STATISTICAL INFERENCE-1

Sampling of variables-small samples, students 't' distribution, Chi-square distribution as a test of goodness of fit. F-Distribution.

Sampling of variables for small samples:

In case of large samples, sampling distribution approaches a normal distribution. But in case of small samples, it is not possible to assume that statistics computed are normally distributed. Therefore a new technique has been introduced for small samples which involves the concept of **degree of freedom (d. f).** It is denoted by γ.

Number of degrees of freedom:

It is the number of values in a set which may be assigned arbitrarily. For example, if $x_1 + x_2 + x_3 = 15$ and we assign any values of two of the variables say x_1 and x_2, then the values of x_3 will be obtained. Hence these two variables x_1 and x_2 are the degree of freedom as they are free and independent choices to find the third one x_3.

If there are n observations, the degree of freedom (d. f) are (n - 1).

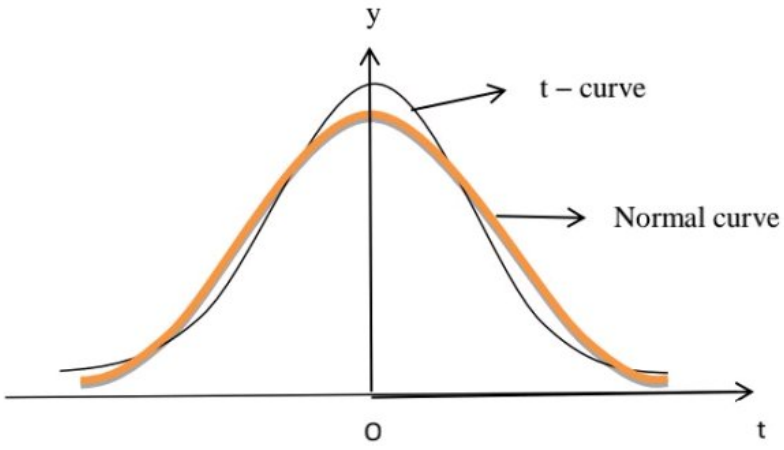
Student's t – distribution:

Consider a small sample of size n, drawn from a normal population with mean μ and standard deviation σ. If $\bar{x}$ and σ_s be the sample mean and standard deviation, then the statistic, "t" is defined as $t = \frac{(\bar{x}-\mu)}{\sigma}\sqrt{n}$ or $t = \left(\frac{\bar{x}-\mu}{\sigma_s}\right)\sqrt{n-1}$, where n -1 denotes the d. f. of t.

If we calculate t for each sample, we obtain the sampling distribution for t. This distribution is called as Student's t – distribution, is given by $y = \frac{y_0}{\left(1+\frac{t^2}{n-1}\right)^{n/2}}$, where y_0 is constant such that the area under the curve is unity.

As $n \to \infty$, $y = \frac{y_0}{e^{t^2/2}}$ $\left[using \lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n = e^a\right]$

i.e. $y = y_0 e^{\left(-\frac{1}{2}\right)t^2}$ which is normal curve. Hence t is normally distributed for large samples.



Scanned by CamScanner

The t – distribution is often used in tests of hypothesis about the mean when the population standard deviation σ is unknown.

Significance Test of a sample mean:

Given a random small sample $x_1, x_2, x_3, \ldots\ldots\ldots, x_n$ from a normal population, we have to test the hypothesis that mean of the population is μ. For this we first calculate $t = \frac{(\bar{x}-\mu)\sqrt{n-1}}{\sigma_s}$, where $\bar{x} = \frac{1}{n}\sum_1^n x_t$, $\sigma_s{}^2 = \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2$ or $\sigma_s{}^2 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2$, then find the value of p for the given d. f. from the table.

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

(ii) If $t > t_{0.01}$, the difference is said to be significant at 1% level of significance.

(iii) If $t < t_{0.05}$, the data is said to be consistent with the hypothesis that μ is the mean of the population.

Problems:

1. A certain stimulus administered to each of 12 patients resulted in the following increases of blood pressure: 5, 2, 8, -1, 3, 0, -2, 1, 5, 0, 4, 6. Can it be concluded that the stimulus will in general be accompanied by an increase in blood pressure.
[From table $t_{0.05} = 2.2$ for d.f 11]

Solution:

Let us assume that stimulus administered does not change the blood pressure. Then, the population mean for change in the blood pressure is $\mu = 0$.

Here n = 12 and d. f = γ = n – 1 = 12-1 = 11

$\therefore\ \bar{x} = \frac{1}{n}\sum x_i = \frac{1}{12}[5 + 2 + 8 - 1 + 3 + 0 - 2 + 1 + 5 + 0 + 4 + 6] = 2.583.$

And $\sigma_s{}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 = \frac{1}{n}\sum x_i{}^2 - (\bar{x})^2.$

$$= \frac{1}{12}[5^2 + 2^2 + 8^2 + (-1)^2 + 3^2 + 0^2 + (-2)^2 + 1^2 + 5^2 + 0^2 + 4^2 + 6^2] - (2.583)^2.$$

$\therefore\ \sigma_s{}^2 = 8.744.$ $\quad\therefore\ \sigma_s = 2.9571.$

Now $t = \frac{(\bar{x}-\mu)}{\sigma_s}\sqrt{n-1} = \frac{(2.583 - 0)}{2.9571}\sqrt{12-1} = 2.897.$ From table, $t_{0.05}$= 2.2 for df 8 = 11.

