4. (a) 1/36, (b) 1/6, (c) 1/4 (d) 5/6
(e) 1/6 (f) 1/6 (g) 1/2

5.

x_i	1	2	3
$f(x_i)$	1/6	1/3	1/2

y_i	1	2	3
$g(y_i)$	1/6	1/3	1/2

X and Y are independent random variables.

6. (a) $P(X \le x) = F_1(x) = \begin{cases} 0, & x < 1 \\ x^2/16, & 0 \le x < 4 \\ 1, & x \ge 4 \end{cases}$

$$P(Y \le y) = F_2(y) = \begin{cases} 0, & y < 1 \\ (y^2-1)/24, & 1 \le y < 5 \\ 1, & y \ge 5 \end{cases}$$

(b) 5/128 and 7/128

7. (a) 1/4 (b) LHS = RHS = 27/64

8. $c = 2/3$

9. (a) 0.46 (b) 0.26

ENGINEERING MATHEMATICS - IV

Module - 5

5.1 Sampling Theory

5.11 Introduction

In our day to day life it becomes quite often necessary to draw some valid and reasonable conclusions concerning a large mass of individuals or things. It becomes practically impossible to examine every individual or the entire group known as *population*. Therefore we may prefer to examine a small part of this population known as a *sample* with the motive of drawing some conclusion about the entire population based on the information/result revealed by the sample. The entire process known as ***statistical inference*** aims in ascertaining maximum information about the population with minimum effort and time. Poll prediction is a good example for statistical inference.

5.12 Random Sampling

A large collection of individuals or attributes or numerical data can be understood as a *population* or *universe*.

A finite subset of the universe is called a ***sample***. The number of individuals in a sample is called a sample size. If the sample size (n) is less than or equal to 30 the sample is said to be small, otherwise it is a large sample.

The process of selecting a sample from the population is called as ***sampling***.

The selection of an individual or item from the population in such a way that each has the same chance of being selected is called as ***random sampling***. Suppose we take a sample of size n from a finite population of size N, then we will have N_{C_n} possible samples. Random sampling is a technique in which each of the N_{C_n} samples has an equal chance of being selected.

Sampling where a member of the population may be selected more than once is called as *sampling with replacement*, on the other hand if a member cannot be chosen more than once is called as *sampling without replacement*.

Simple sampling is a special case of random sampling in which trials are independent and the probability of success is a constant.

The word *statistic* is often used for the random variable or for its values.

5.13 Sampling Distributions

Let us suppose that we have different samples of size n drawn from a population. For each and every sample of size n we can compute quantities like mean, standard deviation etc. Obviously these will not be the same. Suppose we group these characteristics according to their frequencies, the frequency distributions so generated are called *sampling distributions*. These can be distinguished as sampling distribution of mean, standard deviation etc. *The sampling distribution of large samples is assumed to be a normal distribution.*

The standard deviation of a sampling distribution is also called the *Standard Error* (S.E). The reciprocal of the standard error is called *precision*.

Sampling distribution of the means

Sample mean is a statistic and we discuss the sampling distribution of this statistic.

We consider all possible random samples of size n and determine the mean of each one of these samples. We discuss the sampling distribution of the sample means for the two possible types of random sampling (with/without replacment) associated with a finite population.

Case - (1) ***Random sampling with replacement***

Here the items are drawn one by one and are put back to the population before the next draw. If N is the size of the finite population and n is the size of the sample then we have N^n samples.

The mean $\mu_{\bar{x}}$ of the frequency distribution of the sample means will be equal to the population mean (μ).

The variance $\sigma_{\bar{x}}^2$ of the frequency distribution of the sample means will be equal to $\frac{\sigma^2}{n}$ where σ^2 is the variance of the population.

Thus we have, $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}}^2 = \sigma^2/n$

$\sigma_{\bar{x}}$ is also called the *standard error* of the means.

Case - (2) ***Random sampling without replacement***

Here the items are drawn one by one and are not put back to the population before the next draw. In this case there will be ${}^{N}C_{n}$ samples and we have the following results in accordence with the notations used in the *case - (1)*

$$\mu_{\bar{x}} = \mu \; ; \; \sigma_{\bar{x}}^2 = \left[\frac{N-n}{N-1}\right]\frac{\sigma^2}{n} = C\,\frac{\sigma^2}{n}$$

where $C = \dfrac{N-n}{N-1}$ is called the *finite population correction factor.*

Remark : *If N is very large, then C is closer to 1 as we have* $\lim_{N\to\infty} \dfrac{N-n}{N-1} = 1$

In practice if N is large the correction factor can be omitted.

Illustrative Example

Let $\{1, 2, 3\}$ constitute a population. We form the sampling distribution of the sample means in the case of (1) random samples of size 2 with replacement (2) random samples of size 2 without replacement.

>> Here N (*size of the finite population*) $= 3$

Population mean $(\mu) = \dfrac{1+2+3}{3} = 2$

Population variance $(\sigma^2) = \dfrac{1}{3}\left\{(1-2)^2+(2-2)^2+(3-2)^2\right\} = \dfrac{2}{3}$

Case - (i) Random samples of size 2 ($n = 2$) with replacement

The various possible samples are :

(1, 1) (1, 2) (1, 3) ; (2, 1), (2, 2), (2, 3) ; (3, 1), (3, 2), (3, 3)

These are $N^n = 3^2 = 9$ in number.

The mean of these are respectively

1, 1.5, 2, 1.5, 2, 2.5, 2, 2.5, 3

We prepare the frequency distribution of these means where x is the variate and f is the frequency.

x	1	1.5	2	2.5	3
f	1	2	3	2	1

We shall compute the mean and variance of this frequency distribution.

$$\mu_{\bar{x}} = \frac{\sum fx}{\sum f} = \frac{1+3+6+5+3}{9} = 2$$

$$\sigma_{\bar{x}}^2 = \frac{\sum f(x-\mu_{\bar{x}})^2}{\sum f}$$

$$\sigma_{\bar{x}}^2 = \frac{1}{9}\left\{1(1-2)^2+2(1.5-2)^2+3(2-2)^2+2(2.5-2)^2+1(3-2)^2\right\}$$

$$\sigma_{\bar{x}}^2 = \frac{1}{9}(1+0.5+0+0.5+1) = \frac{1}{3}$$

Thus we have $\mu_{\bar{x}} = 2 = \mu$; $\sigma_{\bar{x}}^2 = \frac{1}{3}$ and $\sigma^2/n = \frac{(2/3)}{2} = \frac{1}{3}$

Case - (ii) Random sample of size 2 without replacement.

We have ${}^3C_2 = 3$ samples. The three samples are (1, 2), (2, 3), (3, 1)

The associated means are 1.5 , 2.5 , 2 . Further we have,

$$\text{Mean} = \mu_{\bar{x}} = \frac{1.5+2.5+2}{3} = 2$$

$$\text{Variance} = \sigma_{\bar{x}}^2 = \frac{1}{3}\left\{(1.5-2)^2+(2.5-2)^2+(2-2)^2\right\} = \frac{0.5}{3} = \frac{1}{6}$$

Here we note that $\mu_{\bar{x}} = \mu = 2$

Also $\left[\frac{N-n}{N-1}\right]\frac{\sigma^2}{n} = \left[\frac{3-2}{3-1}\right]\frac{2/3}{2} = \frac{1}{6} = \sigma_{\bar{x}}^2$

Thus $\sigma_{\bar{x}}^2 = \left[\frac{N-n}{N-1}\right]\frac{\sigma^2}{n} = \frac{1}{6}$

Note : *Suppose N is large say 500 and n is small say 5 then the correction factor (C) becomes*

$$C = \frac{N-n}{N-1} = \frac{500-5}{500-1} = \frac{495}{499} = 0.992 \approx 1$$

5.14 Testing of Hypothesis

In order to arrive at a decision regarding the population through a sample of the population we have to make certain assumption referred to as **hypothesis** *which may or may not be true.* Much depends on the framing of hypothesis.

The hypothesis formulated for the purpose of its rejection under the assumption that it is true is called the **Null Hypothesis** denoted by H_0.

Any hypothesis which is complimentary to the null hypothesis is called **Alternative Hypothesis** denoted by H_1.

Examples

(1) To test whether a process B is better than a process A we can formulate the hypothesis as *there is no difference between the process A and B.*

(2) To test whether there is a relationship between two variates we can formulate the hypothesis as *there is no relationship between them.*

- **Errors**

In a test process there can be four possible situations of which two of the situations leads to the two types of errors and the same is presented as follows.

	Accepting the hypothesis	Rejecting the hypothesis
Hypothesis true	Correct decision	Wrong decision (Type I error)
Hypothesis false	Wrong decision (Type II error)	Correct decision

In order to minimize both these types of errors we need to increase the sample size b is further important to note that acceptance or non acceptance of a hypothesis is purely based on the information revealed by the sample and what is indicated by a particular sample may not always be true in respect of the population.

A region which amounts to the rejection of null hypothesis is called critical region or region of rejection.

- **Significance level**

The probability level, below which leads to the rejection of the hypothesis is known as the *significance level.* This probability is conventionally fixed at 0.05 or 0.01 i.e. 5% or 1%. These are called *significance levels.*

We feel confident in rejecting a hypothesis at 1% level of significance than at 5% level of significance. 5% level of significance can also be understood as, the probability of committing errors of either types, (Type I or Type II) is 0.05.

- **Tests of significance and Confidence intervals**

The process which helps us to decide about the acceptance or rejection of the hypothesis is called the *test of significance.*

Let us suppose that we have a normal population with mean μ and S.D σ. If $\bar{x}$ is the sample mean of a random sample of size n the quantity z defined by

$$z = \frac{\bar{x}-\mu}{(\sigma/\sqrt{n})}$$

is called the *Standard Normal Variate (S.N.V).*

From the table of normal areas we find that 95% of the area lies between $z = -1.96$ and $z = +1.96$. In other words we can say with 95% confidence that z liesbetween -1.96 and $+1.96$. Further 5% level of significance is denoted by $z_{0.05}$. Thus we can write the verbal statement in the mathematical form,

$$-1.96 \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le 1.96$$

i.e., $$\frac{-\sigma}{\sqrt{n}}(1.96) \le \bar{x}-\mu \le \frac{\sigma}{\sqrt{n}}(1.96)$$

$$\Rightarrow \quad \mu \le \bar{x} + \frac{\sigma}{\sqrt{n}}(1.96) \text{ and } \bar{x} - \frac{\sigma}{\sqrt{n}}(1.96) \le \mu$$

Thus we can write by combining the two results in the form,

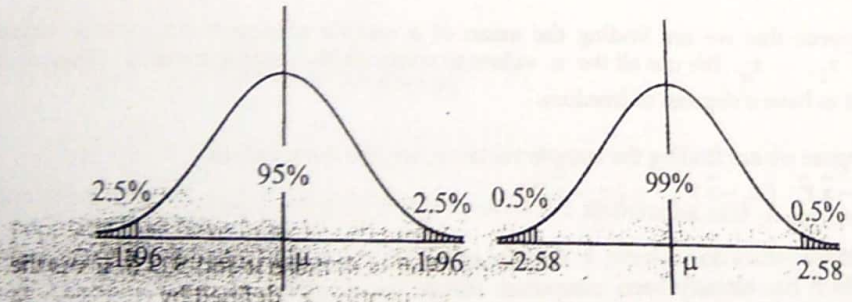
$$\bar{x} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right) \le \mu \le \bar{x} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right) \quad \ldots (2)$$

Similarly, from the table of normal areas 99% of the area lies between -2.58 and $+2.58$. This is equivalent to the form,

$$\bar{x} - 2.58\left(\frac{\sigma}{\sqrt{n}}\right) \le \mu \le \bar{x} + 2.58\left(\frac{\sigma}{\sqrt{n}}\right) \quad \ldots (3)$$

Thus we can say that (2) is the 95% confidence interval and (3) is the 99% confidence interval.

The constants 1.96, 2.58 etc. in the confidence limits are called confidence coefficients denoted by z_c. From confidence levels we can find confidence coefficients and vice-versa.



As reflected in the figure, we can say with 95% confidence that if the hypothesis is true, the value of z for an actual sample lies between -1.96 to 1.96 since the area under the normal curve between these values is 0.95. However if the value of z for random sample lies outside this range we can conclude that the probability of the happening of such an event is only 0.05 if the given hypothesis is true.

The total shaded area 0.05 being, the level of significance of the test, represents the probability of making type-I error. *(rejecting the hypothesis when it should have been accepted)*. The set of values of z outside the range -1.96, 1.96 constitutes the critical (*significant*) region or the region of rejecting the hypothesis whereas the values of z within the same range constitutes the insignificant region or the region of acceptance of the hypothesis.

- **One tailed and two tailed tests**

In our test of acceptance or non acceptance of a hypothesis we concentrated on the value of z on both sides of the mean. This can be categorically stated that the focus of attention lies in the two "tails" of the distribution and hence such a test is called a *two tailed test*.

Sometimes we will be interested in the extreme values to only one side of the mean in which case the region of significance will be a region to one side of the distribution. Obviously the area of such a region will be equal to the level of significance itself. Such a test is called a *one tailed test*.

The critical values of z : -1.96 , 1.96 as already stated with reference to 5% and 1% level of significance can be understood as the values in respect of a two tailed test. However the critical values of z in respect of a one tailed test (*as found in the table of areas under a normal curve*) are $[-1.645, 1.645]$; $[-2.33, 2.33]$ at 5% and 1% level of significance respectively.

The following table will be useful for working problems.

Test	Critical values of z	
	5% level	1% level
One-tailed test	–1.645 or 1.645	–2.33 or 2.33
Two-tailed test	–1.96 and 1.96	–2.58 and 2.58

5.15 Tests of significance for large samples

- **Test of significance of proportions**

In the discussion of probability distributions we have remarked that the normal distribution is the limiting form of the binomial distribution when n is large and neither p nor q is small. Let us suppose that we take N samples, each having n members. Let p be the probability of success of each member and q of failure so that $p + q = 1$. The frequencies of samples with successes 0, 1, 2, . . . n are the terms of the binomial expansion of $N(q+p)^n$. Thus the binomial distribution is regarded as the sampling distribution of the number of successes in the sample.

We know that the mean of this distribution is np and S.D is $\sqrt{npq}$.

Let us consider the proportion of successes.

(1) mean proportion of successes $= \frac{np}{n} = p$

(2) S.D or S.E proportion of successes $= \frac{\sqrt{npq}}{n} = \sqrt{pq/n}$

Let x be the observed number of successes in a sample size of n and $\mu = np$ be the expected number of successes. Let the associated standard normal variate Z be defined by

$$Z = \frac{x-\mu}{\sigma} = \frac{x-np}{\sqrt{npq}}$$

If $|Z| > 2.58$ we conclude that the difference is highly significant and reject the hypothesis. Since p is the probability of success and $\sqrt{pq/n}$ is the S.E proportion of successes, $p \pm 2.58\sqrt{pq/n}$ are the probable limits.

- **Test of significance for difference of means**

Let μ_1 and μ_2 be the mean of two populations.
Let $(\bar{x}_1, \sigma_1); (\bar{x}_2, \sigma_2)$ be the mean and S.D of two large samples of size n_1 and n_2 respectively. We wish to test the null hypothesis H_0 that there is no difference between the population means. That is $H_0 : \mu_1 = \mu_2$.

The statistic for this test is given by

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Also confidence limits for the difference of means of the population are

$$(\bar{x}_1 - \bar{x}_2) \pm z_c\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

We adopt the same procedure for testing the null hypothesis by using one tailed test or two tailed test.

Corollary: If the samples are drawn from the same population then $\sigma_1 = \sigma_2 = \sigma$

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma\sqrt{(1/n_1 + 1/n_2)}}$$

- **Test of significance for difference of properties (attributes) for two samples**

Let p_1 and p_2 be the sample proportions in respect of an attribute corresponding to two large samples of size n_1 and n_2 drawn from two populations.

We wish to test the null hypothesis H_0 that there is no difference between the population with regard to the attribute.

The statistic for this test is given by

$$Z = \frac{p_1 - p_2}{\sqrt{pq\,(1/n_1 + 1/n_2)}} \quad \text{where} \quad p = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} \quad \text{and} \quad q = 1-p$$

We adopt the same procedure for testing the null hypothesis by using one tailed test or two tailed test.

- **Test of significance for small samples**

In the case of large samples, sampling distribution follows a normal distribution which is not true in the case of small samples. We introduce the concept of degrees of freedom for discussing *Student's t distribution*.

- **Degrees of freedom**

The number of degrees of freedom $(d \cdot f)$ usually denoted by ν is the number of values in a set which may be assigned arbitrarily. It can be interpreted as the number of independent values generated by a sample of small size for estimating a population parameter.

Examples : Let us suppose that we need to find 3 numbers whose sum is 25. That is to find a, b, c such that $a+b+c = 25$. We can arbitrarily assign values to any two of the variables a, b, c and hence these are the degrees of freedom. That is to say that $d \cdot f(\nu) = 2$. If there are n observations $d \cdot f$ is equal to $(n-1)$.

