UNIT-4 Sampling Distribution

Introduction

In our day to day life it becomes necessary to draw some valid & reasonable conclusions concerning a large mass of individuals or things. It becomes practically impossible to examine every individual or the entire ~~a small part of this pop~~ group known as population.

Therefore we may prefer to examine a small part of this population known as a sample with the motive of drawing same conclusion about the entire population based on the information or result revealed by the sample. The entire process known as statistical inference aims in ascertaining maximum information about the population with min effort and time

Random sampling.

A large collection of individuals or attributes or numerical data can be understood as a population or universe.

A finite subset of the universe is called a Sample. The no. of individuals in a sample is called a sample size. If the sample size n is less than or equal to 30 the sample is said to be small, otherwise it is a large sample.

The process of selecting a sample from the population is called as Sampling

The selection of an individual or item from the population in such a way that each has the same chance of being selected is called as Random Sampling. Suppose we take a sample of size n from a finite population of size N, then we have NC_n possible samples. Random Sampling is a technique in which each of the NC_n samples has an equal chance of being selected.

Sampling where a member of the population may be selected more than once is called as sampling with replacement, on the other hand if a member cannot be chosen more than once is called as Sampling without replacement

Simple sampling is a special case of random sampling in which trials are independent & the probability of success p is a constant.

The word statistic is often used for the random variable or for its values.

Sampling Distribution

Suppose that we have different samples of size n drawn from a population. For each & every sample of size n we can compute some quantities like mean, standard deviation etc. obviously these will not be the same. Suppose we group these characteristics according to their frequencies, the frequency distributions so generated are called Sampling distributions.

These can be distinguished as sampling distribution of mean, standard deviation etc. The sampling distribution of large samples is assumed to be a normal distribution.

The standard deviation of a sampling distribution is also called the standard error (S.E). The reciprocal of the standard error is called precision.

Sampling distribution of Means.

Suppose we draw all possible samples of a certain size n from a population and find the mean $\bar{x}$ of each of these samples. The frequency distribution of these means is called the sampling distribution of means

It can be proved that for large samples (i.e. for values of $n \geq 30$) the sampling dist^n 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, the sampling distribution of means is a normal distribution even for $n < 30$.

Accordigly, the standard normal variate for the distribution of means is given by

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

The sampling distribution of the sample means for the two possible types of random sampling (with/ without replacement) associated with a finite population are

Statistic: Statistic is a real valued function of the random sample. So statistic is a function of one or more random var. not involving any unknown parameter. Thus statistic is a function of samples observations only & is itself a random var. Therefore a static must have a prob. distribution sample mean and sample variance are two important statistics which are measures of a random sample $X_1, X_2, X_3, \ldots, X_n$ of size n.

$$\text{sample mean} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$ (Measure of central tendency)

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X}^2)}{n-1} = \frac{n\sum X_i^2 - (\sum X_i)^2}{n(n-1)}$$

(measure of variability of data about mean)

Sample standard deviation is the +ve square root of the sample variance.

Degrees of Freedom (dof): of a statistic is a positive integer, denoted by ν, equals to $n-k$ where n is the no. of independent observations of the random sample & k is the no. of population parameters which are calculated using the sample data. Thus dof $\boxed{\nu = n-k}$ is the difference bet^n n(sample size) & k the no. of independent constraints imposed on the observations in the sample.

Central Limit theorem

Whenever n is large, the sampling distribution of $\bar{X}$ is appox normal with mean μ & variance $\frac{\sigma^2}{n}$ regardless of the form a population distribution with mean μ & finite variance σ^2, then the standardized sample mean $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is a random var whose distribution

---

Central Limit theorem

Whenever n is large, the sampling distribution of $\bar{X}$ is appox normal with mean μ & variance $\frac{\sigma^2}{n}$ regardless of the form of the population dist^n. This is established by central limit theorem.

Thm: If $\bar{X}$ is the mean of a sample of size n drawn from a population with mean μ and finite variance σ^2, then the standardized sample mean $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is a random var whose distribution fun^c approaches that of the standard normal distribution $N(Z:0,1)$ as $n \to \infty$

Testing of Hypothesis

On the basis of sample information, we make certain decisions about the population. In making such decisions we make certain assumptions. These assumptions are known as statistical hypothesis

Testing a hypothesis, is a process for deciding whether to accept or reject the hypothesis. The method consists in assuming the hypothesis as correct and then computing the probability is less than a certain preassigned value the hypothesis is rejected.

Null hypothesis

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

Any hypothesis which is complimentary to the null hypothesis is called Alternative hypothesis denoted by H_1.

Ex: 1. To test whether a process B is Better than a process A, we can formulate the hypothesis as there is no difference between the process A & B.

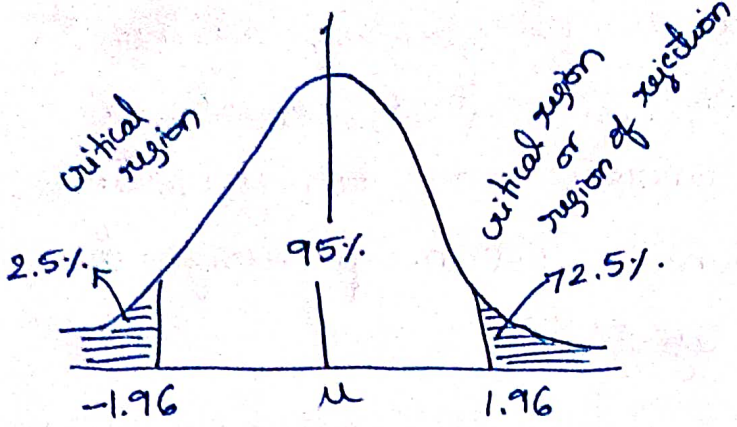
2. To test whether there is a relationship b/w two variates we can formulate the hypothesis as there is no relationship b/w them

