TEST OF HYPOTHESIS – I
ESTIMATIONS

ESTIMATE:

An Estimate is a statement made to find unknown population parameter (μ, σ called parameters, $\bar{X}, S$ are called statics)

ESTIMATOR:

The method (or) procedure (or) rule to determine an unknown parameter is called an Estimator. Sample mean is Estimator for μ.

ESTIMATION:

Process of determining Estimator in such a way that they are close to parameter value. This is known as Estimation.

ESTIMATION CAN BE DONE IN TWO WAYS:

(a) Point Estimation (b) Interval Estimation

(A) POINT ESTIMATION:

* In a point estimation, parameter is estimated by a single numerical value from sample data.
* A point estimator is a static for estimating parameter θ and it will be denoted by $\hat{\theta}$
* An estimator is not expected to estimate the population parameter without error. An estimator should be close to the true value of unknown parameter.
* If $\hat{\theta}$ be an estimator of θ. If expected value of $\hat{\theta}$ *is equal to* θ then estimator is unbiased. i.e.,$E(\hat{\theta}) = \theta \ \ or\ E(static) = parameter$
* S^2 is unbiased estimator of parameter σ^2.

VARIANCE OF POINT ESTIMATOR:

* $\widehat{\theta_1}\ and\ \widehat{\theta_2}$ are two unbiased estimators of same population parameter θ. The estimator whose sampling distribution has the small variance is variance of point estimator i.e.,$\sigma^2_{\theta_1} < \sigma^2_{\theta_2}$
* If we consider all possible unbiased estimator of same parameter then smallest variance is called most efficient estimator.
* An estimator is said to be a good estimator if it is unbiased, consistent and efficient.

(B) INTERVAL ESTIMATION:

Point estimate rarely coincide with quantities. So, instead of point estimation where the parameter to be estimated by a single value, better way of estimation is Interval estimation.

* An interval estimation determines an interval in which parameter value can be expected. Such an interval is called an "Interval Estimation".
* An interval estimator of population parameter θ is on interval of the form $\widehat{\theta_L} < \theta < \widehat{\theta_U}$ where $\widehat{\theta_L}, \widehat{\theta_U}$ are called Upper, Lower Confidence Limits.
* $P(\theta_L < \theta < \theta_U)$ is equal to a positive fractional value. For instant we assign that $P(\theta_L < \theta < \theta_U) = 1 - \alpha\ where\ 0 < \alpha < 1$
  The interval $\widehat{\theta_L} < \theta < \widehat{\theta_U}$ computed from selected sample.
* $(1-\alpha)100\%$ is called confidence interval.
  The fraction $1-\alpha$ is called Confidence Coefficient.
  For example: If $\alpha = 0.05$ then confidence interval is 95%. Similarly for $\alpha = 0.01$, confidence interval is 99%.

BAYESIAN ESTIMATION:

The Bayesian estimation method involving sample information to be combined with prior distribution of μ. This gives posterior distribution of μ which is approximately normal with

Mean $(\mu_1) = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2}$

$S_D\sigma_1 = \sqrt{\frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2}}$

$where\ \bar{x} = mean\ of\ sample,$
$n = size\ of\ sample$
$\sigma^2 = variance\ of\ population,$
$\mu_0 = mean\ of\ prior\ distribution,$
$\sigma_0^2 = variance\ of\ prior\ distribution,$
$\mu_1 = mean\ of\ posterior\ distribution,$
$\sigma = SD\ of\ posterior\ of\ distribution.$

BAYESIAN INTERVAL:

A $(1-\alpha)100\%$ Bayesian interval for μ is represented as

$$\boxed{\left(\mu_1 - Z_{\frac{\alpha}{2}}\sigma_1\right) < \mu < \left(\mu_1 + Z_{\frac{\alpha}{2}}\sigma_1\right)}$$

MAXIMUM ERROR OF ESTIMATE E FOR LARGE SAMPLE:

Since the sample mean $\bar{X}$ estimate very rarely equal to mean of population μ (parameter)
Error is $|\bar{X} - \mu|$
For large sample the random variable
$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is normal variate approximately
We know that $P\left(-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}\right) = 1 - \alpha\ where\ Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$

(FIGURE)

$P\left(-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}\right) = 1 - \alpha$

$P\left(-Z_{\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < Z_{\frac{\alpha}{2}}\right) = 1 - \alpha$

Multiplying each term in the inequality by $\sigma/\sqrt{n}$ and subtracting $\bar{X}$ from each term and multiplying by -1

$\therefore P(\bar{X} - Z_{\frac{\alpha}{2}}(\sigma/\sqrt{n}) < \mu < \bar{X} + Z_{\frac{\alpha}{2}}(\sigma/\sqrt{n}) = 1 - \alpha$

CONFIDENCE INTERVAL FOR μ, σ KNOWN:

If $\bar{X}$ is the mean of a random sample of size n from the population with known variance $\sigma^2, (1-\alpha)100\%$confidence interval for μ is given by $\boxed{\bar{X} \pm Z_{\frac{\alpha}{2}}(\sigma/\sqrt{n})}$

Error $\bar{X} - \mu$ will not exceed $Z_{\frac{\alpha}{2}}(\sigma/\sqrt{n}),\ where\ Z_{\frac{\alpha}{2}}\ is\ the\ Z\ value$ leaving an area of $\alpha/2$ to the right. E is maxim error of estimate E with (1 – α) probability

$* E = Z_{\frac{\alpha}{2}}(\sigma/\sqrt{n})$

SAMPLE SIZE (n):

When $\alpha,\ E, \sigma\ are\ known\ then\ n = \left(\frac{Z_{\frac{\alpha}{2}}\sigma}{E}\right)^2$

When σ is unknown then σ is replaced by S.D. of sample S.

MAXIMUM ERROR OF ESTIMATE FOR SMALL SAMPLE:

For small samples the maximum error of estimate can be taken as

$P\left(-t_{\frac{\alpha}{2}} < T < t_{\frac{\alpha}{2}}\right) = 1 - \alpha \quad where\ T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$

$P\left(-t_{\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{\frac{\alpha}{2}}\right) = 1 - \alpha$

Multiplying each term in the inequality by $S/\sqrt{n}$ and then subtracting $\bar{X}$ from each term and multiply by -1.

FIGURE

$P(\bar{X} - T_{\frac{\alpha}{2}}\left(s/\sqrt{n}\right) < \mu < \bar{X} + T_{\frac{\alpha}{2}}\left(s/\sqrt{n}\right)\ where\ t_{\frac{\alpha}{2}}\ is\ t - value\ with\ V = n - 1$ Degrees of freedom leaving area of $\alpha/2$ to the right.

$E = t_{\frac{\alpha}{2}}\left(S/\sqrt{n}\right) \quad S = SD\ of\ sample; n = sample\ size$

Confidence Interval $\boxed{\bar{X} \pm\ t_{\frac{\alpha}{2}}\left(S/\sqrt{n}\right)}$

PROBLEMS:

In a study of an automobile insurance a random sample of 80 body repair costs had a mean of Rs 472.36 and S.D. of Rs. 62.35. If $\bar{X}$ is used as a point estimate to the true average repair costs. With that confidence we can assert that maximum error doesn't exceed Rs 10.

Sol: Size of sample $n = 80$

The mean of random sample $\bar{X} = Rs\ 472.36$

S.D. $\sigma = 62.35$

Maximum error of estimate $E_{\max error} = Rs. 10$

$$\left\{\begin{array}{c}
\underline{\boldsymbol{For\ large\ samples}} \\
E_{maxerror} = Z_{\frac{\alpha}{2}}\left(\sigma/\sqrt{n}\right) \\
Confidence\ interval\ (\bar{X} - Z_{\frac{\alpha}{2}}\left(\sigma/\sqrt{n}\right),\ \bar{X} + Z_{\frac{\alpha}{2}}\left(\sigma/\sqrt{n}\right) \\
n = \left(\frac{Z_{\frac{\alpha}{2}}\sigma}{E}\right)^2 \\
\underbrace{Z_{\frac{\alpha}{2}}}_{0.02} = 2.33\ for\ 98\%,\quad \underbrace{Z_{\frac{\alpha}{2}}}_{0.025} = 1.96\ for\ 95\%,\ Z_{0.5} = 1.645\ \ 90\%
\end{array}\right\}$$

$$\left\{\begin{array}{c}
\underline{\boldsymbol{for\ small\ samples}} \\
E_m = t_{\frac{\alpha}{2}} S/\sqrt{n}\ \ in\ the\ interval\ (\bar{X} - t_{\frac{\alpha}{2}}\left(s/\sqrt{n}\right),\ \bar{X} + t_{\frac{\alpha}{2}}\left(s/\sqrt{n}\right) \\
\boxed{T = \frac{x-\mu}{S/\sqrt{n}}}
\end{array}\right\}$$

$E_{max} = Z_{\frac{\alpha}{2}}\left(\sigma/\sqrt{n}\right) \ \Rightarrow\ Z_{\alpha/2} = \frac{E_{max}\sqrt{n}}{\sigma} = \frac{10\sqrt{80}}{62.35}$

$Z_{\alpha/2} = 0.9236\ from\ table$

$1-\frac{\alpha}{2}=Z_{\frac{\alpha}{2}}=0.9236 \;\Rightarrow\; \alpha=0.1528$

Consider $(1-\alpha)100\%=84.72\%$

1. Mean of random sample is unbiased estimate of the mean of the population 3, 6, 9, 15, 27.
i) List of all possible sample of size 3 that can be taken without replacement from finite population.
ii) Calculate mean of each of the samples listed in assigning each sample a probability of 1/10. Verify the mean of these $\bar{X}$ is equal to 12 which is equal to the mean of population θ, i.e., $E(\bar{X})=\theta$. Prove that $\bar{X}$ is unbiased.