Suppose that we are finding the mean of a sample of size n comprising values $x_1, x_2, \cdots x_n$. We use all the n values to compute the sample mean $\bar{x}$. Then $\bar{x}$ is said to have n degrees of freedom.

Suppose we are finding the sample variance, we use the n values $(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \cdots (x_n - \bar{x})^2$.

But these values do not have n degrees of freedom as they all depend on a fixed value $\bar{x}$ which has already been computed. Hence the sample variance is said to have $(n-1)\ d \cdot f$. If we compute another statistic based on the sample mean and variance, that statistic is said to have $(n-2)\ d \cdot f$ and so on. In general the number of degrees of freedom $\nu = n - k$ where n is the number of observations in the sample and k is the number of constraints / number of values which are pre determined.

5.16 Student's t Distribution

Sir william Gosset under the pen name '*Student*' derived a theoretical distribution to test the significance of a sample mean where the small sample is drawn from a normal population.

Let $x_i\ (i = 1, 2, \cdots n)$ be a random sample of size n drawn from a normal population with mean μ and variance σ^2. The statistic t is defined as follows.

$$t = \frac{\bar{x}-\mu}{(s/\sqrt{n})} = \frac{\bar{x}-\mu}{s}\sqrt{n}$$

Here $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean.

$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2$ is the sample variance.

$\nu = (n-1)$ denote the number of degrees of freedom of t.

The statistic t follows the Student's t distribution with $(n-1)\ d\cdot f$ having the probability density function

$$y = f(t) = \frac{y_0}{\left[1+t^2/\nu\right]^{(\nu+1)/2}}$$

where y_0 is a constant such that the area under the curve is unity.

Note : 1. *Statistic t is also defined as follows.*

$$t = \frac{\bar{x}-\mu}{\sigma}\sqrt{n-1}$$

2. *The constant y_0 present in p.d.f is given by*

$$y_0 = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma(\nu/2)}$$

so that the p.d.f of the Student's t distribution with ν degrees of freedom is given by

$$y = f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma(\nu/2)}\left[1+\frac{t^2}{\nu}\right]^{-(\nu+1)/2} \qquad ;-\infty < t < \infty$$

3. *If ν is large $(\nu \geq 30)$ the graph of $f(t)$ closely approximates standard normal curve. In other words we can say that t is normally distributed for large samples.*

• Student's t test for a sample mean

We need to test the hypothesis, whether the sample mean $(\bar{x})$ differs significantly from the population mean (μ) / hypothetical value (μ).

We compute $t = \frac{\bar{x}-\mu}{s}\sqrt{n}$ and consider $|t|$.

We also take a note of the value of t for the given $d\cdot f$ from the table of standard values.

If $|t| > t_{.05}$ the difference between $\bar{x}$ and μ is said to be significant at 5% level of significance.

If $|t| > t_{.01}$ the difference is said to be significant at 1% level of significance.

If $|t|$ is less than the table value at a certain level of significance, the data is said to be conformal / consistent with the hypothesis that μ is the mean of the population.

• Confidence limits for the population mean μ

If $t_{.05}$ is the tabulated value of t for $(n-1)\,d\cdot f$ at 5% level of significance, it implies that.

$$P[|t| > t_{.05}] = 0.05$$

$$\Rightarrow \quad P[|t| \leq t_{.05}] = 1-0.05 = 0.95$$

Now consider $|t| \leq t_{.05}$

i.e., $$\left|\frac{\bar{x}-\mu}{s}\sqrt{n}\right| \leq t_{.05}$$

or $$\left|\frac{\bar{x}-\mu}{(s/\sqrt{n})}\right| \leq t_{.05}$$

i.e., $$-t_{.05} \leq \frac{\bar{x}-\mu}{(s/\sqrt{n})} \leq t_{.05}$$

i.e., $$\frac{-s}{\sqrt{n}}t_{.05} \leq \bar{x}-\mu \leq \frac{s}{\sqrt{n}}t_{.05}$$

$$\Rightarrow \quad \mu \leq \bar{x}+\frac{s}{\sqrt{n}}t_{.05} \text{ and } \bar{x}-\frac{s}{\sqrt{n}}t_{.05} \leq \mu$$

Combining these two results we can write in the form

$$\bar{x}-\frac{s}{\sqrt{n}}t_{.05} \leq \mu \leq \bar{x}+\frac{s}{\sqrt{n}}t_{.05}$$

Thus we have 95% confidence limits for μ given by $\bar{x} \pm \frac{s}{\sqrt{n}} t_{.05}$

Similarly 99% confidence limits for μ are given by $\bar{x} \pm \frac{s}{\sqrt{n}} t_{.01}$

Note : *Confidence limits are also called Fiducial limits.*

- **Test of significance of difference between sample means**

Consider two independent samples $x_i (i = 1, 2, \cdots n_1)$ and $y_j (j = 1, 2, \cdots n_2)$ drawn from a normal population.

Let $(\bar{x}, \sigma_x)$ and $(\bar{y}, \sigma_y)$ respectively be the mean and variance of the two samples. Let μ be the population mean and σ be the population variance. We need to test the hypothesis whether the difference between the sample means is significant.

We compute $t = \dfrac{\bar{x} - \bar{y}}{S.\sqrt{1/n_1 + 1/n_2}}$

where $S^2 = \dfrac{1}{n_1 + n_2 - 2} \left\{ \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right\}$

and degrees of freedom $v = n_1 + n_2 - 2$

if $|t| > t_{.05}$ the difference between the sample means is said to be significant at 5% level of significance. Similarly for 1% level of significance also.

Table for values of $|t|$ is given at the end of the book.

5.17 Chi - Square distribution

Chi - Square distribution provides a measure of correspondence between the theoretical frequencies and observed frequencies.

If $0_i (i = 1, 2, \ldots n)$ and $E_i (i = 1, 2, \ldots n)$ respectively denotes a set of observed and estimated frequencies, the quantity chi - square denoted by χ^2 is defined as follows.

$$\chi^2 = \sum_{i=1}^{n} \frac{(0_i - E_i)^2}{E_i} \quad ; \text{ degrees of freedom} = n - 1$$

Note : *If the expected frequencies are less than 10, we group them suitably for computing the value of chi. square.*

- **Chi - Square test as a test of goodness of fit**

It is possible to test the hypothesis about the association of two attributes. We have already discussed the fitting of Binomial distribution, Normal distribution, Poisson distribution to a given data. It is easily possible to find the theoretical frequencies from the distribution of fit.

Chi - Square test helps us to test the goodness of fit of these distributions.

If the calculated value of χ^2 is less than the table value of χ^2 at a specified level of significance the hypothesis is accepted, otherwise the hypothesis is rejected.

WORKED PROBLEMS

Sampling distribution of the means

1. *A population consists of five numbers 2, 3, 6, 8, 11. Consider all possible samples of size 2 which can be drawn with replacement from this population. Find (a) the mean and S.D of the population. (b) the mean and standard deviation of the sampling distribution of means. (c) Considering samples without replacement find the mean and S.D of the sampling distribution of means.*

>> (a) Population mean $\mu = \dfrac{2+3+6+8+11}{5} = 6$

Population variance $\sigma^2 = \dfrac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$

i.e., $\sigma^2 = \dfrac{1}{5} \left\{ (2-6)^2 + (3-6)^2 + (6-6)^2 + (8-6)^2 + (11-6)^2 \right\}$

$\sigma^2 = \dfrac{54}{5} = 10.8$

Thus $\mu = 6$ and $\sigma = \sqrt{10.8}$

(b) Let us consider all samples of size 2 with replacement. $N = 5$, $n = 2$. There will be $N^n = 5^2 = 25$ samples which are as follows.

(2, 2) (2, 3) (2, 6) (2, 8) (2, 11)
(3, 2) (3, 3) (3, 6) (3, 8) (3, 11)
(6, 2) (6, 3) (6, 6) (6, 8) (6, 11)
(8, 2) (8, 3) (8, 6) (8, 8) (8, 11)
(11, 2) (11, 3) (11, 6) (11, 8) (11, 11)

The mean of these samples in the respective order is as follows.

(2, 2.5, 4, 5, 6.5) ; (2.5, 3, 4.5, 5.5, 7)

(4, 4.5, 6, 7, 8.5) ; (5, 5.5, 7, 8, 9.5) (6.5, 7, 8.5, 9.5, 11)

Note: *The mean and S.D can be found as we have done in the case of population or by forming a frequency distribution as follows.*

x	2	2.5	3	4	4.5	5	5.5	6	6.5	7	8	8.5	9.5	11
f	1	2	1	2	2	2	2	1	2	4	1	2	2	1

$$\mu_{\bar{x}} = \frac{\Sigma fx}{\Sigma f} = \frac{150}{25} = 6$$

$$\sigma_{\bar{x}}^2 = \frac{\Sigma fx^2}{\Sigma f} - [\mu_{\bar{x}}]^2 = \frac{1035}{25} - (6)^2 = 5.4$$

Thus $\mu_{\bar{x}} = 6$ and $\sigma_{\bar{x}} = \sqrt{5.4}$

Remark : *We observe that* $\mu_{\bar{x}} = \mu$ *and* $\sigma_{\bar{x}}^2 = \sigma^2/n$ *in this case of random sampling with replacement.*

(c) Let us consider random samples without replacement.

$N_{C_n} = 5_{C_2} = 10$ samples are as follows.

(2, 3) (2, 6) (2, 8) (2, 11) (3, 6) (3, 8)
(3, 11) (6, 8) (6, 11) (8, 11)

The mean of these samples are respectively

2.5, 4, 5, 6.5, 4.5, 5.5, 7, 7, 8.5, 9.5

$$\mu_{\bar{x}} = \frac{1}{10}(2.5+4+5+\cdots+9.5) = \frac{60}{10} = 6$$

$$\sigma_{\bar{x}}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\bar{x}})^2 = \frac{1}{10}(40.5) = 4.05$$

Thus $\mu_{\bar{x}} = 6$ and $\sigma_{\bar{x}} = \sqrt{4.05}$

Remark : *We observe that* $\mu_{\bar{x}} = \mu$. *Also the result*

$\sigma_{\bar{x}}^2 = \dfrac{\sigma^2}{n}\left(\dfrac{N-n}{N-1}\right)$ *can be verified as we have*

$$\text{RHS} = \frac{10.8}{2}\left(\frac{5-2}{5-1}\right) = \frac{10.8}{2} \times \frac{3}{4} = 4.05 = \sigma_{\bar{x}}^2 = \text{LHS}$$

2. *A population consists of 4 numbers 3, 7, 11, 15.*

(a) *Find the mean and S.D of the sampling distribution of means by considering samplings of size 2 with replacement.*

(b) *If N,n denotes respectively the population size and sample size,* σ *and* $\sigma_{\bar{x}}$ *respectively denotes population S.D and S.D of the sampling distribution of means without replacement verify that*

(i) $\sigma_{\bar{x}}^2 = \dfrac{\sigma^2}{n}\left[\dfrac{N-n}{N-1}\right]$

(ii) $\mu_{\bar{x}} = \mu$ *where* $\mu_{\bar{x}}$ *is the mean of this distribution and* μ *is the population mean.*

>> Population mean $\mu = \frac{1}{4}(3+7+11+15) = 9$

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

Thus $\mu = 9$ and $\sigma = \sqrt{20}$

(a) Let us consider samples of size 2 with replacement. They are as follows.

(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)

Sampling means are as follows.

(3, 5, 7, 9); (5, 7, 9, 11); (7, 9, 11, 13): (9, 11, 13, 15)

The frequency distribution of the sampling means is as follows.

x	3	5	7	9	11	13	15
f	1	2	3	4	3	2	1

$$\mu_{\bar{x}} = \frac{\Sigma fx}{\Sigma f} = \frac{144}{16} = 9$$

$$\sigma_{\bar{x}}^2 = \frac{\Sigma fx^2}{\Sigma f} - [\mu_{\bar{x}}]^2 = \frac{1456}{16} - (9)^2 = 10$$

Thus $\mu_{\bar{x}} = 9$ and $\sigma_{\bar{x}} = \sqrt{10}$

Remark : $\mu_{\bar{x}} = \mu$ *and* $\sigma_{\bar{x}}^2 = \sigma^2/n$ *where* $\sigma^2 = 20$, $n = 2$

(b) Let us consider samples without replacement. They are as follows.

(3, 7) (3, 11) (3, 15) (7, 11) (7, 15) (11, 15)

The sampling means are 5, 7, 9, 9, 11, 13

$$\therefore \quad \mu_{\bar{x}} = \frac{1}{6}(5+7+9+9+11+13) = 9 \text{ ; Thus } \mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}}^2 = \frac{1}{6}\left\{(5-9)^2+(7-9)^2+\cdots+(13-9)^2\right\} = \frac{40}{6} = \frac{20}{3}$$

Consider $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left[\frac{N-n}{N-1}\right]$

$$\text{RHS} = \frac{20}{2}\left[\frac{4-2}{4-1}\right] = 10 \times \frac{2}{3} = \frac{20}{3} = \sigma_{\bar{x}}^2 = \text{LHS}$$

3. *The weights of 1500 ball bearings are normally distributed with a mean of 635 gms. and S.D of 1.36 gms. If 300 random samples of size 36 are drawn from this population, determine the expected mean and S.D of the sampling distribution of means if sampling is done*

(a) with replacement (b) without replacement.

>> Here $N = 1500$, $\mu = 635$, $\sigma = 1.36$, $n = 36$

(a) Expected mean $\mu_{\bar{x}} = \mu = 635$

Expected S.D $\sigma_{\bar{x}} = \sqrt{\sigma^2/n} = \frac{\sigma}{\sqrt{n}} = \frac{1.36}{6} = 0.227$

(b) Expected mean $\mu_{\bar{x}} = \mu = 635$

Expected variance $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left[\frac{N-n}{N-1}\right] = \frac{(1.36)^2}{36}\left[\frac{1500-36}{1500-1}\right] = 0.05$

Thus $\sigma_{\bar{x}} = \sqrt{0.05} = 0.224$

4. *Consider the data as in the previous problem. In the case of random sampling with replacement find how many random samples would have their mean (a) between 634.76 gms and 635.24 gms (b) greater than 635.5 gms (c) less than 634.2 gms (d) less than 634.5 gms or more than 635.24 gms.*

>> We assume that the population is a normal population and hence the sampling distribution of means is also taken to be distributed normally.

The standard normal variate $z = \frac{x-\mu}{\sigma}$ in the case of sampling distribution of means is in the equivalent form

$$z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} \text{ where we have } \mu_{\bar{x}} = 635,\ \sigma_{\bar{x}} = 0.227$$

Hence we have $z = \frac{\bar{x}-635}{0.227}$ ...(1)

(a) Probability of a sample having mean between 634.76 and 635.24 is represented by $P(634.76 < \bar{x} < 635.24)$

Now from (1), if $\bar{x} = 634.76$, $z = \frac{-0.24}{0.227} = -1.06$

if $\bar{x} = 635.24$, $z = \frac{0.24}{0.227} = 1.06$

Hence we have to find $P(-1.06 < z < 1.06)$

ie., $= 2P(0 < z < 1.06)$

$= 2\phi(1.06) = 2(0.3554)$ by using tables.

$= 0.7108$

Thus we have corresponding to 300 samples, the expected number of samples having their mean between 634.76 gms and 635.24 gms is given by

$300 \times 0.7108 = 213.24 = \mathbf{213\ samples}$

(b) To find $P(\bar{x} > 635.5) \times 300$

If $\bar{x} = 635.5$ then $z = \frac{635.5-635}{0.227}$ from (1). That is $z = 2.203$

$$\therefore \quad P(z > 2.203) = P(z > 0) - P(0 < z < 2.203)$$

$$= 0.5 - \phi(2.2)$$

$$= 0.5 - 0.4861 = 0.0139$$

Thus $P(\bar{x} > 635.5) \times 300 = 4.17 = \mathbf{4\ samples}$

(c) To find $P(\bar{x} < 634.2) \times 300$

If $\bar{x} = 634.2$ then $z = \frac{634.2-635}{0.227}$ from (1). That is, $z = -3.52$

$$P(z < -3.52) = P(z > 3.52)$$
$$= P(z > 0) - P(0 < z < 3.52)$$
$$= 0.5 - \phi(3.52)$$
$$= 0.5 - 0.4998 = 0.0002$$

Thus $P(\bar{x} < 634.2) \times 300 = 0.06 \approx$ **0 samples**

(d) To find $\left[P(\bar{x} < 634.5) + P(\bar{x} > 635.24) \right] \times 300$

$\bar{x} = 634.5$ then $z = -2.2$

$\bar{x} = 635.24$ then $z = 1.06$; by using (1).

$$P(z < -2.2) + P(z > 1.06)$$
$$= P(z > 2.2) + P(z > 1.06)$$
$$= \{P(z > 0) - P(0 < z < 2.2)\} + \{P(z > 0) - P(0 < z < 1.06)\}$$
$$= \{0.5 - \phi(2.2)\} + \{0.5 - \phi(1.06)\}$$
$$= 1 - \{\phi(2.2) + \phi(1.06)\}$$
$$= 1 - (0.4861 + 0.3554) = 0.1585$$

Multiplying this value by the sample size 300 we get $47.55 \approx$ **48 samples.**

5. *Certain tubes manufactured by a company have mean life time of 800 hours and S.D of 60 hours. Find the probability that a random sample of 16 tubes taken from the group will have a mean life time*
(a) between 790 hours and 810 hours. (b) less than 785 hours.
(c) more than 820 hours. (d) between 770 hours and 830 hours.

