Chi-squared goodness of fit test

We want to test whether there's a significant difference in the preferences of people for three different flavours of ice cream: chocolate, vanilla, and strawberry. We'll collect data from a sample of 200 individuals and record their preferences.

Here's hypothetical data: Chocolate: 80 people; Vanilla: 60 people; Strawberry: 60 people

Now, we want to test **whether these preferences are significantly different** from what we would expect if there were no preference (i.e., if people were equally likely to choose any flavour).

Null hypothesis (H0): there is no difference in preference, meaning each flavour is equally likely to be chosen.

Alternative hypothesis (H1): there is a difference in preference, meaning at least one flavour is more preferred than the others.

We'll use the chi-squared test to analyse this data.

First, we need to calculate the expected frequencies under the assumption of no preference. Since there are three flavours and 200 people, the expected frequency for each flavour is 200/3 = 66.67.

Expected frequencies: Chocolate: 66.67; Vanilla: 66.67; Strawberry: 66.67

Now, we calculate the chi-squared statistic:

$$\chi^2 = \Sigma ((\text{Observed frequency} - \text{Expected frequency})^2 / \text{Expected frequency})$$

For each flavour: Chocolate: $(80 - 66.67)^2 / 66.67 \approx 2.66$; Vanilla: $(60 - 66.67)^2 / 66.67 \approx 0.67$; Strawberry: $(60 - 66.67)^2 / 66.67 \approx 0.67$

Summing these values, we get: $\chi^2 \approx 2.66 + 0.67 + 0.67 \approx 4$

Now, we need to compare this value to the critical value from the chi-squared distribution table with (3-1) = 2 degrees of freedom (since there are 3 categories). Assuming a significance level (α) of 0.05, we find that the critical value is approximately 5.99. CHISQ.INV(0.95,2)

Since our calculated χ^2 value (4) is less than the critical value (5.99), **we fail to reject the null hypothesis.** This means that we do not have sufficient evidence to conclude that there is a significant difference in preferences for the three flavours of ice cream.

Here are some examples where you can use the chi-squared goodness of fit test to test hypotheses:

1. Dice Fairness:

Hypothesis: Are the outcomes of a fair six-sided die statistically consistent with the expected probabilities? Assume a significance level (α) of 0.05.

Data Collection: Roll a fair six-sided die a large number of times and record the frequencies of each outcome (1 through 6).

Number	Frequency
1	80
2	120
3	70
4	130
5	110
6	90

Null Hypothesis (H0): The observed frequencies match the expected probabilities of each outcome (1/6 for each).

Alternative Hypothesis (H1): The observed frequencies do not match the expected probabilities.

2. Marbles in a Bag:

Hypothesis: Do the observed frequencies of different coloured marbles in a bag match the expected frequencies based on a specified distribution? Assume a significance level (α) of 0.01.

Data Collection: Randomly sample a large number of marbles from a bag and record the frequencies of each colour.

Colour	Observed frequency	Expected frequency
Red	30	35%
Green	20	20%
Blue	50	35%
Black	10	10%

Null Hypothesis (H0): The observed frequencies match the expected frequencies based on the specified distribution.

Alternative Hypothesis (H1): The observed frequencies do not match the expected frequencies.

In each of these examples, you would define the expected frequencies based on the null hypothesis, calculate the chi-squared statistic, and compare it to the critical value from the chi-squared distribution to make a conclusion about the goodness of fit of the observed data to the expected distribution.

Chi-squared test for independence

Consider an example where we want to determine if there's an association between smoking habits and gender among a group of individuals. We'll collect data from a sample of 500 people and record whether they are smokers or non-smokers and their gender.

Here's hypothetical data: Among 250 males, 100 are smokers and 150 are non-smokers.

Among 250 females, 50 are smokers and 200 are non-smokers.

We want to test **whether smoking habits are independent of gender.**

Null hypothesis (H0): Smoking habits **are independent** of gender.