Sol: i) The possible samples of size 3 taken from 2, 6, 9, 15, 27 without replacement are ${}^5C_3=10$ samples, i.e.,$(3,6,9),(3,6,15),(3,6,27),(6,9,15),(6,9,27),(3,9,15),$
$(3,9,27),(9,15,27),(6,15,27),(3,15,27)$

ii) Mean of population $\theta=\frac{3+6+9+15+27}{5}=12$

Mean of samples are 6, 8, 12, 10, 14, 9, 13, 17, 16, 15

Probability assigns to each one is 1/10

$\bar{X}$	6	8	12	10	14	9	13	17	16	15
$P(\bar{X})$	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10

$E(\bar{X})=6\times\frac{1}{10}+8\times\frac{1}{10}+12\times\frac{1}{10}+10\times\frac{1}{10}+14\times\frac{1}{10}+9\times\frac{1}{10}+13\times\frac{1}{10}+17\times\frac{1}{10}+16\times\frac{1}{10}+15\times\frac{1}{10}=\frac{120}{10}=12$

$E(\bar{X})=\theta$

$\bar{X}$is unbiased estimate of θ.

2. Suppose that we observe random variable having binomial distribution and get x success in n trials (a) show that $\frac{x}{n}$ is an unbiased estimate of binomial parameter P.
(b) show that $\frac{x+1}{n+2}$is not an unbiased estimate of binomial parameter P.

Sol: If $\hat{\theta}$ is an unbiased estimator of q, if $E\hat{\theta}=q$

a) $E\left(\frac{x}{n}\right)=\frac{1}{n}E(x)=\frac{np}{n}=p \quad \left(\therefore \frac{x}{n}\ \textit{is an unbaised estimator of } p\right)$

b) $E\left(\frac{x+1}{n+2}\right)=E\left(\frac{x}{n+2}+\frac{1}{n+2}\right)(E(ax+b)=aE(x)+b)$

$=\frac{1}{n+2}E(x)+\frac{1}{n+2}=\frac{p+1}{n+2}\neq p \quad \frac{x+1}{n+2}\ \textit{is not an unbiased estimate of } p$

3. Find 95% confidence limits for the mean of a normally distributed population from which the following sample was taken 15, 17, 10, 18, 16, 9, 7, 11, 13, 14

Sol: We have $\bar{X}=\frac{15+17+10+18+16+9+7+11+13+14}{10}=\frac{130}{10}=13$

$S^2=\frac{\sum(x_i-\bar{X})^2}{n-1} \quad =\left[\frac{(15-13)^2+(17-13)^2+(10-13)^2+(18-13)^2+\cdots..+(14-13)^2}{9}\right]=\frac{120}{9}=\frac{40}{3}$

$S=\sqrt{40/3}=3.651.$

Given confidence limit

$(1-\alpha)100=95\%\Rightarrow(1-\alpha)=\frac{95}{100}=0.95\Rightarrow\ \alpha=1-0.95=0.05$

$\frac{\alpha}{2}=0.025,\ \ n=10, \quad t_{\alpha/2}=t_{0.025}=2.262,\ (from\ table)\ \ v=10-1=9$

$t_{\alpha/2}\frac{\sqrt{S^2}}{\sqrt{n}}=\frac{\sqrt{40}}{\sqrt{10}\times\sqrt{3}}\times 2.262=2.6$

Confidence limit $(\bar{X}-t_{\frac{\alpha}{2}}(s/\sqrt{n}),\ \bar{X}+t_{\frac{\alpha}{2}}(s/\sqrt{n})=(13-2.6,13+2.6)=(10.4,15.6)$

4. Ten bearings made by a certain process have a mean diameter of 0.5060 cm with S.D. of 0.0040. Assuming that the data may be taken as a random sample from a normal distribution. Construct a 95% confidence interval for the actual average diameter of bearings.

Sol: Given that sample size $n = 10 < 30$ $\quad(so\ use\ t\ distribution)$

$\bar{X} = 0.5060$

Maximum error estimate = 95%

$t_{\frac{\alpha}{2}} = 2.262$ (from table)

$E = t_{\frac{\alpha}{2}}(\sigma/\sqrt{n}) = \frac{2.262\times0.004}{\sqrt{10}} = 0.00286$

95% confidence internal limits are

$(\bar{X} - t_{\frac{\alpha}{2}}(s/\sqrt{n}),\ \bar{X} + t_{\frac{\alpha}{2}}(s/\sqrt{n})$

$= 0.5060 \pm 0.00286 = (0.5031, 0.5089)$

5. A sample of 11 rats from a central population had an average blood viscosity 3.92 S.D. of 0.6. Estimate the 95% confidence limits for the mean blood viscosity of the population.

Sol: We have $t_{\frac{\alpha}{2}}\ for\ 95\%\ and\ v = n - 1 = 11 - 1 = 10\ is\ 2.23.$

Given $S = S.D. = 0.61,\ n = 11,\ \bar{X} = 3.92$

Confidence interval $(\bar{X} - t_{\frac{\alpha}{2}}(s/\sqrt{n}),\ \bar{X} + t_{\frac{\alpha}{2}}(s/\sqrt{n}) = 3.92 \pm 2.23\left(\frac{0.61}{\sqrt{11}}\right)$

$= 3.92 \pm 0.41 = (3.51, 4.33)$

6. What is the size of smallest sample required to estimate an unknown proportion to within a maximum error of 0.06 with atleast 95% confidence.

Sol: Given maximum error $E = 0.06$

Confidence limit = 95%

$(1 - \alpha)100 = 95\% \Rightarrow 1 - \alpha = 0.95$

$$\left\{\begin{matrix} Value\ of\ Z_{\frac{\alpha}{2}}\ for\ 95\% \\ i.e., Z_{0.025} = 1.96 \\ for\ 90\%, Z_{0.5} = 1.645 \\ for\ 98\%, Z_{0.02} = 2.33 \end{matrix}\right\}$$

$\Rightarrow\ \alpha = 1 - 0.95 = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$

$Z_{\frac{\alpha}{2}} = 1.96\ \ from\ table$

Here p is not given so we take $p = \frac{1}{2}, q = \frac{1}{2}$

$n = \left(\frac{Z_{\frac{\alpha}{2}}}{E}\right)^2 pq \quad = \frac{1}{4}\left(\frac{1.96}{0.06}\right)^2 = 266.78 = 267.$

7. If we can assert with 95% that the maximum error is 0.05 and p = 0.2, find sample size?

Sol: Given $p = 0.2, E = 0.05\ \ \ for\ 95\%\ \ Z_{\frac{\alpha}{2}} = 1.96$

$E = Z_{\alpha/2}\sqrt{\frac{pq}{N}} \Rightarrow 0.05 = 1.96\sqrt{0.2 \times \frac{0.8}{n}} \Rightarrow n = \frac{0.2\times0.8\times(1.96)^2}{(0.05)^2} = 246$

8. Assuming that $\sigma = 20.0$ how large a random sample be taken to assert with probability 0.95 that the sample mean will not differ from true mean by more than 3.0 points?

Sol: Given Max error E = 3.0, and $\sigma = 20.0$

$1 - \alpha = 0.95 \rightarrow 95\%$

$Z_{\frac{\alpha}{2}} = 1.96\ ,\ n = \left(\frac{Z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{1.96\times20}{3}\right)^2 \Rightarrow n = 171$

9. It is desired to estimate the mean number of hours of continuous use until a certain computer will first require repair. If it can be assumed that $\sigma = 48\,hours$, how large a sample be needed so that one will be able to assert with 90% confidence that sample mean is off by at most 10 hours.

Sol: It is given that max error E = 10 hours

$\sigma = 48\ hours$

$Z_{\frac{\alpha}{2}} = 1.645\ (for\ 90\%)$

$n = \left(\frac{Z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{1.645\times 48}{10}\right)^2 = 62.3 \Rightarrow n = 62$

10. A random sample of size 100 has a S.D. = 5. What can you say about maximum error with 95% confidence.

Sol: Given S = 5, n = 100, $Z_{\frac{\alpha}{2}} = 1.96\ \ for\ 95\%$

$Max\ E = Z_{\frac{\alpha}{2}}\left(\sigma/\sqrt{n}\right) = \frac{1.96\times 5}{\sqrt{100}} = 0.98$

11. What is maximum error one can expect to make probability 0.90 when using the mean of a random sample size n = 64 to estimate the mean of population with $\sigma^2 = 2.56$.

Sol: $n = 64,$

$probability\ 1 - \ \alpha = 0.90$

$\Rightarrow \alpha = 1 - 0.90 = 0.10 \Rightarrow \frac{\alpha}{2} = 0.5$

$Z_{\alpha/2} = 1.645$

Max error $E = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1.645 \times \frac{1.6}{\sqrt{64}} = 0.329$

12. A random sample of size 81 was taken whose variance is 20.25 and mean is 32. Construct 98% confidence interval.

Sol: Given$\bar{X} = sample\ mean = 32$

$n = 81,\ \sigma^2 = 20.25 \Rightarrow\ \sigma = 4.5$

$1 - \alpha = 0.98$

$\Rightarrow\ \alpha = 1 - 0.98 = 0.02 \Rightarrow \frac{\alpha}{2} = 0.01$

$Z_{\alpha/2} = 2.33\ \ for\ 98\%$

We know that confidence interval is

$(\bar{X} - Z_{\frac{\alpha}{2}}\left(\sigma/\sqrt{n}\right),\ \bar{X} + Z_{\frac{\alpha}{2}}\left(\sigma/\sqrt{n}\right)\ \left[Z_{\frac{\alpha}{2}}\left(\sigma/\sqrt{n}\right) = 2.33\frac{4.5}{\sqrt{81}} = 1.165\right]$

$= (32 - 1.165,\ 32 + 1.165) = (30.835, 33.165)$

13. A random sample of size 100 is taken from a population with $\sigma = 5.1$. Given that sample mean $\bar{X} = 21.6$ construct a 95% confidence interval for population mean μ.

$\bar{X} = sample\ mean = 21.6,\ Z_{\alpha/2} = 1.96\ (for\ 95\%)$

$n = 100, \sigma = 5.1$

Confidence interval $(\bar{X} - Z_{\frac{\alpha}{2}}\left(\sigma/\sqrt{n}\right),\ \bar{X} + Z_{\frac{\alpha}{2}}\left(\sigma/\sqrt{n}\right)\ \left[Z_{\frac{\alpha}{2}}\left(\sigma/\sqrt{n}\right) = 1.96\frac{5.1}{\sqrt{100}} \cong 1\right]$

$= (21.6 - 1,\ 21.6 + 1) = (20.6,\ 22.6)$

14. The mean and S.D. of population are 11,795 and 14,054. What can one assert with 95% confidence about max error if $\bar{X} = 11{,}795\ and\ n = 50$ also construct 95%. Confidence interval for true mean (or) The mean and S.D. of population are 11,795 and 14054. If n = 50 find 95% confidence interval for mean?

