SAMPLING THEORY

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

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

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

A finite subset of statistical individuals in a population is called a *sample* and the number of individuals in a sample is called the sample size.

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

Sampling is quite often used in our day to day practical life. For example, in a shop we assess the quality of sugar, wheat or any other commodity by taking a handful of it from the bag and then decide to purchase it or not. A housewife normally tests the cooked products to find if they are properly cooked and contain the proper quantity of salt.

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

Consider the following example.

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

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

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

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

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

Further the data can be grouped as below.

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

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

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

Mean:

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

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

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

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

Variance:

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

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

Sampling Distributions:

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

Sampling Distribution of Means:

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

Suppose the population is finite with size N_p. Then $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$ are related to μ and σthrough the following formulae:

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

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

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

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

Problems:

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

   Solution: Given $N_p = 4$

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

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

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

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

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

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

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

   $(13-9)^2\} = \frac{20}{3}$.

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

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

2. The daily wages of 3000 workers in a factory are normally distributed with mean equal to Rs 68 and standard deviation equal to Rs 3. If 80 samples consisting of 25 workers each are obtained, what would be the mean and standard deviation of the sampling distribution of means if sampling were done (a) with replacement (b) without replacement? In how many samples will the mean is likely to be (i) between Rs 66.8 & Rs 68.3 and (ii) less than Rs 66.4?

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

In case of sampling with replacement

$\mu_{\bar{X}} = \mu = 68$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} = \frac{3}{\sqrt{25}} = 0.6$.

In case of sampling without replacement

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

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

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

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

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

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

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

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

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

Exercise:

1. The mean of a certain normal infinite population is equal to the standard error of the distribution of means of samples of size 100 drawn from that population. Find the probability that the mean of a sample of size 25 drawn from the population will be negative.
2. If the mean of an infinite population is 575 with standard deviation 8.3 how large a sample must be used in order that there be one chance in 100 that the mean of the sample is less than 572?

Answers: 1. 0.3085, 2. N = 43.

STATISTICAL DECISIONS

One of the objectives of the sampling theory is to evolve appropriate rules that enable us to make decisions about populations on the basis of the information available in respect of samples. Such decisions are called statistical decisions.

For reaching statistical decisions, we start with some assumptions or guesses about the populations involved. Such assumptions/guesses, which may or may not be true, are called statistical hypothesis.

In many situations we formulate a statistical hypothesis primarily to see whether it can be rejected (nullified). Such a hypothesis is called null hypothesis. A statistical hypothesis which differs from a given hypothesis is called an alternative hypothesis.

Procedures which enable us to decide whether to accept or reject a hypothesis or to determine whether observed samples differ significantly from expected results in regard to the corresponding populations are called Tests of Hypothesis/Tests of Significance or Rules of decision.

STANDARD ERRORS

Type I error: By an error of judgment suppose we reject a hypothesis H when it should be accepted, then we say that a Type I error has occurred.

Type II error: By an error or judgment suppose we accept a hypothesis H when it should be rejected, then we say that a Type II error has occurred.

Problems:

1. To test the hypothesis that a coin is fair, the following rule of decision is adopted. Accept the hypothesis if the number of heads in a sample of 100 tosses is between 40 and 60; reject the hypothesis otherwise. Find the probability of occurrence of Type I error.

   Solution: X – number of heads showing, N = 100 tosses. If the coin is fair, then p = 1/2. Then X is binomially distributed. Thus,

   $\text{Mean} = \mu = Np = 100 * \frac{1}{2} = 50,$

   $\text{Standard deviation} = \sigma = \sqrt{Npq} = \sqrt{100 * \frac{1}{2} * \frac{1}{2}} = 5.$

   The corresponding standard normal variate is; $z = \frac{X-\mu}{\sigma} = \frac{X-50}{\sigma}$

   $$P(40 < X < 60) = P\left(\frac{40-50}{5} < z < \frac{60-50}{5}\right)$$
   $$= P(-2 < z < 2)$$
   $$= F(2) - F(-2)$$

$$= F(2) - (1 - F(2))$$

$$= 0.9773 - 1 + 0.9773 = 0.9546 .$$

The probability of not getting heads out of 100 tosses between 40 and 60 is

$1 - 0.9546 = 0.0454$. i.e., the occurrence of Type I error is 0.0454.