Note:

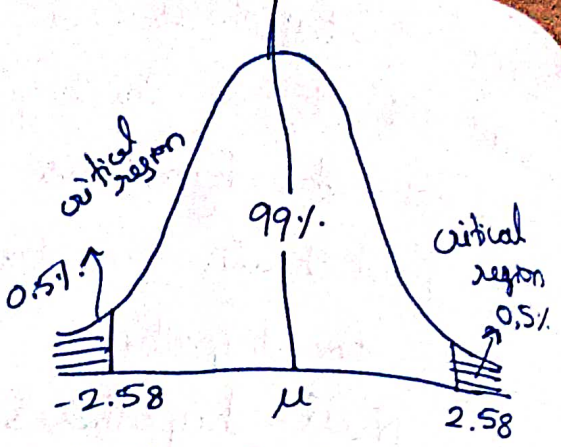
A null hypothesis is a statement about a population parameter (such as μ)

After 7 pages →

The constants 1.96, 2.58 etc in the confidence limits are called confidence coefficients denoted by Z_c. From confidence levels we can find confidence coefficients and viceversa.



5% level of significance



1% level of significance

As reflected in the fig, we can say with 95% confidence that if the hypothesis is true, the value of Z for an actual sample lies betⁿ −1.96 to 1.96. Since the area under the normal curve betⁿ these values is 0.95. However if the value of Z for random sample lies outside this range we can conclude that the prob of the happening of such an event is only 0.05 if the given hypothesis is true

The total shaded area 0.05 being the level of significance of the test, represents the prob of making type-I error. The set of values of Z outside the range −1.96, 1.96 constitutes the critical region or the region of rejecting the hypothesis whereas the values of Z within the same range constitutes the insignificant region or the region of acceptance of the hypothesis

One tailed and two tailed tests

Depending on the nature of the problem. we use a single tail test or double-tail test to estimate the significance of a result.

In double-tail test, the areas of both the tails of the curve respectively the sampling distribution are considered.

In single tail test, only the area on the right of an ordinate is taken into consideration

The foll table will be useful for working problems

Test	Critical values of z	
	5% level	1% level
One-tailed test	−1.645 (or) 1.645	−2.33 (or) 2.33
Two-tailed test	−1.96 and 1.96	−2.58 and 2.58

Significance level

The probability level. below which leads to the rejection of the hypothesis is known as the significance level. This probability is conventionally fixed at 0.05 or 0.01 i.e. 5% or 1%. These are called significance levels.

Test of significance and confidence intervals

The process which helps us to decide about the acceptance or rejection of the hypothesis is called the test of significance.

Let us suppose that we have a normal population with mean μ and S.D σ. If $\bar{x}$ is the sample mean of a random sample of size n, the standard normal variate z is defined by $z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$

From the normal distribution table we find that 95% of the area lies bet^n $z = -1.96$ & $z = 1.96$

In otherwords we can say with 95% confidence that z lies bet^n -1.96 & 1.96. Further 5% level of significance is denoted by $z_{0.05}$. Thus we have

$$-1.96 \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 1.96$$

i.e. $$-\frac{\sigma}{\sqrt{n}}(1.96) \leq \bar{x} - \mu \leq \frac{\sigma}{\sqrt{n}}(1.96)$$

$$\Rightarrow \mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}}(1.96) \quad \& \quad \bar{x} - \frac{\sigma}{\sqrt{n}}(1.96) \leq \mu$$

or $$\bar{x} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

Similary from the table, 99% of the area lies between -2.58 & $+2.58$. Thus

$$\bar{x} - 2.58\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + 2.58\left(\frac{\sigma}{\sqrt{n}}\right)$$

Test of significance for large samples

We know that the binomial dist^n tends to normal for large n, suppose we wish to test the hypothesis that the prob of success in such trial is P. Assuming it to be true, the mean μ and the standard deviation σ of the sampling dist^n of no. of successes are np and $\sqrt{npq}$ resp.

For a normal dist^n, 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 no. of successes in the sample & z is the standard normal variate then

$$z = \frac{x-\mu}{\sigma} = \frac{x-np}{\sqrt{npq}}$$

Thus we have the foll test of significance

i) If $|z| < 1.96$, difference bet^n the observed & expected no. of successes is not significant.

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

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

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.

Soln: H_0: Suppose the coin is unbiased

Then the prob of getting the head in a toss $= \frac{1}{2}$

$\therefore$ expected no. of successes $= \frac{1}{2} \times 400 = 200 = np$

& the observed no. of successes $= 216 = x$

Thus the excess of observed value over expected value

$= 216 - 200 = 16$

Also S.D of simple sampling $= \sqrt{npq} = \sqrt{400 \times \frac{1}{2} \times \frac{1}{2}} = 10$

Hence $Z = \frac{x - np}{\sqrt{npq}} = \frac{16}{10} = 1.6 < 1.96$

As, $Z < 1.96$, the hypothesis is accepted at 5% level of significance.

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?

