Chapter 11: The Chi-Square Distribution

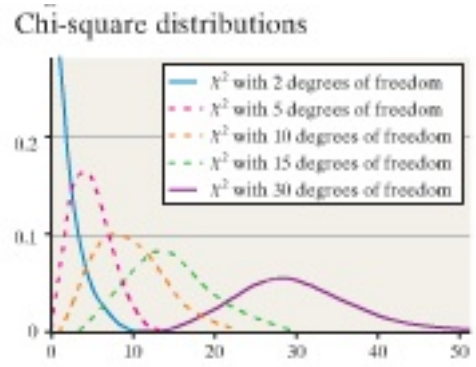
11.1: Facts About the Chi-Square Distribution

In this topic, you will learn an inference technique that applies to a categorical variable with *more than two* categories. One of the most famous and widely used procedures in all of statistics is the chi-square goodness-of-fit test. This procedure assesses how closely sample results conform to a hypothesized model about the proportional breakdown of the various categories.

The Chi-Square Family of Distributions

A **chi-square distribution**:

1. It is a right-skewed distribution.
2. There is a different chi-square curve for each df. The $df = k - 1$
3. The test statistic for any test is always greater than or equal to zero.
4. When $df > 90$, the chi-square curve approximates the normal distribution. For $X \sim \chi^2_{1000}$ the mean, $\mu = df = 1000$ and the standard deviation, $\sigma = \sqrt{2(1000)} = 44.7$. Therefore, $X \sim$ normal $(1000, 44.7)$.
5. The mean, μ, is located just to the right of the peak.



Notation for Chi-Square Statistic Goodness-of-Fit Statistic

The following notation can be used for the chi-square test:

$O =$ observed count in each category.
$E = np =$ expected count in each category.
$k =$ number of categories; $df = k - 1$
$n =$ total sample size
$p =$ probability for each category.

11.2: Goodness-of-Fit Test

A **chi-square goodness-of-fit test** is used to test the hypothesis that an observed frequency distribution fits (or conforms to) some claimed distribution.

Step 1: Determine the Null and Alternative Hypotheses

H_0 : The distribution is the same.
H_a : The distribution is different.

$H_0 : p_1 = p_2 = \ldots = p_n$
H_a : At least one is different.

Step 2: Verify That The Conditions Are Met And State the Level of Significance

1. The data are representative of the population.
2. The sample data consist of frequency counts for each of the different categories.
3. For each category, the expected count is at least 5.

Note: If at least one of these conditions are violated, the results may not be valid.

Step 3: Summarize the Data into an Appropriate Test Statistic

The Chi-Square Statistic for Goodness-of-Fit

- The **observed counts**, O, are the counts in the cells of a one-way table of the sample data.
- The **expected counts**, E, are hypothetical counts that would occur in the cells of the table if the null hypothesis were true. Expected counts are calculated by:

$$\text{Expected } = n * p$$

Expected Counts for Goodness-of-Fit Test

Suppose that we hypothesize that 45% of a population has brown eyes, 35% has blue eyes, and 20% has an eye color other than brown or blue. If we randomly select $n = 200$ people from this population, it probably makes sense to you that we would expect about 45% of $200 = 0.45 \times 200 = 90$ people to have brown eyes if our hypothesis is correct. We also would expect about $0.35 \times 200 = 70$ people to have blue eyes and $0.20 \times 200 = 40$ people to have an eye color other than brown or blue.

The expected count for each category is the number of trials of the experiment times the probability of success in the category, $E = np$.

- The **chi-square statistic** measures the difference between the observed counts and corresponding expected counts. It can be calculated as:

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \sum \frac{(O-E)^2}{E}$$

  Note: Although actual observed counts must be whole numbers, do not round off the expected counts to whole numbers when using them to compute the chi-square statistic.

Step 4: Find the p-Value OR determine the critical value

The **p-value** for the chi-square test is the probability that the calculated chi-square statistic, χ^2, could be as large as it is or larger than it is if the null hypothesis is true.

The **degrees of freedom** are given by $df = k - 1$.

In the eye color example above, there are three categories, so $k = 3$, and the degrees of freedom are $df = 3 - 1 = 2$.

