MARUDHAR KESARI JAIN COLLEGE FOR WOMEN (AUTONOMOUS)

VANIYAMBADI

PG and Department of Mathematics

1st M.COM – Semester - I

E-Notes (Study Material)

Allied Paper -1: Business Mathematics and Operation Research – I **Code: 24PCOE12**
Unit: 3 –Sampling and Hypothesis testing **Sampling- Sampling methods, Sampling error and standard error-relationship between sample size and standard error. Testing hypothesis-testing of means and proportions- large and small samples- z test (15 Hours)**
Learning Objectives: To understand the sampling and hypothesis testing
Course Outcome: Explain the sampling, hypothesis testing, large and small samples

Overview:

Sampling means selecting the group that you will actually collect data from in your research. For example, if you are researching the opinions of students in your university, you could survey a sample of 100 students. In statistics, sampling allows you to test a hypothesis about the characteristics of a population.

A sampling error is a statistical error that occurs when an analyst does not select a sample that represents the entire population of data. As a result, the results found in the sample do not represent the results that would be obtained from the entire population.

Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution. First, a tentative assumption is made about the parameter or distribution. This assumption is called the null hypothesis and is denoted by H_0.

- Sampling
- Sampling methods
- Sampling error
- Standard error
- Sample size
- Testing hypothesis
- Means and Proportions
- Large samples
- Small samples
- Z test

Unit-3 Sampling and Sampling distribution

Definition of sampling:

The process of selecting a sample from a population is called sampling.

In sampling a representative sample or portion of elements of a population or process is selected and then analyzed.

On sample results called sample statistics statistical inferences are made about the population characteristics.

Reasons of Sample Survey:

A censes is a count of all the element in a population. Few example of censes are,

⇒ Population of eligible voters, censes of consumer preference to a particular product, buying habits of adult India. Sum of the reasons to prefer sample survey instead of censes are given below.

* Movement of population element
* Cost and/or Time required to contact whole population.
* destructive nature of certain tests

Movement of population element:

The population of fish, birds, snake, mosquitos etc.. are large and or constantly moving, being born and ~~dieding~~ dying so instead of attempting to count all element of such population, it is desirable to make estimates using techniques essential such as counting birds at a place picked at a random setting at nest at pre determined places, etc.

Cost and time required to contact whole population

Time required to contact the whole population. A Censes involves a complete count of every individual member of the population of interest, such as persons in state, households in town, shops in city, students in a college and so on. Apart from them cost and large amt of resources that are required main problem is the time required to process the data. Hence the results are known ... after the big gap of time.

Destructive nature of certain tests :

The Censes become extremely difficult, if not impossible, when the population of interest is either infinite ... in terms of size (Number) constantly changing, in state of a movement or observation result required destruction.

For Example - Sometimes it is required to test the strength of some manufactured items by applying a stress untill the units breaks.

The amt of stress that result in breakage is the value of the observation that is recorded. If the procedure is applied to entire population, there would be nothing left. This type of testing is called destructive test. and requires the sampling be used in such cases

Sampling and non sampling errors:

Any statistical inference based on sample result may not always be correct because sample results are either based on partial or incomplete analysics of population futures. This error is referred to the sampling error because each sample taken may produce a different estimate of population characteristics compared to those result that could have be optioned by complete enumeration. It is necessary to measure so as to have an extract item about the reliability of sample based estimates of population . Sampling error exists any simplified magnitude must always be simplified in terms of a probability value say 5%..

This acceptable margin of error is then used to produce a confidence in the decision maker to arrive at certain conclusions with the limited data at his disposal. Decision make wish to be 95% or more confident that the range of values of sample result reflect the true characteristics of population or process of interest.

Non-Sampling errors!

Non-Sampling error arise during census as well as sampling surveys due to biases and mistakes such as

* Incorrect enumeration of population members.
* Non random selection of samples.

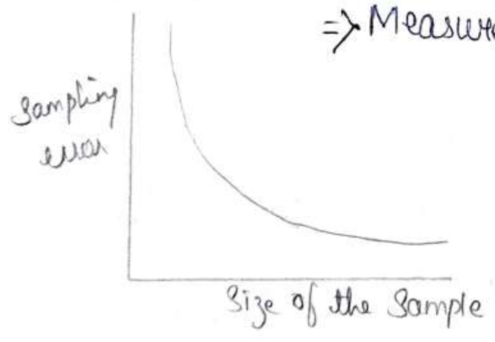
* Use incomplete vague (or) faculty questionnaire for data Collection.

* Wrong editing Coditing and presenting of the responses received through the questionnaire.

Measurement of Sampling Error:

A measure of Sampling error is provided by the standard error of the estimate. Estimation of Sampling error can reduce the element of uncertainty associated with interpretation of data. In most cases, the degree of precision or the level of error, would depend on the size of the sample. The standard error of estimate is inversely proportional to the square root of the sample size. In other words, as the sample size increases, element of error is reduced.

=> Measure of Sampling Error



Population parameters and Sample statistics:

parameters: An exact, but generally unknown measure (or value) which describes the entire population or process characteristics is called a parameter.

For example, Quantities such as mean μ, Variance σ^2, Standard deviation σ, median, mode and proportion P computed from a data set (also called population) are called parameters.

A parameter is usually denoted with letters of the lower case Greek alphabet such as mean μ and standard deviation σ.

Sample statistics :

A measure (or value) found from analysing sample data is called a sample statistic or simply a statistic.

Inferential statistical methods attempt to estimate population parameters using sample statistics.

Sample statistics are usually denoted by Roman letters such as mean $\bar{x}$, standard deviation s, variance s^2 and proportion $\bar{p}$.

(2m) ✗ ∴ Sample statistics - A sample measure, such as mean $\bar{x}$, standard deviation s, proportion $\bar{p}$ and so on.

Estimation Relationship between Sample and Population Measures.

Sample Statistics $(\bar{x}, s, s^2, \bar{p}, n)$ —— Population parameters $(\mu, \sigma, \sigma^2, P, N)$

Estimate

principles of Sampling:

The following are two important principles which determine the possibility of arriving at a valid statistical inference about the features of a population or process:-

i) principle of statistical regularity
ii) principle of Inertia of large numbers.

principle of Statistical regularity:

According to King, "The law of statistical regularity lays down that a moderately large number of items chosen at random from a large group are almost sure on the average to possess the characteristic of the large group."

(This principle is based on the mathematical theory of probability)

=> This principle, emphasises on two factors:

i) Sample size should be large:

As the size of sample increases it becomes more and more representative of parent population and shows its characteristics.

However, in actual practice, large samples are more expansive.

Thus, a balance has to be maintained between the sample size, degree of accuracy desired and financial resourses available.

ii) Samples Must be drawn Randomly:

The random sample is the one in which elements of the population are drawn in a such way that each combination of elements has an equal probability of being selected in the sample.

When the term random sample is used without any specification. It usually refers to a simple random sample.

The selection of samples based on this principle can reduce the amount of efforts required in arriving at a conclusion about the characteristic of a large population.

For Example: TO understand the book buying habit of students in a college, instead of approaching every student, it is easy to talk to a randomly selected group of students to draw the inference about all students in the college.

Principle of Inertia of large numbers:

This principle is a corollary of the principle of statistical regularity and plays a significant role in the sampling theory.

This principle states that, under similar conditions rate, as the sample size (number of observations in a sample) get large enough, the statistical inference is likely to be more accurate and stable.

For Example → If a coin is tossed a large number of times, then relative frequency of occurance of head and equal is expected to be equal.

Sampling Methods :- It is classified into two types

~~Simple Random~~ * probability Sampling

* Non-probability Sampling

Sampling Methods

Probability Sampling
- Simple Random
- Systematic
- Stratified
- Cluster

Non-probability Sampling
- convenience
- Quota
- Judgement
- Snowball.

Probability Sampling: Probability Sampling means that every individual in a population Stands an equal chance of being Selected.

It is that where the samples are selected without any restrictions that is in a random manner. So the results will be appropriate since the samples are selected in an unbiased way.

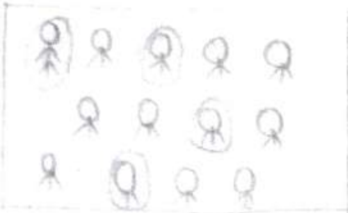
It is otherwise Known as "Random Sampling".

i) Simple random Sampling:

The Simplest type of probability sampling. Researchers take every individual in a population and randomly select their sample, often using some type of computer program or random number generator.

In Simple random Sampling, every unit of the population has got an equal chance of being Selected. Every unit is Selected at random.

Example → It is done by lottery method.



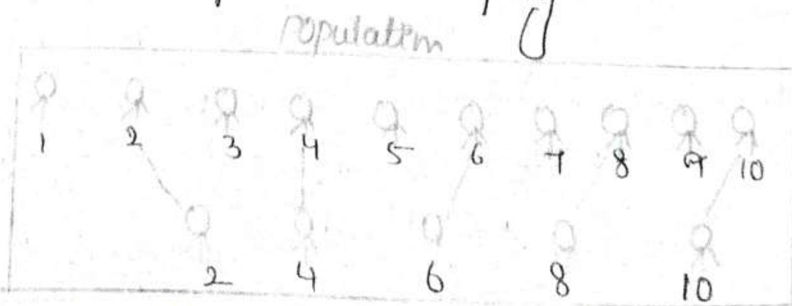
Systematic Random Sampling

Systematic Sampling is similar to Simple random Sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular inter..

In Systematic Sampling, the first unit in the population is selected at random, thereafter every Kth items is Selected in order to have samples at specified intervals.

This may be arranged either in ascending order or descending order or in Some alphabetical or numerical order.

Systematic Sampling



∴ In this case, every Second person is systematically Selected.

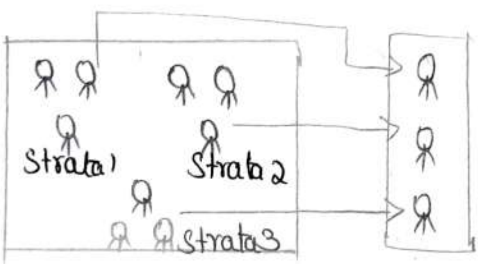
Stratified Random Sampling:

The population is divided into mutually exclusive groups (Such as groups) and random samples are drawn from each group.

In this method the entire heterogeneous population is divided into small sub units known as "Stratas". Based on the relevant characteristics (e.g. gender, age range, income bracket, job role).

These Stratas are homogeneous among themselves with respect to certain common factor or characteristics.

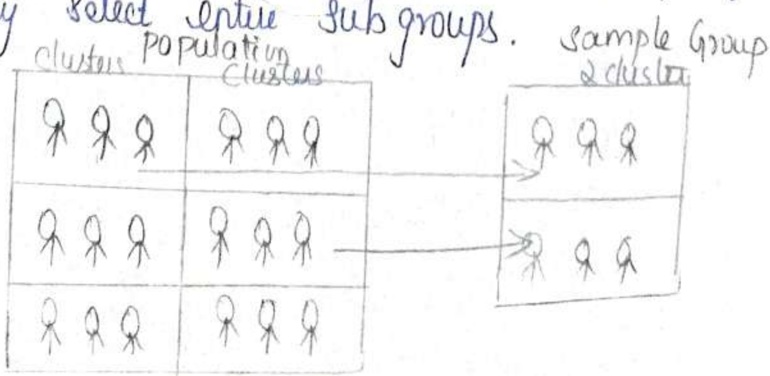
The items / Sampling units are randomly selected from these stratas that together make up the sample.



Cluster (area) Sample: (This method is used when the population size is large)

The population is divided into mutually exclusive group (Such as city blocks) and the researcher draws a sample of the groups to interview.

It also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each sub group, you randomly select entire sub groups.

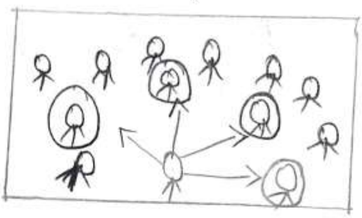


Non-probability Sampling: (Non-Random Sampling)

It is that samples are not selected in a random manner. It is otherwise known as biased sampling or restricted sampling.