Soln: H_0: Suppose the die is unbiased, Then the prob of throwing 5 or 6 with one die $= \frac{1}{3} = p$

The expected no. of successes $= np = \frac{1}{3} \times 9000$

$= 3000$

And the observed value of successes $= 3240$

Thus the excess of observed value over expected value is = 3240 − 3000 = 240

Also S.D of simple sampling = $\sqrt{npq} = \sqrt{9000 \times \frac{1}{3} \times \frac{2}{3}}$

$= 44.72$

Hence $Z = \frac{x - np}{\sqrt{npq}} = \frac{240}{44.72} = \underline{\underline{5.4}}$

As $Z > 2.58$, the hypothesis has to be rejected at 1% level of significance and we conclude that the die is biased.

3) A die is tossed 960 times and it falls with 5 upwards 184 times. Is the die biased?

Let Ho : The die is biased

The probability of falling 5 upwards with one die = $1/6$

The expected number of successes = $\frac{1}{6} \times 960 = 160 = \mu$

Actual no. of success = $184 = x$

$Z = \frac{x - \mu}{\sqrt{npq}} = \frac{184 - 160}{\sqrt{960 \times \frac{1}{6} \times 5/6}} = \frac{24}{11.546} = 2.07 <$ 2.58

As $Z < 2.58$ the hypothesis has to be accepted at 1% level of significance and we conclude that die is biased.

4) 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?

Let H_0 : observed value deviates from the expected value

probability of getting 2, 3, 4 in one die $= \frac{3}{6} = \frac{1}{2} = p$

Here $n = 12 \times 3086 = 37032$

Expected no. of success $= \frac{1}{2} \times 37032 = 18516 = np$

$$Z = \frac{x - \mu}{\sigma} \quad (\mu \to np,\ \sigma \to \sqrt{npq}) = \frac{19412 - 18516}{96.22} = 9.31 > 2.58$$

$\therefore$ Reject the hypothesis

5) Balls are drawn from a bag contains equal no. of black and white balls. Each ball bearing are replaced before drawing another. In 2250 drawing 1018 black and 1232 white balls have been drawn. Do you suspect some bias on the part of the drawer?

H_0 : Some bias on the part of the drawer

$n = 2250$ Black $= 1018$ white $= 1232$

probability of selecting black ball $= \frac{1}{2}$

Mean no. of success $= np = \frac{2250}{2} = 1125$

Variance $= \sqrt{npq} = 23.72$

Actual no. of black balls = 1018

$$\therefore Z = \frac{x-\mu}{\sigma} = \frac{1018-1125}{23.72} = -4.511$$

$|Z| = 4.511 > 2.58$ Reject the hypothesis

6) In a group of 80 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.

Test of significance for large samples

i) Test of significance of proportions

Let us consider the proportion P of successes

Mean proportion of successes = $\mu_p = P$

S.D or S.E proportion of successes = $\sigma_P = \frac{\sqrt{npq}}{n}$

$= \sqrt{\frac{pq}{n}}$

Let x be the observed no. of successes in a sample size of n $\mu = np$ be the expected no. of successes.

Then $Z = \frac{x - \mu}{\sigma} = \frac{x - np}{\sqrt{npq}} = \frac{P - p}{\sqrt{\frac{pq}{n}}}$

If $|Z| > 2.58$ we conclude that the difference is highly significant and reject the hypothesis.

Since p is the prob of succes & $\sqrt{\frac{pq}{n}}$ is the SE proportion of successes, $P \pm 2.58\sqrt{\frac{pq}{n}}$ are the ~~prop~~ probable limits.

1) A sample of 1000 days is taken from meteorological records of a 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?

Soln: p = proportion of foggy days in a sample of 1000 days is $= \frac{120}{1000}$

$p = 0.12$ and $q = 1 - p = 1 - 0.12 = 0.88$

∴ Probable limits of foggy days $= p \pm 2.58\sqrt{\frac{pq}{n}}$

$= 0.12 \pm 2.58\sqrt{\frac{0.12 \times 0.88}{1000}}$

$= 0.12 \pm 2.58\sqrt{0.0001066}$

$= 0.12 \pm 2.58 \times 0.01027$

$= 0.12 \pm 0.02651$

$= 0.14651$ and 0.09349

$= 14.65\%$ & 9.31%

Thus the percentage of foggy days lies b/w 9.34 and 14.65

2) 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 as well as the standard error of the estimate. Find deduce that the percentage of bad apples in the consignment almost certainly lies betn 8.5 and 17.5

Soln: p - proportion of bad apples in a sample of 500

$p = \frac{65}{500} = 0.13$, $q = 1 - p = 0.87$

S.E proportion of bad apples $= \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.13 \times 0.87}{500}}$

$= 0.015$

$\therefore$ Probable limits $= P \pm 2.58\sqrt{\frac{pq}{n}} = 0.13 \pm 2.58\sqrt{\frac{0.13 \times 0.87}{500}}$

$= 0.13 \pm 0.0387$

$= 0.0913$ and 0.1687

$= 9.13\%$ and 16.87%.

$\therefore$ % lies bet^n 8.5 and 17.5

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 ₹250 or less. It is desired to estimate how many out of 18,000 families have a monthly income of ₹250 or less. Within what limits would you place your estimate?

Sol^n: P = Proportion of no. of families in a sample of 600 families.

Here $p = \frac{206}{840} = \frac{103}{420} = 0.245$ and $q = \frac{317}{420} = 0.755$

$\therefore$ Standard error of the population of families having a monthly income of ₹250 or less

$= \sqrt{\left(\frac{pq}{n}\right)} = \sqrt{\frac{103}{420} \times \frac{317}{420} \times \frac{1}{840}} = 0.015$

$= 1.5\%$

Probabilities limits of families having monthly income of Rs. 250 or less are.

