}\mathrm{e}^{\frac{-x}{800}}$

Let X denotes the sales per day. Total sales tax on X items = $\frac{6}{100}$X

Given total sales tax exceeds Rs 30 per day. i.e., $\frac{6}{100}\mathrm{X} > 30.\ \mathrm{i.e.,}\ \mathrm{X} > 500$

Probability of sales tax exceeding Rs 30 = Probability of sales per day exceeding 500

$$= \mathrm{P}(\mathrm{X} > 500) = 1 - \mathrm{P}(\mathrm{X} \leq 500)$$

$$= 1 - \int_0^{500} f(x)dx$$

$$= 1 - \int_0^{500} \frac{1}{800} \mathrm{e}^{\frac{-x}{800}} dx = 0.5353$$

4. After the appointment of a new sales manager the sales in a 2 wheeler showroom is exponentially distributed with mean 4. If 2 days are selected at random what is the probability that (i) on both days, the sales is over 5 units (ii) the sales is over 5 times at least 1 of 2 days.

Solution: Given, Mean = 4. i.e., Mean = $4 = \frac{1}{\lambda} \rightarrow \lambda = \frac{1}{4}$

The probability density function is $f(x) = \lambda \mathrm{e}^{-\lambda x} = \frac{1}{4}\mathrm{e}^{\frac{-x}{4}}$

Let X represents the sales per day

$P(x > 5) = \int_5^{\infty} f(x)dx = \int_5^{\infty} \frac{1}{4} \mathrm{e}^{\frac{-x}{4}} dx = \frac{1}{4}\int_5^{\infty} \mathrm{e}^{\frac{-x}{4}} dx = 0.2865$

Let D = number of days on which sales is over 5 units

(i) $P(D = 2) = n_{C_x} p^n q^{n-x} = 2_{C_2} (e^{-\frac{5}{4}})^2 \left(1 - e^{-\frac{5}{4}}\right)^{2-2} = 0.082$

(ii) $P(D = \text{at least 1 of 2 days}) = P(D = 1) + P(D = 2)$

$2_{C_1} (e^{-\frac{5}{4}})^1 \left(1 - e^{-\frac{5}{4}}\right)^{2-1} + 2_{C_2} (e^{-\frac{5}{4}})^2 \left(1 - e^{-\frac{5}{4}}\right)^{2-2} = 0.4908.$

Exercise:

1. The sales per day in a shop are exponentially distributed with average sale amounting to Rs 100 and net profit is 8%. Find the probability that net profit exceed Rs 30 on 2 consecutive days.
2. Let X and Y have common p.d.f $\alpha \mathrm{e}^{-\alpha x}, 0 < x < \infty, \alpha > 0$. Find the p.d.f of

   (i) 3 + 2X (ii) X – Y.
3. If X has exponential distribution with mean 2, find $P(X < 1|X < 2)$.
4. The life (in years) of a certain electrical switch has an exponential distribution with an average life of 2 years. If 100 of these switches are installed in different systems, find the probability that at most 30 fail during the first year.

Answers: 1. $(e^{-3.75})^2$ 2. $\frac{\alpha}{2}\exp\left(-\frac{\alpha(x-3)}{2}\right), x > 3$, $\frac{\alpha}{2}\exp(-\alpha|x|)$, $\forall\, x$

3. $\frac{(1-\mathrm{e}^{-\lambda})}{(1-\mathrm{e}^{-2\lambda})}$, where $\lambda = \frac{1}{2}$. 4. $P(X \le 30) = \sum_{x=0}^{30} 100_{C_x} (0.606)^x (0.394)^{100-x}$

Normal distribution

The normal distribution was first discovered in 1733 by English mathematician De-Moivre, who obtained this continuous distribution as a limiting case of the binomial distribution and applied it to problems arising in the game of chance.

