One of the principal objective of statistical analysis is to draw inference about population on the basis of data collected by sampling from population. In other words, it is required to draw inference or (to generalise) about population from the sample. The inference to be drawn relates to some parameters of population such as mean, SD, proportion of an attribute in the pop^n.

In this chapter we will develop techniques that enable us to generalise the results of sample to population & check whether these generalisations are valid or not. & the branch of statistics which helps us in arriving the criteria for such decisions is known as "Testing of hypothesis"

* Population & Random Sample:

Population is the aggregate or totality of statistical data forming a subject of investigations i.e it is a collection of objects, observations or measurements either actual or conceptual.

eg: the population of the books in the national library, the population of heights of Indians. etc

A sample is a portion of the population. The process of selection of a sample is called sampling

A Random sample is one in which each member of population has an equal chance of being included in it. There are ${}^{N}C_{n}$ different samples of size n that can be picked up from a population of size N.

SRSWR: In this method a random sample is taken from population with replacement ie after selecting a article from the population after recording its properties, it is replaced back in the population before selecting the next article. so in every draw the probability of selection remains same & it is $\frac{1}{N}$. if N is the population size

SRSWOR:
In this method, a random sample is taken without replacement ie after selecting an article from the population it is kept aside. & next article is selected. so if the population size is N then in the 1st draw chance of selection of any article is $\frac{1}{N}$, in 2nd draw that chance is $\frac{1}{N-1}$. in th 3rd draw the chance is $\frac{1}{N-2}$ & so on

* Parameter and statistic:

The statistical constants of the population such as mean, S.D are called parameters. Parameters are denoted by Greek letters.

The mean ($\bar{x}$), S.D (s) of sample are known as statistic. statistic ~~are~~ is denoted by Roman letter.

Symbols for population & samples

	population	sample
Characteristic	parameter	statistic
Symbols	pop^n size N	sample size n
	pop^n mean μ	sample mean $\bar{x}$
	pop^n S.D σ	sample S.D - S
	pop^n proportion P	sample proportion $\hat{p}$

* Aims of sample:

The pop^n parameters are not known generally. Then sample characteristics are utilised to approximately determine or estimate of the pop^n. Thus statistic is an estimate of parameter. The estimate of mean & S.D of the pop^n is a primary purpose of all scientific experimentation.

~~The~~ logic

$\boxed{\text{Pop}^n\ f(x)}$:

pop^n f(x) is a pop^n whose prob. distribution is f(x). eg:- If f(x) is binomial, poisson or normal then corresponding pop^n is known as b.p, p.p or normal pop^n

Note:

The units selected in two or more samples drawn from a popⁿ are not the same, hence the value of a statistic varies from sample to sample, but the parameter remains the same. In practise, parameter values are not known, and their estimates are obtained by taking sample values. Thus a statistic can be regarded as estimate of a popⁿ parameter.

* Sampling Distribution of a statistic:

Let us consider a popⁿ of size N. If we draw a sample of size n from this population, then the total no. of possible samples is $^{N}C_{n} = \frac{N!}{n!(N-n)!} = k$ (say).

For each of these k samples we compute a statistic (t) (ie sample mean, sample S.D) Then value of statistic t may vary from sample to sample.

Let $t_1, t_2 \ldots t_k$ be the values of t for k samples. Each of these value occurs with a definite probability ie we can treat t as r.v assuming values $t_1, t_2 \ldots t_k$ with some prob. law. This probability distribution of t is known as the sampling distribution of t

* standard error (S.E) of a statistic:

The S.E of a statistic t is the standard deviation of the sample distribution of the statistic ie mean of sampling distribution of t is.

* Sampling distribution of means from infinite Population:

Let the population be infinitely large & having a population mean μ & a population variance σ^2. If x is r v denoting the measurement of characteristic then

Expected value of $x = E(x) = \mu$

variance of $x = var(x) = \sigma^2$

The sample mean $\bar{x}$ is the sum of n random variables viz $x_1, x_2 \dots x_n$ each being divided by n.

Here $x_1, x_2 \dots x_n$ independent r v from infinitely large population.

$\therefore E(x_1) = \mu$ & $var(x_1) = \sigma^2$

$E(x_2) = \mu$ & $var(x_2) = \sigma^2$ & so on.

Finally $E(\bar{x}) = E\left(\frac{x_1 + x_2 + \dots + x_n}{n}\right)$

$= \frac{1}{n}E(x_1) + \frac{1}{n}E(x_2) + \dots + \frac{1}{n}E(x_n)$

$= \frac{\mu}{n} + \frac{\mu}{n} + \dots + \frac{\mu}{n} = \mu$

$$Var(\bar{x}) = var\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right)$$

$$= \frac{1}{n^2} var(x_1) + \frac{1}{n^2} var(x_2) + \cdots + \frac{1}{n^2} var(x_n)$$

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

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

The expected value of the sample mean is same as the population mean. The variance of the sample mean is the variance of population divided by sample size.

