UNIT- I

Population: In statistics, a population is a set of similar items or events which is of interest for some question or experiment. A statistical population can be a group of actually existing objects (e.g. the set of all stars within the Milky Way galaxy) or a hypothetical and potentially infinite group of objects conceived as a generalization from experience (e.g. the set of all possible hands in a game of poker). So, in short, Population is the collection of all individuals or items under consideration in a statistical study. A common aim of statistical analysis is to produce information about some chosen population. In statistical inference, a subset of the population (a statistical sample) is chosen to represent the population in a statistical analysis. If a sample is chosen properly, characteristics of the entire population that the sample is drawn from can be estimated from corresponding characteristics of the sample.

Sample: In statistics and quantitative research methodology, a data sample is a set of data collected and/or selected from a statistical population by a defined procedure. The elements of a sample are known as sample points, sampling units or observations. Typically, the population is very large; making a census or a complete enumeration of all the values in the population is either impractical or impossible. Sample is that part of the population from which information is collected. The sample usually represents a subset of manageable size. Samples are collected and statistics are calculated from the samples so that one can make inferences or extrapolations from the sample to the population. The data sample may be drawn from a population without replacement, in which case it is a subset of a population; or with replacement, in which case it is a multisubset.

Parameters: The various constants such as; mean, variance, correlation coefficient etc. of the population are known as parameters. The parameters are the functions of Population observation. A parameter is usually unknown and is estimated from the

sample. So the inference about some specific unknown parameter is based on a statistic.

Statistic: The estimator which is used to estimate the population is called a statistic. A Statistic is also a constant such as; mean, median, variance, correlation coefficient etc. computed from the sample observations. Hence, a statistic is a function of sample observations and is used to make inference about parameters. The primary focus of most research studies is the parameter of the population, not statistics calculated for the particular sample selected. The sample and statistics describing it are important only in so far as they provide information about the unknown parameters.

Sampling distribution of a statistic:

Suppose we have a population of size 'N' and we are interested to draw a sample of size 'n' from the population. In different time if we draw the sample of size n, we get different samples of different observations i.e. we can get ${}^{N}c_{n}$ possible samples. If we calculate some particular statistic from each of the ${}^{N}c_{n}$ samples, the distribution of sample statistic is called sampling distribution of the statistic. For example if we consider the mean as the statistic, then the distribution of all possible means of the samples is a distribution of the sample mean and it is called sampling distribution of the mean. 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. For example, consider a normal population with mean μ and variance σ². Assume we repeatedly take samples of a given size from this population and calculate the arithmetic mean $\bar{x}$ for each sample – this statistic is called the sample mean. Each sample has its own average value, and the distribution of these averages is called the "sampling distribution of the sample mean". This distribution is normal $N\left(\mu, \frac{\sigma^2}{n}\right)$ (n is the sample size) since the underlying population is normal,

although sampling distributions may also often be close to normal even when the population distribution is not (see central limit theorem).

Standard error:

Standard deviation of the sampling distribution of the statistic t is called standard error of t.

$$\text{i.e.} \quad \text{S.E (t)} = \sqrt{Var(t)}$$

Utility of standard error:

1. It is a useful instrument in the testing of hypothesis. If we are testing a hypothesis at 5% l.o.s and if the test statistic i.e. $|Z| = \left|\frac{t - E(t)}{S.E(t)}\right| > 1.96$ then the null hypothesis is rejected at 5% l.o.s otherwise it is accepted.
2. With the help of the S.E we can determine the limits with in which the parameter value expected to lie.
3. S.E provides an idea about the precision of the sample. If S.E increases the precision decreases and vice-versa. The reciprocal of the S.E i.e. $\frac{1}{S.E}$ is a measure of precision of a sample.
4. It is used to determine the size of the sample.