$= P \pm 2.58\sqrt{\frac{pq}{n}} = P \pm 2.58 \times 0.015$

$= 0.245 \pm 0.0387$

$= 0.2063$ and 0.2837

$= 20.63\%$ and 28.37%

∴ The probable limits in respect of 18,000 families is $0.2063 \times 18{,}000$ and 0.2837×18000

i.e. 3713.4 and 5106.6

Thus 3713 to 5107 families likely to have monthly income of Rs. 250 or less

4) 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 proportions.

Sol^n: H_0 : Sexes should be in equal proportion

H_1 : Sexes should not be in equal proportion

level of significance : 0.01, 0.05

$Z = \frac{x-\mu}{\sigma} = \frac{x-np}{\sqrt{npq}} = \frac{P-p}{\sqrt{\frac{pq}{n}}}$ $\quad \mu = p$ $\quad T = \sqrt{\frac{pq}{n}}$

For $n = 50$, $P = \frac{27}{50} = 0.54$, $p = \frac{1}{2} = 0.5$

$\therefore Z = \frac{0.54 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{50}}} = \frac{0.04}{\sqrt{0.05}} = 0.565 < 1.96$

We accept the null hypothesis series should be in equal proportion.

5) A machine produces 16 ~~important~~ imperfect articles in a sample of 500. After machine is overhauled, it produces 3 imperfect articles in a batch of 100. Has the machine been improved?

Sol^n: Ho : Machine is not improved

$n = 500$ $\quad P = X = \frac{16}{500} = 0.032$, $\quad p = \mu = \frac{3}{100} = 0.03$

$$Z = \frac{X - \mu}{\sigma} = \frac{0.032 - 0.03}{0.0076} = 0.2632$$

$$\sigma = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.03 \times 0.97}{500}} = 0.0076$$

As $Z < 1.96$ we accept the hypothesis
i.e. Machine is not improved.

16) 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?

Ho : Tea and coffee are equally popular in the state
H1 : $P \neq 0.5$ (two tailed)

$n = 500$, p = sample proportion of tea drinker

$= \frac{280}{500} = 0.56$

P_1 = population proportion of tea drinker = $\frac{1}{2} = 0.5 = p$

$q = 0.5$

$$Z = \frac{P - p}{\sqrt{\frac{pq}{n}}}$$

$$Z = \frac{0.56 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{500}}} = \underline{2.68}$$

$Z > 1.96$

We reject H_0 : ie. Tea & coffee are not equally popular in the state.

12) 400 children are chosen in an industrial town and 150 are found to be under weight. Assuming the conditions of simple sampling, estimate the percentage of children who are under weight in the industrial town and assign limits which with in which the percentage probably lies?

$n = 400$

p = proportion of children who are under weight = $\frac{150}{400}$

$= \underline{0.375}$

$p = 0.375$

$q = 1 - 0.375 = 0.625$

SE of proportion = $\sqrt{\frac{pq}{n}} = \sqrt{\frac{0.375 \times 0.625}{400}} = \underline{0.024}$

∴ The probable limits are

$$p \pm 3(SE)$$

$$0.375 \pm 3(0.024)$$

$= (0.303, 0.447) = 30\%$ and 44.7%

The probable limits of children : underweight in the population of 400 is

$0.303 \times 400 = 121 \approx 121$

$0.443 \times 400 = 178.8 \approx 179$

121 to 179 children are underweight

Errors

In sampling theory, valid reference about the population parameter ~~is~~ done on the basis of results of sample

We decide to accept or to reject the population after examining a sample from it. As such, we are liable to commit the following two types of errors

Type I Error : If a hypothesis is rejected while it should have been accepted

Type II Error : If a hypothesis is accepted while it should have been rejected.

	Accepting the hypothesis	Rejecting the hypothesis
Hypothesis true	Correct decision	Wrong decision (Type I error)
Hypothesis false	Wrong decision (Type II error)	Correct decision

The only way to reduce both types of errors is to increase the sample size. It is further important to note that acceptance as non acceptance of a hypothesis is purely based on the information revealed by the sample may not always be true in respect of the population

A region which amounts to the rejection of null hypothesis is called critical region or region of rejection

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 bags had the same defect. Is the difference betn the proportions significant?

Soln: We have $n_1 = 900$ $n_2 = 1600$

H_0: The difference betn the proportion is not significant

$$p_1 = \frac{20}{100} = \frac{1}{5} = 0.2 \quad , \quad p_2 = \frac{18.5}{100} = 0.185$$

$$p_1 - p_2 = 0.015$$

$$\therefore p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{180 + 296}{900 + 1600} = 0.19$$

$$q = 1 - 0.19 = 0.81$$

Thus $e^2 = pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = 0.19 \times 0.81 \left(\frac{1}{900} + \frac{1}{1600}\right) = 0.0017$

$$e = 0.004$$

Also $p_1 - p_2 = \frac{1.5}{100} = 0.015$

$$\therefore Z = \frac{p_1 - p_2}{e} = \frac{0.015}{0.014} = 0.37$$

As $Z < 1.96$, we accept the null hypothesis at 5% level of significance i.e. the difference betn the proportions is not significant.

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

Soln: Here $P_1 = 0.03$ $P_2 = 0.25$ so that $P_1 - P_2 = 0.05$

$$e^2 = \frac{P_1 q_1}{n_1} + \frac{P_2 q_2}{n_2} = \frac{0.3 \times 0.7}{1200} + \frac{0.25 \times 0.75}{900}$$

H_0: Sample proportions are equal $P_1 = P_2$

$$e = 0.0195$$

$$\therefore Z = \frac{P_1 - P_2}{e} = \frac{0.05}{0.0195} = 2.5 > 1.96.$$

$\therefore$ We reject the null hypothesis at 5% level of significance

Hence it is ~~usually~~ unlikely that the real difference will be hidden.

