

MODULE-5-STATISTICAL INFERENCE

Random Sampling

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

A finite subset of the universe is called a *sample*. The number of individuals in a sample is called a *sample size*(n).

If $n \leq 30$ then the sample is said to be *small* otherwise it is a *large sample*.

The selection of an individual or item from the population in such a way that each has the same chance of being selected is called as *random sampling*.

Sampling where a member of the population may be selected more than once is called as *sampling with replacement*.

If a member cannot be chosen more than once is called as *sampling without replacement*.

Parameter and Statistics

The statistical constants of population is mean(μ), standard deviation (σ), correlation (ρ) are called the *parameter*.

The statistical constants drawn from the given population is mean($\bar{x}$), standard deviation (s), correlation (r) are called *statistic*.

Testing of Hypothesis

To reach the decisions about the populations and the basis of sample information, we make certain assumptions about the population involved. Such assumptions which may or may not be true are called *hypothesis*. **OR** Quantitative statement about the population. Explanation on the basis of limited evidences.

Null Hypothesis

The hypothesis formulated for the purpose of its rejection under the assumption that it is true is called the *Null Hypothesis* denoted by H_0 .

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

Type-I and Type-II Errors

Type-I Error: Wrong decision to reject the null hypothesis when it is actually true.

Type-II Error: Wrong decision to accept the null hypothesis when it is actually false.

	Accepting the hypothesis	Rejecting the hypothesis
--	--------------------------	--------------------------

Hypothesis True	Correct decision	Wrong decision (Type-I Error)
Hypothesis False	Wrong decision (Type-II Error)	Correct decision

Significance level

The probability level, below which leads to the rejection of the hypothesis is known as the *significance level*. The probability levels are fixed at 0.05 or 0.01 being 5% or 1%. These are called *Significance level*.

Test of Significance

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

Confidence Intervals

Let us suppose that we have a normal population with mean μ and S.D σ . If $\bar{x}$ is the sample

mean of a random sample size n the quantity z defined by $z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$ is called the *Standard normal variate*.

From the normal distribution table we find the 95% of the area lies between $z = -1.96$ and $z = +1.96$ i.e., 95% confidence interval lies between -1.96 and $+1.96$.

Further 5% level of significance is denoted by $z_{0.05}$. Mathematically we can write

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

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

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

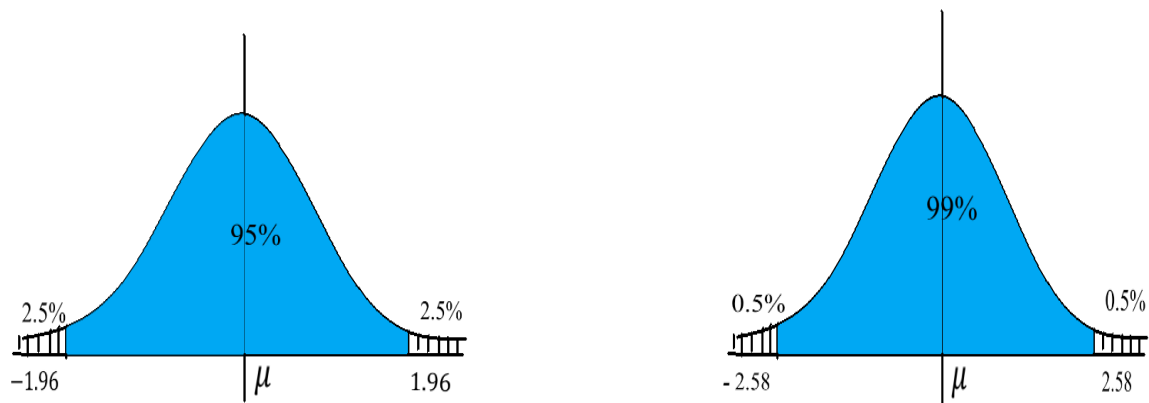
Also we can write combining the two results in the form,

$$\bar{x} - 1.96 \left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + 1.96 \left(\frac{\sigma}{\sqrt{n}}\right) \quad \text{--- (1)}$$

Similarly, from the table of normal areas 99% of the area lies between -2.58 and 2.58. This is equivalent to the form,

$$\bar{x} - 2.58 \left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + 2.58 \left(\frac{\sigma}{\sqrt{n}}\right) \quad \text{--- (2)}$$

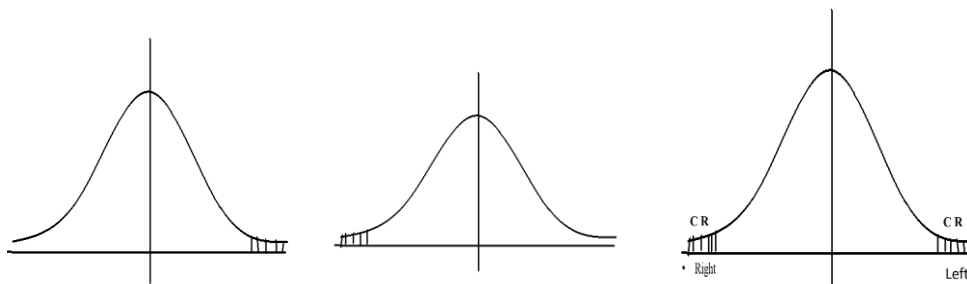
We can say that (1) is the 95% confidence interval and (2) is the 99% confidence interval.



One tailed and two tailed tests

In our test of acceptance or non acceptance of a hypothesis, we concentrated on the value of z on both sides of the mean. This can be categorically stated that, the focus of the attention lies in the two “tails” of the distribution and hence such a test is called a *two tailed test*.