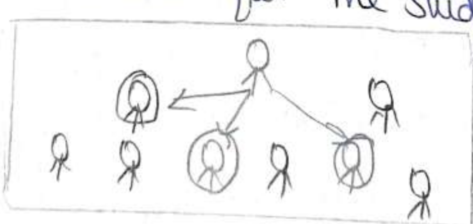
Here the samples are selected according to convenience of the research.

Convenience Sampling: In this method the research simply selects the samples that are easily available and accessible. No extra efforts are taken by the researcher, the samples are selected only based upon the convenience.



Judgemental / purposive Sampling:

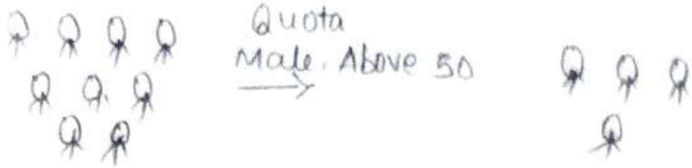
In this method of sampling, the researcher choses samples based upon his/her own judgement. The researcher selects the samples in which his opinion will be best for the study.



Quota Sampling:

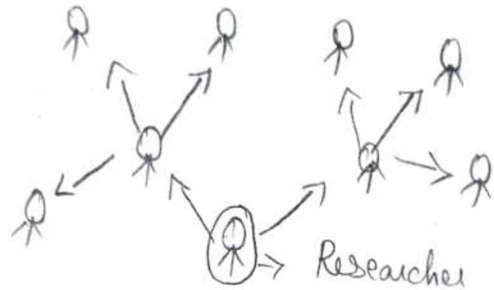
In this method of sampling, quotas in the form of reservation or percentage are established for different classes of population based upon age, gender, nationality, income etc..

A sample is drawn out on the basis of these quotas.



Snowball Sampling:

In this method of sampling, the researcher selects samples first based upon his judgement then chooses according the directions/ advice/ referrals provided by the first sampling unit.



Researcher gets sampling like chain process

Formula :-

(X) Sampling distribution of Mean when population as normal distribution

Population Standard deviation σ is known:

For any given sample of size n taken from a population with mean μ and S.D σ, the sampling distributio of sample statistic such as mean and S.D are defined

- Mean of the distribution of sample means or expected value of the mean

$$\mu_{\bar{x}} = \mu = E(\bar{x})$$

Standard deviation (or) Error of the distribution of sample mean (or) Standard error of mean

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

The value of sample mean $\bar{x}$ is first converted very into value of Z. on the Standard N.Dis to know any single mean value deviates from the mean $\bar{x}$ of sample mean value by using the formula

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

Since $\sigma_{\bar{x}}$ measures the dispersion disposal of values of sample means in the sample distribution of the means

(X) $\bar{x} \pm \sigma_{\bar{x}}$ covers about the middle 68% of the total possibile sample means

$\bar{x} \pm 1.96\,\sigma_{\bar{x}}$ covers about the middle 95% of the total possible sample means (large sample)

The procedure for making statistical inference using Sample Distribution about the population mean $\boxed{\mu}$ ~~and mean~~ based on mean $\bar{x}$ sample mean is summarized has follows

- If the population SD σ value is known and either

a) population distribution is normal

b) population distribution is not normal but the sample size is 'n' large is ($n \geq 30$) then sampling distribution of mean $\mu_{\bar{x}} = \mu$ and SD $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ is ~~and SD~~ very close to standard normal distribution is given by $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

The factor has little effect on reducing the amt of sampling error when the size of sample is less than 5% of population error. But N is large relative to sample size n.

$\sqrt{(N-n)(N-1)}$ is approximately equal to 1

(X) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$ In case of two 'n' values.

1. The mean length of life of a certain Cutting tool 4.15 hour with a S.D 2.5 hours. What is the probability that they are simple random sample of size 50 drawn from this population will have a mean between 40.5 hours & 42 hours

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

$n = 50$

$\mu_{\bar{x}} = \mu = 41.5 \quad \sigma = 2.5$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.5}{\sqrt{50}} = \frac{2.5}{7.071} = 0.3535$$

$P(40.5 > z > 42)$

$$= \left(\frac{\bar{x} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}} > z > \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}} \right)$$

$$= \left(\frac{40.5 - 41.5}{0.3535} > z > \frac{42 - 41.5}{0.3535} \right)$$

$= (-2.82885 > z > 1.41442)$

$= P(z > -2.82885) + P(z < 1.41442)$

$= 0.4976 + 0.4207$

$= 0.9183.$

2

A Continuous manufacturing process produces items who's weights are normal distributed with a mean weight 800g and S.D 300g a random sample of 16 items to be drawn from the process

01. What is the probability that the arithmetic mean of the sample exist 900g interpect the result?

02. Find the value of sample arithmetic mean within which the middle 95% of the sample mean will fall?

Sol

i) $P(Z \geq 900)$

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{300}{\sqrt{16}} = \frac{300}{4} = 75$$

$$= \frac{900-800}{75}$$

$$= \frac{100}{75}$$

$$Z = 1.33$$

$$P(Z \geq 1.33) = 0.5 - 0.4082$$

$$= 0.0918.$$

ii) $Z = 95\% = 1.96 \left[\therefore AL = \frac{1+CL}{2} = \frac{1+0.95}{2} = \frac{1.95}{2} = 0.975 \right.$

$$\left. = 1.9 + 0.06 = 1.96 \right]$$

$$\bar{x}_1 = \mu_{\bar{x}} + Z\sigma_{\bar{x}} = 800 + 1.96(75) = 800 + 147 = 947$$

$$\bar{x}_2 = \mu_{\bar{x}} - Z\sigma_{\bar{x}} = 800 - 1.96(75) = 800 - 147 = 653$$

3. An oil refinary as back up monitors to keep track of the refinary flows continuously and to prevent machine malfunctions from disrupting the process. One particular monitor has an average life of 4300 hours and SD of 730 hrs. In addition to the primary monitor, the refinary has set up two standby units which are duplicates of the primary one. In case of malfunction of one of the monitors another will automatically take over in its place. The operating life of each monitor is independent of the other. 4.

a) What is the probability that a given set of monitors will last at least 13000 hours?

b) Atmost 12630 hrs?

Sol $\mu = 4300$ hrs $\sigma = 730$ $n = 3$

$\mu_{\bar{x}} = \mu = 4300$

S.D $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{730}{\sqrt{3}} = \frac{730}{1.732} = 421.48.$

a) $P(\bar{x} \geq 4333.33) = P\left[\frac{\bar{x}-\mu}{\sigma_{\bar{x}}} \geq \frac{4333.33 - 4300}{421.48}\right]$

$\frac{13000}{3} = 4333.33$

$= P[z \geq 0.08]$

$= 0.5 - 0.0319 = 0.4681$

b) For the set to last at most 12630 hrs, the average life cann't exceed $12630/3 = 4210$ hrs.

$P(\bar{x} \leq 4210) = P\left[\frac{\bar{x}-\mu}{\sigma_{\bar{x}}} \leq \frac{4210-4300}{421.48}\right]$

$= P[z \leq -0.213]$

$= 0.5 - 0.0832$

$= 0.4168$

5. Safal, a tea manufacturing company is interested in determining the consumption rate of tea per household in Delhi. The management believes that yearly consumption per house holder is normally distributed with an unknown mean μ and SD of 1.50 kg

i) If a sample of 25 household is taken to record their consumption of tea for 1 year. What is the probability that the sample mean is within 500 kg of the population mean?

ii) How large a sample must be in order to be 98% certain that the sample mean is within the 500 gram of the population mean.

gr - km ÷ 1000

i)

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.50}{\sqrt{25}} = \frac{1.50}{5} = 0.3$$

$\mu = 500$ g to kg

$$= \frac{500}{1000} = 0.5$$

$P(-t\alpha \leq \bar{x} \leq t\alpha)$ $P(-t\alpha \leq \bar{x} \leq t$

$P(\mu - 0.5 \leq z \leq \mu + 0.5)$ $P(\mu - 0.5 \leq \bar{x}$

$P\left(\frac{-0.5}{0.3} \leq z \leq \frac{0.5}{0.3}\right)$

$P(-1.66 \leq z \leq 1.66)$

$0.4515 + 0.4515$

$= 0.903$

ii) $Z = 98\%$ $\left[AL = \frac{1+CL}{2} = \frac{1+0.98}{2} = \frac{1.98}{2} \right.$

$= 0.99$

$2.3 + 0.03 = 2.33$

$98\% = 2.33 = \dfrac{\mu}{\frac{\sigma}{\sqrt{n}}}$

$2.33 = \dfrac{0.5}{\frac{1.50}{\sqrt{n}}}$

$2.33 = \dfrac{0.5\sqrt{n}}{1.50}$

$2.33 = 0.33\sqrt{n}$

$\dfrac{2.33}{0.33} = \sqrt{n}$

$\sqrt{n} = 7.06$

$n = (7.06)^2$

$= 49.85.$

Sampling Distribution of difference between two Sample means:

The Concept of Sampling distribution of Sample mean introduced earlier in the Chapter Can also be used to Compare a population of size N_1 having mean μ_1 and SD σ_1 with another Similar type of Population of size N_2 having mean μ_2 and SD σ_2

Let $\bar{x}_1$ and $\bar{x}_2$ be the mean of sampling distribution of mean of two population respectively. Then the difference between their Mean values μ_1 and μ_2 Can be estimated by generalizing the formula of Standard normal variable as follows.

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_{\bar{x}_1} - \mu_{\bar{x}_2})}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$ (mean of sampling distribution of difference of two mean).

$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ (Standard error of sampling distribution of two mean)

n_1, n_2 = Independent random samples drawn from first and Second population.

Since random Samples are drawn independently from two population with replacement, therefore the Sampling distribution of the difference of two means $\bar{x}_1 - \bar{x}_2$ will be normal provided sample

Size is sufficiently large

The standard error of sampling distribution of some other statistics is given below.

Sampling distribution	Standard Error & mean	Remarks
Median	$\sigma_{med} = 1.2533 \frac{\sigma}{\sqrt{n}}$ $\mu_{med} = \mu$	For a large sample size $n \geq 30$, the sampling distribution of median approaches normal distribution. This result is true only if the population is normal or approximately normal.
Sample standard deviation	i) $\sigma_s = \frac{\sigma}{\sqrt{2n}}$ ii) $\sigma_s = \sqrt{\frac{\mu_4 - \mu_2^2}{4n\mu_2}}$ iii) $\mu_s = \sigma$	• For a large sample size $n \geq 100$, the sampling distribution is close to normal distribution. • If population is normally distributed, then σ_s is calculated using (i) otherwise (ii) • μ_2 and μ_4 are second and fourth moments, where $\mu_2 = \sigma^2$ and $\mu_4 = 3\sigma^4$

1. Car Stereos Manufacturer A have a mean lifetime of 1400 hrs with a SD of 200 hrs, while those of manufacturer B have mean life-time of 1200 hrs with a SD of 100 hrs. If a random sample of 125 stereos of each manufacturer are tested. What is the probability that manufacturer A's stereos will have a mean lifetime which is atleast a) 160 hrs more than manufacturer B's stereos and b) 250 hrs more than the manufacture B's stereos.

