Titanic

2023-08-05

# Load the required package  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

titanic\_data<- read.csv("C:/Users/fredy/Downloads/Lab3-Data.csv")  
# View the structure of the dataset  
str(titanic\_data)

## 'data.frame': 1309 obs. of 5 variables:  
## $ name : chr "Allen, Miss. Elisabeth Walton" "Allison, Master. Hudson Trevor" "Allison, Miss. Helen Loraine" "Allison, Mr. Hudson Joshua Creighton" ...  
## $ pclass : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ survived: int 1 1 0 0 0 1 1 0 1 0 ...  
## $ sex : chr "female" "male" "female" "male" ...  
## $ age : chr "29" "0.9167" "2" "30" ...

# View the first few rows of the dataset  
head(titanic\_data)

## name pclass survived sex age  
## 1 Allen, Miss. Elisabeth Walton 1 1 female 29  
## 2 Allison, Master. Hudson Trevor 1 1 male 0.9167  
## 3 Allison, Miss. Helen Loraine 1 0 female 2  
## 4 Allison, Mr. Hudson Joshua Creighton 1 0 male 30  
## 5 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) 1 0 female 25  
## 6 Anderson, Mr. Harry 1 1 male 48

# Question 1a: How many cases are there?  
  
n\_cases <- nrow(titanic\_data)  
print(paste("Number of cases in the dataset:", n\_cases))

## [1] "Number of cases in the dataset: 1309"

# Question 1a: What is the identifier variable?  
identifier\_var <- titanic\_data$name  
print("name")

## [1] "name"

# Question 1a: What is/are the categorical and numerical variable(s) in the data?  
categorical\_vars <- c("PCLASS", "SURVIVED", "GENDER")  
numerical\_var <- "AGE"  
print(paste("The categorical variables are:", categorical\_vars))

## [1] "The categorical variables are: PCLASS"   
## [2] "The categorical variables are: SURVIVED"  
## [3] "The categorical variables are: GENDER"

print(paste("The numerical variable is:", numerical\_var))

## [1] "The numerical variable is: AGE"

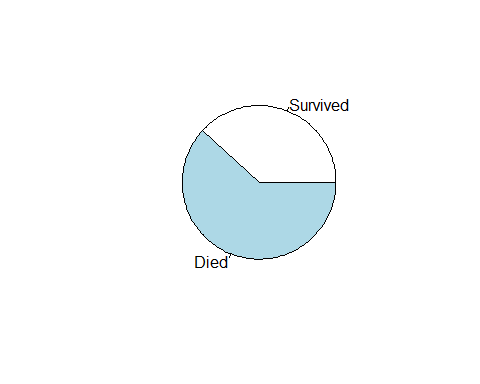
#(b) Is this an observational study or an experiment? Can the results of the study be extended to the population of   
#interest which is all ships colliding with an iceberg? Are causal inferences possible?  
  
#Question 2  
library(ggplot2)  
  
# Question 2(a): What percentage of the passengers survived the disaster?  
# Calculate the proportion of survivors  
# Task 2(a): Calculate the percentage of passengers who survived the disaster and create a pie chart  
total\_passengers <- nrow(titanic\_data)  
survival\_table <- table(titanic\_data$survived)  
survived\_passengers <- survival\_table[2]  
survived\_passengers

## 1   
## 500

percentage\_survived <- (survived\_passengers / total\_passengers) \* 100  
percentage\_survived

## 1   
## 38.1971

# Generate a pie chart  
Died\_passengers <- survival\_table[1]  
labels <- c("Survived","Died")  
piedata <- c(survived\_passengers, Died\_passengers)  
pie(piedata,labels)



# Task 2(b): Hypothesis test  
# Null hypothesis: The majority of passengers will not survive a similar future disaster.  
# Alternative hypothesis: The majority of passengers will survive a similar future disaster.  
  
# Perform the hypothesis test  
alpha <- 0.01  
p\_hat <- percentage\_survived / 100  
z\_test\_stat <- (p\_hat - 0.5) / sqrt((p\_hat \* (1 - p\_hat)) / total\_passengers)  
p\_value <- 2 \* (1 - pnorm(abs(z\_test\_stat)))  
  
# Print the output  
print("Hypothesis Test:")

## [1] "Hypothesis Test:"

print(paste("Z-test statistic:", z\_test\_stat))

## [1] "Z-test statistic: -8.78899149198757"

print(paste("Distribution of the test statistic under the null hypothesis:", "Standard Normal (Z) distribution"))

## [1] "Distribution of the test statistic under the null hypothesis: Standard Normal (Z) distribution"

print(paste("p-value:", p\_value))

## [1] "p-value: 0"

# State your conclusion based on the p-value and alpha  
if (p\_value <= alpha) {  
 print("Reject the null hypothesis. There is evidence to suggest that the majority of passengers will survive a similar future disaster.")  
} else {  
 print("Fail to reject the null hypothesis. There is no significant evidence to suggest that the majority of passengers will not survive a similar future disaster.")  
}

## [1] "Reject the null hypothesis. There is evidence to suggest that the majority of passengers will survive a similar future disaster."

