MODULE-3

Statistical Inference-1: Introduction. Sampling distributions, standard error. Levels of significance. Test of significances. Problems. Levels of significance. Confidence limits. Simple sampling of attributes, Test of significance for large samples. Comparison of large samples.

Population: A population consists of the totality of the observations with which we are concerned.

Examples: Groups of people, animals, or all possible outcomes of some system.

Sampling: A small section selected from the population is called a **sample**, and the process of drawing a sample is called **sampling**. It is essential that a sample must be a random selection.

Simple sampling: A random sampling in which each event has the same probability p of success and the chance of success of different events are independent whether previous trials have been made or not, is known as simple sampling.

Parameters: The statistical constants of the population such as mean (μ), standard deviation (σ) etc. are called the **parameters**.

Statistic: The statistical constants for the sample drawn from the given population such as mean $(\bar{x})$, standard deviation (S) etc. are called the **Statistic**.

Generalization from the sample to population is called **Statistical inference**.

Sampling distribution: Consider all possible samples of size n which can be drawn from a given population at random. Frequency distribution of different means of samples is called sampling distribution of the means. Frequency distribution of different standard deviation of samples is called sampling distribution of the S.D. etc.

Standard error: The standard deviation of the sampling distribution is called standard error.(S.E.)

Thus the standard error of the sampling distribution of the means is called standard error of means.

Precision: The reciprocal of the standard error is called precision.

Statistical hypothesis: To take the decisions about populations on the basis of sample information, we make certain assumptions about the populations, such assumptions are called statistical hypothesis.

Testing a hypothesis: First assume that hypothesis is correct, and then compute the probability of observed sample. If this probability is less than the pre assigned value, then hypothesis is rejected.

Errors:

Type I error: If a hypothesis is rejected while it should have been accepted, then we say that type I error has been committed.

Type II error: If a hypothesis is accepted while it should have been rejected, then we say that type II error has been made.

Null hypothesis: The hypothesis formulated for the sake of rejecting it, under the assumption that it is true, is called null hypothesis and is denoted by H_0 .

Level of significance: The probability level below which the hypothesis is rejected is called level of significance.

Critical region: The region in which a sample value falling is rejected, is known as critical region.

Test of significance: The procedure which enables us to decide whether to accept or reject the hypothesis is called test of significance.

Confidence limits: 95% confidence limits for sample statistic S to estimate μ are $S \pm 1.96\sigma$.
And 99% confidence limits for sample statistic S to estimate μ are $S \pm 2.58\sigma$.

Simple sampling of attributes: The expected value of success in a sample of size n is np, and standard deviation is $\sqrt{npq}$.

Mean proportion of successes $= \frac{np}{n} = p$.

Standard error of proportion of successes$= \sqrt{\frac{pq}{n}}$.

Precision of the proportion of successes$= \sqrt{\frac{n}{pq}}$.

Test of significance for large samples: If x be the observed number of successes in the large sample and z is the standard normal variate then $z = \frac{x-\mu}{\sigma}$.

1. If $|z| < 1.96$, difference between the observed and expected number of successes is not significant.
2. If $|z| > 1.96$, difference is significant at 5% level of significance.
3. If $|z| > 2.58$, difference is significant at 1% level of significance.

Examples:

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: Suppose the coin is unbiased, then the probability of getting the head in each toss is 0.5.

Therefore expected number of successes is $\mu = np = 0.5 \times 400 = 200$.

And the observed value of successes is $x = 216$.

Since $\mu = 200$, $\sigma = \sqrt{npq} = \sqrt{100} = 10$. $z = \frac{x-\mu}{\sigma} = \frac{16}{10} = 1.6 < 1.96$.

And hence difference between the observed and expected number of successes is not significant.

That is the coin is unbiased at 5% level of significance.

2. 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: Suppose the die is unbiased. Then the probability of throwing 5 or 6 in each throw is $p = \frac{1}{3}$.

Therefore expected number of successes is $\mu = np = \frac{9000}{3} = 3000$.

And the observed value of successes is $x = 3240$.

Since $\mu = 3000$, $\sigma = \sqrt{npq} = \sqrt{2000} = 44.7214$. $z = \frac{x-\mu}{\sigma} = \frac{240}{44.7214} = 5.3666 > 2.58$.

And hence difference is significant at 1% level of significance. And hypothesis is rejected at 1% level of significance. That is the die is biased.

