Chi-Square (X^2) Test

By Daanish Qureshi,
Maxim Attiogbe, and
Darshan Krishnaswamy

Why is this test needed?

1. The ***X^2 Goodness of Fit Test*** measures how well matches a population.
2. The ***X^2 Test of Independence*** compares variables in a contingency tables to see if there is an association between them.
3. The ***X^2 Test for Homogeneity*** tests whether or not different rows/columns of data come from the same population

When should this test be used?

- Only with categorical data
  - Gender
  - Color
  - Shape
  - Number for a dice roll

When should this test be used?

Test	Type of Data	Type of Question	Example
Good-ness of Fit	Counts in several categories	Do the data match my model?	M&Ms
Homogeneity	One variable, across multiple populations	Is the variable similar across all populations?	Types of credit cards (variable) across mailings (populations)
Independence	Two variables, one population	Are the two variables independent or not?	Problems and country of origin (variables) within cars (population)

Chi Square Statistic Formula (Stays the same for each test)

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Conditions for any chi-squared test

- The sampling method must be random
- The data must be categorical
- The expected number of samples for each category should be at least 5
- No more than 10% of the total population should be used in the sample to preserve independence between members of the sample

Chi-Square Test for Goodness of Fit

Chi Square Goodness of Fit Test

- Used to test a hypothesis about the distribution of categorical data.
- For example, when rolling a dice many times in a row, a likely hypothesis would be that each number would occur the same number of times
- After the dice has been rolled 360 times and the data has been collected, chi squared goodness of fit can test the hypothesis

Example: The dice rolling problem - Conditions

- First, the conditions must be met
  - Dice rolling is categorical
  - Rolling a dice is a random sampling method
  - Expected ≥5 /category -> ≥ 30 rolls
  - No real “population” for dice rolling (infinite)
- Summary: As long as there are at least 30 rolls, the chi-square test can be applied to this data.

Goodness of Fit Test Example:
The dice rolling problem - Calculation

- Assume the die is rolled 96 times & the results are:
  - 1: 13 times; 2: 17 times
  - 3: 15 times; 4: 17 times
  - 5: 18 times; 6: 16 times
- The expected outcome for each value is 96/6 = 16 times. The chi square formula yields the following:

$$\frac{(13-16)^2}{16} + \frac{(17-16)^2}{16} + \frac{(15-16)^2}{16} + \frac{(17-16)^2}{16} + \frac{(18-16)^2}{16} + \frac{(16-16)^2}{16} = 1$$

Chi Squared Table

- To interpret test results, a table is needed. It uses the chi squared value as well as the number of degrees of freedom to estimate the p-value
  - The number of degrees of freedom in a chi square test is the number of categories minus 1 for a 1-D category set.
  - In the case of the dice problem, this would be 6-1 = 5.

Interpreting the results

Based on the table, the p-value for this experiment was between 0.95 and 0.975, since there were 5 degrees of freedom and the chi square value was between 0.831 and 1.145. Because the p-value is greater than 0.05, we are able to accept the null hypothesis (that each number is equally likely to appear when rolling a dice).

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188

Chi-Square Test for Homogeneity

Homogeneity Test

- Homogeneity means “same distribution”
- Used to test whether the data in separate columns or rows come from the same population
- Done by looking at the differences in data and whether they are consistent

Example of this Test

- Imagine this set of data is given to you
- Degree of freedom is 2 because (r-1)(c-1) = (3-1)(2-1) = 2
- Set a hypothesis to compare to
    - Without a hypothesis and a null hypothesis the data cannot be compared to anything
- Null Hypothesis: Always observed = expected -> no difference
- Alternative Hypothesis: difference between observed and expected
- Performed in the same way as the good fit test

	Viewing Preferences			Total
	Lone Ranger	Sesame Street	The Simpsons	
Boys	50	30	20	100
Girls	50	80	70	200
Total	100	110	90	300

https://stattrek.com/chi-square-test/homogeneity.aspx

Example of this Test - continued

Null Hypothesis: The amount of girls and boys liking each movie will be proportional to the population, roughly 50% each

Alternative: At least one of the movies does not follow the pattern

Expected Frequency of People: $(n_r * n_c) / n$

n_r = population

n_c = total per

n = total

Examples:

LR B: (100 * 100)/300 = 33.33

SS B: (110 * 100)/300 = 36.67

TS B: (90 * 100)/300 = 30

LR G: (100 * 200)/300 = 66.67

SS G: (110 * 200)/300 = 73.33

TS G: (90 * 200)/300 = 60

Example of this Test - continued more

$X^2 = \Sigma[(O - E)^2 / E]$
O - observed
E - expected

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$X^2 = [(50 - 33.33)^2 / 33.33] + [(30 - 36.67)^2 / 36.67]+[(20 - 30)^2 / 30] + \ldots [(70 - 60)^2 / 60]$

