Lab: Inferencia Estadística 27/Ago/2021

Analisis de bases y manipulación:

```
library(readr)
library(tidyverse)
library(lubridate)
#Es muy facil conectar con sistemas de bases de datos con tidyverse
# Leer la base de datos que acabamos de importar
datos <- read_csv("/Volumes/External/carpetas-de-investigacion-pgj-cdmx.csv")
# <- significa igual en R, lo puedo usar para asignar valores etc etc...

#Vamos a aplicar funciones: cuantos datos tengo con tally(.) ?
tally(datos)
#de otra forma usando pipe %>% aplica la funcion del lado derecho al objeto del
#lado izquierdo: base %>% tally()
datos %>% tally()

#Ahora vamos a calcular la media: glimpse() nos resume los datos
datos %>% glimpse()
datos[1,2] # accedemos a la entrada de una matriz
datos$delito # vemos la columna delito

#vamos a hacer que lea la fecha estructurada

datos <- datos %>% mutate(fecha_nueva=ymd_hms(fecha_hechos))


#Vamos a hacer una nueva columna de la base de datos y me voy  quedar con
#Solo la fecha sin la hora.

datos <- datos %>% mutate(fecha=date(fecha_nueva))

#Ahora queremos agrupar por fecha y contar

conteo_delitos <- datos %>% group_by(fecha) %>% tally()

#Calculamos la media
mean(conteo_delitos$n)
media <- conteo_delitos %>% summarise(mean(n))
desv.est <- conteo_delitos %>% summarise(sd(n))
x.barra <- mean(conteo_delitos$n)
n.longi<- length(conteo_delitos$n)
  media
  desv.est
```

```
  x.barra
  n.longi
#Calculo del Skewness=Asimetria
coefAsim <- (1/desv.est^3)*mean((conteo_delitos$n - x.barra)^3)
  coefAsim
#mediana
median(conteo_delitos$n)
conteo_delitos%>% summarise(mean(n), median(n), sum(n), var(n), mad(n), sd(n))

#Dia promedio que ocurren los delitos
datos <- datos %>% mutate(`Día de la semana` = weekdays(fecha))

conteo.dia <- datos %>% group_by(`Día de la semana`) %>% tally()
conteo.dia <- conteo.dia %>% mutate(NumDia= NA)
#conteo.dia <- #conteo.dia %>% mutate(NumDia=if_else(`Día de la semana`=="Friday", 5,
NumDia))
conteo.dia$NumDia<- c(5,1,6,7,4,2,3)

#Los estadísticos
mean(conteo_delitos$n)
mad(conteo_delitos$n) #promedio(|x_i - xbarra|)
median(conteo_delitos$n)
quantile(conteo_delitos$n, 0.25)
?var
n <- length(conteo_delitos$n)
(n-1)/n*var(conteo_delitos$n)


cuantil <- quantile(conteo_delitos$n, 0.25)
cuantil[1] + 20
as.numeric(cuantil[1])

library(moments)
conteo_delitos %>% summarise(IQR(n), kurtosis(n))
IQR(conteo_delitos$n)

ggplot(conteo_delitos) +
  geom_boxplot(aes(x = n))

ggplot(conteo_delitos) +
  geom_boxplot(aes(y = n), color = "#3ba17d", fill = "deepskyblue4") +
  labs(
```

```
    title = "Distribución de las carpetas de investigación de delitos en PGJ",
    subtitle = "Datos de CDMX, diciembre 2018",
    x = "",
    y = "Número de carpetas abiertas por día",
    caption = "Datos de la AIP"
  )

conteo_tipo <- datos %>% filter(str_detect(categoria_delito, "ROBO"))
conteo_tipo <- conteo_tipo %>% group_by(delito) %>% tally()

ggplot(conteo_tipo) +
  geom_col(aes(x = delito, y = n)) +
  theme(axis.text.x = element_text(angle = 90, size = 3)) +
  labs(
    title = "Careptas de investigación iniciadas en PGJ por robo",
    subtitle = "Datos de CDMX, diciembre 2018",
    x = "Tipo de robo",
    y = "Conteo de casos",
    caption = "Datos de la AIP"
  )

setwd("/cloud/project")
ggsave("Grafica_barras.pdf", width = 6, height = 4)

ggplot(datos) +
  geom_point(aes(x = longitud, y = latitud), size = 1,
             color = "purple", alpha = 0.1) + #Alpha = transparencia de 0 a 1
  theme_minimal()

ggplot(conteo_delitos) +
  geom_line(aes(x = fecha, y = n)) +
  geom_point(aes(x = fecha, y = n), color = "red") +
  theme_classic() +
  geom_label(aes(x = dmy("25/12/2018"), y = 425),
             label = "Navidad") +
  labs(
    x = "Fecha de aperturas de la carpeta",
    y = "Número de carpetas abiertas",
    title = "Carpetas de investigación abiertas en PGJ",
    subtitle = "Datos abiertos CDMX, Diciembre 2018",
    caption = "Fuente: AIP"
  )
```

