Estadistica descriptiva I

Alexis Rangel

2022-11-07

### SHORTCUTS

ctrl + alt + k selecciona todas las ocurrencias de una NIVEL variable ctrl + shift + l selecciona la linea o lineas completas CTRL + shift + v para copiar y pegar la celda anterior

INICIANDO LA ESTADISTICA DESCRIPTIVA Y ANALISIS DE DATOS DE LA BASE DE DATOS

setwd("~/Maestria/Curso propedeutico/coloquio/ColoquioMAEA")   
#estableciendo directorio  
library(readr)  
library(ggplot2)  
library(tidyr)  
library(readxl)  
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

library(magrittr)

##   
## Attaching package: 'magrittr'

## The following object is masked from 'package:tidyr':  
##   
## extract

library(glue)  
library(plyr)

## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.  
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:  
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##   
## Attaching package: 'plyr'

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

#file.choose()  
ruta\_db = "./DBS/Default of Credit Card Clients Dataset/UCI\_Credit\_Card.csv"  
  
df0 = read.csv(ruta\_db)  
dfbase = read.csv(ruta\_db)

## ¿De qué trata mi base de datos? -Introducción

####Default Payments of Credit Card Clients in Taiwan from 2005 Esta base de datos cuenta con NIVEL variables numericas y categoricas que describen la situación crediticia de personas con tarjetas de crédito en Taiwan en 2005

¿Qué NIVEL variables vamos a usar? Todas. ID: ID of each client ####\*Categoricas EDUCATION: Gender (1=male, 2=female) MARRIAGE: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown) MARRIAGE: Marital status (1=married, 2=single, 3=others) AGE: Age in years

####\*Numericas LIMIT\_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit PAY\_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above) PAY\_2: Repayment status in August, 2005 (scale same as above) PAY\_3: Repayment status in July, 2005 (scale same as above) PAY\_4: Repayment status in June, 2005 (scale same as above) PAY\_5: Repayment status in May, 2005 (scale same as above) PAY\_6: Repayment status in April, 2005 (scale same as above) BILL\_AMT1: Amount of bill statement in September, 2005 (NT dollar) BILL\_AMT2: Amount of bill statement in August, 2005 (NT dollar) BILL\_AMT3: Amount of bill statement in July, 2005 (NT dollar) BILL\_AMT4: Amount of bill statement in June, 2005 (NT dollar) BILL\_AMT5: Amount of bill statement in May, 2005 (NT dollar) BILL\_AMT6: Amount of bill statement in April, 2005 (NT dollar) PAY\_AMT1: Amount of previous payment in September, 2005 (NT dollar) PAY\_AMT2: Amount of previous payment in August, 2005 (NT dollar) PAY\_AMT3: Amount of previous payment in July, 2005 (NT dollar) PAY\_AMT4: Amount of previous payment in June, 2005 (NT dollar) PAY\_AMT5: Amount of previous payment in May, 2005 (NT dollar) PAY\_AMT6: Amount of previous payment in April, 2005 (NT dollar) default.payment.next.month: Default payment (1=yes, 0=no)

Inicialmente modificamos valores por factores y filtramos NIVEL variables con valores no establecidas en la descrición

names(df0)

## [1] "ID" "LIMIT\_BAL"   
## [3] "SEX" "EDUCATION"   
## [5] "MARRIAGE" "AGE"   
## [7] "PAY\_0" "PAY\_2"   
## [9] "PAY\_3" "PAY\_4"   
## [11] "PAY\_5" "PAY\_6"   
## [13] "BILL\_AMT1" "BILL\_AMT2"   
## [15] "BILL\_AMT3" "BILL\_AMT4"   
## [17] "BILL\_AMT5" "BILL\_AMT6"   
## [19] "PAY\_AMT1" "PAY\_AMT2"   
## [21] "PAY\_AMT3" "PAY\_AMT4"   
## [23] "PAY\_AMT5" "PAY\_AMT6"   
## [25] "default.payment.next.month"

head(df0)