>> By data $\mu = 800$, $\sigma = 60$, $n = 16$

$\therefore \quad \sigma_{\bar{x}} = \sigma/\sqrt{n} = 60/4 = 15$

We have $z = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \dfrac{\bar{x} - 800}{15}$ ... (1)

(a) To find $P(790 < \bar{x} < 810)$

If $\bar{x} = 790$, $z = -0.67$

$\bar{x} = 810$, $z = 0.67$; by using (1).

$$P(-0.67 < z < 0.67) = 2P(0 < z < 0.67)$$
$$= 2\phi(0.67) = 2(0.2486) = 0.4972$$

Thus $P(790 < \bar{x} < 810) = 0.4972$

(b) To find $P(\bar{x} < 785)$

If $\bar{x} = 785$ then $z = -1$ from (1).

$\therefore \quad P(z < -1) = P(z > 1)$

$$= P(z > 0) - P(0 < z < 1)$$
$$= 0.5 - \phi(1) = 0.5 - 0.3413 = 0.1587$$

Thus $P(\bar{x} < 785) = 0.1587$

(c) To find $P(\bar{x} > 820)$

If $\bar{x} = 820$ then $z = 1.33$ from (1)

$\therefore \quad P(z > 1.33) = P(z > 0) - P(0 < z < 1.33)$

ie., $= 0.5 - \phi(1.33) = 0.5 - 0.4082 = 0.0918$

Thus $P(\bar{x} > 820) = 0.0918$

(d) To find $P(770 < \bar{x} < 830)$

If $\bar{x} = 770$ then $z = -2$ and $\bar{x} = 830$ then $z = 2$; from (1).

$\therefore \quad P(-2 < z < 2) = 2P(0 < z < 2) = 2\phi(2) = 2(0.4772) = 0.9544$

Thus $P(770 < \bar{x} < 830) = 0.9544$

Test of significance of proportions

6. *A coin is tossed 1000 times and head turns up 540 times. Decide on the hypothesis that the coin is unbiassed.*

>> Let us suppose that the coin is unbiassed.

p = probability of getting a head in one toss = 1/2

Since $p + q = 1$, $q = 1/2$.

Expected number of heads in 1000 tosses = $np = 1000 \times 1/2 = 500$

Actual number of heads = 540

The difference = $x - np = 540 - 500 = 40$.

Consider $Z = \dfrac{x - np}{\sqrt{npq}} = \dfrac{40}{\sqrt{1000 \times 1/2 \times 1/2}} = 2.53 < 2.58$

Thus we can say that the coin is unbiassed.

7. *Result extracts revealed that in a certain school over a period of five years 725 students had passed and 615 students had failed. Test the hypothesis that success and failure are in equal proportions.*

>> Total number of students = 725 + 615 = 1340

Observed proportion of success = $\frac{725}{1340} = 0.54$

Suppose that success and failure are in equal proportion. Then $p = 1/2$

$\therefore$ difference in proportion = 0.54 − 0.5 = 0.04

S.D proportion of success = $\frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.5 \times 0.5}{1340}} = 0.014$

In terms of proportion $Z = \frac{x - \mu}{\sigma} = \frac{0.04}{0.014} = 2.86 > 2.58$

Thus the hypothesis that success and failure are in equal proportion is rejected.

Aliter : Observed number of successes = 725

Expected number of successes = 1340/2 = 670

$$Z = \frac{x - \mu}{\sqrt{npq}} = \frac{725 - 670}{\sqrt{1340 \times 1/2 \times 1/2}} = 3.005 > 2.58$$

Hence the hypothesis is rejected.

8. *A sample of 900 days was taken in a coastal town and it was found that on 100 days the weather was very hot. Obtain the probable limits of the percentage of very hot weather.*

>> Probability of very hot weather $= p = \frac{100}{900} = \frac{1}{9} \quad \therefore \quad q = \frac{8}{9}$

Probable limits $= p \pm 2.58\sqrt{pq/n}$

$$= 0.111 \pm (2.58)\sqrt{\frac{1}{9} \times \frac{8}{9} \times \frac{1}{900}} = 0.111 \pm 0.027$$

$= 0.084$ and 0.138

Probable limits of very hot weather is 8.4% to 13.8%

9. *In a sample of 500 men it was found that 60% of them had over weight. What can we infer about the proportion of people having over weight in the population ?*

>> Probability of persons having over weight = $p = \frac{60}{100} = 0.6$ & $q = 1 - p = 0.4$

Probable limits = $p \pm 2.58\sqrt{pq/n}$

Probable limits $= 0.6 \pm 2.58\sqrt{\frac{(0.6)(0.4)}{500}}$

Probable limits $= 0.6 \pm 0.0565 = 0.5435$ and 0.6565

Thus the probable limits of people having over weight is 54.35% to 65.65%

10. *A survey was conducted in a slum locality of 2000 families by selecting a sample of size 800. It was revealed that 180 families were illiterates. Find the probable limits of the illiterate families in the population of 2000.*

>> Probability of illiterate families $= p = \frac{180}{800} = 0.225 \quad \therefore \quad q = 0.775$

Probable limits of illiterate families $= p \pm (2.58)\sqrt{pq/n}$

i.e., $= 0.225 \pm (2.58)\sqrt{\frac{(0.225)(0.775)}{800}} = 0.225 \pm 0.038$

$= 0.187$ and 0.263

$\therefore$ the probable limits of illiterate families in the population of 2000 is

$(0.187)\,2000$ and $(0.263)\,2000$

Thus **374 to 526 are probably illiterate families.**

11. *To know the mean weights of all 10 year old boys in Delhi a sample of 225 was taken. The mean weight of the sample was found to be 67 pounds with S.D of 12 pounds. What can we infer about the mean weight of the population ?*

>> Sample mean $(\bar{x}) = 67$, Sample size $n = 225$, S.D $(\sigma) = 12$

95% confidence limits for the mean of the population corresponding to a given sample is $\bar{x} \pm 1.96(\sigma/\sqrt{n})$ and 99% confidence limits for the mean is $\bar{x} \pm 2.58(\sigma/\sqrt{n})$.

We have $\sigma/\sqrt{n} = 12/15 = 0.8$

95% confidence limits : $67 \pm 1.96(0.8) = 65.432$ and 68.568

99% confidence limits : $67 \pm 2.58(0.8) = 64.936$ and 69.064

We can say with 95% confidence that the mean weight of the population lies between 65.4 pounds and 68.6 pounds. Also with 99% confidence we can say that the mean weight lies between 64.9 pounds to 69.1 pounds.

12. *In a hospital 230 females and 270 males were born in a year. Do these figures confirm the hypothesis that sexes are born in equal proportions ?*

>> Total number of births, $n = 230 + 270 = 500$

Observed proportion of females $= \frac{230}{500} = \frac{23}{50} = 0.46.$

assuming that sexes are born in equal proportion, probability of female birth is equal to 1/2.

That is, $p = 1/2$ and hence $q = 1/2$

difference in proportions $= 0.5 - 0.46 = 0.04$

S.D proportion of females $= \dfrac{\sqrt{n\,p\,q}}{n} = \sqrt{p\,q/n}$

$= \sqrt{1/2 \times 1/2 \times 1/500} = 0.0224$

$Z = \dfrac{\text{Difference}}{S.D} = \dfrac{0.04}{0.0224} = 1.786 < 2.58$

Thus we conclude that the figures are conformal with the hypothesis that the sexes are born in equal proportions.

---

13. *In 324 throws of a six faced 'die', an odd number turned up 181 times. Is it reasonable to think that the 'die' is an unbiassed one ?*

Probability of the turn up of an odd number is $p = 3/6 = 1/2$

Hence $q = 1 - p = 1/2$

Expected number of successes $= 1/2 \times 324 = 162$

Observed number of successes $= 181$

difference $= 181 - 162 = 19$

Consider $Z = \dfrac{x - np}{\sqrt{npq}} = \dfrac{19}{\sqrt{324 \times 1/2 \times 1/2}} = \dfrac{19}{9} = 2.11 < 2.58$

Thus we conclude that the die is unbiassed .

---

14. *A 'die' is thrown 9000 times and a throw of 3 or 4 was observed 3240 times. Show that the die cannot be regarded as an unbiassed one.*

Probability of getting 3 or 4 in a single throw is $p = 2/6 = 1/3$ and $q = 1 - p = 2/3$

Expected number of successes $= 1/3 \times 9000 = 3000$

Observed number of successes $= 3240$

difference $= 3240 - 3000 = 240$

Consider $Z = \dfrac{x - np}{\sqrt{npq}} = \dfrac{240}{\sqrt{9000 \times 1/3 \times 2/3}} = \dfrac{240}{\sqrt{2000}} = 5.37$

Since $Z = 5.37 > 2.58$ we conclude that the die is biassed.

---

15. *A sample of 100 days is taken from meteorological records of a certain district and 10 of them are found to be foggy. What are the probable limits of the percentage of foggy days in the district.*

>> $p =$ proportion of foggy days in a sample of 100 days is given by $10/100 = 0.1$

Hence $q = 1 - p = 0.9$

$\therefore$ probable limits of foggy days

$= p \pm 2.58\sqrt{pq/n}$

$= 0.1 \pm 2.58\sqrt{(0.1 \times 0.9)/100}$

$= 0.1 \pm 0.0774 = 0.0226$ and 0.1774

$= 2.26\ \%$ and $17.74\ \%$

Thus the percentage of foggy days lies between 2.26 and 17.74

---

16. *A random sample of 500 apples was taken from a large consignment and 65 were found to be bad. Estimate the proportion of bad apples in the consignment as well as the standard error of the estimate. Also find the percentage of bad apples in the consignment.*

>> $p =$ proportion of bad apples in the sample is given by $65/500 = 0.13$.

Hence $q = 1 - p = 0.87$

S.E proportion of bad apples $= \sqrt{pq/n} = \sqrt{(0.13 \times 0.87)/500} = 0.015$

Probable limits of bad apples in the consignment

$= p \pm 2.58\sqrt{pq/n}$

$= 0.13 \pm 2.58\ (0.015) = 0.13 \pm 0.0387$

$= 0.0913$ and 0.1687

$= 9.13\ \%$ and $16.87\ \%$

Thus the required percentage of bad apples in the consignment lies between 9.13 and 16.87

---

17. *In a locality of 18000 families a sample of 840 families was selected at random. Of these 840 families, 206 families were found to have monthly income of Rs.2500 or less. It was desired to estimate how many of the 18,000 families have monthly income of Rs.2500 or less. Within what limits would you place your estimate.*

>> Proportion of families having monthly income of Rs.2500 or less is given by

$p = 206/840 = 0.245$. Hence $q = 1 - p = 0.755$

S.E proportion $= \sqrt{pq/n} = \sqrt{(0.245 \times 0.755)/840} = 0.015$

Probable limits of families having monthly income of Rs.2500 or less are $p \pm 2.58\sqrt{pq/n}$. That is,

$= 0.245 \pm (2.58)(0.015)$

$= 0.245 \pm 0.0387$

$= 0.2063$ and 0.2837 or 20.63% and 28.37%

Hence the probable limits in respect of 18,000 families is given by

$0.2063 \times 18{,}000$ and $0.2837 \times 18{,}000$

That is 3713.4 and 5106.6 or 3713 and 5107

Thus we say that 3713 to 5107 families are likely to have monthly income of Rs.2500 or less.

Remark : *We have said that the value 2.58 is the confidencce coefficient corresponding to the 99 % confidence level. However we can even take this value to be equal to 3 corresponding to 99.73 % confidence level so that the probable limits can be taken as* $p \pm 3\sqrt{pq/n}$

Accordingly we have the following answers in the previous 3 problems.

In problem - 15, we have, $0.1 \pm 3(0.03) = 0.01$ *to* 0.19 *or* 1% *to* 19%

In problem - 16, we have, $0.13 \pm 3(0.015) = 0.085$ *to* 0.175 *or* 8.5 *to* 17.5%

In problem - 17, we have, $0.245 \pm 3(0.015) = 0.2$ *to* 0.29

Multiplying by 18, 000 we get 3600 to 5220.

---

18. *The mean and S.D of the maximum loads supported by 60 cables are 11.09 tonnes and 0.73 tonnes respectively. Find* **(a)** *95%* **(b)** *99% confidence limits for mean of the maximum loads of all cables produced by the company.*

>> By data $\bar{x} = 11.09$, $\sigma = 0.73$

(a) 95 % confidence limits for the mean of maximum loads are given by

$\bar{x} \pm 1.96(\sigma/\sqrt{n})$

$= 11.09 \pm 1.96(0.73/\sqrt{60})$

$= 11.09 \pm 0.18$

Thus 10.91 tonnes to 11.27 tonnes are the 95 % confidence limits for the mean of maximum loads.

(b) 99 % confidence limits for the mean of maximum loads are given by

$\bar{x} \pm 2.58(\sigma/\sqrt{n})$

$= 11.09 \pm 2.58(0.73/\sqrt{60})$

$= 11.09 \pm 0.24 = 10.85$ and 11.33

Thus 10.85 tonnes to 11.33 tonnes are the 99 % confidence limits for the mean of maximum loads.

---

19. *The mean and S.D of the diameters of a sample of 250 rivet heads manufactured by a company are 7.2642 mm and 0.0058 mm respectively. Find* **(a)** *99 %* **(b)** *98 %* **(c)** *95 %* **(d)** *90 %* **(e)** *50 % confidence limits for the mean diameter of all the rivet heads manufactured by the company.*

>> By data $\bar{x} = 7.2642$, $\sigma = 0.0058$, $n = 250$

Confidence limits for the mean is given by $\bar{x} \pm (z_c)(\sigma/\sqrt{n})$ where z_c is the confidence coefficient corresponding to the confidence level. We have the following from the normal probability tables.

Confidence level	99%	98%	95%	90%	50%
z_c	2.58	2.33	1.96	1.645	0.6745

Now $\dfrac{\sigma}{\sqrt{n}} = \dfrac{0.0058}{\sqrt{250}} = 0.00037$

Confidence limits for various confidence level respectively are as follows.

(a) $7.2642 \pm 2.58(0.00037) = 7.2642 \pm 0.00095$

(b) $7.2642 \pm 2.33(0.00037) = 7.2642 \pm 0.00086$

(c) $7.2642 \pm 1.96(0.00037) = 7.2642 \pm 0.00073$

(d) $7.2642 \pm 1.645(0.00037) = 7.2642 \pm 0.00061$

(e) $7.2642 \pm 0.6745(0.00037) = 7.2642 \pm 0.00025$

---

20. *An unbiassed coin is thrown n times. It is desired that the relative frequency of the appearence of heads should lie between 0.49 and 0.51. Find the smallest value of n that will ensure this result with* **(a)** *95% confidence* **(b)** *90% confidence*

>> p = probability of getting a head $= 1/2$; $q = 1 - p = 1/2$

S.E proportion of heads $= \sqrt{pq/n} = \sqrt{1/4n} = (1/2)\sqrt{n}$

(a) Probable limits for 95% confidence level is given by $p \pm 1.96\sqrt{pq/n}$ which should be 0.51 and 0.49

ie $0.5 \pm 1.96(1/2\sqrt{n}) = 0.51$ or 0.49

ie., $0.5 + \dfrac{1.96}{2\sqrt{n}} = 0.51$ and $0.5 - \dfrac{1.96}{2\sqrt{n}} = 0.49$

$\Rightarrow \dfrac{1.96}{2\sqrt{n}} = 0.01$ or $\sqrt{n} = \dfrac{1.96}{0.02} = 98$

Thus $n = 9604$

(b) Taking the confidence coefficient equal to 1.645 for 90 % confidence level we have as before

$$\frac{1.645}{2\sqrt{n}} = 0.01 \text{ or } \sqrt{n} = \frac{1.645}{0.02} = 82.25 \quad \text{or} \quad n = 6765.0625 \approx 6765 .$$

Thus $n = 6765$

Test of significance of a sample mean

21. *A manufacturer claimed that atleast 95% of the equipment which he supplied to a factory conformed to specifications . An examination of a sample of 200 pieces of equipment revealed that 18 of them were faulty. Test his claim at a significance level of 1% and 5%.*

>> Let p be the probability of success which being the probability of the equipment supplied to the factory conformal to the specifications.

$\therefore$ by data $p = 0.95$ and hence $q = 0.05$

$H_0 : p = 0.95$ and the claim is correct.

$H_1 : p < 0.95$ and the claim is false.

