

① chi-Square Test (χ^2 -test) :

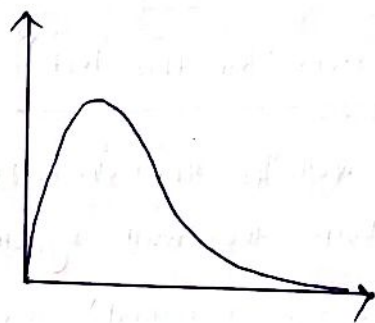
Before going into the deep of this test, let the chi-square distribution be explained. This test is based on χ^2 -distribution.

② chi-square Distribution :

chi-square distribution arises when we do the sum of squares of independent standard normal random variables. and chi-square distribution are getting the shape and spread as Normal distribution when we are going to increase the number of independent standard normal random variables.

Also chi-square distribution is a special case of gamma distribution. It is a continuous distribution —

having pdf as — $f_X(x) = \begin{cases} \frac{(1/2)^{n/2}}{\Gamma(n/2)} e^{-x/2} x^{n/2-1} & 0 < x < \infty \\ 0 & \text{otherwise} \end{cases}$



③ chi-square distribution has only one parameter, that is degree of freedom, which is equals to number of independent standard normal random variables.

④ Mean = n (degree of freedom) and Variance = 2n

☑ This chi-square test is used in various statistical test, one of them is chi-square test.

Let us explain about chi-square test briefly —

⑩ chi-square test is one of the non-parametric statistical test by which we can find out the dependency of two categorical columns or we can ~~state~~ draw conclusion about the equality of theoretical distribution and practical distribution / observed distribution with respect to any categorical column.

With based on those things, chi-square Test can be classified into two types —

- (a) Goodness of Fit (b) Test for Independence.

⑪ Goodness of Fit

This type of chi-square test helps us to find out the relation between observed distribution and theoretical distribution, for any categorical column.

In one word, it helps us to find out that the theoretical distribution and observed distribution, both are same or not.

⑫ How can it be conducted (steps for this test)

① Create the null hypothesis as well as the alternative hypothesis.

② For this test by default the null hypothesis is that both distribution (observed and theoretical) are same.

Alternative will be that they are not same, they are different.

③ Find out the expected value corresponding the categories in that column. (according to the theoretical distribution).

④ Find out the χ^2 -statistics with help of the following formula —

$$\chi^2 = \sum \frac{(\text{Observed Value} - \text{Expected Value})^2}{\text{Expected Value}}$$

④ Means, we need to find out $(\text{Observed Value} - \text{Expected Value}) / \text{Expected Value}$ for each and every categories in that categorical column. Then their sum would be our χ^2 -statistic.

- ④ Based on the ~~p-value~~ χ^2 -statistic and degree of freedom we easily find out the p-value. (also can be done by the method of Region of rejection).
- ⑤ Compare the value with significance level (α). Based on this we can reach our target.

① Numerical Example

- ① Suppose we have a six-sided fair die, and we want to test if the die is indeed fair. We roll the die 60 times and record the number of times each side comes up. We'll use the chi-square Goodness-of-Fit test to determine if the observed frequencies are consistent with a fair die - (i.e. a uniform distribution of the sides).
- Observed frequencies:

| | | |
|--------------------|--------------------|--------------------|
| ① Side 1: 12 times | ② Side 2: 8 times | ③ Side 3: 11 times |
| ④ Side 4: 9 times | ⑤ Side 5: 10 times | ⑥ Side 6: 10 times |

② Answer

⇒ Here our Null Hypothesis (H_0): Observed frequencies are consistent with a fair die.

Alternative hypothesis (H_a): Observed frequencies are consistent with a biased die.

③ Now we need to find out the expected value according to our theoretical distribution (Here uniform distribution).

So, the expected frequencies should be $(\frac{1}{6} \times 60) = 10$ times for each side.

⑩ Now we have to find out the χ^2 -statistics

$$\begin{aligned}\chi^2 &= \frac{(12-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(9-10)^2}{10} + 2 \times \frac{(10-10)^2}{10} \\ &= \frac{4}{10} + \frac{4}{10} + \frac{1}{10} + \frac{1}{10} \\ &= 1\end{aligned}$$

So, we have χ^2 statistic = 1

degree of freedom = $(6-1) = 5$

level of significance = 0.05

Now, from the table we got the value as 11.070 with respect to ~~χ^2 statistic~~ and degree of freedom (5).

level of significance
(0.05)

(Note that we didn't use the P-value).

and we have —

$$1 < 11.070$$

So, we cannot reject our Null Hypothesis.

⑪ For finding P-value we can use the statistical tools in any environment (python, excel etc)

So, P-value corresponding to χ^2 -statistic and df = 5 is 0.962.

So, P-value > level of significance (α) = 0.05

So, we can't reject our Null Hypothesis.

So, for this example, observed frequencies are consistent with a fair die. (Proved)

(All though the strength of evidence is not so strong)

Example - 2

A survey of 800 families in a village with 4 children each revealed the following distribution: —

| | | | | | |
|--------|----|-----|-----|-----|----|
| Girls | 4 | 3 | 2 | 1 | 0 |
| Boys | 0 | 1 | 2 | 3 | 4 |
| Family | 32 | 178 | 290 | 236 | 64 |

Is this data consistent with the result that male and female births are equally probable.