Standard error of sample mean:

Theorem: Show that the standard error of sample mean $(\bar{x})$ of a random sample of size n drawn at random from a population with mean μ and variance σ^2 is $\frac{\sigma}{\sqrt{n}}$ i.e. $S.E(\bar{x}) = \frac{\sigma}{\sqrt{n}}$

Proof: let $x_1, x_2, \dots\dots\dots, x_n$ be a random sample of size n drawn at random from a population with mean μ and varianceσ^2, Therefore we have

$E(x_i) = \mu \ \ and \quad V(x_i) = \sigma^2 \qquad \forall\, i = 1,2, \dots\dots. n \qquad (1)$

The sample mean is given b

$$(\bar{x}) = \frac{1}{n}(x_1 + x_2 + \cdots\dots\dots. + x_n)$$

$$\therefore V(\bar{x}) = V\left(\frac{1}{n}(x_1 + x_2 + \cdots\dots\dots. + x_n)\right)$$

$$= \frac{1}{n^2}\,(V(x_1) + v(x_2) + \cdots\dots\dots. + v(x_n))$$

$$= \frac{1}{n^2}\,(\sigma^2 + \sigma^2 + \cdots\dots\dots. + \sigma^2)$$

$$= \frac{1}{n^2} n\sigma^2$$

$V(\bar{x}) = \frac{\sigma^2}{n} \implies S.E(\bar{x}) = \frac{\sigma}{\sqrt{n}}$

Standard error of Population proportion:

Theorem: Show that the standard error of sample proportion (p) of an attribute in a random sample of size n drawn at random from a population with its population proportion P is $\sqrt{\frac{PQ}{n}}$ where $Q = 1 - P$.

Proof: let X be the number of the persons possessing the given attribute in a random sample of size n drawn at random from a population with its population proportion P, then the distribution of X will be Binomial i.e. $X \sim B(n, P)$. Let p be the proportion of persons possessing the given attribute in a random sample then

$p = \frac{X}{n}$

$E(p) = \frac{E(X)}{n} = \frac{nP}{n} = P \quad [\because E(X) = nP]$

Also $V(p) = V\left(\frac{X}{n}\right) = \frac{nPQ}{n^2} = \frac{PQ}{n} \implies S.E(p) = \sqrt{\frac{PQ}{n}}$ where $Q = 1 - P$.

Statistical Hypotheses And its Types

Hypothesis: A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true. Hypothesis testing refers to the formal procedures used by statisticians to accept or reject statistical hypotheses. The best way to determine whether a statistical hypothesis is true would be to examine the entire population. Since that is often impractical, researchers typically examine a random sample from the population. If sample data are not consistent with the statistical hypothesis, the hypothesis is rejected.

Example: A coin may be tossed 200 times and we may get heads 80 times and tails 120 times, we may now be interested in testing the hypothesis that the coin is unbiased. To take another example we may study the average weight of the 100 students of a particular college and may get the result as 110lb. We may now be interested in testing

the hypothesis that the sample has been drawn from a population with average weight 115lb.

Hypotheses are of two types

1. Null Hypothesis
2. Alternative hypothesis

Null hypothesis:

The hypothesis under verification is known as null hypothesis and is denoted by H_0 and is always set up for possible rejection under the assumption that it is true.

For example, if we want to find out whether extra coaching has benefited the students or not, we shall set up a null hypothesis that "extra coaching has not benefited the students". Similarly, if we want to find out whether a particular drug is effective in curing malaria we will take the null hypothesis that "the drug is not effective in curing malaria".

Alternative hypothesis:

The rival hypothesis or hypothesis which is likely to be accepted in the event of rejection of the null hypothesis H_0 is called alternative hypothesis and is denoted by H_1 or H_a.

For example, if a psychologist who wishes to test whether or not a certain class of people has a mean I.Q. 100, then the following null and alternative hypothesis can be established.

The null hypothesis would be

$$H_0 : \mu = 100$$

Then the alternative hypothesis could be any one of the statements.

$$H_1 : \mu \neq 100$$
$$(or)\, H_1 : \mu > 100$$
$$(or)\, H_1 : \mu < 100$$

Errors in testing of hypothesis:

After applying a test, a decision is taken about the acceptance or rejection of null hypothesis against an alternative hypothesis. The decisions may be four types.

1) The hypothesis is true but our test rejects it.(type-I error)
2) The hypothesis is false but our test accepts it. .(type-II error)
3) The hypothesis is true and our test accepts it.(correct)
4) The hypothesis is false and our test rejects it.(correct)

The first two decisions are called errors in testing of hypothesis.

i.e.1) Type-I error

2) Type-II error

1) **Type-I error:** The type-I error is said to be committed if the null hypothesis (H_0) is true but our test rejects it.

2) **Type-II error:** The type-II error is said to be committed if the null hypothesis (H_0) is false but our test accepts it.

Level of significance:

The maximum probability of committing type-I error is called level of significance and is denoted by α.

