Engineering Mathematics –IV(15MAT41)
Module-V : SAMPLING THEORY and STOCHASTIC PROCESS

By

Dr. K.S.BASAVARAJAPPA
Professor and Head,
Department of Mathematics,
Bapuji Institute of Engineering and Technology,
Davangere-4, Karnataka
E mail : ksbraju@hotmail.com

Module-V : Sampling Theory and Stochastic Process

- **Module-V : Sampling Theory and Stochastic Process**
- **Sampling**
- **Sampling distribution**
- **Standard error**
- **Test of Hypothesis for means and proportions**
- **Confidence limits for means**
- **Student's t – distribution**
- **Chi - square distribution as a test of goodness of fit**

Stochastic processes

- **Stochastic processes**
- **Probability vector**
- **Stochastic matrices, Fixed points**
- **Regular stochastic matrices**
- **Markov chains**
- **Higher transition probability**
- **Simple problems**

Sampling Theory

- Sampling theory is the field of statistics that is involved with the collection, analysis and interpretation of data gathered from random samples of a population under study. The application of sampling theory is concerned not only with the proper selection of observations from the population that will constitute the random sample

- Sampling aims at gathering maximum information about the population with the minimum effort, cost and time. Sampling determines the reliability of these estimates. The logic of the sampling theory is the logic of induction in which we pass from a particular(sample) to general (population). Such a generalization from sample to population is called **Statistical Inference.**

- It also involves the use of probability theory, along with prior knowledge about the population

parameters, to analyse the data from the random sample and develop conclusions from the analysis. The normal distribution, along with related probability distributions, is most heavily utilized in developing the theoretical background for sampling theory.

Meaning and objectives of Sampling

- **Sampling**

  Sampling denotes the selection of part of the aggregate statistical material with a view to obtaining information about the whole
- **population or Universe**

  The aggregate or totality of statistical information on a particular character of all the members covered by an investigation is called population or Universe.
- **sample**

  The selected part which is used to ascertain the characteristics of population is called sample

- **Population Size(N)**

  It is the total number of members of the population denoted by 'N'
- **Sample Size(n)**

  It is the number included in the sample denoted by 'n'
- **Main objectives of sampling**

  To obtain the maximum information about the population with the minimum effort

  To state the limits of accuracy of estimates based on samples

Some important types of Sampling

- **Random Sampling**

  Here sampling must be random, It is the method of selection of a group of units in such a manner that every unit comprising the population has an equal chance of being included in the sample
- **Purposive Sampling**

  When the sample is selected based on individual judgement of the sampler, it is called purposive sampling
- **Stratified Sampling**

  The population is subdivided into several parts called strata and then subsample is chosen from each of them, then all the sub samples combined together give the stratified sample

- **Systematic Sampling**

It involves the selection of sample units at equal intervals after all the units in the population have been arranged in some order

- **Multistage Sampling**

  It refers to a sampling procedure which is carried out in several stages

The sampling distribution of a statistic is the distribution of that statistic, considered as a random variable, when derived from a random sample of size n. It may be considered as the distribution of the statistic for all possible samples from the same population of a given size. The sampling distribution depends on the underlying distribution of the population, the statistic being considered, the sampling procedure employed, and the sample size used.

There is often considerable interest in whether the sampling distribution can be approximated by an asymptotic distribution, which corresponds to the limiting case either as the number of random samples of finite size, taken from an infinite population and used to produce the distribution, tends to infinity, or when just one equally-infinite-size "sample" is taken of that same population.

Sampling Distribution

It is the probability law which the statistic follows if repeated random samples of a fixed size are drawn from a specified population

The probability distribution of the statistic of the mean of x values will be called as sampling distribution of sample mean

- **Two important sampling distributions(Large samples)**

  Sampling distribution of sample mean(mean of x)
- If x denotes the mean of a random sample of size n drawn from a population with mean μ and standard deviation σ, then the sampling distribution of x is approximately a normal distribution with
- Mean = μ and standard deviation standard error of x

Consider all possible samples of size ‘n’ which can be drawn from a given population at random. For example, we can compute the mean. The means of the samples will not be identical. If we group these different means according to their frequencies, the frequency distribution so formed is known as **sampling distribution of the mean.** Similarly we have **sampling distribution of the standard deviation** etc.

STANDARD ERROR

The standard deviation of the sampling distribution is called standard error. Thus the standard error of the sampling distribution of means is called standard error opf means. The reciprocal of the standard error is called precision.

