Part 2.
Chi-squared test

Topics

1. Chi-squared test purpose
2. Types of Chi-squared tests
3. Concepts and terminologies
4. Visualization of frequency data
5. Interpretation of results
6. Practice with examples
7. Advanced topics; Assumptions and limitations

Chi-squared test (Chi-square test, χ^2 test)

- **Name:**

  - This term arises from the symbol of the **Chi-square statistic** (χ^2), where "chi" is a Greek letter. The test is commonly used in statistics to test hypotheses about **the distribution of categorical data**.

1. Purpose

- The Chi-squared test is used

  - To determine **whether there is a significant association** between two categorical variables in a contingency table

  - Or to determine if **a single categorical variable's distribution differs significantly** from an expected distribution.

Gender	Vegetarian	Non-Vegetarian	Total
Male	20	30	50
Female	30	20	50
Total	50	50	100

2. Types of tests

- Two types of Chi-squared test:

  - **Test for Independence**: Determines if there's a relationship between two categorical variables.

  - **Goodness of Fit Test**: Compares the observed frequencies of categories to expected frequencies to see if they differ significantly.

Chi-squared distribution

- **Chi-squared test vs. Pearson's chi-squared test**

  - **Chi-squared test** is a broad term that can refer to any statistical test that uses the Chi-squared distribution, which includes **tests for goodness of fit**, **tests for independence** in contingency tables, and others.

  - **Pearson's Chi-squared test**: This specifically refers to the Chi-squared test developed by Karl Pearson. It is typically used in the context of a contingency table to **test for independence** between two variables.

  - However, it's important to note that there are other types of Chi-squared tests, such as the **Likelihood Ratio Chi-squared test**, which is used in similar contexts but calculates the test statistic differently.

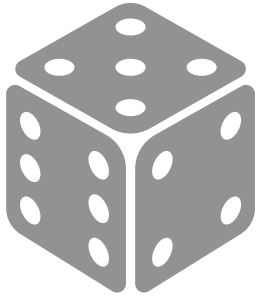
1) Chi-squared Goodness of fit test

- **Purpose**: To determine ***how well a theoretical distribution fits an observed distribution***. In other words, it tests whether the observed frequencies in a categorical dataset significantly deviate from the expected frequencies, which are calculated based on a hypothesized distribution.
- **Applications**: This test is commonly used in cases where you want to see if your data follows a certain distribution, such as a normal distribution, binomial distribution, or any other theoretical distribution. For example, checking whether ***the observed data fit a uniform distribution*** (where all categories are equally likely) (e.g., testing if a die is fair.)

Procedure:

- **Step 1**: Define the null hypothesis, typically stating that the data follows the hypothesized distribution.
- **Step 2**: Calculate the expected frequencies for each category based on the hypothesized distribution.
- **Step 3**: Compute the chi-squared statistic, which is the sum of the squared differences between observed and expected frequencies, divided by the expected frequencies.
- **Step 4**: Compare the chi-squared statistic to a critical value from the Chi-squared distribution (based on the desired significance level and degrees of freedom) to determine whether to reject the null hypothesis.

Example

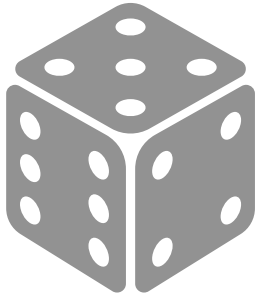


Scenario: We want to test if a fair six-sided die is indeed fair, meaning each face has an equal probability of occurring.

Hypothesis:

- **Null Hypothesis (H0): The observed frequencies of each die face match the expected frequencies of a fair die.**
- **Alternative Hypothesis (H1): The observed frequencies of the die faces differ from what would be expected for a fair die.**

Data

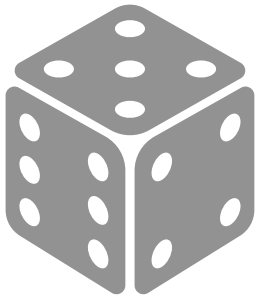


Scenario: We want to test if a fair six-sided die is indeed fair, meaning each face has an equal probability of occurring.

Data:

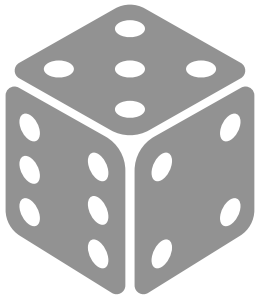
- We roll a fair six-sided die 120 times and record the outcomes.

Die Face	1	2	3	4	5	6
Observed	22	18	19	21	20	20
Expected	20	20	20	20	20	20



Die Face	1	2	3	4	5	6
Observed	22	18	19	21	20	20
Expected	20	20	20	20	20	20

1. **Calculate Expected Frequencies**: Since we're assuming a fair six-sided die, each face has an expected frequency of $\frac{\text{Total rolls}}{\text{Number of sides}} = \frac{120}{6} = 20$.
2. **Set Up the Chi-Square Formula**: The formula for the chi-square statistic is:$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ Where:
   - O_i = Observed frequency for each category (in this case, each die face)
   - E_i = Expected frequency for each category
3. **Calculate the Differences**: For each die face, calculate the difference between the observed and expected frequencies, then square the result.$(O_i - E_i)^2$
4. **Divide by Expected Frequencies**: Divide each squared difference by the corresponding expected frequency.$\frac{(O_i - E_i)^2}{E_i}$
5. **Sum Up the Results**: Sum up all the values obtained in step 4 to get the chi-square statistic.



