24.5.21

SAMPLING AND STATISTICAL INFERENCE :

Sampling is a technique of selecting a subset of a population to make statistical inferences from them and estimate characteristics of the whole population.

Eg (i) Vaccine trials (ii) Poll prediction

Some important terms & definitions :

1 Population : It is the set of all aggregate objects under study (usually denoted by N)

2 Sample It is a subset of the population that is considered to study the behaviour of the population. (The sample size is usually denoted by n)

If $n \geq 30$, it is considered to be a large sample

$n < 30$, $\rightarrow$ small sample

The statistical constants of the population such as Mean, Variance etc are called Parameters (μ, σ^2)

The statistical constants of the sample which is used to estimate the parameters are called Statistics $(\bar{x}$ & $s^2)$

Sampling can be done in different ways but one which is most common is Random Sampling. This means every member of the population has an equal chance of being selected. It can be done in 2 ways

(i) With Replacement : If N is the size of the population & n is the size of sample, then there are N^n possible samples.

(ii) Without Replacement : ${}^{N}C_{n}$ possibilities

<u>Sampling Distribution</u> : A sampling distribution is the frequency distribution of a statistic over many random samples from a single population.

Let us say we have different samples of size n drawn from a population of size N. For each of these samples, we can compute mean, variance etc and they will not be the same. If we group these according to their frequencies, then the frequency distribution so generated is called the <u>Sampling Distribution</u>. In general, we can have a sampling distribution of mean, variance etc.

The standard deviation of the sampling distribution is called the <u>Standard Error</u>.

<u>Eg</u> A population consists of 3 members $\{1, 2, 3\}$. Form a sampling distribution of the mean for random samples of size 2 with replacement

<u>Note that</u>: $N = 3$, $\mu = \frac{1+2+3}{3} = 2$, $\sigma^2 = \frac{1}{N}\Sigma(x-\mu)^2 = \frac{2}{3}$

The random samples of size 2 with replacement are:

$\{(1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3,3)\}$ $\{= N^n = 3^2 = 9\}$

Means of these samples are: 1, 1.5, 2, 1.5, 2, 2.5, 2, 2.5, 3

Sampling distribution of mean is:

x_i	f_i	$f_i x_i$
1	1	1
1.5	2	3
2	3	6
2.5	2	5
3	1	3
	9	18

<u>Mean</u> $\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{18}{9} = 2$ $(\mu = 2)$

<u>Variance</u>: $s^2 = \frac{1}{\Sigma f_i} \Sigma f_i (x_i - \bar{x})^2$

$$= \frac{1}{9}\left[1(1-2)^2 + 2(1.5-2)^2 + 3(2-2)^2 + 2(2.5-2)^2 + 1(3-2)^2\right]$$

$$= \frac{1}{3} \quad \Big| \Rightarrow s^2 = \frac{\sigma^2}{n} = \left(\frac{2}{3}\right)\left(\frac{1}{2}\right) = \frac{1}{3}$$

Central limit Theorem: If we have a population N with mean μ and s.d. σ i.e. $N(\mu, \sigma)$ and we take sufficiently large samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.

Hypothesis: The assumption that we make regarding the parameter of the population is called Hypotheses.

Test of Significance or Test of Hypothesis:

An important aspect of sampling theory is to study or assert that the parameter of the population is the same as the statistic obtained from the random sample.

The process or test that decides whether to accept the hypothesis or not is called the test of significance or test of hypothesis.

In order to arrive at a decision, we want to make certain assumptions known as hypotheses which may or may not be true.

The hypothesis is accepted or rejected based on some statistical tests. We decide whether the sample statistic is significantly different from the parameter at a desired level of significance. Hence these tests are called tests of significance.

There are 2 types of hypotheses:

(1) Null Hypotheses: (Hypotheses of No difference)

The hypothesis which assumes that there is no significant difference between sample statistic and the population parameter is called the Null Hypothesis & it is denoted by Ho.

(ii) <u>Alternate Hypothesis</u> : Under this hypothesis, we assume that there is a significant difference between sample & population aspects

Type I & Type II Errors :

<u>Type I Error</u> : if a true hypothesis is rejected

<u>Type II Error</u> : if a false hypothesis is selected.

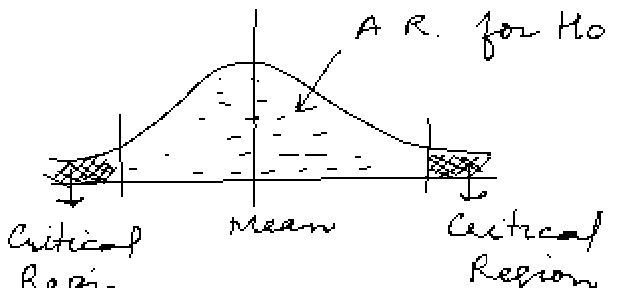
Hypothesis	Decision: Accepted	Decision: Rejected
True	✓	Type I error
False	Type II error	✓

<u>Level of Significance</u> : The probability level below which we reject the hypothesis is called the level of significance (l.o.s). It can be 1%, 2%, 5%, 10% etc. In general, 1% & 5% l.o.s. are considered. For this, we refer to the table of values under the columns 0.01 or 0.05 resp

Critical Region and Acceptance Region:

C.R. → Region that corresponds to the rejection of hypotheses

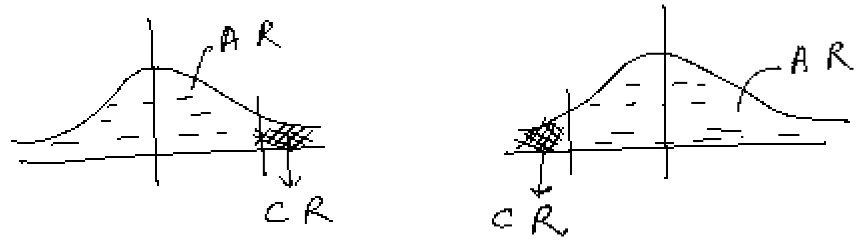
A.R. → " " " " " acceptance " "



One tailed and Two tailed Tests :

A test on statistical hypothesis where the alternate hypothesis is one sided is called one tailed test

Here $H_0 : \mu = \mu_0$; $H_1 : \mu > \mu_0$ or $\mu < \mu_0$

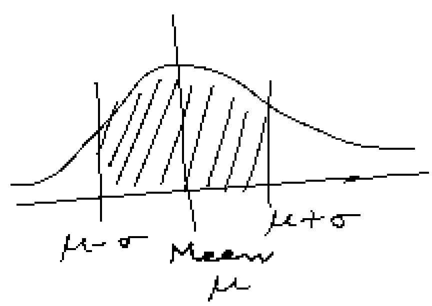


A test on statistical hypothesis where the alternate hypothesis is two sided is called two tailed test.