## ID LIMIT\_BAL SEX EDUCATION MARRIAGE AGE PAY\_0 PAY\_2 PAY\_3 PAY\_4 PAY\_5 PAY\_6  
## 1 1 20000 2 2 1 24 2 2 -1 -1 -2 -2  
## 2 2 120000 2 2 2 26 -1 2 0 0 0 2  
## 3 3 90000 2 2 2 34 0 0 0 0 0 0  
## 4 4 50000 2 2 1 37 0 0 0 0 0 0  
## 5 5 50000 1 2 1 57 -1 0 -1 0 0 0  
## 6 6 50000 1 1 2 37 0 0 0 0 0 0  
## BILL\_AMT1 BILL\_AMT2 BILL\_AMT3 BILL\_AMT4 BILL\_AMT5 BILL\_AMT6 PAY\_AMT1 PAY\_AMT2  
## 1 3913 3102 689 0 0 0 0 689  
## 2 2682 1725 2682 3272 3455 3261 0 1000  
## 3 29239 14027 13559 14331 14948 15549 1518 1500  
## 4 46990 48233 49291 28314 28959 29547 2000 2019  
## 5 8617 5670 35835 20940 19146 19131 2000 36681  
## 6 64400 57069 57608 19394 19619 20024 2500 1815  
## PAY\_AMT3 PAY\_AMT4 PAY\_AMT5 PAY\_AMT6 default.payment.next.month  
## 1 0 0 0 0 1  
## 2 1000 1000 0 2000 1  
## 3 1000 1000 1000 5000 0  
## 4 1200 1100 1069 1000 0  
## 5 10000 9000 689 679 0  
## 6 657 1000 1000 800 0

unique(df0$EDUCATION)

## [1] 2 1 3 5 4 6 0

df0$EDUCATION = factor(df0$EDUCATION, levels = c(1, 2), labels = c('Hombre','Mujer'))  
recognized\_levels\_EDUCATION = c(1, 2, 3, 4, 5, 6)  
  
#filtrando el df0 solo con valores reconocidos para la var MARRIAGE  
df0 = df0[df0$MARRIAGE %in% recognized\_levels\_EDUCATION, ]  
df0$MARRIAGE = factor(df0$MARRIAGE, levels = c(1,2,3,4,5,6), labels = c('graduate school', 'university', 'high school', 'others', 'unknown', 'unknown2'))  
unique(df0$MARRIAGE)

## [1] graduate school university high school   
## Levels: graduate school university high school others unknown unknown2

# MARRIAGE: Marital status (1=married, 2=single, 3=others)  
unique(df0$MARRIAGE)

## [1] graduate school university high school   
## Levels: graduate school university high school others unknown unknown2

recognized\_levels\_marriage = c(1, 2, 3)  
df0 = df0[df0$MARRIAGE %in% recognized\_levels\_marriage, ]  
df0$MARRIAGE = factor(df0$MARRIAGE, levels = recognized\_levels\_marriage, labels = c('married', 'single', 'others'))  
  
diferencia\_df0\_dfbase = dim(dfbase)[1] - dim(df0)[1]  
  
glue('INFORME: Al finalizar los primeros filtrados de valores no especificados de la BD, se eliminaron {diferencia\_df0\_dfbase} registros de un total de {length(dfbase)}')

## INFORME: Al finalizar los primeros filtrados de valores no especificados de la BD, se eliminaron 30000 registros de un total de 25

Manejando proporciones

names(df0)

## [1] "ID" "LIMIT\_BAL"   
## [3] "SEX" "EDUCATION"   
## [5] "MARRIAGE" "AGE"   
## [7] "PAY\_0" "PAY\_2"   
## [9] "PAY\_3" "PAY\_4"   
## [11] "PAY\_5" "PAY\_6"   
## [13] "BILL\_AMT1" "BILL\_AMT2"   
## [15] "BILL\_AMT3" "BILL\_AMT4"   
## [17] "BILL\_AMT5" "BILL\_AMT6"   
## [19] "PAY\_AMT1" "PAY\_AMT2"   
## [21] "PAY\_AMT3" "PAY\_AMT4"   
## [23] "PAY\_AMT5" "PAY\_AMT6"   
## [25] "default.payment.next.month"

attach(df0)  
# ¿cuántos hombre y mujeres hay en la población y que proporción nos indica del total?  
table(EDUCATION)

## EDUCATION  
## Hombre Mujer   
## 0 0

prop.table(table(EDUCATION))