3) A Machine produces 16 imperfect articles in a sample of 500. After machine is overhauled, it produces 3 imperfect articles in a batch of 100. Has the machine been improved?

Soln: H_0: proportions of defectiveness before and after overhaleing are equal $P_1 = P_2$

H_1: proportion of defectiveness has decreased after overhauling $P_1 > P_2$

$$n_1 = 500 \quad , P_1 = \frac{16}{500} = 0.032$$

$$n_2 = 100 \quad , P_2 = \frac{3}{100} = 0.03$$

$$P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2} = \frac{500(0.032) + 100(0.03)}{500 + 100} = 0.0316$$

$$Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.032 - 0.03}{\sqrt{(0.0316)(0.9684)\left(\frac{1}{500} + \frac{1}{100}\right)}}$$

$$= 0.105 < 1.645$$

(one tail test)

From one tail test, we accept the null hypothesis i.e. Machine is not improved.

4) One type of aircraft is found to develop engine trouble in 5 flights and of a total of 100 and another type in 7 flights out of a total of 200 flights. Is there a significant difference in the two types of aircraft so far as engine defects are concerned?

Soln: $n_1 = 100$, $n_2 = 200$

Let p_1 and p_2 be the proportion of defects in the two types of aircrafts $p_1 = \frac{5}{100} = 0.05$

$$p_2 = \frac{7}{200} = 0.035$$

H_0: There is no significant difference bet the two types of aircrafts. $p_1 = p_2$

H_1: $p_1 \neq p_2$

$$p = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} = \frac{5+7}{300} = 0.04$$

$$q = 1 - p = 1 - 0.04$$

$$q = 0.96$$

$e^2 = pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = (0.04)(0.96)(0.01 + 0.05)$

$= 0.0009570$

$e = 0.02387$

$Z = \frac{p_1 - p_2}{e} = 0.628 < 1.96, 2.58$ (two tailed test)

∴ Null hypothesis is accepted at 5% and 1%.

ie. Difference is not significant

5) In a sample of 600 men from a certain city, 450 are smokers. In a sample of 900 men from another city, 450 are smokers. Do the data indicate that cities are significantly different with respect to the habit of smoking among men?

Sol^n: $n_1 = 600$, $n_2 = 900$, $p_1 = \frac{450}{600} = 0.75$

$p_2 = \frac{450}{900} = 0.50$

H_0: There is no significant difference with respect to the habit of smoking in cities $p_1 = p_2$

$H_1 = p_1 \neq p_2$

$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{450 + 450}{1500} = 0.6$, $q = 1 - p = 0.4$

$e^2 = pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = 0.6 \times 0.4\left(\frac{1}{600} + \frac{1}{900}\right)$

$= 0.0006675$

$e = 0.02584$

$Z = \frac{p_1 - p_2}{e} = \frac{0.25}{0.02584} = 9.67 > 1.96$ and 2.58

We reject the null hypothesis at 1% and 5% level or accept H_1

i.e. Difference is significant

Confidence limits for unknown Mean

Let the population from which a random sample of size n is drawn have mean μ and S.D σ

If μ is not known, there will be a range of values of μ for which observed mean $\bar{x}$ of the sample is not significant at any assigned level of probability. The relative deviation of $\bar{x}$ from μ is $\dfrac{(\bar{x}-\mu)}{\sigma/\sqrt{n}}$

If $\bar{x}$ is not significant at 5% level of probability. then $\left|\dfrac{(\bar{x}-\mu)\sqrt{n}}{\sigma}\right| < 1.96$ i.e.

$$\bar{x} - \frac{1.96\sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{1.96\sigma}{\sqrt{n}}$$

Thus 95% confidence or fiducial limits for the mean of the population corresponding to given sample are $\bar{x} \pm \dfrac{1.96\sigma}{\sqrt{n}}$

Similarly, Thus 99% confidence limits for μ are $\bar{x} \pm \dfrac{2.58\sigma}{\sqrt{n}}$

1) A sample of 900 numbers is found to have a mean of 3.4cm. can it be ~~random~~ reasonably regarded as a truly random sample from a large population with mean 3.25cm and S.D 1.61cm.

Soln:- H_0 : The sample is regarded as a true random sample for the population

Here $\bar{x} = 3.4\text{cm}$, $n = 900$, $\mu = 3.25$ and $\sigma = 1.61\text{cm}$

$$\therefore Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{3.4 - 3.25}{1.61/30} = 2.8 > 1.96 \text{ \& } 2.58$$

As $Z > 1.96$, we reject the hypothesis at 5% & 1% level i.e. 2.58 the deviation of the sample mean from the mean of the population is significant at 1% & 5% level of significance. Hence it cannot be regarded as a random sample.

2) An unbiased coin is thrown n times. It is desired that the relative frequency of the apperance of heads should be bet^n 0.49 and 0.51. Find the smallest value of n that will ensure this result will 90% confidence.

Soln: SE of the proportion of heads $= \sqrt{\frac{PQ}{n}} =$

$$= \sqrt{\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{n}} = \frac{1}{2\sqrt{n}}$$

90% of confidence = 45% or 0.45 of the total area under the normal curve on each side of the mean.