If $n \geq 30$, a sample is called large otherwise small. The sampling distribution of large samples is assumed to be normal. Where n is the sample size(number of items)

TESTING HYPOTHESIS

- To reach decisions about populations on the basis of sample information, we make certain assumptions about the populations involved. Such assumptions, which may or may not be true, are called **Statistical hypothesis.** By testing a hypotheis is meant a process for deciding whether to accept or reject the hypothesis.

- The method consists in assuming the hypothesis as correct and then computing the probability of getting the observed sample. If this probability is less than a certain pre- assigned value the hypothesis is rejected.
- The American statistician J. Neyman(1894-1981) and the English statistician E.S.Pearson(1895-1980) - son of Kari Pearson developed a systematic theory of tests around 1930.

ERRORS

- If a hypothesis is rejected while it should have been accepted, we say that a **Type I error** has been estimated. On the other hand, if a hypothesis is accepted while it should have been rejected , we say that a **Type II error** has been made. To reduce the both types of errors is to increase the sample size, if possible.

NULL HYPOTHESIS

The hypothesis formulated for the sake of rejecting it, under the assumption that it is true, is called the null hypothesis and is denoted by H_0. To test whether procedures is better than another, we assume that there is no difference between the procedures. Similarly to test whether there is a relationship between two variates, we take H_0 that there is no relationship.

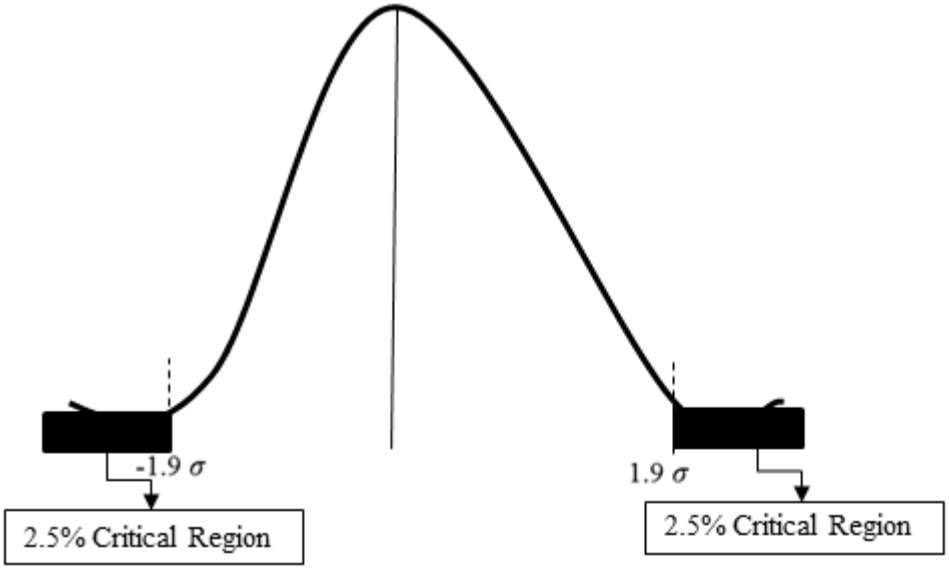
By accepting a null hypothesis, we mean that on the basis of the statistic calculated from the sample, we do not reject the hypothesis. It however, does not imply that the hypothesis is proved

to be true. Nor its rejection implies that it is disproved.

LEVEL OF SIGNIFICANCE

- The probability level below which we reject the hypothesis is known as the **level of significance**. The region in which a sample value falling is rejected, is known as **critical region.** We generally take two critical regions which cover 5% and 1% areas of the normal curve.

- The shaded portion in the figure corresponds to 5% level of significance. Thus the probability of the value of the variate falling in the critical region is the level of significance



TEST OF SIGNIFICANCE

- The procedure which enables us to decide whether to accept or reject the hypothesis is called the test of significance. Test is carried out by whether the differences between the sample values and the population values are so large that they signify evidence against the hypothesis or these differences are so small as to account for fluctuations of sampling.

CONFIDENCE LIMITS

- Let the sampling distribution of a statistic 'S' is normal with mean μ and standard deviation σ. As in the fig.1.the sample statistic 'S' can be expected to lie in the interval $(\mu - 1.96\sigma, \mu + 1.96\sigma,)$ for 95% times i.e.we can be confident of finding μ in the interval $(S - 1.96\sigma, S + 1.96\sigma,)$ in the 95% cases. Because of this, we call $(S - 1.96\sigma, S + 1.96\sigma,)$ the 95%confidence interval for estimation μ .