Die Face	1	2	3	4	5	6
Observed	22	18	19	21	20	20
Expected	20	20	20	20	20	20

1. Calculate differences:

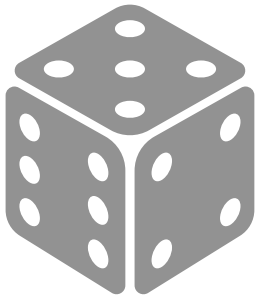
$$(22-20)^2 = 4$$
$$(18-20)^2 = 4$$
$$(19-20)^2 = 1$$
$$(21-20)^2 = 1$$
$$(20-20)^2 = 0$$
$$(20-20)^2 = 0$$

2. Divide by expected frequencies:

$$\frac{4}{20} = 0.2$$
$$\frac{4}{20} = 0.2$$
$$\frac{1}{20} = 0.05$$
$$\frac{1}{20} = 0.05$$
$$\frac{0}{20} = 0$$
$$\frac{0}{20} = 0$$

3. Sum up:

$$\chi^2 = 0.2 + 0.2 + 0.05 + 0.05 + 0 + 0 = 0.5$$

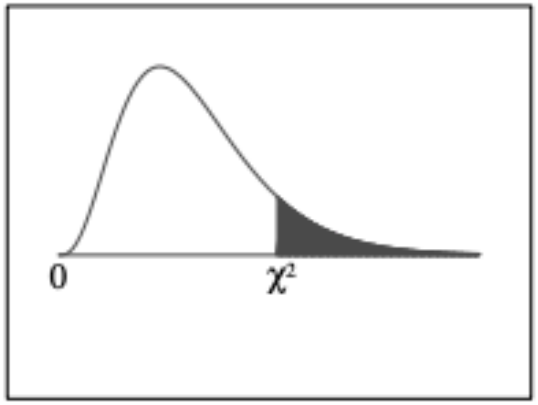


1. **Determine the Degrees of Freedom (df)**: In the case of a goodness of fit test, the degrees of freedom (df) is equal to the number of categories minus 1. **Df = 6-1 = 5**
2. **Consult the Chi-Squared Distribution Table or Software**: You can use statistical software or a chi-squared distribution table to find the critical value corresponding to your calculated chi-squared statistic at the chosen significance level (usually 0.05).

3. Sum up:
$\chi^2 = 0.2 + 0.2 + 0.05 + 0.05 + 0 + 0 = 0.5$

https://www.math.arizona.edu/~jwatkins/chi-square-table.pdf

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	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

2) Chi-squared Independence test

- **Purpose**: The test of independence assesses whether two categorical variables are independent of each other. It examines if the occurrence of **one category is independent of the occurrence of another category.**
- **Applications**: This test is widely used in observational studies where the goal is to discover **associations or relationships between variables**. For example, determining whether there is a relationship between gender and voting preference, or between education level and product choice.

Procedure:

- **Step 1**: Define the null hypothesis, typically stating that the two variables are independent of each other.
- **Step 2**: Create a contingency table showing the frequency distribution of the variables.
- **Step 3**: Calculate expected frequencies for each cell in the contingency table, assuming the null hypothesis is true.
- **Step 4**: Compute the chi-squared statistic in a similar manner as the goodness of fit test, using the observed and expected frequencies.
- **Step 5**: Compare the chi-squared statistic to the critical value from the chi-squared distribution to decide whether to reject the null hypothesis.

Example

Candy Preferences Scenario: Suppose we want to determine if there's a relationship between people's favorite candy preferences and their gender.

Hypothesis:

- **Null Hypothesis (H0): There is no association between candy preference and gender.**
- **Alternative Hypothesis (H1): There is an association between candy preference and gender.**

Data

- We survey 100 people, asking them their favorite candy (chocolate, gummy bears, or lollipops) and their gender (male or female).

	Chocolate	Gummy Bears	Lollipops	Total
Male	20	30	10	60
Female	30	20	20	70
Total	50	50	30	130

Analysis:

- We calculate the expected frequencies assuming independence between candy preference and gender.
- We then perform the chi-squared test to see if the observed frequencies significantly differ from the expected frequencies.
- If the p-value is low (typically below 0.05), we reject the null hypothesis and conclude that there is evidence of an association between candy preference and gender.

Data

- We survey 100 people, asking them their favorite candy (chocolate, gummy bears, or lollipops) and their gender (male or female).

