Instituto Tecnológico de Costa Rica

Proyecto de Estadística

Integrantes:

Jose Ignacio Alfaro Solano

Marco Antonio Espinoza Salas

Warner Fidel Hurtado Laguna

Lester Alexander Trejos Bermúdez

Profesor:

Esteban Ballestero Alfaro

Fecha:

14/4/2021

### Nombre de la base de datos

nycflights13

### Descripción del dataset seleccionado

La base utilizada para el proyecto contiene información sobre todos los vuelos que salían de NUEVA YORK (por ejemplo EWR, JFK Y LGA)a destinos en los Estados Unidos, Puerto Rico y las Islas Vírgenes Americanas) en 2013: 336,776 vuelos en total. Para ayudar a comprender qué provoca retrasos, también incluye otros conjuntos de datos útiles. La base que está siendo utilizada contiene muchas tablas de datos como vuelos, climas, aviones, aeropuertos etc, para el desarrollo del presente proyecto se eligió la columna flights o vuelos.

El paquete flights o vuelos se divide en varias columnas como año, hora de salida, minutos, avión, origen destino y muchos otros. Para el manejo de este proyecto se trabajó con tres columnas respectivas, la de origen, destino y duración en el aire, las cuales fueron utilizadas en pruebas de una y dos poblaciones. Para la prueba se tomaron un origen y un destino en el cual se toma la duración entre ese respectivo origen y destino.

### Librerías necesarias

if (!require(‘devtools’)) install.packages(‘devtools’) devtools::install\_github(‘fhernanb/stests’, force=TRUE)

library(nycflights13)

## Warning: package 'nycflights13' was built under R version 4.0.5

library(PASWR2)

## Warning: package 'PASWR2' was built under R version 4.0.5

## Loading required package: lattice

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.0.4

library(stests)

##   
## Attaching package: 'stests'

## The following object is masked from 'package:PASWR2':  
##   
## z.test

## The following object is masked from 'package:stats':  
##   
## var.test

library(BSDA)

## Warning: package 'BSDA' was built under R version 4.0.5

##   
## Attaching package: 'BSDA'

## The following object is masked from 'package:stests':  
##   
## z.test

## The following objects are masked from 'package:PASWR2':  
##   
## normarea, nsize, ntester, SIGN.test, tsum.test, z.test, zsum.test

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

data("flights")  
  
dataFlight <- flights  
  
  
#if (!require('devtools')) install.packages('devtools')  
#devtools::install\_github('fhernanb/stests', force=TRUE)

### Variables de la base datos

dataFlight

## # A tibble: 336,776 x 19  
## year month day dep\_time sched\_dep\_time dep\_delay arr\_time sched\_arr\_time  
## <int> <int> <int> <int> <int> <dbl> <int> <int>  
## 1 2013 1 1 517 515 2 830 819  
## 2 2013 1 1 533 529 4 850 830  
## 3 2013 1 1 542 540 2 923 850  
## 4 2013 1 1 544 545 -1 1004 1022  
## 5 2013 1 1 554 600 -6 812 837  
## 6 2013 1 1 554 558 -4 740 728  
## 7 2013 1 1 555 600 -5 913 854  
## 8 2013 1 1 557 600 -3 709 723  
## 9 2013 1 1 557 600 -3 838 846  
## 10 2013 1 1 558 600 -2 753 745  
## # ... with 336,766 more rows, and 11 more variables: arr\_delay <dbl>,  
## # carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,  
## # air\_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time\_hour <dttm>

### Resumen de variables seleccionadas

#### La siguiente variable fue utilizada para una y dos poblaciones, en donde se utilizaba la duración en aire de un destino al otro.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Variables | Tipo | Parametros por estima (IC) |
|  | air-time | minutos | Normal |
|  | air-time | minutos | t Student |
|  | air-time | minutos | F |
|  | air-time | minutos | Chi-Cuadrado |
|  |  |  | Exponencial |

## Graficos

Que alguién los haga jaja

## Pruebas de intervalos de confianza para una población

### Distribución normal Z

En el siguiente prueba se procederá a realizar las respectivas pruebas con las distribución z, en la cual se tomará un nivel de confianza del 95%.

