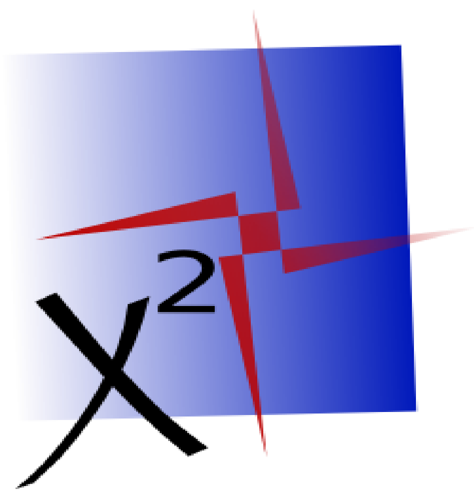
Disclaimer

Chapter 11

CHI-SQUARE AND F-DISTRIBUTIONS



CASE STUDY:

For those who drive, is gender of the driver independent of stopping at a stop sign?

To answer the question, researchers observed a random sample of drivers and documented whether they stopped or didn't stop at a stop sign along with their gender. After several days of observations, an analysis was performed. It was determined that gender of a driver is indeed independent of stopping at a stop sign.

So, what was the methodology used to determine the results?

Well, they had to do the following:

1. Create a contingency table of the variables being analyzed.
2. Perform a hypothesis test to determine if the variables are independent of one another.

In the following chapter, you will learn about these concepts and their applications.

You will also learn how to:

3. Perform an Analysis of Variance (ANOVA).

Section 11.1:

X^2 Test of Independence

Student Learning Outcomes

By the end of the section,

1. **You will conduct a chi-square test of independence.**

In Chapter 10, the relationship between two quantitative variables was examined. In this section, the relationship between two qualitative variables will be explored. When analyzing the relationship between two qualitative variables, researchers often use the term "independent".

Recall, two variables are independent if one variable does not affect the other variable. To analyze whether two qualitative variables are independent of each other, the data must first be organized via a contingency table also known as a cross-tabulation table or two-way table.

DEFINITION

Contingency Table: a table of rows and columns that shows the observed counts of data. The categories of one qualitative variable are represented via the rows and the categories of the other qualitative variable are represented via the columns.

A contingency table with 2 rows and 4 columns would be referred to as a 2 × 4 contingency table. Likewise, a contingency table with 5 rows and 3 columns would be referred to as a 5 × 3 contingency table.

Once a contingency table is created, a **chi-square test of independence** can be performed to test for independence between two qualitative variables.

Note: Chi-square is pronounced "ki"-square.

Performing a chi-square test of independence requires the use of the chi-square distribution.

DEFINITION

Chi-square (X^2) distribution: **The X^2-distribution has the following properties:**

1. **There are different curves for the different degrees of freedom.**
2. **The total area under the curve sums to 1.**
3. **The curve approaches but never touches the x-axis.**
4. **The distribution is skewed right.**
5. **The X^2-values are greater than or equal to 0.**

The chi-square distribution can be seen by viewing the table in the Appendix.

The procedure for conducting a chi-square test of independence is the following:

Procedure for Conducting a Chi-Square Test of Independence

1. **Hypotheses:**

   H_0: The variables are independent.
   H_A: The variables are dependent.

2. **Level of Significance**

3. **Test Statistic:** $\chi^2 = \sum \frac{(O - E)^2}{E}$

   with df = (r – 1)(c – 1)

4. **Critical Value or P-value**

5. **Decision and Conclusion**

Summary of Notations:

- O represents the observed counts
- E represents the expected counts
- r represents the number of rows
- c represents the number of columns

The chi-square test of independence can only be performed when both of the following conditions are met:

- The observed counts are obtained from a random sample.
- The expected counts are greater than or equal to five.

Note: The expected counts are calculated based on the assumption that the null hypothesis is true, which is the assumption that the variables are independent. If the variables are indeed independent, then minimal variation can be expected between the observed counts and expected counts. Thus, the chi-square test statistic will be small (i.e. close to 0). As a result, the rejection of the null hypothesis is improbable. However, if the variables are not independent, the difference between the observed counts and expected counts will be large. Thus, the chi-square test statistic will also be large. As a result, this provides evidence against the null hypothesis which in turn will allow researchers to reject the null hypothesis.

Let's explore an example.

