# Hypothesis Testing

## Hypothesis testing :-

Statistical technique to test some hypothesis about the parent population from which the sample is actually drawn.

## population :

A complete collection of all elements to be studied.

## Sample :

A sub collection of elements drawn from a population.

## Estimation :-

It is to use the statistics obtained from the sample as estimate of the unknown parameters of the population from which the sample is drawn.

" A hypothesis is statistics is simply a quantitative statement about population"

## procedure for Hypothesis Testing :-

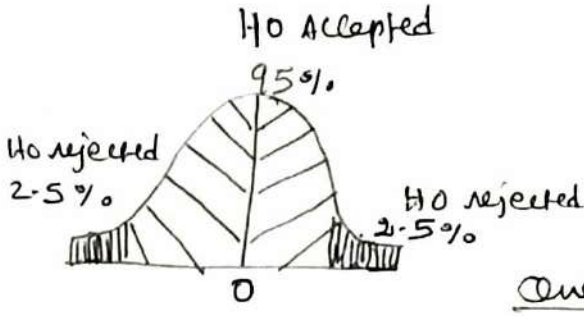
### ① Set up the Hypotheses :

$H_0$ ( Null Hypothesis )

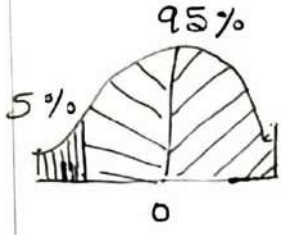
$H_a$ / $H_1$ ( Alternative Hypotheses).

Test of validity of Ho against Ha at certain level of Significant i.e 5%, 1%. etc.

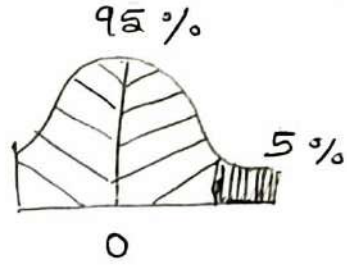
Two tailed




One tailed




One tailed




| Level of Significance | 10% | 5% | 1% | 0.1% |
|---|---|---|---|---|
| One tailed | 1.65% | 1.96 | 2.58 | 3.29 |
| Two tailed | 1.28 | 1.64 | 2.33 | 3.10 |

Critical z value (Tabulated)

## Z value :-

It is a standardized score that describes how many S.D ($\sigma$) an element is from the mean.

③ Setting a test Criterion:

Selection of appropriate probability distribution for the test. i.e., t-test, $\chi^2$ test, F test etc

④ Doing Computation:

All calculation part is carried out.

⑤ making Decision:

we have to draw Statistical conclusions and take decisions.

| Decision | population | Condition |
|---|---|---|
| | Ho is true | Ho is false |
| Accept Ho / Reject Ha | Correct decision | Type II Error |
| Reject Ho / Accept Ha | Type I Error | Correct decision |

# Chi - square Test

## Definition:

"Chi - square test is the test of significance of overall deviation square in the observed and Expected frequencies devided by Expected frequencies."

## Characteristics of $\chi^2$ test:

- The test is based on events or frequencies and not based on mean or S.D. etc,
- The test can be used between one entire set of observed and expected frequencies.
- To draw inferences, this test is applied, especially testing the hypotheses.
- It is a general test and is highly useful in research

## Application of Chi- Square test:

- It is used to test the goodness of fit.
- The test enables to find out whether the difference between the expected and observed value is significant or not.
- If the difference is little then the fit is good, Otherwise the fit is poor

Formula:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

where: O = observed frequencies
E = Expected frequencies
$\Sigma$ = Sum of

Steps:

1. A hypothesis is established i.e. Null hypothesis.
2. Calculate the difference between observed value and expected value $(O-E)$.
3. Square the deviations calculated $(O-E)^2$
4. Divide the $(O-E)^2$ by its expected frequency $\frac{(O-E)^2}{E}$.
5. Add all the values obtained in Step 4

$$\chi^2 = \Sigma \left[\frac{(O-E)^2}{E}\right]$$

6. Find the Chi-Square from $\chi^2$ table at certain level of significance, usually 5 % or 1 % level.

Inference:

- If the calculated value of $\chi^2$ is greater than the table value of $\chi^2$ at certain degree of level of significance, we reject the hypothesis.
- If the calculated value of $\chi^2$ is zero, the observed values and expected values completely coincide.