$\therefore\ t > t_{0.05}$. $\therefore$ The difference between the means μ and $\bar{x}$ is significant at 5% level of significance. Therefore hypothesis is rejected. Hence the stimulus administered will increase in the blood pressure.

Scanned by CamScanner

2. The nine items of a sample have the following values: 45, 47, 50, 52, 48, 47, 49, 53, 51. Does the mean of these differ significantly from the assumed mean of 47.5? (Given $t_{0.05} = 2.26$ for d. f = 8).

Solution:

Here n = 9 and d. f $= \gamma =$ n – 1= 8.

Assume that there is no significant difference between mean of the sample and the assumed mean $\mu = 47.5$ of population.

$\therefore \ \bar{x} = \frac{1}{n}\sum x_i = \frac{1}{9}[45 + 47 + 50 + 52 + 48 + 47 + 49 + 53 + 51] = 49.11.$

And $\sigma_s{}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2.$

$$= \frac{1}{9}[(45 - 49.11)^2 + (47 - 49.11)^2 + (50 - 49.11)^2 + (52 - 49.11)^2 + (48 - 49.11)^2 + (47 - 49.11)^2 + (49 - 49.11)^2 + (53 - 49.11)^2 + (51 - 49.11)^2.$$

$\therefore \ \sigma_s{}^2 = 6.0988. \quad \therefore \ \sigma_s = 2.4695.$

$\therefore \ t = \frac{(\bar{x}-\mu)}{\sigma_s}\sqrt{n-1} = \frac{(49.11 - 47.5)}{2.4695}\sqrt{9-1} = 1.844.$

From the table, $t_{0.05} = 2.31$ for d. f = 8. $\quad \therefore \ t < t_{0.05}$

$\therefore$ The value of t s not significant at 5% level of significance.

$\therefore$ There is no significant difference between $\bar{x}$ and μ. (Hypothesis accepted)

3. A mechanist is making engine parts with axle diameter of 0.7 inch. A random sample of 10 parts shows mean diameter 0.742 inch with a standard deviation of 0.04 inch. On the basis of this sample, would you say that the work is inferior [$t_{0.05} = 2.262$ for d.f = 9]

Solution:

Here $\mu = 0.7, \ \bar{x} = 0.742, \ \sigma_s = 0.04, \ n = 10.$ $\quad \therefore$ df = n – 1 = 9

Assume that the work is not inferior i.e. there is no significant difference between $\bar{x}$ and μ.

$\therefore \ t = \frac{(\bar{x}-\mu)}{\sigma_s}\sqrt{n-1} = \frac{(0.742 - 0.7)}{0.04}\sqrt{10-1} = 3.16.$

From table, $t_{0.05} = 2.262$ for d.f = 9. $\quad \therefore \ t > t_{0.05}$.

$\therefore$ The value of t is significant at 5% level of significance.

$\therefore \ \bar{x}$ differs significantly from μ and hypothesis is rejected. Hence the work is inferior.

4. Find the student t for the following values in a sample of eight: -4, -2, -2, 0, 2, 2, 3, 3; taking the mean of the universe to be zero.

Scanned by CamScanner

Solution:

Here n = 8, $\mu = 0$.

$\therefore\ \bar{x} = \frac{1}{n}\sum x_i = \frac{1}{8}[-4-2-2+0+2+2+3+3] = 0.25.$

And $\sigma_s{}^2 = \frac{1}{n}\sum x_i{}^2 - (\bar{x})^2 = \frac{1}{8}[(-4)^2+(-2)^2+(-2)^2+0^2+2^2+2^2+3^2+3^2] - (0.25)^2.$

$\therefore\ \sigma_s{}^2 = 6.1875. \quad \therefore\ \sigma_s = 2.4875.$

$\therefore\ t = \frac{(\bar{x}-\mu)}{\sigma_s}\sqrt{n-1} = \frac{(0.25-0)}{2.4875}\sqrt{8-1} = 0.2664.$

5. A random sample of 10 boys has the following IQ: 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of the population mean IQ of 100 at 5% level of significance. [use $t_{0.05}(9) = 2.26$]

Solution:

Here $\mu = 100$, n = 10. $\therefore$ df = n – 1 = 10 – 1 = 9.

$\therefore\ \bar{x} = \frac{1}{n}\sum x_i = \frac{1}{10}[70+120+110+101+88+83+95+98+107+100] = 97.2.$

And $\sigma_s{}^2 = \frac{1}{n}\sum(x_i-\bar{x})^2 = \frac{1}{10}(70-97.2)^2 + (120-97.2)^2 + (110-97.2)^2$
$+ (101-97.2)^2 + (88-97.2)^2 + (83-97.2)^2 + (95-97.2)^2$
$+(98-97.2)^2 + (107-97.2)^2 + (100-97.2)^2.$

$\therefore\ \sigma_s{}^2 = 183.36. \quad \therefore\ \sigma_s = 13.54.$

$\therefore\ t = \frac{(\bar{x}-\mu)}{\sigma_s}\sqrt{n-1} = \frac{(97.2-100)}{13.54}\sqrt{10-1} = -0.62. \quad \therefore\ |t| = 0.62.$

From table $t_{0.05} = 2.26$ for d. f = 9. $\quad \therefore |t| < t_{0.05}.$

$\therefore$ The difference between $\bar{x}$ and μ is not significant at 5% level of significance. Therefore hypothesis is accepted. Hence the population mean 100 can be accepted at 5% level of significance.