Three Ways to Determine the p-Value

Here are three different ways to determine the p-value:

- Use statistical software to do the chi-square test. The p-value will be part of the output.
- Use the Excel command =chisq.dist.rt(chi-square value, df). This command will return the area to the right of the value you specify.
- Use the χ^2-table to give a range for the p-value.

Using the χ^2-table to Approximate the p-Value

Look in the corresponding "df" row of the χ^2-table for the degrees of freedom. Scan across that row until you locate approximately where the calculated chi-square test statistic falls.

- If the value of the chi-square statistic falls between two table entries, the p-value is between the two values at the top of those two columns.
- If the value of the chi-square statistic is larger than the entry in the rightmost column (labeled 0.0005), then the p-value is less than 0.0005 ($p < 0.0005$).
- If the value of the chi-square statistic is smaller than the entry in the leftmost column (labeled 0.20), the p-value is greater than 0.20 ($p > 0.20$).

Step 5: Make a decision based on either the p-value or the rejection region

- If p-value $< \alpha$, Reject H_0
- If the test statistic is in the shaded region (rejection region), reject H_0

Step 6: State your conclusion in terms of the problem

Interpret the conclusion in context of the situation. We should also consider the manner in which the data were collected.

Example 1. Find the critical value for a hypothesis test where $\alpha = 0.05$, assuming 15 degrees of freedom.

Example 2. Find the critical value for a hypothesis test where $\alpha = 0.10$, assuming 27 degrees of freedom.

Testing Hypotheses for Goodness-of-Fit Tests

1. Set up **hypothesis**:

   H_0 : The distribution is the same.
   H_1 : The distribution is different.

   OR

   $H_0 : p_1 = p_2 = \ldots = p_n$
   H_1 : At least one is different.

2. Verify the following two **requirements** are satisfied:
   (a) The sample is representative of the population.
   (b) The expected counts are at least five for each category.

3. Compute the **test statistic**
   $$\chi_0^2 = \sum \frac{(O-E)^2}{E}$$
   where $E_i = np$

4. Find the p**-Value**. Use the χ^2-table.
   where $df = k - 1$

4. OR determine the **critical value**, χ^{2*}.
   where $df = k - 1$

5. **Decision Rule:** If the p-value $< \alpha$, reject the null hypothesis.

5. OR **Decision Rule:** If the test statistic is MORE EXTREME than the critical value, reject the null hypothesis.

6. **Conclusion** - State your decision and your conclusion in terms of the problem.
   If you Reject H_0: There is sufficient evidence to conlcude [statement in H_a].
   If you Fail to Reject H_0: There is not sufficient evidence to conlcude [statement in H_a].

Example 3. An urban economist wonders if the distribution of U.S. residents in the United States is different today than it was in 2000. That year, 19.0% of the population of the United States resided in the Northeast, 22.9% resided in the Midwest, 35.6% resided in the South, and 22.5% resided in the West (based on data obtained from the Census Bureau). The economist randomly selects 1500 households in the United States and obtains the frequency distribution shown below.

Region	Observed Counts	Expected Counts
Northeast	274	
Midwest	303	
South	564	
West	359	

Does the evidence suggest that the distribution of residents in the United States is different today from the distribution in 2000 at the $\alpha = 0.05$ level of significance?

Step 1

Step 2 Since all the expected counts are greater than or equal to 5, the requirements for the goodness-of-fit test are satisfied.

Step 3

Step 4

Step 5

Step 6

Step 4

Step 5

Example 4. An obstetrician wants to know whether or not the proportions of children born on each day of the week are the same. She randomly selects 500 birth records and obtains the data shown below. Is there any reason to believe that the day on which a child is born occurs with equal frequency at the $\alpha = 0.01$ level of significance?

Day of Week	Frequency
Sunday	57
Monday	78
Tuesday	74
Wednesday	76
Thursday	71
Friday	81
Saturday	63

Step 1

Step 2

Step 3

Step 4

Step 5

Step 6

Step 4

Step 5

11.3: Tests of Independence

A **contingency table** is a table consisting of frequency counts of categorical data corresponding to two different variables. May also be called a **two-way table**.

Each combination of a row variable category and a column variable category is referred to as a **cell** of the contingency table.

In a **test of independence**, we test the null hypothesis that in a contingency table, the row and column variables are independent.