∴ the corresponding value of $Z = 1.645$ from the tables. Thus $p \mp 1.645\sigma = 0.49$ or 0.51

i.e. $0.5 - 1.645 \cdot \frac{1}{2\sqrt{n}} = 0.49$

$$0.5 + 1.645 \frac{1}{2\sqrt{n}} = 0.51$$

Hence $\frac{1.645}{2\sqrt{n}} = 0.01$ or $\sqrt{n} = \frac{329}{4}$

$$\underline{n = 6765}$$

3) The mean of a certain normal population is equal to the standard error of the distribution of means of samples of size 100 drawn from the population. Find the prob. that the mean of a sample of size 25 from the population will be -ve.

Sol^n: From the data we have $\mu = \sigma_{\bar{x}}$ for $n=100$

we have $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{100}} = \frac{\sigma}{10} = \mu$

Now, consider the dist^n of means of samples of size 25. Let $\bar{X}$ denote the mean of this dist^n. The corresponding Std normal variable is $Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma/\sqrt{25}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{25}}$

$$= \frac{\bar{X} - \sigma/10}{\frac{\sigma}{\sqrt{25}}} = \frac{5\bar{x}}{\sigma} - \frac{1}{2}$$

$$Z = \frac{5\bar{x}}{\sigma} - \frac{1}{2}$$

$\therefore \bar{x} = \frac{\sigma}{5}\left(z + \frac{1}{2}\right)$

$\bar{x} < 0 \Rightarrow \frac{\sigma}{5}\left(z + \frac{1}{2}\right) < 0 \Rightarrow z + \frac{1}{2} < 0$

$\therefore$ The required prob is, $P(\bar{x} < 0) = P\left(z < -\frac{1}{2}\right)$

$= P\left(z > \frac{1}{2}\right)$

$= P(z > 0) - P\left(0 < z < \frac{1}{2}\right)$

$= 0.3085$

0.5.

5) Find the probability that the mean of simple sample of 900 members will be negative.

Soln: $\mu = 0.1$, $\sigma = 2.1$, $n = 900$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - 0.1}{\left(\frac{2.1}{\sqrt{900}}\right)} = \frac{(\bar{x} - 0.1)30}{2.1}$$

$$= (\bar{x} - 0.1)14.2857$$

$$Z = 14.2857\bar{x} - 1.42857$$

$$14.2857\bar{x} = Z + 1.42857$$

$$\bar{x} = \frac{1}{14.2857}(Z + 1.42857)$$

$Z = -1.5$
-1.43

$\bar{x}$ = -ve, when $Z < -1.42857$

$\therefore P(\bar{x} < 0) = P(Z < -1.42857) = 0.076359$

$P(Z > 1.42857)$

6) A sample of 400 items is taken from a normal population whose mean is 4 and variance 4. If the sample mean is 4.45, can the sample be regarded as a simple sample?

Soln: H_0: Sample of be regarded as a simple sample from the population

$n = 400$, $\mu = 4$, $\sigma^2 = 4 \Rightarrow \sigma = 2$, $\bar{x} = 4.45$

$$\therefore Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{4.45 - 4}{\frac{2}{\sqrt{400}}} = \frac{0.45 \times 20}{2} = 4.50$$

As $Z = 4.50 > 2.58$, the deviation of the sample mean from the mean of the population is significant at 1% level of significance. Hence it cannot be regarded as a simple sample.

7) To know the mean weights of all 10 year old boys in Delhi, a sample of 225 is taken. The mean weight of the sample is found to be 67 pounds with a S.D of 12 pounds. Can you draw any inference from it about the mean weight of the population?

Sol^n: $n = 225$, $\bar{x} = 67$, $\sigma = 12$, $\sqrt{n} = \sqrt{225} = 15$

Confidence limits for mean of the population is

$$\left(\bar{x} - 2.58\frac{\sigma}{\sqrt{n}}\right) < \mu < \left(\bar{x} + 2.58\frac{\sigma}{\sqrt{n}}\right)$$

$$\left(67 - 2.58 \times \frac{12}{15}\right) < \mu < \left(67 + 2.58 \times \frac{12}{15}\right)$$

$$(67 - 2.064) < \mu < (67 + 2.064)$$

$$64.936 < \mu < 69.064$$

$\therefore$ Mean weight lies b/w 64.9 and 69.06

8) If the mean breaking strength of copper wire is 575 lbs with a SD of 8.3 lbs. How large a sample must be used in order that there be one chance in 100 that the mean breaking strength of the sample is less than 578 lbs?

Soln: $\mu = 575$, $\sigma = 8.3$, $\bar{x} = 572$

$$P(\bar{x} < 572) = \frac{1}{100} = 0.01$$

$$|Z| = \left| \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \right| = \left| \frac{572 - 575}{\frac{8.3}{\sqrt{n}}} \right| = \frac{3\sqrt{n}}{8.3} \rightarrow (1)$$

Given,

$$P(\bar{x} < 572) = 0.01$$

i.e. $P\left(Z < \frac{3\sqrt{n}}{8.3}\right) = 0.01$

$$\frac{3\sqrt{n}}{8.3} = 2.33$$

From the normal dist^n table $Z = +2.33$

Subs in (1)

$$2.33 = \frac{3}{8.3}\sqrt{n} \Rightarrow \sqrt{n} = 6.446$$

$$\Rightarrow n = 42$$

Test of significance for means of two large samples

a) Suppose two random samples of sizes n_1 & n_2 have been drawn from the same population with S.D σ. We wish to test whether the diff bet^n the sample means $\bar{x}_1 + \bar{x}_2$ is significant or is merely due to fluctuations and sampling.