We choose the one tailed test to determine whether the supply is conformal to the specification.

$$\mu = np = 200 \times 0.95 = 190$$

$$\sigma = \sqrt{npq} = \sqrt{200 \times 0.95 \times 0.05} = 3.082$$

Expected number of equipments according to the specification= 190

Actual number = 182 since 18 out of 200 were faulty

$\therefore$ difference = $190 - 182 = 8$

Now $Z = \dfrac{x - np}{\sqrt{npq}} = \dfrac{8}{3.082} = 2.6$

The value of Z is greater than the critical value 1.645 at 5% level and 2.33 at 1% level of significance. The claim of the manufacturer (*null hypothesis that the claim is correct*) is rejected at 5% as well as at 1% level of significance in accordence with the one tailed test.

22. *It has been found from experience that the mean breaking strength of a particular brand of thread is 275.6 gms with standard deviation of 39.7 gms. Recently a sample of 36 pieces of thread showed a mean breaking strength of 253.2 gms. Can one conclude at a significance level of (a) 0.05 (b) 0.01 that the thread has become inferior?*

>> We have to decide between the two hypothesis

$H_0 : \mu = 275.6$ gms, mean breaking strength

$H_1 : \mu < 275.6$ gms, inferior in breaking strength.

We choose the one tailed test.

Mean breaking strength of a sample of 36 pieces = 253.2

$\therefore$ difference= $275.6 - 253.2 = 22.4$; $n = 36$

$$Z = \frac{\text{difference}}{(\sigma/\sqrt{n})} = \frac{22.4}{39.7/6} = 3.38$$

The value of Z is greater than the critical value of Z = 1.645 at 5% level and 2.33 at 1% level of significance.

Under the hypothesis H_1 that the thread has become inferior is accepted at both 0.05 and 0.01 levels in accordence with one tailed test.

23. *In an examination given to students at a large number of different schools the mean grade was 74.5 and S.D grade was 8. At one particular school where 200 students took the examination the mean grade was 75.9. Discuss the significance of this result from the view point of (a) one tailed test (b) two tailed test at both 5% and 1% level of significance.*

>> $H_0 : \mu = 74.5$ and there is no change in the mean grade.

$H_1 : \mu \neq 74.5$ *i.e.*, $\mu > 74.5$ and $\mu < 74.5$.

$\mu = 74.5$ and mean of a sample of size 200 (n) is 75.9

$\therefore$ difference = $75.9 - 74.5 = 1.4$

$$Z = \frac{\text{difference}}{(\sigma/\sqrt{n})} = \frac{1.4}{8/\sqrt{200}} = 2.475$$

We have the table for the critical values of Z in the case of one and two tailed tests.

Test	$Z_{0.05}$	$Z_{0.01}$
One tailed	± 1.645	± 2.33
Two tailed	± 1.96	± 2.58

The calculated value of Z is more than $Z_{0.05}$, $Z_{0.01}$ in one tailed test as well as $Z_{0.05}$ in two tailed test.

Thus we conclude that the difference in the mean grade is significant in these tests but the same is not significant in the two tailed test at 1% level of significance.

Test of significance of difference between means

24. *In an elementary school examination the mean grade of 32 boys was 72 with a standard deviation of 8, while the mean grade of 36 girls was 75 with a standard deviation of 6. Test the hypothesis that the performance of girls is better than boys.*

>> We have $\bar{x}_B = 72,\ \sigma_B = 8,\ n_B = 32$ [Boys]

$\bar{x}_G = 75,\ \sigma_G = 6,\ n_G = 36$ [Girls]

Consider $Z = \dfrac{(\bar{x}_G - \bar{x}_B)}{\sqrt{\sigma_G^2/n_G + \sigma_B^2/n_B}}$

$= \dfrac{(75-72)}{\sqrt{36/36 + 64/32}} = \dfrac{3}{\sqrt{3}} = \sqrt{3} = 1.73$

$\therefore \quad Z = 1.73 \begin{cases} > Z_{.05} = 1.645 \text{ (one tailed test)} \\ < Z_{.01} = 2.33 \text{ (one tailed test)} \end{cases}$

The difference in the performance of girls and boys in the examination is significant at 5% level but not at 1% level.

25. *A sample of 100 bulbs produced by a company A showed a mean life of 1190 hours and a standard deviation of 90 hours. Also a sample of 75 bulbs produced by a company B showed a mean life of 1230 hours and a standard deviation of 120 hours. Is there a difference between the mean life time of the bulbs produced by the two companies at*
(a) 5% level of significance (b) 1% level of significance.

>> By data $\bar{x}_A = 1190,\ \sigma_A = 90,\ n_A = 100$ [Company A]

$\bar{x}_B = 1230,\ \sigma_B = 120,\ n_B = 75$ [Company B]

Consider $Z = \dfrac{(\bar{x}_B - \bar{x}_A)}{\sqrt{\sigma_B^2/n_B + \sigma_A^2/n_A}}$

$= \dfrac{(1230-1190)}{\sqrt{(120)^2/75 + (90)^2/100}} = 2.42$

$Z = 2.42 \begin{cases} > Z_{.05} = 1.96 \text{ (two tailed test)} \\ < Z_{.01} = 2.58 \text{ (two tailed test)} \end{cases}$

The null hypothesis that there is no difference between the mean life time of bulbs is rejected at 5% level but not at 1% level of significance.

The null hypothesis is rejected at both the levels of significance in a one tailed test as the respective critical values are 1.645 and 2.33.

26. *The mean of two large samples of 1000 and 2000 members are 168.75 cms and 170 cms respectively. Can the samples be regarded as drawn from the same population of standard deviation 6.25 cms. ?*

>> $\bar{x}_1 = 168.75,\ \bar{x}_2 = 170$

$n_1 = 1000,\ n_2 = 2000$

$Z = \dfrac{\bar{x}_2 - \bar{x}_1}{\sigma\sqrt{1/n_1 + 1/n_2}}$

$= \dfrac{1.25}{6.25\sqrt{1/1000 + 1/2000}} = 5.16$

$Z = 5.16$ is very much greater than $Z_{.05} = 1.96$ and also $Z_{.01} = 2.58$. Thus we say that the difference between the sample means is significant and we conclude that **the samples cannot be regarded as drawn from the same population.**

27. *A random sample for 1000 workers in company has mean wage of Rs 50 per day and SD of Rs 15. Another sample of 1500 workers from another company has mean wage of Rs 45 per day and S.D of Rs 20. Does the mean rate of wages varies between the two companies ? Find the 95 % confidence limits for the difference of the mean wages of the population of the two companies.*

>> Company - 1 : $\bar{x}_1 = 50,\ \sigma_1 = 15,\ n_1 = 1000$

Company - 2 : $\bar{x}_2 = 45,\ \sigma_2 = 20,\ n_2 = 1500$

$Z = \dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$

ie $Z = \dfrac{5}{\sqrt{225/1000 + 400/1500}} = 7.1307$

$Z = 7.1307$ is greater than $Z_{.05} = 1.96$ and $Z_{.01} = 2.58$.

Hence we can say that the difference between the mean wages is significant both at 5 % and 1 % levels of significance.

Also 95 % confidence limits for the difference of mean wages is given by

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96 \sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}$$

$$= 5 \pm 1.96\,(0.7012)$$

$$= 5 \pm 1.374$$

$$= 3.626 \text{ and } 6.374 \text{ or } 3.63 \text{ and } 6.37 \text{ approximately.}$$

Thus we can say with 95 % confidence that the difference of population mean of wages between the two companies lies between Rs 3.63 and Rs 6.37

Test of significance for difference of properties.

28. *In an exit poll enquiry it was revealed that 600 voters in one locality and 400 voters from an other locality favoured 55% and 48% respectively a particular party to come to power. Test the hypothesis that there is a difference in the locality in respect of the opinion.*

>> By data, $p_1 = \dfrac{55}{100} = 0.55$ (First locality)

$$p_2 = \frac{48}{100} = 0.48 \text{ (Second locality)}$$

H_0 is the null hypothesis that there is no difference in the locality.

Population proportion $p = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ where $n_1 = 600,\ n_2 = 400$

$$\therefore \quad p = \frac{600\,(0.55) + 400\,(0.48)}{600 + 400} = 0.522$$

Also $q = 1 - p = 0.478$

Consider $Z = \dfrac{p_1 - p_2}{\sqrt{pq\,(1/n_1 + 1/n_2)}}$

i.e., $$Z = \frac{0.55 - 0.48}{\sqrt{(0.522)\,(0.478)\,(1/600 + 1/400)}} = 2.171$$

$$Z = 2.171 \begin{cases} > Z_{.05} = 1.96 \text{ (Two tailed test)} \\ < Z_{.01} = 2.58 \text{ (Two tailed test)} \end{cases}$$

Thus the null hypothesis that there is no difference between the localities is rejected at 5% level but not at 1% level of significance.

29. *One type of aircraft is found to develop engine trouble in 5 flights out of a total of 100 and another type in 7 flights out of a total of 200 flights. Is there a significant difference in the two types of aircrafts so far as engine defects are concerned?*

>> Let p_1 and p_2 be the proportion of defects in the two types of aircrafts

$$\therefore \quad p_1 = \frac{5}{100} = 0.05,\ p_2 = \frac{7}{200} = 0.035$$

H_0 is the null hypothesis that there is no significant difference between the two type of aircrafts.

Combined proportion $= p = \dfrac{5+7}{100+200} = \dfrac{12}{300} = 0.04$ and $q = 1 - p = 0.96$

Consider $Z = \dfrac{p_1 - p_2}{\sqrt{pq\,(1/n_1 + 1/n_2)}}$

$$= \frac{0.05 - 0.035}{\sqrt{(0.04)\,(0.96)\left(1/100 + 1/200\right)}} = 0.625$$

$$Z = 0.625 \begin{cases} < Z_{.05} = 1.96 \text{ (two tailed test)} \\ < Z_{.01} = 2.58 \text{ (two tailed test)} \end{cases}$$

Thus the null hypothesis is accepted both at 5 % and 1 % levels of significance.

30. *Random sample of 1000 engineering students from a city A and 800 from city B were taken. It was found that 400 students in each of the sample were from payment quota. Does the data reveal a significant difference between the two cities in respect of payment quota students?*

>> $n_1 = 1000,\ n_2 = 800$

$$p_1 = \frac{400}{1000} = 0.4\ ;\ p_2 = \frac{400}{800} = 0.5$$

$$\therefore \quad p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{1000\,(0.4) + 800\,(0.5)}{1800} = \frac{4}{9}$$

$$\therefore \quad q = 1 - p = 5/9$$

Let H_0 be the null hypothesis that there is no significant difference between the 2 cities.

Consider $Z = \dfrac{p_2 - p_1}{\sqrt{p\,q\,(1/n_1 + 1/n_2)}}$

$$Z = \frac{0.1}{\sqrt{4/9 \times 5/9\,(1/1000 + 1/800)}} = 4.243$$

$$Z = 4.243 > \begin{cases} Z_{.05} = 1.96 \\ Z_{.01} = 2.58 \end{cases}$$

Thus the hypothesis H_0 is rejected both at 5% and 1% levels of significance.

Student's t distribution / test

31. *Find the student's t for the following variable values in a sample of eight :* $-4, -2, -2, 0, 2, 2, 3, 3$, *taking the mean of the universe to be zero.*

>> $t = \dfrac{\bar{x} - \mu}{s}\sqrt{n}$

By data $\mu = 0$ and we have $n = 8$

$$\bar{x} = \frac{1}{8}(-4-2-2+0+2+2+3+3) = \frac{1}{4} = 0.25$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \frac{1}{7}\{(-4.25)^2 + (-2.25)^2 + (-2.25)^2 + (-0.25)^2 + (1.75)^2 + (1.75)^2 + (2.75)^2 + (2.75)^2\}$$

$$s^2 = \frac{1}{7}(49.5) = 7.07 \qquad \therefore \quad s = 2.66$$

Thus $t = \dfrac{0.25 - 0}{2.66}\sqrt{8} = 0.266$

Note : *The expression for s^2 can also be put in the following form.*

$$s^2 = \frac{1}{n-1}\left\{\sum_1^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2)\right\}$$

$$= \frac{1}{n-1}\left\{\sum_1^n x_i^2 - 2\sum_1^n x_i \cdot \frac{1}{n}\sum_1^n x_i + n\left(\frac{1}{n}\sum x_i\right)^2\right\}$$

$$= \frac{1}{n-1}\left\{\sum_1^n x_i^2 - \frac{2}{n}(\sum x_i)^2 + \frac{1}{n}(\sum x_i)^2\right\}$$

$$\therefore \qquad s^2 = \frac{1}{n-1}\left\{\sum_1^n x_i^2 - \frac{1}{n}(\sum x_i)^2\right\}$$

According to this formula we have in the given example

$$s^2 = \frac{1}{7}\left\{50 - \frac{1}{8}(2)^2\right\} = \frac{1}{7}(49.5) = 7.07$$

Remark : *We can employ this formula when $\bar{x}$ is not an integer.*

32. *A machine is expected to produce nails of length 3 inches. A random sample of 25 nails gave an average length of 3.1 inch with standard deviation 0.3. Can it be said that the machine is producing nails as per specification?* ($t_{0.05}$ *for 24 d.f is 2.064*)

>> By data we have,

$\mu = 3$, $\bar{x} = 3.1$, $n = 25$, $s = 0.3$

$$t = \frac{\bar{x} - \mu}{s}\sqrt{n} = \frac{0.1}{0.3}\sqrt{25} = 1.67 < 2.064$$

Thus the hypothesis that the machine is producing nails as per specification is accepted at 5% level of significance.

33. *Ten individuals are chosen at random from a population and their heights in inches found to be 63, 63, 66, 67, 68, 69, 70, 70, 71, 71. Test the hypothesis that the mean height of the universe is 66 inches.* ($t_{.05} = 2.262$ *for 9 d.f*)

>> We have $\mu = 66$, $n = 10$

$$\bar{x} = \frac{\sum x}{n} = \frac{678}{10} = 67.8$$

$$s^2 = \frac{1}{n-1}\sum(x - \bar{x})^2$$

$$s^2 = \frac{1}{9}\left[(63 - 67.8)^2 + \cdots + (71 - 67.8)^2\right] = 9.067 \qquad \therefore \quad s = 3.011$$

We have $t = \dfrac{\bar{x} - \mu}{s}\sqrt{n} = \dfrac{(67.8 - 66)}{3.011}\sqrt{10} = 1.89 < 2.262$

Thus the hypothesis is accepted at 5% level of significance.

34. *A sample of 10 measurements of the diameter of a sphere gave a mean of 12 and standard deviation 0.15cm. Find the 95% confidence limits for the actual diameter.*

>> By data $n = 10$, $\bar{x} = 12$, $s = 0.15$

Also $t_{.05}$ for $9\,d.f = 2.262$

Confidence limits for the actual diameter is given by

$$\bar{x} \pm \left[\frac{s}{\sqrt{n}}\right] t_{.05} = 12 \pm \frac{0.15}{\sqrt{10}}(2.262) = 12 \pm 0.1073$$

Thus 11.893cm to 12.107cm is the confidence limits for the actual diameter

35. *A certain stimulus administered to each of the 12 patients resulted in the following change in blood pressure. 5, 2, 8, −1, 3, 0, 6, −2, 1, 5, 0, 4. Can it be concluded that the stimulus will increase the blood pressure? ($t_{.05}$ for 11 d.f = 2.201)*

$$\bar{x} = \frac{\sum x}{n} = \frac{31}{12} = 2.5833$$

$$s^2 = \frac{1}{n-1}\sum(x-\bar{x})^2 = \frac{1}{n-1}\left\{\sum x^2 - \frac{1}{n}(\sum x)^2\right\}$$

$$s^2 = \frac{1}{11}\left\{185 - \frac{1}{12}(31)^2\right\} = 9.538 \quad \therefore \quad s = 3.088$$

We have, $t = \dfrac{\bar{x}-\mu}{s}\sqrt{n}$

Let us suppose that the stimulus administration is not accompanied with increase in blood pressure, we can take $\mu = 0$

$$\therefore \quad t = \frac{2.5833 - 0}{3.088}\sqrt{12} = 2.8979 = 2.9 > 2.201$$

Hence the hypothesis is rejected at 5% level of significance. We conclude with 95% confidence that the stimulus in general is accompanied with increase in blood pressure.

36. *A group of boys and girls were given an intelligence test. The mean score, S.D score and numbers in each group are as follows.*

	Boys	Girls
Mean	74	70
SD	8	10
n	12	10

Is the difference between the means of the two groups significant at 5% level of significance ($t_{.05}$ = 2.086 for 20 d.f)

>> We have by data $\bar{x} = 74$, $s_1 = 8$, $n_1 = 12$ [Boys]

$\bar{y} = 70$, $s_2 = 10$, $n_2 = 10$ [Girls]

Also we have $t = \dfrac{\bar{x}-\bar{y}}{s\sqrt{1/n_1 + 1/n_2}}$

where $s^2 = \dfrac{1}{n_1+n_2-2}\left\{\sum_{i=1}^{n_1}(x_i-\bar{x})^2 + \sum_{j=1}^{n_2}(y_j-\bar{y})^2\right\}$

or $s^2 = \dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

Now $s^2 = \dfrac{12(64) + 10(100)}{20} = \dfrac{1768}{20} = 88.4 \quad \therefore \quad s = 9.402 = 9.4$

Hence $t = \dfrac{74-70}{9.4\sqrt{1/12 + 1/10}} = 0.994$

$t = 0.994 < t_{.05} = 2.086$

Thus the hypothesis that there is a difference between the means of the two groups is accepted at 5% level of significance.

