Lab 05 : Test for 2 population proportion

Nilanjana Dey

2023-12-19

library(readr)

## Warning: package 'readr' was built under R version 4.2.3

## INTRODUCTION : We shall conduct the hypothesis testing for equality of 2 sample proportion.

## ABOUT DATASET:

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases.  Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Information about dataset attributes -

Pregnancies: To express the Number of pregnancies

Glucose: To express the Glucose level in blood

BloodPressure: To express the Blood pressure measurement

SkinThickness: To express the thickness of the skin

Insulin: To express the Insulin level in blood

BMI: To express the Body mass index

DiabetesPedigreeFunction: To express the Diabetes percentage

Age: To express the age

Outcome: To express the final result 1 is Yes and 0 is No

##THE DATASET IS TAKEN FROM KAGGLE

diabetes <- read\_csv("C:/Users/NILANJANA/Downloads/diabetes.csv")

## Rows: 768 Columns: 9  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## dbl (9): Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, D...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

## DISPLAYING THE DATASET

View(diabetes)

##DIMENSION OF THE DATASET

dim(diabetes)

## [1] 768 9

## METHOD : We shall split the population into 2 subpopulations based on the no. of pregnancies and our target variable will be ‘Outcome’.

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.2.3

##   
## 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

## The ‘Pregnancies’ column comprises character values which needs to be converted into numeric ones for splitting.

diabetes$Pregnancies = as.numeric(as.character(diabetes$Pregnancies))

## 

## Now, we shall segregate the dataset into 2 population based on the no. of pregnancies a woman had.

## population 1 : comprises the ‘Outcome’ values of those females whose no. of ‘Pregnancies’ <=2

## population 2: comprises the ‘Outcome’ values of those females whose no. of ‘Pregnancies’ >2

pop\_1=filter(diabetes,Pregnancies<=2)  
pop\_1

## # A tibble: 349 × 9  
## Pregnancies Glucose BloodPressure SkinThickness Insulin BMI  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 85 66 29 0 26.6  
## 2 1 89 66 23 94 28.1  
## 3 0 137 40 35 168 43.1  
## 4 2 197 70 45 543 30.5  
## 5 1 189 60 23 846 30.1  
## 6 0 118 84 47 230 45.8  
## 7 1 103 30 38 83 43.3  
## 8 1 115 70 30 96 34.6  
## 9 1 97 66 15 140 23.2  
## 10 2 90 68 42 0 38.2  
## # ℹ 339 more rows  
## # ℹ 3 more variables: DiabetesPedigreeFunction <dbl>, Age <dbl>, Outcome <dbl>

pop\_2=filter(diabetes,Pregnancies>2)  
pop\_2

## # A tibble: 419 × 9  
## Pregnancies Glucose BloodPressure SkinThickness Insulin BMI  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 6 148 72 35 0 33.6  
## 2 8 183 64 0 0 23.3  
## 3 5 116 74 0 0 25.6  
## 4 3 78 50 32 88 31   
## 5 10 115 0 0 0 35.3  
## 6 8 125 96 0 0 0   
## 7 4 110 92 0 0 37.6  
## 8 10 168 74 0 0 38   
## 9 10 139 80 0 0 27.1  
## 10 5 166 72 19 175 25.8  
## # ℹ 409 more rows  
## # ℹ 3 more variables: DiabetesPedigreeFunction <dbl>, Age <dbl>, Outcome <dbl>

target\_var\_1=pop\_1$Outcome

target\_var\_2=pop\_2$Outcome

## DRAWING OF SAMPLE FROM POPULATION 1

sample\_1=sample(pop\_1$Outcome,100,replace=FALSE)

## CONVERTING THE SAMPLE INTO DATAFRAME  
s1=as.data.frame(sample\_1)  
s1

## sample\_1  
## 1 0  
## 2 1  
## 3 0  
## 4 0  
## 5 0  
## 6 0  
## 7 0  
## 8 0  
## 9 0  
## 10 0  
## 11 0  
## 12 0  
## 13 1  
## 14 1  
## 15 0  
## 16 0  
## 17 0  
## 18 0  
## 19 0  
## 20 0  
## 21 0  
## 22 1  
## 23 1  
## 24 1  
## 25 1  
## 26 1  
## 27 0  
## 28 0  
## 29 0  
## 30 0  
## 31 0  
## 32 1  
## 33 0  
## 34 0  
## 35 0  
## 36 0  
## 37 0  
## 38 0  
## 39 0  
## 40 0  
## 41 1  
## 42 1  
## 43 0  
## 44 1  
## 45 0  
## 46 1  
## 47 0  
## 48 1  
## 49 0  
## 50 0  
## 51 0  
## 52 0  
## 53 0  
## 54 1  
## 55 0  
## 56 0  
## 57 0  
## 58 0  
## 59 0  
## 60 1  
## 61 1  
## 62 0  
## 63 0  
## 64 0  
## 65 0  
## 66 0  
## 67 0  
## 68 0  
## 69 0  
## 70 0  
## 71 0  
## 72 1  
## 73 0  
## 74 0  
## 75 0  
## 76 0  
## 77 1  
## 78 1  
## 79 0  
## 80 0  
## 81 0  
## 82 1  
## 83 0  
## 84 1  
## 85 1  
## 86 1  
## 87 0  
## 88 0  
## 89 0  
## 90 0  
## 91 1  
## 92 0  
## 93 1  
## 94 0  
## 95 0  
## 96 0  
## 97 0  
## 98 0  
## 99 0  
## 100 0

