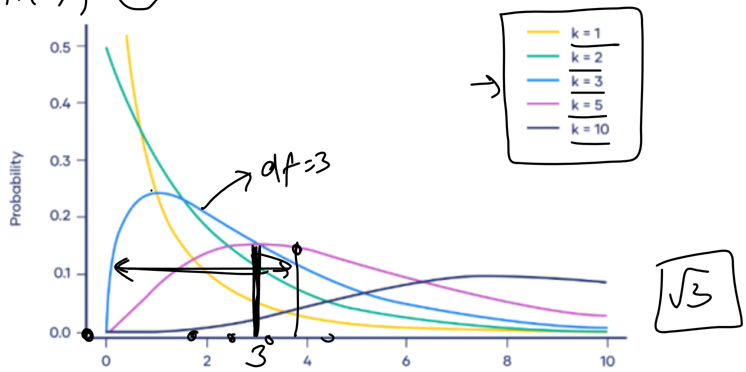
Chi Square Distribution → distribution → continous pd

07 April 2023 10:29

$X \sim N(0,1)$ N

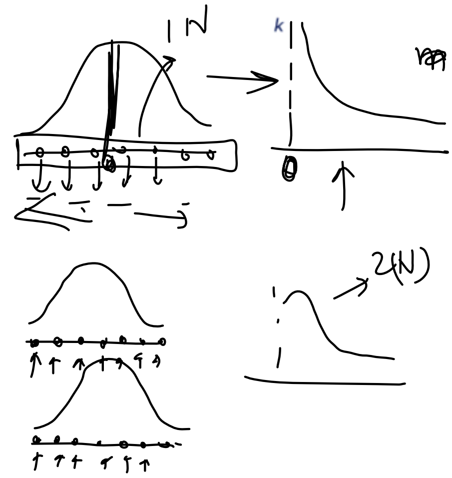


The Chi-Square distribution, also written as χ^2 distribution, is a continuous probability distribution that is widely used in statistical hypothesis testing, particularly in the context of goodness-of-fit tests and tests for independence in contingency tables. It arises when the sum of the squares of independent standard normal random variables follows this distribution.

The Chi-Square distribution has a single parameter, the degrees of freedom (df), which influences the shape and spread of the distribution. The degrees of freedom are typically associated with the number of independent variables or constraints in a statistical problem.

Some key properties of the Chi-Square distribution are:

a. It is a continuous distribution, defined for non-negative values.
b. It is positively skewed, with the degree of skewness decreasing as the degrees of freedom increase.
c. The mean of the Chi-Square distribution is equal to its degrees of freedom, and its variance is equal to twice the degrees of freedom.
d. As the degrees of freedom increase, the Chi-Square distribution approaches the normal distribution in shape.



$\mu = \boxed{K}$

$\sigma^2 = 2K$

χ^2

The Chi-Square distribution is used in various statistical tests, such as the Chi-Square goodness-of-fit test, which evaluates whether an observed frequency distribution fits an expected theoretical distribution, and the Chi-Square test for independence, which checks the association between categorical variables in a contingency table.

chi-square

$\chi^2 = \boxed{Z^2}$ → degree of freedom

→ df=1

$\chi^2 = Z_1^2 + Z_2^2$ → df=2

$\chi^2 = Z_1^2 + Z_2^2 + Z_3^2$ → df=3

$\chi^2 = \sum_{i=1}^{k} Z_k^2$ $\boxed{df = K}$

df ↑

Chi Square Test

07 April 2023 15:03

→ categorical ←

Z-test t-test

continous

χ^2-square

The Chi-Square test is a statistical hypothesis test used to determine if there is a significant association between categorical variables or if an observed distribution of categorical data differs from an expected theoretical distribution. It is based on the Chi-Square (χ^2) distribution, and it is commonly applied in two main scenarios:

1. Chi-Square Goodness-of-Fit Test: This test is used to determine if the observed distribution of a single categorical variable matches an expected theoretical distribution. It is often applied to check if the data follows a specific probability distribution, such as the uniform or binomial distribution.
2. Chi-Square Test for Independence (Chi-Square Test for Association): This test is used to determine whether there is a significant association between two categorical variables in a sample.

