UNIT-IV

SAMPLING AND ESTIMATION

SAMPLING THEORY

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

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

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

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

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

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

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

Consider the following example.

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

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

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. Then $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ are related to μ and σthrough the following formulae:

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

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

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

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

$$Z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}}.$$

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 with and without replacement. Find the mean and standard deviation in the population, and in the sampling distribution of means verify the formulas $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$.

   Solution: Given $N = 4$

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

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

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

   Case i) Sampling with replacement:

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

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

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

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

$= 5$

Standard variation= $\sigma_{\bar{x}} = \sqrt{5}$

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

and

$$\frac{\sigma}{\sqrt{n}} = \frac{\sqrt{20}}{\sqrt{4}} = \sqrt{5} = \sigma_{\bar{x}}$$

Case ii) Sampling without replacement:

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

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

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

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

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

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

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

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

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

In case of sampling with replacement

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

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

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

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

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

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

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

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

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

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

Solution:

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

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

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

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

4. An automobile battery manufacturer claims that its midgrade battery has a mean life of 50 months with a standard deviation of 6 months. Suppose the distribution of battery lives of this particular brand is approximately normal. On the assumption that the manufacturer's claims are true, find (a) the probability that a randomly selected battery of this type will last less than 48 months, (b) the probability that the mean of a random sample of 36 such batteries will be less than 48 months.

Solution:

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

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

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

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

Exercise:

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

   Answers: 1) With replacement: $\mu_{\bar{x}} = 5$ and $\sigma_{\bar{x}} = \frac{\sqrt{13}}{2}$, Without replacement: $\mu_{\bar{x}} = 5$ and $\sigma_{\bar{x}} = \sqrt{\frac{13}{6}}$, 2) With replacement: Mean= 100, SD= 8, Without replacement: mean=100, SD 7.95, 3) With replacement: P=0.46, Without replacement: P=0.47, 4) 0.3085, 5) n = 43. 6) 0.9946

Sampling distribution of mean (σ unknown): t – distribution

Sampling distribution on the assumption that they are normal or approximately normal is valid when the sample size n is large. However, if the sample size n is small, the distribution of various statistics are far from the normality.

For small samples $(n < 30)$, 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.

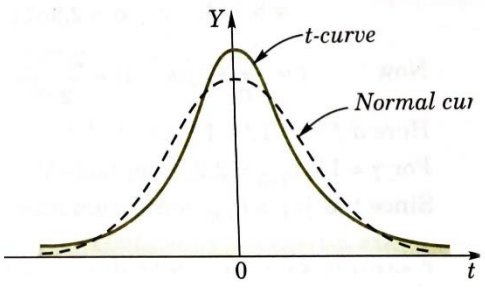
Consider the statistic t defined by $t = \frac{(\bar{x}-\mu)}{s}\sqrt{n}$.

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 Student's t-distribution with probability density function $Y(t) = Y_0\left(1 + \frac{t^2}{\gamma}\right)^{\frac{-(\gamma+1)}{2}}$ with $\gamma = n - 1\ d.f.$ where Y_0 is a constant and $-\infty < t < \infty$. The curve $Y(t)$ 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.

The t-distribution curve is symmetric about the mean 0, unimodal and bell shaped and asymptotic on both sides of t −axis. Thus the t-distribution curve is similar to normal curve. While the variance for normal distribution is 1, the variance for the t-distribution is more than 1 since it depends on the parameter γ. So the t-distribution is more variable. As $n \to \infty$, variance of t-distribution approaches 1. Thus as $\gamma = n - 1 \to \infty$, t-distribution approaches the standard normal distribution.



Note: The n-1 degrees of freedom in the t-distribution account for the fact that one degree of freedom is used to estimate the sample mean when calculating the sample variance. This adjustment allows the t-distribution to provide a better approximation to the actual distribution of sample means, accounting for the uncertainty introduced by estimating population parameters from sample data.

Applications of t-distribution:

- To estimate the population mean from a sample mean when the population standard deviation is unknown and estimated using the sample standard deviation.
- To test the hypothesis of the population mean.
- To test the hypothesis of the difference of two population means of two normal population.

