|  |  |
| --- | --- |
| A picture of a winding road and trees  Chi-Square Test: A Versatile Tool in Data Analysis  Made by-Arabinda Chand | Abstract  The study of genetics involves understanding how traits are passed from parents to offspring. Analyzing observed inheritance patterns often requires statistical tools to determine whether deviations from expected ratios are due to chance or reflect actual genetic mechanisms. The Chi-Square test is a valuable method for comparing observed and expected frequencies in genetic data, helping researchers to validate genetic hypotheses and uncover the complexities of inheritance. This test is particularly useful when dealing with data in the form of frequencies, as is common in genetic studies.  S.E. (Computer Engineering) Sem-4 Roll No.-11  BR HARNE College of Engineering & Technology |

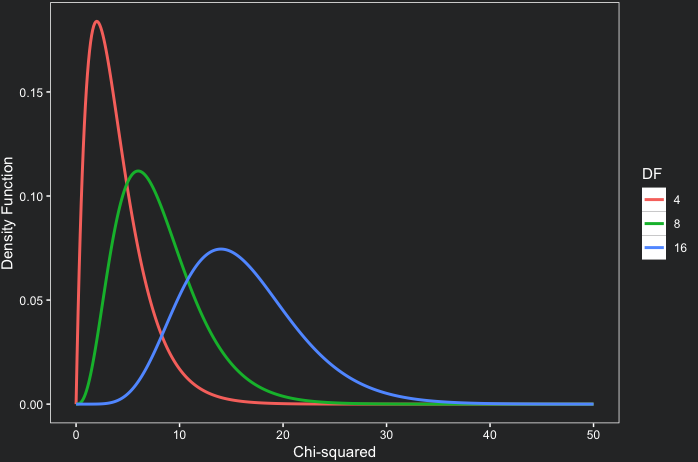
**Contents**

|  |  |
| --- | --- |
| topic | page |
| Introduction |  |
| Concept |  |
| Applications |  |
| Conditions |  |
| Examples |  |
| Conclusion |  |
| Bibliography |  |



**1. Introduction**

The Chi-Square (χ2) test is a versatile and widely used non-parametric statistical test. It is particularly suited for analyzing data in the form of frequencies or data measured at the ordinal or nominal level. The Chi-Square test serves various purposes, including comparing observed and expected results, assessing the independence of attributes, and evaluating the goodness of fit of theoretical distributions.



[This Photo](https://www.codingthepast.com/2023/12/04/Chi-square-distribution-in-R.html) by Unknown Author is licensed under [CC BY-NC](https://creativecommons.org/licenses/by-nc/3.0/)

**2. Foundations of the Chi-Square Test**

**2.1 Chi-Square Distribution**

The Chi-Square distribution is a probability distribution defined by its degrees of freedom. The shape of the Chi-Square distribution is influenced by the degrees of freedom. It is skewed to the right with smaller degrees of freedom and becomes more symmetrical, approximating a normal distribution, as the degrees of freedom increase.

**2.2 Chi-Square Statistic**

The Chi-Square statistic is calculated using the formula:

χ2=∑Ei​(Oi​−Ei​)2​

Where:

* χ2 = Chi-Square statistic
* Oi​ = Observed frequency in category i
* Ei​ = Expected frequency in category i
* ∑ = Summation across all categories

Steps for Chi-square Testing

1) First set a null of hypotheses.

2) Collect the data and find out observed frequency.

3) Find out the expected frequencies by adding all the observed frequencies divided by number of Categories

4) Find out the difference between observed frequencies and expected frequencies.(fo-fe)