6. A random sample of size 25 from a normal population has mean $\bar{x} = 47.5$ and standard deviation $\sigma_s = 8.4$. Does this information refute the claim that mean of the population is $\mu = 42.1$.[use $t_{0.05} = 2.06$ for df = 24]

Solution:

Here n = 25. $\therefore$ df = n – 1 = 25 – 1 = 24.

$t = \frac{(\bar{x}-\mu)}{\sigma_s}\sqrt{n-1} = \frac{(47.5-42.1)}{42.1}\sqrt{25-1} = 3.149.$

Scanned by CamScanner

$t_{0.05} = 2.06$ for d. f = 24. $\therefore \quad t > t_{0.05}$.

$\therefore$ The value of t is significant at 5% level of significance.

Hence refute the claim that mean population is $\mu = 42.1$.

Significance Test of Difference between Sample Means:

Consider two independent samples $x_1, x_2, x_3, \ldots\ldots\ldots, x_{n_1}$ and $y_1, y_2, y_3, \ldots\ldots\ldots, y_{n_2}$ with means $\bar{x}$ and $\bar{y}$ and standard deviations σ_x and σ_y from a normal population with the same variance. To test the hypothesis that the population means μ_1 and μ_2 are the same, first find $t = \frac{\bar{x} - \bar{y}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$,

where $\bar{x} = \frac{1}{n_1}\sum_1^{n_1} x_i$ and $\bar{y} = \frac{1}{n_2}\sum_1^{n_2} y_i$ (Here $\mu_1 - \mu_2 = 0$) and

$\sigma_s{}^2 = \frac{1}{n_1+n_2}\left[\sum_1^{n_1}(x_i - \bar{x})^2 + \sum_1^{n_2}(y_i - \bar{y})^2\right]$ where the variable t follows the t – distribution with $n_1 + n_2 - 2$ degrees of freedom.

If $t > t_{0.05}$, the difference between the sample means is significant at 5% level of significance.

If $t > t_{0.01}$, the difference is significant at 1% level of significance.

If $t < t_{0.05}$, the data is consistent with the hypothesis, that $\mu_1 = \mu_2$

Remark:

If two samples are of the same size and the data are paired, then t is defined by

$t = \frac{(\bar{x}-\mu)}{\sigma_s}\sqrt{n-1}$, where $\sigma_s{}^2 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2$.

x_i = difference of x and y values (i.e corresponding values) of the samples.

$\bar{x}$ = mean of the differences.

Problems:

1. Eleven students were given a test in statistics. They were given a month's further tuition and a second test of equal difficulty was held at the end of it. Do the marks give evidence that the students have benefitted by extra coaching? [use $t_{0.05}(10) = 2.228$].

Boys :	1	2	3	4	5	6	7	8	9	10	11
Marks I Test	23	20	19	21	18	20	18	17	23	16	19
Marks II Test	24	19	22	18	20	22	20	20	23	20	17

Solution:

Here n = 11. $\therefore$ d. f = n – 1= 10.

Let x_i = difference in marks = Marks in II Test – Marks in I Test.

Scanned by CamScanner

$\therefore\ x_i = 1,\ -1,\ 3,\ -3,\ 2,\ 2,\ 2,\ 3,\ 0,\ 4,\ -2.$

$\therefore\ \bar{x} = \frac{1}{n}\sum x_i = \frac{1}{11}[1 - 1 + 3 - 3 + 2 + 2 + 2 + 3 + 0 + 4 - 2] = 1.$

And $\sigma_s{}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 = \frac{1}{11}(1-1)^2 + (-1-1)^2 + (3-1)^2 + (-3-1)^2 + (2-1)^2$
$+(2-1)^2 + (2-1)^2 + (3-1)^2 + (0-1)^2 + (4-1)^2 + (-2-1)^2$

$\therefore\ \sigma_s{}^2 = 4.545. \qquad \therefore\ \sigma_s = 2.13.$

Assume that the students have not been benefitted by extra coaching, then that the mean of the difference between the marks of two tests is zero i.e. $\mu = 0$.

Now $t = \frac{(\bar{x}-\mu)}{\sigma_s}\sqrt{n-1} = \frac{(1-0)}{2.13}\sqrt{11-1} = 1.485.$

From table $t_{0.05}$ = 2.228 for df = 10.

$\therefore\ t < t_{0.05}$, the value of t is not significant at 5% level of significance. Therefore hypothesis is accepted.

Hence there is no evidence that the students have benefitted by extra coaching.

2. From a random sample of 10 pigs fed on diet A, the increases in weight in a certain period were 10, 6, 16, 17, 13, 12, 8, 14, 15, 9 lbs. For another random sample 12 pigs fed on diet B, the increases in the same period were 7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17 lbs. Test whether diets A and B differ significantly as regards their effect on increases in weight? [use $t_{0.05}(20) = 2.09$].