Problems:

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

   $$s^2 = \frac{1}{7}\{(-4-0.25)^2 + (-2-0.25)^2 + (-2-0.25)^2 + (0-0.25)^2 + (2-0.25)^2 + (3-0.25)^2 + (3-0.25)^2\}$$

   $s^2 = 6.6339$. Given $\mu = 0$.

   Then; $t = \frac{\bar{X}-\mu}{s}\sqrt{n} = \frac{0.25-0}{\sqrt{6.6339}}\sqrt{8} = 0.2745$.

2. The following are the I Qs of a randomly chosen sample of 10 boys;
   70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Find the students 't' for the given sample taking the mean of population to be 100.

   Solution: n = 10

   $$\bar{x} = \frac{1}{10}(70+120+110+101+88+83+95+98+107+100)$$
   $$= 97.2$$
   $$s^2 = \frac{1}{9}\{(70-97.2)^2 + (120-97.2)^2 + (110-97.2)^2 + (101-97.2)^2 + (88-97.2) + (83-97.2)^2 + (95-97.2)^2 + (98-97.2)^2 + (107-97.2)^2 + (100-97.2)\}$$
   $$s^2 = 203.73$$
   $$s = \sqrt{203.73} = 14.27$$
   $\mu = 100$ given
   $$t = \frac{\bar{x}-\mu}{s}\sqrt{n} = \frac{97.2-100}{14.27}\sqrt{10} = -0.62$$

3. 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). Find the students 't' for the given sample assuming $\mu = 0$.

Solution: $N = 12$, $n = 11$

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

$$= 2.58$$

$$s^2 = \frac{1}{11}\{(5 - 2.58)^2 + (2 - 2.58)^2 + (8 - 2.58)^2 + (-1 - 2.58)^2 + (3 - 2.58)^2 + (0 - 2.58)^2 + (6 - 2.58)^2 + (-2 - 2.58)^2 + (1 - 2.58)^2 + (5 - 2.58)^2 + (0 - 2.58)^2 + (4 - 2.58)^2\}$$

$$s^2 = 9.538$$

$$s = \sqrt{9.538} = 3.088$$

given

$\mu = 0$ then $t = \frac{\bar{x} - \mu}{s}\sqrt{n} = \frac{(2.58 - 0)}{3.088}\sqrt{12} = 2.89$

4. 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

Find the t for difference in marks assuming the mean $\mu = 0$.

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

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

$$s^2 = \frac{1}{10}\{(1 - 1)^2 + (-1 - 1)^2 + (3 - 1)^2 + (-3 - 1)^2 + (2 - 1)^2 + (2 - 1)^2 + (2 - 1)^2 + (3 - 1)^2 + (0 - 1)^2 + (4 - 1)^2 + (-2 - 1)^2\}$$

$$s^2 = 5$$

$$s = \sqrt{5} = 2.2361$$

Given

$\mu = 0$ then $t = \frac{\bar{x}-\mu}{s}\sqrt{n} = \frac{(1-0)}{2.2361}\sqrt{10} = 1.4142$

Problems:

1. 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). Find the students 't' for the given sample taking the mean of the population to be 0.
2. Nine items of a sample have the following values: 45, 47, 50, 52, 48, 47, 49, 53, 51. Find the students 't' for the given sample taking the mean of the population to be 47.5.

Answers: 1) 2.89, 2) 1.83

Sampling distribution of variance:

The sampling distribution of the variance is the distribution of sample variances, with all samples having same sample size N taken from the same population.

Chi – Square Distribution:

Sampling distribution of some important statistics allow us to learn information about the parameters. Generally, the parameters are the counterpart to the statistics in questions. For example, if an engineer is interested in the population mean resistance of a certain type of resistor, the sampling distribution of mean will be exploited once the sample information is gathered. On the other hand, if the variability in the resistance is to be studied clearly the sampling distribution of variance will be used to know about the parameter, the population variance.

If s^2 is the variance of a random sample of size n taken from a normal population having variance σ^2, then the statistic

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

has a chi-squared distribution with $\nu = n - 1$ degrees of freedom.

The probability density function for χ^2-distribution with ν d.f. is given by

$$f(\chi^2) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} e^{-\frac{\chi^2}{2}}(\chi^2)^{\left(\frac{\nu}{2}\right)-1} \quad ; \quad 0 < \chi^2 < \infty$$

where, $\nu = n - 1$.