# Task 2(c): Calculate the 98% two-sided confidence interval for the population proportion of all non-survivors  
z\_critical <- qnorm(1 - alpha/2)  
margin\_of\_error <- z\_critical \* sqrt((p\_hat \* (1 - p\_hat)) / total\_passengers)  
confidence\_interval <- c(p\_hat - margin\_of\_error, p\_hat + margin\_of\_error)  
  
# Print the confidence interval  
print("Confidence Interval:")

## [1] "Confidence Interval:"

print(confidence\_interval)

## 1 1   
## 0.3473797 0.4165623

# Interpret the confidence interval  
print("We are 98% confident that the true proportion of all non-survivors lies between:" )

## [1] "We are 98% confident that the true proportion of all non-survivors lies between:"

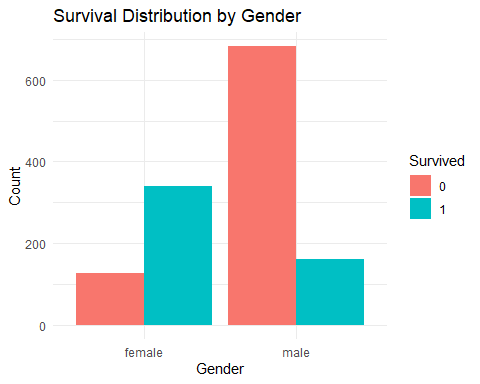
confidence\_interval

## 1 1   
## 0.3473797 0.4165623

# Compare the result from the confidence interval with the conclusion from part (b)  
if (confidence\_interval[1] <= 0.5 && confidence\_interval[2] >= 0.5) {  
 print("The confidence interval includes the proportion 0.5, which means we cannot conclude with 98% confidence that the majority of passengers will not survive a similar future disaster.")  
} else {  
 print("The confidence interval does not include the proportion 0.5, which means we can conclude with 98% confidence that the majority of passengers will not survive a similar future disaster.")  
}

## [1] "The confidence interval does not include the proportion 0.5, which means we can conclude with 98% confidence that the majority of passengers will not survive a similar future disaster."

########Question 3  
# Create a side-by-side bar chart  
ggplot(titanic\_data, aes(x = sex, fill = factor(survived))) +  
 geom\_bar(position = "dodge") +  
 labs(title = "Survival Distribution by Gender",  
 x = "Gender",  
 y = "Count",  
 fill = "Survived") +  
 theme\_minimal()



#(b) Carry out an appropriate two-sample proportion z-test at α = 0.01 to see whether there is a difference in the   
#non-survival rate for females and males. Carry out the appropriate two-sample proportion z-test. Paste the   
#output into your report. State the null and alternative hypotheses in terms of parameters. Report the value of   
#the appropriate test statistic, the distribution of the test statistic under the null hypothesis, and the p-value of the   
#test to answer the question. State your conclusion.   
  
# Load necessary library  
library(stats)  
  
# Subset the data for females and males  
female\_data <- subset(titanic\_data, sex == "female")  
male\_data <- subset(titanic\_data, sex == "male")  
  
# Calculate the proportion of non-survivors for females and males  
female\_table <- table(female\_data$survived)  
prop\_female\_non\_survived <- female\_table[1]/ nrow(female\_data)  
male\_table <- table(male\_data$survived)  
prop\_male\_non\_survived<- male\_table[1]/ nrow(male\_data)  
  
# Perform the two-sample proportion z-test  
z\_test\_result <- prop.test(x = c(male\_table[1], female\_table[1]),  
 n = c(nrow(male\_data), nrow(female\_data)),  
 correct = FALSE, alternative = "two.sided", conf.level = 0.99)  
# Print the z-test result  
print(z\_test\_result)

##   
## 2-sample test for equality of proportions without continuity correction  
##   
## data: c(male\_table[1], female\_table[1]) out of c(nrow(male\_data), nrow(female\_data))  
## X-squared = 365.89, df = 1, p-value < 2.2e-16  
## alternative hypothesis: two.sided  
## 99 percent confidence interval:  
## 0.4729311 0.6000354  
## sample estimates:  
## prop 1 prop 2   
## 0.8090154 0.2725322

