####Ayudantía 3####

####Repaso y ejercicios####

####Martes 26 de abril de 2022####

####Ayudante Santiago Ortúzar####

####Estadística descriptiva-R, FEN-UAH####

####Prof. Alejandro González####

####Siempre limpiar el environment####

rm(list=ls())

####Buscar el directorio desde el que abrimos el script####

getwd()

####Fijar el directorio en el que queremos trabajar####

#setwd()

####Parte 1####

####Librerías####

#install.packages("dplyr")

library(dplyr)

####Base de datos####

data(starwars)

####P 1####

starwars1 <- starwars[starwars$name=="C-3PO"|

starwars$name=="Chewbacca"|

starwars$name=="Darth Vader"|

starwars$name=="Jabba Desilijic Tiure"|

starwars$name=="Luke Skywalker"|

starwars$name=="Obi-Wan Kenobi"|

starwars$name=="R2-D2"|

starwars$name=="Yoda", ]

####P 2####

starwars2 <- starwars1[order(starwars1$height,decreasing = TRUE), ]

####P 3####

rm(starwars)

rm(starwars1)

####P 4####

#Opción 1

starwars3 <- starwars2[starwars2$height>199, ] #seleccionar filas con valores mayores a 199

#Opción 2

starwars3 <- starwars2[starwars2$height>=200, ] #seleccionar filas con valores iguales o mayores a 200

#Opción 3

starwars3 <- starwars2[1:2, ] #seleccionar filas 1 y 2

####P 5####

starwars4 <- starwars2[ , c("name","height","mass","birth\_year",

"sex", "homeworld")]

####P 6####

rm(starwars2)

rm(starwars3)

####P 7 A####

yoda <- starwars4[starwars4$name=="Yoda" , c("name","birth\_year", "homeworld")]

####Concepto de caso perdido####

####P 7 B####

vader <- starwars4[starwars4$name=="Darth Vader", c("name", "height")]

####P 7 C####

jabba <- starwars4[starwars4$name=="Jabba Desilijic Tiure", c("name", "sex", "birth\_year", "mass")]

####P 7 D####

tatooine <- starwars4[starwars4$homeworld=="Tatooine", ]

####P 7 E####

edad\_obiwan <- starwars4$birth\_year[starwars4$name=="Obi-Wan Kenobi"]

edad\_vader <- starwars4$birth\_year[starwars4$name=="Darth Vader"]

edad\_obiwan-edad\_vader

####Parte 2####

####Siempre limpiar el environment####

rm(list=ls())

####Buscar el directorio desde el que abrimos el script####

getwd()

####Fijar el directorio en el que queremos trabajar####

setwd("D:/SANTIAGO/OneDrive/Documentos/2022 Ayudantía UAH A. González Estadística R")

data(swiss) #paquete disponible en R base

View(swiss)

####P 1####

help(swiss)

*#Description: Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.*

*#A data frame with 47 observations on 6 variables, each of which is in percent, i.e., in [0, 100].*

*Switzerland, in 1888, was entering a period known as the demographic transition; i.e., its fertility was beginning to fall from the high level typical of underdeveloped countries.*

*The data collected are for 47 French-speaking “provinces” at about 1888.*

*[,1] Fertility Ig, ‘common standardized fertility measure’*

*[,2] Agriculture % of males involved in agriculture as occupation*

*[,3] Examination % draftees receiving highest mark on army examination*

*[,4] Education % education beyond primary school for draftees.*

*[,5] Catholic % ‘catholic’ (as opposed to ‘protestant’).*

*[,6] Infant.Mortality live births who live less than 1 year.*

####P 2 A####

hist(swiss$Fertility, breaks=10)

####P 2 B####

#Media

mean(swiss$Fertility) #70,1

#Mediana

median(swiss$Fertility) #70,4

#Moda

#Código profesor

table(swiss$Fertility)

which.max(table(swiss$Fertility)) #máximo valor de frecuencia

names(which.max(table(swiss$Fertility))) #nombre asociado a ese valor

#install.packages("DescTools")

library(DescTools)

Mode(swiss$Fertility) #65

####P 2 C####

range(swiss$Fertility)

#¿Qué indica el rango? El rango indica un conjunto de dos valores: el valor mínimo en una distribución y el valor máximo. A nosotros al conocer una distribución en general nos interesa conocer dónde se concentra la masa de datos (para esto sumamos las “medidas de tendencia central”, como media, mediana o moda), pero también información sobre los “extremos” (o “las colas”). El rango nos indica dónde se ubican los valores que están en cada uno de los extremos.

####P 2 D####

#Opción 1

quantile(swiss$Fertility, probs = c(0.25, 0.50, 0.75, 1.0))

#Opción 2

quantile(swiss$Fertility, probs = seq(from = 0.25, to = 1.0, by = 0.25))

#Opción 3

library(Hmisc)

Hmisc::describe(swiss$Fertility)

####P 2 E####

#Describa lo que observa en el histograma.

####Distribución que en algunos aspectos se asemeja a la normal, pero

####Tiene un sesgo a la izquierda

####El grueso de la masa de datos se concentra en entre 70 y 80, mientras que el rango va entre XXX y XXX.

####Esto último lo podemos ver tanto en el histograma como en los cuartiles.

####P 2 F####

hist(swiss$Fertility, breaks=5)

hist(swiss$Fertility, breaks=10)

hist(swiss$Fertility, breaks=25)

####P 2 G####

plot(density(swiss$Fertility))

####P 2 H####

#Recuerde que ahora al código del histograma debe agregar el comando freq=FALSE

hist(swiss$Fertility, breaks=5, freq=FALSE)

lines(density(swiss$Fertility), col="red", lwd=2)

hist(swiss$Fertility, breaks=10, freq=FALSE)

lines(density(swiss$Fertility), col="red", lwd=2)

hist(swiss$Fertility, breaks=25, freq=FALSE)

lines(density(swiss$Fertility), col="red", lwd=2)

####P 3 A####

hist(swiss$Education, breaks=8, freq=FALSE)

lines(density(swiss$Education), col="red", lwd=2)

####P 3 B####

#Utilizando el rango, la media, la mediana, la moda, cuartiles y/o deciles, y cualquier otro

#estadístico que usted considere pertinente,

#además del histograma del ejercicio anterior

#¿qué diría usted sobre la distribución de educación

#en las regiones francesas?

range(swiss$Education)

diff(range(swiss$Education))

mean(swiss$Education)

median(swiss$Education)

table(swiss$Education)

which.max(table(swiss$Education)) #máximo valor de frecuencia

moda.Education <- names(which.max(table(swiss$Education))) #nombre asociado a ese valor

quantile(swiss$Education, probs = c(0.25, 0.50, 0.75, 1.0))

quantile(swiss$Fertility, probs = seq(from = 0.1, to = 1.0, by = 0.1))

####P 4####

hist(swiss$Catholic, breaks=14, freq=FALSE)

lines(density(swiss$Catholic), col="red", lwd=2)

range(swiss$Catholic)

diff(range(swiss$Catholic))

mean(swiss$Catholic)

median(swiss$Catholic)

table(swiss$Catholic)

which.max(table(swiss$Catholic)) #máximo valor de frecuencia

moda.Catholic <- names(which.max(table(swiss$Catholic))) #nombre asociado a ese valor

quantile(swiss$Catholic, probs = c(0.25, 0.50, 0.75, 1.0))

quantile(swiss$Catholic, probs = seq(from = 0.1, to = 1.0, by = 0.1))