- The ends of this interval $(S \pm 1.96)$ are called 95% confidence limits for fiducial limits for S. Similarly $(S \pm 2.58\sigma)$ are 99% confidence limits. The numbers 1.96,2.58 etc. are called confidence coefficients. The values of confidence coefficients corresponding to various levels of significance can be found from the normal curve area.

SIMPLE SAMPLING OF ATTRIBUTES

- The sampling of attributes may be regarded as the selection of samples from a population whose members posses the attribute 'K' or 'not K'. The presence of 'K' may be called a success and its absence a failure.
- Suppose we draw a simple sample of n items. Clearly it is same as series of n independent trials with the same probability 'p' of success.

- The probabilities of 0,1,2,....n success are the terms in the binomial expansions of $(p + p)^n$ where $q = 1 - p$
- We know that the mean of this distribution is 'np'and standard deviation is $\sqrt{npq}$ i.e. the expected value of success in a sample of size n is 'np' and the standard error is $\sqrt{npq}$.

- If we consider the proportion of success, then,
- Mean proportion of success $= \frac{np}{n} = p$
- Standard error of the proportion of successes

$$\sqrt{n \cdot \frac{p}{n} \cdot \frac{q}{n}} = \sqrt{\frac{pq}{n}}$$

- Precision of the proportion of success.$= \sqrt{\frac{n}{pq}}$, which varies as $\sqrt{n}$ since p and q are constants.

Example

- 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 unbiased die?

Solution:

Here n = 9000, x = 3240(observed value of successes)

P(throwing 5 or 6) = 1/6+1/6 = 2/6 =1/3,

we get q = 2/3, therefore

Expected number of successes = (1/3) *9000 = 3000

The standard normal variate(SNV) is,

Z = (x – mean)/S.D = (x – np)/S.D = (x – np)/√npq

= (3240 - 3000)/√9000* 1/3*2/3 = 5.4

Z = 5.4 > 2.58

This shows that the hypothesis is to be rejected at 1% level of significance

Conclusion : The die is biased

Example:

- A sample of 900 items has mean 3.4 and S.D 2.61. Can it be regarded as drawn from a population with mean 3.25 at 5% level of significance?

Solution:

Given n = 900, Sample Mean = 3.4,Population Mean(μ) = 3.25, S.D = 2.61 = √npq, Then the Standard normal variate(SNV) is,

SNV (Z) = (Sample mean – Population mean)(√ (n) / √npq

= (3.4 – 3.25)(√900)/2.61 =1.73

I Z I = 1.73 < 1.96

Conclusion: The given sample can be regarded as one from the population with mean 3.25

Example:

- A sugar factory is expected to sell sugar in 100 kg bags, A sample of 144 bags gives mean as 99 kg and S.D as 4. Can we conclude that the factory is working as per standards?

  Solution:

  Given n = 144, Sample Mean = 99,

  Population Mean(μ) = 100, S.D = 4, Then the Standard normal variate(SNV) is ,

  SNV (Z) = (Sample mean – Population mean)(√ (n) / S.D

  = **(99 – 100)(√144)/4 = - 3**

  I Z I = 3 > 1.96

  Conclusion: Factory is not working as per standards

EXAMPLE

- In a locality containing 18000 families , a sample of 840 families was selected at random. Of these 840 families , 206 families were found to have a monthly income of Rs. 250 or less. It is desired to estimate how many out of 18000 families have a monthly income of Rs. 250 or less. Within what limits would you place your estimate?

  Here, $p = \frac{206}{840} - \frac{103}{420}$ *and* $q - \frac{317}{420}$

Therefore S.E. of these population of families having a monthly income of Rs.250 or less

$$= \sqrt{\frac{pq}{n}} = 0.015 = 1.5\%$$

Hence taking 24.5% to be the estimate of families having a monthly income of Rs.250 or less in the locality, the limits are 20% and 29% approximately.

PROBLEMS

- A coin is tossed 400 times and it turns up head 216 times. Discuss whether the coin may be an unbiased one.
- A die is tossed 960 times and it falls with 5 upwards 184 times. Is the die biased.
- Balls are drawn from a bag containing equal number of black and white balls, each ball being

replaced before drawing another. In 2250 drawings 1018 balck and 1232 while balls have been drawn. Do you suspect some bias on the part of the drawer

- A machine produces 16 imperfect articles in a sample of 500. After machine is overhauled, it produces 3 imperfect articles in a batch of 100. Has the machine been improved.
- One type of aircraft is found to develop engine trouble in 5 flights out of total 100and another type in 7 flights out of a total of 200 flights. Is there a significant difference tin the two types of aircrafts so far as engine defects are concerned.