$X^2 = 8.38 + 1.22 + 3.33 + 4.18 + 0.61 + 1.67$

$X^2 = 19.39$

Example of the Test - Conclusion

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188

Since the degrees of freedom is 2, and the Chi Square Homogeneity test results in a value of 19.39, the corresponding P value is less than 0.005, which means the null hypothesis is rejected, and the alternate hypothesis is accepted. Thus, at least one of the movies does not follow the pattern

Chi-Square Test for Independence

Test for Independence

Performed using the same formula as the other chi-square tests

Null hypothesis: variables are independent so probabilities can be multiplied (expected)

Alternative hypothesis: variables are not independent (chi-squared statistic is too extreme)

Example: Shoe Size and Basketball Ability

Observed Data

	Above Average Ability to Play Basketball	Average Ability to Play Basketball	Below Average Ability to Play Basketball	Total
Shoe Size greater than 8	21	12	13	46
Shoe Size less than 8	18	17	19	54
Total	39	29	32	100

Theoretical data (assuming null hypothesis: no relationship)

	Above Average Ability to Play Basketball	Average Ability to Play Basketball	Below Average Ability to Play Basketball	Total
Shoe Size greater than 8	17.94	13.34	14.72	46
Shoe Size less than 8	21.06	15.66	17.28	54
Total	39	29	32	100

Chi-Square Independence Test Example

- The number of degrees of freedom is (3-1)*(2-1) = 2.
- X^2 test is shown below

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{(21-17.94)^2}{17.94} + \frac{(12-13.34)^2}{13.34} + \frac{(13-14.72)^2}{14.72} + \frac{(18-21.06)^2}{21.06} + \frac{(17-15.66)^2}{15.66} + \frac{(19-17.28)^2}{17.28} = 1.588$$

Interpreting Chi-Square Independence

Because there are 2 degrees of freedom with a chi-square value of 1.588, the p-value is between 0.10 and 0.90. Because this is greater than 0.05, the null hypothesis can be accepted and it can be assumed that shoe size and ability to play basketball are independent.

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188

Reporting Chi-Square Results

Chi-Square results should be reported in the form:

X^2 (degrees of freedom, N = size of sample) = X^2 value, p = p-value

If the exact p-value is not known, put the range of values.

For example, in the dice rolling problem, this would look like this:

$X^2\ (5, N = 96) = 1,\ 0.95 < p < 0.975$

Example of chi-square test used in an journal article

Chi-square Calculator

Success! The contingency table below provides the following information: the observed cell totals, (the expected cell totals) and [the chi-square statistic for each cell].

The Chi-square statistic, P value and statement of significance appear beneath the table. Blue means you're dealing with dependent variables; red, independent.

	Category 1	Category 2	*Marginal Row Totals*
Group 1	36 (31.43) [0.66]	14 (18.57) [1.13]	50
Group 2	30 (34.57) [0.6]	25 (20.43) [1.02]	55
Marginal Column Totals	66	39	105 (Grand Total)

The Chi-square statistic is 3.4177. The P value is 0.064502. This result is *not* significant at p < 0.05.