Sol: Mean of population $\mu = 11795$

S.D. of population $\sigma = 14054$

$\bar{X} = 11{,}795 \ and \ n = 50$

Max error $= Z_{\frac{\alpha}{2}}(\sigma/\sqrt{n})$

For 95% $Z_{\frac{\alpha}{2}} = 1.96$

$E = Z_{\frac{\alpha}{2}}(\sigma/\sqrt{n}) = 1.96 \times \frac{14054}{\sqrt{50}} = 3899$

Confidence interval $(\bar{X} - Z_{\frac{\alpha}{2}}(\sigma/\sqrt{n}),\ \bar{X} + Z_{\frac{\alpha}{2}}(\sigma/\sqrt{n})$

$= (11795 - 3899, 11795 + 3899) = (7896, 15694)$

15. $\bar{X} = 1.02, \sigma = 0.044$ for normal population. Determine 95% confidence interval for actual mean.

Sol: $\bar{X} = 1.02, \sigma = 0.044,\ Z_{\alpha/2}\ for\ 95\% = 1.96$

Confidence interval $(\bar{X} - Z_{\frac{\alpha}{2}}(\sigma/\sqrt{n}),\ \bar{X} + Z_{\frac{\alpha}{2}}(\sigma/\sqrt{n})$

$= \left(1.02 \pm 1.96\left(\frac{0.044}{\sqrt{10}}\right)\right) = (1.02 \pm 0.027) = (1.047,\ 0.993)$

SAMPLING DISTRIBUTION:

POPULATION:

A group of individual objects possessing in an area. No. of observations in the population is defined as size of population (N). A population may be finite or infinite.

SAMPLE:

A finite subset of population is sample. It is useful to know nature of population.
Some types of samples

1. PURPOSIVE SAMPLING:

If the sample elements are selected with definite purpose is called purposive sampling.

2. RANDOM SAMPLING:

A random sample is one in which each sample unit of population has an equal chance of being included.

3. SIMPLE SAMPLING:

In a random sampling in which the chance of selection of member of the population for the sample is independent of the previous selection, the sample is called simple sample.
Simple sample is special type of random sample.

4. STRATIFIED SAMPLING:

The sample which is aggregated of two sample individuals of each stratum is called stratified sample.

Examples : Population of employees may be divided into strata, rural employees and urban employees. After dividing into strata, we select individual from each of stratum.

$$\boxed{\begin{array}{c} \mu \ \rightarrow Mean\ of\ population \\ \sigma \ \rightarrow S.D.of\ population \\ S \ \rightarrow S.D.of\ sample \\ \bar{X} \rightarrow Mean\ of\ random\ sample \end{array}}$$

SAMPLE SIZE (n) :

No. of items in sample is called sample size (n)

SMALL SAMPLE:

If the size of the sample is < 30, the sample is called small sample.

LARGE SAMPLE:

If the size of the sample is ≥ 30 then the sample is called large sample.

SAMPLING DISTRIBUTION OF THE MEAN (σ KNOWN):

STATISTIC:

A function of random variable constitute random sample is called statistic.

POPULATION MEAN:

If $X_1,\ X_2, \ldots\ldots X_n$ represent a random sample of size n then sample mean is defined as $\mu = \sum_{i=1}^{n} \frac{x_i}{n}$

VARIANCE OF POPULATION:

$\sigma = \sum_{i=1}^{n} \frac{(x_i - \bar{X})^2}{n}$

STANDARD ERROR: (S.E.)

The standard deviation of sampling distribution of a statistic is known as Standard error. Standard error $= \frac{\sigma}{\sqrt{n}}$. S.E. plays very important role in the theory of large sample.

CENTRAL LIMIT:

If $\bar{X}$ is the mean of a random sample of size n taken from a population having the mean μ and the finite variance σ^2 then

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

NOTE:

1. Mean of population $\mu = \frac{\sum x}{N},\ S.D = \sqrt{\frac{\sum(x-\mu)^2}{N}}$
2. For a finite population of size N, S.D. $= \sigma, mean = \ \mu, sample\ size\ n$
   Variance $= \sigma^2 = \frac{\sigma^2}{N}\left(\frac{N-n}{N-1}\right)\ (without\ replacement)$
3. $\left(\frac{N-n}{N-1}\right)$is called finite population correction error.
   No. of samples of size n from population of size N
   i) ${}^{N}C_n\ without\ replacement$
   ii) $N^n\ with\ replacement$

SAMPLING DISTRIBUTION OF MEANS:

Suppose the samples are drawn from infinite population (or) sampling is done with replacement

Mean $\mu_x = \frac{\mu+\mu+\mu+\cdots..n}{n} = \frac{n\mu}{n} = \mu$

Variance $\sigma^2 = \frac{\sigma^2+\sigma^2\pm\cdots....\sigma^2}{n} = \frac{\sigma^2}{n}$

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

PROBLEMS:

1. What is the value of correction factor if i) n = 5, N = 200

Sol: $\frac{N-n}{N-1} = \frac{200-5}{200-1} = \frac{195}{199} = 0.98$

ii) $n = 10, N = 1000$

$\frac{N-n}{N-1} = \frac{1000-10}{1000-1} = \frac{990}{999} = 0.001$

2. A population consist of five numbers 2, 3, 6, 8, 11. Consider all possible samples of size two, which can be drawn without replacement from the population.

Find: (a) The mean of the population
(b) S.D. of the population
(c) The mean of the sampling distribution of mean.
(d) The S.D. of sampling distribution of mean.

Sol: The possible samples are $^5C_2 = 10$

They are $(2,3),(2,6),(2,8),(2,11),(3,6),(3,8),(3,11),(6,8),(6,11),(8,11)$

a) The mean of the population

$\mu = \frac{2+3+6+8+11}{5} = \frac{30}{5} = 6$

b) S.D. of the population

$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}} = \sqrt{\frac{(2-6)^2+(3-6)^2+(6-6)^2+(8-6)^2+(11-6)^2}{5}} = \sqrt{\frac{54}{5}} = 3.29$

c) The mean of the sampling distribution

$(2,3) \to \frac{2+3}{2} = 2.5,\ (2,6) \to \frac{2+6}{2} = 4$

Likewise for all other samples means are $5, 6.5, 4.5, 5.5, 7, 7, 8.5, 9.5$

d) S.D. of sampling distribution

$S.D. = \frac{\sigma}{\sqrt{n}} = \frac{3.29}{\sqrt{2}} \cong 2$

TEST OF HYPOTHESIS – I

TESTING A HYPOTHESIS:

A statistical hypothesis is a statement about the parameter of one or more population. Testing on hypothesis is a process for deciding whether to accept or reject the hypothesis.

There are two types of hypothesis

1) Null hypothesis
2) Alternative hypothesis

1. NULL HYPOTHESIS:

Hypothesis of no difference is called Null Hypothesis. A null hypothesis is the hypothesis which is tested for possible rejection under assumption that is true. It is always denoted by H_0.

To test one procedure is better than other we assume that there is no difference between procedure.

Example: 1. To decide whether a computer perform well we form the hypothesis that computer doing well.

2. To estimate parameter μ for example μ_0 is estimated value, we take hypothesis like $\mu = \mu_0$.

2. ALTERNATIVE HYPOTHESIS:

A hypothesis which is complementary to null hypothesis is called Alternative Hypothesis.

For example: $H_0: \mu = \mu_0$

$H_1: \mu > \mu_0 \ or \ \mu_1 < \mu_0 \ \ i.e., \ \mu \neq \mu_0$

ERRORS IN SAMPLING:

To accept or reject the hypothesis always gives rise to some error or risk. There are two types of errors in testing the hypothesis.

i) TYPE I – ERROR:

Reject H_0 when it is true, i.e., rejecting a correct hypothesis.
P (reject H_0 when it is true) = P(type 1 error) = α
For example:H_0:$The\ coin\ is\ unbiased.$
After investigation if H_0 is true then we reject the coin.

ii) TYPE II – ERROR:

Accept H_0 when it is wrong. If a hypothesis is accepted while it should have been rejected we say the type – 1 error has been committed.
P (Accept H_0 when it is wrong) = P(Type II error) = β

The statistical testing of hypothesis aims at limiting. Type – 1 error to (1% or 5%) and to minimize type – 2 error. The only way to reduce both type of errors is to increase the sample size if possible.

CRITICAL REGION:

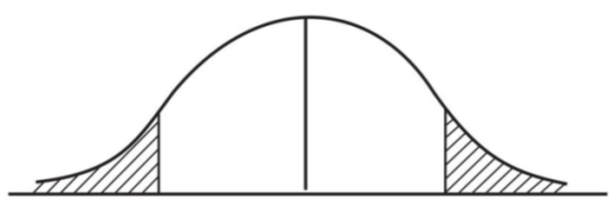
A region corresponding to a statistic in the sample space S which leads to rejection of H_0 called critical region or rejection region. The region which lead to accept H_0 is called acceptance region. In general the region which cover 5% and area of normal curve are rejection area S.

CRITICAL VALUE OR SIGNIFICANT VALUE:

The value of the test statistic which separates the critical region (or rejection region) and the acceptance region is called critical value or significant value. This value depends on level of significance and alternative hypothesis whether it is two tailed.

ONE – TAILED AND TWO – TAILED TESTS:

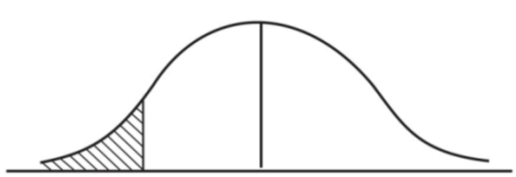
If the alternative hypothesis is of the not equal type i.e., in alternative hypothesis maximum cases $\mu \neq \mu_0, \sigma \neq \sigma_0$. Then critical region lies in both sides of right and left tails of the curve such that the critical region of area $\alpha/2$ lies on right and critical region of area $\alpha/2$ lies on left tail.