Suppose that, in a random experiment, a set of events $E_1, E_2, E_3 \ldots \ldots E_n$ are observed to occur with frequencies $f_1, f_2, f_3, \ldots \ldots f_n$. According a theory based on probability rules, suppose the same event are expected to occur with the frequencies $e_1, e_2, e_3 \ldots \ldots e_n$. Then

$f_1, f_2, f_3, \ldots\ldots f_n$ are called the observed frequencies and $e_1, e_2, e_3 \ldots\ldots e_n$ are called the expected or theoretical frequencies. The statistic χ^2 is defined through the following formula:

$$\chi^2 = \frac{(f_1 - e_1)^2}{e_1} + \frac{(f_2 - e_2)^2}{e_2} + \cdots\ldots\ldots\ldots + \frac{(f_n - e_n)^2}{e_n} = \sum_{i=1}^{n} \frac{(f_i - e_i)^2}{e_i}$$

If N is the total frequency, we should have

$$N = \sum_{i=1}^{n} f_i = \sum_{i=1}^{n} e_i$$

If the expected frequencies are at least equal to 5, then it can be proved that the sampling distribution of statistic χ^2 is approximately identical with the probability distribution of the variate χ^2 whose density function is given by $P(\chi^2) = P_0 \chi^{\nu-2} e^{-\frac{\chi^2}{2}}$, where ν is a positive constant, called the number of degrees of freedom, P_0 is such that the total area under the corresponding probability curve is one.

Applications of χ^2-distribution:

The applications of χ^2-distribution are very wide in statistics. It is used:

- To test the hypothetical value of population variance.
- To test the goodness of fit, that is, to judge whether there is a discrepancy between theoretical and experimental observations.
- To test the independence of two attributes, that is, to judge whether the two attributes are independent.

Problems:

1. A manufacturer of car batteries guarantees that the batteries will last, on average, 3 years with a standard deviation of 1 year. Assuming the battery lifetime follows a normal distribution, find χ^2 for the life time 1.9, 2.4, 3.0, 3.5 and 4.2 years of five of these batteries.

Solution:

Given $\mu = 3$ and $\sigma = 1$.

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

$$\bar{x} = \frac{1.9 + 2.4 + 3.0 + 3.5 + 4.2}{5} = 3$$

$$s^2 = \frac{(1.9-3)^2 + (2.4-3)^2 + (3.0-3)^2 + (3.5-3)^2 + (4.2-3)^2}{5-1} = 0.815$$

$$\chi^2 = \frac{4 \times 0.815}{1} = 3.26$$

2. Pull-strength tests on 5 soldered leads for a semiconductor device yield the following results, in pounds of force required to rupture the bond: 19.8, 12.7, 13.2, 16.9 and 10.6. If the pull-strength of the population follows normal distribution with mean 15 pounds of force and standard deviation of 0.5.

Solution:

Given $\mu = 15$ and $\sigma = 4$.

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

$$\bar{x} = \frac{19.8 + 12.7 + 13.2 + 16.9 + 10.6}{5} = 14.64$$

$$s^2 = \frac{53.892}{5-1} = 13.473$$

$$\chi^2 = \frac{4 \times 13.473}{16} = 3.368$$

3. In 200 tosses of a coin, 118 heads and 82 tails were observed. Find the χ^2.

Solution:

Here the observed frequencies of heads and tails are $f_1 = 118$ and $f_2 = 82$ respectively.

Also $N = \sum f_i = f_1 + f_2 = 200$ =number of trials(tosses).

Expected frequencies of heads and tails in 200 trials are $e_1 = 200 \times \frac{1}{2} = 100$ and $e_2 = 200 \times \frac{1}{2} = 100$ respectively. Also we note that $\sum e_i = 200$.

We have $\chi^2 = \frac{(f_1 - e_1)^2}{e_1} + \frac{(f_2 - e_2)^2}{e_2}$

$$= \frac{(118-100)^2}{100} + \frac{(82-100)^2}{100} = 6.48$$

4. A die is thrown 60 times and the frequency distribution for number appearing on the face x is given by the following table:

x	1	2	3	4	5	6
frequency	15	6	4	7	11	17

Find the statistic χ^2.