The average value of the sample tends to true population mean. If sample size (n) is increased then variance of $\bar{x}$ is, $\frac{\sigma^2}{n}$ get reduced. by taking large value of n, the variance $\left(\frac{\sigma^2}{n}\right)$ of $\bar{x}$ can be made as small as desired.

The S.D $\frac{\sigma}{\sqrt{n}}$ of $\bar{x}$ is also called standard error of $\bar{x}$ It is denoted $\sigma_{\bar{x}}$.

* Sampling from Normal pop^n:-

If $x \sim N(\mu, \sigma^2)$ then it follow that

$\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$ ie if $x_1, x_2 .. x_n$ is a random sample from $N(\mu, \sigma^2)$ pop^n then $dist^r$ of $\bar{x}$ is normal with mean μ & variance σ^2/n

Examples:

(i) The diameter of a component produced on semi automatic machine is known to be distributed normally with a mean of 10 mm & S.D of 0.1 mm. If we pick up a random sample of size 5, what is the prob that the same mean will bet^n 9.95 & 10.05 mm

→ Let x be r.v representing diameter of one component picked up random.

Here

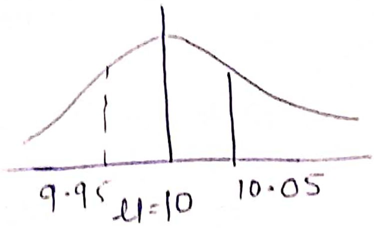
$X \sim N(10, 0.01)$

$\therefore \bar{X} \sim N\left(10, \frac{0.01}{5}\right) \qquad \left[\because \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)\right]$

We have

$P[9.95 \le \bar{x} \le 10.05] = 2\,P[10 \le \bar{x} \le 10.05]$

($\because$ curve is symmetric)



$$= 2\,P\left[\frac{10-\mu}{\sigma/\sqrt{n}} \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le \frac{10.05-\mu}{\sigma/\sqrt{n}}\right]$$

$$= 2\,P\left[0 \le z \le 1.12\right]$$

$$= 2 \times 0.3686 = 0.7372$$

HW

(2) A sample of size 25 is picked up at random from a population which is normally distributed with a mean of 100 & variance of 36. Calculate

a) $P[\bar{x} \le 99]$ (ii) $P[98 \le \bar{x} \le 100]$

→ (a) 0.2023 (ii) 0.4522.

(Hint: $n = 25,\ \mu = 100,\ \sigma^2 = 36 \Rightarrow \underline{\sigma = 6}$

i) $P[\bar{x} \le 99] = P\left[\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le \frac{99-\mu}{\sigma/\sqrt{n}}\right] \rightarrow$

ii) $P[98 \le \bar{x} \le 100] = P\left[\frac{98-\mu}{\sigma/\sqrt{n}} \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le \frac{100-\mu}{\sigma/\sqrt{n}}\right]$

chance or the fluctuations of sampling is called
tests of significance or test of hypothesis

* ~~Tests of significance OR~~ Testing a hypothesis:
on the basis of sample information we make
certain decisions about population. In ~~r~~taking
such decisions we make certain assumptions.
These assumptions are known as statistical
hypothesis. These hypothesis are tested.
Assuming hypothesis correct, we calculate probability
of getting observed sample. If the prob. is less than
certain assigned value, the hypothesis is to be
rejected.

* Null hypothesis & Alternative hypothesis :-
Null hypothesis is best for analysing the
problem. Null hypothesis is the hypothesis of
no difference. Thus we shall persume that there
is no significant difference betn observed value &
expected value. Then we shall test whether this
hypothesis is satisfied by data or not. If hypothesis
is not approved the difference is considered to be
significant. If hypothesis is approved then the
difference would be described as due to sampling

fluctuations. Null hypothesis is denoted by H_0.

Any hypothesis which is complementary to the null hypothesis is called alternative hypothesis & denoted by H_1.

eg: If we want to test the popn mean is some specified value say μ_0, then $H_0: \mu = \mu_0$ then the alternative hypothesis may be any one of the following

$H_1: \mu \neq \mu_0$

$H_1: \mu < \mu_0$, $H_1: \mu > \mu_0$.