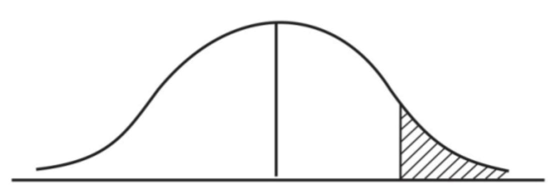
2
Two Tailed

ONE TAILED TEST:

For one tail test critical region is represented at only one side (left or right).
Suppose we want to test null hypothesis $H_0: \mu = \mu_0$
Alternative hypothesis $H_1: \mu < \mu_0 (Left\ tailed)$; $H_1: \mu > \mu_0\ (Right\ tailed)$



Left Tail



Right Tail

WORKING RULE OF TESTING HYPOTHESIS:

Step 1: Define a null hypothesis H_0 in clear terms

Step 2: Define alternative hypothesis H_1 so that one can decide we should use one tailed or two tailed test.

Step 3: Level of significance, i.e., suitable 'α' is selected in advance.

Step 4: Test static Z : Calculate value of Z using $Z = \frac{x-\mu}{\sigma}$ or $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

Step 5: Conclusion: Compare the computed value of Z with critical value Z_α(from table) at given level of significance.

* If $|Z| < Z_\alpha$ (i.e., calculated value of Z is less than table value of Z). It is not significant then we accept the null hypothesis.

* If $|Z| > Z_\alpha$ (i.e., calculated value of Z is greater than table value of Z), then we reject the null hypothesis at that level of significance.

FOR TWO TAILED TEST:

If $|Z| < 1.96$ $accept\ H_0\ at\ 5\%\ level\ of\ significance$
If $|Z| > 1.96$ $reject\ H_0\ at\ 5\%\ level\ of\ significance$
If $|Z| < 2.58$ $accept\ H_0\ at\ 1\%\ level\ of\ significance$
If $|Z| > 2.58$ $reject\ H_0\ at\ 1\%\ level\ of\ significance$

FOR SINGLE TAILED (RIGHT OR LEFT) TEST:

If $|Z| < 1.645$ $accept\ H_0\ at\ 5\%\ level\ of\ significance$
If $|Z| > 1.645$ $reject\ H_0\ at\ 5\%\ level\ of\ significance$
If $|Z| < 2.33$ $accept\ H_0\ at\ 1\%\ level\ of\ significance$
If $|Z| > 2.33$ $reject\ H_0\ at\ 1\%\ level\ of\ significance$

PROBLEMS:

1. A coin was tossed 960 times and returned heads 183 times. Test the hypothesis that coin is unbiased. Use a 0.05 level of significance.

Sol: Here $n = 960, p = \frac{1}{2},\ q = \frac{1}{2}$

	Level of significance		
	1% (0.01)	**5%(0.05)**	**10%(0.1)**
Two tailed (98%)	$\lvert Z_\alpha\rvert = 2.58$	$\lvert Z_\alpha\rvert = 1.96$ (95%)	$\lvert Z_\alpha\rvert = 1.645$
Right tailed	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$ (90%)	$Z_\alpha = 1.28$
Left tailed	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$

$\mu = np$

$\mu = 960 \times \frac{1}{2} = 480$

$\sigma = \sqrt{npq} = \sqrt{(np)q} = \sqrt{480 \times \frac{1}{2}} = \sqrt{240} = 15.49$

$x = number\ of\ success = 183$

1. Null Hypothesis H_0 : $The\ coin\ is\ unbiased$
2. Alternative Hypothesis H_1:$The\ coin\ is\ biased$
3. Level of significance $\alpha = 0.05$ (5%)
4. The test statistic is $Z = \frac{x-\mu}{\sigma} = \frac{183-480}{15.49} = -19.17$

$\therefore |Z| = 19.17$

As $|Z| > 1.96$, a null hypothesis H_0 to be rejected at 5% level of significance. Conclude that coin is biased.

2. A dice is tossed 960 times and it falls with 5 upwards 184 times. Is the dice unbiased at level of significance of 0.01?

Sol: Here $n = 960, p = probability\ of\ getting\ 5\ on\ dice\ is\ \frac{1}{6},\ q = \frac{5}{6}$

$\mu = np \Rightarrow \mu = 960 \times \frac{1}{6} = 160$

$\sigma = \sqrt{npq} = \sqrt{(np)q} = \sqrt{160 \times \frac{5}{6}} = 11.55$

$x = number\ of\ success = 184$

1. Null Hypothesis H_0 : $The\ coin\ is\ unbiased$
2. Alternative Hypothesis H_1:$The\ coin\ is\ biased$
3. Level of significance $\alpha = 0.01$
4. The test statistic is $Z = \frac{x-\mu}{\sigma} = \frac{184-160}{11.55} = 2.08$

$\therefore |Z| = 2.08$

As $|Z| < 2.58$, the null hypothesis H_0has to be accepted at 1% of level of significance and we conclude that dice is unbiased.

<u>Z – Tests :(LARGE SAMPLE TESTS)</u>

There are four large sample tests which are called Z – Tests:

Name of the test	**Null HypothesisH_0**	**Level of significance(α)**	**Test statistic**
1. Test for single mean	$\mu = \mu_0$	5% or 1% or10%	Z=$\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$
2.Test for difference of means	$\mu_1 = \mu_2$	5% or 1% or10%	Z= $\frac{\bar{x}-\bar{y}}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$
3.Test for single proportion	P=0.5	5% or 1% or10%	Z= $\frac{p-P}{\sqrt{\frac{PQ}{n}}}$
4. Test for difference of proportion	$P_1 = P_2$	5% or 1% or10%	Z=$\frac{P_1-P_2}{\sqrt{\frac{pq}{n}}}$

<u>Z-Table values (most important)</u>

Level of significance(α)	**Two tail(α)**	**One tail(2 α)**
5%	**1.96**	**1.645**
1%	**2.58**	**2.33**
2%	**2.33**	**-**
10%	**1.645**	**-**

TEST OF SIGNIFICANCE OF SINGLE MEAN:

Suppose we want to test, we are given a sample of size n has been drawn from population with mean μ, we set up null hypothesis $\bar{x} - \mu$

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

Population S.D. is not known $Z = \frac{\bar{x}-\mu}{S/\sqrt{n}}$

1. According to the norms established for a mechanical aptitude test, persons who are 18 years old have an average height of 73.2 with S.D. 8.6. If 64 randomly selected persons of that age averaged 76.7 test hypothesis μ = 73.2 against alternative hypothesis $\mu > 73.2\ at\ 0.01\ level$
$of\ significance\ on\ n = 4, \mu = 73.2, \bar{x} = 76.7, \sigma = 8.6$

Sol: Given $n = 64, \mu = 73.2, \bar{x} = 76.7, \sigma = 8.6$
1. Null hypothesis $H_0 = \mu = 73.2$
2. Alternative hypothesis$H_1, \mu > 73.2\ (Right\ tailed)$
3. Level of significance $\alpha = 0.01$
4. Test statistic $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{76.7-73.2}{8.6/\sqrt{64}} = 3.25$

Value of Z at 1%, level of significance is 2.33 (One tail).
Z > tabulated Z, Null hypothesis H_0 is rejected.

2. A sample of 64 students with mean weight of 70 kgs can this be regarded as a sample from a population with mean weight 56 kgs and S.D. 25 kgs.

Sol: $n = 64, \mu = 56\ kgs, \bar{x} = 70kgs, \sigma = 25kgs$
1. Null hypothesis H_0 : μ = 56.
2. Alternative hypothesis H_1 : μ≠ 56 (Two tailed)
3. Level of significance: 0.05 (Assumption)
4. Test statistic $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{70-56}{25/\sqrt{64}} = 4.48$; z – table = 1.96.

$|Z| > 1.96\ ; null\ hypothesis -\ H_0\ is\ rejected.$

3. An oceanographer want to check whether depth of the ocean in a certain region is 57.4 fathoms, as had previously been recorded. What can be concluded at the level of significance α = 0.05, if readings taken at 40 random location in the given region yielding a mean of 59.1 fathoms with a S.D. of 5.2 fathoms.

Sol: $n = 40, \mu = 57.4, \bar{x} = 59.1, \sigma = 5.2$
1. Null hypothesis $H_0: \mu = 57.4$
2. Alternative hypothesis $H_1: \mu \neq 57.4$ (Two tail)
3. Level of significance $\alpha = 0.05$
4. Test statistic $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{59.1-57.4}{5.2/\sqrt{40}} = 2.067$

Tabulated value of Z at 5% , level of significance is 1.96 (Two tail).
Hence calculated Z > tabulated Z. The null hypothesis H_0 is rejected.

4. In a random sample of 60 workers, the average time taken by them to get to work is 33.8 minute with a S.D. of 6.1 minutes. Can we reject the null hypothesis μ = 32.6 minutes in favour of alternative null hypothesis μ> 32.6 at α = 0.01 level of significance.

Sol: $n = 60, \mu = 32.6, \bar{x} = 33.8, \sigma = 6.1$
1. Null hypothesis $H_0: \mu = 32.6$
2. Alternative hypothesis $H_1: \mu > 32.6\ (Right\ tailed)$
3. Level of significance $\alpha = 0.01$
4. Test statistic $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{33.8-32.6}{6.1/\sqrt{60}} = 1.5238$

Tabulated value of Z at 0.01 level of significance is 2.33 (0.01 one tail).
Hence calculated Z < tabulated Z. The null hypothesis H_0 is accepted.

5. A sample of 400 items is taken from a population whose S.D. is 10. The mean of sample is 40. Test whether the sample has come from a population with mean 38. Also calculate 95% confidence interval for the population.

Sol: $n = 400, \mu = 38, \bar{x} = 40, \sigma = 10$

1. Null hypothesis $H_0: \mu = 38$
2. Alternative hypothesis $H_1: \mu \neq 38$ (Two tail)
3. Level of significance $\alpha = 0.05$
4. Test statistic $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{40-38}{10/\sqrt{400}} = 4$

$Z = 4 > 1.96$ (z – table = 1.96, 5% two tail)

We reject null hypothesis H_0, i.e., the sample is not from population whose mean is 38.