Here $H_0 : \mu = \mu_0$; H_1 $\mu \neq \mu_0$

Normally, we use 2 tailed test. If we are using a one tailed test, then for the l.o.s. of α, we have to refer to 2α value in the 2 tailed table.

Confidence Intervals : If μ & σ are the mean & s.d of the sampling distribution of a test statistic S, then we can expect to find S in the intervals $\mu-\sigma$ to $\mu+\sigma$, $\mu-2\sigma$ to $\mu+2\sigma$ and $\mu-3\sigma$ to $\mu+3\sigma$ about 68.27 %, 95.45 % & 99.73 % resp.



$x = \mu - \sigma$

$$z = \frac{x-\mu}{\sigma} = \frac{\mu-\sigma-\mu}{\sigma} = -1$$

$$x = \mu+\sigma, \quad z = \frac{\mu+\sigma-\mu}{\sigma} = 1$$

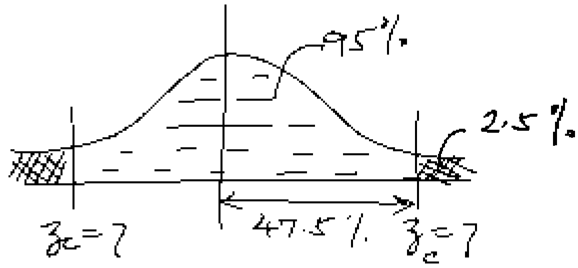
$$P(\mu-\sigma < x < \mu+\sigma) = P(-1 < z < 1) = 2\,P(0 < z < 1) = 2(0.3413)$$
$$= 0.6826$$

For 1% & 5% l.o.s.

For 5% : $A(z) = 0.475$

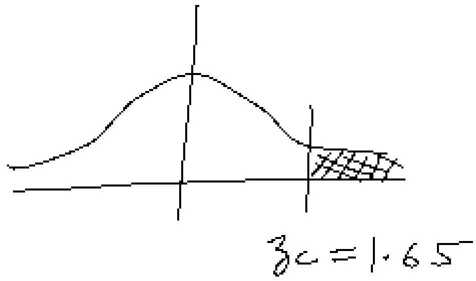
$\Rightarrow z_c = 1.96$

For 1% $z_c = 2.58$



5%	−1.96	1.96
1%	−2.58	2.58

One tailed test :



5%

Test for large samples (Z-test).

Test of Significance for Single Mean :

(to test whether the difference between sample mean & population mean is significant or not)

Under null hypothesis, $z = \dfrac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

where $\bar{x}$ = sample mean

μ = population mean

σ = s.d. of the population

n = sample size

If σ is not known, we use $z = \dfrac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

s = sample s.d.

If the l.o.s. is α and z_α is the critical value, then

$$-z_\alpha < z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_\alpha$$

Also, the limits of the population mean μ are given by

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

For 5% l.o.s. $\quad \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$

For 1% l.o.s $\quad \bar{x} - 2.58 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 2.58 \frac{\sigma}{\sqrt{n}}$

Ex A normal population has a mean of 6.8 & s.d. of 1.5. A sample of 400 members gave a mean of 6.75. Is the difference significant?

Given $\mu = 6.8$, $\sigma = 1.5$, $n = 400$, $\bar{x} = 6.75$

Ho: $\bar{x} = \mu$ i.e. the sample mean & population mean are same

$$|z| = \left| \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \right| = \left| \frac{6.75 - 6.8}{\frac{1.5}{\sqrt{400}}} \right| = |-0.667| = 0.667$$

$$< 1.96$$

$\Rightarrow$ Ho is accepted at 5% l.o.s

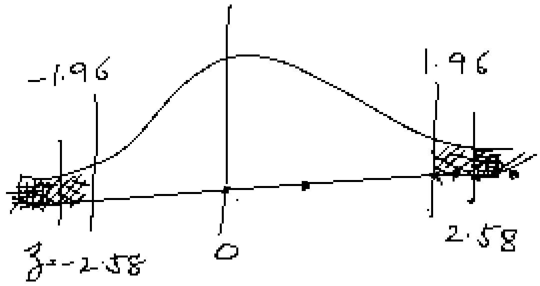
Ex The mean weight obtained from a random sample of size 100 is 64 gm, the s.d. of the weight dist. in the population is 3gm. Can we say that the mean weight of the population is 66 gm at 5% l.o.s?

Also set up 99% Confidence interval for the mean weight of the population.

Given $\bar{x} = 64$, $n = 100$, $\sigma = 3$, $\mu = 66$

Ho: $\mu = \bar{x}$; $\quad z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{64-66}{\frac{3}{\sqrt{100}}} = -6.667$

As $|z| = 6.667 > 1.96 \Rightarrow$ Ho rejected



99% confidence limits are

$$= \bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}}$$

$$= 64 \pm 2.58 \frac{3}{\sqrt{100}} = 64 \pm 0.774$$

$$= (63.226, \ 64.774)$$

HW

Ex. (1) Sugar is packed in bags by an automatic machine with mean contents of a bag as 1.0 kg. A random sample of 36 bags is selected & mean is found to be 0.997 kg. If a s.d. of 0.01 kg is acceptable, determine if the machine needs adjustment

(2) The average zinc concentration recovered from a sample of zinc measurements in 40 locations is found to be 2.54 gm per millilitre. Find the 95% C.I. for the mean zinc concentration in the river assuming the s.d. is 0.32 gm.