* Errors (Type I error & Type II error)

In testing of hypothesis, we decide to accept or reject the hypothesis based on the results obtained by sample. so we may get one of the following 4 cases

Actually	Decision from sample: Reject H_0	Decision from sample: Accept H_0
H_0 is true	Wrong (Type I error)	Correct
H_0 is false	Correct	Wrong (Type II error)

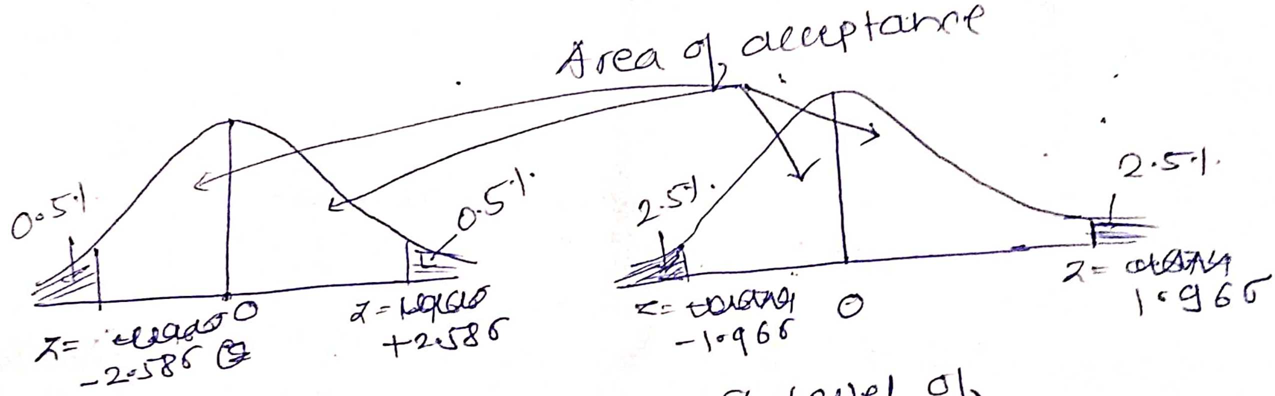
Thus in testing of hypothesis we are likely to commit two types of errors:

Type I error : rejecting H_0 when H_0 is true

Type II error : Accepting H_0 when it is false.

* Level of significance:-

There are two critical regions which cover the 5% & 1% areas of the normal curve. The shaded portion are the critical regions.



1% level of significance

5% level of significance

Thus the probability of the value of variate falling in the critical region is level of significance. If variate falls in the critical area, the hypothesis to be rejected. ie prob. of rejecting null hypothesis when it is true. eg:- LOS 5%. means 5% risk of concluding that diff. exist when there is no actual difference

* Test of significance:-

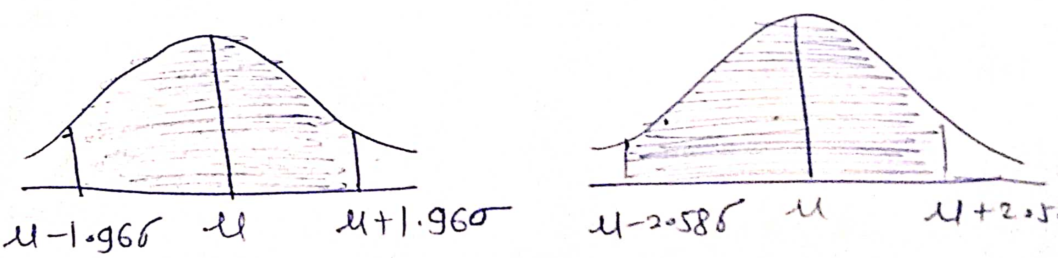
The tests which enables us to decide whether to accept or reject the null hypothesis is called test of significance.

If the difference betn sample values & the population values are so large (lies in the critical area) it is to be rejected.

* Confidence limits :-

$\mu - 1.96\sigma$, $\mu + 1.96\sigma$ are 95% confidence limits as the area betn $\mu - 1.96\sigma$ & $\mu + 1.96\sigma$ is 95%. If a sample statistics lies in the interval $\mu - 1.96\sigma$, $\mu + 1.96\sigma$ we call 95% confidence interval.

Simly $\mu - 2.58\sigma$, $\mu + 2.58\sigma$ is 99% confidence limits as the area betn $\mu - 2.58\sigma$, $\mu + 2.58\sigma$ is 99%. The numbers 1.96 & 2.58 are called confidence ~~limits~~ coeffn



Note: If n > 30, a sample is called large otherwise small. The sampling distribution of large samples is assumed to be normal.

* Test of significance of large sample..

Normal distribution is the limiting case of Binomial distribution when n is large enough. Suppose we wish to test the hypothesis that prob. of success in such trial is p. Assuming it to be true, the mean μ & S.D σ of sampling distribution of number of successes are np & $\sqrt{npq}$ resp.