## DRAWING OF SAMPLE FROM POPULATION 1

sample\_2=sample(pop\_2$Outcome,130,replace=FALSE)

## CONVERTING THE SAMPLE INTO DATAFRAME

s2=as.data.frame(sample\_2)  
s2

## sample\_2  
## 1 0  
## 2 0  
## 3 1  
## 4 0  
## 5 1  
## 6 0  
## 7 0  
## 8 0  
## 9 0  
## 10 1  
## 11 1  
## 12 0  
## 13 1  
## 14 0  
## 15 0  
## 16 0  
## 17 1  
## 18 1  
## 19 1  
## 20 1  
## 21 0  
## 22 0  
## 23 1  
## 24 1  
## 25 0  
## 26 1  
## 27 0  
## 28 1  
## 29 0  
## 30 0  
## 31 1  
## 32 0  
## 33 1  
## 34 1  
## 35 1  
## 36 1  
## 37 1  
## 38 0  
## 39 0  
## 40 1  
## 41 1  
## 42 1  
## 43 0  
## 44 0  
## 45 1  
## 46 0  
## 47 1  
## 48 0  
## 49 0  
## 50 1  
## 51 1  
## 52 0  
## 53 0  
## 54 0  
## 55 0  
## 56 0  
## 57 1  
## 58 0  
## 59 0  
## 60 1  
## 61 1  
## 62 0  
## 63 0  
## 64 0  
## 65 0  
## 66 0  
## 67 1  
## 68 1  
## 69 1  
## 70 0  
## 71 0  
## 72 0  
## 73 0  
## 74 0  
## 75 1  
## 76 1  
## 77 1  
## 78 1  
## 79 1  
## 80 0  
## 81 0  
## 82 1  
## 83 1  
## 84 1  
## 85 1  
## 86 0  
## 87 1  
## 88 1  
## 89 1  
## 90 0  
## 91 0  
## 92 0  
## 93 1  
## 94 0  
## 95 0  
## 96 1  
## 97 1  
## 98 1  
## 99 1  
## 100 0  
## 101 1  
## 102 0  
## 103 1  
## 104 0  
## 105 1  
## 106 1  
## 107 1  
## 108 0  
## 109 1  
## 110 0  
## 111 0  
## 112 0  
## 113 1  
## 114 1  
## 115 1  
## 116 1  
## 117 0  
## 118 0  
## 119 1  
## 120 1  
## 121 0  
## 122 1  
## 123 0  
## 124 0  
## 125 0  
## 126 0  
## 127 1  
## 128 0  
## 129 1  
## 130 1

library(plyr)

## Warning: package 'plyr' was built under R version 4.2.3

## ------------------------------------------------------------------------------  
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

##CREATING A CONTINGENCY TABLE TO COUNT THE NO. OF ‘1’ AND NO. OF ‘0’ FROM SAMPLE 1

x1=count(s1,sample\_1)  
x1

## sample\_1 sample\_1.1 sample\_1.2 sample\_1.3 sample\_1.4 sample\_1.5 sample\_1.6  
## 1 0 0 0 0 0 0 0  
## 2 1 1 1 1 1 1 1  
## sample\_1.7 sample\_1.8 sample\_1.9 sample\_1.10 sample\_1.11 sample\_1.12  
## 1 0 0 0 0 0 0  
## 2 1 1 1 1 1 1  
## sample\_1.13 sample\_1.14 sample\_1.15 sample\_1.16 sample\_1.17 sample\_1.18  
## 1 0 0 0 0 0 0  
## 2 1 1 1 1 1 1  
## sample\_1.19 sample\_1.20 sample\_1.21 sample\_1.22 sample\_1.23 sample\_1.24  
## 1 0 0 0 0 0 0  
## 2 1 1 1 1 1 1  
## sample\_1.25 freq  
## 1 0 74  
## 2 1 26

