Lab-03 : Test for difference between two population mean

Nilanjana Dey

2023-12-05

# I downloaded the PIMA DIABETES dataset from Kaggle  
library(readr)

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

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.

View(diabetes)

# Load dplyr package  
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

# We shall take the 2 populations based on the pregnancy numbers 1 and 2 and their respective BMI.So BMI is our target variable.

# Using filter() we seggregate the rows containing pregnancy 1  
population\_1=filter(diabetes, Pregnancies == '1')  
population\_1

## # A tibble: 135 × 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 1 189 60 23 846 30.1  
## 4 1 103 30 38 83 43.3  
## 5 1 115 70 30 96 34.6  
## 6 1 97 66 15 140 23.2  
## 7 1 146 56 0 0 29.7  
## 8 1 103 80 11 82 19.4  
## 9 1 101 50 15 36 24.2  
## 10 1 73 50 10 0 23   
## # ℹ 125 more rows  
## # ℹ 3 more variables: DiabetesPedigreeFunction <dbl>, Age <dbl>, Outcome <dbl>

#There are 135 observations

#we seggregate the rows containing pregnancy 2  
  
population\_2=filter(diabetes, Pregnancies == '2')  
population\_2

## # A tibble: 103 × 9  
## Pregnancies Glucose BloodPressure SkinThickness Insulin BMI  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 2 197 70 45 543 30.5  
## 2 2 90 68 42 0 38.2  
## 3 2 71 70 27 0 28   
## 4 2 84 0 0 0 0   
## 5 2 141 58 34 128 25.4  
## 6 2 109 92 0 0 42.7  
## 7 2 100 66 20 90 32.9  
## 8 2 112 66 22 0 25   
## 9 2 74 0 0 0 0   
## 10 2 110 74 29 125 32.4  
## # ℹ 93 more rows  
## # ℹ 3 more variables: DiabetesPedigreeFunction <dbl>, Age <dbl>, Outcome <dbl>

#there are 103 observations

# BMI values of first population  
BMI\_1=population\_1$BMI  
BMI\_1

## [1] 26.6 28.1 30.1 43.3 34.6 23.2 29.7 19.4 24.2 23.0 19.6 32.0 24.7 26.5 19.1  
## [16] 20.4 49.7 39.0 26.1 26.6 28.7 22.4 31.2 55.0 33.3 34.5 37.4 40.6 25.2 43.5  
## [31] 27.7 32.0 24.3 40.1 33.2 23.6 34.6 25.4 29.2 30.1 33.6 25.9 33.2 45.6 40.5  
## [46] 25.6 35.1 43.3 23.9 25.9 25.9 32.0 28.7 27.2 33.3 36.5 32.8 37.2 30.8 25.4  
## [61] 25.1 24.3 22.3 32.0 23.7 42.4 34.4 42.4 26.2 27.2 18.2 45.3 34.1 35.0 30.0  
## [76] 24.5 18.2 25.3 21.2 26.7 21.8 41.3 40.7 28.6 18.2 41.3 33.3 32.0 32.8 27.4  
## [91] 29.9 36.9 38.1 37.6 38.6 30.1 35.5 22.5 27.8 36.8 23.1 27.1 27.8 35.8 40.0  
## [106] 19.5 24.0 27.5 19.5 23.4 25.2 33.8 34.2 38.9 42.9 34.8 25.4 28.3 24.1 46.1  
## [121] 35.0 35.5 38.1 29.3 32.8 36.1 39.5 28.5 49.3 46.3 39.0 36.5 37.5 30.1 30.4

# BMI values of 2nd population  
BMI\_2=population\_2$BMI  
BMI\_2

## [1] 30.5 38.2 28.0 0.0 25.4 42.7 32.9 25.0 0.0 32.4 38.5 24.7 31.6 39.6 33.6  
## [16] 21.1 33.8 32.5 30.5 27.3 24.6 29.0 29.7 28.9 29.7 20.4 27.7 31.6 38.2 28.0  
## [31] 24.4 28.7 40.0 27.5 25.2 40.5 25.3 27.0 43.3 28.0 31.6 39.7 32.9 26.6 34.1  
## [46] 35.7 45.5 32.7 34.9 36.2 27.7 42.1 26.0 30.8 22.2 25.6 32.2 28.9 19.6 37.8  
## [61] 33.7 36.8 33.5 30.1 25.2 30.4 27.3 25.0 29.8 26.1 22.6 34.7 42.9 34.7 39.4  
## [76] 28.5 20.1 38.4 24.2 30.8 26.2 31.6 39.4 26.8 38.7 21.8 24.2 24.2 33.2 39.1  
## [91] 23.5 35.9 38.5 34.4 38.0 22.9 30.1 44.5 29.0 23.3 36.6 28.4 36.8

