Lab04

Siddhartha Sinha

2023-12-21

# AIM: Test for proportion of single population using Prop Test and Binom Test

# 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.

# 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(1:nrow(StudentsPerformance), 100, replace = FALSE)  
new = StudentsPerformance[s,]  
count\_val = count(new, gender)  
count\_val

## # A tibble: 2 × 2  
## gender n  
## <chr> <int>  
## 1 female 46  
## 2 male 54

class(new)

## [1] "tbl\_df" "tbl" "data.frame"

Here we get 100 samples using the without replacement procedure and using the “count” function on a “new” variable, we can see that the count of male in the 100 samples in 54, and the rest are females.

# SETTING THE HYPOTHESIS:

Po: Proportion of Sex Male

Assume, the population proportion where, H0: Po = 0.50 vs H1: Po != 0.50

x=54  
n=100  
prop.test(x, n, p = 0.50, alternative = "two.sided", conf.level = 0.95)

##   
## 1-sample proportions test with continuity correction  
##   
## data: x out of n, null probability 0.5  
## X-squared = 0.25, df = 1, p-value = 0.6171  
## alternative hypothesis: true p is not equal to 0.5  
## 95 percent confidence interval:  
## 0.3703535 0.5719775  
## sample estimates:  
## p   
## 0.47

**CONCLUSION FROM THE HYPOTHESIS:**

The test was conducted assuming the null hypothesis (H0) that the proportion of males in the population is 0.51. The alternative hypothesis (H1) suggests that the proportion differs from 0.51.

After conducting the one-sample proportions test, the obtained p-value of 0.6171 indicates that there's insufficient evidence to reject the null hypothesis at a significance level of 0.05. Therefore, based on this sample, it seems that the proportion of males is not significantly different from the assumed population proportion of 0.51.