Are views on abortion independent of age? A recent poll asked senior citizens and teenagers whether they were pro-life or pro-choice when it comes to abortion. The results of the survey are the following:

	Pro-Life	Pro-Choice	Totals
Senior Citizens	196	199	395
Teenagers	239	249	488
Totals	435	448	883

Test whether an individual's opinion regarding abortion is independent of age at the 5% level of significance.

Note: This table has two rows and two columns. Thus, it is a 2 × 2 contingency table. With regards to the individual cell counts, there were 196 senior citizens who were pro-life and 199 senior citizens who were pro-choice. There were 239 teenagers who were pro-life and 249 teenagers who were pro-choice.

Step 1: Hypotheses. For the null and alternative hypotheses, the statements will **always** be the variables are independent and the variables are dependent, respectively. Ensure to specifically identify the two variables when stating the null and alternative hypotheses. Thus, the hypotheses are the following:

H_0: Views regarding abortion and age are independent.
H_A: Views regarding abortion and age are dependent.

Step 2: Level of Significance. The level of significance is 5%. Thus, $\alpha = 0.05$.

Step 3: Test Statistic. First, identify the observed counts (O).

The observed counts (O) are located in the contingency table. They are highlighted below in **red**.

	Pro-Life	Pro-Choice	Totals
Senior Citizens	**196**	**199**	395
Teenagers	**239**	**249**	488
Totals	435	448	883

Second, determine the expected counts (E). They are determined by taking each row total multiplied by the column total, then dividing by the sample size.

$$E = \frac{row\ total \times column\ total}{n}$$

To help compute the expected counts, identify each cell by naming them. The names are highlighted below in **blue**.

	Pro-Life	Pro-Choice	Totals
Senior Citizens	**R1C1**	**R1C2**	395
Teenagers	**R2C1**	**R2C2**	488
Totals	435	448	883

Note: **R1C1** stands for *row 1, column 1* because the cell is in the first row and first column.

R1C2 stands for *row 1, column 2* because the cell is in the first row and second column.

R2C1 stands for *row 2, column 1* because the cell is in the second row and first column.

R2C2 stands for *row 2, column 2* because the cell is in the second row and second column.

The expected counts are highlighted below in **purple**.

	Pro-Life	Pro-Choice	Totals
Senior Citizens	**R1C1 = 194.592**	**R1C2 = 200.408**	395
Teenagers	**R2C1 = 240.408**	**R2C2 = 247.592**	488
Totals	435	448	883

194.592 was found by multiplying the **row total of <u>row 1</u>** by the **column total of <u>column 1</u>**, then dividing by the sample size. Thus, $\frac{395 \times 435}{883} = 194.592$

200.408 was found by multiplying the **row total of <u>row 1</u>** by the **column total of <u>column 2</u>**, then dividing by the sample size. Thus, $\frac{395 \times 448}{883} = 200.408$

240.408 was found by multiplying **the row total of <u>row 2</u>** by the **column total of <u>column 1</u>**, then dividing by the sample size. Thus, $\frac{488 \times 435}{883} = 240.408$

247.592 was found by multiplying the **row total of <u>row 2</u>** by the **column total of <u>column 2</u>**, then dividing by the sample size. Thus, $\frac{488 \times 448}{883} = 247.592$

Third, place both the observed counts and expected counts in the cells.

	Pro-Life	Pro-Choice	Totals
Senior Citizens	**196 – 194.592**	**199 – 200.408**	395
Teenagers	**239 – 240.408**	**249 – 247.592**	488
Totals	435	448	883

Fourth, plug the values into the formula.

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

$$\chi^2 = \frac{(196 - 194.592)^2}{194.592} + \frac{(199 - 200.408)^2}{200.408} + \frac{(239 - 240.408)^2}{240.408} + \frac{(249 - 247.592)^2}{247.592}$$

$$\chi^2 = 0.010 + 0.0099 + 0.0082 + 0.008$$

$$\chi^2 = 0.036$$

Note: The chi-square test statistic will always be positive because squared values are used.

Step 4: Critical Value or P-value.

Critical Value Approach: For this example, α = 0.05 and it's a right-tailed test. To determine the degrees of freedom, find r – 1 and c – 1. Recall for this example, there are 2 rows and 2 columns.

