A red and black logo

Description automatically generated

**Chi-Square**

Traje, Ariel July B.

Rasonado, Ramiel Anthony G.

Macasampon, Donald L.

Domingo, Reynaldo Jr. N.

Franco, Deyvis Felippe P.

|  |  |
| --- | --- |
| Introduction | 2 |
| Learning Objectives |  |
| Equipment and Software |  |
| Safety Guidelines |  |
| Lab Procedures |  |
| Sample Exercises |  |
| Assessment |  |
| Reference Material |  |
| FAQs and Tips |  |
| Glossary |  |
| Appendices |  |
| Contact Information |  |
| Lab Policies |  |

**TABLE OF CONTENTS**

**Introduction**

****The Chi-Square test finds its application across diverse fields, including market research, healthcare, social sciences, and education. In market research, it's utilized to dissect consumer preferences and behaviors among various demographic segments. In healthcare, it aids in scrutinizing the connection between treatment outcomes and patient attributes. Similarly, in the social sciences, it helps unearth relationships between categorical variables in surveys or observational studies. Likewise, in education, it delves into the correlation between student demographics and academic performance or program preferences. This test serves several purposes, including testing independence to discern significant associations between categorical variables, goodness of fit to compare observed and expected distributions, and homogeneity to assess if different populations share similar categorical variable distributions. Its appeal lies in its simplicity as a non-parametric test, not presuming a normal distribution, rendering it adaptable to diverse data types.

There are different applications of the Chi-Square test, and one of these is the Chi-Square Goodness-of-Fit Test, the Chi-Square Test of Association, most commonly known as the Chi-Square Test for Independence, and the Chi-Square Test of Homogeneity. There is also an alternative test when dealing with a small sample size, the Chi-Square test. The Chi-Square Goodness-of-Fit Test is used to compare the observed distribution of a single categorical variable. The Chi-Square of Association is a test to examine the association between the two categorical variables. It also determines whether there is a significant relationship between the variables by comparing the observed frequencies in a contingency table to the frequencies. The chi-square test of homogeneity is a test that is used to compare the distribution of a categorical variable across different populations. It also determines whether the distribution of the categorical variables has significant differences between them. Lastly, the Fisher Exact Test is a statistical test that is used to determine if there are nonrandom associations between the two categorical variables. This is also used as an alternative test to deal with the small sample sizes if the assumptions of the Chi-Square test are not met.

The Chi-Square test, along with its various applications and variations, is a foundation of categorical data analysis, providing researchers with strong insights into the correlations and patterns present in their datasets across a wide range of fields. Its simplicity, adaptability, and capacity to handle non-normal distributions make it an indispensable tool for statistical analysis and hypothesis testing in both academic and practical settings.

**Learning Objectives**

**1. Understanding the Chi-Square:**

- Gain a comprehensive understanding of Chi-Square as a statistical test used to analyze the association between categorical variables.

- Explore its significance in research, including its role in hypothesis testing and identifying patterns in data sets.

**2. Types of Data and Data Organization:**

- Identify various types of categorical data suitable for Chi-square analysis, such as nominal and ordinal data.

- Learn effective strategies for organizing categorical data, creating contingency tables, and preparing data sets for Chi-Square testing.

**3. Chi-Square Tests:**

- Master the application of Chi-Square tests for different purposes:

- Chi-Square test for independence: Assess whether there is a relationship between two categorical variables.

-Determine whether categorical variables have similar distributions across groups using the Chi-Square test.

- Chi-Square test for goodness of fit: Evaluate how well-observed data match expected frequencies based on a theoretical distribution.

- Understand the assumptions underlying each Chi-Square test and know when and how to apply them appropriately in statistical analysis.

**4. Interpreting the Results:**

- Develop proficiency in interpreting Chi-Square test results, including:

- Calculating expected frequencies and degrees of freedom.

- Evaluating p-values and determining statistical significance.

- Interpreting Chi-Square test statistics and understanding their implications for research findings.

**5. Software Proficiency: **

- Acquire hands-on experience using statistical software tools such as R, Python (with libraries like pandas and scipy), SPSS, or Excel for Chi-Square analysis.

- Learn how to input data, perform Chi-Square tests, generate contingency tables, and visualize results using charts and graphs.

**6. Real-world Applications:**

- Apply Chi-Square analysis to real-world scenarios across diverse domains:

- Market research: Analyze consumer preferences and behaviors based on demographic data.

- Social sciences: Investigate relationships between categorical variables in surveys or studies.

- Healthcare: Examine outcomes based on categorical variables like treatment options or patient characteristics.