- If the calculated value of $\chi^2$ is less than table value at certain degree of level of significance, it is said to be non-significant.
- Thus implies that the difference between the observed and expected frequencies may be due to fluctuations in sampling.

Illustration 1: A coin is tossed 100 times of which head comes 60 times and tail 40 times. Could you accept the hypothesis that the coin is normal having no bias for either head or tail.

Solution:

Step: 1: Null hypothesis - i.e. the coin is normal having no bias for either head or tail.

2: level of significance 5 %.

3. Determining expected frequencies (E)

| possibilities | Observed (O) frequencies | Expected frequencies (E) |
|---|---|---|
| Head | 60 | 50 |
| Tail | 40 | 50 |

4. Fixing the degree of freedom $df = n-1$

$n$ = number of events or possibilities

i.e. head and tail $n = 2-1 = 1$

5. calculation: $x^2 = \Sigma \left[ \frac{(O-E)^2}{E} \right]$

| possibilities | Observed frequency (O) | Expected frequency (E) | (O-E) | $(O-E)^2$ | $\frac{(O-E)^2}{E}$ |
|---|---|---|---|---|---|
| Head | 60 | 50 | 60-50 = 10 | = 100 | $\frac{100}{50} = 2.0$ |
| Tail | 40 | 50 | 40-50 = 10 | = 100 | $\frac{100}{50} = 2.0$ |

Calculated $x^2$ value = 4.00 $= 4$

Table value at 5 % level for one degree of freedom is 3.84

Inference: The calculated $x^2$ value is greater than the table value. Therefore hypothesis is rejected. In other words the coin is defective with bias for head.

Illustration 3: when two heterozygous pea plants are crossed. 1600 plants are produced in the F2 generation. out of which 940 are yellow round. 260 are yellow wrinkled, 340 are green round and 60 are green wrinkled. By means of $x^2$ test whether these values are deviated from mendel's dihybrid ratio 9:3:3:1.

Solution:

Step 1: Null hypothesis: There is no difference between Observed values and Mendel's dihybrid ratio 9:3:3:1.

2: level of Significance 5%.

3: Determining expected frequencies (E):
Mendel's dihybrid ratio 9:3:3:1

Yellow Round = 9 Total 1600 $\therefore E = \frac{9}{16} \times 1600 = 900$

yellow wrinkled = 3 " $\therefore E = \frac{3}{16} \times 1600 = 300$

Green Round = 3 " $\therefore E = \frac{3}{16} \times 1600 = 300$

Green wrinkled = $\frac{1}{16}$ Total 1600 $\therefore E = \frac{1}{16} \times 1600 = \frac{100}{1600}$

4: Fixing the df = n-1 = 4-1 = 3

Calculation: $x^2 = \Sigma \left[ \frac{(O-E)^2}{E} \right]$

| Variables | O | E | O-E | $(O-E)^2$ | $\frac{(O-E)^2}{E}$ |
|---|---|---|---|---|---|
| yellow Round | 940 | 900 | 40 | 1600 | 1.77 |
| yellow wrinkled | 260 | 300 | -40 | 1600 | 5.33 |
| Green Round | 340 | 300 | 40 | 1600 | 5.33 |
| Green wrinkled | 60 | 100 | -40 | 1600 | 16.00 |
| | | | | | $\Sigma = 28.43$ |

Calculated $x^2$ value = 28.43

For 3 df, at 5% level of Significance, Table $x^2$ value 7.81

Inference: The calculated $x^2$ value is greater than the table $x^2$ value. Therefore the hypothesis is rejected. In other words There is no real independent assortment or the observed value are deviated from Mendel's dihybrid ratio 9:3:3:1.

