Lab05

Siddhartha Sinha

2023-12-21

# AIM: Test of equality of two population proportion

# INTRODUCTION:

The data set contains information about students, including their race, family background, lunch status and academic performances. The focus here is on exploring the gender distribution within a subset of this data. The analysis involves sampling 100 students and assessing the proportion of males within this sample. A hypothesis test is performed to evaluate whether this sample proportion is significantly different from the assumed population proportion.

# DATASET DESCRIPTION:

library(readxl)  
StudentsPerformance <- read\_excel("C:/Users/Siddhartha/Downloads/StudentsPerformance.xls")

This dataset encompasses diverse demographic attributes, including gender, race, parental education level, lunch type, and test preparation course participation. Academic performance is measured through math, reading, and writing scores. This comprehensive collection enables nuanced exploration of educational outcomes and societal factors, fostering insightful analyses in the realm of education research.

# ANALYSIS:

# FIRST SIX OBSERVATIONS:

head(StudentsPerformance)

## # A tibble: 6 × 8  
## gender `race/ethnicity` parental level of educa…¹ lunch test preparation cou…²  
## <chr> <chr> <chr> <chr> <chr>   
## 1 female group B bachelor's degree stan… none   
## 2 female group C some college stan… completed   
## 3 female group B master's degree stan… none   
## 4 male group A associate's degree free… none   
## 5 male group C some college stan… none   
## 6 female group B associate's degree stan… none   
## # ℹ abbreviated names: ¹​`parental level of education`,  
## # ²​`test preparation course`  
## # ℹ 3 more variables: `math score` <dbl>, `reading score` <dbl>,  
## # `writing score` <dbl>

Here we obtained the first six rows from our StudentsPerformance data set.

# LAST SIX OBSERVATIONS:

tail(StudentsPerformance)

## # A tibble: 6 × 8  
## gender `race/ethnicity` parental level of educa…¹ lunch test preparation cou…²  
## <chr> <chr> <chr> <chr> <chr>   
## 1 male group A high school stan… none   
## 2 female group E master's degree stan… completed   
## 3 male group C high school free… none   
## 4 female group C high school free… completed   
## 5 female group D some college stan… completed   
## 6 female group D some college free… none   
## # ℹ abbreviated names: ¹​`parental level of education`,  
## # ²​`test preparation course`  
## # ℹ 3 more variables: `math score` <dbl>, `reading score` <dbl>,  
## # `writing score` <dbl>

Here we obtained the last six rows from our StudentsPerformance data set.

# DIMENSION OF THE DATASET:

dim(StudentsPerformance)

## [1] 1000 8

Here we can see that our data set contains 1000 rows and 8 columns.

# SUMMARY OF THE DATASET:

summary(StudentsPerformance)

## gender race/ethnicity parental level of education  
## Length:1000 Length:1000 Length:1000   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## lunch test preparation course math score reading score   
## Length:1000 Length:1000 Min. : 0.00 Min. : 17.00   
## Class :character Class :character 1st Qu.: 57.00 1st Qu.: 59.00   
## Mode :character Mode :character Median : 66.00 Median : 70.00   
## Mean : 66.09 Mean : 69.17   
## 3rd Qu.: 77.00 3rd Qu.: 79.00   
## Max. :100.00 Max. :100.00   
## writing score   
## Min. : 10.00   
## 1st Qu.: 57.75   
## Median : 69.00   
## Mean : 68.05   
## 3rd Qu.: 79.00   
## Max. :100.00

Here we can see the minimum value, the first quartile, median (second quartile), third quartile , mean and the maximum value of each of the 8 columns of the data set.

# CREATING TWO POPULATIONS:

# POPULATION 1:  
MALE <- subset(StudentsPerformance, gender == "male")  
MALE

## # A tibble: 482 × 8  
## gender `race/ethnicity` parental level of educ…¹ lunch test preparation cou…²  
## <chr> <chr> <chr> <chr> <chr>   
## 1 male group A associate's degree free… none   
## 2 male group C some college stan… none   
## 3 male group B some college free… none   
## 4 male group D high school free… completed   
## 5 male group C associate's degree stan… none   
## 6 male group D associate's degree stan… none   
## 7 male group A some college stan… completed   
## 8 male group C high school stan… none   
## 9 male group C master's degree free… completed   
## 10 male group D high school stan… none   
## # ℹ 472 more rows  
## # ℹ abbreviated names: ¹​`parental level of education`,  
## # ²​`test preparation course`  
## # ℹ 3 more variables: `math score` <dbl>, `reading score` <dbl>,  
## # `writing score` <dbl>

# POPULATION 2:  
FEMALE <- subset(StudentsPerformance, gender == "female")  
FEMALE

## # A tibble: 518 × 8  
## gender `race/ethnicity` parental level of educ…¹ lunch test preparation cou…²  
## <chr> <chr> <chr> <chr> <chr>   
## 1 female group B bachelor's degree stan… none   
## 2 female group C some college stan… completed   
## 3 female group B master's degree stan… none   
## 4 female group B associate's degree stan… none   
## 5 female group B some college stan… completed   
## 6 female group B high school free… none   
## 7 female group B high school stan… none   
## 8 female group A master's degree stan… none   
## 9 female group C some high school stan… none   
## 10 female group B some high school free… none   
## # ℹ 508 more rows  
## # ℹ abbreviated names: ¹​`parental level of education`,  
## # ²​`test preparation course`  
## # ℹ 3 more variables: `math score` <dbl>, `reading score` <dbl>,  
## # `writing score` <dbl>

# Here we created two populations using the “subset” function, namely “MALE” and “FEMALE”.

# IMPORTING THE LIBRARY:

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

Here we install the “dplyr” (data set plier) package for making data manipulation easier.

# TAKING THE SAMPLE:

s = sample(FEMALE$`test preparation course`, 100, replace = FALSE)  
S1 = as.data.frame(s)  
  
ss = sample(MALE$`test preparation course`, 100, replace = FALSE)  
S2 = as.data.frame(ss)  
  
count(S1, s)

## s n  
## 1 completed 44  
## 2 none 56

count(S2, ss)

## ss n  
## 1 completed 28  
## 2 none 72

Here we get 100 samples using the without replacement procedure and using the “count” function on a ‘s’ and ‘ss’ variables, we can see that the count of female in the 100 samples of the target variable “test preparation course” who completed is 44 and rest is none, while the count of male in 100 samples of target variable “test preparation course” who completed is 28 and the rest is none.

# TESTING THE HYPOTHESIS:

Po: proportion of “COMPLETED” Male test preparation course

P1: proportion of “COMPLETED” Female test preparation course

Assume the population proportion: H0: Po = P1 vs H1: Po != P1 where Success = 30 completed test preparation course in Female; Success = 20 completed test preparation course in Male.

x1=c(44,28)  
n1=c(100,100)  
  
prop.test(x1, n1, alternative = "two.sided", conf.level = 0.95)

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: x1 out of n1  
## X-squared = 2.16, df = 1, p-value = 0.1416  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.02921995 0.22921995  
## sample estimates:  
## prop 1 prop 2   
## 0.3 0.2

**CONCLUSION FROM THE HYPOTHESIS:**

The analysis utilized a sample of 100 observations for both male and female students’ activities. The count within each sample revealed that 28 out of 100 male students and 44 out of 100 female students reported completing test preparation course. Upon performing a two-sample test for equality of proportions with a significance level of 0.05, the p-value obtained was 0.1416. This p-value exceeds the significance level, indicating that there isn’t sufficient evidence to reject the null hypothesis. Consequently, we fail to find significant differences in the proportions of male and female students engaging in completing the test preparation course within this sample. Therefore, based on this analysis, we do not have enough evidence to conclude that there’s a significant difference in test preparation course between male and female students in this specific dataset and population.