Among all the distribution of a continuous random variable, the most popular and widely used one is normal distribution function. Most of the work in correlation and regression analysis, testing of hypothesis, has been done based on the assumption that problem follows a normal distribution function or just everything normal. Also, this distribution is extremely important in statistical applications because of the central limit theorem, which states that "under very general assumptions, the mean of a sample of n mutually Independent random variables (having finite mean and variance) are normally distributed in the limit $n \to \infty$". It has been observed that errors of measurement often possess this distribution.

Definition: A random variable X is said to have a normal distribution with parameters μ (called "mean") and σ^2 (called "variance") if its density function is given by the probability law:

$$f(x;\ \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2}\left\{\frac{x-\mu}{\sigma}\right\}^2\right] \text{ for } -\infty < x < \infty, -\infty < \mu < \infty \text{ and } 0 < \sigma < \infty\,.$$

Examples such as marks scored by students and life span of a product can be included under normal distribution.

Remarks:

1. A random variable X with mean μ and variance σ^2 and following the normal law is expressed by $X\sim N(\mu, \sigma^2)$.
2. If $X\sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma}$, is a standard normal variate with E(Z) = 0 and Var (Z) = 1 and we write $Z \sim N(0, 1)$.
3. The p.d.f of standard normal variate Z is given by $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, -\infty < z < \infty$

   and the corresponding distribution function, denoted by F(z)is given by

   $F(z) = P(Z \leq z) = \int_{-\infty}^{z} \varphi(u)du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\, du.$

   Also, $F(-z) = \int_{-\infty}^{-z} \varphi(u)du = 1 - F(z).$
4. The graph of f(x) is a famous 'bell-shaped' curve. The top of the bell is directly above the mean μ. For large values of σ, the curve tends to flatten out and for small values of σ, it has a sharp peak.

Note: The limiting form of the binomial distribution for large values of n with neither p nor q is very small, is the normal distribution.

Properties of Normal Distribution:

1. All normal curves are bell-shaped.
2. All normal curves are symmetric about the mean μ.
3. The area under an entire normal curve is 1.

4. All normal curves are positive for all x. i.e., $f(x) > 0$ for all x.
5. The shape of any normal curve depends on its mean and the standard deviation.

The probabilities are computed numerically and recorded in a special table called the normal distribution table (the probabilities can also be computed using a standard calculator). Use the following results for the calculation of probabilities.

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

(ii) $P(a < X < b) = F(b) - F(a)$

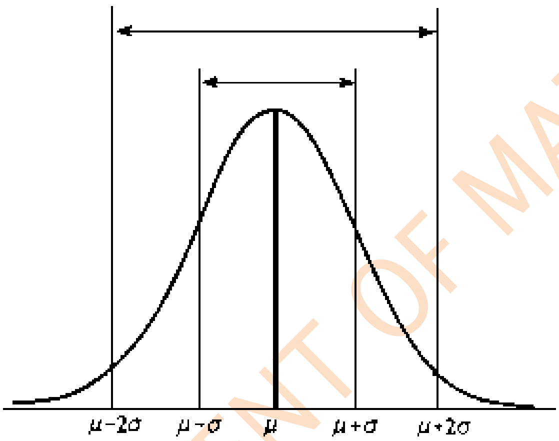
(iii) $P(a < X) = 1 - P(X \leq a) = 1 - F(a)$

(iv) $F(-b) = 1 - F(b)$, where b is positive.

The distributions of some variables including aptitude-tests scores, heights of women/men, have roughly the shape of a normal curve (bell shaped curve)

Normally Distributed Variable

A variable is said to be **normally distributed** or to have a **normal distribution** if its distribution has the shape of a normal curve.



Problems:

2. A sample of 100 battery cells is tested to find the length of life, gave the following results. Mean = 12 hrs. Standard Deviation = 3 hrs. Assuming the data to be normally distributed what % of battery cells are expected to have life (i) more than 15 hrs. (ii) less than 6hrs. (iii) between 10 & 14 hrs .