6. Calculation : $x^2 = \Sigma \left[ \frac{(O-E)^2}{E} \right]$

| O | E | O-E | $(O-E)^2$ | $\frac{(O-E)^2}{E}$ |
|---|---|---|---|---|
| 200 | 300 | −100 | 10000 | 33.33 |
| 280 | 180 | 100 | 10000 | 55.55 |
| 300 | 200 | 100 | 10000 | 50.00 |
| 20 | 120 | −100 | 10000 | 83.33 |
| | | | | $\Sigma = 222.21$ |

Calculated $x^2$ value : 222.21

For 1 df, at 5 % level of Significance
- the table $x^2$ value = 3.84

Inference :

The calculated $x^2$ value (222.21) is greater than the table $x^2$ value (3.84) Therefore the null hypothesis is rejected. In other words the drug is effective in preventing typhoid.

# Analysis of Variance (ANOVA)

- The term ANOVA was first proposed by R.A Fisher.
- Analysis of variance refers to the examination of differences among the samples.
- It is an extremely useful technique concerning research in Biology
- It is used to examine the significance of the difference amongst more than two sample means at the same time.
- The analysis of variance has been classified into
  - One way classification
  - Two-way classification

**Principles:**

We take two estimates of population variance i.e., one based on between samples variance and the other within samples variance. Then these two estimates of population variance are compared with 'F' test as follows

$$F = \frac{\text{Variance between Samples}}{\text{Variance within Samples}}$$

The value of F is to be compared to the F-limit for a given degrees of freedom. If the calculated F value exceeds the F-table value, we can say that there are significant variance between the sample means.

Steps involved in the Analysis are:

Step: 1

Find out the means of each samples

$\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4 \text{---} \bar{x}_k$

Step: 2

Find out the combined mean of the samples

$$\bar{X} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4 \ldots \ldots \bar{X}_k}{\text{No. of Samples}}$$

Step: 3

Sum of Squares between the samples (or) SS-between

$$\therefore \text{SS-between} = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + n_3(\bar{x}_3 - \bar{x})^3 + n_4(\bar{x}_4 - \bar{x})^2 \ldots \ldots n_k(\bar{x}_k - \bar{x})^2$$

$n$ = number of items in the corresponding samples.

Step: 4

Mean square between the samples (or) MS-between.

$$\therefore \text{MS-between} = \frac{\text{SS- between}}{\text{degrees of freedom between the Samples.}}$$

Step: 5

Sum of squares within the samples (or) SS-within

$$\therefore \text{SS within} = \Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2 + \Sigma(x_3 - \bar{x}_3)^2 \Sigma(x_4 - \bar{x}_4)^2 \ldots \ldots \Sigma(x_k - \bar{x}_k)^2$$

Step: 6

Mean square within the samples (or) M-S within.

$$\therefore \text{M-S within} = \frac{\text{SS- within}}{\text{degrees of freedom within the Samples}}$$

Step: 7 Make ANOVA Table:

| Source of Variance | Sum of Square (SS) | Degree of freedom | Mean Square (MS) |
|---|---|---|---|
| Between Sample | | | |
| within Samples | | | |
| Total | | | |

Step: 8 Find out F - value

$$F = \frac{\text{Variance between Samples}}{\text{Variance within Samples}} = \frac{\text{MS between}}{\text{MS within}}$$

If the calculated F- value is less than F-Table value, there is no significant.

## Illustration : ANOVA — one way.

A certain manure was used on four plots of land A, B, C and D. Four beds were prepared in each plot and the manure used. The output of the crop in the beds of plots A, B, C and D is given below

| A | B | C | D |
|---|---|---|---|
| 6 | 15 | 9 | 8 |
| 8 | 10 | 3 | 12 |
| 10 | 4 | 7 | 1 |
| 8 | 7 | 1 | 3 |

using — ANOVA find out whether the difference in the means of the production of crops of the plots is significant or not.