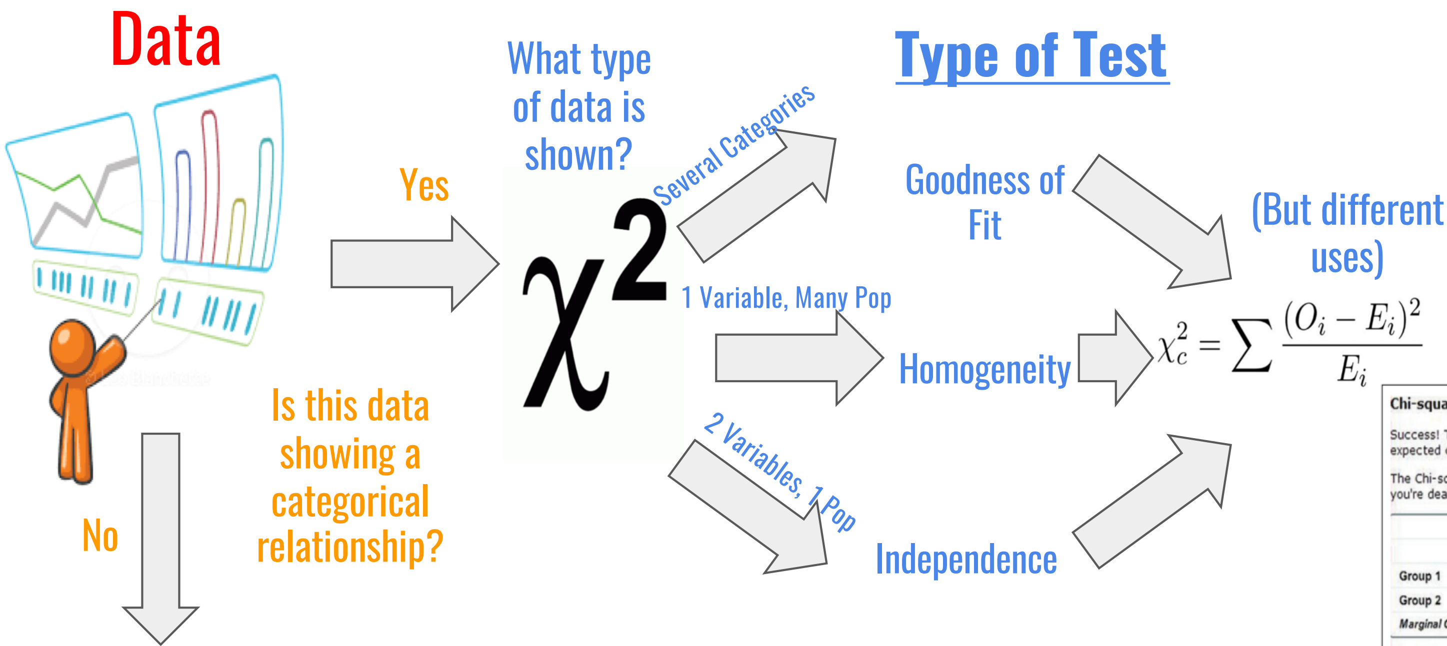
Back Home | Back to Calculators

Figure 3: Chi-square value and *P*-value calculated by online calculator.

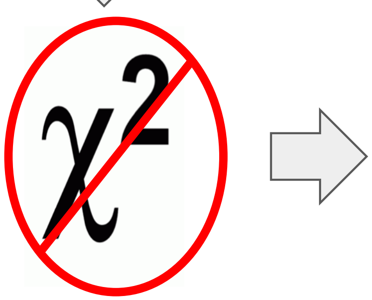
The image above shows the Chi-square value as 3.4177 and its associated *P* value as 0.0645 which is actually greater than *P* value of 0.05, hence no significant difference has been observed. To conclude, there is no association between smoking and lung disease.

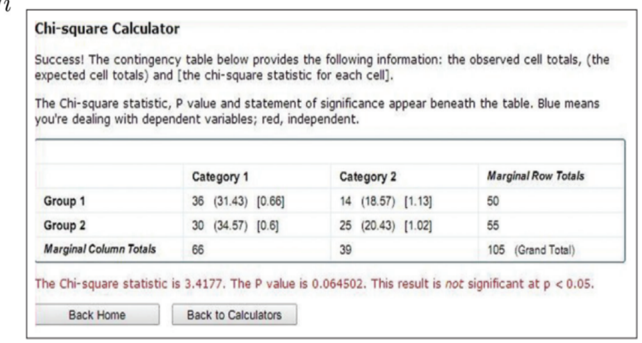
From the article “Chi-square test and its application in hypothesis testing”

The Chi Square test is used for analyzing categorical data and determining if there is a pattern



$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$





Chi-square Calculator

Success! The contingency table below provides the following information: the observed cell totals, (the expected cell totals) and [the chi-square statistic for each cell].

The Chi-square statistic, P value and statement of significance appear beneath the table. Blue means you're dealing with dependent variables; red, independent.

	Category 1	Category 2	*Marginal Row Totals*
Group 1	36 (31.43) [0.66]	14 (18.57) [1.13]	50
Group 2	30 (34.57) [0.6]	25 (20.43) [1.02]	55
Marginal Column Totals	66	39	105 (Grand Total)

The Chi-square statistic is 3.4177. The P value is 0.064502. This result is *not* significant at p < 0.05.

Back Home | Back to Calculators

Figure 3: Chi-square value and *P*-value calculated by online calculator.

Example of how to use it in a journal article

The image above shows the Chi-square value as 3.4177 and its associated P value as 0.0645 which is actually greater than P value of 0.05, hence no significant difference has been observed. To conclude, there is no association between smoking and lung disease.

References

Chi-Square Statistic: How to Calculate It / Distribution. (n.d.). Retrieved from https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/chi-square/

Chi-square tests for categorical data | AP® Statistics. (n.d.). Retrieved from https://www.khanacademy.org/math/ap-statistics/chi-square-tests

Dinov, I. (n.d.). Retrieved from http://socr.ucla.edu/Applets.dir/ChiSquareTable.html

How to Report a Chi-Square Test Result (APA). (n.d.). Retrieved from https://www.socscistatistics.com/tutorials/chisquare/default.aspx

Schoonjans, F. (2019, April 11). Values of the Chi-squared distribution table. Retrieved from https://www.medcalc.org/manual/chi-square-table.php

Stat Trek. (n.d.). Retrieved from https://stattrek.com/chi-square-test/homogeneity.aspx

Statistical Methods Lecture ppt video online download. (n.d.). Retrieved from https://slideplayer.com/slide/1596799/