goodness of fit test

1 categorical col

test for independence

2 categorical diff

60

$\frac{1}{6} \times 60$

	1	2	3	4	5	6
theo	10	10	10	10	10	10
obs	5	15	5	15	5	15

uniform dist

titanic

Pclass	Survived
1	0
2	1
3	

Goodness of Fit Test

07 April 2023 10:29

The Chi-Square Goodness-of-Fit test is a statistical hypothesis test used to determine if the observed distribution of a single categorical variable matches an expected theoretical distribution. It helps to evaluate whether the data follows a specific probability distribution, such as uniform, binomial, or Poisson distribution, among others. This test is particularly useful when you want to assess if the sample data is consistent with an assumed distribution or if there are significant deviations from the expected pattern.

Steps

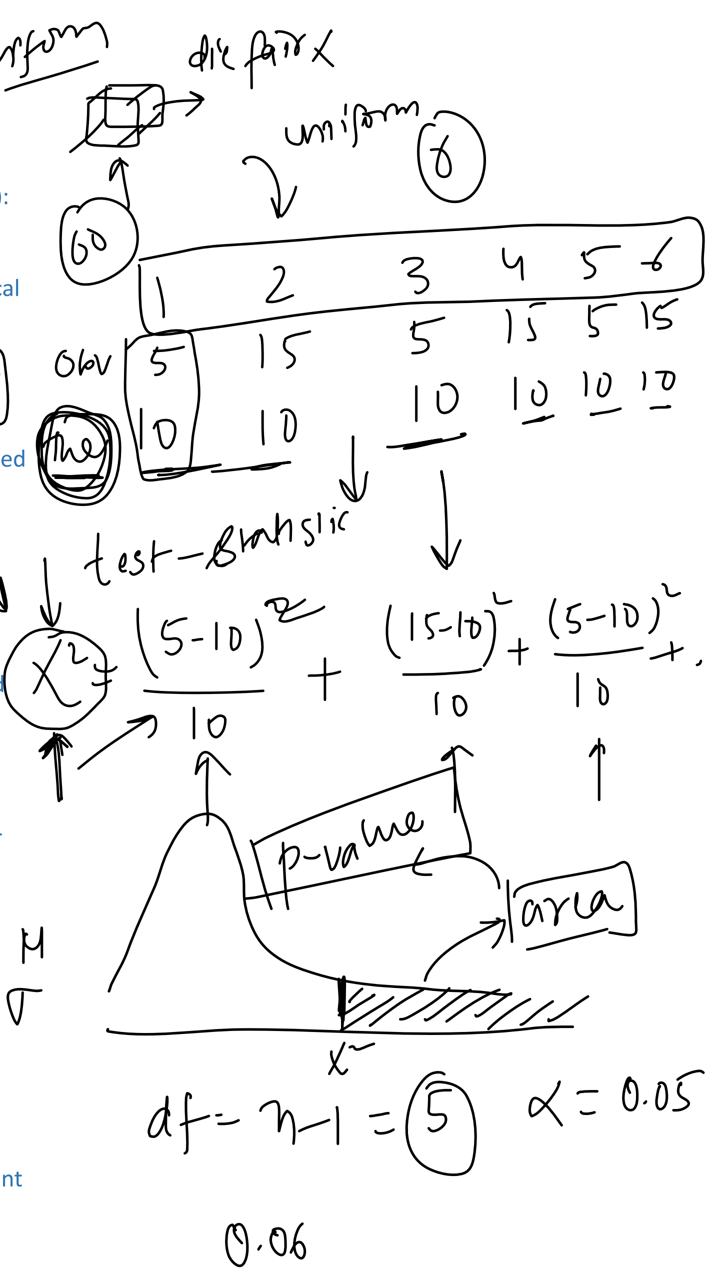
The Chi-Square Goodness-of-Fit test involves the following steps:

- Define the null hypothesis (H0) and the alternative hypothesis (H1):
  - H0: The observed data follows the expected theoretical distribution.
  - H1: The observed data does not follow the expected theoretical distribution.
- Calculate the expected frequencies for each category based on the theoretical distribution and the sample size.
- Compute the Chi-Square test statistic (χ^2) by comparing the observed and expected frequencies. The test statistic is calculated as:

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

uniform dist

- where Oi is the observed frequency in category i, Ei is the expected frequency in category i, and the summation is taken over all categories.
- Determine the degrees of freedom (df), which is typically the number of categories minus one (df = k - 1), where k is the number of categories.
- Calculate the p-value for the test statistic using the Chi-Square distribution with the calculated degrees of freedom.
- Compare the test statistic to the critical value or the p-value



t - test $t = \boxed{S}$

Assumptions

1. Independence: The observations in the sample must be independent of each other. This means that the outcome of one observation should not influence the outcome of another observation.
2. Categorical data: The variable being analysed must be categorical, not continuous or ordinal. The data should be divided into mutually exclusive and exhaustive categories.
3. Expected frequency: Each category should have an expected frequency of at least 5. This guideline helps ensure that the Chi-Square distribution is a reasonable approximation for the distribution of the test statistic. Having small expected frequencies can lead to an inaccurate estimation of the Chi-Square distribution, potentially increasing the likelihood of a Type I error (incorrectly rejecting the null hypothesis) or a Type II error (incorrectly failing to reject the null hypothesis).

potentially increasing the likelihood of a Type I error (incorrectly rejecting the null hypothesis) or a Type II error (incorrectly failing to reject the null hypothesis).

4. Fixed distribution: The theoretical distribution being compared to the observed data should be specified before the test is conducted. It is essential to avoid choosing a distribution based on the observed data, as doing so can lead to biased results.

parametric or non-parametric

The Chi-Square Goodness-of-Fit test is a non-parametric test. Non-parametric tests do not assume that the data comes from a specific probability distribution or make any assumptions about population parameters like the mean or standard deviation.

In the Chi-Square Goodness-of-Fit test, we compare the observed frequencies of the categorical data to the expected frequencies based on a hypothesized distribution. The test doesn't rely on any assumptions about the underlying distribution's parameters. Instead, it focuses on comparing observed counts to expected counts, making it a non-parametric test.

Example 1

07 April 2023 16:26

uniform

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).

H_0: die is fair → uniform

H_1: die is not fair

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

Observed frequencies:

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

expected

	1	2	3	4	5	6	
Obs	12	8	11	9	10	10	→ 60
	10	10	10	10	10	10	

$60 \frac{1}{6}$

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

$\boxed{\chi^2 = 1}$ → chisquare $\quad df = n-1 = 6-1 = \boxed{5}$

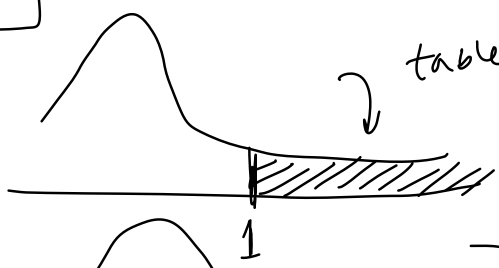
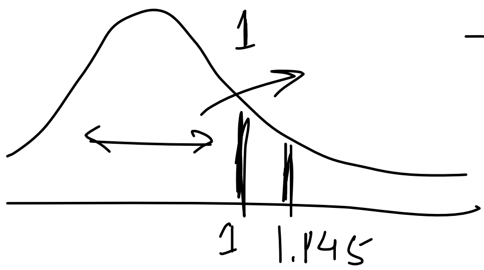


table → python pvalue 0.96

$\alpha = 0.05$



→ pvalue > 0.05

$\boxed{\text{not reject}}$ — fair die

Example 2

07 April 2023 15:26

uniform

Suppose a marketing team at a retail company wants to understand the distribution of visits to their website by day of the week. They have a hypothesis that visits are uniformly distributed across all days of the week, meaning they expect an equal number of visits on each day. They collected data on website visits for four weeks and want to test if the observed distribution matches the expected uniform distribution.