Solution: (i) when x = 15 for given mean = 12 hrs and standard deviation = 3 hrs;

$$P(x > 15) = P\left(\frac{X-\mu}{\sigma} > \frac{15-\mu}{\sigma}\right) = P\left(\frac{X-\mu}{\sigma} > \frac{15-12}{3}\right)$$

$$= P(z > 1)$$

$$= 0.5 - 0.3413$$

$$= 0.1587 = 16\%$$

(ii) When x = 6

$$P(x < 6) = P\left(\frac{X-\mu}{\sigma} < \frac{6-\mu}{\sigma}\right) = P\left(\frac{X-\mu}{\sigma} < \frac{6-12}{3}\right)$$

$= P(z < -2)$

= 0.5 - 0.4772

= 0.0228 = 2.28%

(iii) $P(10 < x < 14) = P\left(\frac{10-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{6-\mu}{\sigma}\right)$

$= P(-0.6667 < z < 0.6667)$

$= 2 * P(0 < z < 0.6667)$

= 2*0.2485 = 0.497 = 50%

3. Find the mean and standard deviation of an examination in which grades 70 and 88 corresponds to standard scores of -0.6 and 1.4 respectively.

Solution: Standard variable $z = \frac{X-\mu}{\sigma}$

Here, $-0.6 = \frac{70-\mu}{\sigma}$ gives $\mu - 0.6\sigma = 70$

and $1.4 = \frac{88-\mu}{\sigma}$ gives $\mu + 1.4\sigma = 88$

by solving the above equations, $\mu = 75.4$ and $\sigma = 9$.

4. The marks X obtained in mathematics by 1000 students is normally distributed with mean 78% and standard deviation 11%. Determine how many students got marks above 90%.

Solution: Here, mean = 78% = 0.78 and standard deviation = 11% = 0.11.

Thus, $z = \frac{X-\mu}{\sigma} = \frac{X-0.78}{0.11}$

For X = 0.9, write $z = \frac{0.9-0.78}{0.11} = 1.09$

$P(x > 0.9) = 1 - P(X \le 0.9) = 1 - P(z \le 1.09)$

$= 1 - 0.86214 = 0.13786$

5. X is a normal variate with mean 30 and standard deviation 5. Find the probabilities that (i) $26 \le X \le 40$ (ii) $X \ge 45$ (iii) $|X - 30| > 5$.

Solution: Given, mean = 30 and standard deviation = 5

Thus, $z = \frac{X-\mu}{\sigma} = \frac{X-30}{5}$

(i) For X = 26, $z = \frac{26-30}{5} = -0.8$ and

For X = 40, $z = \frac{40-30}{5} = 2$

Therefore, $P(26 \le X \le 40) = P(-0.8 \le z \le 2)$

$$= F(2) - F(-0.8) = 0.97725 - 0.21186$$

$$= 0.76539$$

(ii) For X = 45, $z = \frac{45-30}{5} = 3$

$P(X \ge 45) = 1 - P(X \le 45) = 1 - P(z \le 3)$

$$= 1 - F(3) = 1 - 0.99865 = 0.00135$$

(iii) $P(|X - 30| > 5) = 1 - P(|X - 30| \le 5)$

$$= 1 - P(-5 \le X - 30 \le 5)$$

$$= 1 - P(25 \le X \le 35)$$

$$= 1 - P(-1 \le z \le 1)$$

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

= 1- (0.84134 - 0.15866) = 0.31732

Exercise:

1. In a test of 2000 electric bulbs, it was found that the life of a particular make was normally distributed with an average life of 2040 hours and standard deviation of 60 hours. Estimate the number of bulbs likely to burn for (i) more than 2150 hours (ii) less than 1950 hours (iii) more than 1920 hours but less than 2060 hours.
2. Assume that the reduction of a person's oxygen consumption during a period of transcendental meditation (T M) is a continuous random variable X normally distributed with mean 37.6 cc/min and standard deviation 4.6 cc/min. Determine the probability that during a period of T M a person's oxygen consumption will be reduced by (i) at least 44.5 cc/min (ii) at most 35 cc/min (iii) anywhere from 30 cc/min to 40 cc/min/
3. An analog signal received at a detector (measured in micro volts) may be modeled as a Gaussian random variable N (200, 256) at a fixed point in time. What is the probability that the signal will exceed 240 micro volts? What is the probability that the signal is larger than 240 micro volts, given that it is larger than 210 micro volts.