Sometimes we will be interested in the extreme values to only one side of the mean in which the region of significance will be a region to one side of the distribution. Obviously the area of such a region will be equal to the level of significance itself. Such a tail is called a *one tailed test*.



One tailed test

Two tailed test

The following table will be useful for working problems

Test	Critical Value of z	
	5% level	1% level
One-tailed test	-1.645 or 1.645	-2.33 or 2.33
Two-tailed test	-1.96 or 1.96	-2.58 or 2.58

Tests of significance of Proportions

1. Let x be the observed number of successes in a sample size of n and $\mu = np$ be the expected number of successes let the associated standard normal variate Z be defined by $z = \frac{x-\mu}{\sigma} = \frac{x-np}{\sqrt{npq}}$

2. If $|Z| > 2.58$, we conclude that the differences is highly significant and reject the hypothesis .Probable limits $= p \pm 2.58 \sqrt{\frac{pq}{n}}$

3. If $Z < 1.96$, difference between the observed and expected number of successes is not significant.

4. If $Z > 1.96$, difference is significant at 5% level of significance.

Test of hypothesis for means

Let μ_1 and μ_2 be the mean of two populations. Let $(\bar{x}_1, \sigma_1)$; $(\bar{x}_2, \sigma_2)$ be the mean and standard deviation of two large samples of size n_1 and n_2 respectively. To test the null hypothesis H_0 that there is no difference between the population means. i.e., $H_0: \mu_1 = \mu_2$

The static for this is given by $Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

Test of significance of proportions

1. A coin is tossed 1000 times and head turns up 540 times. Decide on the hypothesis that the coin is unbiased.

Sol: Let us suppose that the coin is unbiased

p = Probability of getting a head in one toss = $\frac{1}{2}$

Since $p + q = 1$, $q = \frac{1}{2}$

Expected number of heads in 1000 tosses $= np$

$$= 1000 \times \frac{1}{2} = 500$$

Actual Number of heads $= 540 = x$

Now $x - np = 540 - 500 = 40$

Consider $z = \frac{x-\mu}{\sqrt{npq}} = \frac{x-np}{\sqrt{npq}} = \frac{540-500}{\sqrt{1000 \times 0.5 \times 0.5}}$

$$z = 2.53 < 2.58$$

Thus, we can say that the coin is unbiased.

2. A sample of 900 days was taken in a coastal town and it was found that on 100 days the weather was very hot. Obtain the probable limits of the percentage of very hot weather.

Sol: Probability of very hot weather, $p = \frac{100}{900} = \frac{1}{9}$

$$\therefore q = \frac{8}{9}, n = 900$$

$$\begin{aligned}\text{Probable limits} &= p \pm (2.58)\sqrt{\frac{pq}{n}} \\ &= \frac{1}{9} \pm (2.58)\sqrt{\frac{1}{9} \cdot \frac{8}{9} \cdot \frac{1}{900}} \\ &= 0.111 \pm 0.027 \\ &= 0.084 \text{ and } 0.138\end{aligned}$$

Probability limits of very hot weather is 8.4% to 13.8%

3. In a sample of 500 men it was found that 60% of them had over weight. What can we infer about the proportion of people having over weight in the population?

Sol: Probability of persons having over weight is , $p = \frac{60}{100} = 0.6$

$$\therefore q = 1 - p = 0.4, n = 500$$

$$\begin{aligned}\text{Probable limits} &= p \pm (2.58)\sqrt{\frac{pq}{n}} \\ &= 0.6 \pm (2.58)\sqrt{\frac{(0.6)(0.4)}{500}} \\ &= 0.6 \pm 0.0565\end{aligned}$$

Probable limits = 0. 5435 and 0.6565

Thus, the probable limits of people having over weight is 54.35% and 65.65%

4. A survey was conducted in a slum locality of 2000 families by selecting a sample size 800. It was revealed that 180 families were illiterates. Find the probable limits of the illiterate families in the population of 2000.

Sol: Probability of illiterate families is , $p = \frac{180}{800} = 0.225$

$$\therefore q = 1 - p = 0.775, n = 800$$

$$\begin{aligned}\text{Probable limits} &= p \pm (2.58)\sqrt{\frac{pq}{n}} \\ &= 0.225 \pm (2.58)\sqrt{\frac{(0.225)(0.775)}{800}} \\ &= 0.225 \pm 0.038\end{aligned}$$

Probable limits = 0. 187 and 0.263

The probable limits of the illiterate families in the population of 2000

$$= 0.187 * 2000 \text{ and } 0.263 * 2000$$

= 374 and 526

Thus, 374 to 526 are probably illiterate families.

5. In 324 throws of a six faced 'die', an odd number turned up 181 times. Is it reasonable to think that the 'die' is an unbiased one?

Sol: Probability of the turn up of an odd number , $p = 3/6 = \frac{1}{2}$

Since $p + q = 1$, $q = \frac{1}{2}$

Expected number of successes = np

$$= 324 \times \frac{1}{2} = 162$$

Actual Number of heads = 181 = x

Now $x - np = 181 - 162 = 19$

$$\text{Consider } z = \frac{x - \mu}{\sqrt{npq}} = \frac{x - np}{\sqrt{npq}} = \frac{19}{\sqrt{324 \times 0.5 \times 0.5}} = \frac{19}{9}$$

$$z = 2.11 < 2.58$$

Thus, we can say that the die is unbiased.

6. A sample of 100 days is taken from meteorological records of a certain district and 10 of them are found to be foggy. What are the probable limits of the percentage of foggy days in the district.