95% confidence interval is

$(\bar{X} - 1.96(\sigma/\sqrt{n}),\ \bar{X} + 1.96(\sigma/\sqrt{n})$

$= \left(40 - 1.96\left(\frac{10}{\sqrt{400}}\right),\ 40 + 1.96\left(\frac{10}{\sqrt{400}}\right)\right) = \left(40 - \frac{1.96}{20},\ 40 + \frac{1.96}{20}\right)$

$= (40 - 0.98, 40 + 0.98) = (39.02,\ 40.98)$

6. The mean and S.D. of population are 11795 and 14054 respectively if n = 50, find 95% confidence interval for the mean.

Sol: Given $n = 50, \bar{x} = 11795, \sigma = 14054$

95% confidence interval is

$(\bar{X} - 1.96(\sigma/\sqrt{n}),\ \bar{X} + 1.96(\sigma/\sqrt{n})$

$= \left(11795 - 1.96\left(\frac{14054}{\sqrt{50}}\right),\ 11795 + 1.96\left(\frac{14054}{\sqrt{50}}\right)\right)$

$= (11795 - 3895.6,\ 11795 + 3895.6)$

$= (7899.4,\ 15690.6)$

7. An ambulance service claims that it takes on average less than 10 minutes to reach its destination in emergency calls. A sample of 36 calls has a mean of 11 minutes and the variance of 16 minutes. Test the significance at 0.05 level.

Sol: Given $n = 36, \mu = 10, \bar{x} = 11, \sigma = \sqrt{16} = 4$

1. Null hypothesis $H_0: \mu = 10$
2. Alternative hypothesis $H_1: \mu \neq 10$ (Two tail)
3. Level of significance $\alpha = 0.05$
4. Test statistic $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{11-10}{4/\sqrt{36}} = \frac{6}{4} = 1.5$

Tabulated value of Z at 5% level of significance is 1.96 (Two tail).

Hence calculated Z < tabulated Z. We accept the null hypothesis H_0.

8. It is claimed that a random sample 49 tyres has a mean life of 15200 km. This sample was drawn from population whose mean is 15150 kms and S.D. 1200 kms. Test the significance at 0.05 level.

Sol: Given $n = 49, \mu = 15150, \bar{x} = 15200, \sigma = 1200$

1. Null hypothesis $H_0: \mu = 15150$
2. Alternative hypothesis $H_1: \mu \neq 15150$ (Two tail)
3. Level of significance $\alpha = 0.05$
4. Test statistic $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{15200-15150}{1200/\sqrt{49}} = 0.2917$ (z – table = 1.96, 5% two tail)

Since $|Z| < 1.96$, we accept the null hypothesis H_0.

TEST OF SIGNIFICANCE FOR DIFFERENCE OF MEANS OF TWO LARGE SAMPLES:

Let $\overline{x_1}$ be the mean of random sample of size n_1 drawn from a normal population with mean μ_1 and variance σ_1^2. Let $\overline{x_2}$ be the mean of an independent sample of size n_2 drawn from a population with mean μ_2 and variance σ_2^2.

To test whether significant difference between $\overline{x_1}$ *and* $\overline{x_2}$, we have to use statistic

$Z = \frac{(\overline{x_1}-\overline{x_2})-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$

Where $\delta = \mu_1 - \mu_2$

If $\delta = 0$, *the two populations have same means*

If $\delta \neq 0$, *the two populations are different*

Under $H_0: \mu_1 = \mu_2$, *statistic become* $Z = \frac{(\overline{x_1}-\overline{x_2})}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$

NOTE :

If the samples have been drawn from the population with common S.D. σ then

$\sigma_1^2 = \sigma_2^2 = \sigma^2$

Hence $Z = \frac{(\overline{x_1}-\overline{x_2})}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

PROBLEMS:

1. A sample of students were drawn from two universities and from their weight in kilograms, mean and standard deviation are calculated as shown below. Make test to test of significance of the difference between the means.

	Mean	**S.D.**	**Size of sample**
University A	55	10	400
University B	57	15	100

Sol: $\overline{x_1} = 55, \overline{x_2} = 57, n_1 = 400, n_2 = 100, S_1 = 10, S_2 = 15$

1. Null hypothesis $H_0: \overline{x_1} = \overline{x_2}$ there is no difference.

2. Alternative hypothesis $H_1, : \overline{x_1} \neq \overline{x_2}$

3. Level of significance $\alpha = 0.05$ $(assumed)$

4. Test statistic $Z = \frac{(\overline{x_1}-\overline{x_2})}{\sqrt{\frac{S_1^2}{n_1^2}+\frac{S_2^2}{n_2^2}}} = \frac{55-57}{\sqrt{\frac{100}{400}+\frac{225}{100}}} = -1.26$

Since $|Z| = 1.26 < 1.96$, we accept the null hypothesis H_0 at 5% level of significance.

2. The means of two large samples of sizes 1000 and 2000 members are 67.5 inches and 68.0 inches respectively. Can the samples be regarded as drawn from the same population of S.D. 2.5 inches?

Sol: Given $n_1 = 1000, n_2 = 2000$ *and* $\overline{x_1} = 67.5$ *inches*, $\overline{x_2} = 68$ *inches*

Population S.D., $\sigma = 2.5$ *inches*.

1. Null Hypothesis H_0: The samples have been drawn from the same population of S.D. 2.5 inches.

i.e., $\mu_1 = \mu_2$ *and* $\sigma = 2.5$ *inches*

2. Alternate Hypothesis $H_1: \mu_1 \neq \mu_2$ $(Two\ tail)$

3. α: 5%

4. The test statistic is, $Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} = \frac{67.5-68}{\sqrt{(2.5)^2\left(\frac{1}{1000}+\frac{1}{2000}\right)}} \Rightarrow Z = \frac{-0.5}{0.0968} = -5.16$

5. Conclusion: $|Z| = 5.16 > 1.96$

∴The null hypothesis H_0 is rejected at 5% level of significance. i.e., the samples are not drawn from the same population of S.D. 2.5 inches.

3. The mean yield of wheat from a district A was 210 pounds with S.D. 10 pounds per acre from a sample of 100 plots. In another district the mean yield was 220 pounds with S.D. 12 pounds from a sample of 150 plots. Assuming that the S.D. of the yield in the entire state was 11 pounds, test whether there is any significant difference between the mean yield of crops in the two districts.

Sol: Given $\overline{x_1} = 210, \overline{x_2} = 200, and\ n_1 = 100, n_2 = 150,\ population\ S.D., \sigma = 11$

1. Null Hypothesis H_0: There is no difference between $\overline{x_1} and\ \overline{x_2}\ i.e., H_0:\ \overline{x_1} = \ \overline{x_2}$
2. Alternate Hypothesis $H_1: \overline{x_1} \neq \ \overline{x_2}\ (Two\ tail)$
3. $\alpha : 5\%$
4. The test statistic is $Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} = \frac{210-200}{\sqrt{(11)^2\left(\frac{1}{100}+\frac{1}{150}\right)}} \Rightarrow Z = 7.04178$
5. Conclusion: $|Z| = 7.041 > 1.96$

∴The null hypothesis H_0 is rejected at 5% level of significance. i.e., there is a difference between $\overline{x_1} and\ \overline{x_2}$.

i.e., there is a significant difference between the mean yields of crops in the two districts.

4. In a survey of buying habits, 400 women shoppers are chosen at random per market 'A' located in a certain section of the city. Their average weekly food expenditure is Rs 250 with a S.D. of Rs 40. For 400 women shoppers chosen at random in super market 'B' in another section of the city, the average weekly food expenditure is Rs 220 with a S.D. of Rs 55. Test at 10% level of significance whether the average weekly food expenditure of the two populations of shoppers are equal.

Sol: Given $\overline{x_1} = Rs250, \overline{x_2} = 220, and\ n_1 = 400, n_2 = 400,\ S_1 = Rs\ 40\ and\ S_2 = Rs\ 55$

1. Null Hypothesis H_0:Assume that the average weekly food expenditure of the two populations of shoppers are equal i.e.,$H_0:\ \mu_1 = \mu_2$
2. Alternate Hypothesis $H_1:\ \mu_1 \neq \mu_2$
3. $\alpha : 5\%$
4. The test statistic is $Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{250-220}{\sqrt{\left(\frac{(40)^2}{400}+\frac{(55)^2}{400}\right)}} = \frac{30}{3.4} \Rightarrow Z = 8.82$
5. Conclusion: $|Z| = 8.82 > 2.58$

∴The null hypothesis H_0 is rejected. i.e., the average weekly food expenditure of the two populations of shoppers are not equal.

5. Two types of new cars produced in U.S.A. are tested for petrol mileage, one sample is consisting of 42 cars averaged 15 kmpl while the other sample consisting of 80 cars averaged 11.5 kmpl with population variances as $\sigma_1^2 = 2.0\ and\ \sigma_2^2 = 1.5\ respectively.$ Test whether there is any significance difference in the petrol consumption of these two types of cars (use α = 0.01)

Sol: Let the types of the cars be names as A and B.

Number of cars of type A = 42

Average mileage for $A = \overline{x_1} = 15, Variance = \ \sigma_1^2 = 2.0$

Number of cars of type B = 80

Average mileage for $B = \overline{x_2} = 11.5, Variance = \sigma_2^2 = 1.5$

1. Null Hypothesis $H_0: \mu_1 = \mu_2$
2. Alternate Hypothesis $H_1: \mu_1 \neq \mu_2$
3. α : 5%
4. The test statistic is $Z = \frac{|\overline{x_1} - \overline{x_2}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{|15-11.5|}{\sqrt{\left(\frac{2}{42} + \frac{1.5}{80}\right)}} = \frac{3.5}{\sqrt{0.0476+0.01875}} = \frac{3.5}{\sqrt{0.06635}} = 13.587$

$\Rightarrow Z = 13.587$

5. Conclusion: At 1% significance level,

Since $(z)_{calculate} > 2.58\,(z\ table)$, we reject Null hypothesis H_0at 1% level of significance and conclude that there is a significant difference in petrol consumption.