Thus, r – 1 = 2 – 1 = 1 and c – 1 = 2 – 1 = 1. Last, multiply (r – 1) by (c – 1) = 1 x 1 = 1. Therefore, the degree of freedom is 1. The critical value is 3.8415.

***Note*: A chi-square test of independence will always be a one-tailed test with area in the right tail. Therefore, the critical value will always be positive.**

P-value Approach: Using technology, the p-value is 0.8489.

Step 5: **Decision and Conclusion.** To make a decision, either compare the critical value to the test statistic **OR** compare the p-value to the level of significance.

Critical Value Approach: For this approach, compare the value of the test statistic to the critical value. For a right-tailed test, if the test statistic is greater than the critical value, the null hypothesis can be rejected. For this example, the test statistic of 0.036 is not greater than the critical value of 3.8415. Thus, the null hypothesis cannot be rejected.

***Note*: When conducting a chi-square test of independence, the null hypothesis is always rejected if the test statistic is greater than the critical value. If the test statistic is not greater than the critical value, then the null hypothesis cannot be rejected.**

P-value Approach: The p-value was determined to be 0.8489. By comparing the p-value to the level of significance, the null hypothesis cannot be rejected because the p-value is not $\leq \alpha$.

Since the null hypothesis cannot be rejected, the conclusion is the following:

"There is not sufficient evidence to suggest that views regarding abortion and gender are dependent."

Section 11.2:

Analysis of Variance

Student Learning Outcomes

By the end of the section,

1. **You will conduct an analysis of variance (ANOVA).**

Is there a difference in the mean miles per gallon (mpg) for compact cars, sedans and SUV's? To answer this question, researchers can perform an analysis of variance (ANOVA) to compare the means.

DEFINITION

Analysis of Variance (ANOVA): a statistical procedure used to compare the means of three or more independent populations.

Performing an analysis of variance requires the use of the F-distribution.

DEFINITION

F-distribution: The F-distribution has the following properties:

1. **There is a different curve for each combination of degrees of freedom, DFn (numerator degrees of freedom) and DFd (denominator degrees of freedom).**
2. **The total area under the curve sums to 1.**
3. **The curve approaches but never touches the x-axis.**
4. **The distribution is skewed right.**
5. **The F-values are greater than or equal to 0.**

The F-distribution can be seen by viewing the table in the Appendix.

Note: While viewing the distribution, select the table that corresponds to the desired level of significance. Once the table is selected, identify the numerator degrees of freedom (DFn) via the first row and the denominator (DFd) degrees the freedom via the first column. The intersection of DFn and DFd is the F critical value.

The procedure for conducting an ANOVA is the following:

Procedure for Conducting an Analysis of Variance (ANOVA)

1. **Hypotheses:**

   H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$.
   H_A: Not all the population means are equal.

2. **Level of Significance**

3. **Test Statistic:** $F = \dfrac{MS_B}{MS_W}$

   with DFn = k – 1 and DFd = N – k (where k is the number of samples and N is the sum of the sample sizes.)

4. **Critical Value or P-value**

5. **Decision and Conclusion**

There are 3 key points to keep in mind when performing an ANOVA.

I. An ANOVA can only be performed when the following conditions are met: the samples must be selected from normal populations, the samples must be independent of each other and the samples must have equal variances.

II. The test statistic of an ANOVA is the comparison of the variance between samples (MS_B) to the variance within samples (MS_W). MS_B measures the variation of the sample means. MS_W measures the variation within the samples. In other words, weighted by sample size, MS_B is the sample variance of the sample means and MS_W is the sample mean of the sample variances. If there is no difference between the means, MS_B and MS_W will be equal. Thus, the ratio of MS_B to MS_W will be 1. F test statistics close to 1 suggest that there is not sufficient evidence to reject the null hypothesis. Thus, researchers cannot conclude that not all the population means are equal. However, F test statistics that are significantly greater than 1 means that MS_B is greater than MS_W. As a result, there is some variation between the means. Therefore, it allows researchers to conclude that not all the population means are equal.

III. When the null hypothesis is rejected, researchers can conclude that not all the population means are equal. However, researchers cannot determine which of the population means are different without performing follow-up tests. Performing follow-up tests is beyond the scope of this course.

Let's explore an example.