37. *A sample of 11 rats from a central population had an average blood viscosity of 3.92 with a standard deviation of 0.61. On the basis of this sample, establish 95% fiducial limits for μ the mean blood viscosity of the central population ($t_{.05}$ = 2.228 for 10 d.f)*

>> By data $\bar{x} = 3.92$, $s = 0.61$, $n = 11$

95% fiducial limits for μ are $\bar{x} \pm \dfrac{s}{\sqrt{n}} t_{.05}$

i.e., $= 3.92 \pm \dfrac{0.61}{\sqrt{11}}(2.228)$

$= 3.92 \pm 0.41 = 3.51$ and 4.33

Thus 95% confidence limits for μ are 3.51 and 4.33.

38. *Two types of batteries are tested for their length of life and the following results were obtained.*

Battery A: $n_1 = 10$, $\bar{x}_1 = 500$ *hrs*, $\sigma_1^2 = 100$

Battery B: $n_2 = 10$, $\bar{x}_2 = 500$ *hrs*, $\sigma_2^2 = 121$

Compute Student's t and test whether there is a significant difference in the two means.

>> $$s^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2 - 2}$$

$$s^2 = \frac{(10 \times 100) + (10 \times 121)}{18} = 122.78 \quad \therefore \quad s = 11.0805$$

We have, $t = \dfrac{(\bar{x}_2 - \bar{x}_1)}{s\sqrt{1/n_1 + 1/n_2}}$

$$t = \frac{60}{11.0805\sqrt{0.1+0.1}} = 12.1081 \sim 12.11$$

This value of t is greater than the table value of t for 18 $d \cdot f$ at all levels of significance.

The null hypothesis that there is no significant difference in the two means is rejected at all significance levels.

39. *A group of 10 boys fed on a diet A and another group of 8 boys fed on a different diet B for a period of 6 months recorded the following increase in weights (lbs.)*

Diet A: 5 6 8 1 12 4 3 9 6 10

Diet B: 2 3 6 8 10 1 2 8

Test whether diets A and B differ significantly regarding their effect on increase in weight.

>> Let the variable x correspond to the diet A and y to the diet B.

$$\bar{x} = \frac{\sum x}{n_1} = \frac{64}{10} = 6.4 \; ; \; \bar{y} = \frac{\sum y}{n_2} = \frac{40}{8} = 5$$

$$\sum_1^{n_1} (x-\bar{x})^2 = 102.4 \; ; \; \sum_1^{n_2} (y-\bar{y})^2 = 82$$

$$s^2 = \frac{1}{n_1+n_2-2}\left\{\sum_1^{n_1}(x-\bar{x})^2 + \sum_1^{n_2}(y-\bar{y})^2\right\}$$

$$s^2 = \frac{1}{16}(102.4+82) = \frac{184.4}{16} = 11.525 \quad \therefore \quad s = 3.395$$

Consider $t = \dfrac{\bar{x}-\bar{y}}{s\sqrt{1/n_1 + 1/n_2}}$

$$t = \frac{1.4}{3.395\sqrt{1/10+1/8}} = 0.86505 \simeq 0.87$$

But $t_{.05}$ for 16 $d \cdot f = 2.12$ from the tables. $t = 0.87$ is less than the table value for 16 $d \cdot f$ at 5% level of significance.

Thus we conclude that the two diets do not differ significantly regarding their effect on increase in weight.

40. *Two horses A and B were tested according to the time (in seconds) to run a particular race with the following results.*

Horse A: 28 30 32 33 33 29 34

Horse B: 29 30 30 24 27 29

Test whether you can discriminate between the two horses.

>> Let the variables x and y respectively correspond to horse A and horse B.

$$\bar{x} = \frac{\sum x}{n_1} = \frac{219}{7} = 31.3, \; \bar{y} = \frac{\sum y}{n_2} = \frac{169}{6} = 28.2$$

$$\sum_1^{n_1}(x-\bar{x})^2 = 31.43, \; \sum_1^{n_2}(y-\bar{y})^2 = 26.84$$

$$s^2 = \frac{1}{n_1+n_2-2}\left\{\sum_1^{n_1}(x-\bar{x})^2 + \sum_1^{n_2}(y-\bar{y})^2\right\}$$

$$s^2 = \frac{1}{11}(31.43+26.84) = 5.2973 \quad \therefore \quad s = 2.3016$$

Consider $t = \dfrac{\bar{x}-\bar{y}}{s\sqrt{1/n_1 + 1/n_2}}$

$$t = \frac{(31.3-28.2)}{2.3016\sqrt{1/7+1/6}} = 2.42$$

But $t_{.05} = 2.2$ and $t_{.02} = 2.72$ for 11 $d \cdot f$

$$t = 2.42 \begin{cases} > t_{.05} = 2.2 \\ < t_{.02} = 2.72 \end{cases}$$

The discrimination between the horses is significant at 5% level but not at 2% level of significance.

Chi-Square distribution

41. *A die is thrown 264 times and the number appearing on the face (x) follows the following frequency distribution.*

x	1	2	3	4	5	6
f	40	32	28	58	54	60

Calculate the value of χ^2.

>> The frequencies in the given data are the observed frequencies. Assuming that the dice is unbiassed the expected number of frequencies for the numbers 1, 2, 3, 4, 5, 6 to appear on the face is $\frac{264}{6} = 44$ each. Now the data is as follows :

No. on the dice	1	2	3	4	5	6
Observed frequency (0_i)	40	32	28	58	54	60
Expected frequency (E_i)	44	44	44	44	44	44

$$\chi^2 = \sum \frac{(0_i - E_i)^2}{E_i}$$

$$= \frac{(40-44)^2}{44} + \frac{(32-44)^2}{44} + \cdots + \frac{(60-44)^2}{44}$$

$$= \frac{1}{44}\left[16 + 144 + 256 + 196 + 100 + 256\right] = \frac{968}{44} = 22$$

Thus $\chi^2 = 22$

42. *Five dice were thrown 96 times and the numbers 1, 2 or 3 appearing on the face of the dice follows the frequency distribution as below.*

No. of dice showing 1,2 or 3	5	4	3	2	1	0
Frequency	7	19	35	24	8	3

Test the hypothesis that the data follows a binomial distribution. ($\chi^2_{0.05} = 11.07$ *for 5 d.f*)

>> The data gives the observed frequencies and we need to calculate the expected frequencies.

Probability of a single dice throwing 1, 2 or 3 is $p = 3/6 = 1/2 \quad \therefore \quad q = 1 - p = 1/2$

The binomial distribution of fit is, $N(q+p)^n = 96\left(1/2 + 1/2\right)^5$

The theoretical frequencies of getting 5, 4, 3, 2, 1, 0 successes with 5 dice are respectively the successive terms of the binomial expansion.

They are respectively $96 \times \frac{1}{2^5}, 96 \times 5C_1 \times \frac{1}{2^5}, \ldots 96 \times \frac{1}{2^5}$ or 3, 15, 30, 30, 15, 3.

We have the table of observed and expected frequencies.

0_i	7	19	35	24	8	3
E_i	3	15	30	30	15	3

$$\chi^2 = \sum \frac{(0_i - E_i)^2}{E_i}$$

$$= \frac{16}{3} + \frac{16}{15} + \frac{25}{30} + \frac{36}{30} + \frac{49}{15} + \frac{0}{3} = 11.7$$

$$\chi^2 = 11.7 > \chi^2_{0.05} = 11.07$$

Thus the hypothesis that the data follows a binomial distribution is rejected.

43. *A sample analysis of examination results of 500 students was made. It was found that 220 students had failed, 170 had secured third class 90 had secured second class and 20 had secured first class. Do these figures support the general examination result which is in the ratio*
4 : 3 : 2:1 for the respective categories ($\chi^2_{0.05} = 7.81$ *for 3 d.f*)

>> Let us take the hypothesis that these figures support to the general result in the ratio 4 : 3 : 2 : 1.

The expected frequencies in the respective category are

$$\frac{4}{10} \times 500, \frac{3}{10} \times 500, \frac{2}{10} \times 500, \frac{1}{10} \times 500 \quad \text{or} \quad 200, 150, 100, 50.$$

We have the following table.

O_i	220	170	90	20
E_i	200	150	100	50

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{400}{200} + \frac{400}{150} + \frac{100}{100} + \frac{900}{50}$$

$$\chi^2 = 23.67 > \chi^2_{0.05} = 7.81$$

Thus the hypothesis is rejected.

---

48. *4 coins are tossed 100 times and the following results were obtained. Fit a binomial distribution for the data and test the goodness of fit ($\chi^2_{0.05} = 9.49$ for 4 d.f)*

Number of heads	0	1	2	3	4
Frequency	5	29	36	25	5

>> Referring to Problem-28 in Module - 4, we have obtained the theoretical frequencies equal to 7, 26, 37, 24, 6 respectively.

We have the following table.

O_i	5	29	36	25	5
E_i	7	26	37	24	6

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{4}{7} + \frac{9}{26} + \frac{1}{37} + \frac{1}{24} + \frac{1}{6} = 1.15$$

$$\chi^2 = 1.15 < \chi^2_{0.05} = 9.49$$

Thus the hypothesis that the fitness is good can be accepted.

---

45. *Fit a Poisson distribution for the following data and test the goodness of fit given that $\chi^2_{0.05} = 7.815$ for 3 d.f*

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

>> Referring to Problem-32 in Module - 4 we have obtained the theoretical frequencies equal to 121, 61, 15, 3, 0. Since the last of the expected frequency is 0 we shall club it with the previous one.

We have the following table.

O_i	122	60	15	2+1 = 3
E_i	121	61	15	3+0 = 3

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{1}{121} + \frac{1}{61} + 0 + 0 = 0.025$$

$\chi^2 = 0.025 < \chi^2_{0.05} = 7.815$. The fitness is considered good.

Thus the hypothesis that the fitness is good can be accepted.

---

46. *The number of accidents per day (x) as recorded in a textile industry over a period of 400 days is given below. Test the goodness of fit in respect of Poisson distribution of fit to the given data ($\chi^2_{0.05} = 9.49$ for 4 d.f)*

x	0	1	2	3	4	5
f	173	168	37	18	3	1

>> Referring to the Problem-33 in Module-4, the corresponding theoretical frequencies are 183, 143, 56, 15, 3, 0. We shall club the last two frequencies to have the following table.

O_i	173	168	37	18	3+1 = 4
E_i	183	143	56	15	3+0 = 3

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{100}{183} + \frac{625}{143} + \frac{361}{56} + \frac{9}{15} + \frac{1}{3} = 12.297 \approx 12.3$$

$\chi^2 = 12.3 > \chi^2_{0.05} = 9.49$. The fitness is not good.

Thus the hypothesis that the fitness is good is rejected.

EXERCISES

1. A random sample of size 2 is drawn from the population 3, 4, 5. Find the sampling distribution of the sample mean. (a) with replacement (b) without replacement. Find the sample mean and sample variance in these two cases.
2. 500 ball bearings have a mean weight of 142.30 gms. and S.D of 8.5 gms. Find the probability that a random sample of 100 ball bearings chosen from this group will have a combined weight (a) between 14,061 and 14,175 gms. (b) more than 14,460 gms.
3. The weights of packages received by a department store have a mean of 136 kgs. and a S.D of 22.5 kgs. What is the probability that 25 packages received at random and loaded on an elevator will exceed the safety limit of the elevator quoted as 3720 kgs.
4. A 'die' was thrown 9000 times and a throw of 5 or 6 was obtained 3240 times. On the assumption of random throwing, do the data indicate an unbiassed die ?
5. A sample of 900 days is taken from meteorological records of a certain district and 100 of them are found to be foggy. Find the 99.73% confidence level probable limits of the percentage of foggy days in the district.
6. The mean and S.D marks of a sample of 100 students are 67.45 and 2.92 respectively. Find a) 95% b)99% confidence intervals for estimating the mean marks of the population.
7. The mean of samples of size 1000 and 2000 are 67.5 cms. and 68 cms. respectively. Can the samples be regarded as drawn from the same population of S.D 2.5 cms ?
8. A machine produced 20 defective units in a sample of 400. After over oiling the machine it produced 10 defective in a batch of 300. Has the machine improved due to over oiling ?
9. Ten individuals are chosen at random from a population and their heights in inches are found to be 63, 63, 64, 65, 66, 69, 69, 70, 70, 71. Discuss the suggestion that the mean height of the population is 65 inches given that $t_{.05} = 2.262$ for 9 d.f

10. From a random sample of 10 pigs fed on diet A, the increase in weight in a certain period were 10, 6, 16, 17, 13, 12, 8, 14, 15, 9 lbs. For another random sample of 12 pigs fed on diet B, the increase in weight in the same period were 7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 7 lbs. Test whether diets A and B differ significantly regarding their effect on increase in weight. ($t_{.05}$ for 20 d.f is equal to 2.09)
11. 4 coins were tossed 160 times and the following results were obtained.

No. of heads	0	1	2	3	4
frequency	17	52	54	31	6

Test the goodness of fit of the binomial distribution.
($\chi^2_{0.05} = 9.49$ for 4 d.f)

12. Fit a Poisson distribution for the following data and test the goodness of fit given that $\chi^2_{0.05} = 9.49$ for 4 d.f

x	0	1	2	3	4
f	419	352	154	56	19

ANSWERS

1. (a) 4, 1/3 (b) 4, 1/6
2. (a) 0.2222 (b) 0.0013
3. 0.0023
4. $Z = 5.4$ and the hypothesis is rejected at 1% level of significance.
5. 7.96% to 14.26%
6. (a) 66.88 and 68.02 (b) 66.7 and 68.2
7. $Z = 5.1$; Samples cannot be regarded as drawn from the same population.
8. $Z = 0.4254$. Hypothesis is accepted at 5% level of significance.
9. $t = 2.02$; Hypothesis is accepted at 5% level of significance.
10. $t = 1.6$, difference is not significant. The two diets do not differ significantly regarding increase in weight.
11. Fitness is not good.
12. Fitness is good.

5.2 Stochastic Process

We have already stated that in a random experiment, if a real variable X is associated with every out come then it is called a random variable or stochastic variable. This is equivalent to, having a function on the sample space S and this function is called a random function or a stochastic function. In this article we discuss a stochastic process called the *Markov process* which is such that the generation of the probability distributions depend only on the present state. Before we take up the actual discussion of this Markov process we present some basic definitions and concepts relating to stochastic process.

• Classification of Stochastic Processes

Let S be the sample space of a random experiment and R be the set of all real numbers. A random variable X is a function f from S to R *i.e.*, $X = f(s), s \in S$. We define an *index set* $T \subset R$ indexed by the parameter t such as time. Let us suppose that the value of a random variable defined on S depends on $s \in S$ and $t \in T$. In this context a ***Stochastic process*** is a set of random variables $\{X(t),\ t \in T\}$ defined on S with a parameter t. Here $X_0 = X(0)$ is called as the initial state of the system.

The values assumed by the random variable $X(t)$ are called *states* and the set of all possible values forms the *state space* of the process. If the state space of a stochastic process is discrete then it is called a ***discrete state process*** also called a *chain*.

On the other hand if the state space is continuous then the stochastic process is called a ***continuous state process.***

Similarly if the index set T is discrete then we have a ***discrete parameter process.*** Otherwise (*i.e., when T is a continuous set*) we have a ***continuous parameter process.*** A discrete parameter process is also called a stochastic sequence denoted by $\{X_n\}$, $n \in T$.

The classification of the four different type of stochastic processes are presented in the form of a table.

	Discrete Index Set - T	Continuous Index Set - T
Discrete State Space	Discrete parameter stochastic process (chain)	Continuous parameter stochastic process (chain)
Continuous State Space	Discrete parameter continuous state stochastic process	Continuous parameter continuous state stochastic process

5.21 Definitions

Probability Vector : By a vector we simply mean n tuple of numbers $(v_1, v_2, \ldots v_n)$ where the quantities $v_1, v_2, \ldots v_n$ are called components of the vector.

A vector $v = (v_1, v_2, \ldots v_n)$ is called a *probability vector* if each one of its components are non negative and their sum is equal to unity.

Examples : $u = (1, 0)$; $v = \left(\frac{1}{2}, \frac{1}{2}\right)$, $w = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right)$ are all probability vectors.

Note : *If v is not a probability vector but each one of the v_i ($i = 1$ to n) are non negative then λv is a probability vector where* $\lambda = 1 / \sum_{i=1}^{n} v_i$

For example if $v = (1, 2, 3)$ then $\lambda = 1/6$ and $(1/6, 2/6, 3/6)$ is a probability vector.