6. A simple sample of the height of 6400 Englishmen has a mean of 67.85 inches and a S.D. of 2.56 inches while a simple sample of heights of 1600 Austrians has a mean of 68.55 inches and S.D. of 2.52 inches. Do the data indicate the Austrians are on the average taller than the Englishmen? (use α as 0.01)

Sol: We are given

$n_1 = size\ of\ the\ first\ sample = 6400$

$n_2 = size\ of\ the\ second\ sample = 1600$

$\overline{x_1} = mean\ of\ the\ first\ sample = 67.85$

$\overline{x_2} = mean\ of\ the\ second\ sample = 68.55$

$\sigma_1 = Standard\ deviation\ of\ the\ first\ sample = 2.56$

$\sigma_2 = Standard\ deviation\ of\ the\ second\ sample = 2.52$

1. Null Hypothesis $H_0: \mu_1 = \mu_2$
2. Alternate Hypothesis $H_1: \mu_1 \neq \mu_2$
3. Level of significance :α : 0.05
4. The test statistic is

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{67.85-68.55}{\sqrt{\left(\frac{(2.56)^2}{6400} + \frac{(2.52)^2}{1600}\right)}} = \frac{-0.7}{\sqrt{\frac{6.5536}{6400} + \frac{6.35}{1600}}} = \frac{-0.7}{\sqrt{0.001+0.004}} = \frac{-0.7}{0.0707} = -9.9$$

$\Rightarrow Z = -9.9$

5. Conclusion:$\therefore |Z| = 9.9 > 1.96$

Hence, we reject the Null Hypothesis H_0 at 5% level of significance and conclude that the Austrians are taller than Englishmen.

7. The mean life of a sample of 10 electric bulbs (or motors) was found to be 1456 hours with S.D. of 423 hours. A second sample of 17 bulbs (motors) chosen from a different batch showed a mean life of 1280 hours with S.D. of 398 hours. Is there a significant difference between the means of two batches?

Sol: It is given that

$n_1 = sample\ size\ of\ the\ first\ batch = 10$

$n_2 = sample\ size\ of\ the\ second\ batch = 17$

$\overline{x_1} = mean\ life\ of\ first\ batch = 1456$

$\overline{x_2} = mean\ life\ of\ second\ batch = 1280$

$\sigma_1 = Standard\ deviation\ offirst\ batch = 423$

$\sigma_2 = Standard\ deviation\ of\ second\ batch = 398$

1. Null Hypothesis $H_0: \mu_1 = \mu_2$
2. Alternate Hypothesis $H_1: \mu_1 \neq \mu_2$
3. Level of significance :α : 0.05

4. The test statistic is

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{1456-1280}{\sqrt{\left(\frac{(423)^2}{10} + \frac{(398)^2}{17}\right)}} = \frac{176}{\sqrt{17892.9+9317.88}} = \frac{176}{164.96} = 1.067$$

$\Rightarrow Z = 1.067$

5. Conclusion: Since $Z < Z_\alpha = 1.96$,we accept the Null Hypothesis H_0i.e., no, there is no difference between the mean life of electric bulbs of two batches.

8. The average marks scored by 32 boys is 72 with a S.D. of 8. While that of 36 girls is 70 with a S.D. of 6. Does this indicate that the boys perform better than girls at level of significance 0.05

Sol: Let μ_1 and μ_2 be the means of the two populations.

1. Null Hypothesis $H_0: \mu_1 = \mu_2$
2. Alternate Hypothesis $H_1: \mu_1 > \mu_2$
3. $\alpha : 5\%$
4. The test statistic is

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Here $\bar{x} = 72,\ \bar{y} = 70, \sigma_1 = 8, \sigma_2 = 6, n_1 = 32, n_2 = 36$

$$\therefore Z = \frac{72-70}{\sqrt{\left(\frac{64}{32} + \frac{36}{36}\right)}} = \frac{2}{\sqrt{2+1}} = \frac{2}{\sqrt{3}} = 1.1547 < 1.96$$

5. Conclusion: Since the computed value of Z is less than the table value, we cannot reject the Null hypothesis at 5% level and conclude that the performance of boys and girls is the same.

9. At a certain large university a sociologist speculates that male students spend considerably more money on junk food than do female students. To test her hypothesis, the sociologist randomly selects from the registrar's records the names of 200 students. Of these, 125 are men and 75 are women. The sample mean of the average amount spent on junk food per week by the men is Rs 400 and standard deviation is 100. For the women the sample mean is Rs 450 and the sample standard deviation is Rs 150. Test the difference between the mean at 0.05 level.

Sol: Let μ_1 and μ_2 be the means of the two populations.

1. Let the Null Hypothesis $H_0: \mu_1 = \mu_2$
2. Then the Alternate Hypothesis $H_1: \mu_1 \neq \mu_2$
3. $\alpha : 5\%$

Let us assume that H_0 is true i.e., there is no difference between μ_1 and μ_2.

Given that$n_1 = Number\ of\ men = 125$

$n_2 = Number\ of\ women = 75$

$\overline{x_1} = Mean\ of\ men = 400$

$\overline{x_2} = Mean\ of\ women = 450$

$\sigma_1 = S.D. of\ men = 100$

$\sigma_2 = S.D. of\ women = 150$

Level of significance, $\alpha = 0.05$

4. The test statistic is

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{400-450}{\sqrt{\frac{(100)^2}{125} + \frac{(150)^2}{75}}} = \frac{-50}{\sqrt{80+300}} = \frac{-50}{\sqrt{380}} = \frac{-50}{19.49} = -2.5654$$

5. Conclusion: $\therefore |Z| = 2.5654 > 1.96$ i.e., the difference is highly significant.
Hence, we reject the Null Hypothesis at 5% level of significance and conclude that the two population means are not equal.

10. A company claims that its bulbs are superior to those of its main competitor. If a study showed that a sample of 40 of its bulbs have a mean life time of 647 hrs of continuous use with a S.D. of 27 hours. While a sample of 40 bulbs made by its main competitor had a mean life time of 638 hrs of continuous use with a S.D. of 31 hrs. Test the significance between the difference of two means at 5% level.

Sol: Let μ_1 and μ_2 be the means of the two populations.
1. Let the Null Hypothesis $H_0: \mu_1 = \mu_2$
2. Then the Alternate Hypothesis $H_1: \mu_1 > \mu_2$
3. α : 5%
4. Since the sample sizes are large, the test statistic

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is approximately normally distributed with mean 0 and S.D. 1.
Here $\bar{x} = 647,\ \bar{y} = 638, \sigma_1 = 27, \sigma_2 = 31, n_1 = n_2 = 40$

$$\therefore Z = \frac{647 - 638}{\sqrt{\left(\frac{(27)^2}{40} + \frac{(31)^2}{40}\right)}} = \frac{9}{\sqrt{\frac{729+961}{40}}} = \frac{9}{6.5} = 1.38$$

5. Conclusion: $\therefore |Z| = 1.38 < 1.645$
Since the computed value of Z is less than the table value, we cannot reject the Null Hypothesis at 5% level and conclude that the difference between the two sample means is not significant.

11. Studying the flow of traffic at two busy intersections between 4 pm and 6 pm to determine the possible need for turn signals. It was found that on 40 week days there were on the average 247.3 cars approaching the first intersection from the south which made left turn, while on 30 week days there were on the average 254.1 cars approaching the first intersection from the south made left turns.
The corresponding sample standard deviations are 15.2 and 12. Test the significance between the difference of two means at 5% level.

Sol: Let the average cars in two places beμ_1 and μ_2 respectively.
1. Let the Null Hypothesis $H_0: \mu_1 = \mu_2$
2. Then the Alternate Hypothesis $H_1: \mu_1 \neq \mu_2$
3. α : 5%
Let us assume that H_0 is true i.e., there is no significant difference between μ_1 and μ_2
4. Since the sample sizes are large, the test statistic

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Here $\overline{x_1} = 247.3,\ \overline{x_2} = 254.1, \sigma_1 \cong s_1 = 15.2, \sigma_2 \cong s_2 = 12, n_1 = 40,\ n_2 = 30$

$$\therefore Z = \frac{247.3 - 254.1}{\sqrt{\left(\frac{(15.2)^2}{40} + \frac{(12)^2}{30}\right)}} = \frac{-6.8}{\sqrt{5.776+4.8}} = \frac{-6.8}{\sqrt{10.576}} = \frac{-6.8}{3.2521} = -2.091$$

5. Conclusion: $\therefore |Z| = 2.091 > 1.96$
Since the computed value of Z is greater than the table value, we reject the Null Hypothesis at 5% level and conclude that the two average cars are significantly different i.e., they are not the same in the two busy intersections.

12. In a certain factory there are two independent processes for manufacturing the same item. The average weight in a sample of 700 items produced from one process is found to be 250 gms with a standard deviation of 30 gms while the corresponding figures in a sample of 300 items from the other process are 300 and 40. Is there significant difference between the mean at 1% level.

Sol: Let the average weight in the two independent processes be μ_1 *and* μ_2 respectively.

1. Let the Null Hypothesis $H_0: \mu_1 = \mu_2$
2. Then the Alternate Hypothesis $H_1: \mu_1 \neq \mu_2$
3. α : 1%

Let us assume that H_0 is true i.e., there is no significant difference between μ_1 *and* μ_2

4. The test statistic

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Here $\overline{x_1} = 250,\ \overline{x_2} = 300, \sigma_1 = 30, \sigma_2 = 40, n_1 = 700,\ n_2 = 300$

$$\therefore Z = \frac{250-300}{\sqrt{\left(\frac{900}{700} + \frac{1600}{300}\right)}} = \frac{-50}{\sqrt{\frac{9}{7} + \frac{16}{93}}} = -19.43$$

5. Conclusion: $\therefore |Z| = 19.43 > 2.58$

Since the computed value of Z is greater than the table value, we reject the Null Hypothesis at 1% level and conclude that there is a significant difference between the means.

13. The mean height of 50 male students who participated in sports is 68.2 inches with a S.D. of 2.5. The mean height of 50 male students who have not participated in sport is 67.2 inches with a S.D. of 2.8. Test the hypothesis that the height of students who participated in sports is more than the students who have not participated in sports.

Sol: Let the mean height in the two cases be μ_1 *and* μ_2 respectively.