2. For the purpose of estimating the mean of an infinite population with standard deviation 2.4, the following rule is adopted. The population mean is estimated as 20 if and only if the mean in a sample of size 36 is less than 20.75. Find (a) The probability of occurrence of Type I error (b) The probability of occurrence of Type II error when the population mean is taken as 21.

Solution: $\sigma = 2.4, N = 36, \mu = 20$ iff $\bar{X} < 20.75$

(a) If $\bar{X} < 20.75$ *and* $\mu = 20$, then there is no error.

If $\bar{X} < 20.75$ *and* $\mu \neq 20$, then there is Type I error.

With $\bar{X} = 20.75,\ \mu = 20, \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} = \frac{2.4}{\sqrt{36}} = 0.4$

$z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{20.75 - 20}{0.4} = 1.875$.

$P(z < 1.875) = F(1.875) = 0.9696,\ [\bar{X} < 20.75\ \&\ \mu = 20]$

Then the Type I error is $1 - P(z < 1.875) = 1 - 0.9696 = 0.0304$

(b) with $\bar{X} = 20.75$ and $\mu = 21, z = \frac{20.75 - 21}{0.4} = -0.625$

$P(z < -0.625) = F(-0.625) = 1 - F(0.625) = 1 - 0.7340 = 0.266$

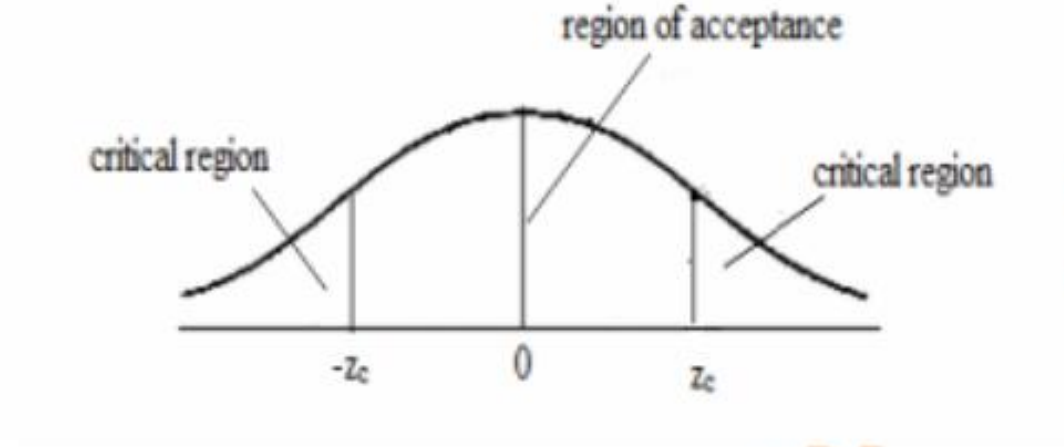
is the probability of occurrence of Type II error when $\mu = 21$.

LEVELS OF SIGNIFICANCE

Suppose that, under a given hypothesis H, the sampling distribution of a statistic S is a normal distribution with mean μ_S and standard deviation σ_S, then $z = \frac{S - \mu_S}{\sigma_S}$ is the standard normal variate associated with S, so that for the distribution of z the mean is 0 (zero) and the standard deviation is 1 (one). Accordingly, for z, the z% confidence interval is $(-z_c, z_c)$. This means that we can be z% confident that, if the hypothesis H is true, then the value of z will lie between $-z_c$ and z_c. This is equivalent to saying that (when H is true) there is $(100 - z)\%$ chance that the value of z lies outside the interval $(-z_c, z_c)$.

If we reject a true hypothesis H on the grounds that the value of z lies outside the interval $(-z_c, z_c)$, we would be making a Type I error and the probability of making this error is $(100 - z)\%$. Here we say that the hypothesis H is rejected at $(100 - z)\%$ level of significance.

The set of values of z that are outside the confidence interval $(-z_c, z_c)$ constitutes what is called the region of rejection or the region of significance or the critical region. The set of values of z that are inside the confidence interval $(-z_c, z_c)$ is called the region of acceptance.



In practice, two levels of significance are employed namely;

(i) 5% level of significance & (ii) 1% level of significance.