## EDUCATION  
## Hombre Mujer   
##

# ¿Cuántos hombres y mujeres hay por nivel educativo y que proporción representan en cada uno?  
table(SEX, EDUCATION)

## < table of extent 0 x 2 >

prop.table(table(SEX, EDUCATION)) # A NIVEL GLOBAL

## EDUCATION  
## SEX Hombre Mujer

prop.table(table(SEX, EDUCATION), margin = 1) # A NIVEL VARIABLE SEX

## EDUCATION  
## SEX Hombre Mujer

prop.table(table(SEX, EDUCATION), margin = 2) # A NIVEL VARIABLE EDUCACIÓN

## EDUCATION  
## SEX Hombre Mujer

"  
Se reconocen los 4 grupos, por educación, más robustos en la población:  
  
 1) Mujeres en universidad, representando el 28.9% de la población  
 i.Representando el 47.8% del total de mujeres  
 2) Mujeres graduadas, representando el 20.8%  
 3) Hombres en universidad, con el 17.95%  
 i.Representando el 45.28% del total de hombres  
 4) Hombres graduados de universidad con el 14.54%  
"

## [1] "\nSe reconocen los 4 grupos, por educación, más robustos en la población:\n\n 1) Mujeres en universidad, representando el 28.9% de la población\n i.Representando el 47.8% del total de mujeres\n 2) Mujeres graduadas, representando el 20.8%\n 3) Hombres en universidad, con el 17.95%\n i.Representando el 45.28% del total de hombres\n 4) Hombres graduados de universidad con el 14.54%\n"

# ¿Cuántos hombres y mujeres hay por estado civil casado y que proporción representan en cada uno?  
table(EDUCATION, MARRIAGE)

## MARRIAGE  
## EDUCATION married single others  
## Hombre 0 0 0  
## Mujer 0 0 0

prop.table(table(EDUCATION, MARRIAGE)) # A NIVEL GLOBAL

## MARRIAGE  
## EDUCATION married single others  
## Hombre   
## Mujer

prop.table(table(EDUCATION, MARRIAGE), margin = 1) # A NIVEL VARIABLE EDUCATIONO

## MARRIAGE  
## EDUCATION married single others  
## Hombre   
## Mujer

prop.table(table(EDUCATION, MARRIAGE), margin = 2) # A NIVEL VARIABLE EDUCACIÓN

## MARRIAGE  
## EDUCATION married single others  
## Hombre   
## Mujer

"  
Por estado civil se pueden el orden de grupos más robustos en la población:  
 1) Las mujeres solteras que representan un 31.4% de la población  
 i. Representando al 52% de las mujeres  
 2) Las mujeres casadas con un 28.28%  
 3) Los hombres solteros con un 21.87%  
 i. Representando al 55.17% de las mujeres  
 4) Los hombres casados con un 17.33%  
  
Por EDUCATIONo, se tiene más registros de personas solteras  
"

## [1] "\nPor estado civil se pueden el orden de grupos más robustos en la población:\n 1) Las mujeres solteras que representan un 31.4% de la población\n i. Representando al 52% de las mujeres\n 2) Las mujeres casadas con un 28.28%\n 3) Los hombres solteros con un 21.87%\n i. Representando al 55.17% de las mujeres\n 4) Los hombres casados con un 17.33%\n\nPor EDUCATIONo, se tiene más registros de personas solteras\n"

# ¿Cuántos Casados y solteros hay por nivel educativo y que proporción representan en cada uno?  
table(EDUCATION, MARRIAGE)

## MARRIAGE  
## EDUCATION married single others  
## Hombre 0 0 0  
## Mujer 0 0 0

prop.table(table(EDUCATION, MARRIAGE)) # A NIVEL GLOBAL

## MARRIAGE  
## EDUCATION married single others  
## Hombre   
## Mujer

prop.table(table(EDUCATION, MARRIAGE), margin = 1) # A NIVEL VARIABLE EDUCATION

## MARRIAGE  
## EDUCATION married single others  
## Hombre   
## Mujer

prop.table(table(EDUCATION, MARRIAGE), margin = 2) # A NIVEL VARIABLE MARRIAGE

## MARRIAGE  
## EDUCATION married single others  
## Hombre   
## Mujer