Sol: Proportion of foggy days in a sample of 100 days is , $p = \frac{10}{100} = 0.1$

$$\therefore q = 1 - p = 0.9 \quad , n = 100$$

$$\begin{aligned} \text{Probable limits} &= p \pm (2.58) \sqrt{\frac{pq}{n}} \\ &= 0.1 \pm (2.58) \sqrt{\frac{(0.1)(0.9)}{100}} \\ &= 0.1 \pm 0.0774 \end{aligned}$$

Probable limits = 0. 0226 and 0.1774

Thus, the percentage of foggy days lies between 2.26% and 17.74%

7. A random sample of 500 apples was taken from a large consignment and 65 were found to be bad. Estimate the proportion of bad apples in the consignment as well as the standard error of the estimate. Also find the percentage of bad apples in the consignment.

Sol: Proportion of bad apples in the sample is given by , $p = \frac{65}{500} = 0.13$

$$\therefore q = 1 - p = 0.87, n = 500$$

$$\text{S.E proportion of bad apples} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.13)(0.87)}{500}} = 0.015$$

$$\begin{aligned}\text{Probable limits} &= p \pm (2.58)\sqrt{\frac{pq}{n}} \\ &= 0.13 \pm (2.58)(0.015) \\ &= 0.13 \pm 0.0387\end{aligned}$$

$$\begin{aligned}\text{Probable limits} &= 0.0913 \text{ and } 0.1687 \\ &= 9.13\% \text{ and } 16.87\%\end{aligned}$$

Thus, the required percentage of bad apples in the consignment lies between 9.13 and 16.87.

8. The mean and standard deviation of the maximum loads supported by 60 cables are 11.09 tonnes and 0.73 tonnes respectively. Find a) 95% b) 99% confidence limits for mean of the maximum loads of all cables produced by the company.

Sol: By data $\bar{x} = 11.09$, $\sigma = 0.73$, $n = 60$

a) 95% confidence limits for mean of the maximum loads are given by

$$\begin{aligned}\bar{x} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right) &= 11.09 \pm (1.96)\left(\frac{0.73}{\sqrt{60}}\right) \\ &= 11.09 \pm 0.18 \\ &= 10.91 \text{ and } 11.27\end{aligned}$$

Thus, 10.91 tonnes to 11.27 tonnes are the 95% confidence limits for the mean of the maximum loads.

b) 99% confidence limits for mean of the maximum loads are given by

$$\begin{aligned}\bar{x} \pm 2.58\left(\frac{\sigma}{\sqrt{n}}\right) &= 11.09 \pm (2.58)\left(\frac{0.73}{\sqrt{60}}\right) \\ &= 11.09 \pm 0.24 \\ &= 10.85 \text{ and } 11.33\end{aligned}$$

Thus, 10.85 tonnes to 11.33 tonnes are the 95% confidence limits for the mean of the maximum loads.

9. An unbiased coin is thrown n times. It is desired that the relative frequency of the appearance of heads should lie between 0.49 and 0.51. Find the smallest value of n that will ensure this result with a) 95% confidence b) 99% confidence

Sol: Probability of getting a head, $p = \frac{1}{2}$

$$q = \frac{1}{2}$$

$$\text{S.E proportion of heads} = \sqrt{\frac{pq}{n}}$$

$$\text{S.E proportion of heads} = \sqrt{\frac{\frac{1}{2} \cdot \frac{1}{2}}{n}} = \frac{1}{2\sqrt{n}}$$

a) Probable limits for 95% confidence level is given by : $p \pm (1.96)\sqrt{\frac{pq}{n}}$, which should lie between 0.51 and 0.49

$$p \pm (1.96)\sqrt{\frac{pq}{n}} = 0.51 \text{ or } 0.49$$

$$0.5 \pm (1.96)\frac{1}{2\sqrt{n}} = 0.51 \text{ or } 0.49$$

$$0.5 \pm \frac{1.96}{2\sqrt{n}} = 0.51 \text{ or } 0.49$$

$$0.5 + \frac{1.96}{2\sqrt{n}} = 0.51$$

$$0.5 - \frac{1.96}{2\sqrt{n}} = 0.49$$

$$\frac{1.96}{2\sqrt{n}} = 0.51 - 0.5$$

$$\frac{1.96}{2\sqrt{n}} = 0.5 - 0.49$$

$$\frac{1.96}{2\sqrt{n}} = 0.01$$

$$\frac{1.96}{2\sqrt{n}} = 0.01$$

$$\sqrt{n} = \frac{1.96}{0.02}$$

$$\sqrt{n} = 98$$

Thus, **n = 9604**

b) Taking the confidence coefficient equal to 1.645 for 90% confidence level we have as before

$$\frac{1.645}{2\sqrt{n}} = 0.01$$

$$\sqrt{n} = \frac{1.645}{0.02}$$

$$\sqrt{n} = 82.25$$

Thus, **n = 6765**

Test of significance of a sample mean

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

Sol: Let p be the probability of success which being the probability of the equipment

supplied to the factory conformal to the specifications.

∴ By data $p = 0.95$ and hence $q = 0.05$

$H_0: p = 0.95$ and the claim is correct.

$H_1: p < 0.95$ and the claim is false.

We choose the one tailed test to determine whether the supply is conformal to the specification.

$$\mu = np = 200 * 0.95 = 190$$

$$\sigma = \sqrt{npq} = \sqrt{200 * 0.95 * 0.05} = 3.082$$

Expected number of equipment's according to the specification =190

Actual number =182 , since 18 out of 200 were faulty.

$$\text{Now, } Z = \frac{x - np}{\sigma}$$

$$Z = \frac{190 - 182}{3.082}$$

$$Z = 2.6$$

The value of Z is greater than the critical value 1.645 at 5% level and 2.33 at 1% level of significance.