SAMPLING VARIABLES

- Let us consider sampling of a variable such as weight, height etc. Each member of the population gives a value of the variable and the population is a frequency distribution of the variable. Thus a random sample of size 'n' from the population is same as selecting 'n' values of the variable from those of the distribution.

SAMPLING DISTRIBUTION OF THE MEAN

- If a population is distributed normally with mean μ and standard deviation σ, then the sample means $\mu_1, \mu_2, \mu_3, \ldots.$ are also distributed normally with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$. Even if the parent population is not normal, the distribution of sample means is nearly normal for large values of 'n'.

CONFIDENCE LIMITS FOR UNKNOWN MEAN

- Let the population from which a random sample of size 'n' is drawn, have mean μ and standard deviation σ. If μ is not known, there will be a range of values of μ for which observed mean $\bar{x}$ of the sample is not significant at any assigned level of probability. The relative deviation $\bar{x}$ of form μ is $\frac{\bar{x}-\mu}{\sqrt{\sigma}}$

- If $\bar{x}$ is not significant at 5% level of probability, then

$$\left|\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}\right| < 1.96 \; ie.$$

$$\bar{x} + 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} - 1.96\frac{\sigma}{\sqrt{n}}$$

- Thus 95% confidence or fiducial limits for the mean of the population corresponding to given sample are $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$
- Similarly 99% confidence limits for μ are $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$

PROBLEMS

- A sample of 400 items is taken from a normal population whose mean is 4 and variance 4. If the sample mean is 4-15, can the samples be regarded as a simple sample.
- A normal population has a mean 0.1 and a S.D of 2.1. Find the probability that the mean of simple sample of 900 members will be negative.
- The density function of a random variable x is f(x) = ke^{-2x^2+10x}. Find the upper 5% point of the distribution of the means of the random sample of size 25 from the

NUMBER OF DEGREES OF FREEDOM

- Number of degrees of freedom is the number of values in a set which may be assigned arbitrarily. If n is the given sample size then,

  degrees of freedom(df) = n-1.

STUDENT'S t – DISTRIBUTION

- Consider a small sample of size 'n', drawn from a normal population with mean μ and S.D σ. If $\bar{x}$ and σ_s be the sample mean and S.D., then the statistic 't' is defined as

$$t = \frac{\bar{x}-\mu}{\sigma_s}\sqrt{n} \text{ or } t = \frac{\bar{x}-\mu}{\sigma_s}\sqrt{\gamma+1}$$

Where γ = n-1 denotes the 'df' of 't'. If we calculate 't' for each sample, we obtain the sampling distribution for t. This distribution known as **Student's t – distribution** is given by,

$$y = \frac{y_0}{\left(1+\frac{t^2}{\gamma}\right)^{\frac{\gamma+1}{2}}}$$

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

Properties of t - Distribution

- This curve is symmetrical about the line t = 0, like the normal curve, since only even powers of t appear in equation. But it is more peaked than the normal curve with same S.D. The t-curve approaches the horizontal axis less rapidly than the normal curve. Also t –curve attains its maximum value at t = 0 so that its mode coincides with the mean.

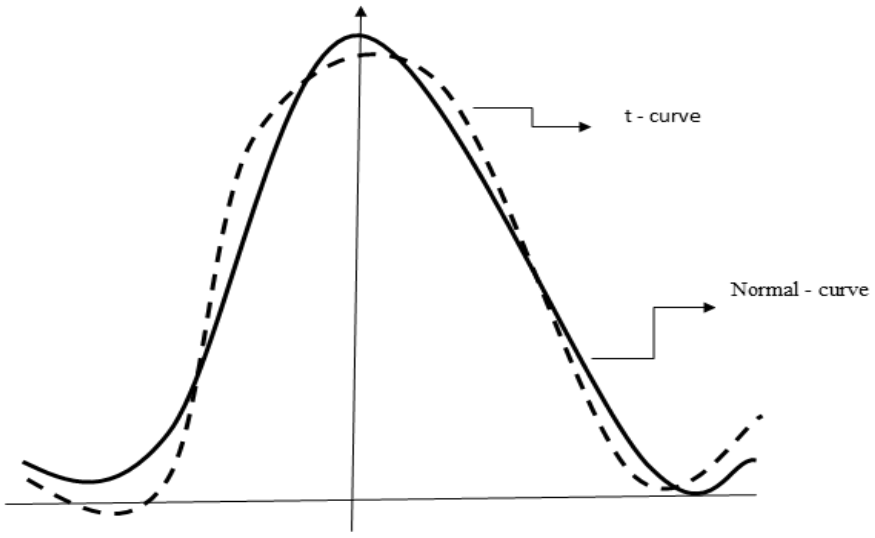


Fig - 2

- The limiting form of t – distribution when $\gamma \to \infty$ is given by $y = y_0 e^{-\left(\frac{1}{2}\right)t^2}$ which is a normal curve. This shows that t is normally distributed for large samples.
- The probability 'P' that the value of t will exceed t_0 is given by

$$P = \int_{t_0}^{\infty} y \, dx$$

- Moments about the mean.