Level of significance	Critical value	Confidence interval
0.05	$z_c = 1.96$	(-1.96, 1.96)
0.01	$z_c = 2.58$	(-2.58, 2.58)

The value of the normal variate z (determined by using $z = \frac{S-\mu_S}{\sigma_S}$) is called the z-score.

A hypothesis is rejected at $\alpha\% = \beta$ level of significance, if the z-score is in the critical region, else the hypothesis is accepted.

Problems:

1. A coin was tossed 400 times and the head turned up 216 times. Test the hypothesis that the coin is unbiased at 5% level of significance.

Solution: X = Number of heads shown (binomial distribution),

N = 400, p = 1/2, q = 1/2.

$\mu_X = Np = 400 * \frac{1}{2} = 200$, $\sigma_X = \sqrt{Npq} = \sqrt{400 * \frac{1}{2} * \frac{1}{2}} = 10$

$X = 216$, $z = \frac{X-\mu_X}{\sigma_X} = \frac{216-200}{10} = 1.6$

For 5% level of significance, $z_c = 1.96$ and 1.6 lies in the interval (-1.96, 1.96). Therefore, the coin is unbiased.

2. Find how many heads in 64 tosses of a coin will ensure its fairness at 0.05 level of significance.

Solution: N = 64, p = 1/2, X = ?

$$\mu_X = Np = 64 * \frac{1}{2} = 32, \sigma_X = \sqrt{Npq} = \sqrt{64 * \frac{1}{2} * \frac{1}{2}} = 4$$

$$z = \frac{X - \mu_X}{\sigma_X} = \frac{X - 32}{4}$$

For 0.05 level of significance $z_c = 1.96$

$-1.96 < \frac{X-32}{4} < 1.96$ or

$-1.96 * 4 + 32 < X < 1.96 * 4 + 32$

$i.e., 24.16 < X < 39.84$ or $25 \leq X \leq 39$

Therefore, number of heads between 25 and 39 ensures the fairness of the coin with 0.05 level of significance.

3. An examination was given to two classes A & B consisting of 40 and 50 students respectively. In class A, the mean mark was 74 with a standard deviation of 8, while in class B the mean mark was 78 with a standard deviation of 7. Is there a significant difference between the performances in the two classes, at the level of significance 0.05? What about the situation at 0.01 level of significance?

Solution: μ_{X1} and μ_2 are means of two classes.

Hypothesis H: There is no difference between the performances of two classes $i.e., \mu_1 = \mu_2$. $\mu_{(\bar{X}_1 - \bar{X}_2)} = \mu_1 - \mu_2 = 0$

$$\sigma_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} = \sqrt{\frac{8^2}{40} + \frac{7^2}{70}} = 1.606$$

The z-score for the differences in means is

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_{(\bar{X}_1 - \bar{X}_2)}}{\sigma_{(\bar{X}_1 - \bar{X}_2)}} = \frac{(74-78)-0}{1.606} = -2.49$$

$z \notin (-1.96, 1.96)$. Therefore rejected at 0.05 level of significance.

i.e., there is difference in performances.

$z \in (-2.58, 2.58)$ is accepted at 0.01 level of significance. i.e., there is no difference in performances.

Exercise:

1. A die was thrown 9000 times and a throw of 5 or 6 was obtained 3240 times. On the assumption of random throwing, does the data indicate an unbiased die at 0.01 level of significance?
2. The mean life-time of a sample of 100 fluorescent tube lights manufactured by a company is found to be 1570 hours with a standard deviation of 120 hours. Test the hypothesis that the mean life-time of the lights produced by the company is 1600 hours.
3. A sample of 900 items is found to have the mean equal to 3.4. Can it be reasonably regarded as a truly random sample from a large population with mean 3.25 and standard deviation 1.61?
4. One type of aircraft is found to develop engine trouble in 5 flights out of 100 flights and another type in 7 flights out of 200 flights. Is there a significant difference in the two types of aircrafts so far as engine defects are concerned?

Answers: 1. The die is biased at 0.01 as $z > z_c$, 2. 0.01 is the level of significance, 3. No, considered sample cannot be regarded as a truly random sample, 4. There is no significant difference between the two types of aircrafts.

t – distribution: Sampling distribution on the assumption that they are normal or approximately normal is valid when the sample size N is large.

For small samples ($N \leq 30$), we consider the t-distribution.