A research firm plans to analyze the effectiveness of three new innovative teaching techniques. To do so, three cohorts of students are randomly selected. Each cohort of students is randomly assigned one of the three new innovative teaching techniques. Data is collected on 5 students from each cohort, which includes their final exam scores. The final exam scores are the following:

Cohort #1	Cohort #2	Cohort #3
80	70	80
85	70	70
90	80	75
80	90	80
90	90	70

Assuming the populations are normally distributed, the samples are independent and the population variances are equal, is there enough evidence to suggest that not all the population mean final exam scores are equal? Test at the 5% level of significance.

Step 1: **Hypotheses.** For the null and alternative hypotheses, the statements will **always** be the population means are equal and not all the population means are equal, respectively. Thus, the hypotheses are the following:

H_0: $\mu_1 = \mu_2 = \mu_3$.
H_A: Not all the population mean final exam scores are equal.

Step 2: **Level of Significance.** The level of significance is 5%. Thus, $\alpha = 0.05$.

Step 3: **Test Statistic.** To compute the test statistic, first find the following:

- The sample mean and variance of each sample.

$$\bar{x} = \frac{\sum x}{n} \qquad s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$

- The mean of all the observations from all samples (i.e. the overall mean).

$$\bar{\bar{x}} = \frac{\sum x}{N}$$

- The sum of squares between the samples.

$$SS_B = \sum n_i(\bar{x}_i - \bar{\bar{x}})^2$$

- The sum of squares within the samples.

$$SS_W = \sum (n_i - 1)s_i^2$$

- The variance between the samples).

$$MS_B = \frac{SS_B}{k-1}$$

- The mean square within the samples (i.e. the variance within the samples).

$$MS_W = \frac{SS_W}{N-k}$$

Recall the data from the example is the following:

Cohort #1	Cohort #2	Cohort #3
80	70	80
85	70	70
90	80	75
80	90	80
90	90	70

The sample mean and variance for each sample is the following:

	Cohort #1	Cohort #2	Cohort #3
Mean	85	80	75
Variance	25	100	25

The overall mean, which is the mean of all the observations from all samples, is 80.

Second, compute MS_B and MS_W.

$$MS_B = \frac{SS_B}{k-1} = \frac{\sum n_i(\bar{x}_i - \bar{\bar{x}})^2}{k-1}$$

$$MS_B = \frac{5(85-80)^2 + 5(80-80)^2 + 5(75-80)^2}{3-1}$$

$$MS_B = \frac{250}{2} = 125$$

$$MS_W = \frac{SS_W}{N-k} = \frac{\sum (n_i - 1)s_i^2}{N-k}$$

$$MS_W = \frac{(5-1)25 + (5-1)100 + (5-1)25}{15-3}$$

$$MS_W = \frac{600}{12} = 50$$

Finally, plug the values into the formula to compute the test statistic.

$$F = \frac{MS_B}{MS_W} = \frac{125}{50} = 2.5$$

Note: The F test statistic will always be greater than or equal to 0.

Step 4: Critical Value or P-value.

Critical Value Approach: For this example, α = 0.05 and it's a right-tailed test. To determine the degrees of freedom, determine DFn and DFd, where DFn = k – 1 and DFd = N – k. Recall for this example, there are 3 groups and a total of 15 subjects. Thus, DFn = 2 and DFd = 12. The critical value is 4.1028. *Tip: If there are no entries for the desired degrees of freedom, use the closest value.*

***Note*: An ANOVA will always be a one-tailed test with area in the right tail. Therefore, the critical value will always be positive.**

P-value Approach: Using technology, the p-value is 0.1237.

Step 5: Decision and Conclusion. To make a decision, either compare the critical value to the test statistic **OR** compare the p-value to the level of significance.

Critical Value Approach: For this approach, compare the value of the test statistic to the critical value. Recall, the F test statistic will always be greater than or equal to 0. For this example, the test statistic of 2.5 is not greater than the critical value of 4.1028. Thus, the null hypothesis cannot be rejected.

***Note*: When performing an ANOVA, the null hypothesis is always rejected if the test statistic is greater than the critical value. If the test statistic is not greater than the critical value, then the null hypothesis cannot be rejected.**

P-value Approach: The p-value was determined to be 0.1237. By comparing the p-value to the level of significance, the null hypothesis cannot be rejected because the p-value is not $\leq \alpha$.

Since the null hypothesis cannot be rejected, the conclusion is the following:

"There is not sufficient evidence to suggest that not all the population mean final exam scores are equal."