# Drawing a sample of size 60 from population 1  
sample\_1=sample(BMI\_1,60,replace=FALSE)  
sample\_1

## [1] 19.5 39.0 37.6 25.9 40.0 46.3 32.8 27.8 22.5 29.7 49.3 25.4 19.1 27.2 29.2  
## [16] 21.8 45.6 46.1 27.4 30.4 30.1 39.5 33.6 39.0 43.5 33.2 33.3 34.1 36.5 23.6  
## [31] 18.2 35.0 38.6 36.9 31.2 24.7 26.1 34.6 35.0 33.8 18.2 32.8 19.5 40.5 32.0  
## [46] 27.8 30.0 40.7 28.5 40.1 42.4 38.1 32.0 34.4 25.3 34.6 35.5 38.9 28.7 32.0

#Drawing a sample of size 50 from population 2  
sample\_2=sample(BMI\_2,50,replace=FALSE)  
sample\_2

## [1] 19.6 30.1 30.4 31.6 45.5 26.0 25.0 38.2 33.7 28.0 24.4 25.0 24.7 30.5 29.0  
## [16] 38.0 24.2 26.1 38.5 34.9 42.1 43.3 28.0 27.3 24.2 28.9 29.7 38.5 27.0 34.1  
## [31] 26.6 35.7 33.2 33.6 31.6 32.7 28.0 34.4 27.5 28.5 28.4 38.7 39.4 42.7 27.7  
## [46] 25.2 38.2 29.0 21.1 0.0

# S.D of BMI values from population 1  
sd(BMI\_1)

## [1] 7.548474

# S.D of BMI values from population 2  
sd(BMI\_2)

## [1] 7.483683

##TESTING OF HYPOTHESIS

#Two-tailed H0:MEAN\_1=MEAN\_2 VS H1:MEAN\_1!=MEAN\_2

#Performing the 2 sample z test  
library(BSDA)

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

## Loading required package: lattice

##   
## Attaching package: 'BSDA'

## The following object is masked from 'package:datasets':  
##   
## Orange

z.test(x=sample\_1,y=sample\_2,  
alternative = "two.sided",  
mu = 0,  
sigma.x = sd(BMI\_1),  
sigma.y = sd(BMI\_2),  
conf.level = 0.95  
)

##   
## Two-sample z-Test  
##   
## data: sample\_1 and sample\_2  
## z = 1.3978, p-value = 0.1622  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.8087391 4.8307391  
## sample estimates:  
## mean of x mean of y   
## 32.585 30.574

#Since the p-value is greater than 0.05, we accept the H0.Hence the means of both the populations are similar.

#Two-tailed H0:MEAN\_1=MEAN\_2 VS H1:MEAN\_1>MEAN\_2

library(BSDA)  
z.test(x=sample\_1,y=sample\_2,  
alternative = "greater",  
mu = 0,  
sigma.x = sd(BMI\_1),  
sigma.y = sd(BMI\_2),  
conf.level = 0.95  
)

##   
## Two-sample z-Test  
##   
## data: sample\_1 and sample\_2  
## z = 1.3978, p-value = 0.08108  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## -0.3553996 NA  
## sample estimates:  
## mean of x mean of y   
## 32.585 30.574

#Since the p-value is greater than 0.05, we accept the H0.Hence the means of both the populations are similar.

#Two-tailed H0:MEAN\_1=MEAN\_2 VS H1:MEAN\_1<MEAN\_2

library(BSDA)  
z.test(x=sample\_1,y=sample\_2,  
alternative = "less",  
mu = 0,  
sigma.x = sd(BMI\_1),  
sigma.y = sd(BMI\_2),  
conf.level = 0.95  
)

##   
## Two-sample z-Test  
##   
## data: sample\_1 and sample\_2  
## z = 1.3978, p-value = 0.9189  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## NA 4.3774  
## sample estimates:  
## mean of x mean of y   
## 32.585 30.574

#Since the p-value is greater than 0.05, we accept the H0.Hence the means of both the populations are similar.

# HENCE WE CAN CONCLUDE THAT THERE IS NO DIFFERENCE BETWEEN THE BMI VALUES OF WOMEN WHO HAD 1 PREGNANCY AND BMI VALUES OF WOMEN WHO HAD 2 PREGNANCIES.