#En la siguiente asignación se toman las n cantidad de vuelos que salen de LGA y tienen como destino JFK, y me guarda en var1 sólo los vuelos con dichas características.  
var1 <- dataFlight[dataFlight$origin == "LGA" | dataFlight$dest == "JFK", ]  
  
#En esta variable se trabaja con el tiempo en el aire de el origen y destino seleccionado anterionrmente  
var1 <- na.omit(var1$air\_time)  
  
#Se procede a aplicar la función de PASWR2, en la cual se le pasan como parametros la lista de tiempos, la desviación estandar y por último el nivel de significancia.  
z.test(x = var1, sigma.x = sd(var1), conf.level = 0.95)$conf.int

## [1] 117.5214 118.1302  
## attr(,"conf.level")  
## [1] 0.95

Al realizar los respectivos cálculos de la distribución normal z con la ‘n’ cantidad de tiempos de vuelos con origen LGA y destino a JFK se obtuvo un intervalo de confianza de: ]117.5214, 118.1302[.

### Distribución T

En el siguiente prueba se procederá a realizar las respectivas pruebas con las distribución T, en la cual se tomará un nivel de confianza del 95%. Esta distribución debe tener un n < 30, Por lo que se tomarón los primeros 25 datos de la columna de air-time para dicha prueba.

#En la siguiente asignación se toman las n cantidad de vuelos que salen de LGA y tienen como destino JFK, y me guarda en var1 sólo los vuelos con dichas características.  
var1 <- dataFlight[dataFlight$origin == "LGA" | dataFlight$dest == "JFK", ]  
  
#En esta variable se trabaja con el tiempo (air-time) en el aire de el origen y destino seleccionado anterionrmente el cual se van a utlizar las primeras 25 posiciones  
var1 <- na.omit(var1$air\_time[1: 25])  
  
#Se procede a aplicar la función T, en la cual se le pasan como parametros la lista de tiempos y por último el nivel de significancia.  
t.test(var1, conf.level = 0.95)$conf.int

## [1] 134.9679 181.4321  
## attr(,"conf.level")  
## [1] 0.95

-

Al realizar los respectivos cálculos de la distribución T con 25 datos de tiempos de vuelo, de origen LGA a destino JFK se obtuvo un intervalo de confianza de: ]134.9679, 180.2621[.

### Una proporción

Para realizar el respectivo cálculo de la proporción de tiempos de vuelo se tomaron todos los vuelos no mayores o igual 100, en la cual se tomará un nivel de confianza del 95%.

#En la siguiente asignación se toman las n cantidad de vuelos que salen de LGA y tienen como destino JFK, y me guarda en var1 sólo los vuelos con dichas características.  
var1 <- dataFlight[dataFlight$origin == "LGA" | dataFlight$dest == "JFK", ]  
  
#En la siguiente variable se asignan los éxitos obtenidos de la condioción n <= 100  
exitos <- var1[which(na.omit(var1$air\_time) <= 100), ]  
  
#En esta variable se trabaja con el tiempo en el aire de el origen y destino seleccionado anterionrmente  
var1 <- na.omit(var1$air\_time)  
  
  
#Se obtine la cantidad de exitos de la variable anterior exitos  
totalExitos <- length(exitos$air\_time)  
totalExitos

## [1] 35797

#Para obtener la proporción se procede a ingresar el total de éxitos, el total de todas las variables y por último el nivel de confianza  
prop.test(totalExitos, length(var1), conf.level = 0.95)$conf.int

## [1] 0.3509888 0.3568926  
## attr(,"conf.level")  
## [1] 0.95

#Al realizar los respectivos calculos de la proporcion con la cantidad de tiempos de vuelos de origen LGA a destino JFK menores a 100 minutos se obtuvo un intervalo de confianza de: ]0.3509888, 0.3568926[.

### Varianza

En la última prueba de una población se procede a realizar los cálculos con un intervalo de confianza del 95%, en la cual se utiliza la catidad total de la la variable air-time para una mejor estamación.

#En la siguiente asignación se toman las n cantidad de vuelos que salen de LGA y tienen como destino JFK, y me guarda en var1 sólo los vuelos con dichas características.  
var1 <- dataFlight[dataFlight$origin == "LGA" | dataFlight$dest == "JFK", ]  
  
#En esta variable se trabaja con el tiempo en el aire de el origen y destino seleccionado anterionrmente  
var1 <- na.omit(var1$air\_time)  
  
#En la siguiente función de stests se procede a igresar la lista de los tiempos y luego el nivel de confianza  
var.test(x = var1, conf.level = 0.95)$conf.int

## [1] 2419.025 2461.562  
## attr(,"conf.level")  
## [1] 0.95

* Al realizar los respectivos cálculos de la variancia con un intervalo de confianza del 95% para los tiempos de vuelos de origen LGA a destino JFK se obtuvo un intervalo de confianza de: ]2419.025, 2461.562[