4. In an examination it is laid down that a student passes if he secures 30 percent or more marks. He is placed in the first, second or third division according as he secures 60% or more marks, between 45% to 60% marks and marks between 30% and 45% respectively. He gets distinction in case he secures 80% or more marks. It is noticed from the result that 10% of the students failed in the examination, whereas 5% of them obtained distinction. Calculate the percentage of students placed in the second division. (Assume normal distribution of marks).

Answers: 1. 0.0336 (67 bulbs), 0.0668 (134 bulbs), 0.6065 (1213). 2. 0.0668, 0.2877, 0.649
3. 0.0062, 0.02335 4. 34%.

SAMPLING THEORY

In a statistical investigation the interest usually lies in the assessment of the general magnitude and the study of variation with respect to one or more characteristics relating to individuals belonging to a group. This group of individuals under study is called *population or universe*. Thus in statistics, population is an aggregate of objects, animate or inanimate, under study. The population may be finite or infinite.

It is obvious that, for any statistical investigation complete enumeration of the population is rather impracticable. For example, if we want to have an idea of the average per capita (monthly) income of the people in India, we will have to enumerate all the earning individuals in the country which is rather a very difficult task.

If the population is infinite, complete enumeration is not possible. Also if the units are destroyed in the course of inspection, (*e.g.*, inspection of crackers, explosive materials, etc.), 100% inspection, though possible, is not at all desirable. But even if the population is finite or the inspection is not destructive, 100% inspection is not taken recourse to because of multiplicity of causes, *viz.*, *administrative* and financial implications, time factor, etc., and we take the help of *sampling*.

Size of the population is the number of objects or observations in the population and is denoted by N. A finite subset of statistical individuals in a population is called a *sample* and the number of individuals in a sample is called the sample size. Size of the sample is denoted by n. If $n \geq 30$, the sampling is said to be *large sampling*. If $n < 30$, the sampling is said to be *small sampling*.

For the purpose of determining population characteristics, instead of enumerating the entire population, the individuals in the sample only are observed. Then the sample characteristics are utilized to approximately determine or estimate the population. This method is called the *statistical inference*. For example, on examining the sample of a particular stuff we arrive at a decision of purchasing or rejecting that stuff. The error involved in such approximation is known *sampling error* and is inherent and unavoidable in any and every sampling scheme. But sampling results in considerable gains, especially in time and cost not only in respect of making observations of characteristics but also in the subsequent handling of the data.

Sampling is quite often used in our day to day practical life. For example, in a shop we assess the quality of sugar, wheat or any other commodity by taking a handful of it from the bag and then decide to purchase it or not. A housewife normally tests the cooked products to find if they are properly cooked and contain the proper quantity of salt. Statistical measures or constants obtained from the population such as population mean, variance etc., are called the *Parameters*. Similarly, statistical quantities computed from sample such as sample mean, sample variance etc., are known as *statistics.*

Types or Sampling: Some of the commonly known and frequently used types of sampling are: (i) Purposive sampling (ii) Random sampling (iii) Stratified sampling, (iv) Systematic sampling.

Consider the following example.

Suppose the following are the marks of 30 students in a test carrying 10 marks; the marks are arranged, say, according to the roll number of the students.

2, 4, 0, 5, 8, 6, 4, 1, 3, 5, 3, 3, 2, 4, 7, 7, 3, 2, 0, 4, 6, 8, 7, 1, 8, 1, 4, 5, 6, 7.