All the moments of odd order about the mean are zero, due to its symmetry about the line t = 0.

Even order moments about the mean are,

$$\mu_r = \frac{\gamma}{\gamma-2} , \mu_4 = \frac{3\gamma^2}{(\gamma-2)(\gamma-4)} \ \ldots$$

$$\mu_{2r} = \frac{1.3.5 \ldots (2r-1)\gamma}{(\gamma-2)(\gamma-4)\ldots(\gamma-2r)}$$

SIGNIFICANE TEST OF A SAMPLE MEAN

- Given a random small sample $x_1, x_2, \ldots . x_n$ from a normal population, we have to test the hypothesis that mean of the population is μ. For this we calculate

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

$$\bar{x} = \frac{1}{n}\sum_1^n x_i , \sigma_s = \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2$$

- Then find the value of p for the given df (degrees of freedom) from the table.
- If the calculated value of **t > t at 0.05**, then the difference between mean and μ is said to be significant at 5% level of significance.
- If calculated value of **t > t at 0.01**, then the difference is said to be significant at 1% level of significance.
- If calculated value of t < t at 0.05, then the difference is said to be consistent with the hypothesis that μ is the mean of the population.

Example

- The nine items of sample have the following values : 45,47,50,52,48,47,49,53,51. Does the mean of these differ significantly from the assumed mean of 47.5?

We find the mean and S.D of the sample as follows,

x	d = x – 48	d^2
45	-3	9
47	-1	1
50	2	4
52	2	4
48	0	0
47	-1	1
49	1	1
53	5	25
51	3	9
Total	10	66

Taking A = 48, n = 9, We find the mean and S.D as

Mean of the sample = A + (∑d)/9

= 48 + 10/9 = 49.1

Variance = (∑d *d)/9 – ((∑d)/9)((∑d)/9)

= 498/81 = 6.15

Then S.D (σ) = 2.48

For the df = 9-1 = 8, the table value = 2.31

Therefore

t = (Sample mean – Population mean)(√ n) / S.D

= (49.1 – 47.5)(√900)/2.48 =1.83

I t I = 1.83

t(calculated value) = 1.83 < t(table value) = 2.31

Conclusion:

This implies that, the value of t is significant at 5% level of significance

The test provides evidence against the population mean being 47.5

Example:

- A certain stimulus is administered to each of the 12 patients give the increase in blood pressure as x: 5,2,8,-1,3,0,-2,1,5,0,4,6. Can it be concluded that the stimulus Will in general be accompanied by an increase in blood pressure?

Solution:

Given n = 12, Population Mean(μ) = 0,

Sample Mean = (5+2+8-1+3+0-2+1+5+0+4+6)/12

= 2.583,

S.D = √∑((x – mean)**2)/n-1 = 2.9571

Then by Student t - test,

t = (Sample mean – Population mean)(√ (n) / S.D

= (2.583 – 0)(√12)/2.9571 = 2.897

I t I = 2.897 > 2.201 at t - 0.05(Table value at df=11)

Conclusion:

The stimulus does not increase the blood pressure

Example:

- Eleven students were given a test in one subject, they were given a month's further tuition and a second test of equal difficulty was held at the end of it. Do the marks give evidence that the students have benefitted by extra coaching? Given that,

Students : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

Marks of I test : 23, 20, 19, 21, 18, 20,18, 17, 23, 16, 19

Marks of II test : 24, 19, 22, 18, 20, 22,20, 20, 23, 20, 17

Solution: We compute the mean and the standard deviation of the difference between the marks of the two tests as under:

Let d = marks of test I – marks of test-2, then

Mean of d = 11/11 = 1

Then the variance =((∑(d-mean of d)**2))/n-1)

= 50/10 = 5,

then the S.D = 2.24

Assuming that students have not been benefitted by extra coaching, it implies that the mean of the difference between the marks of the two tests is zero ($\mu = 0$)

Then by student t-test

t = ((mean of d - μ)/S.D) √n

= (1-0) (√11)/ 2.24

= 1.48 < t = 2.228 at t-0,05

Conclusion:

- Here t = 1.48 < t = 2.228 at t-0,05, then the value of t is not significant at 5% level of significance
- The test provides no evidence that the students have benefitted by extra coaching.