If the samples are independent, then the standard error e of the diff of their means is given by

$$\boxed{e^2 = e_1^2 + e_2^2}$$

where $e_1 = \frac{\sigma}{\sqrt{n_1}}$, $e_2 = \frac{\sigma}{\sqrt{n_2}}$ are the SE's of the two samples

$$e^2 = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]$$

$$e = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Hence $$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

is normally distributed with mean zero & S.D 1.

Test of significance (n_1, n_2 being large)

a) If $Z > 1.96$, then the diff is significant at 5% level of significance. If $Z > 3$, it is highly probable that either the samples have not been drawn from the same population or the sampling is not simple.

b) If the samples are known to be drawn from different populations with means μ_1, μ_2 and standard deviations σ_1 & σ_2. Then the S.E of their means is given by $$e = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Assuming that the two populations have the same mean (i.e $\mu_1 = \mu_2$) the diffrene of the means of the samples will be normally distributed with mean zero and S.D e Now the same procedure of test of significance is applied.

3) The means of two large samples of 1000 and 2000 members are 168.75 cms and 170 cms resp can the samples be regarded as drawn from the same population of standard deviation 6.25 cm.

Soln: $n_1 = 1000$, $n_2 = 2000$, $\bar{x}_1 = 168.75$, $\bar{x}_2 = 170$

Ho: Samples are drawn from same population

$$\sigma = 6.25$$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} = \frac{1.25}{0.009375} = 5.16 > 1.96 \text{ \& } 2.58$$

The difference bet^n the means of samples is very much greater than 1.96 and is therefore significant

Thus the samples are not drawn from same population

4) If 60 new entrants in a given university are found to have a mean ht of 68.60 inches & 50 seniors a mean ht of 69.51 inches, is the evidence conclusive that the mean ht of the seniors is greater than that of the new entrants? Assume the S.D of ht to be 2.48 inches.

Soln: $n_1 = 60$, $\bar{x}_1 = 68.60$, $n_2 = 50$, $\bar{x}_2 = 69.51$, $\sigma = 2.48$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} = \frac{0.91}{2.48\left(\frac{1}{50} + \frac{1}{60}\right)} = \frac{0.91}{0.0909} = 10.011 > 1.96$$

the difference is very much greater than 1.96 and is there significant. Thus the mean ht of seniors is not greater than that of new entrants.

1. The means of simple samples of sizes 1000 and 2000 are 67.5 and 68.0 cm respectively can the samples be regarded as drawn from the same population of S.D 2.5 cm.

Sol^n: We have $\bar{x}_1 = 67.5$, $\bar{x}_2 = 68.0$; $n_1 = 1000$

$n_2 = 2000$

on the hypothesis, that the H_0: samples are drawn from same population of S.D $\sigma = 2.5$, we get

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{67.5 - 68.0}{2.5\sqrt{\frac{1}{1000} + \frac{1}{2000}}} = \frac{0.5}{2.5 \times 0.0387}$$

$$= 5.1$$

Hence the diff b/t the samples means i.e. 5.1 is very much greater than 1.96 and is therefore significant. Thus the samples (cannot) can't be regarded as drawn from the same population.

2. A sample of height of 6400 soldiers has a mean of 67.85 inches and a SD of 2.56 inches while a simple sample of hts of 1600 sailors has a mean of 68.55 inches and a S.D of 2.52 inches. Do the data indicate that the sailors are on the avg taller than soldiers?

Soln: Here $\bar{x}_1 = 67.85$, $\bar{x}_2 = 68.55$, $\sigma_1 = 2.56$

$\sigma_2 = 2.52$, $n_1 = 6400$, $n_2 = 1600$

Ho : There is no significant diff in height

∴ SE of the diff of the mean heights is

$$e = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(2.56)^2}{6400} + \frac{(2.52)^2}{1600}}$$

$$= \sqrt{0.001024 + 0.003969} \simeq 0.005$$

Also difference bet^n the mean $= \bar{x}_2 - \bar{x}_1 = 0.7$

which $> 10e$. This is highly significant. Hence the data indicates that the sailor are on the average taller than the soldiers

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{e} = \frac{0.7}{0.005} = 140 > 2.58$$

Sampling of variables – Small samples :

In case of large samples, sampling distribution approaches a normal distribution & values of sample statistic are considered best estimates of the parameters in a population. It will no longer be possible to assume that statistics computed from samples are normally distributed. As such, a new technique has been devised for small samples which involves the concept of "degree of freedom"

Number of degrees of freedom is the no. of values in a set which may be assigned arbitrarily. For instance, if $x_1 + x_2 + x_3 = 15$ and we assign any value of two of the vars (say x_1, x_2), then the values of x_3 will be known. The two vars are therefore, free & independent choices for finding the third. Hence there are the degrees of freedom. If there are n observations, the degrees of freedom are $(n-1)$. In other words, while finding the mean of a small sample one degree of freedom is used up and $(n-1)$ degrees of freedom are left to estimate the population variance.