Notation for Chi-Square Statistic for Two-Way Tables

The following notation can be used for the chi-square test:

$O =$ observed count for each cell.
$E =$ expected count for each cell.
$R =$ total count for the cells in each row.
$C =$ total count for the cells in each column.
$n =$ total count for the entire table.
$E = \dfrac{RC}{n} =$ expected count for each cell.
$r =$ number of rows.
$c =$ number of columns.

Step 1: Determine the Null and Alternative Hypotheses

H_0 : The two variables are independent.
H_a : The two variables are dependent.
These phrases are also acceptable for the null hypothesis: *not associated* and *not related*.

Step 2: Verify That The Conditions Are Met And State the Level of Significance

Necessary Conditions

1. The sample is representative of the population.
2. The sample data consist of frequency counts in a two-way table.
3. For every cell in the contingency table, the expected frequency E is at least 5.

Note: If at least one of these conditions are violated, the results may not be valid.

Step 3: Summarize the Data into an Appropriate Test Statistic

The Chi-Square Statistic for Two-Way Tables

- The **observed counts** are the counts in the cells of a contingency table of the sample data.

- The **expected counts** are hypothetical counts that would occur in the cells of the table if the null hypothesis were true. Expected counts are calculated by:

$$\text{Expected } = \frac{\text{Row total } \times \text{ Column total}}{\text{Total } n}$$

- The **chi-square statistic** measures the difference between the observed counts and corresponding expected counts. It can be calculated as:

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Observed } - \text{ Expected})^2}{\text{Expected}} = \sum \frac{(O-E)^2}{E}$$

Interpreting the Expected Counts

The expected counts in a chi-square analysis are the counts that would be expected to occur if the null hypotheses were true. Three important conditions:

- The expected counts have the same row and column totals as the observed counts.
- The pattern of row percentages is identical for all rows of expected counts.
- The pattern of column percentages is identical for all columns of expected counts.

Note: Although actual observed counts must be whole numbers, do not round off the expected counts to whole numbers when using them to compute the chi-square statistic.

Step 4: Find the p-Value OR determine the critical value

The **p-value** for the chi-square test is the probability that the calculated chi-square statistic, χ^2, could be as large as it is or larger than it is if the null hypothesis is true.

The **degrees of freedom** are given by $df = (\text{Rows} - 1)(\text{Columns} - 1)$, where "Rows" indicates the number of rows in the table, and "Columns" indicates the number of columns in the table.

$$\text{Degrees of Freedom } = df = (r-1)(c-1)$$

Using the χ^2-table to Approximate the p-Value

Look in the corresponding "df" row of the χ^2-table for the degrees of freedom. Scan across that row until you locate approximately where the calculated chi-square test statistic falls.

- If the value of the chi-square statistic falls between two table entries, the p-value is between the two values at the top of those two columns.
- If the value of the chi-square statistic is larger than the entry in the rightmost column (labeled 0.0005), then the p-value is less than 0.0005 ($p < 0.0005$).
- If the value of the chi-square statistic is smaller than the entry in the leftmost column (labeled 0.20), the p-value is greater than 0.20 ($p > 0.20$).

Step 5: Make a decision based on either the p-value or the rejection region

- If p-value $< \alpha$, Reject H_0
- If the test statistic is in the shaded region (rejection region), reject H_0

Step 6: State your conclusion in terms of the problem

Interpret the conclusion in context of the situation. We should also consider the manner in which the data were collected.

Suppose we were testing whether there is a relationship between smoking (yes or no) and drinking alcohol (never, occasionally, often). Here are various ways to write the possible conclusions:

Ways to write the conclusion "do not reject the null hypothesis"

- There is not sufficient evidence to conclude smoking and drinking alcohol are independent.
- The relationship between smoking and drinking alcohol is not statistically significant.
- The proportions of smokers who never drink, drink occasionally, and drink often are not significantly different from the proportions of nonsmokers who do so.
- There is insufficient evidence to conclude that there is a relationship in the population between smoking and drinking alcohol.

Ways to write the conclusion "reject the null hypothesis"

- There is sufficient evidence to conclude smoking and drinking alcohol are dependent.
- There is a statistically significant relationship between smoking and drinking alcohol.
- The proportions of smokers in the population who never drink, drink occasionally, and drink often are not all the same as the proportions of nonsmokers who do so.
- Smokers have significantly different drinking behavior than nonsmokers.