Sol

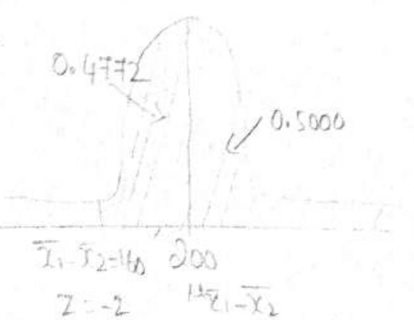
Manufacturer A : $\mu_1 = 1400$ hrs $\sigma_1 = 200$ hrs $n_1 = 125$

Manufacturer B : $\mu_2 = 1200$ hrs $\sigma_2 = 100$ hrs $n_2 = 125$

$$\mu_{\bar{x}_1-\bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2 = 1400-1200$$
$$= 200$$

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(200)^2}{125} + \frac{(100)^2}{125}}$$
$$= \sqrt{80+320} = \sqrt{400} = 20$$

a) $$P[(\bar{x}_1-\bar{x}_2) \geq 160] = P\left[z \geq \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sigma_{\bar{x}_1-\bar{x}_2}}\right]$$
$$= P\left[z \geq \frac{160-200}{20}\right]$$
$$= P[z \geq -2]$$
$$= 0.5000 + 0.4772$$
$$= 0.9772$$



Hence, the probability is very high that the mean lifetime of the stereos of A is 160 hours more than that of B.

b) $P[(\bar{x}_1-\bar{x}_2 \geq 250] = P\left[z \geq \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sigma_{\bar{x}_1-\bar{x}_2}}\right]$

$= P\left[z \geq \frac{250-200}{20}\right]$

$= P[z \geq 2.5]$

$= 0.5000 - 0.4938$

$= 0.0062$

∴ Hence the probability is very less that the mean life time of the stereos of A is 250 hrs more than that of B.

2. The particular brand of ball bearings weights 0.5 kg with a SD of 0.02 kg. What is the probability that two lots of 1000 ball bearings each will differ in weight by more than 2grms?

Sol) Lot 1: $\mu_{\bar{x}_1} = \mu_1 = 0.50$ kg $\sigma_1 = 0.02$ kg $n_1 = 1000$

Lot 2: $\mu_{\bar{x}_2} = \mu_2 = 0.50$ kg $\sigma_2 = 0.02$ kg $n_2 = 1000$

$\mu_{\bar{x}_1-\bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2 = 0.50 - 0.50$

$= 0$

$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{(0.02)^2}{1000} + \frac{(0.02)^2}{1000}}$

$= 0.000895$

A difference of 2 gms in two lots is equivalent to a difference of $2/1000 = 0.002$ kg.

It is possible if $\bar{x}_1 - \bar{x}_2 \leq 0.002$ (or) $\bar{x}_1 - \bar{x}_2 \geq -0.002$

$P[-0.002 \leq \bar{x}_1 - \bar{x}_2 \leq 0.002]$

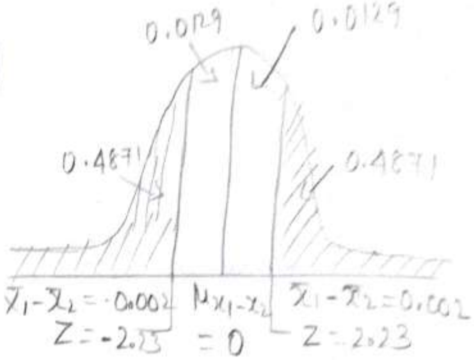
$$= P\left[\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sigma_{\bar{x}_1-\bar{x}_2}} \le Z \le \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sigma_{\bar{x}_1-\bar{x}_2}}\right]$$

$$= P\left[\frac{-0.002}{0.000895} \le Z \le \frac{0.002}{0.000895}\right]$$

$$= P[-2.23 \le Z \le 2.23]$$

$$= 2[0.5000 - 0.4871]$$

$$= 0.0258.$$



Sampling distribution of sample proportion:

- $$\bar{p} = \frac{\text{Elements of sample having characteristics, } x}{\text{Sample size, } n}$$

With the same logic of sampling distribution of mean, the sampling distribution of sample proportions with mean $\mu_{\bar{p}}$ and SD (also called standard error $\sigma_{\bar{p}}$ is given by

$$\mu_{\bar{p}} = p \text{ and } \sigma_{\bar{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$$

If the sample size n is large ($n \ge 30$) the sampling distribution of $\bar{p}$ can be approximated by a normal distribution. The approximation will be adequate if

$$np \ge 5 \text{ and } n(1-p) \ge 5$$

It may be noted that the sampling distribution of the proportion would actually follow binomial distribution because population is binomially distributed.

The mean and SD (error) of the sampling distribution of proportion are valid for a finite population in which sampling is with replacement. However, for finite population in which sampling is done without replacement, we have

$$\mu_{\bar{p}} = P \text{ and } \sigma_{\bar{p}} = \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}$$

Under the same guidelines as mentioned in previous sections, for a large sample size $n (\geq 30)$ the sampling distribution of proportion is closely approximated by a normal sample distribution with mean and SD as stated above. Hence to standard sample proportion $\bar{p}$, the standard normal variable

$$z = \frac{\bar{p} - \mu_{\bar{p}}}{\sigma_{\bar{p}}} = \frac{\bar{p} - p}{\sqrt{p(1-p)/n}}$$

is approximately the standard normal distribution.

Sampling distribution of the difference of two proportions:

1.

Two population of size N_1 and N_2 are given. Sample size n_1 from first population, Compute Sample proportion $\bar{p}_1$ and SD $\sigma_{\bar{p}_1}$. Then Sample size n_2 from 2nd population, compute Sample proportion $\bar{p}_2$ and SD $\sigma_{\bar{p}_2}$

For all combinations of these samples from these population, we can obtain a sampling distribution of the difference $\bar{p}_1 - \bar{p}_2$ of Samples proportion. Such a distribution is "called Sampling distribution of difference of two proportion.

The mean and SD of this distribution are given by

$$\mu_{\bar{p}_1 - \bar{p}_2} = \mu_{\bar{p}_1} - \mu_{\bar{p}_2} = P_1 - P_2$$

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\sigma^2_{\bar{p}_1} + \sigma^2_{\bar{p}_2}} = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

for probability

If sample size n_1 and n_2 are large, that is, $n_1 \geq 30$ and $n_2 \geq 30$, then the Sampling distribution of difference of propoition is closely approximated by a normal distribution.

1. A manufacturer of watches has determined from experience that 3 % of the watches he produces are defective. If a random sample of 300 watches is examined. What is the probability that the proportion defective is between 0.02 and 0.035?

$$\mu_{\bar{p}} = P = 0.03 \quad \bar{P}_1 = 0.02 \quad \bar{P}_2 = 0.035 \quad n = 300$$

Standard error proportion given by

$$\sigma_{\bar{p}} = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{0.03(1-0.0.3)}{300}}$$

$$= \sqrt{0.000097} = 0.0098$$

Calculating the desired probability

$$P[0.02 \le \bar{P} \le 0.035] = P\left[\frac{\bar{P}_1 - P}{\sigma_{\bar{p}}} \le Z \le \frac{\bar{P}_2 - P}{\sigma_{\bar{p}}}\right]$$

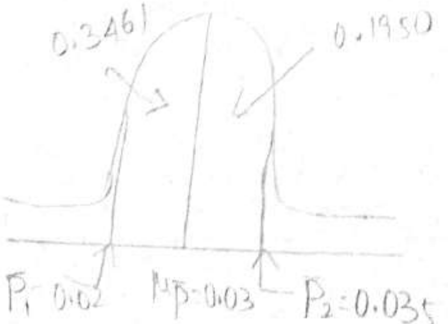
$$= P\left[\frac{0.02 - 0.03}{0.0098} \le Z \le \frac{0.035 - 0.03}{0.0098}\right]$$

$$= P[-1.02 \le Z \le 0.51]$$

$$= P(Z \ge -1.02) + P(Z \le 0.51)$$

$$= 0.3461 + 0.1950$$

$$= 0.5411$$



2. Few years back, a policy was introduced to give loan to unemployed engineers to start their own business. out of 100,000 unemployed engineers, 60,000 accept the policy and got the loan. A Sample of 100 unemployed engineers is taken at the time of allotment of loan. What is the probability that sample proportion would have exceeded 50% acceptance?

$\frac{60{,}000}{100000}$

Sol

$$\mu_{\bar{p}} = p = 0.60 \quad N = 100000 \quad n = 100$$

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

$$= \sqrt{\frac{0.60 \times 0.40}{100}}\sqrt{\frac{1{,}00{,}000-100}{1{,}00{,}000-1}}$$

$$= \sqrt{0.0024}\sqrt{0.9990}$$

$$= 0.0489 \times 0.9995 = 0.0488$$

The probability that sample proportion would have exceeded 50% acceptance is given by

$$P(x \geq 0.50) = P\left[z \geq \frac{\bar{p}-p}{\sigma_{\bar{p}}}\right]$$

$$= P\left[z \geq \frac{0.50-0.60}{0.0489}\right]$$

$$= P[z \geq -2.04]$$

$$= 0.5000 + 0.4793$$

$$= 0.9793$$

3 Ten percent of machines produced by Company A are defective and 5% of those produced by Company B are defective. A random sample of 250 machines is taken from Company A and a random sample of 300 machines from Company B. What is the probability that the difference in sample proportions is less than or equal to 0.02?

Sol

$P_1 = 10/100 = 0.1 \quad P_2 = 5/100 = 0.05$

$$\mu_{\bar{P}_1 - \bar{P}_2} = \mu_{\bar{P}_1} - \mu_{\bar{P}_2} = P_1 - P_2 \qquad n_1 = 250$$

$$= 0.1 - 0.05 \qquad n_2 = 300$$

$$= 0.05$$

$$\mu_{\bar{P}_1 - \bar{P}_2} = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}} = \sqrt{\frac{0.10 \times 0.90}{250} + \frac{0.05 \times 0.95}{300}}$$

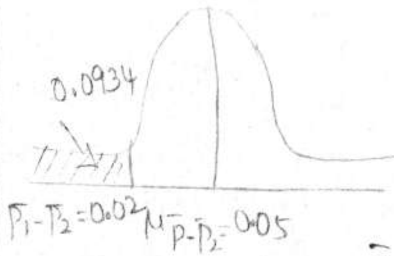
$$= \sqrt{\frac{0.90}{250} + \frac{0.0475}{300}} = \sqrt{0.00052}$$

$$= 0.0228$$

Desired probability of difference in sample proportions is given by

$$P[(\bar{P}_1 - \bar{P}_2) \le 0.02] = P\left[z \le \frac{(\bar{P}_1 - \bar{P}_2) - (\bar{P}_1 - \bar{P}_2)}{\sigma_{\bar{P}_1 - \bar{P}_2}}\right]$$

$$= P\left[z \le \frac{0.02 - 0.05}{0.0228}\right]$$



$$= P[z \le -1.32]$$

$$= 0.5000 - 0.4066$$

$$= 0.0934$$

Hypothesis :- Hypothesis is an assumption which may be true or false.

Types: Null hypothesis
Alternative hypothesis.

NULL HYPOTHESIS :- Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true. It is denoted by H_0.

$$H_0 : \mu = \mu_0$$

ALTERNATIVE HYPOTHESIS :- Any hypothesis which is complementary to the null hypothesis is called as the alternative hypothesis and It is denoted by H_1

i) $H_1 : \mu \neq \mu_0$ ($\mu >$ or $\mu < \mu_0$)

ii) $H_1 : \mu > \mu_0$

iii) $H_1 : \mu < \mu_0$

Setting of alternative hypothesis is very important. Since it enable us to decide whether to use single tail (right or left) or two tail test.

Types of errors in hypothesis testing:

Type I error : The error of rejecting Ho when it is true.

Type II error : The error of accepting when Ho it is false.

$P\{$Reject Ho when it is true$\} = P\{$Reject Ho$/$Ho$\}$
$= \alpha$

$P\{$Accept H_0 when it is wrong$\} = P\{$Accept Ho$/H_1\}$
$= \beta$

then α, β are called the size of type-I error and type II error, respectively

		Decision from theory	
		Reject H_0	Accept H_0
True Statement	H_0 true	Type I error (wrong	✓ (Correct)
	H_0 false	✓ (correct)	Type II error (wrong

Level of Significance:

The probability of type I error is known as level of significance and It is denoted by α. The level of significance is usually employed in testing of hypothesis are 5% and 1%.