##CREATING A CONTINGENCY TABLE TO COUNT THE NO. OF ‘1’ AND NO. OF ‘0’ from SAMPLE 2

x2=count(s2,sample\_2)  
x2

## sample\_2 sample\_2.1 sample\_2.2 sample\_2.3 sample\_2.4 sample\_2.5 sample\_2.6  
## 1 0 0 0 0 0 0 0  
## 2 1 1 1 1 1 1 1  
## sample\_2.7 sample\_2.8 sample\_2.9 sample\_2.10 sample\_2.11 sample\_2.12  
## 1 0 0 0 0 0 0  
## 2 1 1 1 1 1 1  
## sample\_2.13 sample\_2.14 sample\_2.15 sample\_2.16 sample\_2.17 sample\_2.18  
## 1 0 0 0 0 0 0  
## 2 1 1 1 1 1 1  
## sample\_2.19 sample\_2.20 sample\_2.21 sample\_2.22 sample\_2.23 sample\_2.24  
## 1 0 0 0 0 0 0  
## 2 1 1 1 1 1 1  
## sample\_2.25 sample\_2.26 sample\_2.27 sample\_2.28 sample\_2.29 sample\_2.30  
## 1 0 0 0 0 0 0  
## 2 1 1 1 1 1 1  
## sample\_2.31 sample\_2.32 sample\_2.33 sample\_2.34 sample\_2.35 sample\_2.36  
## 1 0 0 0 0 0 0  
## 2 1 1 1 1 1 1  
## sample\_2.37 sample\_2.38 sample\_2.39 sample\_2.40 sample\_2.41 sample\_2.42  
## 1 0 0 0 0 0 0  
## 2 1 1 1 1 1 1  
## sample\_2.43 sample\_2.44 sample\_2.45 sample\_2.46 sample\_2.47 sample\_2.48  
## 1 0 0 0 0 0 0  
## 2 1 1 1 1 1 1  
## sample\_2.49 sample\_2.50 sample\_2.51 sample\_2.52 sample\_2.53 sample\_2.54  
## 1 0 0 0 0 0 0  
## 2 1 1 1 1 1 1  
## sample\_2.55 sample\_2.56 sample\_2.57 sample\_2.58 sample\_2.59 sample\_2.60  
## 1 0 0 0 0 0 0  
## 2 1 1 1 1 1 1  
## sample\_2.61 sample\_2.62 sample\_2.63 sample\_2.64 freq  
## 1 0 0 0 0 65  
## 2 1 1 1 1 65

## TWO-TAILED TEST

H0 : P1=P2

VS

H1 : P1!=P2

x=c(26,65)  
n=c(100,130)  
prop.test(x, n, alternative = "two.sided", conf.level = 0.99)

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: x out of n  
## X-squared = 3.6791, df = 1, p-value = 0.0551  
## alternative hypothesis: two.sided  
## 99 percent confidence interval:  
## -0.29798892 0.03798892  
## sample estimates:  
## prop 1 prop 2   
## 0.27 0.40

##CONCLUSION : The sample estimates for both the populations are 0.27 and 0.40. The p-value is 0.0551 which indicates the acceptance of the H0,i.e, there is equality of proportion of the no. of females patients having diabetes irrespective of their no. of pregnancies.

## RIGHT-TAILED TEST

H0 : P1=P2

VS

H1 : P1 > P2

x=c(27,52)  
n=c(100,130)  
prop.test(x, n, alternative = "greater", conf.level = 0.99)

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: x out of n  
## X-squared = 3.6791, df = 1, p-value = 0.9725  
## alternative hypothesis: greater  
## 99 percent confidence interval:  
## -0.2825752 1.0000000  
## sample estimates:  
## prop 1 prop 2   
## 0.27 0.40

##CONCLUSION : The sample estimates for both the populations are 0.27 and 0.40. The p-value is 0.9725 which indicates the acceptance of the H0,i.e, there is equality of proportion of the no. of females patients having diabetes irrespective of their no. of pregnancies.

## LEFT-TAILED TEST

H0 : P1=P2

VS

H1 : P1 < P2

x=c(27,52)  
n=c(100,130)  
prop.test(x, n, alternative = "less", conf.level = 0.99)

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: x out of n  
## X-squared = 3.6791, df = 1, p-value = 0.02755  
## alternative hypothesis: less  
## 99 percent confidence interval:  
## -1.00000000 0.02257518  
## sample estimates:  
## prop 1 prop 2   
## 0.27 0.40

##CONCLUSION : The sample estimates for both the populations are 0.27 and 0.40. The p-value is 0.02755 which indicates the rejection of the H0,i.e, the sample proportion from population 1 is less than the sample proportion of population 2. This also implies that the no. of female patients having diabetes with no. of pregnancies less than 3 is less than the no. of female patients with no. of pregnancies greater than equal to 3.