① Null-hypothesis: (H_0): Equally probable

Alternative hypothesis (H_a): Not equally probable.

② We have to find the expected value of families —

So, our data is following the binomial distribution. And according to theoretical assumption $p(\text{male}) = p(\text{female}) = 1/2$.

So, for girl = 4, boys = 0.

$$\text{the number of families would be} = \left\{ {}^4C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4 \right\} \times 800 \\ = 50$$

For, girl = 3 boys = 1

$$= \left\{ {}^4C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 \right\} \times 800 \\ = 200$$

For girls = 2 boys = 2

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

For girls = 1 boys = 3

$$= \left\{ {}^4C_3 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 \right\} \times 800 = 200$$

For girls = 0 boys = 4

$$= 50 \text{ (similarly)}$$

③ Now we have to find our χ^2 -statistic —

$$\chi^2\text{-statistic} = \frac{(32-50)^2}{50} + \frac{(178-200)^2}{200} + \frac{(190-200)^2}{300} + \frac{(236-200)^2}{200} + \frac{(64-50)^2}{50}$$

$$= 6.48 + 2.42 + 0.33 + 6.48 + 3.92$$

$$= 19.63$$

and we have degrees of freedom = 4.

Now, with respect to χ^2 -statistic and df, we got 9.488 with level of significance = 0.05

Now, as χ^2 -statistics (19.63) > 9.488

so, we reject the Null hypothesis.

② With respect to P-value it is also shown that we should reject our null hypothesis —

as we got p-value as ~~0.00059~~ 0.00059

which is less than level of significance (α) = 0.05.

So, our alternative hypothesis is true.

So, Male and Female birth are not equally probable. (Proved)

③ Test of Independence :

With this type of chi-square distribution, we find out the dependency relationship between two columns.

In other word, it tests if any two ~~category~~ columns are independent or not.

④ Steps for this test :

① Create the null hypothesis as well as alternative hypothesis.

② Create the contingency table with the observed frequencies for each combination of the categories of two variables.

③ Calculate the expected frequencies of each cell in the contingency table assuming that the null hypothesis is true.

④ Compute the χ^2 -statistic and degree of freedom.

⑤ Based on that we draw conclusion about our hypothesis.

⑥ Two conditions to be proved our null hypothesis —

Ⓐ (Statistic Value) < (level of significance's area)

Ⓑ P value > level of significance

We can use any of these.

⑦ Example 9

A researcher wants to investigate, if there is an association between the level of education (categorical variable) and the preference for a particular type of exercise (categorical variable) among a group of 150 individuals. The researcher collects data and creates the following contingency table.

| Education | Exercise type | | | Total |
|-------------|---------------|---------|----------|-------|
| | Yoga | Running | Swimming | |
| High school | 15 | 20 | 10 | 45 |
| B.Sc | 20 | 30 | 15 | 65 |
| Msc or Phd | 5 | 15 | 20 | 40 |
| Total | 40 | 65 | 45 | 150 |

⑧ Null hypothesis (H_0): There is no association between Education and the preference for a particular type of exercise.

Alternative hypothesis (H_A): There is

Now I have to construct a contingency table based on that the Null hypothesis is true.

Now, the first cell should be the number of student, whose education level is high school, choose to do yoga.

Now, as the education level and Exercise type are independent so, the probability of such people should be —

$$\left(\text{probability (of education level = high school)} \right) \times \text{probability (yoga)} \\ = \left(\frac{45}{150} \times \frac{40}{150} \right)$$

Now, the frequency of such student would be —

$$\left(\frac{45}{150} \times \frac{40}{150} \times 150 \right) = 12$$

Similarly we can find out the expected frequencies for all the cell. and the cell looks like —

| Education | Exercise type | | | Total |
|-------------|---------------|---------|----------|-------|
| | Yoga | Running | Swimming | |
| High school | 12 | 20 | 13 | 45 |
| B.sc | 17 | 28 | 20 | 65 |
| Msc. or Phd | 11 | 17 | 12 | 40 |
| Total | 40 | 65 | 45 | 150 |

Now we have to find out χ^2 -statistic —

$$\begin{aligned} \chi^2\text{-statistic} &= \frac{(15-12)^2}{12} + \frac{(20-20)^2}{20} + \frac{(10-13)^2}{13} + \frac{(20-17)^2}{17} + \frac{(30-28)^2}{28} \\ &\quad + \frac{(15-20)^2}{20} + \frac{(5-11)^2}{11} + \frac{(15-17)^2}{17} + \frac{(20-12)^2}{12} \\ &= 0.75 + 0.69 + 0.53 + 0.11 + 1.25 + 3.27 \\ &\quad + 0.23 + 5.33 \\ &= 12.19 \end{aligned}$$

$$\text{Degree of Freedom} = \left(\text{categories}_{\text{Education}} - 1 \right) \left(\text{categories}_{\text{Exercise}} - 1 \right)$$

$$= (3 - 1)(3 - 1) = 4.$$

The value with respect to level of significance (0.05) and degree of freedom (4) is — 9.488.

clearly it is shown that χ^2 -statistic \neq 9.488

so, we reject the null hypothesis.

If we calculate the p-value with respect to χ^2 -statistic and df we got — 0.0159. Which is less than α (level of significance).

so, we reject our null hypothesis.

so, there is association between Education ~~and~~ level and preference of exercise types. (Proved)