Note that If $|z| < z_c$, H_0 is accepted

$|z| > z_c$, H_0 is rejected

Z-test for Difference between Two Means :

Let $\bar{x}_1, \bar{x}_2$ be the means of 2 samples of sizes n_1 & n_2 from populations with means μ_1, μ_2 and s.d.'s σ_1, σ_2.

Then $z = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ is a standard normal variate

Under the null hypothesis $\mu_1 = \mu_2$,

$$\Rightarrow \quad z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \quad \text{———} \; (*)$$

Remarks (i) If σ_1^2 & σ_2^2 are unknown, then $\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$ is used as an estimate of S.E. of $\bar{x}_1 - \bar{x}_2$

& then $(*) \Rightarrow$
$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

(ii) If the two populations have the same variance σ^2, then
$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

Ex Two random samples of 100 students each of two schools A & B were drawn. The CGPA of students from school A had mean 2.82 with s.d. 0.63 and that of school B had mean 2.43 with s.d. 0.65. Does the data indicate any difference in the mean CGPA from two schools?

Given $n_1 = n_2 = 100$, $\bar{x}_1 = 2.82$, $s_1 = 0.63$

$\bar{x}_2 = 2.43$, $s_2 = 0.65$

Ho : $\mu_1 = \mu_2$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{2.82 - 2.43}{\sqrt{\frac{(0.63)^2}{100} + \frac{(0.65)^2}{100}}} = 4.308$$

As z_c (5 % l.o.s.) = 1.96

z_c (1% l.o.s.) = 2.58

Since $z > z_c$ $\Rightarrow$ Ho is rejected

$\Rightarrow$ There is a difference in the mean CGPA of 2 schools

Ex. The mean heights in two large samples of 1000 and 2000 men are 67.5 inches & 68.0 inches resp. Can the samples be regarded as drawn from the same normal population with s.d 2.5 inches.

Ho : $\mu_1 = \mu_2$ $$z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{67.5 - 68.0}{2.5\sqrt{\frac{1}{1000} + \frac{1}{2000}}} = -5.16$$

$|z| = 5.16$ $> z_c$ (for 1% & 5% l.o.s) $\Rightarrow$ Ho rejected

Ex. A random sample of 50 electric light tubes of type A gave mean of 1400 hrs & s.d. 200 hrs and type B " " " 1200 hrs & s.d. 100 hrs.

Is there any significant difference between the two types of tubes?

Here $n_1 = n_2 = 50$

$H_0 : \mu_1 = \mu_2$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{1400 - 1200}{\sqrt{\frac{(200)^2}{50} + \frac{(100)^2}{50}}} = 6.32$$

As $z > z_c$ (for 1% & 5% l.o.s), H_0 is rejected

Ex. The mean marks of 2 sections are as given:

Section A : 32 students have average 72 with s.d. 8

section B . 36 " " " 75 with s.d. 6

Can we say that section B is better than section A?

$H_0 : \mu_1 = \mu_2$ i.e. there is no significant diff between the 2 sections

$(H_1 : \mu_2 > \mu_1)$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{72 - 75}{\sqrt{\frac{(8)^2}{32} + \frac{(6)^2}{36}}} = -1.73$$

$|z| = 1.73$ $\qquad z_c$ (5% l.o.s, one tailed) $= 1.65$

z_c (1% l.o.s, one tailed) $= 2.33$

$\Rightarrow H_0$ is accepted at 1% l.o.s., rejected at 5% l.o.s.

26.5.21

Test of Significance for Single Proportion :

This test is used to find a significant difference between proportion of the sample and that of the population.

$$z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \quad \text{———} \quad (*)$$

where p = observed (sample) proportion of success

P = population proportion of success

$Q = 1-P$ = " " " failure

n = sample size

Remarks:

1. The probable limits for the observed proportion of successes are; $P \pm z_\alpha \sqrt{\frac{PQ}{n}}$; z_α = significant value of z at l.o.s. α

2. If P is not known, the limits for the proportion in the population are : $p \pm z_\alpha \sqrt{\frac{pq}{n}}$

Ex. A coin was tossed 400 times & the heads turned up 220 times. Test the hypothesis that the coin is unbiased.

Ho : Coin is unbiased i.e. $P = 0.5$, $Q = 1-0.5 = 0.5$

$p = \frac{220}{400} = \frac{11}{20}$; Under Ho $(*) \Rightarrow$ $z = \dfrac{\frac{11}{20} - \frac{1}{2}}{\sqrt{\frac{(\frac{1}{2})(\frac{1}{2})}{400}}} = 2$

$z_c = \begin{cases} 1.96 \text{ at } 5\% \text{ l.o.s} \\ 2.58 \text{ at } 1\% \text{ l.o.s} \end{cases}$ → Accept Ho at 1% l.o.s.
Reject Ho at 5% l.o.s

Ex In a sample of 1000 people in a state, 540 are rice eaters & the rest are wheat eaters. Can we assume at 1% & 5% l.o.s that both rice & wheat are equally popular in the state?

Ho : Rice & wheat are equally popular i.e. $P = \frac{1}{2}$

$p = \frac{540}{1000} = 0.54$, $q = 1 - 0.54 = 0.46$, $n = 1000$

$$z = \frac{0.54 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{1000}}} = 2.5298 \qquad z_c = \begin{cases} 1.96 & \text{at } 5\% \text{ l.o.s} \\ 2.58 & \text{at } 1\% \text{ l.o.s} \end{cases}$$

$\Rightarrow$ Accept Ho at 1% l.o.s. ($z < z_c$)
Reject Ho at 5% " ($z > z_c$)

Ex During testing in a sample of 300 chips, 10 are found to be **defective**. Can the manufacturer's claim that 2% of the chips are defective be accepted?