3. 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 3000 or less. It is desired to estimate how many out of 18,000 families have a monthly income of Rs 3000 or less. Whiten what limits would you place your estimate in 1% level of significance?

Solution: Here $p = \frac{206}{840} = \frac{103}{420}$, $q = \frac{317}{420}$.

$\therefore$ standard error of the population of families having monthly income of Rs 3000 or less s

$= \sqrt{\frac{pq}{n}} = \sqrt{\frac{103\times317}{420\times420\times840}} = 0.0148 = 1.48\%$.

Since $= \frac{206}{840}$, Mean proportion of successes is 24.52% .

Limits are $(24.52 \pm 2.58 \times 1.48)$ That is 20.7% to 28.34 %.

Therefore 3726 to 5101 families are expected to have monthly income of Rs 3000 or less.

Exercise:

1. A die is tossed 960 times and it falls with 5 upwards 184 times. Is the die biased?
2. 12 dice are thrown 3086 times and a throw of 2, 3, 4 is reckoned as a success. Suppose that 19142 throws of 2, 3, 4 have been made out. Do you think that this observed value deviates from the expected value? If so, can the deviation from the expected value be due to fluctuations of simple sampling?
3. Balls are drawen from a bag containing equal number of black and white balls. Each ball being replaced before drawing another. In 2250 drawings 1018 black and 1232 white balls have been drawn. Do you suspect some bias on the part of drawer ?
4. A sample of 1000 days is taken from meteorological records of certain district and 120 of them are found to be foggy. What are the probable limits to the percentage of foggy days in the district?
5. In a group of 50 first cousins there were found to be 27 males and 23 females. Ascertain if the observed proportions are inconsistent with the hypothesis that the sexes should be in equal proportion.
6. A random sample of 500 apples was taken from a large consignment and 65 were found to be bad. Estimate the proportion of the bad apples in the consignment and standard error of the estimate. Deduce that the percentage of bad apples in the consignment is between 8.5 and 17.5 .
7. 400 children are chosen in an industrial town and 150 are found to be underweight. Assuming the conditions of simple sampling, estimate the percentage of children who are underweight in the industrial town and assign limits within which the percentage probably lies.

8. In a sample of 500 people from a state 280 take tea, and rest take coffee. Can we assume that tea and coffee are equally popular in the state at 5% level of significance?

Comparison of large samples: Two large samples of sizes n_1 and n_2 are taken from two populations giving mean proportion of successes are p_1 and p_2 respectively.

1. If the proportions are similar in the two populations,

Then common mean proportion of successes is $p = \frac{n_1p_1+n_2p_2}{n_1+n_2}$.

If e be the standard error of the difference between p_1 and p_2 , then $e^2 = \frac{pq}{n_1} + \frac{pq}{n_2}$.

2. If the proportions are not same in the two populations,

Then $e^2 = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}$

$\therefore z = \frac{p_1 \sim p_2}{e}$.

And if $z > 2.58$, the difference between p_1 and p_2 is real one.

If $z < 1.96$, the difference may be due to fluctuations of simple sampling.

If $1.96 < z < 2.58$, the difference is significant at 5% level of significance.

Examples:

1. In a city A 20% of a random sample of 900 school boys had a certain slight physical defect. In another city B, 18.5% of a random sample of 1600 school boys had the same defect. Is the difference between the proportions significant?

Ans: Given that $n_1 = 900, \quad n_2 = 1600, \quad p_1 = 0.2, \ p_2 = 0.185$

$$\therefore \ p = \frac{n_1p_1+n_2p_2}{n_1+n_2} = 0.19, \qquad q = 1 - p = 0.81.$$

$$e^2 = \frac{pq}{n_1} + \frac{pq}{n_2} = 0.00027 \quad \Longrightarrow \quad e \approx 0.016.$$

$$\therefore z = \frac{p_1 \sim p_2}{e} = \frac{0.015}{0.016} = 0.093 < 2,$$ The difference between the proportions is not significant.

2. In two large populations there are 30% and 25% respectively of fair haired people. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations?

Ans: Given that $p_1 = 0.3, \ p_2 = 0.25$ and for $n_1 = 1200, \quad n_2 = 900,$

$$e^2 = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2} = 0.00038 \qquad \Longrightarrow \quad e \approx 0.0195.$$

$$\therefore z = \frac{p_1 \sim p_2}{e} = \frac{0.05}{0.0195} \approx 2.5,$$ The difference between the proportions is significant.

And hence it is unlikely that the difference will be hidden.

Exercise:

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