Example:

- A machinist is making engine parts with axel diameter of 0.7 inches. A random sample of 10 parts shows mean diameter as 0.742 and S.D as 0.04 inches. On the basis of this sample, would you say that the work is inferior?

Solution:

Given n = 10,

Population Mean(μ) = 0.7,

Sample Mean = 0.742,

S.D = 0.04

Then by student t - test

t = ((mean of d - μ)/S.D) √n

= (0.742-0.7) (√10)/ 0.04

= 3.16

Conclusion:

- Here t = 3.16 > 2.262 at t→0.05(table value), then the hypothesises rejected
- Hence the work is inferior

Chi – square distribution(X^2):

If **O** is the observed frequency and **E** is the expected frequency of a class interval , Then

Chi-square is defined by the relation

$$X^2 = \sum ((O - E)^{**}2) / E$$

Where the summation extends to all class intervals.

Goodness of Fit

- The value of chi-square is used to test whether the deviations of the observed frequencies from the expected frequencies are significant or not. It is also used to test how will a set of observations fit a given distribution.
- Chi-square therefore provides a test of goodness of fit and may be used to examine the validity of some hypothesis about an observed frequency distribution

- As a test of goodness of fit, it can be used to study the correspondence between theory and fact.

Procedure to test significance and goodness of fit

- Set up a null hypothesis and evaluate
- $x^2 = \sum ((O - E)^{**}2) / E$
- Find the df and read the corresponding values of X **2 at a prescribed significance level from Chi-square table

- From the Chi-square table , we can also find the probability p corresponding to the calculated values of Chi-square for the given df.
- If $p < 0.05$, The observed value of Chi-square is significant at 5% level of significance
- If $p < 0.01$, The value of significance is at 1% level.
- If $p > 0.05$, It is a good fit and the value is not significant.

Example: In experiments on pea breeding, the following frequencies of seeds were obtained.

Round and yellow	Wrinkled and yellow	Round and green	Wrinkled and green	total
315	101	108	32	556

Theory predicts that the frequencies should be in proportions 9:3:3:1. Examine the correspondence between theory and experiment.

$$\underset{313}{\frac{9}{16} \times 556} \qquad \underset{104}{\frac{3}{16} \times 556} \qquad \underset{104}{\frac{3}{16} \times 556} \qquad \underset{35}{\frac{1}{16} \times 556}$$

$\therefore \; x^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 0.51 < x^2_{0.05} = 7.815$ for 3 d.f

Hence there is a very high degree of agreement between theory and experiments.

Problem:

A die is thrown 264 times and the number appearing on the face(χ) the following frequency distribution

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

Calculate χ^2 .

Solution : Frequency in the given table are observed frequency.

Assuming that the die is unbiased

The expected number of frequencies for the numbers 1,2,3,4,5,6 to appear on the face is 44 each

Then the data is as fallows

x	1	2	3	4	5	6
Observed frequency	40	32	28	58	54	60
Expected frequency	44	44	44	44	44	44

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 22$$

The table value of Chi-square for 5 df is 11.07, then calculated value 22 > 11.07, the hypothesis is rejected

Problem :

- Fit a poisson distribution to the fallowing data and test for its goodness of fit at level of significance 0.05

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

Solution: Mean = 0.4049

Theoretical frequencies are $100 \times \frac{e^{-m} m^r}{r!}$

x	0	1	2	3	4
f	404.9	366	165.4	49.8	11.3

The calculated value of X**2 = 5.78

The table value of X**2 at 0.05 is 7.82,

Therefore 5.78 < 7.82

It is in good agreement

Problem:

Fit a poisson distribution to the fallowing data and test for its goodness of fit given that $x^2_{0.05} =$

7.815 for d.f = 4

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

Poisson distribution to fit a data is $N \times \frac{e^{-m}m^r}{r!} = 200 \times \frac{e^{-m}m^r}{r!}$

$= 121,61,15,3,0$

$$m = np = \frac{\sum fx}{N} = \frac{1}{2}$$

Therefore the new table is

x	0	1	2	3	4
Observed frequency	122	60	15	2	1
Expected frequency	121	61	15	3	0

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 0.025 < x^2{}_{0.05} = 7.815$$

There the hypothesis for goodness of fit is accepted.

Example:

- A set of similar coins is tossed 320 times then, the following results are obtained : No. of heads: 0,1,2,3,4,5 and getting head up as Frequency(f): 6,27,72,112,71,32 .