Information of this type is called raw (or an unclassified) statistical data. The individual numbers present in the data are called the items or the observations in the data. Denote them by $x_1, x_2, x_3, \ldots$. The information can be put in the form of a table called the table of discrete frequency distribution.

x_i	f_i
0	2
1	3
2	3
3	4
4	5
5	3
6	3
7	4
8	3

The entries in the first column are called the variables x_i and the entries in the second column are called the frequencies f_i .

Further the data can be grouped as below.

Marks class-intervals	No of students f_i
0 - 2	8
3 - 5	12
6 - 8	10

The table of the above type is called a table of grouped frequency distribution. The entries in the first column are called the class-intervals (or classes) and the entries in the second column are the frequencies.

While analyzing statistical data, it is generally observed that the items or the frequencies cluster around some central value of the variable. Such a central value is called a measure of central tendency of the data. The mean (or average) is one such measure.

Mean:

1. For a raw data consisting of 'n' items $x_1, x_2, x_3, \dots x_n$, the arithmetic mean or mean is defined by the formula

$$\text{Mean} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum x_i}{n}$$

2. For a frequency distribution (x_i, f_i) , the mean is defined by the formula

$$\text{Mean} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \cdots + f_nx_n}{f_1 + f_2 + f_3 + \cdots + f_n} = \frac{\sum f_ix_i}{\sum f_i}$$

Variance:

1. For a raw data, the variance is defined by Variance $= \frac{1}{n}\sum(x_i - \text{mean})^2$.
2. For a frequency distribution the variance is defined by

$$\text{Variance} = \frac{\sum f_i(x_i - \bar{x})^2}{\sum f_i} = \frac{\sum x_i^2 f_i}{\sum f_i} - \bar{x}^2.$$

Sampling Distributions:

Given a population, suppose we consider a set of samples of a certain size drawn from the population. For each sample, suppose we compute a statistic (such as the mean, standard deviation, etc). These statistics will vary from one sample to the other sample. Suppose, we group these different statistics according to their frequencies and form a frequency distribution. The frequency distribution so formed is called a sampling distribution. The standard deviation of a sampling distribution is called its standard error. The standard error is used to assess the difference between the expected values and observed values.

Sampling Distribution of Means:

Consider a population for which the mean is μ and the standard deviation is σ. Suppose we draw a set of samples of a certain size n from this population and find the mean $\bar{x}$ of each of these samples. The frequency distribution of these means is called a sample distribution of means.

Suppose the population is finite with size N. Then $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ are related to μ and σ through the following formulae:

$$\mu_{\bar{X}} = \mu\,,\ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \times \frac{\sqrt{N-n}}{\sqrt{N-1}}$$

If the population is infinite (or if the sampling is finite with replacement), the formula is given as;

$$\mu_{\bar{X}} = \mu\,,\ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

It can be proved that for samples of large size or for samples with replacement, the sampling distribution of means is approximately a normal distribution for which the sample mean $\bar{x}$ is the random variable. If the population itself is normally distributed, then the sampling distribution of means is a binomial distribution even for samples of small size. Accordingly, the standard normal variate for the sampling distribution of means is given by $Z = \frac{X-\mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X}-\mu}{\sigma_{\bar{X}}}$.

Problems:

1. A population consists of four numbers 3, 7, 11, 15. Consider all possible samples of size 2 which can be drawn from this population with and without replacement. Find the mean and standard deviation in the population, and in the sampling distribution of means verify the formulas $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$.