Solution:

Here $n_1 = 10,\ n_2 = 12. \quad \therefore df = n_1 + n_2 - 2 = 10 + 12 - 2 = 20.$

$\therefore\ \bar{x} = \frac{1}{n}\sum x_i = \frac{1}{10}[10 + 6 + 16 + 17 + 13 + 12 + 8 + 14 + 15 + 9] = 12.$

$\bar{y} = \frac{1}{n}\sum y_i = \frac{1}{12}[7 + 13 + 22 + 15 + 12 + 14 + 18 + 8 + 21 + 23 + 10 + 17] = 15.$

And $\sum(x_i - \bar{x})^2 = (10-12)^2 + (6-12)^2 + (16-12)^2 + (17-12)^2 + (13-12)^2$
$+(12-12)^2 + (8-12)^2 + (14-12)^2 + (15-12)^2 + (9-12)^2 = 120.$

$\sum(y_i - \bar{y})^2 = (7-15)^2 + (13-15)^2 + (22-15)^2 + (15-15)^2 + (12-15)^2$
$+ (14-15)^2 + (18-15)^2 + (8-15)^2 + (21-15)^2 + (23-15)^2$
$+(10-15)^2 + (17-15)^2 = 314.$

$\therefore\ \sigma_s{}^2 = \frac{1}{n_1+n_2-2}[\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2] = \frac{1}{10+12-2}[120 + 314] = 21.1.$

Scanned by CamScanner

$\therefore \quad \sigma_s = 4.65.$

Assuming that the samples do not differ in weight, then $\mu_1 - \mu_2 = 0$.

$\therefore \quad t = \frac{\bar{x}-\bar{y}}{\sigma_s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{12-5}{4.65\sqrt{\frac{1}{10}+\frac{1}{12}}} = -1.6.$

$\therefore \quad |t| = 1.6$ but $t_{0.05} = 2.09$ for df = 20.

$\therefore \quad |t| < 1.6$. Hence difference between the sample means is not significant.

Hence the two diets do not differ significantly as regards their effect on increase in weight.

Chi – Square (χ^2) Test:

Suppose a coin is tossed 100 times, then we expect theoretically that head will appear 50 times and tail 50 times. But this never happens in practice. i.e., the results obtained in an experiment do not agree with the theoretical results. The magnitude of discrepancy between observation and theory is given by the quantity χ^2 (chi - square). If $\chi^2 = 0$, then the observed and theoretical frequencies completely agree.

As the value of χ^2 increases, the discrepancy between the observed and theoretical frequencies increases.

Definition:

If $O_1, O_2, \ldots \ldots \ldots, O_n$ be the set of observed (experimental) frequencies and $E_1, E_2, \ldots \ldots, E_n$ be the corresponding set of expected (theoretical) frequencies, then χ^2 is defined as $\chi^2 = \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2} + \cdots + \frac{(O_n-E_n)^2}{E_n}$. i.e. $\chi^2 = \sum_{i=1}^{n} \frac{(O_i-E_i)^2}{E_i}$ with n – 1 degrees freedom where n = the total frequency = $\sum O_i = \sum E_i$

Chi – Square distribution:

If $x_1, x_2, \ldots \ldots, x_n$ be n independent normal variate with mean zero and standard deviation unity, then $x_1^2 + x_2^2 + \cdots + x_n^2$, is a random variate having χ^2 – distribution with number of degrees of freedom.

The equation of the χ^2 – curve is $y = y_0\, e^{-\frac{x^2}{2}}(\chi^2)^{\left(\frac{\gamma-1}{2}\right)}$, where $\gamma = n - 1$.

Goodness of fit:

The value χ^2 is used to test whether the deviations of the observed (experimental) frequencies from the expected (theoretical) frequencies are significant or not. It is also used to test how will a set of observations fit a given distribution, therefore χ^2 provides a test of goodness of fit and may be used to examine the validity of some hypothesis about an observed frequency distribution.

Scanned by CamScanner

(i) If $\chi^2 > \chi^2_{0.05}$, the observed value of χ^2 is significant at 5% level of significance. (Hypothesis is rejected)

(ii) If $\chi^2 > \chi^2_{0.01}$, the value is significant at 1% level.

(iii) If $\chi^2 < \chi^2_{0.05}$, it is a good fit and value is not significant (Hypothesis is accepted).

Problems:

1. In experiments on pea breeding, the following frequencies of seeds were obtained

Round and Yellow	Wrinkled and Yellow	Round and Green	Wrinkled and Green	Total
315	101	108	32	556

Theory predicts that the frequencies should be in proportions 9:3:3:1. Examine the correspondence between theory and experiment. [use $\chi^2_{0.05} = 7.815$ for $\gamma = 3$]

Solution:

Here n = 4. ∴ d. f = γ = n – 1 = 4 – 1 = 3.

Observed frequencies are: $O_1 = 315,\ O_2 = 101,\ O_3 = 108,\ O_4 = 32.$

Given proportions of the frequencies is 9:3:3:1. ∴ sum = 16.

∴ The corresponding theoretical(expected) frequencies are

$$E_1 = \frac{9}{16} \times 556 = 313,\ E_2 = \frac{3}{16} \times 556 = 104,\ E_3 = \frac{3}{16} \times 556 = 104,\ E_4 = \frac{1}{16} \times 556 = 35.$$

$$\therefore\ \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(315-313)^2}{313} + \frac{(101-104)^2}{104} + \frac{(108-104)^2}{104} + \frac{(32-35)^2}{35}.$$

$$\therefore\ \chi^2 = \frac{4}{313} + \frac{9}{104} + \frac{16}{104} + \frac{9}{35} = 0.51.$$

From Table, for $\gamma = 3,\ \chi^2_{0.05} = 7.815.$

$\therefore\ \chi^2 < \chi^2_{0.05}$.