Solution:

The observed frequencies from the table are

$$f_1 = 15, f_2 = 6, f_3 = 4, f_4 = 7, f_5 = 11, f_6 = 17.$$

The expected frequencies are $e_1 = 60 \times \frac{1}{6} = 10 = e_2 = e_3 = e_4 = e_5 = e_6$.

Using the formula $\chi^2 = \sum_{i=1}^{n} \frac{(f_i - e_i)^2}{e_i} = \frac{1}{10}\{(15-10)^2 + (6-10)^2 + (4-10)^2 + (7-10)^2 + (11-10)^2 + (17-10)^2\}$=13.6

5. Fit a binomial distribution to the following data:

x_i	0	1	2	3	4	5
f_i	2	14	20	34	22	8

Find the corresponding theoretical estimates for f_i. Also find the statistic χ^2.

Solution:

Here $\sum f_i = 100$

Mean of the given distribution is

Mean $= \frac{\sum x_i f_i}{\sum f_i} = \frac{284}{100} = 2.84$

For binomial distribution Mean=np

$$\Rightarrow p = \frac{Mean}{n} = \frac{2.84}{5} = 0.568$$

$P(x) = {}^nCx\, p^x q^{1-x} = {}^5C_x (0.568)^x (0.432)^{1-x}$

Substituting $x = 0,1,2,3,4,5$ in the above, we get

$$P(0) = 1.505, P(1) = 10.4512, P(2) = 26.01, P(3) = 34.199, P(4) = 22.483,$$
$$P(5) = 5.912$$

Hence the expected frequencies are

$$e_1 = 1.505, e_2 = 10.4512, e_3 = 26.01, e_4 = 34.199, e_5 = 22.483,$$
$$e_6 = 5.912$$

Sum of the observed frequencies is $\sum f_i = 100$ whereas the sum of the expected frequencies is $\sum e_i = 100.5602$. In order to have $\sum f_i = \sum e_i$, we modify the theoretical frequencies by subtracting 0.5602 from e_3 where the discrepancy between the theoretical and observed frequencies is largest. Thus we take the modified

$$e_3 = 26.01 - 0.5602 = 25.4498$$

We further note that the first of the expected frequencies, namely $e_1 = 1.505$ is less than 5. Therefore we club with e_2 to get $e_1 + e_2 = 11.9562$. Correspondingly we club the observed frequencies f_1 and f_2 to get $f_1 + f_2 = 16$

Now we find $\chi^2 = \sum_{i=1}^{n} \frac{(f_i - e_i)^2}{e_i} = \frac{(16-11.9562)^2}{11.9562} + \frac{(20-25.4498)^2}{25.4498} + \frac{(34-34.199)^2}{34.199} + \frac{(22-22.483)^2}{22.483} + \frac{(8-5.912)^2}{5.912} = 3.28$

Problems:

1. A maker of a certain brand of low-fat cereal bars claims that the average saturated fat content is 0.5 gram and standard deviation 0.05. Find χ^2 of a random sample of 8 cereal bars of this brand with the saturated fat content 0.6, 0.7, 0.7, 0.3, 0.4, 0.5 0.4 and 0.2.

2. Fit a Poisson distribution to the following data, obtain the theoretical frequencies and find the χ^2.

x_i	0	1	2	3	4
f_i	419	352	154	56	19

3. The following table gives the number of road accidents that occurred in a large city during the various days of a week. Calculate χ^2 for the above data.

Day	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Total
No. of accidents	14	16	8	12	11	9	14	84

4. Fit a normal distribution to the following data of weights of 100 students of a certain college, obtain the theoretical frequencies and hence find χ^2 for the above data.

Weights (Kgs)	60-62	63-65	66-68	69-71	72-74
Frequency	5	18	42	27	8

5. Fit a binomial distribution to the following data:

x_i	0	1	2	3	4	5
f_i	38	144	342	287	164	25

Find the corresponding theoretical estimates for f_i. Also find the statistic χ^2.

Answers: 1) 2.52, 2)5.8494, 3) 4.17, 4) 0.262

Estimation

Point estimation of a parameter is a statistical estimation where the parameter is estimated by a single number (or value) from sample data.