Solution :

Step 1 : Find out the means of each samples.

| Sample I ($x_1$) | Sample II ($x_2$) | Sample III ($x_3$) | Sample IV ($x_4$) |
|---|---|---|---|
| 6 | 15 | 9 | 8 |
| 8 | 10 | 3 | 12 |
| 10 | 4 | 7 | 1 |
| 8 | 7 | 1 | 3 |
| Tot : 32 | 36 | 20 | 24 |
| $\bar{x}$ 8 | 9 | 5 | 6 |

Step : 2 Find out the Combined mean of the samples.

$$\bar{X} = \frac{\bar{X_1} + \bar{X_2} + \bar{X_3} - \bar{X_4}}{\text{NO. of samples}} = \frac{8+9+5+6}{4}$$

$$= \frac{28}{4} = 7 \qquad \bar{X} = 7$$

Step:3 Sum of Squares between the samples
(or) SS- between

$\therefore$ SS between $= n_1(\bar{x}_1-\bar{x})^2 + n_2(\bar{x}_2-\bar{x})^2 + n_3(\bar{x}_3-\bar{x})^2 + n_4(\bar{x}_4-\bar{x})^2$

$= 4(8-7)^2 + 4(9-7)^2 + 4(5-7)^2 + 4(6-7)^2$

$= 4(1)^2 + 4(2)^2 + 4(-2)^2 + 4(-1)^2$

$= 4(1) + 4(4) + 4(4) + 4(1)$

$= 4 + 16 + 16 + 4$

$= 40$

Step : 4- Mean Square between the samples
(or) MS- between

$\therefore$ MS-between $= \dfrac{\text{SS\_between}}{\text{degree of freedom between the samples}}$

There are four samples so the degrees of freedom are $4-1 = 3$

$\therefore$ MS- between $= \dfrac{40}{3} = 13.33$

Step : 5 Sum of Squares within the Samples (or) SS-within.

$\therefore$ SS- within $= \Sigma(x_1-\bar{x}_1)^2 + \Sigma(x_2-\bar{x}_2)^2 + \Sigma(x_3-\bar{x}_3)^2 + \Sigma(x_4-\bar{x}_4)^2$

| $x_1$ | $(x_1-\bar{x}_1)$ $\bar{x}_1=8$ | $(x_1-\bar{x}_1)^2$ | $x_2$ | $(x_2-\bar{x}_2)$ $\bar{x}_2=9$ | $(x_3-\bar{x}_3)^2$ |
|---|---|---|---|---|---|
| 6 | $6-8=-2$ | 4 | 15 | $15-9=6$ | 36 |
| 8 | $8-8=0$ | 0 | 10 | $10-9=1$ | 1 |
| 10 | $10-8=2$ | 4 | 4 | $4-9=-5$ | 25 |
| 8 | $8-8=0$ | 0 | 7 | $7-9=-2$ | 4 |
| | | 8 | | | 66 |

| $x_3$ | $(x_3-\bar{x}_3)$ $\bar{x}_3=5$ | $(x_3-\bar{x}_3)^2$ | $x_4$ | $(x_4-\bar{x}_4)$ $\bar{x}_4-6$ | $(x_4-\bar{x}_4)^2$ |
|---|---|---|---|---|---|
| 9 | $9-5=4$ | 16 | 8 | $8-6=2$ | 4 |
| 3 | $3-5=-2$ | 4 | 12 | $12-6=6$ | 36 |
| 7 | $7-5=2$ | 4 | 1 | $1-6=-5$ | 25 |
| 1 | $1-5=-4$ | 16 | 3 | $3-6=-3$ | 9 |
| | | 40 | | | 74 |

$\therefore$ SS-within $= 8+66+40+74$

$= 188$

Step :6 - Mean Square within the samples (or) MS- within

$\therefore$ MS-within $=$ $\frac{\text{SS-within}}{\text{degree of freedom within the samples}}$

There are 16 items within the 4 samples

$\therefore$ degrees of freedom $16-4=12$

$\therefore$ MS- within $= \frac{188}{12} = \underline{15.66}$

Step : 7 make ANOVA Table :

| Source of Variance | Sum of Squares SS | Degree of freedom (df) | Mean Square (MS) |
|---|---|---|---|
| Between Sample | 40 | 3 | 13.33 |
| within Sample | 188 | 12 | 15.66 |
| Total | 228 | 15 | |

Step : 8 Find out F value

$$F = \frac{\text{Variance between Samples}}{\text{Variance within Samples}} = \frac{13.33}{15.66} = \underline{\underline{0.851}}$$

The table value of F for $V_1 = 3$ and $V_2 = 15$ at 5% level of significance $= \underline{\underline{3.49}}$

Step : 9 Inference:

The calculated value (0.351) is lesser than the table value (3.49). Therefore the difference in the means of the production of crops of the plots is not significant.