Since the calculated value of χ^2 is very much less than $\chi^2_{0.05}$, there is a very high degree agreement between theory and experiment.

2. A die was thrown 60 times and the following frequency distribution was observed

Faces:	1	2	3	4	5	6
f:	15	6	4	7	11	17

Test whether the die is unbiased? [use $\chi_{0.05} = 11.07$ for df = 5].

Scanned by CamScanner

Solution:

Here $N = \sum f_i = 15 + 6 + 4 + 7 + 11 + 17 = 60$.

If the die is unbiased, then every number from 1 to 6 has equal probability $\frac{1}{6}$ of appearing on the face. Therefore, the expected frequency of each of these numbers appearing on the face in 60 throws is $60 \times \frac{1}{6} = 10$. Thus the expected frequencies are: $E_1 = E_2 = E_3 = E_4 = E_5 = E_6 = 10$.

The observed frequencies are: $O_1 = 15,\ O_2 = 6,\ O_3 = 4,\ O_4 = 7,\ O_5 = 11,\ O_6 = 17$.

$\therefore\ \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$.

$\therefore\ \chi^2 = \frac{(15-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(4-10)^2}{10} + \frac{(7-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(17-10)^2}{10} = 13.6.$

Here n = 6. $\therefore$ Number of d. f = n – 1 = 5 and $\chi^2_{0.05} = 11.07$ for d. f = 5.

$\therefore\ \chi^2 > \chi^2_{0.05}$. Therefore the hypothesis is rejected at 5 % level of significance.

NOTE:

Binomial distribution Fit = $N(p+q)^n$, $N = \sum f_i$ = sum of frequencies.
Number of df for binomial distribution = n – 1, for Poisson distribution n – 2 and for normal distribution = n – 3.

3. A set of five similar coins is tossed 320 times and result is

No. of heads	0	1	2	3	4	5
Frequency	6	27	72	112	71	32

Test the hypothesis that the data follow a binomial distribution.
[use $\chi^2_{0.05} = 11.07$ for df = 5].

Solution:

Here n = 6. $\therefore$ d. f = n – 1 = 5.

P = Probability of getting head = $\frac{1}{2}$, q = Probability of getting tail = $\frac{1}{2}$.

Therefore the theoretical frequencies of getting 0, 1, 2, 3, 4, 5 heads are successive terms of the expansion $N(p+q)^x = N(p+q)^5$, where $N = \sum f_i$.

$\therefore$ N = 6 + 27 + 72 + 112 + 71 + 32 = 320.

$\therefore\ N(p+q)^5 = N[p^5 + 5c_1p^4q + 5c_2p^3q^2 + 5c_3p^2q^3 + 5c_4pq^4 + q^5]$.

$= Np^5 + 5c_1Np^4q + 5c_2Np^3q^2 + 5c_3Np^2q^3 + 5c_4Npq^4 + Nq^5$.

Therefore the theoretical frequencies are:

Scanned by CamScanner

$E_1 = Np^5 = 320\left(\frac{1}{2}\right)^5 = 10,$

$E_2 = 5c_1 Np^4q = 5 \times 320 \times \left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right) = 50,$

$E_3 = 5c_2 Np^3q^2 = 10 \times 320 \times \left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 = 100,$

$E_4 = 5c_3 Np^2q^3 = 10 \times 320 \times \left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3 = 100,$

$E_5 = 5c_4 Npq^4 = 5 \times 320 \times \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^4 = 50,$

$E_6 = Nq^5 = 320\left(\frac{1}{2}\right)^5 = 10.$ Clearly $\sum E_i = 320.$

Take $O_1 = 6, O_2 = 27, O_3 = 72, O_4 = 112, O_5 = 71, O_6 = 32.$ Clearly $\sum O_i = 320.$

$\therefore \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(6-10)^2}{10} + \frac{(27-50)^2}{50} + \frac{(72-100)^2}{100} + \frac{(112-100)^2}{100} + \frac{(71-50)^2}{50} + \frac{(32-10)^2}{10} = 78.68.$

From Table, $\chi^2_{0.05} = 11.05$ for df = 5.

$\therefore \chi^2 > \chi^2_{0.05}$ is calculated χ^2 much greater than $\chi^2_{0.05}$.

$\therefore$ The hypothesis that the data follows the binomial distribution is rejected.

4. Fit a Poisson distribution to the following data and test for its goodness of fit at level of significance 0.05. [use $\chi^2_{0.05}(3) = 7.82$].