The claim of the manufacturer (null hypothesis that claim is correct) is rejected at 5% as well as 1% level of significance in accordance with the one tailed test.

2. It has been found from experience that the mean breaking strength of a particular brand of thread is 275.6gms with standard deviation of 39.7gms. Recently a sample of 36 pieces of thread showed a mean breaking strength of 253.2gms. Can one conclude at a significance level of a) 0.05 b) 0.01 that the thread has become inferior ?

Sol: We have to decide between the two hypothesis,

$H_0: \mu = 275.6 \text{ gms}$, mean breaking strength.

$H_1: \mu < 275.6 \text{ gms}$, inferior in breaking strength.

We choose the one tailed test.

Mean breaking strength of a sample of 36 pieces=253.2

∴ Difference =275.6 – 253.2=22.4 ; n = 36

$$\text{Now, } Z = \frac{\text{Difference}}{(\sigma/\sqrt{n})}$$

$$Z = \frac{22.4}{(39.7/\sqrt{36})}$$

$$Z = 3.38$$

The value of Z is greater than the critical value of $Z=1.645$ at 5% level and 2.33 at 1% level of significance.

Under the hypothesis H_1 that the thread has become inferior is accepted at both 0.05 and 0.01 levels in accordance with one tailed test.

Test of significance of difference between means

1. In an elementary school examination the mean grade of 32 boys was 72 with a standard deviation of 8, while the mean grade of 36 girls was 75 with a standard deviation of 6.

Test the hypothesis that the performance of girls is better than boys.

Sol: We have , $\bar{x}_B = 72$, $\sigma_B = 8$, $n_B = 32$ [Boys]

$$\bar{x}_G = 75 , \sigma_G = 6 , n_G = 36 \quad \text{[Girls]}$$

$$\text{Consider, } Z = \frac{\bar{x}_G - \bar{x}_B}{\sqrt{\frac{\sigma_G^2}{n_G} + \frac{\sigma_B^2}{n_B}}}$$

$$Z = \frac{75 - 72}{\sqrt{\frac{36}{36} + \frac{64}{32}}}$$

$$Z = \sqrt{3}$$

$$Z = 1.73$$

$$Z = 1.73 \begin{cases} > Z_{0.05} = 1.645 \text{ (one tailed test)} \\ < Z_{0.01} = 2.33 \text{ (one tailed test)} \end{cases}$$

The difference in the performance of girls and boys in the examination is significant at 5% level but not at 1% level.

2. A sample of 100 bulbs produced by a company A showed a mean life of 1190 hours and a standard deviation of 90 hours. Also a sample of 75 bulbs produced by a company B showed a mean life of 1230 hours and a standard deviation of 120 hours. Is there a difference between the mean life time of the bulbs produced by the two companies at
a) 5% level of significance b) 1% level of significance.

Sol: By data $\bar{x}_A = 1190$, $\sigma_A = 90$, $n_A = 100$ [Company A]

$$\bar{x}_B = 1230 , \sigma_B = 120 , n_B = 75 \quad \text{[Company B]}$$

$$\text{Consider, } Z = \frac{\bar{x}_B - \bar{x}_A}{\sqrt{\frac{\sigma_B^2}{n_B} + \frac{\sigma_A^2}{n_A}}}$$

$$Z = \frac{1230 - 1190}{\sqrt{\frac{(120)^2}{75} + \frac{(90)^2}{100}}}$$

$$Z = 2.42$$

$$Z = 2.42 \begin{cases} > Z_{0.05} = 1.96 \text{ (Two tailed test)} \\ < Z_{0.01} = 2.58 \text{ (Two tailed test)} \end{cases}$$

The null hypothesis that there is no difference between the mean life time of bulbs is rejected at 5% level but not at 1% level of significance.

The null hypothesis that there is no difference between the mean life time of bulbs is rejected at both levels of significance in a one tailed test as the respective critical values are 1.645 and 2.33.

3. The mean life of two large samples of 1000 and 2000 members are 168.75 cms and 170 cms respectively. Can this be regarded as drawn from the sample population of standard deviation 6.25 cms ?

Sol: By data $\bar{x}_1 = 168.75$, $n_1 = 1000$

$$\bar{x}_2 = 170, \quad n_2 = 2000, \quad \sigma = 6.25$$

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

$$Z = \frac{170 - 168.75}{6.25 \sqrt{\frac{1}{1000} + \frac{1}{2000}}}$$

$$Z = 5.16$$

$Z = 5.16$ is very much greater than $Z_{0.05} = 1.96$ and $Z_{0.01} = 2.58$. Thus we say that the difference between the sample means is significant and we conclude that the samples cannot be regarded as drawn from the same population.

4. A random sample for 1000 workers in company has mean wage of Rs.50 per day and S.D of Rs.15. Another sample of 1500 workers from another company has mean wage of Rs.45 per day and S.D of Rs.20. Does the mean rate of wages varies between the two companies? Find the 95% confidence limits for the difference of the wages of the population of the two companies?

Sol: By data $\bar{x}_1 = 50$, $\sigma_1 = 15$, $n_1 = 1000$ [Company 1]

$$\bar{x}_2 = 45, \quad \sigma_2 = 20, \quad n_2 = 1500 \quad [\text{Company 2}]$$

$$\text{Consider, } Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$Z = \frac{50 - 45}{\sqrt{\frac{15^2}{1000} + \frac{20^2}{1500}}}$$

$$Z = 7.1307$$

The value of $Z = 7.1307$ is greater than $Z_{0.05} = 1.96$ and $Z_{0.01} = 2.58$. Hence we can say that the difference between the mean wages is significant both at 5% and 1% level of significance.