- Utilize Chi-Square results to make data-driven decisions, draw actionable insights, and contribute to evidence-based decision-making processes.

**7. Critical Thinking and Experimental Design:**

- Develop critical thinking skills to formulate meaningful hypotheses that can be tested using Chi-Square analysis.

- Design experiments or observational studies that leverage Chi-Square as a powerful analytical tool to analyze categorical data and draw valid conclusions.

These learning objectives will enable you to become proficient in Chi-Square analysis as well as gain a broader understanding of statistical analysis principles, data interpretation, and critical thinking skills applicable to a variety of research and analytical purposes.

**Equipment and Software**

1. **Laptop**

A portable computer that can be easily carried around. It's a device designed for personal use and can perform various functions such as browsing the internet, creating documents, playing games, etc.

1. **Computer**

A computer that can be used and can perform various functions such as browsing the internet, creating documents, playing games, etc.

1. **RStudio**

RStudio is used in data analysis to import, access, transform, explore, plot, and model data to make predictions.

* 1. **RStudio Version**

We have used the R 4.3.3 version for Rpackages are extensions to the R statistical programming language.

**Safety Guidelines**

1. Only students who are currently enrolled in computer subjects shall be allowed to use the computing laboratory.
2. Eating, drinking, talking loud, loitering, vandalism, wearing of ball caps, wearing of earrings (male), seat-in, use of personal earphone and headphones inside the laboratory are also prohibited.
3. Students shall not be allowed to change computer settings, install any software or computer games, or alter some components of the system unit and opening of other programs.
4. The computer area should be kept clean, the computer unit shall be properly logout after use and/or return to desktop display and return chair to its proper place before leaving the laboratory.
5. Request for installation of new software should be done prior to the start of classes.

**Lab Procedures**

* First, you need to open your RStudio version 4.3.3 in your laptop/computer, if happen that you don’t have the RStudio in your computer, you may download it here <https://posit.co/download/rstudio-desktop/>.

**A screenshot of a computer

Description automatically generated**

* You need to download some packages in your RStudio like “readxl” so that you can access your file in your RStudio with the extension “.xlsx”, and many more.

**A screenshot of a computer

Description automatically generated**

* Import the set of data that you want to have, for you to be able to do that you need to go here Import Dataset > Browse> “select the data that has an extension of “.xlsx”>IMPORT.

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

* Open **Console** because console is where ****you can type and execute immediately your R commands. The **Script Editor** is where you can write, edit, and save R scripts. The **Environment tab** displays information about your R workspace, including variables and datasets. The **Files** **tab** allows you to navigate your file system and manage files.
* Perform your statistical analysis using the appropriate syntax.

A screenshot of a computer

Description automatically generated

* **When dealing with RStudio, there are some troubleshooting tips that you need to know when dealing with the issues:**
  + **Case Sensitivity:** RStudio is indeed case sensitive, meaning that variable names, function names, and other elements must be typed with the correct capitalization.
  + **Check for Data Issues:** Check for missing values, incorrect data types, or anomalies in your dataset.
  + **Double-check the syntax** of your code, especially punctuation, parentheses, and quotation marks. **NOTE**: RStudio is **CASE SENSITIVE**
  + **Search Online Resources:** Utilizing online resources such as forums, documentation, and tutorials can be helpful for finding solutions to coding problems or understanding R concepts.
  + **Keep Software Up to Date:** Ensure that your software applications, packages, or libraries are up to date. Some errors are not because of the block of codes but because of the version. You need to have a RStudio that is up to date to avoid errors.
  + **Restarting** the software or resetting your workspace can resolve transient issues caused by temporary glitches or memory issues.
  + **Seek Assistance:** If the problem has not been resolved do not hesitate to ask your colleague or professor and explain the problem clearly and provide relevant details for better assistance.
* Lastly, you can be able to run your block of code correctly. Just follow this steps for you to be able to run a RStudio.

**Sample Exercises**

**Chi-Square Test of Association or Independence**

> library(readxl)

> FinalProj <- read\_excel("C:/Users/Acer/Downloads/FinalProj.xlsx")

> View(FinalProj)

> data <- data.frame(Sex = FinalProj$Sex, Preference = FinalProj$Program)

> result <- chisq.test(data)

Warning message:

In chisq.test(data) : Chi-squared approximation may be incorrect

> print("Chi-Square Test of Association")

[1] "Chi-Square Test of Association"

> print(result)

Pearson's Chi-squared test

data: data

X-squared = 44.582, df = 299, p-value = 1

**Chi-Square Of Goodness-Of-Fit Test**

> library(readxl)