Ho: $p = P$ i.e. $p = 0.02$

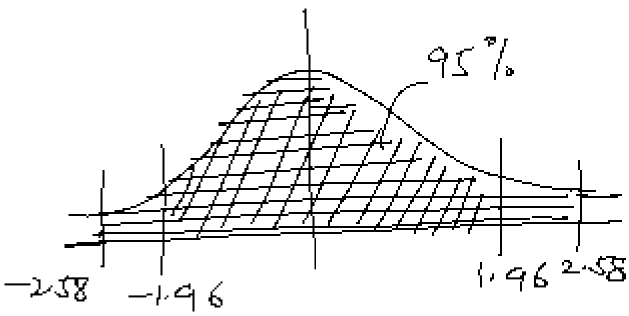
$n = 300$, $p = \frac{10}{300} = \frac{1}{30}$, $\quad P = 0.02$, $Q = 1 - 0.02 = 0.98$

$$z = \frac{\frac{1}{30} - 0.02}{\sqrt{\frac{(0.02)(0.98)}{300}}} = 1.6495 \quad < z_c \; (= 1.96 \text{ at } 5\% \text{ l.o.s})$$

$\Rightarrow$ Ho is **accepted**.

Ex. HW. ① A manufacturer claims that only 4% of his products are defective. A random sample of 600 products contained 36 defectives. Test the manufacturer's claim.

② A die is thrown 9000 times and a throw of 3 or 4 is observed 3240 times. Can the die be regarded as unbiased w.r.to 95% confidence interval?



$z \in (-2.58 \text{ to } 2.58)$ cover 99%

Test of Significance for Difference of Proportions:

$$z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where } P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}, \quad Q = 1 - P$$

Ex A machine produced 16 defectives in a batch of 500. After overhauling, it produced 3 defectives in a batch of 100. Has the machine improved?

Given $n_1 = 500$, $p_1 = \frac{16}{500}$, $n_2 = 100$, $p_2 = \frac{3}{100}$

$H_0: p_1 = p_2$ $\qquad H_1: p_2 < p_1$ (one tailed test)

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{(500)\left(\frac{16}{500}\right) + (100)\left(\frac{3}{100}\right)}{500 + 100} = \frac{19}{600}$$

$$\Rightarrow Q = 1 - P = 1 - \frac{19}{600} = \frac{581}{600}$$

$$z = \frac{\frac{16}{500} - \frac{3}{100}}{\sqrt{\left(\frac{19}{600}\right)\left(\frac{581}{600}\right)\left\{\frac{1}{500} + \frac{1}{100}\right\}}} = 0.10426$$

$z_c = 1.65$ 5% l.o.s.

Inference ?? $\Rightarrow$ H_0 is accepted

$\Rightarrow$ The machine hasn't improved with overhauling

Tests of Significance of Small Samples :

Degrees of Freedom : D.F. are the number of independent values that a statistical analysis can estimate.

I.e. it is the number of values that are free to vary as we estimate parameters.

Typically, the d.f. is equal to the sample size minus the number of parameters we need to calculate.

Tests for Small Samples. $n < 30$

Student's t-test :

(Test of significance of the mean of a small sample)

(Given by GOSSET)

let x_i $(i = 1, 2, \dots, n)$ be a random sample of size n from a normal population with mean μ & variance σ^2.

Ho : There is no significant difference between the sample mean $(\bar{x})$ & the population mean $(\mu$

Then
$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \quad \text{where } s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \text{——} (*)$$

s^2 is an unbiased estimate of the population variance σ^2 & it follows t-distribution with $(n-1)$ d.f

Confidence limits for μ : (by t-test)

C. limits for μ for α l.o.s $= \bar{x} \pm \frac{s}{\sqrt{n}} t_\alpha$

Ex. A random sample of size 16 has mean 53. The sum of squares of the deviation from mean is 135. Can this sample be regarded as taken from a population having 56 as mean? Also obtain 95% & 99% confidence limits for the mean of the population.

H_0 : sample mean = population mean $\boxed{H_1 \quad \mu \neq 56}$

Given : $n = 16, \ \bar{x} = 53, \ \mu = 56, \ \Sigma (x-\bar{x})^2 = 135$

$$s^2 = \frac{1}{n-1} \Sigma (x-\bar{x})^2 = \frac{1}{16-1} (135) = 9, \Rightarrow s = 3$$

$$t = \frac{53-56}{\frac{3}{\sqrt{16}}} = -4, \quad |t| = 4, \quad d.f. = 16-1 = 15$$

$t_{0.05}(15 d.f.) = 2.131$

As $t > t_c \Rightarrow H_0$ is rejected

95% C. I. $= \bar{x} \pm \frac{s}{\sqrt{n}} (2.131)$

$$= 53 \pm \frac{3}{\sqrt{16}} (2.131) = (51.4018, 54.5982)$$

99% CI $= \bar{x} \pm \frac{s}{\sqrt{n}} (2.947) = 53 \pm \frac{3}{\sqrt{16}} (2.947)$

$$= (50.7898, 55.2102)$$

Ex A sample of 20 items has mean 42 units and s d 5 units. Test the hypothesis that it is a random sample from a normal population with mean 45 units

Given $n=20$, $\bar{x}=42$, $S=5$, $(\mu=45)$, $d.f.=20-1=19$

$S \downarrow$ (Sample s.d.)

Ho: $\mu=45$

$$\boxed{s^2=\frac{n}{n-1}S^2} \Rightarrow s^2=\frac{20}{20-1}(5)^2=26.3157$$

$$s=5.1298$$

$$t=\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}=\frac{42-45}{\frac{5.1298}{\sqrt{20}}}=-2.6153$$

$\Rightarrow |t|=2.6153$, $t_{0.05}(19\,d.f.)=2.093$

As $t>t_{0.05}(19\,d.f)$ $\Rightarrow$ Ho is rejected

H.W check at 1% l.o.s

Ex A machine is expected to produce nails of
H.W length 3 cm. A random sample of 25 nails gave an average length of 3.1 cm with s.d 0.3 cm. Can it be said that the machine is producing nails as per specification at 5% l.o.s ?

Ex A random sample of 10 students have I.Q.'s 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Does this data support the assumption of a population mean I.Q. of 100?