1. Let the Null Hypothesis $H_0: \mu_1 = \mu_2$
2. Then the Alternate Hypothesis $H_1: \mu_1 \neq \mu_2$
3. α : 1%

Let us assume that H_0 is true i.e., there is no significant difference between μ_1 *and* μ_2

4. The test statistic is

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Here $\overline{x_1} = 68.2,\ \overline{x_2} = 67.2, \sigma_1 = 2.5, \sigma_2 = 2.8, n_1 = 50,\ n_2 = 50$

$$\therefore Z = \frac{68.2-67.2}{\sqrt{\left(\frac{(2.5)^2}{50} + \frac{(2.8)^2}{50}\right)}} = \frac{1}{\sqrt{\frac{6.5+7.84}{50}}} = 1.88$$

5. Conclusion: $\therefore |Z| = 1.88 < 1.96$

Therefore accept H_0.

TESTING OF HYPOTHESIS – II

TEST OF SIGNIFICANCE FOR SINGLE PROPORTION:

Suppose a large sample of size n taken from a normal population. To test the significant difference between sample proportion p and population proportion P we write the statistic

$Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}}$ $n\ is\ sample\ size, P = \frac{x}{n}, Q = 1 - P$

NOTE:

1. Limits for population proportion P are given by $p \pm 3\sqrt{\frac{pq}{n}}$

2. Confidence interval for proportion P for large samples at α. Level of significance is $\left(P - Z_{\frac{\alpha}{2}}\sqrt{\frac{PQ}{n}}\ ,\ P + Z_{\frac{\alpha}{2}}\sqrt{\frac{PQ}{n}}\right)$; $Q = 1 - P$

PROBLEMS:

1. A manufacture claimed that at least 95% of the equipment which he supplied to a factory conformed to specifications. An examination of a sample 200 pieces of equipment revealed that 18 were faulty. Test his claim at 5% level of significance.

Sol: Given sample size n = 200

No. of pieces confirming to specifications = $200 - 18 = 182$.

p = Proportion of pieces conforming specification $= \frac{182}{200} = 0.91$

P = proportion of population $= \frac{95}{100} = 0.95$

Q = $1 - P = 1 - 0.95 = 0.05$

1. Null hypothesis $H_0: P = 95\%$.
2. Alternative hypothesis $H_1: P < 0.95$
3. Level of significance = 0.05
4. Test statistic $Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.91-0.95}{\sqrt{0.95\times\frac{0.005}{200}}} = -2.59$

Since in alternative hypothesis is left tailed, tabulated value of Z at 5% level of significance is 1.645. $|Z| = 2.59 > 1.645$, we reject H_0.

2. In a sample of 1000 people in Karnataka 540 are rice eater and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state at 1% level.

Sol: Given sample size n = 1000

p = Sample proportion of rice eaters$= \frac{540}{1000} = 0.54$

P = rice eaters$= \frac{1}{2}$; Q = $1 - P = 1 - \frac{1}{2} = \frac{1}{2}$

1. $H_0: Both\ rice\ and\ wheat\ eaters\ are\ equal$
2. $H_1 : p \neq 0.5$
3. Level of significance : 0.01
4. $Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.54-0.5}{\sqrt{0.5\times\frac{0.5}{1000}}} = 2.532$

$|Z| = 2.532 < 2.58$, ∴ H_0 accepted i.e., both rice and wheat are equally popular in the State.

3. In a big city, 325 men out of 600 men were found to be smokers. Does this information support the conclusion that the majority of men in this city are smokers?

Sol: Given n = 600

No. of smokers= 325.

p = Sample proportion $= \frac{325}{600} = 0.5417$

P = Proportion of smokers in city$= \frac{1}{2} = 0.5$

$Q = 1 - P = 1 - 0.5 = 0.5$

1. Null hypothesis $H_0 : Smokers = Non\ smokers\ P = 0.5$
2. Alternative hypothesis $H_1 : p > 0.5\ (right\ tailed)$
3. $\alpha = 0.05$
4. $Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.5417-0.5}{\sqrt{0.5\times\frac{0.5}{600}}} = 2.04$

$|Z| = 2.04 > 1.645$, ∴ H_0 is rejected.

4. A dice was thrown 9000 times and of these 3220 are 3 or 4 in this consistent with that the dice was unbiased?

Sol: Given n = 9000

p = Proportion of success of getting 3 or 4 in 9000 throws $= \frac{3220}{9000} = 0.3578$

P = Population proportion of success$= \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3} = 0.333$

$Q = 1 - P = 1 - 0.333 = 0.667$

1. $H_0 : The\ dice\ is\ unbiased$
2. $H_1 : p \neq \frac{1}{3}\ (two\ tailed\ biased)$
3. Level of significance$= 0.05$
4. $Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = 4.94$

$|Z| > 1.96$, ∴dice is biased (accept H_1).

5. In a random sample of 12.5 cool drinkers, 68 said they prefer thums up to pepsi test null hypothesis P = 0.5 against alternative hypothesis p > 0.5

Sol: Given n = 125, x = 68, $p = \frac{x}{n} = \frac{68}{125} = 0.544$

$P = \frac{1}{2}$; $Q = 1 - P = 1 - \frac{1}{2} = \frac{1}{2}$

1. Null hypothesis$H_0 : p = 0.5$
2. $H_1 : p > 0.5$
3. Level of significance $= 0.05$
4. $Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = 0.9839 \Rightarrow |Z| < 1.645$, ∴accept H_0.

6. Experience had shown that 20% of a manufactured product is of the top quality. In one day production of 400 articles only 50 are top quality. Test the hypothesis at 0.05 level.

Sol: Given n = 400, x = 50, $p = \frac{x}{n} = \frac{50}{400} = 0.125$

$P = 0.2$;

$Q = 1 - P = 1 - 0.2 = 0.8$

1. $H_0 : p = 0.2$
2. $H_1 : p \neq 0.2$
3. $\alpha = 0.05$
4. $Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = -3.75$

$|Z| = 3.75 > 1.96$, ∴reject null hypothesis H_0.

7. Among 900 people in a state 90 are found to be chapati eaters. Construct 99% confidence interval for the population.

Sol: $x = 90, n = 900, P = \frac{90}{900} = \frac{1}{10} = 0.1; Q = 0.9$

Now $\sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.1\times0.9}{900}} = 0.01$

Confidence interval is $\left(P - Z_{\frac{\alpha}{2}}\sqrt{\frac{PQ}{n}}\ ,\ P + Z_{\frac{\alpha}{2}}\sqrt{\frac{PQ}{n}}\right)$

$= (0.1 - 0.03, 0.1 + 0.03) = (0.07, 0.13)$

8. In a random sample of 160 workers exposed to a certain amount of radiation, 24 experienced some ill effects. Construct 99% confidence interval corresponding to true percentage.

Sol: $x = 24, n = 160, P = \frac{24}{160} = 0.15; Q = \ 0.85$

Now $\sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.15\times0.85}{160}} = 0.028$

Confidence interval is $\left(P - Z_{\frac{\alpha}{2}}\sqrt{\frac{PQ}{n}}\ ,\ P + Z_{\frac{\alpha}{2}}\sqrt{\frac{PQ}{n}}\right)$

$= (0.15 - 3\times 0.028, 0.15 + 3\times 0.028) = (0.065, 0.234)$

9. 20 people were attacked by a disease and only 18 survived. Will you reject the hypothesis that survival rate if attacked by this disease 85% in favour of the hypothesis that is more at 5% level.

Sol: Given n = 20, x = 18, $p = \frac{x}{n} = \frac{18}{20} = 0.9$

$\text{P} = 0.85;\ \text{Q} = \ 1 - P = 1 - 0.85 = 0.15$

1. $H_0 : P = 0.85$
2. $H_1 : p > 0.85$
3. $\alpha = 0.05$
4. $Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = 0.625$

$|Z| = 0.625 < 1.645$, ∴accept H_0.

10. A social worker believes that fewer than 25% of the couples in a certain area have ever used any form of birth control. A random sample of 120 couples was contacted. Twenty of them said that they have used test the belief of the social worker at 0.05 level.

Sol: Given n = 120, x = 120, $p = \frac{x}{n} = \frac{20}{120} = \frac{1}{6}$

$\text{P} = 0.25;\ \text{Q} = \ 1 - P = 1 - 0.25 = 0.75$

1. Null hypothesis $H_0 : P = 0.25\ (left\ tailed\ test)$
2. Alternative hypothesis$H_1 : p < 0.25$
3. Level of significance $\alpha = 0.05$
4. The test statistic is $Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{\frac{1}{6}-0.25}{\sqrt{\frac{0.25\times0.75}{120}}} = -2.107$

$|Z| = 2.107 > 1.96$, ∴we reject null hypothesis H_0.

11. A manufacture claims that only 4% of his products are defective. A random sample of 500 were taken among which 100 were defective test hypothesis at 0.05 level.

Sol: Given n = 500, x = 100, $p = \frac{x}{n} = 0.2$

$\text{P} = 0.04;\ \text{Q} = \ 1 - P = 1 - 0.04 = 0.96$

1. Null hypothesis $H_0 : P = 0.04$
2. Alternative hypothesis $H_1 : p > 0.04\ (right\ tailed\ test)$
3. Level of significance $\alpha = 0.05$

4. The test statistic is $Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.2-0.04}{\sqrt{\frac{0.04\times0.96}{500}}} = -\frac{0.16}{0.00876} = -18.26$

$|Z| = 18.26 > 1.645$, $\therefore$ we reject the null hypothesis H_0.

DIFFERENCE OF PROPORTION:

Suppose two large samples of size $n_1\ and\ n_2$ are taken respectively from two different populations. To test the significant difference between sample proportions $p_1\ and\ p_2$

$Z = \frac{p_1-p_2}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\ \ where\ p = \frac{n_1p_1+n_2p_2}{n_1+n_2} = \frac{x_1+x_2}{n_1+n_2};\ \ q = 1-p$

TEST OF SIGNIFICANCES:

1. If $|Z| < 1.96\ accept\ , difference\ is\ not\ significant\ at\ 5\%\ level$
2. If $|Z| > 1.96\ reject\ at\ 5\%\ level$
3. If $|Z| > 2.58\ accept\ at\ 1\%\ level$

PROBLEMS:

1. A manufacturer of electronic equipment subjects samples two completing brands of transistors to an accelerated performance test. If 45 of 180 transistors of the first kind and 34 of 120 transistors of the second kind fail the test what can be conclude at the level of significance 0.05 about difference between corresponding sample proportion.