> FinalProj <- read\_excel("C:/Users/Acer/Downloads/FinalProj.xlsx")

> View(FinalProj)

> obsFreq <- c(FinalProj$yearLevel)

> expFreq <- c(FinalProj$BirthMonth)

> result <- chisq.test(obsFreq, p = expFreq / sum(expFreq))

Warning message:

In chisq.test(obsFreq, p = expFreq/sum(expFreq)) :

Chi-squared approximation may be incorrect

> print("Chi-Square Of Goodness-Of-Fit Test")

[1] "Chi-Square Of Goodness-Of-Fit Test"

> print(result)

Chi-squared test for given probabilities

data: obsFreq

X-squared = 383.28, df = 299, p-value = 0.0007015

**Chi-Square Test Of Homogeneity**

> library(readxl)

> FinalProj <- read\_excel("C:/Users/Acer/Downloads/FinalProj.xlsx")

> View(FinalProj)

> cHomo <- table(FinalProj$Sex, FinalProj$Program)

> result <- chisq.test(cHomo)

Warning message:

In chisq.test(cHomo) : Chi-squared approximation may be incorrect

> print("Chi-Square Test Of Homogeneity")

[1] "Chi-Square Test Of Homogeneity"

> print(result)

Pearson's Chi-squared test

data: cHomo

X-squared = 37.393, df = 3, p-value = 3.799e-08

**Fisher’s Exact Test**

> library(readxl)

> FinalProj <- read\_excel("C:/Users/Acer/Downloads/FinalProj.xlsx")

> View(FinalProj)

> cont\_table <- table(FinalProj$Sex, FinalProj$Program)

> result <- fisher.test(cont\_table)

> print(result)

Fisher's Exact Test for Count Data

data: cont\_table

p-value = 2.998e-07

alternative hypothesis: two.sided

**Assessment**

**Multiple Choice Questions:**

**1. What is the Chi-square test primarily used for?**

A) Comparing means of two independent groups

B) Testing the association between two categorical variables

C) Determining the correlation coefficient between two continuous variables

D) Assessing the normality of a distribution

**2. When conducting a Chi-square test of independence, what is the null hypothesis?**

A) There is no difference between the observed and expected frequencies

B) There is no association between the two categorical variables

C) There is no linear relationship between the variables

D) There is no significant difference in means between the groups

**True or False Questions:**

**3. A Chi-square test can only be used when dealing with categorical data.**

A) True

B) False

**4. The Chi-square test assumes that the observations are independent of**

**each other.**

A) True

B) False

**RStudio Chi-square Problem:**

5. Suppose you have survey data from 200 individuals regarding their favorite ice cream flavors. The data is represented in the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Flavor | Chocolate | Vanilla | Strawberry | Total |
| Preference | 70 | 60 | 70 | 200 |

Using RStudio, perform a Chi-square test to determine if there is a significant difference in preference for ice cream flavors.

**Reference Material**

**GeeksforGeeks. “Chi-Square Test in R.” GeeksforGeeks, 19 Dec. 2023,**

[www.geeksforgeeks.org/chi-square-test-in-r](http://www.geeksforgeeks.org/chi-square-test-in-r).

**A red and yellow logo

Description automatically generated**

**“Chi-square Test of Independence in R.”** Stats and R,

statsandr.com/blog/chi-square-test-of-independence-in-r.

**“Inference for Categorical Data (Chi-square Tests) | Khan Academy.”** Khan

Academy, [www.khanacademy.org/math/statistics-probability/inference-categorical-data-chisquare-tests](http://www.khanacademy.org/math/statistics-probability/inference-categorical-data-chisquare-tests).

**StatQuest with Josh Starmer.** “Fisher’s Exact Test and the Hypergeometric

Distribution.” YouTube, 13 Mar. 2017, [www.youtube.com/watch?v=udyAvvaMjfM](http://www.youtube.com/watch?v=udyAvvaMjfM)

**Home**- **RDocumentation**. [www.rdocumentation.org](http://www.rdocumentation.org).

**DataCamp. Hypothesis Testing in R Course**. 25 Sept. 2023,

[www.datacamp.com/courses/hypothesis-testing-in-r](http://www.datacamp.com/courses/hypothesis-testing-in-r)

**FAQs and Tips**

1. **Frequently Asked Questions (FAQs):**
   * **Q1: What should I do if I encounter a "Chi-squared approximation may be incorrect" warning message?**

○ A: This warning message can occur when the assumptions of the Chi-Square test are violated, such as small sample sizes or expected cell frequencies less than 5. Consider adjusting your analysis or seeking guidance from your instructor.