Method of Maximum Likelihood

One of the best methods of obtaining a point estimation of a parameter is the method of maximum likelihood. This technique was developed in the 1920s by famous British statistician Sir R. A. Fisher. As the name implies, the estimator will be the value of the parameter that maximizes the likelihood function.

In the case of a discrete random variable, the interpretation of the likelihood function is simple. The likelihood function of the sample $L(\theta)$ is just the probability.

$$P(X_1 = x_1, \quad X_2 = x_2, \dots\dots\dots\dots\dots\dots X_n = x_n)$$

That is, $L(\theta)$ is just the probability of obtaining the sample values $x_1, x_2, \dots\dots x_n$. Therefore, in the discrete case, the maximum likelihood estimator is an estimator that maximizes the probability of occurrence of the sample values.

Example1: **Bernoulli Distribution MLE**

Let X be a Bernoulli random variable. The probability mass function is

$$f(x;p) = \begin{cases} p^x(1-p)^{1-x}, & x = 0,1 \\ 0, & \text{otherwise} \end{cases}$$

where p is the parameter to be estimated. The likelihood function of a random sample of size n is

$$L(p) = p^{x_1}(1-p)^{1-x_1}\, p^{x_2}(1-p)^{1-x_2} \dots\dots\dots\dots\dots\dots\dots p^{x_n}(1-p)^{1-x_n}$$

$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}\,(1-p)^{n-\sum_{i=1}^{n} x_i}$$

We observe that if $\hat{p}$ maximizes $L(p)$, $\hat{p}$ also maximizes $\ln L(p)$. Therefore.

$$\ln L(p) = \left(\sum_{i=1}^{n} x_i\right) \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)$$

Now,

$$\frac{d \ln L(p)}{dp} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p}$$

Equating this to zero and solving for p yields $\hat{p} = \left(\frac{1}{n}\right)\sum_{i=1}^{n} x_i$. Therefore, the maximum likelihood estimator of p is

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Example 2: **Poisson Distribution MLE**

Let X be a Poisson variate with probability mass function

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

distributed with unknown μ and known variance σ^2 . The likelihood function of a random sample of size n, say $X_1, X_2, \dots\dots X_n, is$

$$L(x_1, x_2, x_3 \dots x_n; \lambda) = \prod_{i=1}^{n} f(x:\lambda) = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x!}$$

Now

$$\ln L(x_1, x_2, x_3 \ldots x_n; \lambda) = -n\lambda + ln\lambda \sum_{i=1}^{n} x_i - ln \prod_{i=1}^{n} x!$$

And $\frac{\partial \ln L(x_1,x_2,x_3 \ldots x_n;\lambda)}{\partial \lambda} = -n + \sum_{i=1}^{n}(\frac{x_i}{\lambda})$

Equating this last result to zero and solving for λ yields

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Conclusion: The sample mean is the maximum likelihood estimator of λ.

Example 3: **Exponential Distribution MLE**

Let X be exponentially distributed with parameter λ. The likelihood function of a random sample of size n, say, $X_1, X_2, \ldots \ldots X_n, is$

$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$

The log likelihood is.

$\ln L(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$

Now

$\frac{d \ln L(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$

and upon equating this last result to zero we obtain

$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$

Conclusion: Thus, the maximum likelihood estimator of λ is the reciprocal of the sample mean.

Example 4: **Normal Distribution MLEs for μ and σ^2**

Let X be normally distributed with mean μ and variance σ^2 , where both μ and σ^2 are unknown. The likelihood function for a random sample of size n is.

$$L(\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}=\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

And

$$\ln L(\mu,\sigma^2)=-\left(\frac{n}{2}\right)\ln(2\pi\sigma^2)-\left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^{n}(x_i-\mu)^2$$

Now

$$\frac{\partial \ln L(\mu,\sigma^2)}{\partial\mu}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0$$

$$\frac{\partial \ln L(\mu,\sigma^2)}{\partial\sigma^2}=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2=0.$$

The solutions to the above equations yield the maximum likelihood estimators.

$$\hat{\mu}=\bar{x}\qquad\qquad \widehat{\sigma^2}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Video Links:

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

https://www.youtube.com/watch?v=mlQRIoBErso

Disclaimer: The content provided is prepared by department of Mathematics for the specified syllabus by using reference books mentioned in the syllabus. This material is specifically for the use of RVCE students and for education purpose only.