#(c) Find a 99% two-sided confidence interval for the difference in the population proportion of all non-survivors between females and males:  
se\_diff<- sqrt((prop\_male\_non\_survived \* (1 - prop\_male\_non\_survived)) / nrow(male\_data) + (prop\_female\_non\_survived \* (1 - prop\_female\_non\_survived)/nrow(female\_data)))  
# Calculate the difference in proportions  
prop\_diff <- prop\_male\_non\_survived - prop\_female\_non\_survived  
# Calculate the margin of error for the confidence interval  
moe <- qnorm(0.995) \* se\_diff  
# Calculate the lower and upper bounds of the confidence interval  
ci\_lower <- prop\_diff - moe  
ci\_upper <- prop\_diff + moe  
  
# Print the confidence interval  
cat("99% Confidence Interval for the Difference in Proportions:", ci\_lower, "to", ci\_upper, "\n")

## 99% Confidence Interval for the Difference in Proportions: 0.4729311 to 0.6000354

#########Question 4  
cont\_table<- table(titanic\_data$survived, titanic\_data$pclass)  
cont\_table

##   
## 1 2 3  
## 0 123 158 528  
## 1 200 119 181

# (i) What percent of all passengers were third-class passengers and survived?  
percent\_third\_class\_survived <- (cont\_table[2, 3] / sum(cont\_table[2, ])) \* 100  
percent\_third\_class\_survived

## [1] 36.2

# (ii) What percent of the survivors were in third class?  
percent\_survivors\_in\_third\_class <- (cont\_table[2, 3] / sum(cont\_table[, 3])) \* 100  
percent\_survivors\_in\_third\_class

## [1] 25.52891

# (iii) What percent of the third-class passengers survived?  
percent\_third\_class\_passengers\_survived <- (cont\_table[2, 3] / sum(cont\_table[, 3])) \* 100  
percent\_third\_class\_passengers\_survived

## [1] 25.52891

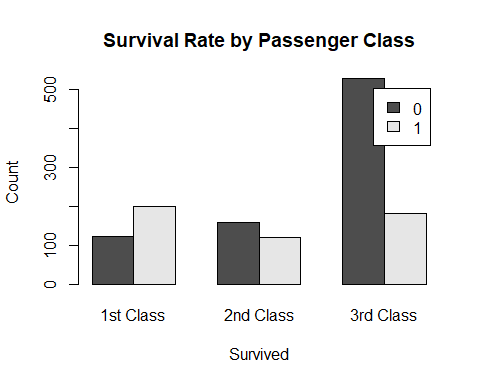
# (iv) In which of the three passenger classes were passengers most likely to have survived?  
most\_likely\_survived\_class <- which.max(cont\_table[2,])  
most\_likely\_survived\_class

## 1   
## 1

# (v) In which of the three passenger classes were the passengers least likely to have survived?  
least\_likely\_survived\_class <- which.min(cont\_table[2,])  
least\_likely\_survived\_class

## 2   
## 2

# (b) Side-by-side bar chart showing the survival rate for each passenger class  
barplot(cont\_table, beside = TRUE, legend = TRUE,   
 names.arg = c("1st Class", "2nd Class", "3rd Class"),   
 main = "Survival Rate by Passenger Class",   
 xlab = "Survived", ylab = "Count")



# (c) Hypothesis test to examine the relationship between survival and passenger class  
test\_result <- chisq.test(cont\_table)  
test\_statistic <- test\_result$statistic  
p\_value <- test\_result$p.value  
p\_value

## [1] 1.720826e-28

# (d) Hypothesis test to see whether the distributions of survival for each level of passenger class are different  
# Null Hypothesis: The distributions of survival for each level of passenger class are the same.  
# Alternative Hypothesis: The distributions of survival for each level of passenger class are different.  
# Output the results  
cat("Percent of all passengers who were third-class and survived:", percent\_third\_class\_survived, "%\n")

## Percent of all passengers who were third-class and survived: 36.2 %

cat("Percent of survivors who were in third class:", percent\_survivors\_in\_third\_class, "%\n")

## Percent of survivors who were in third class: 25.52891 %

cat("Percent of third-class passengers who survived:", percent\_third\_class\_passengers\_survived, "%\n")

## Percent of third-class passengers who survived: 25.52891 %

cat("Passenger class most likely to have survived:", most\_likely\_survived\_class, "\n")

## Passenger class most likely to have survived: 1

cat("Passenger class least likely to have survived:", least\_likely\_survived\_class, "\n")

## Passenger class least likely to have survived: 2

cat("\nChi-squared test result:\n")

##   
## Chi-squared test result:

print(test\_result)

##   
## Pearson's Chi-squared test  
##   
## data: cont\_table  
## X-squared = 127.86, df = 2, p-value < 2.2e-16

cat("Test Statistic:", test\_statistic, "\n")

## Test Statistic: 127.8592

cat("P-value:", p\_value, "\n")