   Solution: Given $N = 4$

   $$\text{Mean} = \mu = \frac{(3+7+11+15)}{4} = 9$$

   $$\text{Variance} = \sigma^2 = \frac{1}{4}\{(3-9)^2 + (7-9)^2 + (11-9)^2 + (15-9)^2\} = 20$$

   $$\text{Standard deviation} = \sigma = \sqrt{20}$$

Case i) Sampling with replacement:

Possible samples of size two which can be drawn with replacement are (3, 3), (3, 7), (3, 11), (3, 15), (7, 3), (7, 7), (7, 11), (7, 15), (11, 3), (11, 7), (11, 11), (11, 15), (15, 3), (15, 7), (15, 11), (15, 15). The means of these 16 samples are 3, 5, 7, 9, 5, 7, 9, 11, 7, 9, 11, 13, 9, 11, 13, 15 respectively. The corresponding frequency distribution is

$\bar{x}_i$	3	5	7	9	11	13	15
f_i	1	2	3	4	3	2	1

Mean= $\mu_{\bar{X}} = \frac{\Sigma f_i \bar{x}_i}{\Sigma f_i} = \frac{3+10+21+36+33+26+15}{16} = 9$

Variance= $\sigma_{\bar{X}}^2 = \frac{\Sigma f_i (\bar{x}_i - \mu_{\bar{X}})^2}{\Sigma f_i} = \frac{1}{16}[3(3-9)^2 + 5(5-9)^2 + 3(7-9)^2 +$
$3(11-9)^2 + 2(13-9)^2 + (15-9)^2]$

$= 10$

Standard variation= $\sigma_{\bar{X}} = \sqrt{10}$

$\Rightarrow \mu_{\bar{X}} = \mu$

and

$$\frac{\sigma}{\sqrt{n}} = \frac{\sqrt{20}}{\sqrt{2}} = \sqrt{10} = \sigma_{\bar{X}}$$

Case ii) Sampling without replacement:

Possible samples of size two which can be drawn without replacement is (3, 7), (3, 11), (3, 15), (7, 11), (7, 15), (11, 15). The mean of these 6 samples are 5, 7, 9, 9, 11, 13 respectively. For this distribution,

Mean $= \mu_{\bar{X}} = \frac{(5+7+9+9+11+13)}{6} = 9.$

Variance $= \sigma_{\bar{X}}^2$

$= \frac{1}{6}\{(5-9)^2 + (7-9)^2 + (9-9)^2 + (9-9)^2 + (11-9)^2 +$
$(13-9)^2\} = \frac{20}{3}.$

Standard deviation $= \sigma_{\bar{X}} = \frac{\sqrt{20}}{\sqrt{3}}.$

$\frac{\sigma}{\sqrt{n}} \frac{\sqrt{N-n}}{\sqrt{N-1}} = \frac{\sqrt{20}}{\sqrt{2}} \frac{\sqrt{4-2}}{\sqrt{4-1}} = \frac{\sqrt{20}}{\sqrt{3}} = \sigma_{\bar{X}}$. Also, $\mu_{\bar{X}} = \mu$.

2. The daily wages of 3000 workers in a factory are normally distributed with mean equal to Rs 68 and standard deviation equal to Rs 3. If 80 samples consisting of 25 workers each are obtained, what would be the mean and standard deviation of the sampling distribution of

means if sampling were done (a) with replacement (b) without replacement? In how many samples will the mean is likely to be (i) between Rs 66.8 & Rs 68.3 and (ii) less than Rs 66.4?

Solution: Given $N = 3000\,, \mu = 68\,, \sigma = 3\,, n = 25\,.$

In case of sampling with replacement

$\mu_X = \mu = 68$ and $\sigma_{\bar{X}} \;= \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{25}} = 0.6\,.$

In case of sampling without replacement

$\mu_{\bar{X}} = \mu = 68$ and $\sigma_{\bar{X}} \;= \frac{\sigma}{\sqrt{n}}\frac{\sqrt{N-n}}{\sqrt{N-1}} = \frac{3}{\sqrt{25}}\frac{\sqrt{3000-25}}{\sqrt{3000-1}} = 0.5976\; \approx 0.6\,.$

Since the population is normally distributed, the sampling distribution of means is also taken as normally distributed. The standard normal variate associated with the sample mean $\bar{x}$ is

$Z = \frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{x}-68}{0.6}$

$P(66.8 < \bar{x} < 68.3) = P\left(\frac{66.8-68}{0.6} < z < \frac{68.3-68}{0.6}\right) = P(-2 < z < 0.5)$

$F(0.5) - F(-2) = F(0.5) - 1 + F(2) = 0.6915 - 1 + 0.9773 = 0.6688\,.$

In a sample of 80; $= 0.6687 \times 80 \approx 53\,.$

$P(\bar{X} < 66.4) = P\left(z < \frac{66.4-68}{0.6}\right) = P(z < -2.67)$

$= 1 - F(2.67) = 1 - 0.9962 = 0.0038\,.$

Thus, in a sample of 80 = 0.0038 × 80 = 0.3040.