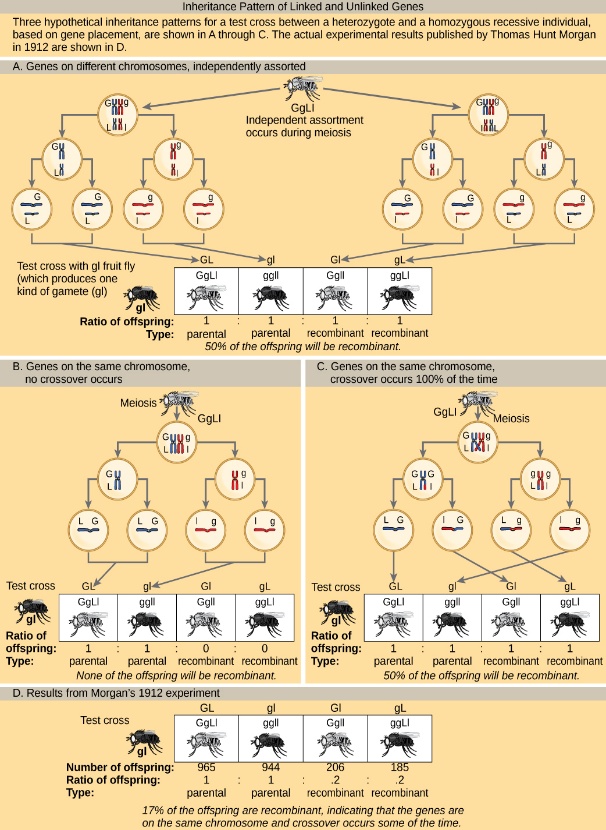
5) Find out the square of the difference between observed and expected Frequency. (fo-fe)2

6) Divide it by expected frequency. (fo-fe)2/fe . You will get a quotient

7) Find out the sum of these quotients.

8) Determine the degree of freedom and find out the critical value of χ2 from table.

9) Compare the calculated and table value of χ2 and use the following decision rule. Accept null hypothesis if χ2 is less than critical value given in table. Reject the null hypothesis if calculated value of χ2 is more than what is given in the table under .05 or .01 significance levels.



[This Photo](https://louis.pressbooks.pub/generalbiology1leclab/chapter/chromosomal-theory-and-genetic-linkage/) by Unknown Author is licensed under [CC BY](https://creativecommons.org/licenses/by/3.0/)

**3. Applications of the Chi-Square Test**

**3.1 Chi-Square Test for Goodness of Fit**

**This test determines how well observed data fits an expected distribution (e.g., normal, binomial, Poisson). It compares the observed sample distribution with the expected probability distribution.**

* **Null Hypothesis (H0​): There is no significant difference between the observed and expected values.**
* **Alternative Hypothesis (H1​): There is a significant difference between the observed and expected values.**
* **The calculated χ2 value is compared to a critical value from the Chi-Square distribution table to determine if the null hypothesis should be rejected.**

**3.2 Chi-Square Test for Independence of Attributes**

**This test examines whether two categorical variables (attributes) are independent or related. It helps determine if the observed differences in sample proportions are significant or due to chance.**

* **Null Hypothesis (H0​): The attributes are independent.**
* **Alternative Hypothesis (H1​): The attributes are dependent (related).**

**Example: A survey of consumer preferences for a product across different geographical regions.**

**3.3 Chi-Square Test for Population Variance**

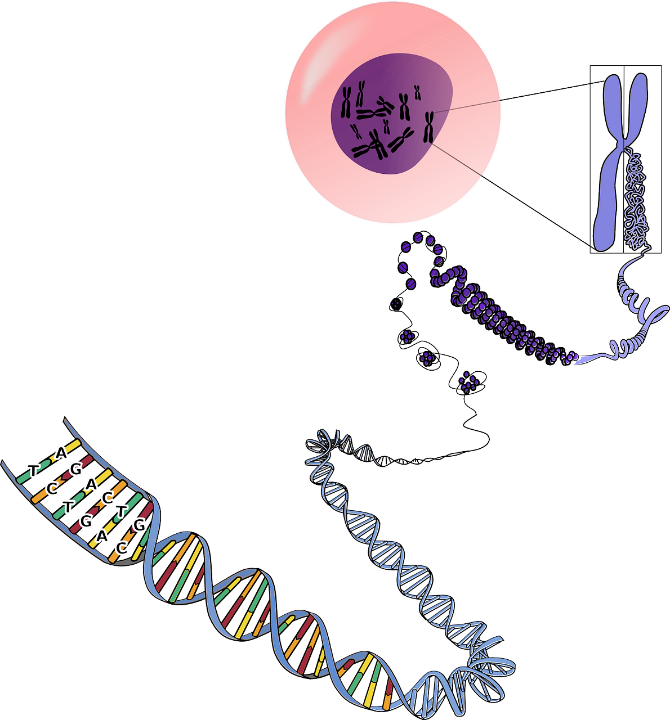
**The Chi-Square test can also be used to test hypotheses about the variance of a single population.**

* **This is relevant in situations where it's important to assess the variability of a characteristic, such as the consistency in the quality of manufactured products.**
* **It assumes that the population follows a normal distribution.**

**3.3 Chi-Square Test for Population Variance**

The Chi-Square test can also be used to test hypotheses about the variance of a single population.