Stochastic Matrix : A square matrix $P = (p_{ij})$ having every row in the form of a probability vector is called a *stochastic matrix.*

Examples : (i) Identity matrix (I) of any order.

$$I_{(2)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} ; \quad I_{(3)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

(ii) $\begin{bmatrix} 1/2 & 1/2 \\ 0 & 1 \end{bmatrix}$

(iii) $\begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \end{bmatrix}$

Regular Stochastic Matrix : A stochastic matrix P is said to be a *regular stochastic matrix* if all the entries of some power P^n are positive.

Example : $A = \begin{bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{bmatrix}$

Consider $A^2 = \begin{bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix}$

$\therefore$ A is a regular stochastic matrix ($n = 2$)

• Properties of a Regular Stochastic Matrix

The following properties are associated with a regular stochastic matrix P of order n.

1. (a) P has a *unique fixed point* $x = (x_1, x_2, \cdots x_n)$ such that $xP = x$

   (b) P has a unique fixed probability vector $v = (v_1, v_2, \cdots v_n)$ such that $vP = v$ where $v_i = \dfrac{x_i}{\sum_{i=1}^{n} x_i}$

2. $P^2, P^3, \cdots$ approaches the matrix V whose rows are each the fixed probability vector v.

3. If u is any probability vector then the sequence of vectors $uP, uP^2, \cdots$ approaches the unique fixed probability vector v.

WORKED PROBLEMS

47. *If* $A = \begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix}$ *is a stochastic matrix and* $v = [v_1, v_2]$ *is a probability vector, show that* vA *is also a probability vector.*

>> By data $a_1 + a_2 = 1, b_1 + b_2 = 1, v_1 + v_2 = 1$.

$$\therefore \quad vA = \begin{bmatrix} v_1, & v_2 \end{bmatrix} \begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix} = \begin{bmatrix} v_1 a_1 + v_2 b_1, & v_1 a_2 + v_2 b_2 \end{bmatrix}$$

We have to prove that $(v_1 a_1 + v_2 b_1) + (v_1 a_2 + v_2 b_2) = 1$

LHS $= v_1 (a_1 + a_2) + v_2 (b_1 + b_2) = v_1 \cdot 1 + v_2 \cdot 1 = v_1 + v_2 = 1$

Thus vA **is also a probability vector.**

---

48. *Prove with reference to two second order stochastic matrices that their product is also a stochastic matrix.*

>> Let $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ and $B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$ be two stochastic matrices. Hence we have,

$$\left.\begin{aligned} a_{11} + a_{12} = 1 \;;\; b_{11} + b_{12} = 1 \\ a_{21} + a_{22} = 1 \;;\; b_{21} + b_{22} = 1 \end{aligned}\right\} \quad \ldots (i)$$

$$\therefore \quad AB = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11} b_{11} + a_{12} b_{21}, & a_{11} b_{12} + a_{12} b_{22} \\ a_{21} b_{11} + a_{22} b_{21}, & a_{21} b_{12} + a_{22} b_{22} \end{bmatrix}$$

We have to show that,

$$a_{11} b_{11} + a_{12} b_{21} + a_{11} b_{12} + a_{12} b_{22} = 1 \quad \ldots (ii)$$

and $$a_{21} b_{11} + a_{22} b_{21} + a_{21} b_{12} + a_{22} b_{22} = 1 \quad \ldots (iii)$$

LHS of (ii) can be written as,

$$a_{11} (b_{11} + b_{12}) + a_{12} (b_{21} + b_{22})$$

ie., $= a_{11} \cdot 1 + a_{12} \cdot 1 = a_{11} + a_{12} = 1$, by using (i).

LHS of (iii) can be written as,

$$a_{21} (b_{11} + b_{12}) + a_{22} (b_{21} + b_{22}) = a_{21} \cdot 1 + a_{22} \cdot 1 = 1$$

Thus AB **is a stochastic matrix.**

Remark : *In particular we can say that* A^n $(n = 1, 2, 3, \ldots)$ *are all stochastic matrices.*

---

49. *If* A *is a square matrix of order* n *whose rows are each the same vector* $a = (a_1, a_2, \cdots a_n)$ *and if* $v = (v_1, v_2, \cdots v_n)$ *is a probability vector, prove that* $vA = a$

>> By data we have, $A = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \\ a_1 & a_2 & \cdots & a_n \\ \cdots & \cdots & \cdots & \cdots \\ a_1 & a_2 & \cdots & a_n \end{bmatrix}$

and $v_1 + v_2 + \cdots v_n = 1$

Consider vA as a matrix product.

$$vA = \begin{bmatrix} v_1, & v_2, \cdots v_n \end{bmatrix} \begin{bmatrix} a_1 & a_2 & \cdots & a_n \\ a_1 & a_2 & \cdots & a_n \\ \cdots & \cdots & \cdots & \cdots \\ a_1 & a_2 & \cdots & a_n \end{bmatrix}$$

$$= \left[v_1 a_1 + v_2 a_1 + \cdots + v_n a_1,\ v_1 a_2 + v_2 a_2 + \cdots + v_n a_2,\ \cdots v_1 a_n + v_2 a_n + \cdots v_n a_n \right]$$

$$= \left[a_1 (v_1 + v_2 \cdots + v_n),\ a_2 (v_1 + v_2 + \cdots + v_n), \cdots a_n (v_1 + v_2 + \cdots + v_n) \right]$$

$$= \left[a_1, a_2, \cdots a_n \right] = a \text{ since } v_1 + v_2 + \cdots + v_n = 1$$

Thus $vA = a$ as required.

---

50. *Find the unique fixed probability vector of the regular stochastic matrix*

$$A = \begin{bmatrix} 3/4 & 1/4 \\ 1/2 & 1/2 \end{bmatrix}$$

>> We have to find $v = (x, y)$ where $x + y = 1$ such that $vA = v$

$$\Rightarrow \begin{bmatrix} x, & y \end{bmatrix} \begin{bmatrix} 3/4 & 1/4 \\ 1/2 & 1/2 \end{bmatrix} = \begin{bmatrix} x, & y \end{bmatrix}$$

i.e., $$\left[\frac{3}{4}x + \frac{1}{2}y,\ \frac{1}{4}x + \frac{1}{2}y \right] = \left[x,\ y \right]$$

$$\Rightarrow \quad \frac{3}{4}x + \frac{1}{2}y = x \qquad \ldots (i)$$

$$\frac{1}{4}x + \frac{1}{2}y = y \qquad \ldots (ii)$$

We can solve either of the two equations by using $y = 1 - x$.

Using $y = 1 - x$ in (i) we have, $\frac{3}{4}x + \frac{(1-x)}{2} = x$

or $\quad 3x + 2 - 2x = 4x \quad \therefore \quad x = 2/3$

Hence $y = 1 - x = 1/3$ and $v = (x, y) = \left(2/3,\ 1/3\right)$

Thus $(2/3,\ 1/3)$ is the unique fixed probability vector.

---

51. *Find the unique fixed probability vector for the regular stochastic matrix*

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1/6 & 1/2 & 1/3 \\ 0 & 2/3 & 1/3 \end{bmatrix}$$

>> We have to find $v = (x, y, z)$ where $x + y + z = 1$ such that $vA = v$

$$\Rightarrow \quad \begin{bmatrix} x, & y, & z \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 1/6 & 1/2 & 1/3 \\ 0 & 2/3 & 1/3 \end{bmatrix} = \begin{bmatrix} x, & y, & z \end{bmatrix}$$

i.e., $$\left[\frac{y}{6},\ x + \frac{y}{2} + \frac{2z}{3},\ \frac{y}{3} + \frac{z}{3} \right] = \left[x,\ y,\ z \right]$$

$$\Rightarrow \quad \frac{y}{6} = x,\ x + \frac{y}{2} + \frac{2z}{3} = y,\ \frac{y}{3} + \frac{z}{3} = z$$

i.e., $\quad y = 6x,\ 6x + 3y + 4z = 6y,\ y + z = 3z$

i.e., $\quad y = 6x,\ 6x - 3y + 4z = 0,\ y - 2z = 0$

Using $y = 6x$ and $z = 1 - x - y = 1 - x - 6x = 1 - 7x$ in $6x - 3y + 4z = 0$ we have,

$$6x - 18x + 4 - 28x = 0 \quad \therefore \quad x = 1/10$$

Hence $y = 6/10,\ z = 3/10$

Thus the required unique fixed probability vector v is given by

$$v = \left(1/10,\ 6/10,\ 3/10\right)$$

---

52. *With reference to the stochastic matrix A in Example-24, verify the property that the sequence A^2, A^3, A^4 approaches the matrix whose rows are each the fixed probability vector.*

>> We have $A = \begin{bmatrix} 3/4 & 1/4 \\ 1/2 & 1/2 \end{bmatrix}$ and we have obtained in Problem-50 the fixed probability vector $v = (2/3,\ 1/3)$.

Let B be the matrix whose each row is v.

i.e., $$B = \begin{bmatrix} 2/3 & 1/3 \\ 2/3 & 1/3 \end{bmatrix}.$$

Consider $A = \frac{1}{4}\begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix}$

Now $A^2 = \frac{1}{16}\begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix}\begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix} = \frac{1}{16}\begin{bmatrix} 11 & 5 \\ 10 & 6 \end{bmatrix} = \begin{bmatrix} 0.6875 & 0.3125 \\ 0.625 & 0.375 \end{bmatrix}$

$$A^3 = \frac{1}{64}\begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix}\begin{bmatrix} 11 & 5 \\ 10 & 6 \end{bmatrix} = \frac{1}{64}\begin{bmatrix} 43 & 21 \\ 42 & 22 \end{bmatrix} = \begin{bmatrix} 0.671875 & 0.328125 \\ 0.65625 & 0.34375 \end{bmatrix}$$

$$A^4 = \frac{1}{256}\begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix}\begin{bmatrix} 43 & 21 \\ 42 & 22 \end{bmatrix} = \frac{1}{256}\begin{bmatrix} 171 & 85 \\ 170 & 86 \end{bmatrix} \approx \begin{bmatrix} 0.67 & 0.33 \\ 0.66 & 0.34 \end{bmatrix}$$

Each row of A^4 is approaching $v = (2/3, 1/3) \approx (0.67, 0.33)$

53. *Find the unique fixed probability vector of the regular stochastic matrix*

$$P = \begin{bmatrix} 0 & 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/4 & 1/4 \\ 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{bmatrix}$$

>> We have to find $v = (a, b, c, d)$ where $a+b+c+d = 1$ such that $vP = v$

$$\Rightarrow \quad [a, b, c, d] \begin{bmatrix} 0 & 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/4 & 1/4 \\ 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{bmatrix} = [a, b, c, d]$$

i.e., $$\left[\frac{b}{2}+\frac{c}{2}+\frac{d}{2}, \frac{a}{2}+\frac{c}{2}+\frac{d}{2}, \frac{a}{4}+\frac{b}{4}, \frac{a}{4}+\frac{b}{4}\right] = [a, b, c, d]$$

$$\Rightarrow \quad \frac{1}{2}(b+c+d) = a \text{ or } b+c+d = 2a$$

$$\frac{1}{2}(a+c+d) = b \text{ or } a+c+d = 2b$$

$$\frac{1}{4}(a+b) = c \text{ or } a+b = 4c$$

$$\frac{1}{4}(a+b) = d \text{ or } a+b = 4d$$

By using $b+c+d = 1-a$ and $a+c+d = 1-b$

(i) and (ii) respectively becomes $1-a = 2a$ and $1-b = 2b$

$\therefore \quad a = 1/3$ and $b = 1/3$

Hence we have from (iii) and (iv),

$4c = 2/3$ and $4d = 2/3$

$\therefore \quad c = 1/6$ and $d = 1/6$

Thus $v = (1/3, 1/3, 1/6, 1/6)$ is the required unique fixed probability vector.

---

54. *Show that (a, b) is a fixed point of the stochastic matrix $P = \begin{bmatrix} 1-b & b \\ a & 1-a \end{bmatrix}$ What is the associated fixed probability vector ?*

Hence write down the fixed probability vector of each of the following matrices.

$$P_1 = \begin{bmatrix} 1/3 & 2/3 \\ 1 & 0 \end{bmatrix}, \quad P_2 = \begin{bmatrix} 1/2 & 1/2 \\ 2/3 & 1/3 \end{bmatrix}, \quad P_3 = \begin{bmatrix} 7/10 & 3/10 \\ 8/10 & 2/10 \end{bmatrix}$$

>> Let $x = (a, b)$ and consider the matrix product

$$xP = [a, b] \begin{bmatrix} 1-b & b \\ a & 1-a \end{bmatrix}$$

$$= [a(1-b)+ba, \; ab+b(1-a)] = [a, b]$$

Thus $xp = x \quad \therefore \quad x = (a, b)$ **is a fixed point of P.**

Also $v = (a/a+b, \; b/a+b)$ **is the required fixed probability vector of P.**

Comparing P_1, P_2, P_3 with P we have respectively

$a = 1, b = 2/3$; $a = 2/3, b = 1/2$; $a = 8/10, b = 3/10$

$a+b = 5/3$; $a+b = 7/6$; $a+b = 11/10$

The corresponding fixed probability vectors of P_1, P_2, P_3 be respectively denoted by v_1, v_2, v_3 where we have in general

$v = (a/a+b, \; b/a+b)$

Thus $v_1 = (3/5, 2/5)$; $v_2 = (4/7, 3/7)$; $v_3 = (8/11, 3/11)$

are the required fixed probability vectors of P_1, P_2, P_3 in the respective order.

---

55. *If $P_1 = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}$ and $P_2 = \begin{bmatrix} 1-b & b \\ a & 1-a \end{bmatrix}$ show that P_1, P_2 and P_1P_2 are stochastic matrices.*

>> In P_1 we have $(1-a)+a = 1$ and $b+(1-b) = 1$

In P_2 we have $b+(1-b) = 1$ and $a+(1-a) = 1$

Thus P_1 **and** P_2 **are stochastic matrices.**

Now $$P_1P_2 = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix} \begin{bmatrix} 1-b & b \\ a & 1-a \end{bmatrix}$$

$$= \begin{bmatrix} (1-a)(1-b)+a^2, & (1-a)b+a(1-a) \\ b(1-b)+a(1-b), & b^2+(1-b)(1-a) \end{bmatrix} = \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix} \text{ (say)}$$

We shall show that $a_1+b_1 = 1$ and $a_2+b_2 = 1$

Now $a_1+b_1 = (1-a)(1-b) + (1-a)b + a^2 + a(1-a)$

$= (1-a)\{1-b+b\} + a\{a+1-a\}$

$= 1-a+a = 1. \quad \therefore \quad a_1+b_1 = 1$

Also $a_2 + b_2 = b(1-b) + b^2 + a(1-b) + (1-b)(1-a)$

$= b\{(1-b) + b\} + (1-b)\{a + (1-a)\}$

$= b + 1 - b = 1$. Thus $a_2 + b_2 = 1$

Thus $P_1 P_2$ is a stochastic matrix.

---

56. *Show that* $P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix}$ *is a regular stochastic matrix. Also find the associated unique fixed probability vector.*

>> Consider $P^2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix}\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \end{bmatrix}$

$$P \cdot P^2 = P^3 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix}\begin{bmatrix} 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}$$

$$P \cdot P^3 = P^4 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix}\begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{bmatrix} = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/4 & 1/4 & 1/2 \\ 1/4 & 1/2 & 1/4 \end{bmatrix}$$

$$P \cdot P^4 = P^5 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix}\begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/4 & 1/4 & 1/2 \\ 1/4 & 1/2 & 1/4 \end{bmatrix} = \begin{bmatrix} 1/4 & 1/4 & 1/2 \\ 1/4 & 1/2 & 1/4 \\ 1/8 & 3/8 & 1/2 \end{bmatrix}$$

We observe that in P^5 all the entries are positive.

Thus P is a regular stochastic matrix.

Next we have to find $v = (a, b, c)$ where $a + b + c = 1$ such that $vP = v$

$$\Rightarrow \quad [a, b, c]\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix} = [a, b, c]$$

$$\left[\frac{c}{2}, a + \frac{c}{2}, b\right] = [a, b, c]$$

$$\frac{c}{2} = a, \quad a + \frac{c}{2} = b, \quad b = c$$

Using $c = 2a$ and $b = c = 2a$ in $a + b + c = 1$ we get

$5a = 1$ or $a = 1/5$ *Hence* $b = c = 2a = 2/5$

Thus $(1/5, 2/5, 2/5)$ is the required unique fixed probability vector of P.

5.22 Markov Chains

A stochastic process which is such that the generation of the probability distribution depend only on the present state is called a *Markov process.*
If this state space is discrete (*finite or countably infinite*) we say that the process is a discrete state process or chain. Then the Markov process is known as a *Markov chain.*

Further if the state space is continuous, the process is called a continuous state process.

We explicitly **define a Markov chain** as follows.

Let the outcomes X_1, X_2, ... of a sequence of trials satisfy the following properties.

