Lab 7 – Chi-Square Goodness of Fit

R Abhijit Srivathsan

# Introduction

The Goodness-of-Fit test is an essential statistical tool for evaluating how closely observed data aligns with a specified distribution or hypothesis. This test compares the frequency distribution of a sample to the expected frequencies based on a theoretical model. Among its variations, the Chi-Square Goodness-of-Fit test is widely used for categorical data, helping identify discrepancies between observed and expected values. It plays a crucial role in validating claims such as election outcomes or analyzing categorical data in business and social research. Additionally, the Chi-Square Test for Independence examines relationships between categorical variables, determining whether two attributes are associated or independent.

# Objectives

* Evaluating the accuracy of reported election results using the Goodness-of-Fit test.
* Analyzing passenger class distribution in the Titanic dataset to assess alignment with an equal distribution assumption.
* Analyzing passenger class distribution in the Titanic dataset to assess alignment with an equal distribution assumption.

##OBJECTIVE 1  
library(ggplot2)  
# Reported percentages from the election  
reported\_percentages = c(41.5, 25.7, 32.8)  
reported\_counts = round(reported\_percentages / 100 \* 123) # Expected counts based on 123 voters  
# Simulated data  
set.seed(42) # For reproducibility  
study\_votes = sample(c("Party 1", "Party 2", "Others"),size = 123,replace = TRUE,prob = c(0.415, 0.257, 0.328));study\_votes

## [1] "Party 2" "Party 2" "Party 1" "Party 2" "Others" "Others" "Others"   
## [8] "Party 1" "Others" "Others" "Others" "Others" "Party 2" "Party 1"  
## [15] "Others" "Party 2" "Party 2" "Party 1" "Others" "Others" "Party 2"  
## [22] "Party 1" "Party 2" "Party 2" "Party 1" "Others" "Party 1" "Party 2"  
## [29] "Others" "Party 2" "Others" "Party 2" "Party 1" "Others" "Party 1"  
## [36] "Party 2" "Party 1" "Party 1" "Party 2" "Others" "Party 1" "Others"   
## [43] "Party 1" "Party 2" "Others" "Party 2" "Party 2" "Others" "Party 2"  
## [50] "Others" "Party 1" "Party 1" "Party 1" "Party 2" "Party 1" "Party 2"  
## [57] "Others" "Party 1" "Party 1" "Others" "Others" "Party 2" "Party 2"  
## [64] "Others" "Party 2" "Party 1" "Party 1" "Party 2" "Others" "Party 1"  
## [71] "Party 1" "Party 1" "Party 1" "Others" "Party 1" "Others" "Party 1"  
## [78] "Party 1" "Others" "Party 1" "Others" "Party 1" "Party 1" "Others"   
## [85] "Party 2" "Others" "Party 1" "Party 1" "Party 1" "Party 1" "Others"   
## [92] "Party 1" "Party 1" "Party 2" "Party 2" "Others" "Party 1" "Others"   
## [99] "Party 2" "Others" "Others" "Party 1" "Party 1" "Party 1" "Party 2"  
## [106] "Party 2" "Others" "Others" "Others" "Party 1" "Others" "Party 2"  
## [113] "Party 2" "Others" "Others" "Others" "Party 1" "Party 1" "Others"   
## [120] "Party 2" "Party 1" "Party 1" "Others"

# Observed counts  
observed\_counts = table(study\_votes);observed\_counts

## study\_votes  
## Others Party 1 Party 2   
## 44 46 33

# Chi-Square Goodness-of-Fit Test  
chi\_square\_test = chisq.test( x = observed\_counts, p = reported\_percentages / 100);chi\_square\_test # Convert percentages to proportions

##   
## Chi-squared test for given probabilities  
##   
## data: observed\_counts  
## X-squared = 8.8589, df = 2, p-value = 0.01192

# Bar plot for comparison of observed vs expected  
expected\_data = data.frame( Party = names(observed\_counts),Count = as.numeric(reported\_counts),Type = "Expected");expected\_data

## Party Count Type  
## 1 Others 51 Expected  
## 2 Party 1 32 Expected  
## 3 Party 2 40 Expected

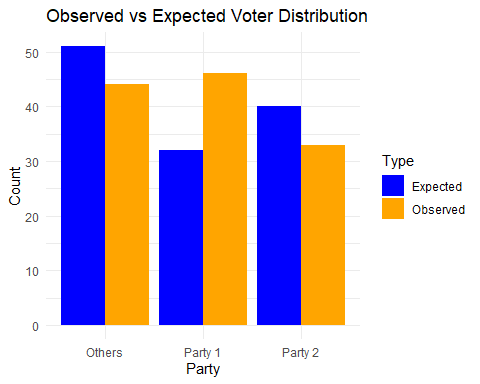
observed\_data = data.frame(Party = names(observed\_counts),Count = as.numeric(observed\_counts),Type = "Observed");observed\_data