3. Let $\overline{x}$ be the mean of a random sample of size 50 drawn from a population with mean 112 and standard deviation 40. Find (a) the mean and standard deviation of $\overline{x}$, (b) the probability that $\overline{x}$ assumes a value between 110 and 114 (c) the probability that $\overline{x}$ assumes a value greater than 113.

Solution: $n = 50$, $\mu = 112$, $\sigma = 40$

$\mu_{\bar{X}} = \mu = 112$, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{50}} = 5.6569$ and $z = \frac{\bar{x}-\mu_{\bar{X}}}{\sigma_{\bar{X}}}$

$P(110 < \overline{x} < 114) = P(-0.35 < z < 0.35) = 0.6368 - 0.3632 = 0.2736$

$P(\overline{x} > 113) = P(z > 0.18) = 1 - P(z \leq 0.18) = 1 - 0.5714 = 0.4286$

4. An automobile battery manufacturer claims that its midgrade battery has a mean life of 50 months with a standard deviation of 6 months. Suppose the distribution of battery lives of this particular brand is approximately normal. On the assumption that the manufacturer's claims are true, find (a) the probability that a randomly selected

battery of this type will last less than 48 months, (b) the probability that the mean of a random sample of 36 such batteries will be less than 48 months.

Solution: $n = 36$, $\mu = 50$, $\sigma = 6$

$\mu_{\overline{x}} = \mu = 50$, $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{36}} = 1$

Using $z = \frac{\overline{x}-\mu}{\sigma}$ $P(\overline{x} < 48) = P(z < -0.33) = 0.3707$

Using $z = \frac{\overline{X}-\mu_{\overline{X}}}{\sigma_{\overline{X}}}$ $P(\overline{X} < 48) = P(z < -2) = 0.0228$

Exercise:

1. A population consists of four numbers 1, 5, 6, 8. Consider all possible samples of size 2 which can be drawn from this population with and without replacement. Find the mean and standard deviation in the population, and in the sampling distribution of means verify the formulas $\mu_{\overline{X}}$ and $\sigma_{\overline{X}}$.
2. The mean and the standard deviation of a normally distributed population of size 250 are 100 and 16 respectively. What are the mean and the standard deviation of the sampling distribution of means for random samples of size 4 drawn with replacement and without replacement?
3. With reference to the above Problem No2, what is the probability that the sample mean lies between 95 and 105 for a sample of size 4 drawn with and without replacement?
4. The mean of a certain normal infinite population is equal to the standard error of the distribution of means of samples of size 100 drawn from that population. Find the probability that the mean of a sample of size 25 drawn from the population will be negative.
5. If the mean of an infinite population is 575 with standard deviation 8.3 how large a sample must be used in order that there be one chance in 100 that the mean of the sample is less than 572?
6. Suppose that the number of customers entering a grocery shop each day over a five-year period is a random variable with mean 100 and standard deviation of 10. Then what is the probability that randomly selected 30-day period is between 95 and 105?

Answers: 1) With replacement: $\mu_{\bar{x}} = 5$ and $\sigma_{\bar{x}} = \frac{\sqrt{13}}{2}$, Without replacement: $\mu_{\bar{x}} = 5$ and $\sigma_{\bar{x}} = \sqrt{\frac{13}{6}}$,

2) With replacement: Mean= 100, SD= 8, Without replacement: mean=100, SD 7.95,

3) With replacement: P=0.46, Without replacement: P=0.47

4) 0.3085

5) n = 43.