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

Solution:

Here n = 5, mean $m = \frac{\sum x_i f_i}{N}$, where $N = \sum f_i = 1000.$

$\therefore \quad m = \frac{0(419)+1(352)+2(154)+3(56)+4(19)}{419+352+154+56+19} = 0.904.$

The Poisson distribution fit is $N.\frac{e^{-m}m^x}{x!}$, $\quad x = 0, 1, 2, 3, 4.$

$\therefore$ The theoretical frequencies are $= \frac{1000 \times e^{-0.904}(0.904)^x}{x!}.$

$\therefore \quad E_1 = \frac{1000 \times e^{-0.904}(0.904)^0}{0!} = 40.49, \quad E_2 = \frac{1000 \times e^{-0.904}(0.904)^1}{1!} = 36.6,$

$E_3 = \frac{1000 \times e^{-0.904}(0.904)^2}{2!} = 16.54, \quad E_4 = \frac{1000 \times e^{-0.904}(0.904)^3}{3!} = 4.98,$

$E_5 = \frac{1000 \times e^{-0.904}(0.904)^4}{4!} = 1.12.$

Scanned by CamScanner

Here $E_1 + E_2 + E_3 + E_4 + E_5 = 997.4 \neq N$. Difference is $1000 - 997.4 = 2.6$.

Divide it by 2. $\therefore \frac{2.6}{2} = 1.3$. Add 1.3 to E_1 and E_5 to make sum of $E_i's$ to 1000. i.e. N.

$\therefore$ $E_1 = 40.49 + 1.3 = 41.79$ and $E_5 = 1.12 + 1.3 = 2.42$ such that $N = \sum E_i = 1000$.

$\therefore$ $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$. Here $O_1 = 419$, $O_2 = 352$, $O_3 = 154$, $O_4 = 56$, $O_5 = 19$.

$\therefore$ $\chi^2 = \frac{(419-40.49)^2}{40.49} + \frac{(352-36.6)^2}{36.6} + \frac{(154-16.54)^2}{16.54} + \frac{(56-4.98)^2}{4.98} + \frac{(19-1.12)^2}{1.12}$.

$\therefore$ $\chi^2 = 5.748$.

Here while finding the theoretical distribution, we used two constraints such as mean(m) and N

$\therefore$ df = 5 – 2 = 3. From table, $\chi^2_{0.05} = 7.82$ for df = 3.

$\therefore$ $\chi^2 < \chi^2_{0.05}$

$\therefore$ The agreement between the fact and theory is good and hence the Poisson distribution can be fitted to the data.

5. Fit a normal distribution to the following data of weights of 100 students of Delhi university and test the goodness of fit.

Weight (KG):	**60-62**	**63-65**	**66-68**	**69-71**	**72-74**
Frequencies:	**5**	**18**	**42**	**27**	**8**

[From table $\chi^2_{0.05} = 5.99$ for df $\gamma = 2$.

Solution:

Here N = $\sum f_i = 5 + 18 + 42 + 27 + 8 = 100$ and n = 5 $\therefore$ d.f = n - 3 = 2

We first write the frequency table for continuous class intervals:

Class Interval:	59.5 - 62.5	62.5 – 65.5	65.5 – 68.5	68.5 – 71.5	71.5 – 74.5
Mid pint of C.I.: (x)	61	64	67	70	73
f:	5	18	42	27	8

Mean m = $\frac{1}{N}\sum f_i . x_i = \frac{1}{100}(6745) = 67.45$.

$\sigma^2 = \frac{1}{N}\sum f_i (x_i - \text{m})^2 = \frac{1}{100}(852.75) = 8.5275$.

Scanned by CamScanner

$\therefore \quad \sigma = \sqrt{8.5275} = 2.9202.0$

Table of Calculation:

Class boundaries (x)	$Z=\frac{x-m}{\sigma}=\frac{x-67.45}{2.92}$	Area under normal curve from O to Z. i.e. P(Z)		Area for each class (A)	Expected frequencies (f_e =N x A)
59.5	-2.72	0.4967			
			(-)	0.0413	$E_1 = 4.13$
62.5	-1.70	0.4554			
			(-)	0.2068	$E_2 = 20.68$
65.5	-0.67	0.2486			
			(+)	0.3892	$E_3 = 38.92$
68.5	0.36	0.1406			
			(-)	0.2771	$E_4 = 27.71$
71.5	1.39	0.4177			
			(-)	0.0743	$E_5 = 7.43$
74.5	2.41	0.4920			

[A is obtained by subtracting the successive areas in the third column when the corresponding values of Z have the same sign and adding them when the Z values have opposite signs]

Here $O_1 = 5$, $O_2 = 18$, $O_3 = 42$, $O_4 = 27$, $O_5 = 8$.

$$\therefore \chi^2 = \sum \frac{(Oi-Ei)^2}{Ei} = \frac{(5-4.13)^2}{4.13} + \frac{(18-20.68)^2}{20.68} + \frac{(42-38.92)^2}{38.92} + \frac{(27-27.71)^2}{27.71} + \frac{(8-7.43)^2}{7.43}$$

$$\therefore \chi^2 = 0.1833 + 0.3473 + 0.2437 + 0.0182 + 0.0437 = 0.8362$$

From table $\chi^2_{0.05} = 5.99$ for df $\gamma = 2$. $\therefore \chi^2 < \chi^2_{0.05}$ Hence fit is good.