H_0: $\mu = \bar{x}$

Given $n = 10$, $\bar{x} = \frac{1}{10}[70+120+110+ \cdots +100] = 97.2$

$$s^2 = \frac{1}{n-1}\sum_i (x_i-\bar{x})^2 = \frac{1}{10-1}\left[(70-97.2)^2+(120-97.2)^2+ \cdots + (100-97.2)^2\right]$$

$$= \frac{1}{9}[1833.6] = 203.7333$$

$\Rightarrow s = 14.2735$

$$\therefore \quad t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{97.2-100}{\frac{14.2735}{\sqrt{10}}} = -0.6203$$

$\Rightarrow$ $|t| = 0.6203$ $\qquad t_{0.05}(9\ d.f.) = 2.262$

As $|t| < t_{0.05}$, $\Rightarrow$ H_0 is accepted

Ex. The life time of electric bulbs for a random sample of 10 from a large consignment gave the following table.

Item :	1	2	3	4	5	6	7	8	9	10
life in 1000 hrs :	4.2	4.6	3.9	4.1	5.2	3.8	3.9	4.3	4.4	5.6

Can we accept the hypothesis that the average life time of bulbs is 4500 hrs?

H_0: $\bar{x} = \mu$ ie av. life of bulbs is 4500 hrs

$$\bar{x} = \frac{1}{10}\left[1.2 + 4.6 + --- \; 5.6\right] \times 1000 = 4100 \text{ hrs.}$$

$$s^2 = \frac{1}{10-1}\left[(1.2-4.1)^2 + (4.6-4.1)^2 + \cdots + (5.6-4.1)^2\right](1000)^2$$

$$= 1.3844 \times 10^6$$

$$\Rightarrow \quad s = 1.1789 \times 10^3 = 1178.9$$

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{4100 - 4500}{\frac{1178.9}{\sqrt{10}}} = -1.0729 \quad , \quad |t| = 1.0729$$

$$t_{0.05}(9 \text{ d.f.}) = 2.262$$

As $|t| < t_{0.05}(9\text{ d f})$ $\Rightarrow$ Ho is accepted

Ex H.W ① The height of 12 men from a city is

70, 67, 62, 68, 61, 68, 70, 64, 65, 69, 71, 68 inches

Is it reasonable to believe that the average height of men in the city is 65 inches?

② The survival time in days of a random sample of 9 mice given a vaccination are

11.7, 10.5, 11.2, 12.9, 12.7, 10.3, 10.4, 10.9, 11.3 days

Does the result suggest a mean survival time of 12.5 days? Construct a 99% C.I.

Two Sample t-test :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_x^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad ; \quad s_x^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

where s_1^2 & s_2^2 are the estimates of the variances of the I & II populations.

$\bar{x}_1, \bar{x}_2 \rightarrow$ means of the 2 samples

The t-statistic given above follows (n_1+n_2-2) degrees of freedom.

Ex Two samples showed the following results:

Sample A : 44 44 56 46 47 47 58 53 49 55

Sample B : 35 38 37 32 40 39 36 41

Test if the averages of the two populations is the same.

$H_0 : \mu_1 = \mu_2 \qquad (H_1 : \mu_1 \neq \mu_2)$

$\bar{x}_1 = 49.9 \quad , \quad \bar{x}_2 = 37.25$

$$s_1^2 = \frac{1}{n_1-1}\sum_i (x_{i1} - \bar{x}_1)^2$$

$$= \frac{1}{9}\left[(44-49.9)^2 + (44-49.9)^2 + (56-49.9)^2 + \cdots + (55-49.9)^2\right]$$

$$= 26.7666$$

$$s_2^2 = \frac{1}{7}\left[(35-37.25)^2 + \cdots + (41-37.25)^2\right] = 8.5$$

$$\therefore \quad s_x^2 = \frac{9(26.7666) + 7(8.5)}{10+8-2} = 18.775$$

$$\therefore \quad t = \frac{49.9 - 37.25}{\sqrt{(18.775)\left(\frac{1}{10}+\frac{1}{8}\right)}} = 6.155$$

<u>As</u> $t_{0.05}$ (16 d.f.) $= 2.120$

$\Rightarrow$ <u>Ho is rejected</u>

<u>Ex.</u> A group of 9 boys were fed on diet A & another

H.W. group of 8 boys were fed on diet B for a period of 6 months, the following increase in weights were then recorded. Test whether diets A & B are significantly different in terms of increase in weight

Diet A :	5	6	8	1	10	4	3	9	6
Diet B :	2	3	6	8	9	1	2	7	

F - Distribution :

becomes relevant when we want to compare the variances of two samples

let $X = \{x_1, x_2, \ldots, x_n\}$, $Y = \{y_1, y_2, \ldots, y_m\}$ be the values of two independent random samples drawn from the same normal population, with variance σ^2.

Then
$$F = \frac{s_1^2}{s_2^2} \quad \text{where } s_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$
$$\& \quad s_2^2 = \frac{1}{m-1} \sum_{j=1}^{m} (y_j - \bar{y})^2$$

$\bar{x}$ & $\bar{y}$ are the sample means of X & Y

s_1^2 & s_2^2 are the unbiased estimates of the population variance

Remark: We always consider the larger of the two values of s_1^2 & s_2^2 in the numerator

Ex. Two samples of sizes 8 & 9 give a sum of squares of deviations from their respective means equal to 91 inches^2 & 160 inches^2. Can these be regarded as drawn from the normal populations with the same variance?

Given $n = 8$, $m = 9$

$\Sigma (x - \bar{x})^2 = 91$, $\Sigma (y - \bar{y})^2 = 160$

$s_1^2 = \frac{1}{8-1}(91) = 13$; $s_2^2 = \frac{1}{9-1}(160) = 20$

H_0 : $\sigma_1^2 = \sigma_2^2$ i.e. the two samples come from the same normal pop with var. σ^2

$$F = \frac{s_2^2}{s_1^2} = \frac{20}{13} = 1.5385 \quad \text{with } (9-1, 8-1) \text{ d.f.}$$
is. $(8, 7)$ d.f.

$$F_{0.05}(8, 7 \text{ d.f}) = 3.73 \qquad \begin{pmatrix} \nu_1 = 8 \\ \nu_2 = 7 \end{pmatrix}$$

As $F < F_{0.05}(8, 7 \text{ d.f})$

$\Rightarrow$ Ho is accepted.