(i) Each outcome belong to the finite set (*state space*) of the outcomes $\{a_1, a_2, \ldots a_m\}$

(ii) The outcome of any trial depend at most upon the outcome of the immediate preceeding trial.

Probability p_{ij} is associated with every pair of states (a_i, a_j) that a_j occurs immediately after a_i occurs. Such a stochastic process is called a *finite Markov chain* These probabilities (p_{ij}) which are non zero real numbers are called *transition probabilities* and they form a square matrix of order m called the *transition probability matrix* (*t.p.m*) denoted by P.

i.e., $$P = [p_{ij}] = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{bmatrix}$$

With each state a_i there corresponds the i^{th} row of transition probabilities $p_{i1}, p_{i2}, \ldots p_{im}$. It is evident that the elements of P have the following properties

$$\text{(i)} \quad 0 \le p_{ij} \le 1 \quad \text{(ii)} \sum_{\substack{j=1 \\ (i = 1, 2, 3, \ldots m)}}^{m} p_{ij} = 1$$

The above two properties satisfy the requirement of a stochastic matrix and hence we conclude that the transition matrix of a Markov chain is a stochastic matrix.

Illustrative Examples for writing t.p.m of a Markov chain

1. *A person commutes the distance to his office everyday either by train or by bus. Suppose he does not go by train for two consecutive days, but if he goes by bus the next day he is just as likely to go by bus again as he is to travel by train.*

The state space of the system is $\{$ train (t), bus $(b)\}$
The stochastic process is a Markov chain since the outcome of any day depends only on the happening of the previous day. The t · p · m is as follows.

$$P = \begin{array}{c} \\ t \\ b \end{array} \begin{array}{c} \begin{array}{cc} t & b \end{array} \\ \begin{bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{bmatrix} \end{array}$$

The first row of the matrix is related to the fact that the person does not commute two consecutive days by train and is sure to go by bus if he had travelled by train. The second row of the matrix is related to the fact that if the person had commuted in bus on a particular day he is likely to go by bus again or by train. Thus the probabilities are equal to 1/2.

2. *Three boys A, B, C are throwing ball to each other. A always throws the ball to B and B always throws the ball to C. C is just as likely to throw the ball to B as to A.*

State space $= \{A, B, C\}$ and the t · p · m P is as follows.

$$P = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix} \end{array}$$

- **Higher transition probabilities**

The entry p_{ij} in the transition probability matrix P of the Markov chain is the probability that the system changes from the state a_i to a_j in a single step. That is $a_i \to a_j$

The probability that the system changes from the state a_i to the state a_j in exactly n steps is denoted by $p_{ij}^{(n)}$

That is $a_i \to a_{r_1} \to a_{r_2} \to \cdots \to a_{r_{n-1}} \to a_j$

The matrix formed by the probabilities $p_{ij}^{(n)}$ is called the 'n - step transition matrix' denoted by $P^{(n)}$

$[p^{(n)}] = [p_{ij}^{(n)}]$ is obviously a stochastic matrix.

It can be proved that the n step transition matrix is equal to the n^{th} power of P.

That is $p^{(n)} = P^n$

Let P be the t.p.m of the Markov chain and let $p = (p_i) = (p_1, p_2, \cdots p_m)$ be the probability distribution at some arbitrary time. Then pP, $pP^2 \cdots pP^n$ respectively are the probabilities of the system after one step, two steps, ... n steps.

Let $p^{(0)} = [p_1^{(0)}, p_2^{(0)}, \cdots p_m^{(0)}]$ denote the initial probability distribution at the start of the process and let $p^{(n)} = [p_1^{(n)}, p_2^{(n)}, \cdots p_m^{(n)}]$ denote the n^{th} step probability distribution at the end of n steps. Thus we have

$p^{(1)} = p^{(0)}P$, $p^{(2)} = p^{(1)}P = p^{(0)}P^2$, $\cdots p^{(n)} = p^{(0)}P^n$

Illustrations

1. Let us consider the t.p.m of the earlier illustrated Example-1

$$P = \begin{array}{c} \\ t \\ b \end{array} \begin{array}{c} \begin{array}{cc} t & b \end{array} \\ \begin{bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{bmatrix} \end{array} = \begin{bmatrix} P_{tt} & P_{tb} \\ P_{bt} & P_{bb} \end{bmatrix}$$

We shall find P^2 and P^3

$$P^2 = \begin{bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix} = \begin{bmatrix} P_{tt}^{(2)} & P_{tb}^{(2)} \\ P_{bt}^{(2)} & P_{bb}^{(2)} \end{bmatrix}$$

$$P^3 = \begin{bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{bmatrix}\begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix} = \begin{bmatrix} 1/4 & 3/4 \\ 3/8 & 5/8 \end{bmatrix} = \begin{bmatrix} P_{tt}^{(3)} & P_{tb}^{(3)} \\ P_{bt}^{(3)} & P_{bb}^{(3)} \end{bmatrix}$$

$P_{tb}^{(2)} = 1/2$ means that the probability that the system changes from the state t to b in exactly two steps is 1/2

$P_{bt}^{(3)} = 3/8$ means that the probability that the system changes from the state b to t in exactly 3 steps is 3/8.

Next let us create an initial probability distribution for the start of the process. Let us suppose that the person rolled a 'die' and decided that he will go by bus if the number appeared on the face is divisible by 3.

$\therefore \quad p(b) = 2/6 = 1/3$ and $p(t) = 2/3$

That is $p^{(0)} = (2/3, 1/3)$ is the initial probability distribution.

Now $p^{(2)} = p^{(0)}P^2 = \begin{bmatrix} 2/3, & 1/3 \end{bmatrix}\begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix} = \begin{bmatrix} 5/12, & 7/12 \end{bmatrix}$

$$p^{(3)} = p^{(0)} P^3 = \left[2/3,\ 1/3\right]\begin{bmatrix} 1/4 & 3/4 \\ 3/8 & 5/8 \end{bmatrix} = \left[7/24,\ 17/24\right]$$

$$p^{(3)} = \left[7/24,\ 17/24\right] = \left[p_1^{(3)},\ p_2^{(3)}\right]$$

This is the probability distribution after 3 days.

That is, probability of travelling by train after 3 days = 7/24

probability of travelling by bus after 3 days = 17/24

2. Let us consider the t.p.m of the earlier illustrated Example - 2.

$$P = \begin{array}{c} \\ A \\ B \\ C \end{array}\begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix} \end{array}$$

Referring to Problem - 56, we have $P^5 = \begin{bmatrix} 1/4 & 1/4 & 1/2 \\ 1/4 & 1/2 & 1/4 \\ 1/8 & 3/8 & 1/2 \end{bmatrix}$

Supposing that C was the person having the ball first then $p^{(0)} = (0, 0, 1)$

Consider $p^{(5)} = p^{(0)} P^5 = [0, 0, 1]\begin{bmatrix} 1/4 & 1/4 & 1/2 \\ 1/4 & 1/2 & 1/4 \\ 1/8 & 3/8 & 1/2 \end{bmatrix}$

$$p^{(5)} = \left[1/8,\ 3/8,\ 1/2\right] = \left[p_A^{(5)},\ p_B^{(5)},\ p_C^{(5)}\right]$$

This implies that after 5 throws the probability that the ball is with A is 1/8, the ball with B is 3/8, the ball with C is 1/2.

- **Stationary distribution of regular Markov chains**

A Markov chain is said to be regular if the associated transition probability matrix P is regular.

If P is a regular stochastic matrix of the Markov chain, then the sequence of n step transition matrices $P^2, P^3, \cdots P^n$ approaches the matrix V whose rows are each the unique fixed probability vector v of P.

We have $p^{(n)} = p^{(0)} P^n$ where, $p^{(n)} = \left[p_1^{(n)}, p_2^{(n)} \cdots p_m^{(n)}\right]$

Further as $n \to \infty$, $p_i^{(n)} = v_i$ where $i = 1, 2, 3, \cdots m$.

This is called the *stationary distribution* of the markov chain and $v = (v_1, v_2, \cdots v_m)$ is called the stationary (*fixed*) probability vector of the Markov chain.

Referring to the Illustrative Example - 2, the t.p.m of the Markov chain is

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix}$$

and by Worked Problem - 56 the unique fixed probability vector of P is $(1/5, 2/5, 2/5)$.

Hence we conclude that in the long run $(n \to \infty)$ A will have thrown the ball 20% of the time, while B and C will have thrown the ball 40% of the time.

Note: A Markov chain is said to be *irreducible* if every state can be reached from every other state in a finite number of steps. That is to say that $p_{ij}^{(n)} > 0$ for some $n \geq 1$.

This is equivalent to saying that a ***Markov chain is irreducible if the associated transition probability matrix is regular.***

- **Absorbing state of a Markov chain**

In a Markov chain the process reaches to a certain state after which it continues to remain in the same state. Such a state is called an *absorbing state* of the Markov chain. In an absorbing state the transition probabilities p_{ij} are such that

$$p_{ij} = 1 \text{ for } i = j \text{ and } p_{ij} = 0 \text{ otherwise.}$$

Thus a state a_i of the Markov chain is absorbing if the i^{th} row of the t.p.m has 1 on the principal diagonal and zeroes elsewhere.

Examples :

1. $P = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \end{array}\begin{array}{c} \begin{array}{ccc} a_1 & a_2 & a_3 \end{array} \\ \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \\ 0 & 1/4 & 3/4 \end{bmatrix} \end{array}$ The state a_2 is absorbing.

2. $P = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{array}\begin{array}{c} \begin{array}{cccc} a_1 & a_2 & a_3 & a_4 \end{array} \\ \begin{bmatrix} 1/4 & 1/4 & 1/2 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$ The states a_3 and a_4 are absorbing.

WORKED EXAMPLES

57. *The transition matrix P of a Markov chain is given by* $\begin{bmatrix} 1/2 & 1/2 \\ 3/4 & 1/4 \end{bmatrix}$ *with the initial probability distribution* $p^{(0)} = (1/4, 3/4)$. *Define and find the following.*

(i) $p_{21}^{(2)}$ *(ii)* $p_{12}^{(2)}$ *(iii)* $p^{(2)}$ *(iv)* $p_1^{(2)}$

(v) the vector $p^{(0)} P^n$ *approches.*

(vi) the matrix P^n *approaches.*

>> (i) $p_{21}^{(2)}$ is the probability of moving from state a_2 to state a_1 in 2 steps. This can be obtained from the 2 - step transition matrix P^2

$$P^2 = \begin{bmatrix} 1/2 & 1/2 \\ 3/4 & 1/4 \end{bmatrix}\begin{bmatrix} 1/2 & 1/2 \\ 3/4 & 1/4 \end{bmatrix} = \begin{bmatrix} 5/8 & 3/8 \\ 9/16 & 7/16 \end{bmatrix} = \begin{bmatrix} p_{11}^{(2)} & p_{12}^{(2)} \\ p_{21}^{(2)} & p_{22}^{(2)} \end{bmatrix}$$

$\therefore \quad p_{21}^{(2)} = 9/16$

(ii) $p_{12}^{(2)}$ is the probability of moving from state a_1 to a_2 in two steps $\quad p_{12}^{(2)} = 3/8$

(iii) $p^{(2)}$ is the probability distribution of the system after 2 steps.

$$p^{(2)} = p^{(0)} p^2 = \begin{bmatrix} 1/4, & 3/4 \end{bmatrix}\begin{bmatrix} 5/8 & 3/8 \\ 9/16 & 7/16 \end{bmatrix} = \begin{bmatrix} \frac{37}{64}, & \frac{27}{64} \end{bmatrix}$$

That is $p^{(2)} = [37/64, 27/64] = [p_1^{(2)}, p_2^{(2)}]$.

(iv) $p_1^{(2)}$ is the probability that the process is in the state a_1 after 2 steps. Hence $p_1^{(2)} = 37/64$.

(v) The vector $p^{(0)} P^n$ approaches the unique fixed probability vector of P and we shall find the same.

Let $v = (x, y)$ where $x + y = 1$ and we must have $vP = v$

That is $[x, y]\begin{bmatrix} 1/2 & 1/2 \\ 3/4 & 1/4 \end{bmatrix} = [x, y]$

$x/2 + 3y/4 = x$ and $x/2 + y/4 = y$

Using $y = 1 - x$ the first equation becomes

$$\frac{x}{2} + \frac{3(1-x)}{4} = x \text{ or } 2x + 3(1-x) = 4x \ \therefore \qquad x = 3/5$$

since $x = 3/5$, $y = 2/5$

The vector $p^{(0)} P^n$ approaches the vector $(3/5, 2/5)$

(vi) P^n approaches the matrix V whose rows are each the fixed probability vector of P.

P^n approaches the matrix $\begin{bmatrix} 3/5 & 2/5 \\ 3/5 & 2/5 \end{bmatrix}$

58. *The $t \cdot p \cdot m$ of a Markov chain is given by*

$$P = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \\ 1/4 & 1/2 & 1/4 \end{bmatrix}$$

and the initial probability distribution is $p^{(0)} = (1/2, 1/2, 0)$

Find $p_{13}^{(2)}$, $p_{23}^{(2)}$, $p^{(2)}$ and $p_1^{(2)}$

>> First let us find the two step transition matrix P^2

$$P^2 = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \\ 1/4 & 1/2 & 1/4 \end{bmatrix}\begin{bmatrix} 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \\ 1/4 & 1/2 & 1/4 \end{bmatrix} = \begin{bmatrix} 3/8 & 1/4 & 3/8 \\ 1/2 & 0 & 1/2 \\ 11/16 & 1/8 & 3/16 \end{bmatrix}$$

$\therefore \quad p_{13}^{(2)} = 3/8$ and $p_{23}^{(2)} = 1/2$

$$p^{(2)} = p^{(0)} P^2 = \begin{bmatrix} 1/2, & 1/2, & 0 \end{bmatrix}\begin{bmatrix} 3/8 & 1/4 & 3/8 \\ 1/2 & 0 & 1/2 \\ 11/16 & 1/8 & 3/16 \end{bmatrix}$$

$$= \begin{bmatrix} 7/16, & 1/8, & 7/16 \end{bmatrix}$$

$\therefore \quad p^{(2)} = (7/16, 1/8, 7/16)$ and $p_1^{(2)} = 7/16$

59. *Prove that the Markov chain whose $t \cdot p \cdot m$ is*

$$P = \begin{bmatrix} 0 & 2/3 & 1/3 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{bmatrix}$$ *is irreducible.*

Find the corresponding stationary probability vector.

>> We shall show that P is a regular stochastic matrix. For convenience we shall write the given matrix in the form

$$P = \frac{1}{6}\begin{bmatrix} 0 & 4 & 2 \\ 3 & 0 & 3 \\ 3 & 3 & 0 \end{bmatrix}$$

Consider $P^2 = \frac{1}{36}\begin{bmatrix} 0 & 4 & 2 \\ 3 & 0 & 3 \\ 3 & 3 & 0 \end{bmatrix}\begin{bmatrix} 0 & 4 & 2 \\ 3 & 0 & 3 \\ 3 & 3 & 0 \end{bmatrix} = \frac{1}{36}\begin{bmatrix} 18 & 8 & 12 \\ 9 & 21 & 6 \\ 9 & 12 & 15 \end{bmatrix}$

Since all the entries in P^2 are positive we conclude that the t.p.m P is regular.

Hence the Markov chain having t.p.m P is irreducible.

Next we shall find the fixed probability vector of P.

If $v = (x, y, z)$ we shall find v such that $vP = v$ where $x+y+z = 1$.

That is $[x, y, z]\cdot\frac{1}{6}\begin{bmatrix} 0 & 4 & 2 \\ 3 & 0 & 3 \\ 3 & 3 & 0 \end{bmatrix} = [x, y, z]$

$\Rightarrow \quad \frac{1}{6}[3y+3z,\ 4x+3z,\ 2x+3y] = [x, y, z]$

$\Rightarrow \quad 3y+3z = 6x\ ;\ 4x+3z = 6y\ ;\ 2x+3y = 6z$

Solving these by using $x+y+z = 1$ we obtain
$x = 1/3,\ y = 10/27,\ z = 8/27$

Thus $v = \left(1/3,\ 10/27,\ 8/27\right)$ is the required stationary probability vector.

---

60. *A habitual gambler is a member of two clubs A and B. He visits either of the clubs everyday for playing cards. He never visits club A on two consecutive days. But, if he visits club B on a particular day, then the next day he is as likely to visit club B or club A. Find the transition matrix of this Markov chain. Also,*

(a) *show that the matrix is a regular stochastic matrix and find the unique fixed probability vector.*

(b) *if the person had visited club B on Monday, find the probability that he visits club A on Thursday.*

>> The transition matrix P of the Markov chain is formulated as follows.

$$P = \begin{matrix} \\ A \\ B \end{matrix}\begin{matrix} \begin{matrix} A & B \end{matrix} \\ \begin{bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{bmatrix} \end{matrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

The first row corresponds to the fact that he never goes to club A on two consecutive days which implies that he is sure to visit club B. The second row corresponds to the fact that if he goes to B on a particular day he visits B or A on the following day. Probability of going to A is $1/2$ and probability of going to B is also $1/2$.