## PROBABILITY THEORY :

↳ Probability theory a branch of mathematics concerned with the analysis is random phenomena. The outcome of a random event cannot be determined before it occurs, but it may be any one of several possible outcomes. The actual outcome is considered to be determined by chance.

↳ It is the most crucial link between the population and its variables, allows us to draw inferences on the population based on the sample observations.

↳ The three probability distributions useful in medicine / health care are,

1. Normal distributions
2. Binomial distributions
3. poisson distributions.

## 1. NORMAL DISTRIBUTIONS :

↳ It is defined as a continuous frequency distribution of infinite range. The normal distribution is a descriptive model that describes real world situations.

↳ Some distributions of data, such as the bell curve are symmetric

↳ This means that the right and the left of the distribution are perfect mirror images of one another.

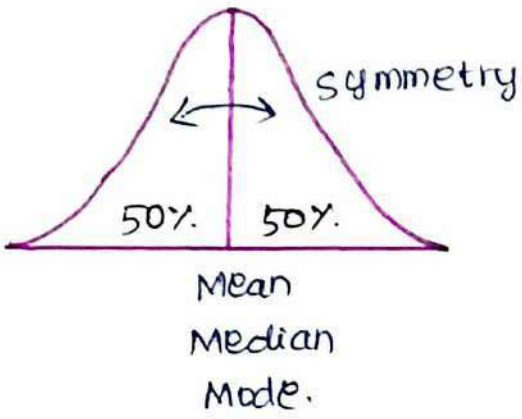
↳ Not every distribution of data is symmetric

↳ sets of data that are not symmetric are said to be asymmetric.

↳ The measure of asymmetric a distribution can be is called skewness.

↳ The mean, median and mode are all measures of the center of a set of data.

↳ The skewness of the data can be determined by these quantities are related to one another.




The normal distribution has

↳ Mean = Median = Mode

↳ Symmetry about the center

↳ 50% of values less than the mean and 50% greater than the mean.

It is often called a "Bell curve" because it looks like a bell.

## Many things closely follow a Normal distribution

↳ Heights of people

↳ Size of things produced by machines

↳ Errors in measurements

↳ Blood pressure

↳ Marks on a test.

## Normal distribution Mathematical formula

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

↳ $\pi = 3.14159$

↳ $e = 2.71828$.

↳ $\sigma = 1$. (standard deviation)

## Features of Normal distribution:

↳ Normal distributions are symmetric around their mean.

↳ The mean, median, and mode of a normal distribution are equal

↳ The area under the normal curve is equal to 1.0

↳ Normal distributions are denser in the center and less dense in the tails

↳ Normal distributions are defined by two parameters, the mean ($\mu$) and the standard deviation ($\sigma$)

↳ 68% of the area of a normal distribution is within one standard deviations of the mean.

↳ Approximately 95% of the area of a normal distribution is written within two standard deviation of the mean.

## Importance:

↳ Many dependent variables are commonly assumed to be normally distribution in the population.

↳ If a variable is approximately normally distributed we can make inferences about values of that variable.

## Properties of a Normal distribution:

↳ The mean, mode and median are all equal

↳ The curve is symmetric at the center (ie around the mean, $\mu$)

↳ Exactly half of the values are to the left of center.

↳ The total area under the curve is 1.

2. BINOMINAL DISTRIBUTION :

↳ Binomial distribution was discovered by James Bernoulli in 1738.

↳ This is discrete probability distribution.

↳ A discrete probability distribution (applicable to the scenarios the set of possible outcomes is discrete such as a coin toss or a roll of dice) can be encoded by a discrete list as the probabilities of the out comes, known as a probability mass function.