Also 95% confidence limits for the difference of mean wages is given by

$$\begin{aligned}(\bar{x}_1 - \bar{x}_2) \pm 1.96 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} &= 5 \pm 1.96(0.7012) \\&= 5 \pm 1.374 \\&= 3.626 \text{ and } 6.374\end{aligned}$$

Thus we can say with 95% confidence that the difference of population mean of wages between the two companies lies between Rs.3.63 and Rs.6.37.

Student's t distribution/test

We need to test the hypothesis, whether the sample mean $\bar{x}$ differs significantly from the population mean or hypothetical value μ .

We compute, $t = \frac{\bar{x} - \mu}{s} \sqrt{n}$ and consider $|t|$.

Here, $\bar{x} = \frac{\sum x}{n}$ and $s^2 = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2$

$v = n - 1$, denote the number of degrees of freedom 't'.

If $|t| > t_{0.05}$, the difference between $\bar{x}$ and μ is said to be significant at 5% level of significance.

If $|t| > t_{0.01}$, the difference between $\bar{x}$ and μ is said to be significant at 1% level of significance.

If $|t|$ is less than the table value at a certain level of significance, the data is said to be conformal/consistent with the hypothesis that μ is the mean of the population.

We have 95% confidence limits for μ is given by $\bar{x} \pm \frac{s}{\sqrt{n}} t_{0.05}$

Similarly 99% confidence limits for μ is given by $\bar{x} \pm \frac{s}{\sqrt{n}} t_{0.01}$

Note: 1. Confidence limits are also called **Fiducial limits**.

2. Test of significance for difference between sample means is

$$t = \frac{\bar{x} - \bar{y}}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{Where, } s^2 = \frac{1}{n_1+n_2-2} \left\{ \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right\}$$

And degrees of freedom, $v = n_1 + n_2 - 2$

Problems

1. Find the student's t for the following variable values in a sample of eight:

-4, -2, -2, 0, 2, 2, 3, 3 taking the mean of the universe to be zero.

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

By data $\mu = 0$ and we have $n = 8$

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

$$\bar{x} = \frac{1}{8} (-4 - 2 - 2 + 0 + 2 + 2 + 3 + 3)$$

$$\bar{x} = \frac{1}{4} = 0.25$$

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

$$s^2 = \frac{1}{(8-1)} \{ (-4.25)^2 + (-2.25)^2 + (-2.25)^2 + (-0.25)^2 + (1.75)^2 + (1.75)^2 + (2.75)^2 + (2.75)^2 \}$$

$$s^2 = \frac{1}{(7)} (49.5)$$

$$s^2 = 7.07$$

$$\therefore s = 2.66$$

$$\text{Thus, } t = \frac{0.25 - 0}{2.66} \sqrt{8}$$

$$t = 0.266$$

2. A machine is expected to produce nails of length 3 inches. A random sample of 25 nails gave an average length of 3.1 inch with standard deviation 0.3. Can it be said that the machine is producing nails as per specification? ($t_{0.05}$ for 24 d.f is 2.064)

Sol: By data, we have $\mu = 3$, $\bar{x} = 3.1$, $n = 25$, $s = 0.3$

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

$$t = \frac{3.1 - 3}{0.3} \sqrt{25}$$

$$t = 1.67 < 2.064$$

Thus, the hypothesis that the machine is producing nails as per specification is accepted at 5% level of significance.

3. Ten individuals are chosen at random from a population and their heights in inches are found to be 63, 63, 66, 67, 68, 69, 70, 70, 71, 71. Test the hypothesis that the mean height of the universe is 66 inches. ($t_{0.05} = 2.262$ for 9 d. f.).

Sol: We have $\mu = 66, n = 10$

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

$$\bar{x} = \frac{678}{10}$$

$$\bar{x} = \mathbf{67.8}$$

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

$$s^2 = \frac{1}{9} [(63 - 67.8)^2 + (63 - 67.8)^2 + \dots + (71 - 67.8)^2]$$

$$s^2 = 9.067$$

$$s = \mathbf{3.011}$$

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

$$t = \frac{(67.8 - 66)}{3.011} \sqrt{10}$$

$$t = \mathbf{1.89} < \mathbf{2.262}$$

Thus the hypothesis is accepted at 5% level of significance.

4. A sample of 10 measurements of the diameter of a sphere gave a mean of 12 cm and a standard deviation 0.15 cm. Find the 95% confidence limits for the actual diameter.

Sol: By data, $n = 10$, $\bar{x} = 12$, $s = 0.15$

Also $t_{0.05} = 2.262$ for 9 d. f

Confidence limits for the actual diameter is given by $\bar{x} \pm \left[\frac{s}{\sqrt{n}} \right] t_{0.05}$

$$\bar{x} \pm \left[\frac{s}{\sqrt{n}} \right] t_{0.05} = 12 \pm \left[\frac{0.15}{\sqrt{10}} \right] (2.262) = 12 \pm 0.1073$$

Thus 11.893 cm to 12.107 cm is the confidence limits for the actual diameter.

5. A certain stimulus administered to each of the 12 patients resulted in the following change in blood pressure 5, 2, 8, -1, 3, 0, 6, -2, 1, 5, 0, 4. Can it be concluded that the stimulus will increase the blood pressure? ($t_{0.05}$ for 11 d. f is 2.201)

Sol: $\bar{x} = \frac{\sum x}{n}$

$$\bar{x} = \frac{31}{12}$$

$$\bar{x} = 2.5833$$

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

$$s^2 = \frac{1}{11} [(5 - 2.58)^2 + (2 - 2.58)^2 + \dots + (4 - 2.58)^2]$$

$$s^2 = 9.538$$

$$s = 3.088$$

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

Let us suppose that the stimulus administration is not accompanied with increase in blood pressure, we can take $\mu = 0$.

$$t = \frac{(2.5833 - 0)}{3.088} \sqrt{12}$$

$$t = 2.8979 > 2.201$$

Hence the hypothesis is rejected at 5% level of significance. We conclude with 95% confidence that the stimulus in general is accompanied with increase in blood pressure.