Sol: We have $n_1 = 180, x_1 = 45, x_2 = 34, n_2 = 120$

$p_1 = \frac{x_1}{n_1} = \frac{45}{180} = 0.25,\ p_2 = \frac{x_2}{n_2} = \frac{34}{120} = 0.283$

$p = \frac{n_1p_1+n_2p_2}{n_1+n_2} = \frac{x_1+x_2}{n_1+n_2} = \frac{45+34}{180+120} = \frac{79}{300} = 0.263$

$q = 1-p = 1-0.263 = 0.737$

1. Null hypothesis $H_0: p_1 = p_2$
2. Alternative hypothesis $H_1: p_1 \neq p_2$ (Two tail)
3. Level of significance $\alpha = 0.05$
4. The test statistic is $Z = \frac{p_1-p_2}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = 0.647$ (z – table = 1.96 (Two tail) 5%)

$|Z| < 1.96$, $\therefore$ we accept the null hypothesis H_0.

2. Two large populations there are 30%, 25% respectively of fair haired people. Is this difference likely to be hidden in sample of 1200 and 900 respectively from the population.

Sol: $n_1 = 1200,\ n_2 = 900$

$p_1 = proportion\ of\ fair\ haired\ people\ in\ the\ first\ population = \frac{30}{100} = 0.3,$

$p_2 = proportion\ of\ fair\ haired\ people\ in\ the\ second\ population = \frac{25}{100} = 0.25$

1. Null hypothesis $H_0: p_1 = p_2$
2. Alternative hypothesis $H_1: p_1 \neq p_2$ (Two tail)
3. The test statistic is $Z = \frac{p_1-p_2}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}}$

$q_1 = 1-p_1 = 1-0.3 = 0.7$

$q_2 = 1-p_2 = 1-0.25 = 0.75$

$Z = 2.55$

Since $Z > 1.96$, $\therefore$ we reject the null hypothesis H_0.

NOTE:

Suppose the population proportions $p_1\ and\ p_2$ are given and $p_1 \neq p_2$. If we want to test the hypothesis that the differences $(P_1 - P_2)$ in population proportions is likely to be hidden in simple samples of size $n_1\ and\ n_2$ from populations respectively then

$Z = \frac{(p_1-p_2)-(P_1-P_2)}{\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}}$

If sample proportion is not known

$|Z| = \frac{(P_1-P_2)}{\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}}$

PROBLEMS:

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

Sol: $n_1 = 1200,\ n_2 = 900$

$p_1 = proportion\ of\ fair\ haired\ people\ in\ the\ first\ population = \frac{30}{100} = 0.3,$

$p_2 = proportion\ of\ fair\ haired\ people\ in\ the\ second\ population = \frac{25}{100} = 0.25$

$q_1 = 1 - p_1 = 1 - 0.3 = 0.7$

$q_2 = 1 - p_2 = 1 - 0.25 = 0.75$

1. Null hypothesis $H_0: p_1 = p_2$

2. Alternative hypothesis $H_1 : p_1 \neq p_2$ (Two tail)

3. α= 0.05

4. The test statistic is $Z = \frac{p_1-p_2}{\sqrt{\frac{p_1q_1}{n_1}+\frac{p_2q_2}{n_2}}} = \frac{0.05}{0.0195} = 2.55$

$Z = 2.55$; Since $Z > 1.96$, ∴ we reject the null hypothesis H_0.

2. On the basis of their total scores 200 candidates of a civil service examination are divided into two groups the upper 30% and the remaining 70%. Consider the first question of examination among the first group, 40 had the correct answer, whereas among second group 80 had the correct answer. On the basis of these results, can one conclude that the first question is not good at discriminating ability of the type being examined.

Sol: $n_1 = 60,\ n_2 = 140, x_1 = 40, x_2 = 80$

$p_1 = \frac{x_1}{n_1} = \frac{40}{60} = \frac{2}{3} = 0.667,$

$p_2 = \frac{x_2}{n_2} = \frac{80}{140} = 0.571$

$P = \frac{40+80}{60+140} = 0.6$

$Q = 1 - P = 1 - 0.6 = 0.4$

1. Null hypothesis $H_0: p_1 = p_2$

2. Alternative hypothesis $H_1 : p_1 \neq p_2$ (Two tail)

3. Level of significance = 0.05 (assumed)

4. The test statistic is $Z = \frac{p_1-p_2}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.096}{0.0750} = 1.27$

$|Z| < 1.96$, ∴ we accept the null hypothesis H_0 at 5% level of significance, i.e., first question is good.

3. A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Find the percentage of bad pineapples in the consignment.

Sol: $n = 500, p = \frac{65}{500} = 0.13; q = 1 - p = 0.87$

We know that the limits for population proportion

P are given by $\left(P \pm 3\sqrt{\frac{pq}{n}}\right)$

$= 0.13 \pm 3\sqrt{\frac{0.13\times0.87}{500}} = (0.13 \pm 0.045) = (0.085, 0.175)$

Percentage of bad pineapples lies between 8.5 and 17.5.

PRACTICE PROBLEM:

4. Random samples of 400 men and 600 women were asked whether they would like to have a fly over near their residence. 200 men and 325 women were in favour of the proposal test the hypothesis that the proportion of men and women in favour of proposal.

Sol: $p_1 = \frac{200}{400} = 0.5$; $p_2 = \frac{325}{600} = 0.541, P = 0.525, Q = 0.475$

$Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}}$; $Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$; $Z = \frac{(p_1-p_2)-(P_1-P_2)}{\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}}$

5. A cigarette manufacturing firm claims that its brand, A line of cigarettes outsells its brand B by 8%. If it found that 42 out of a sample 200 smokers prefer brand A and 18 out of another sample of 100 smokers prefer brand B, test whether the 8% difference is a valid claim.

Sol: Given $n_1 = 200, n_2 = 100, P_1 = \frac{42}{200} = 0.21,\ P_2 = \frac{18}{100} = 0.18$

$P_1 - P_2 = 8\% = 0.08,\ p = \frac{42+18}{300} = 0.2;\ \ q = 1 - 0.2 = 0.8$

1. Null hypothesis $H_0: P_1 - P_2 = 0.08$. Assume 8% difference in the sale of two brands of cigarettes is valid claim.
2. Alternative hypothesis $H_1: P_1 - P_2 \neq 0.08$ (Two tail)
3. Level of significance 0.05 or 5%.
4. $Z = \frac{(p_1-p_2)-(P_1-P_2)}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.03-0.08}{\sqrt{0.2\times0.8\left(\frac{1}{200}+\frac{1}{100}\right)}} = -\frac{0.05}{0.0489} = -1.02$

$|Z| = 1.02 < 1.96$ So, accept null hypothesis H_0.

Key points:

Z – Tests:

There are four large sample tests which are called Z – Tests:

Name of the test	Null HypothesisH_0	Level of significance(α)	Test statistic
1. Test for single mean	$\mu = \mu_0$	5% or 1% or10%	$Z=\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$
2.Test for difference of means	$\mu_1 = \mu_2$	5% or 1% or10%	$Z= \frac{\bar{x}-\bar{y}}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$
3.Test for single proportion	P=0.5	5% or 1% or10%	$Z= \frac{p-P}{\sqrt{\frac{PQ}{n}}}$
4. Test for difference of proportion	$P_1 = P_2$	5% or 1% or10%	$Z=\frac{P_1-P_2}{\sqrt{\frac{pq}{n}}}$

For every test by default α is 5%.

Z-Table values (most important)

Level of significance(α)	Two tail(α)	One tail(2 α)
5%	**1.96**	**1.645**
1%	**2.58**	**2.33**
2%	**2.33**	**-**
10%	**1.645**	**-**

PRACTICE PROBLEMS:

TEST FOR SINGLE MEAN:

1. An ambulance service claims that it takes on the average 8.9 minutes to reach its destination in emergency calls. To check this claim the agency which licences ambulance services has them timed on 50 emergency calls getting mean of 9.3 minutes with a S.D. of 1.6 minutes what can they conclude at level of significance 0.05?

2. A labour management discussion it was brought up that workers at a certain large plant take on the average 32.6 minutes to get to work. If a random sample of 60 workers took on the average 33.8 minutes with a standard deviation of 6.1 minutes can we reject the null hypothesis $\mu > 32.6$ in favour of the alternative hypothesis $\mu > 32.6$

TEST OF DIFFERENCE OF MEANS:

3. To test the claim that the resistance of electric wire can be reduced by more than 0.050 ohm by alloying 32 values obtained for standard wire yielded a mean of 0.136 ohm and a S.D. of 0.004 ohm and 32 values obtained for alloyed wire yielded a mean of 0.083 ohm with a S.D. of 0.005 ohm. At the 0.05 level of significance does this support claim.

4. The average marks stored by 32 boys is 72 with a S.D. of 8, while that for 36 girls is 70 with a S.D. of 6. Does this indicate that the boys perform better than girls at level of significance 0.05?

TEST FOR SINGLE PROPORTION:

5. A study designed to investigate whether certain detonators used with explosives in coal mining meet the requirement that atleast 90% will ignite the explosive when charged. It is found that 174 of 200 detonators function properly. Test the null hypothesis $P = 0.9$ against the alternative hypothesis $P \neq 0.9$ at the 0.05 level of significance.

6. 20 people were attacked by disease and only 18 survived. Will you reject the hypothesis that the survival rate if attacked by this disease is 85% in favour of the hypothesis that is more at 5% level.

TEST FOR DIFFERENCE OF MEANS:

7. Among the items produced by a factory out of 800, 65 were defective in another sample out of 300, 40 were defective, test the significance between the differences of two proportions at 1% level?

8. If 120 out of 200 patients suffering from a certain disease are cured by alopathy and 240 out of 500 patients are cured by homeopathy, is there reason enough to believe that allopathy is better than in curing disease? Use $\alpha = 0.05$ level of significance?

****