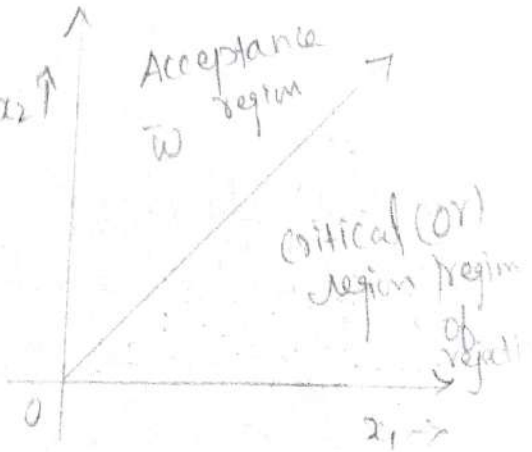
Critical Region:

A region in the sample space S which amounts to rejection of H_0 is termed as critical region or Region of rejection. It w is the critical region and it $t = t(x_1, x_2 \dots x_n)$ is the value of the statistic based on a random sample of size n, then

$$P(t \in W / H_0) = \alpha$$

$$P(t \in \overline{W} / H_1) = \beta$$

where $\overline{w}$, the complementary set of w, is called the acceptence region.



One tailed test :

In any test, the critical region is represented by a portion of a area under the probability curve of the sampling distribution of the test statistic.

A test of any statistical hypothesis where the alternative hypothesis is one tailed (right or left) is called a one tailed.

Ex → A test for testing the mean of a population. $H_0 : \mu = \mu_0$

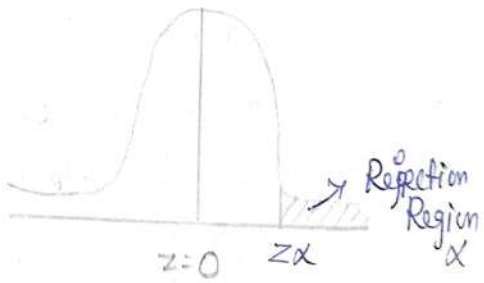
Against the alternative hypothesis

$H_1 : \mu > \mu_0$ (right tailed)

$H_1 : \mu < \mu_0$ (left tailed)

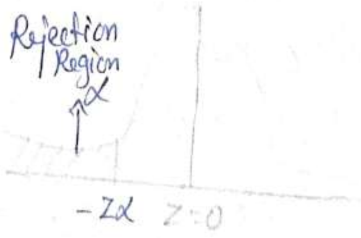
Right tailed test

$(P(Z > Z_\alpha) = \alpha$



Left tailed test

$P(Z < -Z\alpha) = \alpha$



Two-tailed test :

A test of statistical hypothesis where the alternative hypothesis is two tailed, such that

$H_0 : \mu = \mu_0$ against alternative hypothesis

$H_1 : \mu \neq \mu_0$ $(\mu > \mu_0, \mu < \mu_0)$

is Known as two tailed test and in such a case the Critical region is by portion of the area lying in both the tails of the probability curve of the test Statistics

Two tailed test

$$\boxed{\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}}$$

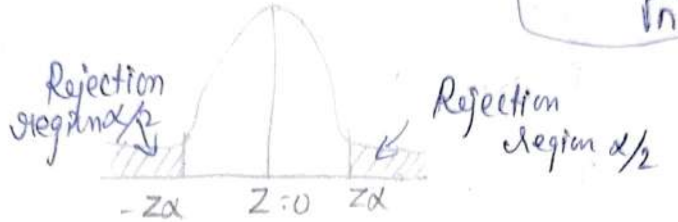


Table value of Z

Critical Values	Level of significance		
	1%	5%	10%
Two tailed test	$Z_\alpha = 2.58$	$Z_\alpha = 1.96$	$Z_\alpha = 1.645$
One tailed test			
Right	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$

Test of Significance for large sample:-

When the sample of size is larger that type of sample is called large sample (i.e) $n \geq 30$ No. of. Sample observation is greater 30.

Types: 1) Test of significance of Single mean or Specified mean.

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$\bar{x}$ → sample mean

μ → population mean

σ → S.D

n → sample of size.

2) Testing of significance for difference of mean (or) double means:

Let $\bar{x}_1, \bar{x}_2$ be the mean of a sample of size n_1, n_2 from a population with mean μ_1, μ_2 and variance σ_1^2, σ_2^2

Thus since sample size are large

$$\bar{x}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$$

$$\bar{x}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Also $\bar{x}_1 - \bar{x}_2$ being the difference of two normal variates. The value of standard normal variate corresponding to $\bar{x}_1 - \bar{x}_2$ is given by

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

Under the null hypothesis $H_0 : \mu = \mu_0$

(i.e) There is no significant difference between the sample means. We get

$$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2)$$
$$= \mu_1 - \mu_2 = 0$$
$$V(\bar{x}_1 - \bar{x}_2) = V(\bar{x}_1) + V(\bar{x}_2)$$
$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

the covariance term vanishes, since the sample means $\bar{x}_1$ and $\bar{x}_2$ are identified.

Thus under $H_0 : \mu_1 = \mu_2$ the test statistic becomes

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Note: If $\sigma_1^2 = \sigma_2^2 = \sigma^2$

(i.e) If the samples have been drawn from the population with common S.D σ

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1)$$

3) Test of significance of single proportion or Specified proportion.

If X is the no. of success in n - indentified trials with Constant probability P of success for each trial

$E(X) = nP$ and $V(X) = nPQ$, where $Q = 1-P$ is the probability failure.

It has been proved that for large n, the binomial distribution tends to normal dist. Hence large n, $X \sim N(nP, nPQ)$

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} \qquad Z = \frac{X - nP}{\sqrt{nPQ}} \sim N(0,1)$$

and we can apply the normal test.

Remarks:

1) In a Sample Size n, let X be the no. of persons possessing the given attribute.

observed proportion of successes $= \frac{X}{n} = p$

$$\therefore E(p) = E(X/n) = 1/n \, E(X)$$

$$= 1/\cancel{n} \, (\cancel{n}P)$$

$$E(p) = 1$$

Thus the sample proportion 'p' given an unbiased estimate of the population proportion $\hat{P}$

$$V(p) = V(X/n) = 1/n^2 \, V(X)$$

$$= \frac{1}{n^2} nPQ \qquad \therefore V(p) = \frac{PQ}{n}$$

$$SE(p) = \sqrt{\frac{PQ}{n}}$$

Since X and consequently $\frac{X}{n}$ is asymptotically normal for large n, the normal test for the proportion of successes becomes

$$Z = \frac{p - E(p)}{SE(p)} = \frac{p - P}{\sqrt{PQ/n}} \sim N(0,1)$$

2) If we have sampling from a finite population of size N, then

$$SE(p) = \sqrt{\left(\frac{N-n}{N-1}\right)\frac{PQ}{n}}$$

3) Since the probable limits for a normal variate X are $E(X) \pm 3\sqrt{V(X)}$, the probable limits for the observed proportion of successes are

$$E(p) \pm 3SE \qquad P \pm 3\sqrt{PQ/n}$$

If P is not known then taking p (the sample proportion) as an estimate of P, the probable limits for the proportion in the population are $p \pm 3\sqrt{\frac{pq}{n}}$

However, the limits for P at level of significance α are given by $p \pm Z_\alpha \sqrt{\frac{pq}{n}}$

Where Z_α is the significant value of Z loss α

In particular 95% confidence limits for population are given by $p \pm 1.96\sqrt{\frac{pq}{n}}$

99% confidence limits for population are given by $p \pm 2.58\sqrt{\frac{pq}{n}}$

A) Test of significance for difference of proportion:

Let X_1, X_2 be the no. of persons possessing the given attribute A in random samples of sizes n_1 and n_2 from the two population respectively.

Then sample proportions are given by

$$p_1 = \frac{X_1}{n_1} \qquad p_2 = \frac{X_2}{n_2}$$

If P_1 and P_2 are the population proportions;

$$E(A) = E\left(\frac{X_1}{n}\right) \cdot E(p_2) = E\left(\frac{X_2}{n_2}\right)$$

$$= \frac{1}{n} E(X_1) \qquad = \frac{1}{n_2}(E X_2)$$

$$= \frac{1}{n_1} n_1 P_1 \qquad = \frac{1}{n_2} n_2 P_2$$

$$= P_1 \qquad = P_2$$

$$V(p_1) = \frac{P_1 Q_1}{n_1} \text{ and } V(p_2) = \frac{P_2 Q_2}{n_2}$$

Since for large samples p_1 and p_2 are asymptotically normally distributed, $(p_1 - p_2)$ is also normally distributed. Then the standard variable corresponding to the difference $(p_1 - p_2)$ is given by

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{V(p_1 - p_2)}} \sim N(0,1)$$

Under Null hyp : $H_0 : P_1 = P_2$

(i.e) There is no significant difference between the sample proportions.

$$E(p_1 - p_2) = E(p_1) - E(p_2)$$

$$= P_1 - P_2$$

$$= 0$$

Also $V(p_1 - p_2) = V(p_1) + V(p_2)$

The co-Vairance term $cov(p_1, p_2)$ Vanishes, since sample proportions are independent

$$\therefore V(p_1 - p_2) = \frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}$$

$$= PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Since under:

$$H_0 : P_1 = P_2 = P$$
$$Q_1 = Q_2 = Q$$

Under $H_0 : P_1 = P_2$

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

In general, we do not have any information as to the proportion of A's in the population from which the sample have been taken.

Under $H_0 : P_1 = P_2 = P$ (say)

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$$

The estimate is unbiased, since

$$E(\hat{P}) = \frac{1}{n_1 + n_2} E(n_1 p_1 + n_2 p_2)$$

$$= \frac{1}{n_1 + n_2} (n_1 E(p_1) + n_2 E(p_2))$$

$$= \frac{1}{n_1 + n_2} (n_1 P_1 + n_2 P_2)$$

$$= P_1 + P_2 \qquad \therefore P_1 = P_2 = P$$

$$= P$$

Small
Sample Samples:

When the no. of sample size is below 30 (i.e) $n < 30$ that type of sample is called small sample

(x) Z tests are used when we have large sample sizes ($n > 30$), whereas

T-tests are most helpful with a smaller sample size ($n < 30$)

Right & Left tailed test

α	Z
0.10	± 1.283
0.05	± 1.645
0.025	± 1.960
0.010	± 2.326
0.005	± 2.576
0.001	± 3.090
0.0001	± 3.719

Two-tailed test

α	Z
0.20	1.282
0.10	1.645
0.05	1.960
0.010	2.576
0.001	3.291
0.0001	3.819

$$Z = \frac{\bar{p}_1 - \bar{p}_2}{S\bar{p}_1 - \bar{p}_2}$$

$$\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2}$$

Two tailed test:

Two Critical Value CV_1 & CV_2 One of each tail of the Sampling distribution is Computed as

Known Sigma (σ)	Unknown Sigma (σ)
⇒ Normal Population: Any sample size 'n' ⇒ Any population: Large Sample size 'n' $CV_1 = \mu_0 - Z_{\frac{\alpha}{2}} \sigma_{\bar{x}}$ $CV_2 = \mu_0 + Z_{\frac{\alpha}{2}} \sigma_{\bar{x}}$ Where $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	⇒ Any population: Large Sample size 'n' $CV_1 = \mu_0 - Z_{\frac{\alpha}{2}} S_{\bar{x}}$ $CV_2 = \mu_0 + Z_{\frac{\alpha}{2}} S_{\bar{x}}$ $S_{\bar{x}} = \frac{S}{\sqrt{n}}$

Decision rule:

Rejected H_0 When $\bar{x} \leq CV_1$ or $\bar{x} \geq CV_2$

Accept H_0 when $CV_1 < \bar{x} < CV_2$

Left tailed test:

The Critical Value for left tail of the Sampling distribution is Compared as

Known Sigma (σ)	Unknown Sigma (σ)
Normal Population: Any Sample Size 'n' Any population: Large Sample Size 'n'	Any population: Large Sample size 'n'

$CV = \mu_0 - Z_\alpha \sigma_{\bar{x}}$ $\qquad$ $CV = \mu_0 - Z_\alpha S_{\bar{x}}$

Decision rule:

Reject H_0 where $\bar{x} \leq CV$

Accept H_0 where $\bar{x} > CV$

Right tailed test:

The critical value for right tail of the sampling distribution is compared as

Known sigma (σ)	Unknown sigma (σ)
Normal population: Any sample size 'n'	Any population: Large sample size 'n'
Any population Large sample size 'n'	
$CV = \mu_0 + Z_\alpha \sigma_{\bar{x}}$	$CV = \mu_0 + Z_\alpha S_{\bar{x}}$

Decision rule:

Reject H_0 where $\bar{x} \geq CV$

Accept H_0 where $\bar{x} < CV$

1. Individual filling of income tax returns prior to 30 June had an average refund of 1200 considered the population of last minute fillers who fill the returns during the last week of june. For a random sample of 400 individuals who filled return between 25 and 30 june the sample mean was refund Rs 1054 and the sample S.D was Rs. 1600. Using 5% level of significance, test the belief that the individuals who ~~weight~~ wait until the last week of june to fill the returns than the early the fillers.

Sol $H_0: \mu \geq 1200$ and $H_1: \mu < 1200$

$n = 400$ ~~σ~~ $S = 1600$ $\bar{x} = 1054$ $\alpha = 5\%$

$$Z = \frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}} = \frac{1054-1200}{\frac{1600}{\sqrt{400}}} = -\frac{146}{80}$$

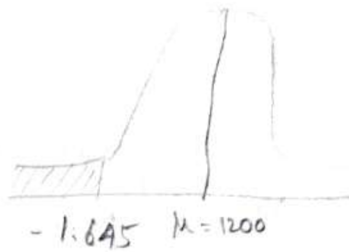
$$= -1.825$$

$$CV = \mu_0 - Z_\alpha \sigma_{\bar{x}}$$

$$= 1200 - 1.645\left(\frac{1600}{\sqrt{400}}\right)$$

$$= 1200 - 131.6$$

$$= 1068.4$$



Since $\bar{x}\,(=1054) < CV\,(=1068.4)$, the null hypothesis H_0 is rejected.

2. A packaging device is set to fill detergent power packs with a mean weight of 5 kg with a SD 0.21 kg. The weight of packs assumed to be normally distributed the weights of packs is known to ~~drifted~~ drified upwards over a period time due to machine, which is not tolerable. A random sample of 100 pack is taken and weight. This sample as a mean weight of 5.03 kg. Can be conclude ie the mean weight produced by the machine has increased? Use a 5% level of significance

$n = 100 \quad \bar{x} = 5.03 \quad s = 0.21$

$\mu = 5$

$$Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$= \frac{5.03 - 5}{\frac{0.21}{\sqrt{100}}}$$

$$= \frac{0.03}{0.021} = 1.4285$$

$$CV = \mu_0 + Z_\alpha \sigma_{\bar{x}}$$

$$= 5 + 1.645\left(\frac{0.21}{\sqrt{100}}\right)$$

$\mu = 5 \quad 1.428$

$$= 5 + 0.0345$$

$$= 5.0345$$

$\bar{x} > CV$ H_0 is Rejected

$\bar{x} < CV$ H_0 is accepted

$\bar{x} < CV$

$5.03 < 5.034$ ∴ H_0 is accepted.

3. The mean life time of sample 400 flourescent produced by a company is found to be 1600 hours with a SD of 150 hours test the hypothesis that the mean life time of the bulbs produced in general is higher than the mean life of 1570 hours at $\alpha =$ 0.01 level of significance. $s/\sqrt{n}$

$n = 400 \quad \bar{x} = 1600 \quad \mu = 1570 \quad s = 150$

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{1600 - 1570}{\frac{150}{\sqrt{400}}} = \frac{30}{7.5} = 4$$

$CV = \mu_0 + Z_\alpha S_{\bar{x}}$

$= 1570 + 2.33(7.5)$

$= 1570 + 17.475$

$= 1587.475$

$\bar{x} > CV$

$1600 > 1587.4$

H_0 is rejected.

$\mu = 1570 \quad 4$

7. A continuous manufacturing process of steel rods is said to be in state of control and produces acceptable rods if the mean diameter of all rods produced is 2 inches. Although the process SD exhibits stability overtime with SD $\sigma = 0.01$ inch. The process mean may vary due to operator error or problems of process adjustment. Periodically, random samples of 100 rods are selected to determine whether the process is producing acceptable rods. If the result of a test indicates that the process is out of control, it is stopped and the source of trouble is sought. Otherwise, it is allowed to continue operating. A random sample of 100 rods is selected resulting in a mean of 2.1 inches. Test the hypothesis to determine whether the process be continued.

$H_0 : \mu = 2$ inches (continue process)

$H_1 : \mu \neq 2$ inches (stop the process)

$n = 100 \quad \bar{x} = 2.1$

$\sigma = 0.01 \quad \mu = 2 \quad \alpha = 0.01$

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2.1 - 2}{\frac{0.01}{\sqrt{100}}} = \frac{0.1}{0.001} = 100$$

Since Z = 100 Value is more than its Critical Value $Z_{\alpha/2} = 2.58$ at $\alpha = 0.01$ The H_0 is rejected Thus stop the process in order to determine the Source of trouble.

$$CV_1 = \mu_0 - Z_{\frac{\alpha}{2}} \sigma_{\bar{x}}$$

$$= \mu_0 - Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = 2 - 2.58 \times \left(\frac{0.01}{\sqrt{100}}\right)$$

$$= 2 - 0.003 = 1.997$$

$$CV_2 = \mu_0 + Z_{\frac{\alpha}{2}} \sigma_{\bar{x}}$$

$$= 2 + 2.58 \times \left(\frac{0.01}{\sqrt{100}}\right)$$

$$= 2 + 0.003 = 2.003$$

Since $\bar{x} (= 2.1) \geq CV_2 (= 2.003)$, the null hypothesis is rejected.

5. An Ambulance Service claims that it takes, on the average 8.9 minutes to reach its destination in emergency calls. To check on this claim, the agency which licenses ambulance Services has then timed on 50 emergency calls, getting a mean of 9.3 minutes with a SD of 1.8 minutes. Does this constitute evidence that the figure claimed is too low at the 1 percent significance level?

$$H_0 : \mu = 8.9, \quad H_1 : \mu \neq 8.9$$

$n = 50 \quad \bar{x} = 9.3 \quad s = 1.8$

$$Z = \frac{\bar{x} - \mu}{S_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{9.3 - 8.9}{\frac{1.8}{\sqrt{50}}} = \frac{0.4}{0.254} = 1.574$$

Since Z = 1.574 is less than its Critical value $Z_{\alpha/2} = \pm 2.58$ $\alpha = 0.01$. H_0 is accepted.

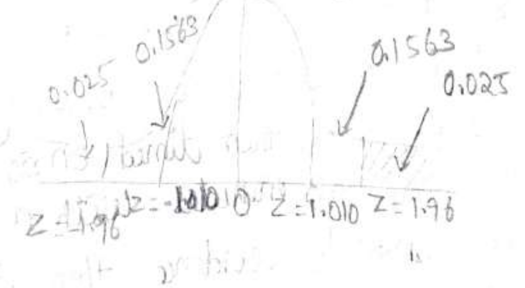
6. An Auto Company decided to introduce a new Six cylinder Car whose mean petrol Consumption is claimed to be lower than that of the existing auto engine. It was found that the mean petrol Consumption for 50 Cars was 10 km per litre with a SD of 3.5 km per litre. Test for the Company at 5% level of significance, the claim that in the new petrol consumption is 9.5 km per litre on the average.

$H_0: \mu = 9.5$ Km/litre & $H_1: \mu \neq 9.5$ km/litre

$\bar{x} = 10$ $n = 50$ $s = 3.5$ and $Z_{\alpha/2} = 1.96$ $\alpha = 0.05$

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{10 - 9.5}{\frac{3.5}{\sqrt{50}}} = 1.010$$

Since $Z = 1.010$ is less than its critical value $Z_{\alpha/2} = 1.96$ at $\alpha = 0.05$ level of significance. Ho is accepted



The p-value is the area to the right as well as left of the calculated value of z-test statistic (two-tailed) since $Z_{cal} = 1.010$, then the area to its right is $0.5 - 0.3437 = 0.1563$

Since it is two tailed $2(0.1563) = 0.3126$.

Since $0.3126 > 0.05$ Ho is accepted

7 (8) A firm believes that the tyres produced by process A on an average last longer than tyres produced by process B. To test this belief, random samples of tyres produced by the two processes were tested and the results are:

Process	Sample size	Average	S D
A	50	22400	1000
B	50	21800	1000

Is there evidence at a 5% level of significance that the firm is correct in its belief?

Sol $H_0 : \mu_1 = \mu_2$ (or) $\mu_1 = \mu_2 = 0$

$H_1 : \mu_1 \neq \mu_2$

$\bar{x}_1 = 22400 \quad \bar{x}_2 = 21800 \quad \sigma_1 = \sigma_2 = 1000\text{Km}$

$n_1 = n_2 = 50.$

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - x_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$= \frac{22400 - 21800}{\sqrt{\frac{(1000)^2}{50} + \frac{(1000)^2}{50}}} = \frac{600}{\sqrt{20{,}000 + 20{,}000}} = \frac{600}{200} = 3$$

Since $Z = 3$ is more than its critical value $Z_{\frac{\alpha}{2}} = \pm 1.645$ at $\alpha = 0.05$ level of significance. Ho is rejected. Hence we can conclude that the tyres produced by process A last longer than those produced by process B.

P-Value approach :-

$$\text{p-value} = p(Z > 3.00) + p(Z < -3.00) = 2P(Z > 3)$$

$$= 2(0.500 - 0.49[illegible])$$

$$= 0.0026$$

Since p-value of 0.0026 is less than [illegible] significance level $\alpha = 0.05$ [illegible]

8) An experiment was conducted to compare the mean time in days required to recover from a common cold for person given daily dose of 4 mg of vitamin C versus those who were not given a vitamin supplement. Suppose that 35 adults were randomly selected for each treatment category and that the mean recovery times and SD for the two grps.

	Vitamin C	No Vitamin Supplement
Sample size	35	35
Sample mean	5.8	6.9
Sample SD	1.2	2.9

Test the hypothesis that the use of vitamin C reduces the mean time required to recover from a common cold and its complications, at the level of significance $\alpha = 0.05$.

$H_0 : (\mu_1 - \mu_2) \leq 0 \quad H_1 : (\mu_1 - \mu) > 0$

$n_1 = 35 \quad \bar{x}_1 = 5.8 \quad s_1 = 1.2 \quad n_2 = 35 \quad \bar{x}_2 = 6.9 \quad s_2 = 2.9$

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{5.8 - 6.9}{\sqrt{\frac{(1.2)^2}{35} + \frac{(2.9)^2}{35}}} = \frac{-1.1}{\sqrt{0.041 + 0.240}} = \frac{-1.1}{0.530}$$

$Z_{cal} < Z_\alpha$

$-2.0754 < 1.645 \qquad = -2.605.$

Using a one-tailed test with significance level $\alpha = 0.05$, the critical value is $Z\alpha = 1.645$. Since $Z < Z\alpha$ ($= 1.645$) the H_0 is rejected. Hence we can conclude that the use of vitamin C does not reduce the mean time required to recover from the common cold.

9 The Educational testing Service Conducted a study to investigate difference between the scores of female and male students on the mathematics Appitude test. The study identified a random sample of 562 female and 852 male students who had achieved the same high score on the mathematics portion of the test. That is, the female and male students viewed as having similar high ability in mathematics. The verbal scores for the 2 samples are given below.

	Female	Male
Sample mean	547	525
Sample S.D	83	78

Do the data support the conclusion that given population of female and male students with similiar high ability in mathematics, the female students will have a significantly high verbal ability? Test at $\alpha = 0.05$ significance level. What is you conclusion?

$$H_0: (\mu_1 - \mu_2) \geq 0 \quad H_1: (\mu_1 - \mu_2) < 0$$

$n_1 = 562$ $\bar{x}_1 = 547$ $S_1 = 83$ $n_2 = 852$ $\bar{x}_2 = 525$

$S_2 = 78$ $\alpha = 0.05$

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{547 - 525}{\sqrt{\frac{(83)^2}{562} + \frac{(78)^2}{852}}} = \frac{22}{\sqrt{12.258 + 7.140}}$$

$$= \frac{22}{\sqrt{19.398}} = \frac{22}{4.404} = 4.995$$

Using a one tailed test with $\alpha = 0.05$ significance level, the critical value $Z_\alpha = \pm 1.645$.