## P-value: 1.720826e-28

##### Question 5  
# Load the required library  
library(dplyr)  
# filter the data for children under 12  
# Assuming your dataset is named 'titanic\_data', filter the data for children under 12  
child\_data <- titanic\_data %>% filter(age < 12)  
child\_data

## name pclass survived sex age  
## 1 Allison, Master. Hudson Trevor 1 1 male 0.9167  
## 2 Carter, Master. William Thornton II 1 1 male 11  
## 3 Becker, Master. Richard F 2 1 male 1  
## 4 Caldwell, Master. Alden Gates 2 1 male 0.8333  
## 5 Hamalainen, Master. Viljo 2 1 male 0.6667  
## 6 Laroche, Miss. Louise 2 1 female 1  
## 7 Mallet, Master. Andre 2 1 male 1  
## 8 Richards, Master. George Sibley 2 1 male 0.8333  
## 9 West, Miss. Barbara J 2 1 female 0.9167  
## 10 Aks, Master. Philip Frank 3 1 male 0.8333  
## 11 Andersson, Miss. Sigrid Elisabeth 3 0 female 11  
## 12 Baclini, Miss. Eugenie 3 1 female 0.75  
## 13 Baclini, Miss. Helene Barbara 3 1 female 0.75  
## 14 Danbom, Master. Gilbert Sigvard Emanuel 3 0 male 0.3333  
## 15 Dean, Master. Bertram Vere 3 1 male 1  
## 16 Dean, Miss. Elizabeth Gladys Millvina 3 1 female 0.1667  
## 17 Goodwin, Master. Sidney Leonard 3 0 male 1  
## 18 Goodwin, Master. William Frederick 3 0 male 11  
## 19 Goodwin, Miss. Jessie Allis 3 0 female 10  
## 20 Hassan, Mr. Houssein G N 3 0 male 11  
## 21 Johnson, Miss. Eleanor Ileen 3 1 female 1  
## 22 Klasen, Miss. Gertrud Emilia 3 0 female 1  
## 23 Nakid, Miss. Maria (Mary) 3 1 female 1  
## 24 Panula, Master. Eino Viljami 3 0 male 1  
## 25 Peacock, Master. Alfred Edward 3 0 male 0.75  
## 26 Rice, Master. Albert 3 0 male 10  
## 27 Sandstrom, Miss. Beatrice Irene 3 1 female 1  
## 28 Skoog, Master. Karl Thorsten 3 0 male 10  
## 29 Thomas, Master. Assad Alexander 3 1 male 0.4167  
## 30 van Billiard, Master. Walter John 3 0 male 11.5  
## 31 Van Impe, Miss. Catharina 3 0 female 10

# Create a contingency table for survival and passenger class for children  
contingency\_table <- table(child\_data$survived, child\_data$pclass)  
contingency\_table

##   
## 1 2 3  
## 0 0 0 13  
## 1 2 7 9

Survival\_rate\_for\_firstclass= contingency\_table[2,1]/sum(contingency\_table[,1])   
Survival\_rate\_for\_firstclass

## [1] 1

Survival\_rate\_for\_secondtclass <- contingency\_table[2,2] / sum(contingency\_table[,2])   
Survival\_rate\_for\_secondtclass

## [1] 1

Survival\_rate\_for\_thirdclass <- contingency\_table[2,3] / sum(contingency\_table[,3])   
Survival\_rate\_for\_thirdclass

## [1] 0.4090909

#(b) The emergency protocol “women and children first” was followed during evacuation to ensure that women and   
#children were the first to board lifeboats. Was the rule effective in saving the lives of women and children   
#aboard the Titanic? Using α = 0.01, carry out the appropriate test to answer the question. State the null and   
#alternative hypotheses. Paste the output into your report. Report the value of the appropriate test statistic, the   
#distribution of the test statistic under the null hypothesis, and the p-value of the test to answer the question.   
#State your conclusion. (Ignore men with no age listed.)  
# For this test, we'll exclude men with no age listed from the analysis.  
# Step 2: Create a new variable "WOMEN\_CHILD" to identify women and children passengers  
titanic\_data$WOMEN\_CHILD <- ifelse(titanic\_data$sex == "female" | titanic\_data$age < 12, 1, 0)  
# Step 3: Create a 2x2 contingency table for survival and "WOMEN\_CHILD" variable  
table\_survival\_women\_children <- table(titanic\_data$survived, titanic\_data$WOMEN\_CHILD)  
# Step 4: Perform the chi-square test for independence  
chi\_square\_test <- chisq.test(table\_survival\_women\_children)  
# Step 5: Print the results  
print(chi\_square\_test)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: table\_survival\_women\_children  
## X-squared = 308.26, df = 1, p-value < 2.2e-16