* + **Q2: How can I determine which Chi-Square test to use for my data?**

○ A: Choose the appropriate Chi-Square test based on the nature of

your research question and the type of categorical variables involved. Consult statistical resources or discuss with your instructor for guidance.

* + **Q3: Is Chi-Square analysis suitable for all types of data?**

○ A: Chi-Square analysis is primarily used for categorical data. While it can be applied to various research scenarios, ensure that your data meet the assumptions of Chi-Square tests before proceeding with analysis.

1. **Common Issues and Solutions:**

* Issue: Incorrect data import or formatting errors in RStudio.

○ Solution: Double-check the file path, file format, and data structure to ensure compatibility with RStudio. Refer to the provided guidelines for troubleshooting tips.

* **Issue: Inaccurate results or warnings during Chi-Square analysis.**

○ Solution: Review the assumptions of Chi-Square tests and verify the integrity of your data. Consider alternative analysis methods or consult with peers or instructors for assistance.

1. **Tips for Efficient Lab Work:**

* **Tip: Organize your data effectively before conducting Chi-Square analysis.**

○ Ensure clear labeling of variables, proper categorization of data, and removal of any unnecessary information to streamline the analysis process.

* **Tip: Familiarize yourself with statistical software functionalities. **

○ Practice using RStudio or other statistical software tools to import data, perform Chi-Square tests, and interpret results efficiently. Explore online tutorials and resources for additional ****guidance.

1. **Best Practices for Data Analysis:**

* **Practice: Conduct exploratory data analysis (EDA) before performing Chi-Square tests.**

○ Explore data distributions, identify outliers or missing values, and assess the overall quality of your dataset to ensure robust analysis outcomes.

* **Best Practice: Document your analysis procedures and results thoroughly.**

○ Keep detailed records of data manipulation steps, statistical analyses performed, and interpretations made to facilitate reproducibility and transparency in your research.

1. **Optimizing Statistical Interpretation:**

* **Tip: Interpret Chi-Square test results in the context of your research question.**

○ Consider the practical significance of statistical findings and relate them to the broader objectives of your study or experiment.

* **Tip: Utilize visualization techniques to enhance result interpretation.**

○ Create charts, graphs, or contingency tables to visually represent Chi-Square analysis outcomes and facilitate clearer communication of findings.

1. **Continuous Learning and Improvement:**

* Engage in continuous learning by exploring additional resources on Chi-Square analysis, statistical methods, and data interpretation.
* Seek feedback from instructors, peers, or online communities to identify areas for improvement and enhance your proficiency in statistical analysis.

****

**Glossary**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**TERM DEFINITIONS**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Chi-Square Analysis** A statistical method utilized to examine the association between categorical variables within data sets. It assesses whether there is a significant relationship between the variables and aids in identifying patterns or trends.

**Categorical Data** Data that can be grouped into categories or distinct groups based on qualitative characteristics. This type of data is often used in Chi-Square analysis and includes nominal (unordered) and ordinal (ordered) categories.

**Contingency Table** A table used to organize categorical data for Chi-Square analysis. It displays the frequency distribution of variables and enables comparison between different categories.

**Hypothesis Testing** A fundamental concept in statistics that involves making inferences about population parameters based on sample data. It helps validate assumptions and draw conclusions from observed data.

**Degrees of Freedom** In Chi-Square analysis, degrees of freedom represent the number of independent observations in a sample minus the number of parameters estimated from the sample data. It affects the distribution of the Chi-Square test statistics.

**Expected Frequencies** The frequencies that would be observed in each category if there was no association between the variables being studied. Expected frequencies are compared to observed frequencies in Chi-Square analysis to assess goodness of fit.

**Statistical Significance** In Chi-Square analysis, statistical significance is determined by comparing the calculated Chi-Square statistic to a critical value or by evaluating the p-value associated with the test.

**Software Proficiency** It involves skills in data input, manipulation, analysis, and visualization using relevant statistical functions and libraries.