Ex Two independent samples of sizes 7 & 6 have the following values:

Sample A :	28	30	32	33	33	29	34	(X)
Sample B :	29	30	30	24	27	29		(Y)

Examine if samples have been taken from normal populations having the same variance.

Ho: $\sigma_1^2 = \sigma_2^2$

Means of the samples: $\bar{x} = 31.286$, $\bar{y} = 28.167$

$$s_1^2 = \frac{1}{7-1} \sum_i (x_i - 31.286)^2 = 5.238$$

$$s_2^2 = \frac{1}{6-1} \sum_j (y_j - 28.167)^2 = 5.367$$

$$F = \frac{5.367}{5.238} = 1.0245 \quad (5, 6 \text{ d.f.}) \qquad \begin{pmatrix} \nu_1 = 5 \\ \nu_2 = 6 \end{pmatrix}$$

$$F_{0.05}(5, 6 \text{ d.f.}) = 4.39$$

As $F < F_{0.05}(5, 6 \text{ d.f})$ $\Rightarrow$ Ho is accepted

Ex. Can we conclude that the two population variances are equal for the following data of students who passed out from 'state' or 'private' universities

H.W

State : 8350 8260 8130 8340 8070

Private : 7890 8140 7900 7950 7840 7920

Ex. Two random samples of size 9 & 13 have s.d.'s 2.1 & 1.8 respectively. Can these be considered to have been drawn from the same normal population?

Given $n = 9$, $S_1 = 2.1$; $m = 13$, $S_2 = 1.8$

$$s_1^2 = \frac{n}{n-1} S_1^2 \qquad ; \qquad s_2^2 = \frac{m}{m-1} S_2^2$$

$$= \frac{9}{8}(2.1)^2 \qquad\qquad = \frac{13}{12}(1.8)^2$$

$$= 4.961 \qquad\qquad = 3.51$$

$$F = \frac{4.961}{3.51} = 1.41 \quad (8, 12 \text{ d.f.})$$

$$F_{0.05}\ (8, 12 \text{ d.f.}) = 2.85$$

$H_0 : \sigma_1^2 = \sigma_2^2$

As $F_{calculated} < F_{0.05}\ (8, 12 \text{ d.f.})$

$\Rightarrow$ H_0 is accepted

Ex. A teacher teaches in two sections A & B for the same subject. Section A has 16 students & B has 25 students, In an exam, though there was no significant difference in mean grades, class A has a s.d of 9 whereas B has 12.

Can we conclude that variability in B is higher than A?

<u>Given</u> $n = 16,\ S_1 = 9$; $m = 25,\ S_2 = 12$

<u>Ho</u> $\sigma_1^2 = \sigma_2^2$

$s_1^2 = \frac{16}{16-1}(9)^2$ $\qquad$ $s_2^2 = \frac{25}{25-1}(12)^2$

$= 86.4$ $\qquad$ $= 150$

$$F = \frac{s_2^2}{s_1^2} = \frac{150}{(86.4)^2} = 1.736 \quad (24,\ 15\ d.f.)$$

$F_{0.05}\ (24, 15\ d.f.) = 2.29$

As $F_{calculated} < F_{0.05}\ (24, 15\ d.f.)$

$\Rightarrow$ Ho is <u>accepted</u>

<u>Ex</u> The daily wages of skilled labour in two
H.W.
cities is given below.

<u>City</u>	<u>Size of Sample</u>	<u>S.D.</u>
A	16	25
B	13	32

Can we say that there is more variability in city B for daily wages?

Chi Square Test:

(Non-parametric test of significance)

It is used for testing:

(i) Independence of Attributes

(ii) Goodness of Fit

$$\chi_i^2 = \frac{(O_i - E_i)^2}{E_i}, \quad \chi^2 = \sum_i \chi_i^2$$

where O_i – observed frequency

E_i – expected frequency

χ^2 follows Chi-square dist. with $(r-1)(c-1)$ d.f.

Ex In a health survey, 400 individuals were asked if they were vaccinated against flu and whether they were subsequently attacked by flu. The following data was obtained. Test whether vaccination prevents flu.

	Vaccinated	Not vaccinated	Total
Attacked by flu	60	85	145
Didn't get flu	190	65	255
Total	250	150	400

H_0: The attributes are independent

ie vaccination does not prevent flu.

If the attributes are independent,

then

obs: $$P(A \cap B) = P(A)\,P(B)$$

$$P(A \cap B) = \frac{60}{400} = 0.15$$

$$P(A)\,P(B) = \left(\frac{250}{400}\right)\left(\frac{145}{400}\right) = 0.2265$$

How to calculate the expected values:

vaccinated, attacked by flu $= \left(\frac{250}{400}\right)\left(\frac{145}{400}\right) \times 400$

$= 90.625$

To calculate χ^2

Cell i	Observed freq. (O_i)	Expected frequency (E_i)	$\chi_i^2 = \frac{(O_i - E_i)^2}{E_i}$
1	60	$\frac{250 \times 145}{400} = 90.625$	10.349
2	85	$\frac{150 \times 145}{400} = 54.375$	17.248
3	190	$\frac{255 \times 250}{400} = 159.375$	5.885
4	65	$\frac{255 \times 150}{400} = 95.625$	9.808
			$\sum_i \chi_i^2 = 43.29$

$\chi^2 = 43.29$ $\quad r = 2,\ c = 2$

$(r-1)(c-1)$ d.f. $= 1$ d.f.

$\chi^2_{0.05}(1\text{ d.f.}) = 3.84$

As $\chi^2_{\text{calculated}} > \chi^2_{0.05}(1\text{ d.f})$

$\Rightarrow$ H_0 is rejected

$\Rightarrow$ Catching flu is not indept of vaccination

i.e. Vaccination prevents flu.

Ex. Perform Chi-square test of independence of attributes for the following data.