6. Obtain the equation of the normal curve that may be fitted to the data and test the goodness of fit.

x:	4	6	8	10	12	14	16	18	20	22	24	Total
f(x):	1	7	15	22	35	43	38	20	13	5	1	200

[From table $\chi^2_{0.05} = 15.51$ for $\gamma = 8$]

Solution:

Here N = 200, n = 11. $\therefore$ d.f = n - 3 = 8.

$$m = \frac{1}{N} f_i . x_i = \frac{2770}{200} = 13.85.$$

$$\sigma^2 = \frac{1}{N} \sum fi\,(\varphi i - m)^2 = \frac{1}{200}(2935.5) = 14.6775 \therefore \sigma = 3.83$$

Scanned by CamScanner

Table for Calculation:

Class interval	Class boundary (x)	$Z = \frac{x-m}{\sigma} = \frac{\varphi - 13.85}{3.83}$	Area under normal curve from O to Z	Area for each class(A)	Expected frequencies (f = N × A)
	3	-2.63	0.4976		
3 - 5			(-)	0.0081	$E_1 = 1.62$
	5	-2.31	0.4895		
5 - 7			(-)	0.0271	$E_2 = 5.42$
	7	-1.78	0.4624		
7 - 9			(-)	0.0663	$E_3 = 13.26$
	9	-1.26	0.3961		
9 - 11			(-)	0.1258	$E_4 = 25.16$
	11	-0.74	0.2703		
11 - 13			(-)	0.1833	$E_5 = 36.66$
	13	-0.22	0.0870		
13 - 15			(+)	0.2049	$E_6 = 40.98$
	15	0.30	0.1179		
15 - 17			(-)	0.1759	$E_7 = 35.18$
	17	0.82	0.2938		
17 - 19			(-)	0.1160	$E_8 = 23.2$
	19	1.34	0.4098		
19 - 21			(-)	0.0587	$E_9 = 11.74$
	21	1.86	0.4685		
21 - 23			(-)	0.0228	$E_{10} = 4.56$
	23	2.38	0.4923		
23 - 25			(-)	0.0068	$E_{11} = 1.36$
	25	2.91	0.4981		

$$\chi^2 = \sum \frac{(Oi-Ei)^2}{Ei}$$

$\therefore \chi^2 = 0.2372+0.4605+0.2283+0.3968+0.0751+0.0995+0.2260+0.4413+0.1352+0.0424+0.0952$

$\therefore \chi^2 = 2.4375.$ From table $\chi^2_{0.05} = 2.73$ for df $\gamma = 8$.

$\therefore \chi^2 < \chi^2_{0.05}$. $\therefore$ The fit is quite good at 5% level.

Scanned by CamScanner

F – DISTRIBUTION:

[Introduced by English statistician Prof. R A Fisher]

Let $x_1, x_2, \ldots\ldots\ldots, x_{n_1}$ and $y_1, y_2, \ldots\ldots\ldots, y_{n_2}$ be the values of two independent random samples drawn from the normal populations having equal variances σ^2.

Let $\overline{x}$ and $\bar{y}$ be the sample means and $S_1^2 = \frac{1}{n_1 - 1} \sum_1^{n_1} (x_i - \bar{x})^2$,

$S_2^2 = \frac{1}{n_2 - 1} \sum_1^{n_2} (y_i - \bar{y})^2$ be the sample variances. Then we define F by

$F = \frac{S_1^2}{S_2}, \quad (S_1^2 > {S_2}^2)$ [Larger value must be in the numerator]

This gives F – distribution (Also known as variance ratio distribution)

With $\gamma_1 = n_1 - 1$ and $\gamma_2 = n_2 - 1$ degrees of freedom.

Test of Significances:

(i) If $F > F_{0.05}$ (for df γ_1 and γ_2), the population variances are significantly different at 5% level of significances. Hence hypothesis is rejected.

(ii) If $F > F_{0.01}$, the population variances are significantly different at 1% level of significances. Hence hypothesis is rejected.

(iii) If $F < F_{0.05}$, the population variances are not significantly different at 5% level of significances. i.e., hypothesis is accepted.

Problems:

1. Two sample sizes 9 and 8 give the sum of squares of deviations from their respective means equal to 160 $inches^2$ and 91 $inches^2$ respectively. Can these be regarded as drawn from some normal population? [$F_{0.05}$ =3.73 for $\gamma_1 = 8$, $\gamma_1 = 7$]

Solution:

Given $n_1 = 9$, $n_2 = 8$. $\therefore$ The degrees of freedom $\gamma_1 = n_1 - 1 = 8$, $\gamma_2 = n_2 - 1 = 7$.

$\therefore$ Assume that $\sigma^2 = {\sigma_2}^2$

Also given $\sum (x_i - \overline{x})^2 = 160$ and $\sum (y_i - \overline{y})^2 = 91$.

$\therefore \quad S_1^2 = \frac{1}{n-1} \sum (x_i - \overline{x})^2 = \frac{160}{8} = 20$ and $S_2^2 = \frac{1}{n-1} \sum (y_i - \overline{y})^2 = \frac{91}{7} = 13$.

Scanned by CamScanner

$\therefore$ $F = \frac{S_1^2}{S_2^2} = \frac{20}{13} = 1.54$. We have $F_{0.05} = 3.73$ for df $\gamma_1 = 8$, $\gamma_2 = 7$.

$\therefore$ $F < F_{0.05}$. $\therefore$ The population variances are not significantly different. Thus, the two samples can be regarded as drawn from the two normal populations with the same variances.

2. Measurements on the length of a copper wire were taken is two experiments A and B as below:

A's measurements: 12.29, 12.25, 11.86, 12.13, 12.44, 12.78, 12.77, 11.90, 12.47.