Testing Hypotheses for Tests of Independence AND Tests of Equal Proportions

1. Set up **hypothesis**:

   H_0 : The two variables are independent in the population.
   H_a : The two variables are dependent in the population.

2. Verify the following two **requirements** are satisfied:
   (a) The sample is representative of the population.
   (b) The sample data consist of frequency counts in a two-way table.
   (c) For every cell in the contingency table, the expected frequency E is at least 5.

3. Compute the **test statistic**

$$\chi_0^2 = \sum \frac{(O-E)^2}{E}$$

   where $E = \dfrac{(\text{Row Total})(\text{Column Total})}{\text{Grand Total}}$

4. Find the ***p*-Value**. Use the χ^2-table.
   where $df = (r-1)(c-1)$

4. OR determine the **critical value**, χ^{2*}.
   where $df = (r-1)(c-1)$

5. **Decision Rule:** If the p-value $< \alpha$, reject the null hypothesis.

5. OR **Decision Rule:** If the test statistic is MORE EXTREME than the critical value, reject the null hypothesis.

6. **Conclusion** - State your decision and your conclusion in terms of the problem.
   (See previous page)

Example 5. Is there a relationship between marital status and happiness? In the 2006 General Social Survey, 2985 randomly sampled individuals were asked their level of happiness and marital status. The results are shown below. Does the sample evidence suggest that one's happiness depends on one's marital status? Use the $\alpha = 0.05$ level of significance.

		Marital Status			
		Married	Widowed	Divorced/Separated	Never Married
Happiness	Very Happy	600	63	112	144
		(435.024)	(78.815)	(180.413)	(224.747)
	Pretty Happy	720	142	355	459
		(793.363)	(143.737)	(329.024)	(409.876)
	No Too Happy	93	51	119	127
		(184.613)	(33.447)	(76.563)	(95.377)

Step 1

Step 2 Since all of the expected frequencies are greater than 5 and we have a SRS, the requirements for the test of independence are satisfied.

Step 3

Step 4

Step 5

Step 6

Step 4

Step 5

Example 6. An obstetrician wants to learn whether the amount of prenatal care and the wantedness of the pregnancy are associated. He randomly selects 939 women who had recently given birth and asks them to disclose whether their pregnancy was intended, unintended, or mistimed. In addition, they were to disclose when they started receiving prenatal care, if ever. Let $\alpha = 0.05$. The results of the survey are as follows:

	Months Pregnant Before Prenatal Care Began		
Wantedness of Pregnancy	Less Than 3 Months	3 to 5 Months	More Than 5 Months (or never)
Intended	593	26	33
Unintended	64	8	11
Mistimed	169	19	16

Step 1

Step 2 Expected Counts are:

Wantedness of Pregnancy	Less Than 3 Months	3 to 5 Months	More Than 5 Months (or never)
Intended			
Unintended			
Mistimed			

Only one expected count is less than 5, so we need to be cautious. Otherwise, the requirements for the test of independence are satisfied.

Step 3

Step 4

Step 5

Step 6

Step 4

Step 5

11.5: Comparison of the Chi-Square Tests

You have seen the χ^2 test statistic used in three different circumstances. The following bulleted list is a summary that will help you decide which χ^2 test is the appropriate one to use.

- **Goodness-of-Fit:** Use the goodness-of-fit test to decide whether a population with an unknown distribution "fits" a known distribution. In this case there will be a single qualitative survey question or a single outcome of an experiment from a single population. Goodness-of-Fit is typically used to see if the population is uniform (all outcomes occur with equal frequency), the population is normal, or the population is the same as another population with a known distribution. The null and alternative hypotheses are:

  H_0: The population fits the given distribution.
  H_a: The population does not fit the given distribution.

- **Homogeneity:** Use the test for homogeneity to decide if two populations with unknown distributions have the same distribution as each other. In this case there will be a single qualitative survey question or experiment given to two different populations. The null and alternative hypotheses are:

  H_0: The two populations follow the same distribution.
  H_a: The two populations have different distributions.

- **Independence:** Use the test for independence to decide whether two variables (factors) are independent or dependent. In this case there will be two qualitative survey questions or experiments and a contingency table will be constructed. The goal is to see if the two variables are unrelated (independent) or related (dependent). The null and alternative hypotheses are:

  H_0: The two variables (factors) are independent.
  H_a: The two variables (factors) are dependent.