## Party Count Type  
## 1 Others 44 Observed  
## 2 Party 1 46 Observed  
## 3 Party 2 33 Observed

combined\_data = rbind(expected\_data, observed\_data);combined\_data

## Party Count Type  
## 1 Others 51 Expected  
## 2 Party 1 32 Expected  
## 3 Party 2 40 Expected  
## 4 Others 44 Observed  
## 5 Party 1 46 Observed  
## 6 Party 2 33 Observed

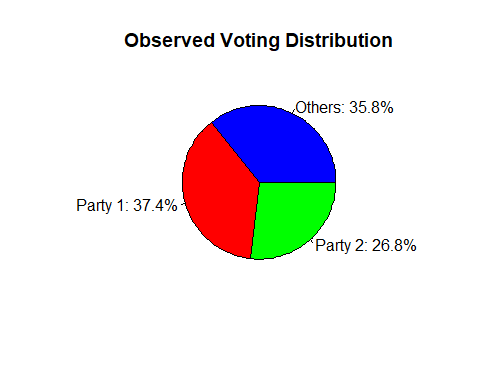
ggplot(combined\_data, aes(x = Party, y = Count, fill = Type)) +geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Observed vs Expected Voter Distribution",x = "Party", y = "Count") +  
 scale\_fill\_manual(values = c("Expected" = "blue", "Observed" = "orange")) +theme\_minimal()



# Pie chart for observed data  
observed\_percentages = round(prop.table(observed\_counts) \* 100, 1);observed\_percentages

## study\_votes  
## Others Party 1 Party 2   
## 35.8 37.4 26.8

pie(observed\_counts,labels = paste0(names(observed\_counts), ": ", observed\_percentages, "%"),  
 main = "Observed Voting Distribution",col = c("blue", "red", "green"))



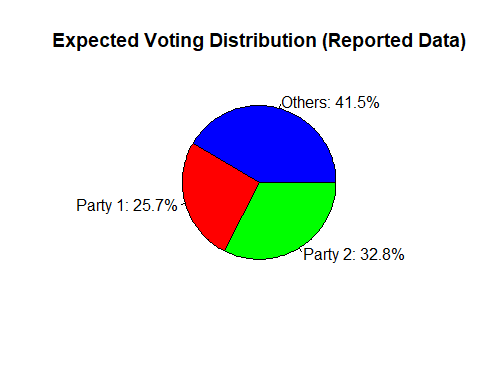
# Pie chart for reported percentages  
pie(reported\_counts,labels = paste0(names(observed\_counts), ": ", reported\_percentages, "%"),  
 main = "Expected Voting Distribution (Reported Data)",col = c("blue", "red", "green"))

H0: The reported voting percentages in a general election match the actual voter responses in the sample.

H1: The reported voting percentages in a general election do not match the actual voter responses in the sample.

Inference:

* p-value = 0.01192
* Since p-value < 0.05, we reject H0 and conclude that the reporting percentages in a general election DO NOT MATCH the actual voter responses in the sample



H0: The reported voting percentages in a general election match the actual voter responses in the sample.

H1: The reported voting percentages in a general election do not match the actual voter responses in the sample.

Inference:

* p-value = 0.01192
* Since p-value < 0.05, we reject H0 and conclude that the reporting percentages in a general election DO NOT MATCH the actual voter responses in the sample

##OBJECTIVE 2  
#install.packages('titanic')  
library(titanic)  
data("titanic\_train")  
titanic\_data = titanic\_train  
head(titanic\_data)

## PassengerId Survived Pclass  
## 1 1 0 3  
## 2 2 1 1  
## 3 3 1 3  
## 4 4 1 1  
## 5 5 0 3  
## 6 6 0 3  
## Name Sex Age SibSp Parch  
## 1 Braund, Mr. Owen Harris male 22 1 0  
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female 38 1 0  
## 3 Heikkinen, Miss. Laina female 26 0 0  
## 4 Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35 1 0  
## 5 Allen, Mr. William Henry male 35 0 0  
## 6 Moran, Mr. James male NA 0 0  
## Ticket Fare Cabin Embarked  
## 1 A/5 21171 7.2500 S  
## 2 PC 17599 71.2833 C85 C  
## 3 STON/O2. 3101282 7.9250 S  
## 4 113803 53.1000 C123 S  
## 5 373450 8.0500 S  
## 6 330877 8.4583 Q