Eye colour	Hair colour: Black	Hair colour: Grey	Total
Black	40	20	60
Blue	20	30	50
Brown	60	30	90
	120	80	$\boxed{200}$

Ho: The attributes are independent

Class (i)	O_i	E_i	$\chi_i^2 = \frac{(O_i - E_i)^2}{E_i}$
1	40	$\frac{60 \times 120}{200} = 36$	0.4444
2	20	$\frac{60 \times 80}{200} = 24$	0.6667
3	20	$\frac{50 \times 120}{200} = 30$	3.3333
4	30	$\frac{50 \times 80}{200} = 20$	5.0
5	60	$\frac{90 \times 120}{200} = 54$	0.6667
6	30	$\frac{90 \times 80}{200} = 36$	1.0
			$\boxed{\chi^2 = 11.1111}$

$$d.f. = (r-1)(c-1)$$

$$= (3-1)(2-1) = 2 \text{ d.f}$$

$$\chi^2_{0.05}(2 \text{ d.f.}) = 5.99$$

As $\chi^2_{cal} > \chi^2_{0.05}(2 \text{ d.f})$

$\Rightarrow$ Ho is rejected

$\Rightarrow$ Eye & hair colours are dependent

Chi-Square Test for Goodness of Fit:

Ex A die is thrown 276 times and the results are given in the table

No. appeared on the die :	1	2	3	4	5	6
Frequency :	44	52	42	50	49	39

Test if the die is unbiased.

Ho: The die is unbiased

The expected frequency for each number under the null hypotheses is equal to $\frac{276}{6} (= 46)$

No. on the die	Observed freq (O_i)	Expected Freq (E_i)	$\chi_i^2 = \frac{(O_i - E_i)^2}{E_i}$
1	44	46	0.0869
2	52	46	0.7826
3	42	46	0.3478
4	50	46	0.3478
5	49	46	0.1956
6	39	46	1.0652
			$\chi^2 = 2.8260$

d.f $= 5$

$\chi^2_{0.05}$ (5 d.f.) $= 11.1$

As $\chi^2_{calculated} < \chi^2_{0.05}$ (5 d.f.) $\Rightarrow$ Ho is accepted

ie. die is unbiased.

Ex.
H.W The demand for a particular spare part in a shop was found to vary day to day. In a sample study, the following information was obtained

Days :	Mon	Tue	Wed	Thu	Fri	Sat
No. of parts : (Oi)	124	125	110	120	126	115

Test if the demand depends on the day of the week.

Hint $E_i = \dfrac{124+125+ \cdots 115}{6} = \underline{120}$

d.f. $\underline{5}$

Ex A sample analysis of examination results of 500 students was done 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 that the general examination result is in the ratio 4:3:2:1 for the various categories?

Expected Freq.

Failed $= \frac{4}{10} \times 500 = 200$

III Class $= \frac{3}{10} \times 500 = 150$

II Class $= \frac{2}{10} \times 500 = 100$

I Class $= \frac{1}{10} \times 500 = 50$

Category	O_i	E_i	χ_i^2
Fail	220	200	2
III C	170	150	2.667
II C	90	100	1
I C	20	50	18

$\chi^2 = 23.667$

[Ho : Results are in the ratio 4:3:2:1]

d.f. = 3 ; $\chi^2_{0.05}(3 d.f.) = 7.81$

→ Ho is rejected

Ex A survey of 800 families with 4 children each revealed the following distribution:

No. of boys :	0	1	2	3	4
No of girls :	4	3	2	1	0
No. of families :	32	178	290	236	64

Is the result consistent with the hypothesis that male & female births are equally probable?

Ho: Male & female births are equally probable

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

Expected freq $\quad P(0\text{ boy}, 4\text{ girls}) = {}^4C_0\, p^0 q^{4-0} = \left(\frac{1}{2}\right)^4$

Exp. freq for $(0B, 4G) = 800 \times \left(\frac{1}{2}\right)^4 = 50$

$P(1B, 3G) = {}^4C_1\, p^1 q^{4-1} = 4\left(\frac{1}{2}\right)^4 = \frac{1}{4}$

Exp freq for $(1B, 3G) = 800 \times \frac{1}{4} = 200$

" " " $(2B, 2G) = 800 \times {}^4C_2 \left(\frac{1}{2}\right)^4 = 300$

" " " $(3B, 1G) = 800 \times \frac{1}{4} = 200$

" " " $(4B, 0G) = 800 \times \frac{1}{16} = 50$

No. of boys in the family	O_i	E_i	$\chi_i^2 = \frac{(O_i - E_i)^2}{E_i}$
0	32	50	6.48
1	178	200	2.42
2	290	300	0.3333
3	236	200	6.48
4	64	50	3.92

$\boxed{\chi^2 = 19.633}$

d.f. = 4

$\chi^2_{0.05}$ (4 d.f.) = 9.49

As $\chi^2_{calculated} > \chi^2_{0.05}$ (4 d.f.)

$\Rightarrow$ Ho is rejected

2.6.21

Chi Square Test for Goodness of Fit for Poisson Distribution:

Remark: If a test statistic is being calculated by making use of the given data (e.g. μ in Poisson dist.), then degrees of freedom associated with a sample of size n is $n-2$ (& not $n-1$) since we lose 1 degree of freedom for the calculation of a parameter.

Ex. Fit a Poisson distribution for the given data and test the goodness of fit.

x	0	1	2	3	4
f	419	352	154	56	19

, $\Sigma f = 1000$

Ho: The given data follows Poisson distribution.