* This is relevant in situations where it's important to assess the variability of a characteristic, such as the consistency in the quality of manufactured products.
* It assumes that the population follows a normal distribution



**4. Conditions and Limitations**

**4.1 Conditions for Applying the Chi-Square Test**

* **Random Sample: Data must be from a random sample to ensure results reflect the population.**
* **Large Sample Size: The sample size must be large enough (generally ~50 or more) to avoid Type-II errors and use raw counts, not percentages.**
* **Adequate Cell Sizes: Expected frequencies in each cell should be at least 5 to prevent overestimation of the Chi-Square statistic.**
* **Independence: Observations within the sample must be independent of each other.**
* **Final Values: Data must be categorized for the Chi-Square test to be applicabl**

**4.2 Yates' Correction**

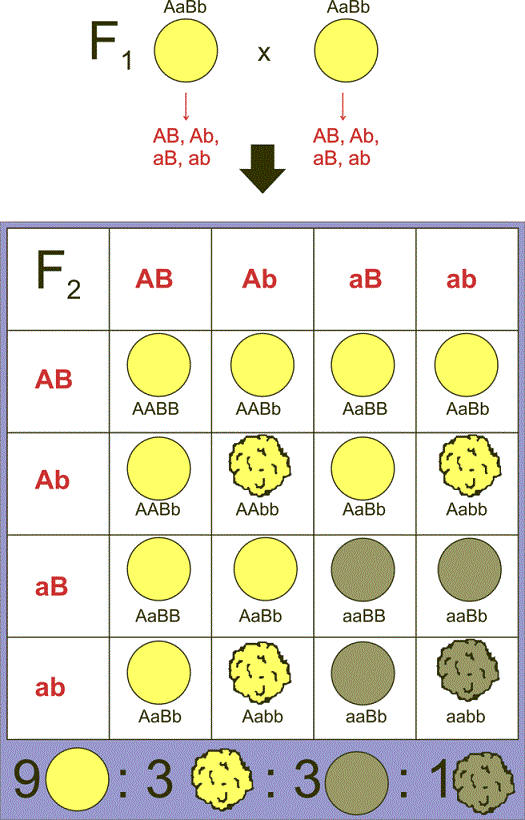
Yates' correction is applied to the Chi-Square test when dealing with 2x2 contingency tables and small expected frequencies (generally, when any expected frequency is less than 5). It involves adjusting the formula to:

χ2=∑Ei​(∣Oi​−Ei​∣−0.5)2​

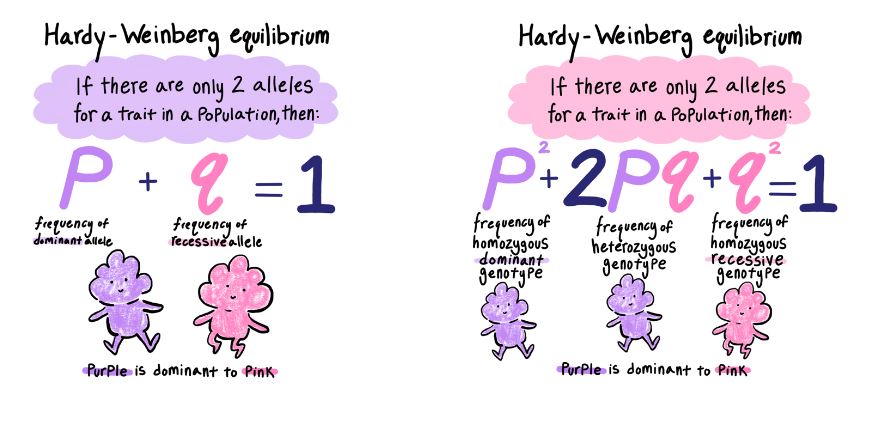
This correction helps to make the Chi-Square test statistic more closely approximate the continuous Chi-Square distribution.

**4.3 Limitations**

* The Chi-Square test indicates associations but does not measure the strength or nature of the association.
* It is sensitive to sample size.