Let N be the sample size, $\bar{X}$ and μ be respectively the sample mean and the population mean, and S be the sample standard deviation.

Consider the statistic t defined by $= \frac{(\bar{X}-\mu)}{S}\sqrt{N-1}$.

Suppose we obtain a frequency distribution of t by computing the value of t for each of a set of samples of size N drawn from a normal or a nearly normal population. The sampling distribution so obtained is called the students t-distribution.

The curve represented by the function $Y(t) = Y_0\left(1+\frac{t^2}{N-1}\right)^{\frac{-N}{2}}$ where Y_0 is a constant is called the t-curve.

The constant Y_0 is generally chosen in such a way that the total area under the curve is equal to unity.

For large values of N the function $Y(t)$ reduces to the standard normal distribution density function $\varphi(t) = \left(\frac{1}{\sqrt{2\pi}}\right)e^{\frac{-t^2}{2}}$ and the t-curve becomes the standard normal curve.

Setting $r = N - 1$, we have

$$t = \frac{\bar{X} - \mu}{S}\sqrt{r} \text{ and } Y(t) = Y_0\left(1 + \frac{t^2}{r}\right)^{\frac{-(r+1)}{2}}$$

The quantity $r = N - 1$ is called the number of degrees of freedom for the statistic t.

For population means, the confidence limits are given by $\bar{X} \pm t_C\left(\frac{S}{\sqrt{r}}\right)$, where $\pm t_C$ are the critical values or confidence coefficients whose values depend on the level of significance desired and the sample size.

For a specified r, $t_\beta(r)$ is the value of t_C at $\beta = \alpha\%$ level of significance. At the β level of significance, the confidence interval for the t-score is $(-t_\beta(r), t_\beta(r))$.

Problems:

1. For a random variable of 16 values with mean 41.5 and the sum of the squares of the deviations from the mean equal to 135 and are drawn from a normal population. Find the 95% confidence limits and the confidence interval, for the mean of the population.

   Solution: $N = 16, r = N - 1 = 15, \bar{X} = 41.5$

   $\sum S^2 = 135$. Therefore $S^2 = \frac{135}{N} = \frac{135}{16} = 8.4375$

   For 95% confidence level, the significance level is 0.05. Hence the required confidence limits are

   $$\bar{X} \pm t_C\left(\frac{S}{\sqrt{r}}\right) = 41.5 \pm t_{0.05}(r) * \frac{\sqrt{8.4375}}{\sqrt{15}}$$

   $$= 41.5 \pm 2.13 * \frac{\sqrt{8.4375}}{\sqrt{15}}$$

   $$= 43.1, 39.9$$

   Therefore, the confidence interval is (39.9, 43.1).

2. A mechanist is making engine parts with axle diameter of 0.7 inch. A random sample of 10 parts showed a mean length of 0.742 inch, with a standard deviation of 0.04 inch. On the basis of this sample, can it be concluded that the work is inferior at 5% level of significance?

   Solution: $N = 10, r = N - 1 = 9, \bar{X} = 0.742, S = 0.04, \mu = 0.7$

   Hypothesis: $\mu = 0.7$, work is not inferior

   $$t = \frac{\bar{X} - \mu}{S}\sqrt{r} = \frac{0.742 - 0.7}{0.04}\sqrt{9} = 3.16$$

   For $r = 9, t_{0.05}(r) = 2.26$ and $3.16 \notin (-2.26, 2.26)$.

   Therefore the hypothesis is rejected. At 5% level of significance, the work is inferior.

3. Find the students 't' for the following values in a sample of eight: -4, -2, -2, 0, 2, 2, 3, 3 taking the mean of the population to be zero.

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

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

$S^2 = 6.1875$. Given $\mu = 0$.

Then; $t = \frac{\bar{x}-\mu}{s}\sqrt{r} = \frac{0.25-0}{\sqrt{6.1875}}\sqrt{8-1} = 0.2659$.

4. The following are the I Qs of a randomly chosen sample of 10 boys; 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Does this data support the hypothesis that the population mean of I Qs is 100 at 5% level of significance?