Alternative hypothesis (H1): Smoking habits **are dependent** on gender.

We'll use the chi-squared test for independence to analyse this data.

First, let's create a contingency table:

	SMOKERS	NON-SMOKERS	**TOTAL**
MALE	100	150	
FEMALE	50	200	
TOTAL			

Now, we'll calculate the expected frequencies assuming independence:
- Expected frequency for male smokers: (250 * 150) / 500 = 75
- Expected frequency for male non-smokers: (250 * 350) / 500 = 175
- Expected frequency for female smokers: (250 * 150) / 500 = 75
- Expected frequency for female non-smokers: (250 * 350) / 500 = 175

Next, we'll calculate the chi-squared statistic:

$\chi^2 = \Sigma$ ((Observed frequency - Expected frequency)² / Expected frequency)

For each cell: $(100 - 75)^2 / 75 \approx 8.33$; $(150 - 175)^2 / 175 \approx 3.57$; $(50 - 75)^2 / 75 \approx 8.33$; $(200 - 175)^2 / 175 \approx 3.57$

Summing these values, we get: $\chi^2 \approx 8.33 + 3.57 + 8.33 + 3.57 \approx 23.8$

Now, we need to compare this value to the critical value from the chi-squared distribution table with (2-1)(2-1) = 1 degree of freedom (since there are 2 categories for both smoking habit and gender). Assuming a significance level (α) of 0.05, the critical value is approximately 3.84.

CHISQ.INV(0.95, 2) = 3.84

Since our calculated χ^2 value (23.8) is greater than the critical value (3.84), **we reject the null hypothesis**. This indicates that **there is a significant association between smoking habits and gender among the population.**

Here are a few more examples where you can apply the chi-squared test for independence to test hypotheses:

1. Educational Attainment and Employment Status:

Hypothesis: Is there a relationship between educational attainment (e.g., high school diploma, bachelor's degree, master's degree) and employment status (e.g., employed, unemployed, student)? Assume a significance level (α) of 0.05.

Data Collection: Survey a sample of individuals and record their educational attainment and current employment status.

	EMPLOYED	UNEMPLOYED	STUDENT	**TOTAL**
HIGH	100	50	70	
BACHELOR	120	40	50	
MASTER	80	20	30	
TOTAL				

Null Hypothesis (H0): Educational attainment and employment status are independent.

Alternative Hypothesis (H1): Educational attainment and employment status are dependent.

2. Customer Satisfaction and Product Type:

Hypothesis: Is there an association between customer satisfaction (e.g., satisfied, neutral, dissatisfied) and the type of product purchased (e.g., electronics, clothing, food)? Assume a significance level (α) of 0.01.

Data Collection: Gather feedback from customers who purchased different types of products and record their satisfaction levels.

	ELECTRONICS	CLOTHING	FOOD	**TOTAL**
SATISFIED	50	40	30	
NEUTRAL	40	30	10	
DISSATISFIED	30	20	10	
TOTAL				

Null Hypothesis (H0): Customer satisfaction and product type are independent.

Alternative Hypothesis (H1): Customer satisfaction and product type are dependent.

3. Preferred Social Media Platform and Age Group:

Hypothesis: Is there an association between preferred social media platform (e.g., Facebook, Instagram) and age group (e.g., teenagers, young adults, middle-aged adults)? Assume a significance level (α) of 0.05.

Data Collection: Survey a sample of individuals across different age groups and record their preferred social media platforms.

	TEENAGERS	YOUNG ADULTS	OTHERS	**TOTAL**
FACEBOOK	25	50	60	
INSTAGRAM	60	30	35	
TOTAL				

Null Hypothesis (H0): Preferred social media platform and age group are independent.

Alternative Hypothesis (H1): Preferred social media platform and age group are dependent.

In each of these examples, you would collect data, create a contingency table, calculate expected frequencies assuming independence, compute the chi-squared statistic, and compare it to the critical value from the chi-squared distribution to make a conclusion about the relationship between the variables.