Observed frequencies (number of website visits per day of the week for four weeks):

- Monday: 420
- Tuesday: 380
- Wednesday: 410
- Thursday: 400
- Friday: 410
- Saturday: 430
- Sunday: 390

	mon	tue	wed	thu	fri	sat	sun
Obs	420	380	410	400	410	430	390
Exp	405	405	405	405	405	405	405

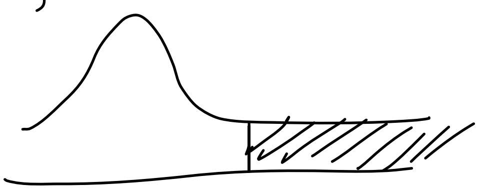
H_0: uniform dis

H_a: Not uniform

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(420-405)^2 + (}{405}$$

$$\chi^2 = \boxed{11}$$

$$df = n-1 = 7-1 = 6$$



11

$\boxed{0.08}$ — p-value

$0.09 < \alpha \rightarrow$

Example 3

07 April 2023 15:27

survey → village → 800 families → 4 children

A survey of $\boxed{800}$ families in a village with 4 children each revealed the following distribution:

binomial

# girls	4	3	2	1	0
# boys	$\boxed{0}$	$\boxed{1}$	2	3	$\boxed{4}$
	1	2	3	4	5
# families	$\boxed{32}$	$\boxed{178}$	$\boxed{290}$	$\boxed{236}$	$\boxed{64}$

theoritical

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

$\boxed{\text{Son}}$

$\boxed{P(S) = P(d) = \frac{1}{2}}$ → 51%

$H_0: \boxed{P(m) = P(f) = \frac{1}{2}}$ → binomial $\boxed{p = \frac{1}{2}}$

$H_a: P(m) \neq P(f)$

$p, q = 1-p$ → $n = 4, \; p = \frac{1}{2}$

$nC_x \, p^x (1-p)^{n-x}$

$P(0) = 4C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4 = \frac{1}{16} \times 800 = 50$

$P(1) = 4C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 = \frac{4!}{\cancel{3!}} \times \frac{1}{16} = \frac{1}{4} \times 800 = 200$

$P(2) = 4C_2 \left(\frac{1}{2}\right)^4 = \frac{4!}{2!\,2!} \times \frac{1}{16} = \frac{4 \times 3}{\cancel{2}} \times \frac{1}{16}$

$\frac{6}{\cancel{16}} \times \cancel{800}$

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

4×1

	0	1	2	3	4
Obs	32	178	290	236	64
Theo	50	200	300	200	50

$$\chi^2 = \frac{(32-50)^2}{50} + \frac{(178-200)^2}{200} + \frac{(290-300)^2}{300} + \frac{(236-200)^2}{200} + \frac{(64-50)^2}{50}$$

$$= \frac{324}{50} + \frac{484}{200} + \frac{100}{300} + \frac{1296}{200} + \frac{196}{50}$$

$$= 6.2 + 2.3 + 0.33 + 6.2 + 3.9$$

$$\chi^2 = 18.93 \qquad df = 5 - 1 = \boxed{4}$$

$0.00081 < \alpha \, (0.05)$

reject the Null hypothesis

Python Case Study

07 April 2023 15:27

1	2	3	
216	184	491	obs
297	297	270	exp

$\chi^2 =$

Test for Independence

07 April 2023 10:29

The Chi-Square test for independence, also known as the Chi-Square test for association, is a statistical test used to determine whether there is a significant association between two categorical variables in a sample. It helps to identify if the occurrence of one variable is dependent on the occurrence of the other variable, or if they are independent of each other.

The test is based on comparing the observed frequencies in a contingency table (a table that displays the frequency distribution of the variables) with the frequencies that would be expected under the assumption of independence between the two variables.