Solution: N = 10

$$\bar{X} = \frac{1}{10}(70+120+110+101+88+83+95+98+107+100) = 97.2$$

$$S^2 = \frac{1}{10}\{(70-97.2)^2 + (120-97.2)^2 + (110-97.2)^2 + (101-97.2)^2 + (88-97.2) + (83-97.2)^2 + (95-97.2)^2 + (98-97.2)^2 + (107-97.2)^2 + (100-97.2)\}$$

$S^2 = 183.36$

$S = \sqrt{183.36} = 13.54$

$\mu = 100$ given

$t = \frac{\bar{x}-\mu}{s}\sqrt{r} = \frac{97.2-100}{13.54}\sqrt{9} = -0.62$ and

$t_{0.05}(r) = t_{0.05}(9) = 2.26$

$t = -0.62$, $|t| = 0.62 < t_{0.05}(9) = 2.26$

Therefore, the hypothesis is accepted.

5. A certain stimulus administered to each of 12 patients resulted in the following change in blood pressure: 5, 2, 8, -1, 3, 0, 6, -2, 1, 5, 0, 4 (in appropriate units). Can it be concluded that, on the whole, the stimulus will change the blood pressure.

Use $t_{0.05}(11) = 2.201$.

Solution: N = 12, r = 11

$$\bar{X} = \frac{1}{12}(5+2+8-1+3+0+6-2+1+5+0+4) = 2.58$$

$$S^2 = \frac{1}{12}\{(5-2.58)^2 + (2-2.58)^2 + (8-2.58)^2 + (-1-2.58)^2 + (3-2.58)^2 + (0-2.58)^2 + (6-2.58)^2 + (-2-2.58)^2 + (1-2.58)^2 + (5-2.58)^2 + (0-2.58)^2 + (4-2.58)^2\}$$

$S^2 = 8.743$

$S = \sqrt{8.743} = 2.96$

Let the hypothesis be: the stimulus does not change the blood pressure then

$\mu = 0$ then $t = \frac{\bar{X}-\mu}{S}\sqrt{r} = \frac{(2.58-0)}{2.96}\sqrt{11} = 2.89$

$2.89 > t_{0.05}(11) = 2.201$

Therefore, we reject the hypothesis. i.e., the stimulus is likely to change the blood pressure.

6. Eleven school boys were given a test in mathematics carrying a maximum of 25 marks. They were given a month's extra coaching and a second test of equal difficulty was held thereafter. The following table gives the marks in the two tests.

Boy	1	2	3	4	5	6	7	8	9	10	11
I Test Marks	23	20	19	21	18	20	18	17	23	16	19
II Test Marks	24	19	22	18	20	22	20	20	23	20	17

Do the marks give evidence that the students have benefited by extra coaching?

Use 0.05 level of significance.

Solution: The difference in the marks is given by 1, -1, 3, -3, 2, 2, 2, 3, 0, 4, -2.

$\bar{X} = \frac{1}{11}(1-1+3-3+2+2+2+3+0+4-2) = 1$

$$S^2 = \frac{1}{11}\{(1-1)^2 + (-1-1)^2 + (3-1)^2 + (-3-1)^2 + (2-1)^2 + (2-1)^2 + (2-1)^2 + (3-1)^2 + (0-1)^2 + (4-1)^2 + (-2-1)^2\}$$

$S^2 = 4.545$

$S = \sqrt{4.545} = 2.1319$

Let the hypothesis be: the students have not been benefited by extra coaching

i. e., $\mu = 0$ then $t = \frac{\bar{X}-\mu}{S}\sqrt{r} = \frac{(1-0)}{2.1319}\sqrt{10} = 1.485$

$t_{0.05}(10) = 2.23$ and $1.485 < 2.23$

Therefore, we accept the hypothesis. i.e., the students have not been benefited by extra coaching.

Chi – Square Distribution: In all random trails there exists some discrepancy between the expected (theoretical) frequencies and the observed frequencies, in general. The discrepancy is analysed through a statistic test called the chi-square, denoted by χ^2.

Suppose that, in a random experiment, a set of values $E_1, E_2, \dots, E_n$ are observed to occur with frequencies $f_1, f_2, \dots, f_n$. According to a theory based on probability rules, suppose the same events are expected to occur with frequencies $e_1, e_2, \dots, e_n$. Then, $f_1, f_2, \dots, f_n$ are called observed frequencies and $e_1, e_2, \dots, e_n$ are called expected or theoretical frequencies.