[This Photo](https://biologia-geologia.com/biologia2/9223_tercera_ley_de_mendel_o_principio_de_la_combinacion_independiente.html) by Unknown Author is licensed under [CC BY-SA](https://creativecommons.org/licenses/by-sa/3.0/)



**5. Applications examples**

The Chi-Square test is used in genetics to:

* **Verify Mendelian Ratios:** Check if observed phenotypic ratios in genetic crosses align with expected Mendelian ratios.
* **Analyze Pedigree Charts:** Chi-square analysis can be used to verify the results of pedigree analysis.
* **Testing for Independence of Traits:** The Chi-Square test can also determine if two traits are inherited independently, as Mendel's law of independent assortment suggests.
* **Analyzing Genetic Linkage:** It can be used to investigate if genes are linked on the same chromosome, causing deviations from expected inheritance patterns.
* **Population Genetics:** The Chi-Square test helps analyze allele and genotype frequencies in populations, examining concepts like Hardy-Weinberg equilibrium.

A)Application: Verifying Mendelian Ratios

Mendel's laws form the foundation of genetics. The Chi-Square test can be used to verify if experimental results align with the ratios predicted by these laws.

**Example: Monohybrid Cross**

In a monohybrid cross, where we examine the inheritance of a single trait, a classic example is Mendel's study of pea plants. Let's simulate a similar experiment.

**Scenario:**

Imagine we cross two heterozygous pea plants (Tt) for the stem height trait, where T (tall) is dominant and t (short) is recessive. According to Mendelian genetics, we expect a 3:1 phenotypic ratio of tall to short plants in the offspring (F2 generation).

**Observed Data:**

Suppose we grow 100 plants and observe the following:

* Tall: 78
* Short: 22

**Expected Data:**

Based on the 3:1 ratio, we expect:

* Expected Tall = (3/4) \* 100 = 75
* Expected Short = (1/4) \* 100 = 25

**Chi-Square Calculation:**

χ2=75(78−75)2​+25(22−25)2​=759​+259​=0.12+0.36=0.48

**Degrees of Freedom and Critical Value:**

We have two categories (tall and short), so the degrees of freedom (df) = number of categories - 1 = 2 - 1 = 1.

Using a Chi-Square table and a significance level of 0.05, the critical value for 1 degree of freedom is 3.841.

**Interpretation:**

Our calculated χ2 value (0.48) is less than the critical value (3.841). Therefore, we fail to reject the null hypothesis. This suggests that the observed data does not significantly deviate from the expected 3:1 ratio, supporting Mendel's law of segregation.

b)**Application in Bioinformatics: Testing for Hardy-Weinberg Equilibrium**

One fascinating application of the Chi-Square test in bioinformatics is checking for **Hardy-Weinberg Equilibrium (HWE)**. This principle states that in a large, randomly mating population, the allele and genotype frequencies will remain constant from generation to generation in the absence of disturbing influences.

Let's consider a gene with two alleles: A and a. There are three possible genotypes: AA, Aa, & aa.

Suppose we observe the following genotype frequencies in a population of 200 individuals:

|  |  |
| --- | --- |
| Genotype | Observed Frequency (Oi​) |
| AA | 45 |
| Aa | 90 |
| aa | 65 |
| Total | **200** |

To test for HWE, we first need to calculate the allele frequencies:

* Frequency of allele A (p) = (2 \* number of AA individuals + number of Aa individuals) / (2 \* total number of individuals) p=2×200(2×45)+90​=40090+90​=400180​=0.45
* Frequency of allele a (q) = (2 \* number of aa individuals + number of Aa individuals) / (2 \* total number of individuals) q=2×200(2×65)+90​=400130+90​=400220​=0.55 (Note that p+q=0.45+0.55=1)

Now, under Hardy-Weinberg Equilibrium, the expected genotype frequencies are:

* Expected frequency of AA = p2×total population=(0.45)2×200=0.2025×200=40.5
* Expected frequency of Aa = 2pq×total population=2×0.45×0.55×200=0.495×200=99
* Expected frequency of aa = q2×total population=(0.55)2×200=0.3025×200=60.5

|  |  |  |
| --- | --- | --- |
| Genotype | Observed Frequency (Oi​) | Expected Frequency (Ei​) |
| AA | 45 | 40.5 |
| Aa | 90 | 99 |
| aa | 65 | 60.5 |

**Calculating the Chi-Square Statistic**

Now we can calculate the Chi-Square statistic:

χ2=40.5(45−40.5)2​+99(90−99)2​+60.5(65−60.5)2​

χ2=40.5(4.5)2​+99(−9)2​+60.5(4.5)2​

χ2=40.520.25​+9981​+60.520.25​

χ2≈0.5+0.818+0.335

χ2≈1.653

**Interpreting the Result**

To determine if this χ2 value is statistically significant, we need to compare it to a critical value from the Chi-Square distribution table. The degrees of freedom (df) for this test are the number of categories minus 1. In this case, we have 3 genotypes, so df=3−1=2.