# Observed counts: Passenger class (Pclass)  
observed\_counts = table(titanic\_data$Pclass)  
# Expected proportions (e.g., assume equal distribution across classes)  
expected\_proportions = rep(1/3, 3) # Hypothesis: Equal distribution among 3 classes  
expected\_counts = sum(observed\_counts) \* expected\_proportions  
# Perform Chi-Square Goodness-of-Fit Test  
chi\_square\_test = chisq.test(x = observed\_counts,p = expected\_proportions,rescale.p = TRUE)  
# Display results  
print("Observed Counts:")

## [1] "Observed Counts:"

print(observed\_counts)

##   
## 1 2 3   
## 216 184 491

print("Expected Counts:")

## [1] "Expected Counts:"

print(expected\_counts)

## [1] 297 297 297

print(chi\_square\_test)

##   
## Chi-squared test for given probabilities  
##   
## data: observed\_counts  
## X-squared = 191.8, df = 2, p-value < 2.2e-16

H0: Passenger class (Pclass) are equally distributed in the Titanic dataset

H1: Passenger class (Pclass) are not equally distributed in the Titanic dataset

Inference:

* Observed Counts: First class = 216, Second class = 184, Third class = 491.
* Expected Counts: 297 for all classes.
* Chi-Square Test: p-value < 2.2e-16
* Since p-value < 0.05, we reject H0 and confirm that the passenger classes are not equal among all the classes in the Titanic dataset.

##OBJECTIVE 3  
data=read.csv("C:\\Users\\abhij\\Downloads\\supermarket\_sales - Sheet1.csv")  
head(data)

## Invoice.ID Branch City Customer.type Gender Product.line  
## 1 750-67-8428 A Yangon Member Female Health and beauty  
## 2 226-31-3081 C Naypyitaw Normal Female Electronic accessories  
## 3 631-41-3108 A Yangon Normal Male Home and lifestyle  
## 4 123-19-1176 A Yangon Member Male Health and beauty  
## 5 373-73-7910 A Yangon Normal Male Sports and travel  
## 6 699-14-3026 C Naypyitaw Normal Male Electronic accessories  
## Unit.price Quantity Tax.5. Total Date Time Payment cogs  
## 1 74.69 7 26.1415 548.9715 1/5/2019 13:08 Ewallet 522.83  
## 2 15.28 5 3.8200 80.2200 3/8/2019 10:29 Cash 76.40  
## 3 46.33 7 16.2155 340.5255 3/3/2019 13:23 Credit card 324.31  
## 4 58.22 8 23.2880 489.0480 1/27/2019 20:33 Ewallet 465.76  
## 5 86.31 7 30.2085 634.3785 2/8/2019 10:37 Ewallet 604.17  
## 6 85.39 7 29.8865 627.6165 3/25/2019 18:30 Ewallet 597.73  
## gross.margin.percentage gross.income Rating  
## 1 4.761905 26.1415 9.1  
## 2 4.761905 3.8200 9.6  
## 3 4.761905 16.2155 7.4  
## 4 4.761905 23.2880 8.4  
## 5 4.761905 30.2085 5.3  
## 6 4.761905 29.8865 4.1

# Select categorical variables for testing independence. Example: City and Customer  
city\_cust\_table=table(data$City,data$Customer);city\_cust\_table

##   
## Member Normal  
## Mandalay 165 167  
## Naypyitaw 169 159  
## Yangon 167 173

# Perform Chi-Square Test for Independence  
chi\_square\_test=chisq.test(city\_cust\_table)  
# Display the test results  
print("Chi-Square Test Results:")

## [1] "Chi-Square Test Results:"

print(chi\_square\_test)

##   
## Pearson's Chi-squared test  
##   
## data: city\_cust\_table  
## X-squared = 0.41881, df = 2, p-value = 0.8111

print("Expected Counts:")

## [1] "Expected Counts:"

print(chi\_square\_test$expected)

##   
## Member Normal  
## Mandalay 166.332 165.668  
## Naypyitaw 164.328 163.672  
## Yangon 170.340 169.660

H0: City & Customer type are independent

H1: City & Customer type are not independent

Inference:

* Observed Counts: A group of numbers on a white background

  Description automatically generated
* Expected Counts:A number and a number of numbers

  Description automatically generated with medium confidence
* Chi-Square test: p-value = 0.8111
* Since p-value < 0.05, we reject H0 and claim that the City & Customer type are not independent.