α = P (Committing Type-I error)

= P (H_0 is rejected when it is true)

This can be measured in terms of percentage i.e. 5%, 1%, 10% etc…….

Power of the test:

The probability of rejecting a false hypothesis is called power of the test and is denoted by $1-\beta$.

Power of the test =P (H_0 is rejected when it is false)

= 1- P (H_0 is accepted when it is false)

= 1- P (Committing Type-II error)

= 1- β

- A test for which both α and β are small and kept at minimum level is considered desirable.
- The only way to reduce both α and β simultaneously is by increasing sample size.
- The type-II error is more dangerous than type-I error.

Critical region:

A statistic is used to test the hypothesis H_0. The test statistic follows a known distribution. In a test, the area under the probability density curve is divided into two regions i.e. the region of acceptance and the region of rejection. The region of rejection is the region in which H_0 is rejected. It indicates that if the value of test statistic lies in this region, H_0 will be rejected. This region is called critical region. The area of the critical region is equal to the level of significance α. The critical region is always on the tail of the distribution curve. It may be on both sides or on one side depending upon the alternative hypothesis.

One tailed and two tailed tests:

A test with the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta \neq \theta_0$, it is called a two tailed test. In this case the critical region is located on both the tails of the distribution.

A test with the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta > \theta_0$ (right tailed alternative) or $H_1 : \theta < \theta_0$ (left tailed alternative) is called one tailed test. In this case the critical region is located on one tail of the distribution.

$H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ ------- right tailed test

$H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$ ------- left tailed test

Test statistic:

The test statistic is defined as the difference between the sample statistic value and the hypothetical value, divided by the standard error of the statistic.

i.e. test statistic $Z = \dfrac{t - E(t)}{S.E(t)}$

Procedure for testing of hypothesis:

1. Set up a null hypothesis i.e. $H_0 : \theta = \theta_0$.
2. Set up a alternative hypothesis i.e. $H_1 : \theta \neq \theta_0$ or $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$
3. Choose the level of significance i.e. α.
4. Select appropriate test statistic Z.
5. Select a random sample and compute the test statistic.
6. Calculate the tabulated value of Z at α% l.o.s i.e. Z_α.
7. Compare the test statistic value with the tabulated value at α% l.o.s. and make a decision whether to accept or to reject the null hypothesis.

Large sample tests:

The sample size which is greater than or equal to 30 is called as large sample and the test depending on large sample is called large sample test.

The assumption made while dealing with the problems relating to large samples are

Assumption-1: The random sampling distribution of the statistic is approximately normal.

Assumption-2: Values given by the sample are sufficiently closed to the population value and can be used on its place for calculating the standard error of the statistic.

Large sample test for single mean (or) test for significance of single mean:

For this test

The null hypothesis is $H_0 : \mu = \mu_0$

against the two sided alternative $H_1 : \mu \neq \mu_0$

where μ is population mean

μ_0 is the value of μ

Let $x_1, x_2, x_3, ..., x_n$ be a random sample from a normal population with mean μ and variance σ^2

i.e. if $X \sim N(\mu, \sigma^2)$ then $\bar{x} \sim N\left(\mu, \sigma^2/n\right)$, Where $\bar{x}$ be the sample mean

Now the test statistic $Z = \dfrac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\bar{x} - E(\bar{x})}{S.E(\bar{x})} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Note: if the population standard deviation is unknown then we can use its estimate s, which will be calculated from the sample. $s = \sqrt{\dfrac{1}{n-1}\sum (x - \bar{x})^2}$.

Large sample test for difference between two means:

If two random samples of size n_1 and n_2 are drawn from two normal populations with means μ_1 and μ_2, variances ${\sigma_1}^2$ and ${\sigma_2}^2$ respectively

Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means for the first and second populations respectively

Then $\bar{x}_1 \sim N\left(\mu_1, {\sigma_1}^2 / n_1\right)$ and $\bar{x}_2 \sim N\left(\mu_2, {\sigma_2}^2 / n_2\right)$

Therefore $\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}\right)$

For this test

The null hypothesis is $H_0 : \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$

against the two sided alternative $H_1 : \mu_1 \neq \mu_2$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E(\bar{x}_1 - \bar{x}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S.E(\bar{x}_1 - \bar{x}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}} \sim N(0,1)$$ [since $\mu_1 - \mu_2$ =0 from H_0]

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Note: If σ_1^2 and σ_2^2 are unknown then we can consider S_1^2 and S_2^2 as the estimate value of σ_1^2 and σ_2^2 respectively..