Since $Z = 4.995$ is more than the critical value $Z_\alpha = 1.645$ H_0 is rejected. Hence we conclude that there is no sufficient evidence to declare that difference between verbal ability of female & male students is significant.

10. In a Sample of 1000, the mean is 17.5 and the SD is 2.5. In another Sample of 800, the mean is 18 and SD is 2.7. Assuming that the Samples are Independent discuss whether the two Samples Could have Come from a population which have the Same SD.

$$H_0: \sigma_1 = \sigma_2 \text{ and } H_1: \sigma_1 \neq \sigma_2$$

Given

$$\sigma_1 = 2.5 \quad n_1 = 1000 \quad \sigma_2 = 2.7 \quad n_2 = 800$$

Standard error

$$\sigma_{\sigma_1 - \sigma_2} = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$$

$$= \sqrt{\frac{(2.5)^2}{2(1000)} + \frac{(2.7)^2}{2(800)}} = \sqrt{\frac{(2.5)^2}{2000} + \frac{(2.7)^2}{1600}}$$

$$= \sqrt{\frac{6.25}{2000} + \frac{7.29}{1600}} = 0.0876$$

$$Z = \frac{\sigma_1 - \sigma_2}{\sigma_{\sigma_1 - \sigma_2}} = \frac{2.7 - 2.5}{0.0876} = \frac{0.2}{0.0876} = 2.283$$

Since $Z = 2.283$ is more than its critical value $Z = 1.96$ at $\alpha = 5\%$ Ho is rejected. Hence we conclude that the two Samples have not come from a population which has the Same SD.

11. The Mean production of wheat from a sample of 100 field is 200 lbs per acre with a SD of 10lbs. Another Sample of 150 fields gives the mean at 220 lbs per acre with a SD of 12lbs. Assuming the SD of the universe as 11 lbs, find at 1% level of significance, whether the two results are consistent.

$H_0: \sigma_1 = \sigma_2$ and $H_1: \sigma_1 \neq \sigma_2$

$\sigma_1 = \sigma_2 = 11 \quad n_1 = 100 \quad n_2 = 150$

$$\sigma_{\sigma_1 - \sigma_2} = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}} = \sqrt{\frac{11}{2(100)} + \frac{11}{2(150)}} = \sqrt{\frac{11}{200} + \frac{11}{300}}$$

$0.055 \quad 0.0367$

$$= \sqrt{\frac{\sigma^2}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{\frac{(11)^2}{2}\left(\frac{1}{100} + \frac{1}{150}\right)}$$

$$= 1.004$$

$$Z = \frac{\sigma_1 - \sigma_2}{\sigma_{\sigma_1 - \sigma_2}} = \frac{10 - 12}{1.004} = \frac{-2}{1.004} = -1.992$$

Since $Z = -1.992$ is more than its critical value of $Z = -2.58$ at $\alpha = 0.01$ Ho is accepted.

Hence we conclude that the two results are likely to be consistent.

Hypothesis testing for single population proportion

$$\bar{p} = \frac{\text{Number of successes in the Sample}}{\text{Sample size}} = \frac{x}{n}$$

The Value of this Statistics is compared with a hypothesized population proportion p_0 so as to arrive at a conclusion about the hypothesis.

The three forms of null hypothesis and alternative hypothesis pertaining to the hypothesized population proportion p are as follows.

Null hypothesis	Alternative hypothesis.
$H_0 : p = p_0$	$H_1 : p \neq p_0$ (Two tailed test)
$H_0 : p \geq p_0$	$H_1 : p < p_0$ (left tailed test)
$H_0 : p \leq p_0$	$H_1 : p > p_0$ (Right tailed test)

To Conduct a test of a hypothesis, it is assumed that the Sampling distribution of a proportion follows a Standardized normal distribution.

$\bar{p} \rightarrow$ Sample proportion

$\sigma_{\bar{p}} \rightarrow$ Standard deviation.

Test Statistic $$z = \frac{\bar{p} - p_0}{\sigma_{\bar{p}}} = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

$$= \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$

Decision rule: Reject H_0 when.

One tailed test	two-tailed test
• $Z_{cal} > Z_\alpha$ or $Z_{cal} < -Z_\alpha$ when $H_1: P < P_0$	• $Z_{cal} > Z_{\frac{\alpha}{2}}$ or $Z_{cal} < -Z_{\frac{\alpha}{2}}$
• p-value $< \alpha$	

Hypothesis testing for difference between two population propotions:-

· Two independent populations each having proportion and SD of an attribute be as follows.

Population	proportion	SD
1	P_1	$\sigma_{\bar{P}_1}$
2	P_2	$\sigma_{\bar{P}_2}$

The hypothesis testing concepts developed in the previous section can be extended to test whether there is any difference between the proportions of these populations.

The null hypothesis that there is no difference between two population proportions is stated as.

$$H_0: P_1 = P_2 \text{ or } P_1 - P_2 = 0 \text{ and } H_1: P_1 \neq P_2$$

The sampling distribution of difference in sample proportions $\bar{P}_1 - \bar{P}_2$ is based on the assumption that the difference between two population proportions. $P_1 - P_2$ is normally distributed. The SD (or) error of sampling distribution of $P_1 - P_2$ is given by

$$\sigma_{\bar{P}_1 - \bar{P}_2} = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

$$q_1 = 1 - P_1 \quad q_2 = 1 - P_2$$

Where the difference $\bar{P}_1 - \bar{P}_2$ between sample proportions of two independent simple random samples is the point estimator of the difference between two population proportions. $E(\bar{P}_1 - \bar{P}_2) = P_1 - P_2$

$$= \frac{(\bar{P}_1 - \bar{P}_2) - (P_1 - P_2)}{\sigma_{\bar{P}_1 - \bar{P}_2}} = \frac{\bar{P}_1 - \bar{P}_2}{\sigma_{\bar{P}_1 - \bar{P}_2}}$$

Pooled estimate $\bar{p} = \dfrac{n_1 \bar{P}_1 + n_2 \bar{P}_2}{n_1 + n_2}$

The Z-test statistic is then restated as

$$Z = \frac{\bar{P}_1 - \bar{P}_2}{S_{\bar{P}_1 - \bar{P}_2}} \; ; \; S_{\bar{P}_1 - \bar{P}_2} = \sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

1. An auditor claims that 10% of customer's ledger A/c are carrying mistakes of posting and balancing. A random sample of 600 was taken to test the accuracy of posting and balancing and 45 mistakes were found. Are these sample results consistent with the claim of the audit? Use 5% level of significance.

cap P (2nd)

$H_0 : P = 0.10 \quad H_1 : P \neq 0.10$ (two-tailed test)

(1st)

small p $\quad \bar{p} = \frac{45}{600} = 0.075 \quad n = 600 \quad \alpha = 5\%$

$$Z = \frac{\bar{p} - P_0}{\sigma_{\bar{p}}} = \frac{0.075 - 0.10}{\sqrt{\frac{0.10 \times 0.90}{600}}} = \frac{-0.025}{0.0122}$$

$\sigma_{\bar{p}} \rightarrow \sqrt{\frac{PQ}{n}}$

$= -2.049. \quad -2.0 < -1.96$

3

Cal

Since $Z = (-2.049)$ is less than its critical value $Z_\alpha (= -1.96)$ at $\alpha = 0.05$. H_0 is rejected. Hence, we conclude that the claim of the auditor is not valid.

2. A manufacturer claims that at least 95% of the equipments which he supplied to a factory conformed to the specification. An examination of the sample of 200 pieces of equipment revealed that 18 were faulty. Test the claim of the manufacturer.

$H_0: P \geq 0.95 \quad H_1: P < 0.95$ (left-tailed test)

$\bar{P}$ = percent of pieces conforming the specification

$= 1 - (18/100) = 0.91$

$n = 200 \quad \alpha = 0.05$

$$Z = \frac{\bar{P} - P_0}{\sigma_{\bar{P}}} = \frac{0.91 - 0.95}{\sqrt{\frac{0.95 \times 0.05}{200}}} = \frac{-0.04}{0.015}$$

$$= -2.67$$

Since $Z = -2.67$ is less than its critical value ($Z_\alpha = -1.96$) at $\alpha = 0.05$. H_0 is rejected. Hence we conclude that the proportion of equipments conforming to specification is not 95%.

3 A Company is considering two different television advertisement for promotion of a new product. Management believes that advertisement A is more effective than advertisement B. Two test market areas with virtually identical consumer characteristics are selected advertisement A is used in one area and advertisement B in the other area. In a random sample of 60 customers who saw advertisement A, 18 had tried the product. In a random sample of 100 customers who saw advertisement B, 22 had tried the product. Does this indicate that advertisement A is more effective than advertisement B, if a 5% level of significance is used?

$H_0 : P_1 = P_2$ $\quad H_1 : P_1 > P_2$ (Right tailed test)

$n_1 = 60$ $\quad \bar{P}_1 = 18/60 = 0.30$ $\quad n_2 = 100$ $\quad \bar{P}_2 = 22/100 = 0.22$

$\alpha = 0.05$

$$Z = \frac{(\bar{P}_1 - \bar{P}_2) - (P_1 - P_2)}{S_{\bar{P}_1 - \bar{P}_2}} = \frac{\bar{P}_1 - \bar{P}_2}{S_{\bar{P}_1 - \bar{P}_2}} \; ; \; P_1 = P_2$$

(2) $S_{\bar{P}_1 - \bar{P}_2} = \sqrt{\bar{P}(1-\bar{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$; $q = 1 - p$

$$= \sqrt{0.25 \times 0.75 \left(\frac{1}{60} + \frac{1}{100}\right)}$$

$$= \sqrt{0.1875 \left(\frac{160}{600}\right)} = 0.0707$$

(1) $$\bar{P} = \frac{n_1 \bar{P}_1 + n_2 \bar{P}_2}{n_1 + n_2} = \frac{60(18/60) + 100(22/100)}{60 + 100}$$

$$= \frac{18 + 22}{160} = \frac{40}{160} = 0.25$$

$$Z = \frac{0.30 - 0.22}{0.0707} = \frac{0.08}{0.0707} = 1.131$$

Since $Z = 1.131$ is less than its critical value $Z_\alpha = 1.645$ at $\alpha = 0.05$. H_0 is accepted.

Hence we conclude that there is no significance difference in the effectiveness of the two advertisements.

A) In a Simple random sample of 600 men taken from a big city, 400 are found to be smokers. In another Simple random Sample of 900 men taken from another City 450 are Smokers. Do the data indicate that there is a significant difference in the habit of Smoking in the two Cities?

$$H_0 : P_1 = P_2 \qquad H_1 : P_1 \neq P_2 \text{ (Two-tailed test)}$$

$$n_1 = 600 \quad \bar{P}_1 = \frac{400}{600} = 0.667 \quad n_2 = 900 \quad \bar{P}_2 = \frac{450}{900} = 0.50$$

$$\alpha = 0.05$$

$\frac{0.21}{60} + \frac{0.1716}{100}$

$0.0035 \quad 0.031716$

$$Z = \frac{(\bar{P}_1 - \bar{P}_2) - (P_1 - P_2)}{S_{\bar{P}_1 - \bar{P}_2}} = \frac{\bar{P}_1 - \bar{P}_2}{S_{\bar{P}_1 - \bar{P}_2}} \qquad P_1 = P_2$$

$0.036636 (0.0167 + 0.01)$

$$S_{\bar{P}_1 - \bar{P}_2} = \sqrt{\bar{P}(1-\bar{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \qquad q = (1-p)$$

$$= \sqrt{0.567 \times 0.433 \left(\frac{1}{600} + \frac{1}{900}\right)}$$

$$= \sqrt{0.245 (0.002)}$$

$$= 0.026$$

$$\bar{P} = \frac{n_1 \bar{P}_1 + n_2 \bar{P}_2}{n_1 + n_2} = \frac{600(400/600) + 900(450/900)}{600 + 900}$$

$$= \frac{400 + 450}{1500} = \frac{850}{1500}$$

$$= 0.567$$

$$Z = \frac{0.667 - 0.500}{0.026} = \frac{0.167}{0.026} = 6.423$$

Since $Z = 6.423$ is greater than its Critical Value $Z_{\frac{\alpha}{2}} = 2.58$ at $\frac{\alpha}{2} = 0.025$. H_0 is rejected. Hence we conclude that there is a significant difference in the habit of Smoking in two cities.