Let us define the statistic χ^2 through the following formula;

$$\chi^2 = \frac{(f_1-e_1)^2}{e_1} + \frac{(f_2-e_2)^2}{e_2} + \cdots + \frac{(f_n-e_n)^2}{e_n} = \sum_{k=1}^{n} \frac{(f_k-e_k)^2}{e_k} \qquad (1)$$

If N is the total frequency, then $N = \sum_{k=1}^{n} f_k = \sum_{k=1}^{n} e_k$.

If the expected frequencies are at least equal to 5, it can be proved that the sampling distribution of the statistic χ^2 is approximately identical with the probability distribution of the variate χ^2 whose density function is given by

$$P(\chi^2) = P_0 \chi^{\gamma-2} e^{\left(\frac{-\chi^2}{2}\right)}$$

where γ is a positive constant, called the number of degrees of freedom, and P_0 is a constant depending on γ such that the total area under the corresponding probability curve is one. The probability distribution for which $P(\chi^2)$ is given by $P_0 \chi^{\gamma-2} e^{\left(\frac{-\chi^2}{2}\right)}$ is called the chi-square distribution with γ degrees of freedom.

In practice, expected frequencies are computed on the basis of a hypothesis H_0. If under this hypothesis the value of χ^2 computed with the use of the formula (1) is greater than some critical value χ^2_C, we would conclude that the observed frequencies differ significantly from the expected frequencies and would reject H_0 at the corresponding level of significance , C. Otherwise, we would accept it or at least not reject it. This procedure is called the chi-square test of hypothesis or significance.

Generally the chi-square test is employed by taking C= = 0.05 or 0.01. The number of degrees of freedom γ is determined by the formula $\gamma = n - m$, where n is the number of frequency pairs (f_1, e_1), m is the number of quantities that are needed in the calculation of e_1.

Goodness of Fit: When a hypothesis H_0 is accepted on the basis of the chi-square test, we say that the expected frequencies calculated on the basis of H_0 form a good fit for the given

frequencies. When H_0 is rejected, we say that the corresponding expected frequencies do not form a good fit.

Problems:

1. In 200 tosses of a coin, 118 heads and 82 tails were observed. Test the hypothesis that the coin is fair at 0.05 and 0.01 levels of significance.

   Solution: Observed frequencies $f_1 = 118, f_2 = 82$.

   Expected frequencies $e_1 = 200 * \frac{1}{2} = 100$ and $e_2 = 200 * \frac{1}{2} = 100$

   $$\chi^2 = \frac{(f_1 - e_1)^2}{e_1} + \frac{(f_2 - e_2)^2}{e_2} = \frac{(118 - 100)^2}{100} + \frac{(82 - 100)^2}{100} = 6.48$$

   $$n = 2, N = \sum f_i = 200, m = 1, \gamma = n - m = 2 - 1 = 1$$

   $$\chi^2_{0.05}(1) = 3.84, \chi^2_{0.01}(1) = 6.64$$

   $\chi^2 < \chi^2_{0.01}$ and $\chi^2 > \chi^2_{0.05}$

   Therefore the hypothesis is accepted at 0.01 and rejected at 0.05.

2. The following table gives the number of road accidents that occurred in a large city during the various days of a week. Test the hypothesis that the accidents are uniformly distributed over all the days of the week.

Day	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Total
No. of accidents	14	16	08	12	11	9	14	84

   Solution: N = 84, no of days = 07. So, $e_i = \frac{84}{7} = 12$.

   $$\chi^2 = \left\{\frac{(14-12)^2}{12} + \frac{(16-12)^2}{12} + \frac{(08-12)^2}{12} + \frac{(12-12)^2}{12} + \frac{(11-12)^2}{12} + \frac{(09-12)^2}{12} + \frac{(14-12)^2}{12}\right\}$$

   $$n = 7, m = 1, \gamma = 7 - 1 = 6$$

   $$\chi^2_{0.05}(6) = 12.59, \quad \chi^2_{0.01}(6) = 16.81$$

   $\chi^2 < \chi^2_{0.05}$ and $\chi^2 = \chi^2_{0.01}$

   The hypothesis is accepted. i.e., the accidents are distributed uniformly over the week.

3. A survey of 240 families with 3 children each revealed the distribution shown in the following table. Is the data consistent with the hypothesis that male and female births are equally probable? Use χ^2 test at 0.01 & 0.05 levels.