6. A group of boys and girls were given an intelligence test. The mean score, S.D score and numbers in each group are as follows.

	Boys	Girls
Mean	74	70
S D	8	10
n	12	10

Is the difference between the means of the two groups significant at 5% level of significance ($t_{0.05}$ for 20 d. f is 2.086).

Sol: By data , $\bar{x} = 74$, $s_1 = 8$, $n_1 = 12$ [Boys]

$\bar{y} = 70$, $s_2 = 10$, $n_2 = 10$ [Girls]

We have $t = \frac{\bar{x} - \bar{y}}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

Where, $s^2 = \frac{1}{n_1 + n_2 - 2} \left\{ \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right\}$

Or $s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

Now $s^2 = \frac{12(64) + 10(100)}{12 + 10 - 2}$

$$s^2 = \frac{1768}{20}$$

$$s^2 = 88.4$$

$$s = 9.402$$

Hence $t = \frac{74 - 70}{(9.402) \sqrt{\frac{1}{12} + \frac{1}{10}}}$

$$t = 0.9936 < 2.086$$

Thus the hypothesis that there is a difference between the means of the two groups is accepted at 5% level of significance.

7. A sample of 11 rats from a central population had an average blood viscosity of 3.92 with standard deviation of 0.61. On the basis of this sample, establish 95% fiducial limits for μ , the mean blood viscosity of the central population. ($t_{0.05} = 2.228$ for 10 d. f).

Sol: By data $\bar{x} = 3.92$, $s = 0.61$, $n_1 = 11$

$$95\% \text{ fiducial limits for } \mu = \bar{x} \pm \left[\frac{s}{\sqrt{n}} \right] t_{0.05}$$

$$\bar{x} \pm \left[\frac{s}{\sqrt{n}} \right] t_{0.05} = 3.92 \pm \left[\frac{0.61}{\sqrt{11}} \right] (2.228)$$

$$= 3.92 \pm 0.41$$

$$= 3.51 \text{ and } 4.33$$

Thus 95% confidence limits for μ are 3.51 and 4.33.

8. Two types of batteries are tested for their length of life and the following results were obtained.

$$\text{Battery A : } n_1 = 10, \bar{x}_1 = 500 \text{ hrs, } s_1^2 = 100$$

$$\text{Battery B : } n_2 = 10, \bar{x}_2 = 500 \text{ hrs, } s_2^2 = 121$$

Compute student's t and test whether there is a significant difference in the two means.

Sol: By data Battery A : $n_1 = 10$, $\bar{x}_1 = 500$ hrs, $s_1^2 = 100$

Battery B : $n_2 = 10$, $\bar{x}_2 = 560$ hrs, $s_2^2 = 121$

$$\text{Now, } s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

$$s^2 = \frac{(10 \times 100) + (10 \times 121)}{10 + 10 - 2}$$

$$s^2 = 122.78$$

$$s = 11.0806$$

$$\text{We have } t = \frac{\bar{x}_2 - \bar{x}_1}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t = \frac{60}{(11.0806) \sqrt{\frac{1}{10} + \frac{1}{10}}}$$

$$t = 12.1081 > 2.101 \quad (t_{0.05} = 2.101 \text{ for } 18 \text{ d.f.})$$

This value of t is greater than the table value of t for 18 d.f. at all levels of significance.

The null hypothesis that there is no significant difference in the two means is rejected at all significance levels.

9. A group of 10 boys fed on a diet A and another group of 8 boys fed on a different diet B for a period of 6 months recorded the following increase in weight (lbs.)

Diet A : 5 6 8 1 12 4 3 9 6 10

Diet B : 2 3 6 8 10 1 2 8

Test whether diets A and B differ significantly regarding their effect on increase in weight.

Sol: Let the variable x correspond to the diet A and y to the diet B.

$$\bar{x} = \frac{\sum x}{n_1}$$

$$\bar{y} = \frac{\sum y}{n_2}$$

$$\bar{x} = \frac{64}{10}$$

$$\bar{y} = \frac{40}{8}$$

$$\bar{x} = 6.4$$

$$\bar{y} = 5$$

$$\sum_{i=1}^{n_1} (x_i - \bar{x})^2 = 102.4 \quad ; \quad \sum_{j=1}^{n_2} (y_j - \bar{y})^2 = 82$$

$$s^2 = \frac{1}{n_1 + n_2 - 2} \left\{ \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right\}$$

$$s^2 = \frac{1}{16} \{102.4 + 82\}$$

$$s^2 = \frac{184.4}{16} = 11.525$$

$$s = 3.395$$

$$\text{We have } t = \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t = \frac{1.4}{(3.395) \sqrt{\frac{1}{10} + \frac{1}{8}}}$$

$$t = 0.869 < 2.12 \text{ for 16 d.f}$$

Thus we conclude that the two diets do not differ significantly regarding their effect on increase in weight.

10. Two horses A and B were tested according to the time (in seconds) to run a particular race with the following results.

Horse A : 28 30 32 33 33 29 34

Horse B : 29 30 30 24 27 29

Test whether you can discriminate between the two horses.