Looking up the critical value for a Chi-Square distribution with 2 degrees of freedom at a common significance level (e.g., α=0.05), we find it to be approximately 5.991.

Since our calculated χ2 value (1.653) is less than the critical value (5.991), we **fail to reject the null hypothesis**. This means that the observed genotype frequencies do not significantly deviate from the expected frequencies under Hardy-Weinberg Equilibrium. In other words, the population appears to be in HWE for this gene, suggesting that factors like mutation, gene flow, non-random mating, genetic drift, or natural selection are not significantly affecting the allele and genotype frequencies for this particular gene in this population.

C)**Application in Bioinformatics: Investigating Gene Linkage**

One compelling application of the Chi-Square test in bioinformatics lies in investigating **gene linkage**. Genes located close together on a chromosome tend to be inherited together more often than genes that are far apart. By analyzing the inheritance patterns of different traits, we can use the Chi-Square test to determine if genes are likely linked.

Consider a study examining the inheritance of two genes in fruit flies: one affecting body color (Gray or Black) and another affecting wing shape (Normal or Vestigial). Suppose we perform a cross and observe the following offspring phenotypes:

|  |  |
| --- | --- |
| Phenotype | Observed Number (Oi​) |
| Gray body, Normal wings | 90 |
| Gray body, Vestigial wings | 15 |
| Black body, Normal wings | 10 |
| Black body, Vestigial wings | 85 |
| Total | **200** |

If the genes for body color and wing shape were inherited independently (not linked), we would expect a 1:1:1:1 phenotypic ratio in the offspring of a dihybrid cross. This would mean an expected frequency of 200/4 = 50 for each phenotype.

|  |  |  |
| --- | --- | --- |
| Phenotype | Observed Number (Oi​) | Expected Number (Ei​) |
| Gray body, Normal wings | 90 | 50 |
| Gray body, Vestigial wings | 15 | 50 |
| Black body, Normal wings | 10 | 50 |
| Black body, Vestigial wings | 85 | 50 |

**Calculating the Chi-Square Statistic**

Now, we calculate the Chi-Square statistic to see if the observed frequencies significantly deviate from these expected frequencies: χ2=50(90−50)2​+50(15−50)2​+50(10−50)2​+50(85−50)2​ χ2=50(40)2​+50(−35)2​+50(−40)2​+50(35)2​

χ2=501600​+501225​+501600​+501225​

χ2=32+24.5+32+24.5

χ2=113

**Interpreting the Result**

To determine the significance of this calculated χ2 value, we compare it to a critical value from the Chi-Square distribution table. The degrees of freedom (df) for this test are calculated as (number of rows - 1) \* (number of columns - 1). In this case, we have 2 body colors and 2 wing shapes, so df=(2−1)×(2−1)=1.

Looking up the critical value for a Chi-Square distribution with 1 degree of freedom at a significance level of α=0.05, we find it to be approximately 3.841.

Our calculated χ2 value (113) is significantly greater than the critical value (3.841). Therefore, we **reject the null hypothesis** of independent assortment. This strong statistical evidence suggests that the genes for body color and wing shape in these fruit flies are likely linked and are not inherited independently

**6. Conclusion**

The Chi-Square test is a powerful statistical tool with broad applications. It is used for analyzing categorical data, comparing observed and expected frequencies, and testing various statistical hypotheses. Understanding its conditions and limitations is essential for proper interpretation and valid conclusions.



References

* Trueman biology by KN bhatia
* Genetics & Evolutionary biology by M. Abbas
* Probability & hypothesis testing(MCO-3)
* Statistical inference by Diwakar Shukla
* Significance of difference of frequency(BPC-004)





Karl Pearson(1857-1936),statistician made chi-square in 1900

**Thank you**