

What are the four assumptions of a Chi-Square test?

Authored by
stats writer

May 5, 2024

RECOMMENDED CITATION

stats writer (2024). *What are the four assumptions of a Chi-Square test?*.

PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=143063>

The Chi-Square test is a statistical tool used to determine the relationship between two categorical variables. In order to conduct this test, there are four main assumptions that must be met. Firstly, the data being used must be in the form of frequencies or counts, rather than percentages or proportions. Secondly, the observations must be independent of each other, meaning that one observation should not influence another. Thirdly, the expected frequency for each category should be at least 5, to ensure the accuracy of the test. Lastly, the categories being compared should be mutually exclusive, meaning that an observation cannot fall into more than one category. These assumptions must be met in order for the Chi-Square test to provide accurate and meaningful results.

The Four Assumptions of a Chi-Square Test

A is used to determine whether or not there is a significant association between two categorical variables.

This test makes four assumptions:

Assumption 1: Both variables are categorical.

It's assumed that both variables are categorical. That is, both variables take on values that are names or labels.

Examples of categorical variables include:

Marital status ("married", "single", "divorced") Political preference ("republican", "democrat", "independent") Smoking status ("smoker", "non-smoker")

Assumption 2: All observations are independent.

It's assumed that every observation in the dataset is independent. That is, the value of one observation in the dataset does not affect the value of any other observation.

Assumption 3: Cells in the contingency table are mutually exclusive.

It's assumed that individuals can only belong to one cell in the contingency table. That is, cells in the table are mutually exclusive - an individual cannot belong to more than one cell.

Assumption 4: Expected value of cells should be 5 or greater in at least 80% of cells.

It's assumed that the expected value of cells in the contingency table should be 5 or greater in at least 80% of cells and that no cell should have an expected value less than 1.

The following example shows how to check each of these four assumptions in practice.

Example: Checking the Assumptions of a Chi-Square Test

Suppose we want to know whether or not gender is associated with political party preference.

We take a simple random sample of 500 voters and survey them on their political party preference. The following table shows the results of the survey:

	Republican	Democrat	Independent	Total
Male	120	90	40	250
Female	110	95	45	250
Total	230	185	85	500

Assumption 1: Both variables are categorical.

This assumption is easy to verify. We can see that the two variables in the contingency table are both categorical:

Gender: This variable can only take on two categories - Male or Female.
Political Party Preference: This variable can take on three categories - Republican, Democrat, or Independent.

Assumption 2: All observations are independent.

The only way to check this assumption is to verify that

each individual included in this dataset was surveyed independently of every other individual.

If we used a (like simple random sampling) then this assumption is likely met.

Assumption 3: Cells in the contingency table are mutually exclusive.

We can verify that this assumption is met by checking that no individual has been counted in more than one cell.

Assuming each individual in the dataset was only surveyed once, this assumption should be met because it's not possible for an individual to be, say, a Male Republican *and* a Female Democrat simultaneously.

Assumption 4: Expected value of cells should be 5 or greater in at least 80% of cells.

We can use the following formula to calculate the expected values for each cell in the contingency table:

Expected value = (row sum * column sum) / table sum.

For example, the expected value for Male Republicans is: $(230 \times 250) / 500 = 115$.

We can repeat this formula to obtain the expected value for each cell in the table:

	Republican	Democrat	Independent	Total
Male	115	92.5	42.5	250
Female	115	92.5	42.5	250
Total	230	185	85	500

We can see that no cell in the table has an expected value less than 5, so this assumption is met.

Once we've verified that the four assumptions are met, we can then use to perform a Chi-Square Test of Independence:

	Group 1	Group 2	Group 3	Group 4	Group 5
Category 1	120	90	40		
Category 2	110	95	45		
Category 3					
Category 4					
Category 5					

CALCULATE

X² Test Statistic: **0.864035**

p-value: **0.649198**

The p-value of the test is 0.649198. Since this p-value is not less than .05, we do not have sufficient evidence to say that there is an association between gender and political party preference.

The following tutorials explain how to perform a Chi-Square Test of Independence in different statistical software: