Single Proportion: Exercise 1 and 2 Assignment

Garridos, Charlene P.

2024-02-16

Contents

1 R Library

```
library(ggplot2)
library(forcats)
```

2 Exercise 1

2.1 Problem

A local school board claims that at least 75% of the parents in the district are in favor of extending the school day by 30 minutes to incorporate additional physical education time. A group of parents skeptical of this claim decides to conduct a survey. They randomly select 120 parents from the district, and 81 express support for extending the school day.

Tasks:

1. Compute the 99% CI using
    a. Score statistics
    b. Wald statistics
    c. Compare the two intervals
2. Test the claim using
    a. Wald
    b. Score
    c. Likelihood Ratio

```
# Given Data
n <- 120
p.sample <- 0.675
pi.null <- 0.75
```

2.2 Computation of 99% CI (Score and Wald Statistics)

2.2.1 Score Statistics

Since the confidence interval is 99%, the $\alpha = 0.01$

```
# let's create a function that compute the score z statistics
# score confidence interval, and test the H0:pi=pi.0
score.test.prop <- function(n,p,pi.0=0.75,alpha=0.01,confidence=T,
                      alternative = c("lesser","greater", "not equal")){

  # compute statistic
  dif <- p-pi.0
  se = sqrt(pi.0*(1-pi.0)/n)
  test.stat <- (dif)/(se)
```

```
  # compute confidence interval
  if(confidence){
    z_ <- abs(qnorm(alpha/2))
    lower.limit <- max(c(0,p - (z_*se)))
    upper.limit <- min(c(p + (z_*se),1))
    confidence.interval <- c(lower.limit,upper.limit)
    names(confidence.interval) <- c("lower limit", "upper limit")
  }

  # test HO
  if(alternative =="not equal"){
    p.value <- (1- pnorm(abs(test.stat)))*2
  }
  else if(alternative =="lesser"){
    p.value <- pnorm(test.stat)
  }
  else if(alternative =="greater"){
    p.value <- pnorm(abs(test.stat))
  }
  else{
    stop("Specify the correct alternative!")
  }

  decision <- ifelse(p.value > alpha,
                     "Do not reject HO",
                     "Reject HO")

  cat("Testing Population Proportion Using Score Test\n\n")
  cat("Ho:The population proportion is greater than or equal to", pi.0,"\n")
  cat("Score Test Statistic:", test.stat,"\t","p-value:",p.value,"\n")
  cat((1-alpha)*100,"CI:",confidence.interval,"\n" )
  cat("Decision:", decision)

  return(list("Score" = test.stat,"p.value"=p.value,
              "CI"=confidence.interval, "decision"=decision))
}
```

2.2.2 Wald Statistics

```
# let's create a function that compute the wald z statistics
# wald confidence interval, and test the HO:pi=pi.0
wald.test.prop <- function(n,p,pi.0=0.75,alpha=0.01,confidence=T,
                       alternative = c("lesser","greater", "not equal")){

  # compute statistic
  dif <- p-pi.0
  se = sqrt(p*(1-p)/n)
  test.stat <- (dif)/(se)

  # compute confidence interval
  if(confidence){
```

```
    z_ <- abs(qnorm(alpha/2))
    lower.limit <- max(c(0,p - (z_*se)))
    upper.limit <- min(c(p + (z_*se),1))
    confidence.interval <- c(lower.limit,upper.limit)
    names(confidence.interval) <- c("lower limit", "upper limit")
  }

  # test HO
  if(alternative =="not equal"){
    p.value <- (1- pnorm(abs(test.stat)))*2
  }
  else if(alternative =="lesser"){
    p.value <- pnorm(test.stat)
  }
  else if(alternative =="greater"){
    p.value <- pnorm(abs(test.stat))
  }
  else{
    stop("Specify the correct alternative!")
  }

  decision <- ifelse(p.value > alpha,
                     "Do not reject HO",
                     "Reject HO")

  cat("Testing Population Proportion Using Wald Test\n\n")
  cat("Ho:The population proportion is greater than or equal to to", pi.0,"\n")
  cat("Wald Test Statistic:", test.stat,"\t","p-value:",p.value,"\n")
  cat((1-alpha)*100,"CI:",confidence.interval,"\n" )
  cat("Decision:", decision)

  return(list("Wald.stat" = test.stat,"p.value"=p.value,
              "CI"=confidence.interval, "decision"=decision))
}
```

2.2.3 Comparison of Two Intervals

```
wald.res <- wald.test.prop(n=n,p=p.sample,pi.0 = pi.null, alternative = "not equal")
```

```
## Testing Population Proportion Using Wald Test
##
## Ho:The population proportion is greater than or equal to to 0.75
## Wald Test Statistic: -1.754116    p-value: 0.07941063
## 99 CI: 0.5648664 0.7851336
## Decision: Do not reject HO
```

```
score.res <- score.test.prop(n=n,p=p.sample,
                             pi.0 = 0.75, alternative = "not equal")
```

```
## Testing Population Proportion Using Score Test
```

```
##
## Ho:The population proportion is greater than or equal to 0.75
## Score Test Statistic: -1.897367   p-value: 0.05777957
## 99 CI: 0.5731814 0.7768186
## Decision: Do not reject H0
```

Discussion

The confidence interval of Wald Test is

$$CI_{\text{Wald}} : (0.5648664, 0.7851336)$$

The confidence interval of Score Test is

$$CI_{\text{Score}} : (0.5731814, 0.7768186)$$

The intervals are quite close in terms of their lower and upper bounds. The difference between the lower bound and upper bound for both the Wald Test and Score Test is 0.2202672 and 0.2036372, respectively. Hence, both intervals cover a similar range of values for the population proportion and are narrow. If an interval is narrow, it gives a more precise estimate. In other words, the margin of error is smaller. Therefore, the slight difference between the lower and upper bounds of both intervals indicates reduced uncertainty in the estimation.

Moreover, the hypothesized value of 0.75 falls within both confidence intervals. This indicates that there is no significant difference between the hypothesized value and the estimated population proportion, as both intervals include the value of 0.75. Thus, there is no evidence to refute the claim made by the local school board that at least 75% of the parents in the district are in favor of extending the school day by 30 minutes to incorporate additional physical education time.

2.3 Test the Claim

2.3.1 Wald Test

```
wald.res <- wald.test.prop(n=n,p=p.sample,pi.0 = pi.null, alternative = "not equal")
```

```
## Testing Population Proportion Using Wald Test
##
## Ho:The population proportion is greater than or equal to to 0.75
## Wald Test Statistic: -1.754116    p-value: 0.07941063
## 99 CI: 0.5648664 0.7851336
## Decision: Do not reject H0
```

Discussion

The results show that the Wald Test Statistic is -1.754116, with a corresponding p-value of 0.07941063. This indicates that there's approximately a 7.94% chance of observing a test statistic as extreme as -1.754116, assuming that the true population proportion is indeed 0.75.

Since the p-value = 0.07941063 is greater than the level of significance of 0.01, we fail to reject the null hypothesis. Thus, there is no enough evidence to reject the null hypothesis. Therefore, at least 75% of the parents in the district are in favor of extending the school day by 30 minutes to incorporate additional physical education time.

2.3.2 Score Test

```
score.res <- score.test.prop(n=n,p=p.sample,
                             pi.0 = 0.75, alternative = "not equal")
```

```
## Testing Population Proportion Using Score Test
##
## Ho:The population proportion is greater than or equal to 0.75
## Score Test Statistic: -1.897367   p-value: 0.05777957
## 99 CI: 0.5731814 0.7768186
## Decision: Do not reject H0
```

Discussion

The Score Test Statistics is -1.897367 and the p-value is 0.05777957. The p-value suggests that there's about a 5.78% chance of observing a test statistic as extreme as -1.897367, given that the true population proportion is indeed 0.75.

We cannot reject the null hypothesis since the p-value of 0.05777957 is greater than the level of significance of 0.01. Thus, there is insufficient evidence to reject the null hypothesis. In conclusion, at least 75% of the district's parents support extending the school day by 30 minutes to provide for more physical education time.

2.3.3 Likelihood Ratio

```
# let's create a function that compute the Likelihood H0:pi=pi.0
LR.test.prop <- function(n,p,x=NULL,pi.0=0.75,alpha=0.01,
                         alternative = c("lesser","greater", "not equal")){

  if(is.null(x)){
    x <- round(n*p,0)
  }


  # compute statistic
  likelihood.0 <- ((pi.0)^(x))*((1-pi.0)^(n-x))
  likelihood.1 <- ((p)^(x))*((1-p)^(n-x))
  test.stat <- -2*log(likelihood.0/likelihood.1)

  # test H0
  if(alternative =="not equal"){
    alpha <- alpha/2
    p.value1 <- (1- pchisq(test.stat,df=1))
    p.value2 <- pchisq(test.stat,df=1)
    p.value <- min(c(p.value1, p.value2))
      decision <- ifelse(p.value > alpha,
                         "Do not reject H0",
                         "Reject H0")
  }
  else if(alternative =="lesser"){
    p.value <- pchisq(test.stat,df=1)
```

```
    decision <- ifelse(p.value > alpha,
                       "Do not reject H0",
                       "Reject H0")
  }
  else if(alternative =="greater"){
    p.value <- 1- pchisq(test.stat,df=1)
    decision <- ifelse(p.value > alpha,
                       "Do not reject H0",
                       "Reject H0")
  }
  else{
    stop("Specify the correct alternative!")
  }



  cat("Testing Population Proportion Using Likelihood Ratio Test\n\n")
  cat("Ho:The population proportion is greater than or equal to", pi.0,"\n")
  cat("Likelihood Ratio Test Statistic:", test.stat,"\t","p-value:",p.value,"\n")
  cat("P-value is compare to", alpha,"\n")
  cat("Decision:", decision)

  return(list("LR.statistics" = test.stat,"p.value"=p.value, "decision"=decision))
}
```

```
LR.test <- LR.test.prop(n=n,p=p.sample,pi.0 = pi.null, alternative = "not equal")
```

```
## Testing Population Proportion Using Likelihood Ratio Test
##
## Ho:The population proportion is greater than or equal to 0.75
## Likelihood Ratio Test Statistic: 3.396009     p-value: 0.06535436
## P-value is compare to 0.005
## Decision: Do not reject H0
```

Discussion

The Likelihood Ratio Test Statistics is 3.396009 and the p-value is 0.06535436. Based on the results, we fail to reject the null hypothesis because the p-value is greater than the significance level of 0.01. This means that there is not enough evidence to conclude that the population proportion is less than 0.75. Therefore, at least 75% of the district's parents support extending the school day by 30 minutes to provide for more physical education time.

3 Exercise 2

3.1 Problem

Using the `Adult dataset`, do the following

1. Visualize the distribution the income.
2. Estimate the proportion of the population having income at most $50k per year.

a. point estimate
b. 97% CI

3. Test the hypothesis that less than 50% of the population are earning more than $50k per year at 3% level of significance.

3.2 Dataset

3.2.1 Load the Dataset

```
# Specify the file path relative to the working directory
file_path <- "/Users/User/Downloads/adult.csv"

# Read the CSV file
demographic_info <- read.csv(file_path, sep = ",", header=TRUE,
                             col.names = c("age", "workclass", "fnlwgt",
                                           "education", "educational-num",
                                           "marital-status", "occupation",
                                           "relationship", "race", "sex",
                                           "capital-gain", "capital-loss",
                                           "hours-per-week", "native-country","income"))

# View the data
head(demographic_info)
```

```
##   age  workclass fnlwgt     education educational.num      marital.status
## 1  25    Private 226802          11th              7       Never-married
## 2  38    Private  89814       HS-grad              9  Married-civ-spouse
## 3  28  Local-gov 336951    Assoc-acdm             12  Married-civ-spouse
## 4  44    Private 160323  Some-college             10  Married-civ-spouse
## 5  18          ? 103497  Some-college             10       Never-married
## 6  34    Private 198693          10th              6       Never-married
##           occupation   relationship   race     sex capital.gain capital.loss
## 1  Machine-op-inspct      Own-child  Black    Male            0            0
## 2    Farming-fishing        Husband  White    Male            0            0
## 3    Protective-serv        Husband  White    Male            0            0
## 4  Machine-op-inspct        Husband  Black    Male         7688            0
## 5                  ?      Own-child  White  Female            0            0
## 6      Other-service  Not-in-family  White    Male            0            0
##   hours.per.week native.country  income
## 1             40  United-States  <=50K.
## 2             50  United-States  <=50K.
## 3             40  United-States   >50K.
## 4             40  United-States   >50K.
## 5             30  United-States  <=50K.
## 6             30  United-States  <=50K.
```

3.2.2 Data Preparation

```
# str() function help inspect the structure and levels of the 'income' column
str(demographic_info$income)
```

```
##  chr [1:16281] " <=50K." " <=50K." " >50K." " >50K." " <=50K." " <=50K." ...
```

```
# Convert 'income' column form the dataset to a factor.
income.factor <- factor(demographic_info$income)
```

```
# str() function help inspect the structure and levels of the converted 'income' column.
str(income.factor)
```

```
##  Factor w/ 2 levels " <=50K.," >50K.": 1 1 2 2 1 1 1 2 1 1 ...
```

3.3 Visualization

```
# Display the distribution of levels in 'income.factor' by frequency.
knitr::kable(table(income.factor))
```

income.factor	Freq
<=50K.	12435
>50K.	3846

```r
# Visualize the distribution of the income using bar plot
ggplot(data.frame(income = income.factor), aes(x = income)) +
  geom_bar( fill = c("darkolivegreen2",  "darkolivegreen4")) +
  labs(title = "Frequency Distribution of Income", x = "Income", y = "Frequency")
```

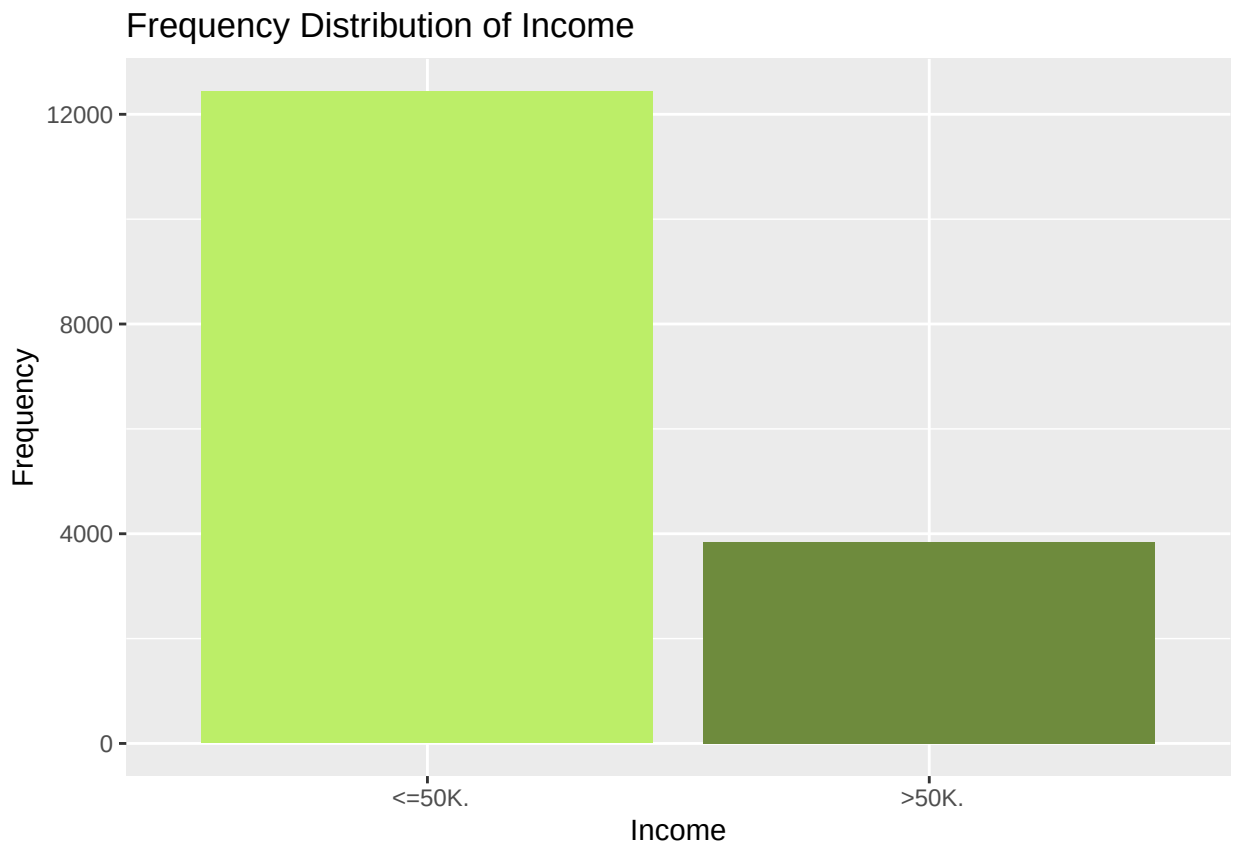

Figure 1: The Distribution of Income By Frequency

Interpretation

Figure 1 depicts the frequency distribution of the income of adults earning at most 50,000 dollars and greater than 50,000 dollars. The dataset comprises 12,435 individuals whose income is less than or equal to 50,000 dollars. Additionally, there are 3,846 individuals whose income exceeds 50,000 dollars.

3.4 Proportion Estimation

Estimate the proportion of the population having income at most $50k per year.

3.4.1 Point Estimate

```r
# table() function to calculate frequencies of income
income.freq <- table(income.factor)
```

```r
# prop.table() function to calculate relative frequencies
income.relativefreq <- prop.table(income.freq)

# A data frame for the tables
income <- data.frame(income.freq = names(income.freq), Frequency = as.vector(income.freq),
                     Relative.Frequency = as.vector(income.relativefreq))
income
```

```
##   income.freq Frequency Relative.Frequency
## 1      <=50K.     12435          0.7637737
## 2       >50K.      3846          0.2362263
```

Interpretation

The table displays the frequency and the relative frequency of adults who earn $\leq 50K$ and $> 50K$. For incomes at most 50000 dollars, there are 12,435 occurrences which is approximately 76.38% of the total. For incomes greater than 50000 dollars, there are 3,846 occurrences which approximately 23.62% of the total. Majority of the population earn $> 50K$.

Using the results above, we can now proceed to extracting proportion of population with income $\leq 50k$ per year.

```r
# A dataframe extracting the proportion of population with income at most $50k per year
income <- data.frame(income.factor = c("<=50K", ">50K"),
                     relative.frequency = c(0.7637737, 0.2362263))

# Calculate the proportion of the population with income at most $50k per year
income.atmost50k <- income$relative.frequency[income$income.factor == "<=50K"]

# Print the proportion of the population with income at most $50k per year
cat("The proportion of population with income at most $50k per year:", income.atmost50k, "\n")
```

```
## The proportion of population with income at most $50k per year: 0.7637737
```

3.4.2 97% CI

```r
# Calculate the total frequency
total.freq <- sum(income.freq)
total.freq
```

```
## [1] 16281
```

Taking the obtained results as shown in the 'income' table

```r
# total frequency
n <- 16281
# Frequency of population earning at most $50K each year
x <- 12435
# proportion of population of adults earning at most $50K each year (x/n)
p.sample <- 0.7638
# null hypothesis
pi.null <- 0.5
```

3.4.2.1 Wald Test

Since the confidence interval is 97%, the $\alpha = 0.03$

```
# let's create a function that compute the wald z statistics
# wald confidence interval, and test the H0:pi=pi.0
wald.test.prop <- function(n,p,pi.0=0.5,alpha=0.03,confidence=T,
                      alternative = c("lesser","greater", "not equal")){

  # compute statistic
  dif <- p-pi.0
  se = sqrt(p*(1-p)/n)
  test.stat <- (dif)/(se)

  # compute confidence interval
  if(confidence){
    z_ <- abs(qnorm(alpha/2))
    lower.limit <- max(c(0,p - (z_*se)))
    upper.limit <- min(c(p + (z_*se),1))
    confidence.interval <- c(lower.limit,upper.limit)
    names(confidence.interval) <- c("lower limit", "upper limit")
  }

  # test H0
  if(alternative =="not equal"){
    p.value <- (1- pnorm(abs(test.stat)))*2
  }
  else if(alternative =="lesser"){
    p.value <- pnorm(test.stat)
  }
  else if(alternative =="greater"){
    p.value <- pnorm(abs(test.stat))
  }
  else{
    stop("Specify the correct alternative!")
  }

  decision <- ifelse(p.value > alpha,
                     "Do not reject H0",
                     "Reject H0")

  cat("Testing Population Proportion Using Wald Test\n\n")
  cat("Ho:The population proportion is less than or equal to", pi.0,"\n")
  cat("Wald Test Statistic:", test.stat,"\t","p-value:",p.value,"\n")
  cat((1-alpha)*100,"CI:",confidence.interval,"\n" )
  cat("Decision:", decision)

  return(list("Wald.stat" = test.stat,"p.value"=p.value,
              "CI"=confidence.interval, "decision"=decision))
}
```

```
wald.res <- wald.test.prop(n=n,p=p.sample,pi.0 = pi.null, alternative = "not equal")
```

```
## Testing Population Proportion Using Wald Test
##
```

```
## Ho:The population proportion is less than or equal to 0.5
## Wald Test Statistic: 79.2475      p-value: 0
## 97 CI: 0.7565762 0.7710238
## Decision: Reject HO
```

Discussion

The results of the Wald Test indicate that the test statistic is 79.2475, with a corresponding p-value of 0. This extremely low p-value suggests strong evidence against the null hypothesis.

Since the p-value = 0 is less than the level of significance of 0.03, we reject the null hypothesis. Thus, there is sufficient evidence to conclude that the population proportion of individuals earning more than $50k per year is greater than 0.5.

3.4.3 Hypothesis Testing

Test the hypothesis that less than 50% of the population are earning more than $50k per year at 3% level of significance.

```r
# let's create a function that compute the score z statistics
# score confidence interval, and test the HO:pi=pi.O
score.test.prop <- function(n,p,pi.0=0.5,alpha=0.03,confidence=T,
                      alternative = c("lesser","greater", "not equal")){

  # compute statistic
  dif <- p-pi.0
  se = sqrt(pi.0*(1-pi.0)/n)
  test.stat <- (dif)/(se)

  # compute confidence interval
  if(confidence){
    z_ <- abs(qnorm(alpha/2))
    lower.limit <- max(c(0,p - (z_*se)))
    upper.limit <- min(c(p + (z_*se),1))
    confidence.interval <- c(lower.limit,upper.limit)
    names(confidence.interval) <- c("lower limit", "upper limit")
  }

  # test HO
  if(alternative =="not equal"){
    p.value <- (1- pnorm(abs(test.stat)))*2
  }
  else if(alternative =="lesser"){
    p.value <- pnorm(test.stat)
  }
  else if(alternative =="greater"){
    p.value <- pnorm(abs(test.stat))
  }
  else{
    stop("Specify the correct alternative!")
  }
```

```
  decision <- ifelse(p.value > alpha,
                     "Do not reject H0",
                     "Reject H0")

  cat("Testing Population Proportion Using Score Test\n\n")
  cat("Ho:The population proportion is less than", pi.0,"\n")
  cat("Score Test Statistic:", test.stat,"\t","p-value:",p.value,"\n")
  cat((1-alpha)*100,"CI:",confidence.interval,"\n" )
  cat("Decision:", decision)

  return(list("Score" = test.stat,"p.value"=p.value,
              "CI"=confidence.interval, "decision"=decision))
}
```

```
score.res <- score.test.prop(n=n,p=p.sample,
                             pi.0 = 0.5, alternative = "greater")
```

3.4.3.1 Score Inference

```
## Testing Population Proportion Using Score Test
##
## Ho:The population proportion is less than 0.5
## Score Test Statistic: 67.32019    p-value: 1
## 97 CI: 0.7552963 0.7723037
## Decision: Do not reject H0
```

Discussion

The Score Test Statistics is 67.32019 and the p-value is 1. It suggests that the observed data is consistent with the null hypothesis that less than 50% of the population earns more than $50,000 per year. Since the p-value = 1 is greater than the level of significance of 0.03, we cannot reject the null hypothesis. This suggests that there is insufficient evidence to conclude that less than 50% of the population earns more than 50000 dollars per year. Hence, we can conclude that less than 50% of the population earns more than 50000 dollars per year.

4 Reference

- Estimating a Population Mean
- Confidence intervals