Sol: Let the variables x and y respectively correspond to horse A and horse B.

$$\bar{x} = \frac{\sum x}{n_1} = \frac{219}{7} = 31.3$$

$$\bar{y} = \frac{\sum y}{n_2} = \frac{169}{6} = 28.2$$

$$\sum_{i=1}^{n_1} (x_i - \bar{x})^2 = 31.43$$

$$\sum_{j=1}^{n_2} (y_j - \bar{y})^2 = 26.84$$

$$s^2 = \frac{1}{n_1 + n_2 - 2} \left\{ \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right\}$$

$$s^2 = \frac{1}{11} (31.43 + 26.84)$$

$$s^2 = 5.2973$$

$$\therefore s = 2.3016$$

$$\text{We have } t = \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t = \frac{31.3 - 28.2}{(2.3016) \sqrt{\frac{1}{7} + \frac{1}{6}}}$$

$$t = 2.42 = \begin{cases} > t_{0.05} = 2.2 \\ < t_{0.02} = 2.72 \end{cases}$$

The discrimination between the horses is significant at 5% level but not at 2% level of significance.

Chi-Square distribution

Chi-square distribution provides a measure of correspondence between the theoretical frequencies and observed frequencies.

If $O_i (i = 1, 2, \dots, n)$ and $E_i (i = 1, 2, \dots, n)$ respectively denotes a set of observed and estimated frequencies, the quantity **chi-square** denoted by χ^2 is defined as follows.

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} ; \quad \text{degrees of freedom} = n - 1$$

Note: If the expected frequencies are less than 10, we group them suitably for computing the value of chi. Square.

Chi-square test as a test of goodness of fit

It is possible to test the hypothesis about the association of two attributes. We already discussed the fitting of Binomial distribution, Normal distribution, Poisson distribution to a given data. It is easily possible to find the theoretical frequencies from the distribution of fit.

Chi-square test helps us to test the goodness of fit of these distributions.

If the calculated value of χ^2 is less than the value of χ^2 at a specified level of significance, the hypothesis is accepted. Otherwise the hypothesis is rejected.

Problems

1. A dice is thrown 264 times and the number appearing on the face (x) follows the following frequency distribution.

x	1	2	3	4	5	6
f	40	32	28	58	54	60

Calculate the value of χ^2 .

Sol: The frequencies in the given data are the observed frequencies. Assuming that the dice is unbiased, the expected number of frequencies for the numbers 1, 2, 3, 4, 5, 6 to appear on the face is $\frac{264}{6} = 44$ each.

Now the data is as follows:

No. on the dice	1	2	3	4	5	6
Observed frequency (O_i)	40	32	28	58	54	60

Expected frequency (E_i)	44	44	44	44	44	44
------------------------------	----	----	----	----	----	----

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{(40 - 44)^2}{44} + \frac{(32 - 44)^2}{44} + \dots + \frac{(60 - 44)^2}{44}$$

$$\chi^2 = \frac{1}{44} [16 + 144 + 256 + 196 + 100 + 256]$$

$$\chi^2 = \frac{968}{44}$$

Thus, $\chi^2 = 22$

2. Five dice were thrown 96 times and the numbers 1, 2 or 3 appearing on the face of the dice follows the frequency distribution as below.

No. of dice showing 1,2 or 3	5	4	3	2	1	0
Frequency	7	19	35	24	8	3

Test the hypothesis that the data follows a binomial distribution. ($\chi_{0.05}^2 = 11.07$ for 5 d. f)

Sol: The data gives the observed frequencies and we need to calculate the expected frequencies.

Probability of a single dice throwing 1, 2 or 3 is $p = \frac{3}{6} = \frac{1}{2}$

$$\therefore q = 1 - p = \frac{1}{2}$$

From the binomial distribution, $P(x) = {}^nC_x p^x q^{(n-x)}$

The theoretical frequencies of getting 5, 4, 3, 2, 1, 0 successes with 5 dice are respectively the successive terms of the binomial expansion.

They are respectively $96 * {}^5C_0 * \frac{1}{2^5}$, $96 * {}^5C_1 * \frac{1}{2^5}$, $96 * {}^5C_2 * \frac{1}{2^5}$, $96 * {}^5C_3 * \frac{1}{2^5}$, $96 * {}^5C_4 * \frac{1}{2^5}$, $96 * {}^5C_5 * \frac{1}{2^5}$ and they are 3,15,30,30,15,3.

We have the table of observed and expected frequencies.

O_i	7	19	35	24	8	3
-------	---	----	----	----	---	---

E_i	3	15	30	30	15	3
-------	---	----	----	----	----	---

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{16}{3} + \frac{16}{15} + \frac{25}{30} + \frac{36}{30} + \frac{49}{15} + \frac{0}{3}$$

$$\chi^2 = 11.7 > \chi_{0.05}^2 = 11.07$$

Thus the hypothesis that the data follows a binomial distribution is rejected.

3. A sample analysis of examination results of 500 students was made. It was found that 220 students had failed, 170 had secured third class 90 had secured second class and 20 had secured first class. Do these figures support the general examination result which is in the ratio 4:3:2:1 for the respective categories ($\chi_{0.05}^2 = 7.81$ for 3 d.f).

Sol: Let us take the hypothesis that these figures support to the general result in the ratio 4:3:2:1

The expected frequencies in the respective category are

$$\frac{4}{10} * 500, \frac{3}{10} * 500, \frac{2}{10} * 500, \frac{1}{10} * 500 \text{ or } 200, 150, 100, 50$$

We have the following table.

O_i	220	170	90	20
E_i	200	150	100	50

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{400}{200} + \frac{400}{150} + \frac{100}{100} + \frac{900}{50}$$

$$\chi^2 = 23.67 > \chi_{0.05}^2 = 7.81$$

Thus the hypothesis is rejected.

4. 4 coins are tossed 100 times and the following results were obtained. Fit a binomial distribution for the data and calculate the theoretical frequencies ($\chi_{0.05}^2 = 9.49$ for 4 d.f).