B's measurements: 12.39, 12.46, 12.34, 12.22, 11.98, 12.46, 12.23, 12.06.

Test whether B's measurements are more accurate than A's (The reading taken in both cases being unbiased) [$F_{0.05}$= 3.73 and $F_{0.01}$= 6.84 for df $\gamma_1 = 8$, $\gamma_2 = 7$]

Solution:

Here $n_1 = 9$, $n_2 = 8$ $\therefore$ degrees of freedom $\gamma_1 = n_1 - 1 = 8$, $\gamma_2 = n_2 - 1 = 7$.

Assume that the two populations have the same variances $(\sigma_1^2 = \sigma_2^2)$

$\bar{x} = \frac{1}{n1} \sum x_i = \frac{110.89}{9} = 12.32$, $\bar{y} = \frac{1}{n_2} \sum y_i = \frac{98.14}{8} = 12.27$

$S_1^2 = \frac{1}{n_1 - 1} \sum (x_i - \bar{x})^2$

$= \frac{1}{8}[0.0009 + 0.0049 + 0.2116 + 0.0361 + 0.0144 + 0.2116 + 0.2025 + 0.1764 + 0.0225]$

$\therefore$ $S_1^2 = \frac{1}{8}[0.8809] = 0.1101$.

$S_2^2 = \frac{1}{n_2 - 1} \sum (y_i - \bar{y})^2$

$= \frac{1}{7}[0.0144 + 0.0361 + 0.0049 + 0.0025 + 0.0841 + 0.0361 + 0.0016 + 0.0441]$

$\therefore$ $S_2^2 = \frac{1}{7}[0.2238] = 0.03197 = 0.032$

$\therefore$ $F = \frac{S_1^2}{S_2^2} = \frac{0.1101}{0.032} = 3.4406$. $(S_1^2 > S_2^2)$

$\therefore$ From table $F_{0.05} = 3.73$ and $F_{0.01} = 6.84$ for df $\gamma_1 = 8$, $\gamma_2 = 7$)

$\therefore$ $F < F_{0.05}$ and $F < F_{0.01}$.

$\therefore$ The result is insignificant at 5% and 1% level.

Hence there is no reason to say that B's measurements are more accurate than those of A's.

Scanned by CamScanner

3. In two independent samples of sizes 8 and 10 the sum of the squares of deviations of the sample values from the respective sample means were 84.4 and 102.6. Test whether the difference of variances of the populations is significant or not.
[From table $F_{0.05} = 3.25$ for (7, 9) df]

Solution:

Given $n_1 = 8, n_2 = 10$. ∴ The df are $\gamma_1 = n_1 - 1 = 7$ and $\gamma_2 = n_2 - 1 = 9$.

$\sum (x_i - \bar{x})^2 = 84.4$ and $\sum (y_i - \bar{y})^2 = 102.6$.

Assume the there is no significant difference between population variances. i.e., ${\sigma_1}^2 = {\sigma_2}^2$

$${S_1}^2 = \frac{1}{n_1 - 1}\sum (x_i - \bar{x})^2 = \frac{84.4}{7} = 12.0571$$

$${S_2}^2 = \frac{1}{n_2 - 1}\sum (y_i - \bar{y})^2 = \frac{102.6}{9} = 11.4$$

$$\therefore \quad F = \frac{{S_1}^2}{{S_2}^2} = \frac{12.0571}{11.4} \qquad ({S_1}^2 > {S_2}^2)$$

$$\therefore \quad F = 1.0576 \text{ (From table } F_{0.05} = 3.25 \text{ for } (7,9) \text{ df)}$$

$$\therefore \quad F < F_{0.05} \qquad \therefore \text{ Hypothesis is accepted.}$$

∴ There is no significant difference between the variance of the populations.

4. Two independent sample of sizes 7 and 6 had the following values:

Sample A	**28**	**30**	**32**	**33**	**31**	**29**	**34**
Sample B	**29**	**30**	**30**	**24**	**27**	**28**	**-**

Examine whether the samples have been drawn from normal populations having the same variance.

Solution:

Here $n_1 = 7,\ n_2 = 6,$ ∴ d.f are $\gamma_1 = n_1 - 1 = 6\ , \gamma_2 = n_2 - 1 = 5.$

Assume that the normal populations having the same variance i.e., ${\sigma_1}^2 = {\sigma_2}^2$.

$$\bar{x} = \frac{1}{n_1}\sum x_i = \frac{217}{7} = 31 \quad \text{and} \quad \bar{y} = \frac{1}{n_2}\sum y_i = \frac{168}{6} = 28$$

$${S_1}^2 = \frac{1}{n_1 - 1}\sum (x_i - \bar{x})^2 = \frac{28}{6} = 4.6667$$

$${S_2}^2 = \frac{1}{n_2 - 1}\sum (y_i - \bar{y})^2 = \frac{26}{5} = 5.2$$

$$\therefore \quad F = \frac{{S_2}^2}{{S_1}^2} = \frac{5.2}{4.6667} \qquad ({S_2}^2 > {S_1}^2)$$

Scanned by CamScanner

$\therefore$ $F = 1.1144$. From table $F_{0.05} = 4.39$ for $(5,6)$ df

$\therefore$ $F < F_{0.05}$ $\therefore$ Hypothesis is accepted.

$\therefore$ The samples have been drawn from the normal population with same variance.

Scanned by CamScanner