****

**Appendices**

**Appendix A: Chi-Square Distribution Table**

This table provides critical values of the Chi-Square distribution for various degrees of freedom and significance levels (typically 0.05 and 0.01).

|  |  |  |  |
| --- | --- | --- | --- |
| **Degrees of Freedom** | **Significance Level** | **Critical Value (0.05)** | **Critical Value (0.01)** |
| 1 | 0.05 | 3.841 | 6.635 |
| 2 | 0.05 | 5.991 | 9.210 |
| 3 | 0.05 | 7.815 | 11.345 |
| 4 | 0.05 | 9.488 | 13.277 |
| 5 | 0.05 | 11.070 | 15.086 |

**Appendix B: Example Contingency Table**

This table illustrates a hypothetical contingency table used for Chi-Square analysis, showing the distribution of respondents' preferences for two different products across different age groups.

|  |  |  |
| --- | --- | --- |
| **Age Group** | **Product A** | **Product B** |
| 18-25 | 30 | 20 |
| 26-35 | 25 | 35 |
| 36-45 | 15 | 25 |
| 46-55 | 10 | 15 |

**Appendix C: Chi-Square Test Calculations**

This section presents detailed calculations for conducting Chi-Square tests manually. It includes formulas and step-by-step procedures for calculating Chi-Square test statistics, expected frequencies, and degrees of freedom.

**Appendix D: Chi-Square Analysis Software Guides**

Comprehensive guides and tutorials for performing Chi-Square analysis using statistical software packages such as R, Python (with pandas and scipy libraries), SPSS, and Excel. These guides cover data input, test execution, interpretation of results, and visualization techniques.

**Appendix E: Additional Charts and Diagrams**

This section includes supplementary charts, graphs, and diagrams relevant to Chi-Square analysis. Examples may include bar charts illustrating frequency distributions or scatter plots depicting relationships between categorical variables.

**Appendix F: References**

A comprehensive list of additional references, including textbooks, academic articles, and online resources, for further exploration of Chi-Square analysis and related statistical methods. These references provide supplementary information and guidance for researchers interested in delving deeper into the topic.

**Contact Information**

****

**Name:  
Contact Number:  
Email:**

**Name:  
Contact Number:  
Email:**

**Name:  
Contact Number:  
Email:**

**Name:  
Contact Number:  
Email:**

**Name:  
Contact Number:  
Email:**

Donald L. Macasampon  
09565395921  
d.macasampon.539122@umindanao.edu.ph

Ramiel Anthony G. Rasonado   
09076989698  
r.rasonado.534426@umindanao.edu.ph

Reynaldo Jr. Domingo   
09855259431  
r.domingo.534871@umindanao.edu.ph

Deyvis Felippe P. Franco  
09276019979  
d.franco.536439@umindanao.edu.ph

Ariel July B. Traje  
09552677446  
a.traje.535724@umindanao.edu.ph

**Lab Policies**

**Lab Hours:** The laboratory is available for research study purposes during the following hours:

* Monday to Thursday: 9:00 AM to 6:00 PM
* Friday: 9:00 AM to 4:00 PM
* Closed on weekends and university holidays.

**Rules and Expectations for Student Behavior:**

1. **Respectful Conduct**: All students engaging in research activities are expected to demonstrate professionalism and respect towards fellow researchers, instructors, and support staff.
2. **Noise Level:** Maintain an environment conducive to focused research. Conversations should be kept to a minimum to minimize distractions for other researchers.
3. **Equipment Handling:** Exercise caution when handling laboratory equipment and materials. Any damage caused by mishandling will be the responsibility of the researcher involved.
4. **Food and Beverage Policy:** Consumption of food and beverages is strictly prohibited within the laboratory to prevent contamination of research materials and equipment.
5. **Cell Phone Usage:** Mobile devices should be silenced or turned off during research sessions to avoid disruption. Phone calls should be taken outside the laboratory area.
6. **Attendance and Punctuality:** Adherence to scheduled research sessions is mandatory. In the event of unavoidable absence, researchers are required to inform the project supervisor or instructor in advance.
7. **Submission Deadlines:** Research tasks and reports must be completed and submitted according to the specified deadlines to ensure timely progress of the study.
8. **Collaboration Guidelines:** Collaboration among researchers is encouraged, provided that everyone contributes independently to the research objectives. Plagiarism or unauthorized use of others' work will not be tolerated.
9. **Safety Protocols:** Strict adherence to safety protocols is essential to prevent accidents and ensure the well-being of all researchers. Any safety concerns or incidents should be reported immediately to the project supervisor or laboratory manager.
10. **Cleanliness Standards:** Researchers are responsible for maintaining cleanliness and organization within their respective workstations. All research materials should be properly stored, and work areas should be tidied up at the end of each session.
11. **Prohibited Items:** Researchers are prohibited from bringing unauthorized materials or personal belongings into the laboratory without prior approval.
12. **Exit Protocol:** Before leaving the laboratory, researchers must ensure that all equipment is switched off, research materials are properly stored, and workstations are left clean and orderly.

Adherence to these lab policies is crucial for the smooth operation of the research study and the safety and well-being of all involved. Failure to comply with these policies may result in disciplinary action and could impact the progress and outcom****es of the research study.