Steps

1. State the null hypothesis (H0) and alternative hypothesis (H1):
   - H0: There is no association between the two categorical variables (they are independent).
   - H1: There is an association between the two categorical variables (they are dependent).
2. Create a contingency table with the observed frequencies for each combination of the categories of the two variables.
3. Calculate the expected frequencies for each cell in the contingency table assuming that the null hypothesis is true (i.e., the variables are independent).
4. Compute the Chi-Square test statistic:

   $\chi^2 = \Sigma [(O_{ij} - E_{ij})^2 / E_{ij}]$

   where O_{ij} is the observed frequency in each cell and E_{ij} is the expected frequency.
5. Determine the degrees of freedom: df = (number of rows - 1) * (number of columns - 1)
6. Obtain the critical value or p-value using the Chi-Square distribution table or a statistical software/calculator with the given degrees of freedom and significance level (commonly α = 0.05).
7. Compare the test statistic to the critical value or the p-value to the significance level to decide whether to reject or fail to reject the null hypothesis. If the test statistic is greater than the critical value, or if the p-value is less than the significance level, we reject the null hypothesis and conclude that there is a significant association between the two variables.

Assumptions

1. Independence of observations: The observations in the sample should be independent of each other. This means that the occurrence of one observation should not affect the occurrence of another observation. In practice, this usually implies that the data should be collected using a simple random sampling method.
2. Categorical variables: Both variables being tested must be categorical, either ordinal or nominal. The Chi-Square test for independence is not appropriate for continuous variables.
3. Adequate sample size: The sample size should be large enough to ensure that the expected frequency for each cell in the contingency table is sufficient. A common rule of thumb is that the expected frequency for each cell should be at least 5. If some cells have expected frequencies less than 5, the test may not be valid, and other methods like Fisher's exact test may be more appropriate.
4. Fixed marginal totals: The marginal totals (the row and column sums of the contingency table) should be fixed before the data is collected. This is because the Chi-Square test for independence assesses the association between the two variables under the assumption that the marginal totals are fixed and not influenced by the relationship between the two variables.

Pclass ↔ Survived

df

$(2-1)(3-1)$

$1 \times 2 = 2$

Observed

	pclass 1	2	3	
0	23	49	111	183
1	100	200	300	600
	123	249	411	

Expected

	pclass 1	2	3
0	25	50	100
1	100	250	350

>5

$\frac{(23-25)^2}{25} +$

χ^2

$\boxed{\chi^2} \rightarrow$ p-value < α

Example 1

07 April 2023 17:50

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

Observed ?

	Exercise Type			
Education	Yoga	Running	Swimming	Total
High School	15	20	10	45
Bachelor's	20	30	15	65
Master's or PhD	5	15	20	40
Total	40	65	45	150

edu ⟷ exercise (independent)

H_0: they are independent X

H_1: they are associated

need a contingency

$\frac{45 \times 40}{150}$ $\frac{65 \times 45}{150}$ × $\frac{45 \times 45}{150}$

expected

	Yoga	run	swim
High	12	19	13.5
Bach	17	28	20
phd	10	17	12

$$\frac{(15-12)^2}{12} + \frac{(20-19)^2}{20} + \frac{(10-13.5)^2}{10} + \_\_ \; \_\_ \; \_\_$$

$\chi^2 = 9.95$

p-value 0.04 < α

reject Null

df = (3-1)(3-1) = 4

Python Case Study

07 April 2023 15:06

Applications in Machine Learning

07 April 2023 17:30

1 Feature selection: Chi-Square test can be used as a filter-based feature selection method to rank and select the most relevant categorical features in a dataset. By measuring the association between each categorical feature and the target variable, you can eliminate irrelevant or redundant features, which can help improve the performance and efficiency of machine learning models.

2. Evaluation of classification models: For multi-class classification problems, the Chi-Square test can be used to compare the observed and expected class frequencies in the confusion matrix. This can help assess the goodness of fit of the classification model, indicating how well the model's predictions align with the actual class distributions.

3. Analysing relationships between categorical features: In exploratory data analysis, the Chi-Square test for independence can be applied to identify relationships between pairs of categorical features. Understanding these relationships can help inform feature engineering and provide insights into the underlying structure of the data.

4. Discretization of continuous variables: When converting continuous variables into categorical variables (binning), the Chi-Square test can be used to determine the optimal number of bins or intervals that best represent the relationship between the continuous variable and the target variable.

5. Variable selection in decision trees: Some decision tree algorithms, such as the CHAID (Chi-squared Automatic Interaction Detection) algorithm, use the Chi-Square test to determine the most significant splitting variables at each node in the tree. This helps construct more effective and interpretable decision trees.