As $\mu = \frac{\sum fx}{\sum f} = \frac{(0)419 + (1)352 + (2)(154) + 3(56) + 4(19)}{419 + 352 + 154 + 56 + 19}$

$= 0.904$

As $P(x) = \frac{e^{-\mu}\mu^{x}}{x!} = \frac{e^{-0.904}(0.904)^{x}}{x!}$

$P(x=0) = \frac{e^{-0.904}(1)}{(1)} = 0.4049$

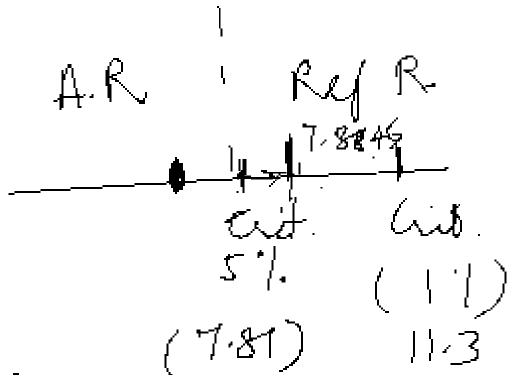
Expected freq for $x=0$ is equal to
$(0.4049)(1000) = 404.9$

x	O_i	E_i	$\chi_i^2 = \frac{(O_i - E_i)^2}{E_i}$
0	419	404.9466	0.4877
1	352	366.0717	0.5408
2	154	165.4644	0.7943
3	56	49.8599	0.7564
4	19	11.2683	5.3056
			$\chi^2 = 7.8846$

$d.f = 5 - 2 = 3$

$\chi^2_{0.05}(3\,d.f.) = 7.81$

$\chi^2_{0.01}(3\,d.f) = 11.3$



Reject H_0 at 5%, Accept at 1% l.o.s

Ex: The number of accidents per day as recorded in a city over a period is given below. Test if the data follows Poisson distribution

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

$$\mu = \frac{\Sigma fx}{\Sigma f} = \frac{0(173)+1(168)+ \cdots + 5(1)}{173+168+37+18+3+1} = 0.7825$$

$\Sigma f = 400$

As $P(x) = \frac{e^{-\mu}\mu^x}{x!} = \frac{e^{-0.7825}(0.7825)^x}{x!}$;

$P(x=0) =$; Exp freq. for $(x=0) =$

$P(x=1) =$ " $(x=1) = 143.12$

$P(x=2) =$

$P(x=3) =$

$P(x=4) =$

$P(x=5) =$

Ho: The data follows Poisson distribution

x_i	O_i	E_i	χ_i^2
0	173	182.9046	
1	168	143.1228	
2	37	55.9968	
3	18	14.6058	
4	3	2.8573	
5	1	0.4472	

d.f. $= 6-1-1 = 4$

$\chi^2_{0.05}(4\ d.f.) = 9.49$

$\Rightarrow$ Ho rejected

$\chi^2 = 12.7811$

Alternate Solution

x	O_i	E_i	χ_i^2
0			
1			
2			
3			
≥ 4	4	$[1-(\ )]\times 400$	..

$d.f. = 4-1=3$

$\chi^2_{0.05}(3\,d.f.) = \ldots$

$\chi^2 = \_\_$

H_0 :

Ex The table below gives the number of students passed & failed by three examiners A, B, C in an exam. Test the hypothesis that the proportion of students failed by the 3 examiners are equal

	A	B	C	Row Total
Passed	50	47	56	153
Failed	5	14	8	27
Column total	55	61	64	G.T=180

H_0: The proportion of students failed by examiners A, B & C are equal

$d.f. = (3-1)(2-1) = 2$

i	O_i	E_i	χ_i^2
1	50	$\frac{153 \times 55}{180} = 46.75$	
2	47	51.85	
3	56	54.4	
4	5	8.25	
5	14	9.15	
6	8	9.6	

$$\chi^2 = 4.8441$$

$$\chi^2_{0.05}\ (2\ d.f.) = 5.99$$

As $\chi^2_{calculated} < \chi^2_{0.05}(2\ d.f.)$

$\Rightarrow$ H_0 is accepted

Ex (H.W) The table below gives the number of good & bad parts produced by each of the three shifts in a factory. Test if the production of bad parts is independent of the shift in which they were produced

Shift	Good parts	Bad parts
Day	960	40
Evening	940	50
Night	950	45

(Similar to the previous one)

Ex. A set of 5 identical coins is tossed 320 times and the results are shown in the table

No. of heads :	0	1	2	3	4	5
Obs. Frequency(O_i) :	16	27	72	112	71	22
Expected(E_i) Freq. :						

Test the goodness of fit for Binomial dist.

Ho : Binomial dist. is a good fit for the data

$p = \frac{1}{2}$, $q = \frac{1}{2}$, $n = 5$

$$\left[B(n, p, x) = {}^5C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{5-x} = {}^5C_x \left(\frac{1}{2}\right)^5 \right] — (*)$$

x_i	O_i	E_i	$\chi_i^2 = \frac{(O_i - E_i)^2}{E_i}$
0	16	10	3.6
1	27	50	10.58
2	72	100	7.84
3	112	100	1.44
4	71	50	8.82
5	22	10	14.4

$\chi^2 = 46.68$

(*) ⇒ For $x = 0$, ${}^5C_0 \left(\frac{1}{2}\right)^5 = \frac{1}{2^5} =$; Freq $(x=0) = \frac{320}{2^5}$
$= 10$

$x = 1$ ${}^5C_1 \left(\frac{1}{2}\right)^5 = 50$

$x = 2$, ${}^5C_2 \left(\frac{1}{2}\right)^5 =$

$\chi^2_{0.05}$ (5 d.f.) = 11.1 ⇒ Ho is rejected

Ex. In 90 throws of a die, face 1 turned up 9 times, face 2 or 3 turned 27 times, face 4 or 5 turned 36 times & 6 turned up 18 times. Test at 5% l.o.s. & 1% l.o.s. if the die is fair

H_0: Die is fair

Face turned	O_i	E_i	χ_i^2
1	9	$\frac{1}{6}\times 90 = 15$	
2 or 3	27	$\frac{2}{6}\times 90 = 30$	
4 or 5	36	$\frac{2}{6}\times 90 = 30$	
6	18	$\frac{1}{6}\times 90 = 15$	

$\boxed{d.f. = 3}$ $\qquad \chi^2 = 4.5$

$\chi^2_{0.05}(3\ d.f.) = 7.81$, $\chi^2_{0.01}(3\ d.f.) = 11.3$

$\Rightarrow$ H_0 is accepted