↳ If 'x' is a discrete random variable with probability mass function.

BINOMIAL DISTRIBUTION FORMULA :

↳ $p(x) = n\,C_x\, p^x (1-p)^{n-x}$.

$x = 0, 1, 2, 3 \ldots\ldots n$

$q = 1-p$, then 'x' is a binomial variate and the distribution of 'x' is called binominal distribution.

↳ The word "binomial" literally means "two numbers". A binomial distribution for a random variable x is one in there are only two possible outcomes, success and failure, for a finite number of trails.

ASSUMPTION FOR BINOMIAL DISTRIBUTION :

↳ For each trial there are only two possible outcomes on each trial, S (Success), F (Failure)

↳ The number of trials 'n' is finite

↳ For each trial are independent the outcome of a trial is not affected by the outcomes of any

other trial.

↳ The probability of success, p is constant from trial to trial.

APPLICATIONS OF BINOMIAL DISTRIBUTIONS:

↳ Binomial distribution describe the possible number of times that a particular event will occur in sequence of observations.

↳ They are used when we want to know about the occurrence of an event, not its magnitude

↳ Common uses of binomial distributions in business include quality control industrial engineers are interested in the proportion of defections.

↳ Also used extensively for medical

↳ It is also used in military applications.

3. POISSION DISTRIBUTION:

↳ Poisson distribution is a discrete probability distribution and it is widely used in statistical work. This distribution was developed by distribution was developed by French mathematician Dr. Simon Denis poisson in 1837 and the distribution is named after him.

↳ The poisson Distribution is used in those situations the probability of happening of an event is small (i.e the event rarely occurs).

## POISSION DISTRIBUTION FORMULA :

↳ Poisson distribution is defined and given by the following probability function.

↳ $P(X=x) = \frac{e^{-m}}{x} \times m^x$

↳ When $P(X=x)$ = probability of obtaining x number of success.

↳ $m = np$ = parameter of distribution

↳ $e = 2.7183$ base of natural logarithms.

↳ poisson distribution is a limiting from of the binomial distribution in n, the number of trials becomes very large & P, the probability of success of the event is very very small.

## Use of poisson distribution :

↳ Control limits of numbers of tablets rejected from an online Metal Detector during tablet Compression cycle.

↳ Microbial counts in raw materials, products, and water for pharmaceutical use.

↳ Control limits for numbers of containers rejected from visual inspection of sterile production batches

↳ Alert limits for microbial levels in cleanroom environment

↳ Release limits for microbial counts in non sterile products.

$$P(X) = \frac{e^{-m} \cdot m^x}{x!}$$

- $x = 1, 2, 3, 4 \ldots n$
- $e = 2.7183$ (The base of natural logarithms)
- $m$ = The mean of poisson distribution i.e the average number of occurrence of an event.

## Condition under which Poisson Distribution is used:

- The random variables x should be discrete
- A dichotomy exists, i.e happening of the event must be of two alternatives such as sucess & failure.
- Applicable in those cases the number of trails n is very large and the probablity of success p is very small but mean $np = m$ is finite.
- Statistical Independence is assumed

## Characteristics of poisson Distribution

- Poisson distribution is a discrete distribution
- It depends mainly on the value of the mean m.
- This distribution is positively skewed to the left with the increase in the value of the mean m, the distribution shift to the right and skewness demenished
- If n is large & p is small this distribution gives a close approximation binomial distribution since the arithmetic mean of poisson is same as that Binomial.

↳ Poisson distribution has only one parameter. i.e m, the arithmetic mean. Thus the entire distribution can be determined once the arithmetic mean is known.

↳ The poisson distribution is a discrete distribution with a single parameter, m. As m increases, the distribution shifts to the right.

↳ All the poission distribution is skewed to right This is the reason the poisson probability distributions has been called the probability of distribution of rare events.

Conclusion:

↳ In conclusion We can say that the poisson Distribution is useful in rare events the probability of success (p) is very small and probility of fulure (q) is very large and value of n is very large.