Hypothesis testing for a binomial proportion

The Sampling of traits or attributes is considered.

$$Z = \frac{\text{Sample estimate} - \text{Expected value}}{\text{Standard error of estimate}}$$

$$= \frac{x - np}{\sqrt{npq}}$$

The Z-test Statistic for determining the magnitude of the difference between the number of successes in a Sample and the hypothesized (expected) number of Successes in the population.

D: N-1

Small Samples:

When the no. of. Small size below ($n < 30$) that type of sample is called Small Sample. In these, Small Sample we have to study 3 different tests.

* T-test
* F-test
* chi-square test] next sem

T-test

The t-statistic is defined by $t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$

Where S is the estimation of SD of population

$$S^2 = \frac{n}{n-1} s^2$$

Where s is the SD of Sample, SD of the population i.e $S.Error = \frac{S}{\sqrt{n}}$

Uses of t-test :-

* It is used to test whether specific value is the population mean when the given sample is small sample and the population deviation is not known.

* It is also used to test the significance difference between means of two population based on two sample of size n_1 and n_2. When the SD of the population are not known and also

Sample are drawn independent.

* It is also used to test the significant difference between the mean and (obj) pair observation

Types of Small Samples :

* Test for Specified mean
* Test for double mean
* Paired test

Test for specified Mean [procedure for Testing]

Null hypothesis $H_0 : \mu = \mu_0$

Alternative hypothesis $H_1 : \mu \neq \mu_0$ (Two-tailed)

$H_1 : \mu > \mu_0$ (Right tailed)

$H_1 : \mu < \mu_0$ (Left tailed)

Level of Significance :

5% or 1%.

Test Statistics :-

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$\bar{x}$ - Sample mean

μ - population mean

$s \rightarrow$ S.D of population

$n \rightarrow$ Sample Size

Degrees of freedom :

$\rightarrow$ gama

df $\leftarrow \gamma = n-1$

Inference :

If the calculated value of

"t" less than table Value of "t."

Ho is accepted

If the Calculated value of 't' greater than table Value of 't'. Ho is rejected.

2.

1. The Average breaking Strength of Steel rods is Specified to be 18.5 thousand kg. for this Sample of 14 rods was tested. The mean and SD Obtained are 17.85 and 1.955 respectively. Test the Significance of deviation.

$\boxed{\frac{\alpha}{2}}$

Given: ② $\mu = 18.5$ $n = 14$ $\bar{x} = 17.85$

$S.D = 1.955$ $\alpha = 5\%$

$= 0.05$

$\frac{\alpha}{2} = \frac{0.05}{2}$

$= 0.025$

① Null hypothesis $H_0 : \mu = \mu_0$

$H_0 : \mu = 18.5$

$H_1 : \mu \neq \mu_0$ $H_1 : \mu \neq 18.5$

$df = n-1$

$= 14-1$

$= 13$

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} = \frac{17.85 - 18.5}{\frac{1.955}{\sqrt{14}}}$$

$$= \frac{-0.65}{\frac{1.955}{3.741}}$$

$$= \frac{-0.65}{0.5228}$$

$= -1.245$ (calculated Value)

$t_{\frac{\alpha}{2}} = -2.160$ $t_{cal} = -1.245$ Value is more than its critical value. $t_{\frac{\alpha}{2}} = -2.160$ at $\alpha = 0.05$ (0.025) and df = 13. H_0 is accepted.

2. An automobile tyre manufacturer claims that the average life of a particular grade of tyre is more than 20,000 km when used under normal conditions. A random sample of 16 tyres was tested and a mean and SD of 22000 Km and 5000 Km respectively were computed. Assuming the life of the tyres in Km to be approximately normally distributed, decide whether the manufacturer's claim is valid.

$H_0 : \mu \geq 20{,}000$

$H_1 : \mu < 20{,}000$

$n = 16$, $\bar{x} = 22{,}000$ $s = 5000$

$df = n-1$

$= 16-1 = 15$

$\alpha = 5\% = 0.05$

$t_{\alpha} = 15 \rightarrow 0.05 \rightarrow$ table Value

$t_{\alpha} = 1.753$

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{22000 - 20{,}000}{\frac{5000}{\sqrt{16}}}$$

$$= \frac{2000}{\frac{5000}{4}} = \frac{2000}{1250}$$

$$= 1.60$$

$t_{cal} = 1.6$ Value is less than its Critical Value $t_\alpha = 1.753$, $\alpha = 0.05$ $d.f = 15$,

Ho is accepted

$1.6 < 1.753$

3. A fertilizer mixing machine is set to give 12kg of nitrate for every 100kg of fertilizer. Ten bags of 100 kg each are examined. The percentage of nitrate so obtained is 11, 14, 13, 12, 13, 12, 13, 14, 11, 12. Is there reason to believe that the machine is defective?

$H_0 : \mu = 12$

$H_1 : \mu \neq 12$

$n = 10$ $\quad d.f = n-1$ $\quad \alpha = 0.05$

$= 10-1$

$= 9$

$\bar{x} = \frac{\Sigma x}{n}$

$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

$= \sqrt{\frac{\Sigma d^2}{n-1} - \frac{(\Sigma d)^2}{n(n-1)}}$

Variable x	Deviation $d = x - 12$	d^2
11	-1	1
14	2	4
13	1	1
12	0	0
13	1	1
12	0	0
13	1	1
14	2	4
11	-1	1
12	0	0
$\Sigma x = 125$	$\Sigma d = 5$	$\Sigma d^2 = 13$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{125}{10} = 12.5$$

$$S = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}} = \sqrt{\frac{\Sigma d^2}{n-1} - \frac{(\Sigma d)^2}{n(n-1)}}$$

$$= \sqrt{\frac{13}{10-1} - \frac{(5)^2}{10(10-1)}}$$

$$= \sqrt{\frac{13}{9} - \frac{25}{10 \times 9}}$$

$$= \sqrt{\frac{13}{9} - \frac{25}{90}} = 1.08$$

Using the t-test

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} = \frac{12.5 - 12}{\frac{1.08}{\sqrt{10}}} = \frac{0.50}{0.341}$$

$$= 1.466$$

$t_{cal} = 1.466$ value is less than its critical value $t_{\frac{\alpha}{2}} = 2.262$ at $\frac{\alpha}{2} = 0.025$ and $df = 9$. the null hypothesis H_0 is accepted.

4. A random sample of size 16 has the sample mean 53. The sum of the squares of deviation taken from the mean value is 150. Can this sample be regarded as taken from the population having 56 as its mean? Obtain 95% and 99% confidence limits of the sample mean.

$H_0: \mu = 56$

$H_1: \mu \neq 56$

$n = 16, \quad df = n-1 \qquad \bar{x} = 53.$

$= 16-1$

$= 15$

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{150}{15}}$$

$= 3.162$

$$\frac{s}{\sqrt{n}} = \frac{3.162}{\sqrt{16}} = 0.7905$$

95% Confidence limit

$$= \bar{x} \pm t_{0.05} \frac{s}{\sqrt{n}}$$

t table value

$= 53 \pm 2.13 (0.7905)$

$= 53 \pm 1.683.$

99%

$\bar{x} \pm t_{0.01} \frac{s}{\sqrt{n}}$

table value is ~ 0.01

$= 53 \pm 2.947 (0.7905)$

$= 53 \pm 2.33.$

Hypothesis testing for difference of two population Means (Independent Samples)

For comparing the mean values of two normally distributed population we draw independent random samples of sizes n_1 and n_2 from the two populations. If μ_1 and μ_2 are the mean values of two populations then our aim is to estimate the value of the difference $\mu_1 - \mu_2$ between mean values of the two populations.

The sampling distribution of $\bar{x}_1 - x_2$ has the following properties.

Expected Value

$$E(\bar{x}_1 - x_2) = E(\bar{x}_1) - E(x_2)$$

$$= \mu_1 - \mu_2$$

Variance:

$$Var(\bar{x}_1 - \bar{x}_2) = Var(x_1) + Var(\bar{x}_2)$$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

If the population SD σ_1 and σ_2 are known, then the large sample interval estimation can also be used for the small sample case. But if these are unknown, then these are estimated by the sample SD s_1 and s_2. It is needed if sampling distribution is not normal even if sampling is done from two normal populations. Thus t-distribution is used to develop a small sample interval estimate for $\mu_1 - \mu_2$.

<u>Population variances are unknown but equal.</u>

If population variances σ_1^2 & σ_2^2 are unknown but equal that is, both populations have exactly the same shape and $\sigma_1^2 = \sigma_2^2 = \sigma^2$. the standard error of the difference in two sample means $\bar{x}_1 - \bar{x}_2$ can be written as. 1.

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$= \sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

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

$$s_1 = \sqrt{\frac{\Sigma(x_1 - \bar{x}_1)^2}{n_1 - 1}} \quad \& \quad s_2 = \sqrt{\frac{\Sigma(x_2 - \bar{x}_2)^2}{n_2 - 1}}$$

$$S^2 = \frac{\Sigma(x_1-\bar{x}_1)^2 + \Sigma(x_2-\bar{x}_2)^2}{n_1+n_2-2}$$

$$S = \sqrt{\frac{\Sigma(x_1-\bar{x}_1)^2 + \Sigma(x_2-\bar{x}_2)^2}{n_1+n_2-2}}$$

t-test statistic is defined as

$$t = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{S\,\bar{x}_1-\bar{x}_2}$$

$$= \frac{\bar{x}_1-\bar{x}_2}{S}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$$

Degrees of freedom

$$n_1+n_2-2$$

1. In a test given to two groups of students, the marks obtained are as follows.

First group 18 20 36 50 49 36 34 49 41

Second group 29 28 26 35 30 44 46

Examine the significance of the difference between the arithmetic Mean of the marks secured by the students of the above two groups.

Sol

$H_0 : \mu_1 - \mu_2 = 0$

$\mu_1 = \mu_2$ &

$H_1 : \mu_1 \neq \mu_2$ (Two-tailed test)

Apply t test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Calculation of sample mean $\bar{x}_1, \bar{x}_2$ and pooled sample standard deviation are

First Group x_1	$x_1 - \bar{x}_1 = x_1 - 37$	$(x_1 - \bar{x}_1)^2$	second grp x_2	$(x_2 - \bar{x}_2) = x_2 - 34$	$(x_2 - \bar{x}_2)^2$
18	-19	361	29	-5	25
20	-17	389	28	-6	36
36	-1	1	26	-8	64
50	13	169	35	1	1
49	12	144	30	-4	16
36	1	1	44	10	100
34	3	9	46	12	144
49	12	144			
41	4	16			
$\Sigma x_1 = 333$	$\Sigma(x_1 - \bar{x}_1) = 0$	$\Sigma(x_1 - \bar{x}_1)^2 = 1234$	$\Sigma x_2 = 238$	$\Sigma(x_2 - \bar{x}_2) = 0$	$\Sigma(x_2 - \bar{x}_2)^2 = 386$

$$\bar{x}_1 = \frac{\Sigma x_1}{n_1} = \frac{333}{9} = 37$$

$$x_2 = \frac{\Sigma x_2}{n} = \frac{238}{7} = 34$$

$$S = \sqrt{\frac{\Sigma(x_1-\bar{x}_1)^2 + \Sigma(x_2-\bar{x}_2)^2}{n_1+n_2-2}}$$

$$= \sqrt{\frac{1234+386}{9+7-2}}$$

$$= \sqrt{\frac{1620}{14}}$$

$$= 10.76$$

t-test statistic we get.

$$t = \frac{\bar{x}_1-\bar{x}_2}{S}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$$

$$= \frac{37-34}{10.76}\sqrt{\frac{9\times 7}{9+7}}$$

$$= \frac{3}{10.76}\sqrt{\frac{63}{16}}$$

$$= \frac{3}{10.76} \times 1.9843$$

$$= 0.551$$

$df = n_1+n_2-2$

$= 9+7-2 = 14$

$t_{cal} = 0.551$

$\alpha = 5\% = 0.05$

$\frac{\alpha}{2} = 0.025$

$t_{\frac{\alpha}{2}} = 2.145$

$0.551 < 2.145$

H_0 is accepted.