No. of children	3 B 0 G	2 B 1 G	1 B 2 G	0 B 3 G
No. of families	37	101	84	18

Solution: Let p = probability of male births and q = probability of female births.

Under the hypothesis male and female births are equally probable p = 1/2, q = 1/2.

Among three children, the probabilities that x children are boys is given by

$$3C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{3-x}$$

Therefore, $P(3B) = 3C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^{3-3} = \frac{1}{8}$; $P(2B) = 3C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^1 = \frac{3}{8}$

$P(1B) = 3C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^2 = \frac{3}{8}$; $P(0B) = 3C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^{3-3} = \frac{1}{8}$

Therefore, among 240 families, the expected number of families with

3 B is $e_1 = 240 * P(3B) = 240 * \frac{1}{8} = 30$

2 B is $e_2 = 240 * P(2B) = 240 * \frac{3}{8} = 90$

1 B is $e_3 = 240 * P(1B) = 240 * \frac{3}{8} = 90$

0 B is $e_4 = 240 * P(0B) = 240 * \frac{1}{8} = 30$

From the table, $f_1 = 37, f_2 = 101, f_3 = 84, f_4 = 18$

$$\chi^2 = \left\{\frac{(37-30)^2}{30} + \frac{(101-90)^2}{90} + \frac{(84-90)^2}{90} + \frac{(18-30)^2}{30}\right\} = 8.1773.$$

$\gamma = 4 - 1 = 3$, $\chi^2_{0.05}(3) = 7.82$, $\chi^2_{0.01}(3) = 11.34$

$\chi^2 < \chi^2_{0.05}$ therefore consistent and $\chi^2 > \chi^2_{0.01}$ therefore inconsistent.

Problems:

1. A set of five identical coins are tossed 320 times and the result is shown in the following table:

No. of heads	0	1	2	3	4	5
Frequency	6	27	72	112	71	32

Test the hypothesis that the data follows a binomial distribution associated with a fair coin. Ans: the hypothesis is rejected.

2. The manufacturer of a certain make of electric bulbs claims that his bulbs have a mean life of 25 months with a standard deviation of 5 months. Random samples of 6 such bulbs have the following values: Life of bulbs in months: 24, 20, 30, 20, 20, and

18. Can you regard the producer's claim to valid at 1% level of significance? (Given that $t_{tab} = 4.032$ corresponding to $\gamma = 5$). Ans: the hypothesis is accepted.

3. A certain stimulus administered to each of the 13 patients resulted in the following increase of blood pressure: 5, 2, 8,-1, 3, 0, -2, 1, 5, 0, 4, 6, 8. Can it be concluded that the stimulus, in general, be accompanied by an increase in the blood pressure?
   Ans: the hypothesis is accepted.
4. The life time of electric bulbs for a random sample of 10 from a large consignment gave the following data: 4.2, 4.6, 3.9, 4.1, 5.2, 3.8, 3.9, 4.3, 4.4, 5.6 (in 4,000 hours). Can we accept the hypothesis that the average life time of bulbs is 4, 000 hours?
   Ans: Yes the mean life of time bulbs is about 4, 000 hours.
5. The following table gives the frequency of occupancy of the digits 0, 1, …, 9 in the last place in four logarithm of numbers 10-99. Examine if there is any peculiarity.

Digits:	0	1	2	3	4	5	6	7	8	9
Frequency:	6	16	15	10	12	12	3	2	9	5

   Ans: H_0 is rejected. i.e., there is no peculiarity between the digits.

6. The sales in a supermarket during a week are given below. Test the hypothesis that the sales do not depend on the day of the week, using a significant level of 0.05.

Days:	Mon	Tue	Wed	Thu	Fri	Sat
Sales (in 1000 Rs)	65	54	60	56	71	84

   Ans: H_0 is accepted.

Video Links:

1. Probability distributions

https://www.youtube.com/watch?v=c06FZ2Yq9rk
https://www.youtube.com/watch?v=N-IVFB8RIfo
https://www.youtube.com/watch?v=d5iAWPnrH6w
https://www.youtube.com/watch?v=yiXLH7EXri8

2. Sample theory

https://www.youtube.com/watch?v=zK70Fc_HHmg
https://www.youtube.com/watch?v=mlORIoBErso