Number of heads	0	1	2	3	4
Frequency	5	29	36	25	5

Sol: Let x denote the number of heads and f the corresponding frequency. Since the data is in the form of a frequency distribution we shall first calculate the mean.

$$\text{Mean, } \mu = \frac{\sum fx}{\sum f} = \frac{(0)(5)+(1)(29)+(2)(36)+(3)(25)+(4)(5)}{5+29+36+25+5} = \frac{196}{100} = 1.96$$

From binomial distribution, $\mu = np$. Here $n=4$

$$\text{Hence } 4p = 1.96$$

$$p = 0.49$$

$$q = 0.51$$

From binomial distribution,

$$P(x) = {}^nC_x p^x q^{n-x}$$

$$P(x) = {}^4C_x (0.49)^x (0.51)^{4-x}$$

Since 4 coins were tossed, expected(theoretical) frequencies are obtained from

$$F(x) = 100P(x)$$

$$F(x) = 100.4c_x (0.49)^x (0.51)^{4-x} \text{ where } x = 0,1,2,3,4.$$

$$F(0) = 100.4c_0 (0.49)^0 (0.51)^4 = 6.765 \cong 7$$

$$F(1) = 100.4c_1 (0.49)^1 (0.51)^3 = 25.999 \cong 26$$

$$F(2) = 100.4c_2 (0.49)^2 (0.51)^2 = 37.47 \cong 37$$

$$F(3) = 100.4c_3 (0.49)^3 (0.51)^1 = 24.0004 \cong 24$$

$$F(4) = 100.4c_4 (0.49)^4 (0.51)^0 = 5.765 \cong 6$$

Thus the required theoretical frequencies are **7,26,37,24,6**.

We have the following table

O_i	5	29	36	25	5
E_i	7	26	37	24	6

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{4}{7} + \frac{9}{26} + \frac{1}{37} + \frac{1}{24} + \frac{1}{6}$$

$$\chi^2 = 1.15 < \chi_{0.05}^2 = 9.49$$

Thus the hypothesis that the fitness is good can be accepted.

5. Fit a poisson distribution for the following data and test the goodness of fit given that $\chi^2_{0.05} = 7.815$ for 3 d.f

x	0	1	2	3	4
f	122	60	15	2	1

Sol: Let x denote the number of heads and f the corresponding frequency. Since the data is in the form of a frequency distribution we shall first calculate the mean.

$$\text{Mean, } \mu = \frac{\sum fx}{\sum f} = \frac{(0)(122)+(1)(60)+(2)(15)+(3)(2)+(4)(1)}{122+60+15+2+1} = \frac{100}{200} = 0.2$$

From poisson distribution, $\mu = m$.

$$\therefore m = 0.2$$

$$\text{Wkt, } P(x) = \frac{m^x e^{-m}}{x!}$$

$$\text{Let } F(x) = 200 * P(x)$$

$$F(x) = 200 * \frac{m^x e^{-m}}{x!}$$

$$F(x) = 200 * \frac{(0.5)^x e^{-(0.5)}}{x!} \quad \text{But } e^{-(0.5)} = 0.6065$$

$$F(x) = 121.3 \frac{(0.5)^x}{x!}$$

$$F(0) = 121, F(1) = 61, F(2) = 15, F(3) = 3, F(4) = 0$$

Thus the required theoretical frequencies are 121,61,15,3,0.

Since the last of the expected frequency is 0 we shall club it with the previous one.

We have the following table

O_i	122	60	15	2+1=3
E_i	121	61	15	3+0=3

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{1}{121} + \frac{1}{61} + 0 + 0$$

$$\chi^2 = 0.025 < \chi^2_{0.05} = 7.815$$

Thus the hypothesis that the fitness is good can be accepted.

6. The number of accidents per day (x) as recorded in a textile industry over a period of 400 days is given below. Test the goodness of fit in respect of Poisson distribution of fit to the given data ($\chi^2_{0.05} = 9.49$ for 4 d. f.).

x	0	1	2	3	4	5
f	173	168	37	18	3	1

Sol: Let x denote the number of heads and f the corresponding frequency. Since the data is in the form of a frequency distribution we shall first calculate the mean.

$$\text{Mean, } \mu = \frac{\sum fx}{\sum f} = \frac{(0)(173)+(1)(168)+(2)(37)+(3)(18)+(4)(3)+(5)(1)}{173+168+37+18+3+1} = \frac{313}{400} = 0.7825$$

From poisson distribution, $\mu = m$.

$$\therefore m = 0.7825$$

$$\text{Wkt, } P(x) = \frac{m^x e^{-m}}{x!}$$

$$\text{Let } F(x) = 400 * P(x)$$

$$F(x) = 400 * \frac{m^x e^{-m}}{x!}$$

$$F(x) = 400 * \frac{(0.7825)^x e^{-(0.7825)}}{x!} \quad \text{But } e^{-(0.7825)} = 0.4573$$

$$F(x) = 182.92 \frac{(0.7825)^x}{x!}$$

$$F(0) = 183, F(1) = 143, F(2) = 56, F(3) = 15, F(4) = 3, F(5) = 0$$

Thus the required theoretical frequencies are 183,143,56,15,3,0.

Since the last of the expected frequency is 0 we shall club it with the previous one.

We have the following table

O_i	173	168	37	18	3+1=4
E_i	183	143	56	15	3+0=3

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{100}{183} + \frac{625}{143} + \frac{361}{56} + \frac{9}{15} + \frac{1}{3}$$

$$\chi^2 = 12.297 > \chi_{0.05}^2 = 9.49$$

Thus the hypothesis that the fitness is good can be rejected.