Student's t - Distribution

Consider a small sample of size n, drawn from a natural population with mean μ and S.D σ. If $\bar{x}$ and σ_s be the sample mean and S.D. then the static, 't' is defined as

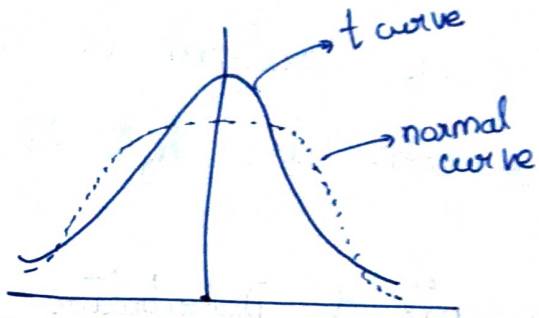
$$t = \frac{\bar{x} - \mu}{\sigma} \sqrt{n} \quad \text{or} \quad \boxed{t = \frac{\bar{x} - \mu}{\sigma_s} \sqrt{(n-1)}}$$

where $\nu = n-1$ denotes the degree of freedom (df) of t. If we calculate t for each sample, we obtain the sampling dist^n for t. This dist^n known as student's t-distribution is given by $y = \frac{y_0}{\left(1 + \frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}} \longrightarrow (1)$

where y_0 is const such that the area under the curve i.e. units

Properties of t-distribution

1. This curve is symmetrical about the line $t=0$, like the normal curve, since only even powers of t appear in (1). But it is more peaked than the normal curve with the same S.D. The t-curve approaches the horizontal axis less rapidly than the normal curve. Also t-curve attains its maximum value at $t=0$, so that its mode coincides with the mean.



2) The limiting form of t-distribution when $v \to \infty$ is given by $y = y_0 e^{-\frac{1}{2}t^2}$ which is a normal curve. This shows that t is normally distributed for large samples.

3) The probability P that the value of t will exceed t_0 is given by $P = \int_{t_0}^{\infty} y \, dx$

4) Moments about the mean: All the moments of odd order the mean are zero, due to its symmetry about the line $t=0$

Even order moments about the mean are

$$\mu_2 = \frac{V}{V-2}, \quad \mu_4 = \frac{3V^2}{(V-2)(V-4)}$$

The t-distribution is often used in tests of hypothesis about the mean when the popul^{n} std deviation σ is unknown

Significance test of a sample Mean

Given a random sample $x_1, x_2, x_3, \ldots x_n$ from a normal population we have to test the hypothesis that mean of the populatn is μ. For this, we first calculate $t = \dfrac{(\bar{x} - \mu)\sqrt{n-1}}{\sigma_s}$

$$\sigma_s^2 = \frac{1}{n}\sum x_i^2 - (\bar{x})^2$$

where $\bar{x} = \frac{1}{n}\sum_1^n x_i$, $\quad \sigma_s^2 = \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2$

Then find the value of for the given df from the table. If the calculated value of $t > t_{0.05}$, the difference betn $\bar{x}$ and μ is said to be significant at 5% level of significance. If $t > t_{0.01}$, the difference is said to be significant at 1% level of significance.

If $t < t_{0.05}$, the data is said to be consistent with the hypothesis that μ is the mean of the population.

1) A certain stimulus administered to each of 12 patients resulted in the following increases of blood pressure: 5, 2, 8, -1, 3, 0, -2, 1, 5, 0, 4, 6. Can it be concluded that the stimulus will in general be accompanied by an increase in blood pressure.

Soln: Let us assume that the stimulus administered to all the 12 patients will increase the B.P. Taking the population to be normal with mean $\mu = 0$ & S.D σ

$$\bar{x} = \frac{5+2+8-1+3+0-2+1+5+0+4+6}{12} = 2.583$$

$$\sigma_s^2 = \frac{\Sigma d^2}{n} - \bar{d}^2 = \frac{1}{12}[5^2+2^2+8^2+1^2+3^2+0^2+2^2+1^2 + 5^2+0^2+4^2+6^2] - [2.583]^2$$

$$\sigma_s^2 = 8.744 \qquad \therefore \sigma_s = 2.971$$

Now $t = \frac{\bar{x}-\mu}{\sigma_s}\sqrt{n-1} = \frac{2.583-0}{2.9571}\sqrt{(12-1)} = 2.897$

Here d.f $\gamma = 12-1 = 11$

For $\gamma = 11$, $t_{0.05} = 2.2$ from the table

Since $|t| > t_{0.05}$ (i.e. $2.897 > 2.2$) our assumption is rejected i.e. the stimulus doesnot increase the B.P

2) The nine items of a sample have the following values: 45, 47, 50, 52, 48, 47, 49, 53, 51. Does the mean of these differ significantly from the assumed mean of 47.5?

Soln: We find the mean and S.D of the sample as follows

x	$d = x-48$	d^2
45	-3	9
47	-1	1
50	2	4
52	4	16
48	0	0
47	-1	1
49	1	1
53	5	25
51	3	9
Total	10	66

$\bar{x}$ = mean $= 48 + \frac{\Sigma d}{9} = 48 + \frac{10}{9} = 49.1$

$\sigma_s^2 = \frac{\Sigma d^2}{9} - \left(\frac{\Sigma d}{9}\right)^2 = \frac{66}{9} - \frac{100}{81} = \frac{494}{81}$

$\sigma_s = 2.47$

Hence $t = \frac{\bar{x} - \mu}{\sigma}\sqrt{(n-1)} = \frac{49.1 - 47.5}{2047}\sqrt{8} = \underline{\underline{1.83}}$

Here d.f $V = 9-1 = 8$

For $V=8$, we get from table, $t_{0.05} = 2.31$

As $t < t_{0.05}$, the value of t is not significant at 5% level of significance between $\bar{x}$ and μ.

Thus the test provides no evidence against the population mean being 47.5.

3) The mechinist is making engine parts with axle diameter of 0.7 inch. A random sample of 10 parts shows mean diameter 0.742 inch with a S.D of 0.04 inch on the basis of this sample, would you say that the work is inferior?