  Test the hypothesis that "the data follows the binomial distribution using chi – square test".

Solution: Probability of getting head up by tossing a single coin p(H) = ½= p(say), then,

probability of not getting head up by tossing a single coin

q(H) = ½ = q(say).

Then by fitting the binomial distribution for getting no. of heads 0,1,2,3,4,5 will be

N *p(r) = N* [q + p]**r,

Here N = ∑f = 6 + 27 + 72 + 112 + 71 + 32 = 320

N *p(r) = 320* [1/2 + 1/2]**5, where r = 0,…5 , we get

N *p(r) = 10 , 50, 100, 100, 50, 10 = Expected frequencies(E)

Then by chi- square test we get,

X**2 = ∑ ((O – E)**2)/E

= ((6-10)**2)/10 + ((27-50)**2)/50 +…. ((32-10)**2)/10

X**2 = 78.68

we have from the table value of X**2 at 0.05 = 11.07 for 5 df

Hence 78.68 > 11.07

Conclusion: Data follow the binomial distribution is **rejected**

Example:

- Genetic theory states that children having one parent of blood type A and the other type B will always be of one of three types A, AB and B and that the proportion of three types will be on an average be as 1: 2:1. A report states that out of 300 children having one A parent and B parent then, 30% were found to be type A, 45% were found to be type AB and the remainder type B, Test the hypothesis by X**2 test that "the observed results support genetic theory"(the table value of X**2 at 0.05 = 5.991 for 2 df).

Solution:

Observed frequencies of the given types are:

Type A : 30% of 300 children = 30 * 300/100 = 90

Type AB : 45% of 300 children = 45 * 300/100 = 135

Type B : 25% of 300 children = 25 * 300/100 = 75

Then we consider observed frequencies(O) as

O : 90, 135, 75

But the given genetic ratio is 1: 2: 1, then

Total = 1 + 2 + 1 = 4

Observed frequencies of the given types are:

Type A : = 1 * 300/4 = 75

Type AB : = 2 * 300/4 = 150

Type B : = 1 * 300/4 = 75

Then we consider Expected frequencies(E) as

E : 75, 150, 75

Then, by chi - square test we get

X**2 = $\sum$ ((O – E)**2)/E

= ((90-75)**2)/75 + ((135-150)**2)/150 + ((75-75)**2)/75

X**2 = 4.5

we have from the table value of X**2 at 0.05 = 5.991 for 2 df

Hence 4.5 < 5.991

Conclusion: the hypothesis that the observed results **support** genetic theory.

STOCHASTIC PROCESSES AND MARKOV CHAINS

- Stochastic process
- Probability vector
- Stochastic matrices, Fixed points
- Regular stochastic matrices
- Markov chains
- Higher transition probability
- Simple problems

Stochastic process :

- In probability theory and related fields, a stochastic or random process is a mathematical object usually defined as a collection of random variables. Historically, the random variables were associated with or indexed by a set of numbers, usually viewed as points in time, giving the interpretation of a stochastic process representing numerical values of some system randomly changing over time.

Such situations include the following:

- growth of a bacterial population,
- an electrical current fluctuating due to thermal noise
- The movement of a gas molecule.
- Jobs arrive at random points in time
- Random events of receiving the telephone calls
- Tossing the coin to expect the out comes as head or tail

Stochastic processes are widely used as **mathematical models** of systems and phenomena that appear to vary in a random manner.

They have applications in many disciplines including sciences such as

- Biology,
- Chemistry,
- Ecology,
- Neuroscience
- Physics

In technology and Engineering fields such as,

- Image processing
- Signal processing
- Information theory
- Computer science
- Cryptography
- Telecommunications

Furthermore, seemingly random changes in financial markets have motivated the extensive use of stochastic processes in finance.

Introduction:

A stochastic or random process can be defined as a collection of random variables that is indexed by some mathematical set, meaning that each random variable of the stochastic process is uniquely associated with an element in the set.

The set used to index the random variables **is called the index set.**
Historically, the index set was some subset of the real line, such as the natural numbers, giving the index set the interpretation of time.
Each random variable in the collection takes values from the same mathematical space **known as the state space**.

This state space can be,
for example : the integers, the real line or n-dimensional Euclidean space.

An increment is the amount that a stochastic process changes between two index values, often interpreted as two points in time. A stochastic process can have many outcomes, due to its randomness, and a single outcome of a stochastic process is called(among other names) a sample function or realization.