	Chocolate	Gummy Bears	Lollipops	Total
Male	20	30	10	60
Female	30	20	20	70
Total	50	50	30	130

Expected Frequencies (under the assumption of independence):

To calculate the expected frequency for each cell, we use the formula:

$$E_{ij} = \frac{\text{Row Total}_i \times \text{Column Total}_j}{\text{Grand Total}}$$

For example, the expected frequency for males who prefer chocolate is:

$$E_{\text{Male, Chocolate}} = \frac{60 \times 50}{130} = \frac{3000}{130} \approx 23.08$$

Expected Frequencies:

	Chocolate	Gummy Bears	Lollipops	Total
Male	23.08	23.08	13.84	60
Female	26.92	26.92	16.16	70
Total	50	50	30	130

Now, let's compute the chi-squared statistic using the formula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Expected Frequencies:

	Chocolate
Male	23.08
Female	26.92
Total	50

Observed Frequencies:

	Chocolate
Male	20
Female	30
Total	50

For "Male Chocolate":

- Observed frequency ($O_{\text{Male, Chocolate}}$) = 20
- Expected frequency ($E_{\text{Male, Chocolate}}$) ≈ 23.08

Now, we use the formula to calculate the contribution to the chi-squared statistic for this cell:

$$\frac{(O_{\text{Male, Chocolate}} - E_{\text{Male, Chocolate}})^2}{E_{\text{Male, Chocolate}}}$$

$$= \frac{(20 - 23.08)^2}{23.08}$$

$$= \frac{(-3.08)^2}{23.08}$$

$$\approx \frac{9.4864}{23.08}$$

$$\approx 0.411$$

Now, let's compute the chi-squared statistic using the formula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

For "Male Chocolate":

- Observed frequency ($O_{\text{Male, Chocolate}}$) = 20
- Expected frequency ($E_{\text{Male, Chocolate}}$) ≈ 23.08

Now, we use the formula to calculate the contribution to the chi-squared statistic for this cell:

$$\frac{(O_{\text{Male, Chocolate}} - E_{\text{Male, Chocolate}})^2}{E_{\text{Male, Chocolate}}}$$

$$= \frac{(20 - 23.08)^2}{23.08}$$

$$= \frac{(-3.08)^2}{23.08}$$

$$\approx \frac{9.4864}{23.08}$$

$$\approx 0.411$$

Summing up these contributions:

$$0.411 + 2.399 + 1.472 + 0.427 + 2.524 + 1.497 \approx 8.73$$

$$df = (r-1) \times (c-1)$$

Where:

- r = number of rows in the contingency table (excluding the total row)
- c = number of columns in the contingency table (excluding the total column)

In our example:

- $r = 2$ (Number of rows: Male and Female)
- $c = 3$ (Number of columns: Chocolate, Gummy Bears, and Lollipops)

$$df = (2-1) \times (3-1) = 1 \times 2 = 2$$

Summing up these contributions:

$$0.411 + 2.399 + 1.472 + 0.427 + 2.524 + 1.497 \approx 8.73$$

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	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

Practice

Example

Data:

Suppose the company provides the following expected color ratios for regular size packs:

- Brown: 30%
- Red: 20%
- Yellow: 20%
- Green: 10%
- Blue: 10%
- Orange: 10%

Hypothesis:

- **Null Hypothesis (H0): The observed distribution of M&M's chocolate colors in regular size packs matches the expected color ratios provided by the company.**
- **Alternative Hypothesis (H1): The observed distribution of M&M's chocolate colors in regular size packs differs from the expected color ratios provided by the company.**

Example

Data:

Suppose the company provides the following expected color ratios for regular size packs:

- Brown: 30%
- Red: 20%
- Yellow: 20%
- Green: 10%
- Blue: 10%
- Orange: 10%

Using the formula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

We collect data from 100 regular size M&M's packs and record the counts of each color.

Color	Observed Count	Expected Count (based on company ratio)
Brown	35	30
Red	15	20
Yellow	20	20
Green	10	10
Blue	12	10
Orange	8	10

=> Goodness of fit test

Calculation

Using the formula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

For example, for the color Brown:

$$\chi^2_{\text{Brown}} = \frac{(35-30)^2}{30} = \frac{25}{30} = 0.833$$

Repeat this calculation for each color:

- For Red: $\chi^2_{\text{Red}} = \frac{(15-20)^2}{20} = \frac{25}{20} = 1.25$
- For Yellow: $\chi^2_{\text{Yellow}} = \frac{(20-20)^2}{20} = \frac{0}{20} = 0$
- For Green: $\chi^2_{\text{Green}} = \frac{(10-10)^2}{10} = \frac{0}{10} = 0$
- For Blue: $\chi^2_{\text{Blue}} = \frac{(12-10)^2}{10} = \frac{4}{10} = 0.4$
- For Orange: $\chi^2_{\text{Orange}} = \frac{(8-10)^2}{10} = \frac{4}{10} = 0.4$

Now, sum up these contributions:

$$\chi^2 = 0.833 + 1.25 + 0 + 0 + 0.4 + 0.4 = 2.883$$