(a) Now consider $P^2 = \begin{bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix}$

Since all the entries of P^2 are positive P is a regular stochastic matrix.

We shall find the unique fixed probability vector. That is to find

$v = (x, y)$ such that $vP = v$ where $x+y = 1$

i.e., $\quad \begin{bmatrix} x, & y \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{bmatrix} = \begin{bmatrix} x, & y \end{bmatrix}$

or $\quad \begin{bmatrix} \frac{y}{2}, & x + \frac{y}{2} \end{bmatrix} = \begin{bmatrix} x, & y \end{bmatrix}$

$\Rightarrow \quad \frac{y}{2} = x\ ;\ x + \frac{y}{2} = y$. But $y = 1 - x$

$\therefore \quad \frac{1-x}{2} = x$ or $x = \frac{1}{3}\ \therefore\ y = \frac{2}{3}$

Thus $v = (1/3, 2/3)$

(b) Let us suppose Monday as day 1, then Thursday will be 3 days after Monday. Given that the person had visited club B on Monday the probability that he visits club A after 3 days is equivalent to finding $a_{21}^{(3)}$ from P^3.

Now $\quad P^3 = P^2 \cdot P = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 1/2 & 1/2 \end{bmatrix} = \begin{bmatrix} 1/4 & 3/4 \\ 3/8 & 5/8 \end{bmatrix}$

$\therefore \quad a_{21}^{(3)} = 3/8$. **Thus the required probability is 3/8.**

---

61. *A student's study habits are as follows. If he studies one night, he is 70% sure not to study the next night. On the other hand if he does not study one night, he is 60% sure not to study the next night. In the long run how often does he study ?*

>> The state space of the system is $\{A,\ B\}$ where A : Studying B : Not studying.

The associated transition matrix P is as follows.

$$P = \begin{matrix} \\ A \\ B \end{matrix}\begin{matrix} \begin{matrix} A & B \end{matrix} \\ \begin{bmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \end{bmatrix} \end{matrix}$$

In order to find the happening in the long run we have to find the unique fixed probability vector v of P. That is to find

$v = (x, y)$ such that $vP = v$ where $x + y = 1$

i.e., $\begin{bmatrix} x, & y \end{bmatrix}\begin{bmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \end{bmatrix} = \begin{bmatrix} x, & y \end{bmatrix}$

i.e., $\begin{bmatrix} 0.3x+0.4y, & 0.7x+0.6y \end{bmatrix} = \begin{bmatrix} x, & y \end{bmatrix}$

$\Rightarrow 0.3x+0.4y = x \; ; \; 0.7x+0.6y = y$

Using $y = 1-x$ in the first of the equations we have

$0.3x+0.4(1-x) = x$ or $1.1x = 0.4 \quad \therefore \quad x = 4/11$

Since $x = 4/11, \; y = 7/11, \; v = (4/11 \text{ and } 7/11) = (p_A, p_B)$

Thus we conclude that in the long run the student will study 4/11 of the time or 36.36% of the time.

62. *A man's smoking habits are as follows. If he smokes filter cigarettes one week, he switches to non filter cigarettes the next week with probability 0.2. On the other hand, if he smokes non filter cigarettes one week there is a probability of 0.7 that he will smoke non filter cigarettes the next week as well. In the long run how often does he smoke filter cigarettes ?*

>> The state space of the system is $\{A, \; B\}$ where

A : Smoking filter cigarettes, B : Smoking non filter cigarettes

The associated transition matrix is as follows.

$$P = \begin{matrix} \\ A \\ B \end{matrix}\begin{matrix} \begin{matrix} A & B \end{matrix} \\ \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{bmatrix} \end{matrix} = \begin{bmatrix} 8/10 & 2/10 \\ 3/10 & 7/10 \end{bmatrix} = \frac{1}{10}\begin{bmatrix} 8 & 2 \\ 3 & 7 \end{bmatrix}$$

We have to find the unique fixed probability vector, $v = (x, y)$ such that $vP = v$ where $x+y = 1$

i.e., $\begin{bmatrix} x, & y \end{bmatrix} \cdot \frac{1}{10}\begin{bmatrix} 8 & 2 \\ 3 & 7 \end{bmatrix} = \begin{bmatrix} x, & y \end{bmatrix}$

i.e., $\begin{bmatrix} 8x+3y, & 2x+7y \end{bmatrix} = \begin{bmatrix} 10x, & 10y \end{bmatrix}$

$\Rightarrow \quad 8x+3y = 10x, \; 2x+7y = 10y$

Using $y = 1-x$ in the first equation, we get,

$8x+3(1-x) = 10x$

$x = 3/5 \quad \therefore \quad y = 2/5$

Hence $v = (x, y) = (3/5, \; 2/5) = (p_A, p_B)$

In the long run, he will smoke filter cigarettes 3/5 or 60% of the time.

63. *Each year a man trades his car for a new car in 3 brands of the popular company Maruti Udyog limited. If he has a 'Standard' he trades it for 'Zen'. If he has a 'Zen' he trades it for a 'Esteem'. If he has a 'Esteem' he is just as likely to trade it for a new 'Esteem' or for a 'Zen' or a 'Standard' one. In 1996 he bought his first car which was Esteem.*

(i) Find the probability that he has

(a) 1998 Esteem *(b) 1998 Standard*

(c) 1999 Zen *(d) 1999 Esteem.*

(ii) In the long run, how often will he have a Esteem?

>> The state space of the system is $\{A, \; B, \; C\}$ where

A : Standard B : Zen C : Esteem.

The associated transition matrix is as follows.

$$P = \begin{matrix} \\ A \\ B \\ C \end{matrix}\begin{matrix} \begin{matrix} A & B & C \end{matrix} \\ \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \end{matrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

(i) With 1996 as the first year, 1998 is to be regarded as 2 years after and 1999 as 3 years after.

We need to compute P^2 and P^3

$$P^2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \\ 1/9 & 4/9 & 4/9 \end{bmatrix}$$

$$P^3 = \begin{bmatrix} 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \\ 1/9 & 4/9 & 4/9 \end{bmatrix}\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/9 & 4/9 & 4/9 \\ 4/27 & 7/27 & 16/27 \end{bmatrix}$$

(a) 1998 Esteem $= a_{33}^{(2)} = 4/9$

(b) 1998 Standard $= a_{31}^{(2)} = 1/9$

} with reference to P^2

(c) 1999 Zen $= a_{32}^{(3)} = 7/27$

(d) 1999 Esteem $= a_{33}^{(3)} = 16/27$

} with reference to P^3

(ii) We have to find the unique fixed probability vector $v = (x, y, z)$ such that $vP = v$ where $x+y+z = 1$

i.e., $\begin{bmatrix} x, & y, & z \end{bmatrix}\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} = \begin{bmatrix} x, & y, & z \end{bmatrix}$

$$\text{i.e.,} \quad [x,\ y,\ z] \cdot \frac{1}{3}\begin{bmatrix} 0 & 3 & 0 \\ 0 & 0 & 3 \\ 1 & 1 & 1 \end{bmatrix} = [x,\ y,\ z]$$

$$\text{i.e.,} \quad [z,\ 3x+z,\ 3y+z] = [3x,\ 3y,\ 3z]$$

$$\Rightarrow \quad z = 3x,\ 3x+z = 3y,\ 3y+z = 3z$$

Consider $3x+z = 3y$; Using $z = 3x$ and $y = 1-x-z$ we get $6x = 3(1-x-z)$ or $6x = 3-3x-3z$ or $18x = 3 \ \therefore \ x = 1/6$

Hence we obtain $y = 1/3,\ z = 1/2$

$$\therefore \quad v = [x,\ y,\ z] = [1/6,\ 1/3,\ 1/2] = [p^A,\ p^B,\ p^C]$$

In the long run, probability of he having Esteem is $p^{(C)} = 1/2$

Thus in the long run in 50% of the time he will have Esteem.

---

64. *Three boys A, B, C are throwing ball to each other. A always throws the ball to B and B always throws the ball to C. C is just as likely to throw the ball to B as to A. If C was the first person to throw the ball find the probabilities that after three throws*

(i) A has the ball (ii) B has the ball (iii) C has the ball

>> State space $= \{A, B, C\}$ and the associated t·p·m is as follows.

$$P = \begin{array}{c} \\ A \\ B \\ C \end{array}\begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix} \end{array}$$

Intially if C has the ball, the associated initial probability vector is given by $p^{(0)} = (0,\ 0,\ 1)$

Since the probabilities are desired after three throws we have to find $p^{(3)} = p^{(0)} P^3$

Refering to the Problem - 56, $P^3 = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}$

$$\therefore \quad p^{(3)} = p^{(0)} P^3 = \left[\frac{1}{4},\ \frac{1}{4},\ \frac{1}{2}\right] = \left[p_A^{(3)},\ p_B^{(3)},\ p_C^{(3)}\right]$$

Thus after three throws the probability that the ball is with A is 1/4, with B is 1/4 and with C is 1/2.

---

65. *Two boys B_1, B_2 and two girls G_1, G_2 are throwing ball from one to the other. Each boy throws the ball to the other boy with probability 1/2 and to each girl with probability 1/4. On the otherhand each girl throws the ball to each boy with probability 1/2 and never to the other girl. In the long run how often does each receive the ball.*

>> State space $= \{B_1, B_2, G_1, G_2\}$ and the associated t·p·m P is as follows.

$$P = \begin{array}{c} \\ B_1 \\ B_2 \\ G_1 \\ G_2 \end{array}\begin{array}{c} \begin{array}{cccc} B_1 & B_2 & G_1 & G_2 \end{array} \\ \begin{bmatrix} 0 & 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/4 & 1/4 \\ 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \end{bmatrix} \end{array}$$

We need to find the fixed probability vector $v = (a,\ b,\ c,\ d)$

such that $vP = v$

Refering to Problem - 53. We have $v = (1/3,\ 1/3,\ 1/6,\ 1/6)$

Thus we can say that in the long run **each boy receives the ball 1/3 of the time and each girl 1/6 of the time.**

---

66. *A gambler's luck follows a pattern. If he wins a game, the probability of winning the next game is 0.6. However if he loses a game, the probability of losing the next game is 0.7. There is an even chance of gambler winning the first game. If so*

(a) What is the probability of he winning the second game?

(b) What is the probability of he winning the third game?

(c) In the long run, how often he will win?

>> State space $\{\text{Win } (W),\ \text{Lose } (L)\}$ and the associated t·p·m is as follows.

$$P = \begin{array}{c} \\ W \\ L \end{array}\begin{array}{c} \begin{array}{cc} W & L \end{array} \\ \begin{bmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{bmatrix} \end{array} = \frac{1}{10}\begin{bmatrix} 6 & 4 \\ 3 & 7 \end{bmatrix}$$

Probability of winning the first game is 1/2.

$\therefore$ initial probability vector $p^{(0)} = (1/2,\ 1/2)$

(a) Now $p^{(1)} = p^{(0)} P = \frac{1}{2}[1,\ 1] \cdot \frac{1}{10}\begin{bmatrix} 6 & 4 \\ 3 & 7 \end{bmatrix} = \frac{1}{20}[9,\ 11]$

Hence $p^{(1)} = [9/20,\ 11/20] = [p^{(W)},\ p^{(L)}]$

Thus the probability of he winning the second game is 9/20.

(b) $p^{(2)} = p^{(1)} P = \frac{1}{20}[9, 11] \cdot \frac{1}{10}\begin{bmatrix} 6 & 4 \\ 3 & 7 \end{bmatrix} = \frac{1}{200}[87, 113]$

Hence $p^{(2)} = [87/200, 113/200] = [p^{(W)}, p^{(L)}]$

Thus **the probability of he winning the third game is 87/200.**

(c) We hall find the fixed probability vector

$v = (x, y)$ such that $vP = v$ where $x + y = 1$

That is $[x, y] \cdot \frac{1}{10}\begin{bmatrix} 6 & 4 \\ 3 & 7 \end{bmatrix} = [x, y]$

$\Rightarrow \quad 6x + 3y = 10x, \; 4x + 7y = 10y$

or $\quad 3y = 4x$ and by using $y = 1 - x$ we get

$3(1 - x) = 4x \quad \therefore \quad x = 3/7$ and $y = 4/7$

Hence $v = [3/7, 4/7] = [p^{(W)}, p^{(L)}]$

Thus **in the long run he wins 3/7 of the time.**

EXERCISES

1. Identify the probability vectors from the following.

   (a) $(2/5, 3/5)$ (b) $(0, -1/3, 4/3)$

   (c) $(1/3, 0, 1/6; 1/2, 1/3)$

   (d) $(1/3, 0, 1/6, 1/2)$ (e) $(0.1, 0.2, 0.3, 0.4)$

2. Find the associated probability vector to each of the following tuples.

   (a) $(1, 3, 5)$ (b) $(4, 0, 1, 2)$

   (c) $(1/2, 2/3, 0, 2, 5/6)$

3. $A = [a_{ij}]$ is a stochastic matrix of order $n \times n$ and $v = (v_1, v_2, \ldots v_n)$ is a probability vector, show that vA is also a probability vector.

4. If A and B are two stochastic matrices of order 3×3, prove that AB is also a stochastic matrix.

5. Show that $(cf + ce + de, \; af + bf + ae, \; ad + bd + bc)$ is a fixed point of the stochastic matrix

$$P = \begin{bmatrix} 1-a-b & a & b \\ c & 1-c-d & d \\ e & f & 1-e-f \end{bmatrix}$$

6. Show that the following matrix P is a regular stochastic matrix and also find its unique fixed probability vector.

$$P = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \end{bmatrix}$$

7. Given the t·p·m $P = \begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \end{bmatrix}$ with initial probability distribution $p^{(0)} = (1/3, 2/3)$, find the following.

   (a) $p_{21}^{(3)}$ (b) $p^{(3)}$ (c) $p_2^{(3)}$

8. A software engineer goes to his office everyday by motorbike or by car. He never goes by bike on two consecutive days, but if he goes by car on a day then he is equally likely to go by car or by bike the next day. Find the t·p·m of the Markov chain. If car is used on the first day of the week find the probability that after 4 days (a) bike is used (b) car is used.

9. A salesman's territory consists of 3 cities A, B, C. He never sells in the same city for 2 consecutive days. If he sells in city A, then the next day he sells in city B. However if he sells in either B or C, then the next day he is twice as likely to sell in city A as in the other city. In the long run how often does he sell in each of the cities?

10. Show that the Markov chain with t·p·m given by

$$P = \frac{1}{10}\begin{bmatrix} 6 & 2 & 2 \\ 1 & 8 & 1 \\ 6 & 0 & 4 \end{bmatrix}$$

is irreducible. Find the corresponding stationary probability vector.

ANSWERS

1. (a), (d), (e) are probability vectors.
2. (a) (1/9, 3/9, 5/9) (b) (4/7, 0, 1/7, 2/7)
   (c) (1/8, 1/6, 0, 1/2, 5/24)
6. (4/11, 4/11, 3/11)
7. (a) 7/8 (b) (11/12, 1/12) (c) 1/12
8. $\begin{matrix} & B & C \\ B & 0 & 1 \\ C & 1/2 & 1/2 \end{matrix}$ (a) 5/16 (b) 11/16
9. $v = (0.4, 0.45, 0.15)$. In the long run he sells 40% of the time in city A, 45% of the time in B, 15% of the time in C
10. $v = (4/10, 4/10, 2/10)$

BEATING THE MEMORY

[Formulae, Properties and Results to be remembered from all the modules at a glance]

Module - 1

Numerical Methods

Formulae for solving the initial value problem :

$$\frac{dy}{dx} = f(x, y);\ y(x_0) = y_0.$$ To compute $y(x_0+h)$.

➢ *Taylor's series formula*

$$y(x) = y(x_0)+(x-x_0)y'(x_0)+\frac{(x-x_0)^2}{2!}y''(x_0)+\cdots$$

➢ *Modified Euler's formula* [M.E.F]

Taking $x_1 = x_0+h$ and $y_1 = f(x_1)$

$$y_1^{(0)} = y_0+hf(x_0, y_0) \ldots \text{[Initial approx. / Euler's formula]}$$

$$y_1^{(1)} = y_0+\frac{h}{2}\left[f(x_0, y_0)+f(x_1, y_1^{(0)})\right] \ldots \text{[First approx./ M.E.F]}$$

$$y_1^{(2)} = y_0+\frac{h}{2}\left[f(x_0, y_0)+f(x_1, y_1^{(1)})\right] \ldots \text{[Second approx. / M.E.F]}$$

➢ *Runge - Kutta formula*

$$y(x_0+h) = y_0+\frac{1}{6}[k_1+2k_2+2k_3+k_4] \text{ where}$$

$$k_1 = hf(x_0, y_0)$$

$$k_2 = hf\left(x_0+\frac{h}{2}, y_0+\frac{k_1}{2}\right)$$

$$k_3 = hf\left(x_0+\frac{h}{2}, y_0+\frac{k_2}{2}\right)$$

$$k_4 = hf(x_0+h, y_0+k_3)$$