For normal distribution only 5% of the members lie outside $\mu \pm 1.96\sigma$ while only 1% of the members lie outside $\mu \pm 2.58\sigma$

If x is the observed number of successes in sample & z is std normal variate then

$$\cancel{z = \frac{x - np}{\sigma}} \qquad z = \frac{x - \mu}{\sigma}$$

First we have to find value of z. Test of significance depends upon value of z.

(i) If $|z| < 1.96$, the difference betn the observed & expected number of successes is not significant at 5% level of significance.

(ii) If $|z| > 1.96$, the difference is significant at 5% level of significance

(iii) If $|z| > 2.58$, the difference is significant at 1% level of significance

(iv) If $|z| < 2.58$, the difference betn observed & expected no. of success is not significant at 1% LOS.

① A coin was tossed 400 times & the head turned up 216 times. Test the hypothesis that the coin is unbiased at 5% LOS

→ Suppose the coin is unbiased.

Then P(getting head in a toss) = $\frac{1}{2}$

Expected no. of success = $400 \times \frac{1}{2} = 200$

& observed value of success = 216

Thus

Excess of observed value over expected value

$= 216 - 200 = 16$

Also S.D of simple sampling = $\sqrt{npq}$

$= \sqrt{400 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 10$

Hence

$$z = \frac{x - np}{\sqrt{npq}} = \frac{16}{10} = 1.6$$

As $z < 1.96$, the hypothesis is accepted at 5% LOS

ie we conclude that the coin is unbiased at 5% LOS.

② A die was thrown 9000 times & a throw of 5 or 6 was obtained 3240 times. On the assumption of random throwing, do the data indicate an unbiased die?

Suppose die is unbiased.

Then $P(\text{throw of 5 or 6}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$

The expected no. of success $= \frac{1}{3} \times 9000 = 3000$

observed value of success $= 3240$

Thus excess of observed value over expected value $= 3240 - 3000$

$= 240$

& $SD = \sqrt{npq} = \sqrt{9000 \times \frac{1}{3} \times \frac{2}{3}} = 44.72$

Hence

$z = \cancel{240} \quad \frac{x - np}{\sqrt{npq}} = \frac{240}{44.72} = 5.4$ nearly

As $z > 2.58$, the hypothesis is to be rejected at 1% LOS & we conclude that the die is biased.

(3)

* sampling distribution of proportion :-

* Suppose a simple sample of n items is drawn from the population. It is same as a series of n independent trials with prob. p of success.

The probabilities of 0, 1, 2 ... n success are the terms in the binomial expansion of $(p+q)^n$

* The sampling of attributes may be regarded as the selection of sample from popⁿ whose members possess the attribute K or not K. The presence of K may called success & its absence as failure

Here mean = np , SD = $\sqrt{npq}$

Let us consider the proportion of successes

(a) mean proportion of success = $\frac{np}{n} = p$

(b) S.D (S.E) of proportion of success = $\sqrt{\frac{npq}{n^2}} = \sqrt{\frac{pq}{n}}$

(c) Precision of proportion of success = $\frac{1}{S.E} = \sqrt{\frac{n}{pq}}$

(1) A group of scientific mens reported 1705 sons & 1527 daughters. Do these figures conform to to the hypothesis that the sex ratio is ½

→ The total no. of observation = 1705 + 1527
= 3232

The no. of sons = 1705

∴ observed male ratio = $\frac{1705}{3232}$ = 0.5175

on the given hypothesis the male ratio = 0.5

Thus,
the difference betn the observed ratio & theoretical ratio = 0.5275 − 0.5
= 0.0275

The S.D of proportion = $\sqrt{\frac{pq}{n}} = \sqrt{\frac{\frac{1}{2}\cdot\frac{1}{2}}{3232}}$ = 0.0088

Here $z = \frac{0.0275}{0.0088} = 3.125$

As $z > 2.58$,
Hence it can be definitely said that the fig given donot confirm to given hypothesis.

* Estimation of the parameters of the population:

The mean, S.D etc of the popⁿ are known as parameters. They are denoted by μ & σ. Their estimates are based on the sample values. The mean & S.D of a sample are denoted by $\bar{x}$ & S resp. Thus statistic is an estimate of the parameter. There are two types of estimates.

1) Point estimation.

An estimate of a popⁿ parameter given by single no. is called point estimation of parameter,

eg $$(S.D)^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$

ii) Interval estimation:

An interval in which popⁿ parameter may be expected to lie with given degree of confidence. The intervals are

i) $\bar{x} \pm 1.96\,\sigma_s$, $\bar{x} \pm 2.58\,\sigma_s$ are 95% & 99% confidence of limits for μ

where $\bar{x}$ & σ_s are mean & S.D of sample