Large sample test for single standard deviation (or) test for significance of standard deviation:

Let $x_1, x_2, x_3, ..., x_n$ be a random sample of size n drawn from a normal population with mean μ and variance σ^2,

for large sample, sample standard deviation s follows a normal distribution with mean σ and variance $\sigma^2/2n$ i.e. $s \sim N\left(\sigma, \sigma^2/2n\right)$

For this test

The null hypothesis is $H_0 : \sigma = \sigma_0$

against the two sided alternative $H_1 : \sigma \neq \sigma_0$

Now the test statistic $Z = \dfrac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{s - E(s)}{S.E(s)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{s - \sigma}{\sigma/\sqrt{2n}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Large sample test for difference between two standard deviations:

If two random samples of size n_1 and n_2 are drawn from two normal populations with means μ_1 and μ_2, variances ${\sigma_1}^2$ and ${\sigma_2}^2$ respectively

Let s_1 and s_2 be the sample standard deviations for the first and second populations respectively

Then $s_1 \sim N\left(\sigma_1, {\sigma_1}^2 \big/ 2n_1\right)$ and $\bar{x}_2 \sim N\left(\sigma_2, {\sigma_2}^2 \big/ 2n_2\right)$

Therefore $s_1 - s_2 \sim N\left(\sigma_1 - \sigma_2, \frac{{\sigma_1}^2}{2n_1} + \frac{{\sigma_2}^2}{2n_2}\right)$

For this test

The null hypothesis is $H_0 : \sigma_1 = \sigma_2 \Rightarrow \sigma_1 - \sigma_2 = 0$

against the two sided alternative $H_1 : \sigma_1 \neq \sigma_2$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{(s_1 - s_2) - E(s_1 - s_2)}{S.E(s_1 - s_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(s_1 - s_2) - (\sigma_1 - \sigma_2)}{S.E(s_1 - s_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(s_1 - s_2)}{\sqrt{\frac{{\sigma_1}^2}{2n_1} + \frac{{\sigma_2}^2}{2n_2}}} \sim N(0,1)$$ [since $\sigma_1 - \sigma_2$ =0 from H_0]

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Large sample test for single proportion (or) test for significance of proportion:

Let x is number of success in n independent trails with constant probability p, then x follows a binomial distribution with mean np and variance npq.

In a sample of size n let x be the number of persons processing a given attribute then the sample proportion is given by $\hat{p} = \frac{x}{n}$

Then $E(\hat{p}) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x) = \frac{1}{n}np = p$

And $V(\hat{p}) = V\left(\frac{x}{n}\right) = \frac{1}{n^2}V(x) = \frac{1}{n^2}npq = \frac{pq}{n}$

$S.E(\hat{p}) = \sqrt{\frac{pq}{n}}$

For this test

The null hypothesis is $H_0 : p = p_0$

against the two sided alternative $H_1 : p \neq p_0$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\hat{p} - E(\hat{p})}{S.E(\hat{p})} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Large sample test for single proportion (or) test for significance of proportion:

let x_1 and x_2 be the number of persons processing a given attribute in a random sample of size n_1 and n_2 then the sample proportions are given by $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$

Then $E(\hat{p}_1) = p_1$ and $E(\hat{p}_2) = p_2 \Rightarrow E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$

And $V(\hat{p}_1) = \frac{p_1 q_1}{n_1}$ and $V(\hat{p}_2) = \frac{p_2 q_2}{n_2} \Rightarrow V(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$

$S.E(\hat{p}_1) = \sqrt{\frac{p_1 q_1}{n_1}}$ and $S.E(\hat{p}_2) = \sqrt{\frac{p_2 q_2}{n_2}} \Rightarrow S.E(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$

For this test

The null hypothesis is $H_0 : p_1 = p_2$

against the two sided alternative $H_1 : p_1 \neq p_2$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\hat{p}_1 - \hat{p}_2 - E(\hat{p}_1 - \hat{p}_2)}{S.E(\hat{p}_1 - \hat{p}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{S.E(\hat{p}_1 - \hat{p}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} \sim N(0,1) \quad \text{Since } p_1 = p_2 \text{ from } H_0$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

When p is not known p can be calculated by $p = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$ and $q = 1 - p$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0