6) 0.9946

Sampling distribution of differences of means

In some situations, we may be interested to draw the inference about the differences of two population means. For example, two companies of bulbs are produced sam type of bulbs and one may be interested to know which one is better.

If we have two independent populations with means μ_1 and μ_2 and variance σ_1^2 and σ_2^2, and if $\bar{X}_1$ and $\bar{X}_2$ are the sample mean of two independent samples of sizes $n_1$and n_2 from these populations, then sampling distribution of the differences of means $\bar{X}_1 - \bar{X}_2$ is approximately normally distributed with mean and variance

Mean: $\mu_{(\bar{X}_1 - \bar{X}_2)} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$

Variance: $\sigma^2_{(\bar{X}_1 - \bar{X}_2)} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

The distribution of $Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ is approximately standard normal.

Sampling distribution of sample proportion

When our interest is to study the proportion of population possess certain characteristics, for example, the proportion of heart patients admitted in hospitals. In such situation, we study the inference about the population proportion.

Mean $\mu_{\hat{P}} = p$

Standard deviation $\sigma_{\hat{P}} = \sqrt{\frac{pq}{n}} = \left[\frac{p(1-p)}{n}\right]^{\frac{1}{2}}$

The distribution of $Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$ is approximately standard normal if n is large or $np \geq 5$ and $np(1-p) \geq 5$.

Problems:

1. The television picture tubes of manufacturer A have a mean lifetime of 6.5 years and a standard deviation of 0.9 year, while those of manufacturer B have a mean lifetime of 6.0 years and a standard deviation of 0.8 year. What is the probability that a random sample of 36 tubes from manufacturer A will have a mean lifetime that is at least 1 year more than the mean lifetime of a sample of 49 tubes from manufacturer B?

Solution: Given

Population A	Population B
$\mu_1 = 6.5$	$\mu_2 = 6$
$\sigma_1 = 0.9$	$\sigma_2 = 0.8$
$n_1 = 36$	$n_2 = 49$

To find $P(\bar{X}_1 - \bar{X}_2 \geq 1)$

$$\mu_{(\bar{X}_1 - \bar{X}_2)} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2 = 0.5$$

$$\sigma^2_{(\bar{X}_1 - \bar{X}_2)} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = 0.03556$$

$$\sigma_{\bar{X}_1 - \bar{X}_2} = 0.1885$$

Using $z = \frac{\bar{X} - \mu_{(\bar{X}_1 - \bar{X}_2)}}{\sigma_{\bar{X}_1 - \bar{X}_2}}$ $P\left(\bar{X}_1 - \bar{X}_2 \geq 1\right) = P(z \geq 2.6476) = 0.0040$

2. A machine produces a large number of items of which 15% are found to be defective. If a random sample of 200 items is taken from the population and sample proportion is calculated then find (i) Mean and standard error of sampling distribution of proportion. (ii) The probability that less than or equal to 12% defectives are found in the sample.

Solution: Given population proportion, $p = 0.15$

Sample size $n = 200$

(i) Mean $\mu_{\hat{P}} = p = 0.15$

Standard deviation $\sigma_{\hat{P}} = \sqrt{\frac{pq}{n}} = \left[\frac{p(1-p)}{n}\right]^{\frac{1}{2}} = \sqrt{\frac{0.15 \times 0.85}{200}} = 0.0252$

(ii) The distribution of $Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$ is approximately standard normal.

$$P\left(\hat{P} \leq 0.12\right) = P(z \leq -1.1904) = 0.1169$$

Video Links:

https://www.youtube.com/watch?v=82Ad1orN-NA
https://www.youtube.com/watch?v=c06FZ2Yq9rk
https://www.youtube.com/watch?v=N-IVFB8Rlfo
https://www.youtube.com/watch?v=d5iAWPnrH6w
https://www.youtube.com/watch?v=vjXLH7FXrj8