A stochastic process can also be regarded as a collection of random variables defined on a common probability space
(Ω , F , P), indexed by some set T, all take values in the same mathematical space S , which must be measurable with respect to some σ.
In other words, for a given probability space (Ω , F , P) and a measurable space (S , Σ), a stochastic process is a collection of S -valued random variables, which can be written as:
$\{ X (t) : t \in T \}$

Probability Vector

A probability vector is a vector (a column or row matrix) which is non-negative and all elements adding upto unity. The elements of a probability vector give us the outcomes which can occur, of a discrete random variable and the vector as a whole represents the probability mass function of that random variable.

Probability vector is a convenient, compact notation for denoting the behavior of a discrete random variable.

Probability Vector with each one of its components as non – negative is denoted by,
V = (V1 , V2, …….. Vn)
The sum is equal to unity,

$\sum$ Vi = 1, i = 1 to n, where Vi ≥ 0.

Stochastic Matrix

A square matrix p = [Pi j] with every row in the form of probability vector is called stochastic matrix

or

p = [Pi j] is a square matrix with each row being a probability vector

V = [3/4, 0, -1/2] → Not a probability vector
u = [-1/2, 3/2, 1] → Not a probability vector
w = [1/2, 1/2, 0] → a probability vector

$V = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{bmatrix}$ → is a Stochastic Matrix

$V = \begin{bmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{1}{3} & \frac{1}{4} \end{bmatrix}$ → is not a Stochastic Matrix

Regular Stochastic Matrix

A stochastic matrix p is said to be a regular stochastic matrix if all the entries of some power pn

are positive.

Example-1

$A = \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$

is a Stochastic Matrix since row I gives 0 + 1 = 1

also row II gives 1/2 + 1/2 = 1

Then for regular stochastic matrix if all the entries of some power p^n are positive.

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

Then

row I gives 1/2 + 1/2 = 1 and

row II gives 1/4 + 3/4 = 1

Therefore A is a regular stochastic matrix

- **Example**

  $A = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}$ is a Regular Stochastic Matrix

Properties of regular stochastic matrix p of order n

- **P has unique fixed point x = (x1 , x2, …….. xn) such that ,**
  x p = x
- **p has unique fixed probability vector**
  v = (v1 , v2, …….. vn) such that,
  v p = v
- **p2 , p3, …….. pn approaches the matrix v whose rows are each the fixed probability vector v**

Fixed probability vectors

Given a regular stochastic matrix P of order m if there exists a probability vector Q of order m such that Q P = Q, then Q is called a fixed probability vector of P

where Q = [x, y] with x + y = 1 for 2*2 matrix

Q = [x, y, z] with x + y + z = 1 for 3*3 matrix and so on.

Markov Chains

A discrete time process **{xn , n = 0,1,2,...}** with discrete state space **xn ∈ {0,1,2,...}** is a Markov chain if it has the Markov property**:**

P[Xn+1=j|Xn=i,Xn−1=in−1,...,X0=i0]= P[Xn+1=j|Xn=i]

In words, "the past is conditionally independent of the future given the present state of the process" or "given the present state, the past contains no additional information on the future evolution of the system."

The Markov property is common in probability models because, by assumption, one supposes that the important variables for the system being modelled are all included in the state space.

We consider **homogeneous Markov chains** for which **P[Xn+1=j|Xn=i] = P[X1=j|X0=i].**

Example: physical systems. If the state space contains the masses, velocities and accelerations of particles subject to Newton's laws of mechanics, the system in Markovian (but not random!)

Example:

- **speech recognition**. Context can be important for identifying words. Context can be modelled as a probability distribution for the next word given the most recent k words. This can be written as a Markov chain whose state is a vector of k consecutive words.

- **Epidemics.**

  Suppose each infected individual has some chance of contacting each susceptible individual in each time interval, before becoming removed (recovered or hospitalized). Then, the number of infected and susceptible individuals may be modelled as a Markov chain.

Transition state

A state I is said to be recurrent state if the system in this state at some step and there is a chance that it will now return to that state.

Higher Transition probability

The entry P_{ij} In the transition probability matrix p of the Markov chain in the probability that the system changes from the state Ai to Aj in n steps. j

Define Pij = PXn+1=j|Xn=i. Let P = [Pij] denote the (possibly infinite) transition matrix of the one-step transition probabilities.

Write P2 ij =P∞ k=0 PikPkj, corresponding to standard matrix multiplication.

P2 ij = X k

PXn+1=k|Xn=iPXn+2=j|Xn+1=k= X k

PXn+2=j,Xn+1=k|Xn=i (via the Markov property. Why?) = P[k Xn+2=j,Xn+1=k Xn=i = PXn+2=j|Xn=i