2. The manager of a Courier Service believes that packets delivered at the end of the month are heavier than those delivered early in the month. As an experiment he weighted a random sample of 20 packets at the beginning of the month. He found that the mean weight was 5.25 Kgs with a SD of 1.20 kgs. Ten packets randomly selected at the end of the month had a mean weight of 4.96 kgs and a SD of 1.15 Kgs at the level of significance 0.05, Can it be concluded that the packets delivered at the end of the month weight more?

$n_1 = 20 \quad \bar{x}_1 = 5.25 \quad S_1 = 1.20$

$n_2 = 10 \quad \bar{x}_2 = 4.96 \quad S_2 = 1.15.$

$$S = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$$

$$= \sqrt{\frac{19 \times (5.25)^2 + 9(4.96)^2}{20+10-2}}$$

$$= \sqrt{\frac{19 \times 27.56 + 9 \times 24.60}{28}}$$

$$= \sqrt{26.60}$$

$$= 5.16$$

t test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$= \frac{5.25 - 4.96}{5.16} \sqrt{\frac{20 \times 10}{20 + 10}}$$

$$= \frac{0.29}{5.16} \sqrt{\frac{200}{30}}$$

$$= 0.056 \times 2.58$$

$$= 0.145$$

$\alpha = 0.01$, $df = 28$ $t_{cal} = 0.145$

$t_\alpha = 1.701$ (table value)

$t_{cal} <$ table value

$0.145 < 1.701$

H_0 is accepted.

<u>Hypothesis testing for difference of two population means (dependent samples)</u>

When two samples of the same size are paired so that each observation in one sample associated with any particular observation in the second sample, the sampling procedure to collect the data and then test the hypothesis is called matched samples.

t-test is called paired t-test becomes

n = number of paired observations

$df = n-1$, degrees of freedom.

$\bar{d}$ = Mean of the difference between paired (or related) observation

n = number of pairs of differences.

S_d = Sample Standard of the distribution of the difference between the paired (or related observation)

$$= \frac{\sum(d-\bar{d})^2}{\frac{S_d}{\sqrt{n}}}$$

Where n = number of paired observation

$df = n-1$, degrees of freedom.

$\bar{d}$ = number of pairs of differences

n = number of pairs of differences.

S_d = Sample standard deviation of the distribution of the difference between the paired (or related observations)

$$= \sqrt{\frac{\sum(d-\bar{d})^2}{n-1}} = \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}}$$

$H_0 : \mu_d = 0$

$H_1 : \mu_d > 0$ (or) $\mu_d < 0$ (one tailed test)

$\mu_d \neq 0$ (Two tailed test)

Confidence Interval:

The Confidence Interval estimate of the difference between two population means is given by

$$\bar{d} \pm t_{\frac{\alpha}{2}} \frac{S_d}{\sqrt{n}}$$

$t_{\alpha/2}$ = Critical value of t-test Statistic at n-1 degrees of freedom and α level of significance.

If the claimed value of null hypothesis H_0 lies within the confidence interval, then H_0 is accepted. Otherwise rejected.

Problems:

1. The HRD manager wishes to see if there has been any change in the ability of trainess after a specific training programme. The trainers take an aptitude test before the start of the programme and an equivalent one after they have completed it. The scores recorded are given below. Has any change taken place at 5% significance level?

Trainee	A	B	C	D	E	F	G	H	I
Score before training	75	70	46	68	68	43	55	68	77
Score after training	70	77	57	60	79	64	55	77	76

Sol

$H_0 : \mu d = 0$

$H_1 : \mu d \neq 0$

Trainee	Before training	After training	Difference the scores	d^2
A	75	70	5	25
B	70	77	-7	49
C	46	57	-11	121
D	68	60	8	64
E	68	79	-11	121
F	43	64	-21	441
G	55	55	0	0
H	68	77	-9	81
I	77	76	1	1
			$\Sigma d = -45$	$\Sigma d^2 = 903$

$$\bar{d} = \frac{\Sigma d}{n} = \frac{-45}{9} = -5$$

$$s_d = \sqrt{\frac{\Sigma d^2}{n-1} - \frac{(\Sigma d)^2}{n(n-1)}}$$

$$= \sqrt{\frac{903}{9-1} - \frac{(-45)^2}{9(9-1)}}$$

$$= \sqrt{\frac{903}{8} - \frac{2025}{72}}$$

$$= \sqrt{112.875 - 28.13} = 9.21$$

Applying t-test Statistic, We have

$$t = \frac{\bar{d} - \mu_d}{\frac{S_d}{\sqrt{n}}} = \frac{-5-0}{\frac{9.21}{\sqrt{9}}}$$

$$= \frac{-5}{3.07} = -1.63$$

$t_{cal} = -1.63$ is more than its Critical

Value, $t_{\frac{\alpha}{2}} = -2.31$

$df = n-1 = 9-1 = 8$ ↳ table Value $\alpha = \frac{0.05}{2}$

$\frac{\alpha}{2} = 0.025$ $= 0.025$

$8 \rightarrow 0.025$

$-1.63 < -2.31$ $= 2.306$

$= 2.31$

H_0 is accepted

2. 12 Students were given intensive Coaching and 5 tests were Conducted in a month. The Score of tests 1 and 5 are given below.

No. of Students :	1	2	3	4	5	6	7	8	9	10	11	12
Marks in 1st test :	50	42	51	26	35	42	60	41	70	55	62	38
Marks in 5th test :	62	40	61	35	30	52	68	51	84	63	72	50

Do the data indicate any improvement in the Scores obtained in tests 1 & 5.

Sol

No.of Students	Marks in 1st test	Marks in 5th test	Difference in marks	d^2
1	50	62	-12	144
2	42	40	+2	4
3	51	61	-10	100
4	26	35	-9	81
5	35	30	5	25
6	42	52	-10	100
7	60	68	-8	64
8	41	51	-10	100
9	70	84	-14	196
10	55	63	-8	64
11	62	72	-10	100
12	38	50	-12	144
			-100	1122

$$\bar{d} = \frac{\Sigma d}{n} = \frac{100}{12} = 8.3 \qquad \frac{-96}{12} = -8$$

$$Sd = \sqrt{\frac{\Sigma d^2}{n-1} - \frac{(\Sigma d)^2}{n(n-1)}}$$

$$= \sqrt{\frac{1122}{11} - \frac{(100)^2}{12(12-1)}} \qquad \frac{9216}{132}$$

$$= \sqrt{102 - 75.75} \qquad 69.81$$

$$\sqrt{32.19}$$

$$= 5.408.$$

$$= 5.6736$$

t - test statistics

$$t = \frac{\bar{d} - \mu d}{\frac{Sd}{\sqrt{n}}} = \frac{8.3}{\frac{5.408}{\sqrt{12}}} \quad \sqrt{12} = 3.4641 \qquad = \frac{-8}{\frac{5.6736}{\sqrt{12}}}$$

$$= \frac{8.3}{1.56115} \qquad = \frac{-8}{1.6398}$$

$$= 5.3165$$

$$= 4.878$$

$t_{cal} = 5.3165$ is more than its critical value $t_{\frac{\alpha}{2}} = 2.201$

$df = 11 \rightarrow \frac{\alpha}{2} = \frac{0.05}{2} = 0.025$

$5.3165 > 2.201$

Ho is rejected.

$\bar{d} = \frac{\Sigma d}{n}$

$t = \frac{\bar{d} - \mu d}{\frac{Sd}{\sqrt{n}}}$

.

Additional Resources:

https://www.youtube.com/watch?v=VK-rnA3-41c .

Practice Questions:

Section – A

1. Define Sampling.
2. Define Standard error.
3. Define Sampling distribution.
4. Define degrees of freedom.
5. Define non sampling errors.
6. State null hypothesis
7. State alternative hypothesis
8. Explain one tailed test.
9. Explain two tailed test.
10. What is meant by proportion.
11. Write the uses of t-distribution

Section – B

1. Explain about the choice of sampling method.
2. Distinguish between population, sample distribution and sampling distribution.
3. Explain about the Sampling methods
4. Explain about the principles of sampling.
5. The mean length of life of a certain cutting tool is 41.5 hours with a standard deviation of 2.5hours. What is the probability that a simple random sample of size 50 drawn from this population will have a mean between 40.5 hours and 42 hours?
6. Safal, a tea manufacturing company is interested in determining the consumption rate of tea per household in Delhi. The management believes that yearly consumption per household is normally distributed with an unknown mean μ and standard deviation of 1.50kg (a) If a sample of 25 household is taken to record their consumption of tea for one year, What is the probability that the sample mean is within 500gms of the population mean ? (b) How large a sample must be in order to be 98 percent certain that the sample mean is within 500gms of the population mean?
7. The particular brand of ball bearings weighs 0.5kg with a standard deviation of 0.02kg. What is the probability that two lots of 1000 ball bearings each will differ in weight by more than 2gms.
8. An experiment was conducted to compare the mean time in days required to recover from a common cold for person given daily dose of 4mg of Vitamin C versus those who were not given a vitamin supplement. Suppose that 35 adults were randomly deviations for the two groups were as follows:

	Vitamin C	No Vitamin Supplement
Sample size	35	35
Sample mean	5.8	6.9
Sample standard deviation	1.2	2.9

9. An auditor claims that 10 percent of customers ledger accounts are carrying mistakes of posting and balancing. A random sample of 600 was taken to test the accuracy of

posting and balancing and 45 mistakes were found. Are these sample results consistent with the claim of the auditor? Use 5 percent level of significance.

Section – C

1. Write briefly about the sampling methods?
2. A continuous manufacturing process produces items whose weights are normally distributed with a mean weight of 800gms and a standard deviation of 300gms. A random sample of 16 items is to be drawn from the process. (a) What is the probability that the arithmetic mean of the sample exceeds 900gms? Interpet the results. (b) Find the values of the sample arithmetic mean within which the middle 95 percent of all sample mean will fall.
3. Car stereos of manufacturer A have a mean lifetime of 1400 hours with a standard deviation of 200 hours, while those of manufacturer B have a mean lifetime of 1200 hours with a standard deviation of 100 hours. If a random sample of 125 stereos of each manufacturer are tested, what is the probability that manufacturer A's stereos will have a mean lifetime which is atleast (a) 160 hours more than manufacturer B's stereos and (b) 250 hours more than the manufacturer B's stereos?
4. Explain about the Sampling distribution of difference between two sample mean.
5. Explain about the Sampling distribution of difference of two proportions.
6. Ten percent of machines produced by company A are defective and five percent of those produced by company B are defective. A random sample of 250 machines is taken from company A and a random sample of 300 machines from company B.What is the probability that the difference in sample proportion is less than or equal to 0.02?
7. A firm believes that the tyres produced by process A on an average last longer than tyres produced by process by B. To test this belief, random samples of tyres produced by the two processes were tested and the results are:

Process	Sample size	Average Lifetime (in km)	Standard deviation (in km)
A	50	22,400	1000
B	50	21,800	1000

8. 12 students were given intensive coaching and 5 tests were conducted in a month. The scores of tests 1 and 5 are given below

No.of Students	1	2	3	4	5	6	7	8	9	10	11	12
Marks in 1^{st} test	50	42	51	26	35	42	60	41	70	55	62	38
Marks in 5^{th} test	62	40	61	35	30	52	68	51	84	63	72	50

.Do the data indicate any improvement in the scores obtained in tests 1 and 5.

9. An auto company decided to introduce a new six cylinder car whose mean petrol consumption is claimed to be lower than the existing auto engine. It was found that the mean petrol consumption for 50 cars was 10km per litre. Test for the company at 5 percent level of significance, the claim that in the new car petrol consumption is 9.5 km per litre on the average.

.

.

References:

J.K. Sharma, Business Statistics- Pearson Education.

Gupta P.K andvDr.D.S.Hira D.S.,(2022) " Operation Research", 7th Edition, S.Chand, Noida(U P)