Lab: Inferencia Estadística 27/Ago/2021

ITAM

```
#EJEMPLO
año.primaria <- sample(c(1,2,3), 1000, replace = T)

#No tiene sentido un boxplot con estos datos:
ggplot() +
  geom_boxplot(aes(x = año.primaria)) +
  geom_point(aes(x = año.primaria, y = 0), position = "jitter")

ggplot() +
  geom_bar(aes(x = año.primaria))

#Tabla de contingencias x= Alcaldia, y=Año de inicio
table(datos$alcaldia_hechos, datos$ao_inicio)
addmargins(table(datos$alcaldia_hechos, datos$ao_inicio))

#Tabla de frecuencias
table(datos$alcaldia_hechos, datos$ao_inicio) %>% prop.table() %>% addmargins()


data("mtcars")
datos.coches<- mtcars
  ggplot(datos.coches) +
    geom_point(aes(x=wt, y=mpg), color= "red") + theme_bw()

x<- datos.coches$wt
y<- datos.coches$mpg

#Calculamos el coseno con producto punto/ norma
  # %*% es la multiplicacion de HADAMARD

producto.punto <- t(x) %*% y
  coseno <- producto.punto/(sqrt(sum(x^2))*sqrt(sum(y^2)))
  coseno

#Un coseno de 1 significaria paralelos i.e que la relación entre x, y es LINEAL.
#Un coseno de 0 significa que son ortogonales, no podemos describir de manera
  #lineal la relación x,y

#Esto es el coseno pero de x- media, y-media (centralizado)
cor(x, y, method="pearson")
```

Lab: Inferencia Estadística 27/Ago/2021

ITAM

```
cor(x,y, method = "spearman")

calidad_alimentos <- factor(c("malo", "bueno", "bueno", "regular", "bueno",
                              "bueno"), order= TRUE, levels= c("malo", "regular",
                                                               "bueno"))
calidad_servicio <- factor(c("1 estrella", "4 estrellas", "5 estrellas", "2 estrellas", "5 estrellas",
                             "4 estrellas"), order= TRUE, levels= c("1 estrella", "2 estrellas",
                                                                    "3 estrellas", "4 estrellas", "5 estrellas"))
# Falta comentar esto
ggplot()+geom_point(aes(x=calidad_alimentos, y=calidad_servicio,
                        size= calidad_alimentos, color=calidad_servicio))

cor(as.numeric(calidad_alimentos), as.numeric(calidad_servicio), method= "spearman")
cor(as.numeric(calidad_alimentos), as.numeric(calidad_servicio), method= "kendall")

#Corremos la regresión lineal
ggplot(datos.coches)+ geom_point(aes(x= wt, y= mpg))+
                      geom_smooth(aes(x= wt, y= mpg), method= "lm",
                      formula= y ~ x, se=FALSE)+ theme_minimal()

hora.dia<- factor(c("MedioDía", "Anochecer", "Medianoche", "Amanecer"), order=TRUE, levels
= c("Amanecer", "MedioDía", "Anochecer", "Medianoche"))
cantidad.aves<-factor(c(200,100,0,100), order= TRUE)
cor(as.numeric(hora.dia), as.numeric(cantidad.aves), method= "spearman")

ggplot()+geom_point(aes(x=hora.dia, y=cantidad.aves,
                        size= hora.dia, color=cantidad.aves))
a<-c(-1/2,1/2,3/2,-3/2)
b<-c(7/4,-1/4,-5/4,-1/4)
producto.punto1 <- t(a) %*% b
producto.punto1
(sqrt(sum(a^2))*sqrt(sum(b^2)))
```

Lab: Inferencia Estadística 27/Ago/2021

Crear Bases de Datos:

```
library(readr)
library(tidyverse)
library(lubridate)

nsim<-1000 #Cantidad de simulaciones
N<- 10000 #Tamaño de población
n<- 100 #tamaño de la muestra por cada simulación

base.datos<- data.frame(x=rpois(N,lambda= 4))#Esta es la U en nuestra notacion
media.U <-mean(base.datos$x)

media.muestra<- rep(NA,nsim)
#Ciclo a través de las 1000 simulaciones
for(i in 1:nsim){
  muestra<- sample(base.datos$x, n, replace=FALSE)#Muestreo simple sin remplazo
media.muestra[i] <- mean(muestra)}#cada uno de estos es un xbarra}

mean(media.muestra)
media.U


ggplot()+
  geom_histogram(aes(x=media.muestra, y=..density..), fill="red")+
  geom_vline(aes(xintercept=media.U), linetype= "dashed")

#Se puede ver que estamos aproximando demasiado bien la media de U
#Ahora vamos a calcular la varianza de la xbarra_s
var.xbarra<- ((1-n/N)/n)*var(base.datos$x)

var.muestral.xbarra<- rep(NA,nsim)
for(i in 1:nsim){
  muestra <- sample(base.datos$x, n, replace=FALSE)
  var.muestral.xbarra[i]<- ((1-n/N)/n)*var(muestra)
}

ggplot()+
  geom_density(aes(x=var.muestral.xbarra), color="red")+
  geom_vline(aes(xintercept=var.xbarra))

x.data<-c(1,2,3)
datos.nuevos<- c()
for (i in 1:length(x.data)){
```

```
  datos.nuevos<- c(datos.nuevos, x[i]^i)
}
mean(datos.nuevos)
```

DENLE UNA PASADA A:

- r-coder.com
- cookbook-r.com
- kaggle.com
- rpubs.com

- Cursos de R que creo que están buenos:
- Curso UNAM, antes era gratuito, ahora ya no sé: https://www.coursera.org/learn/intro-data-science-programacion-estadistica-r (Enlaces a un sitio externo.)
- En particular, a mi me gustan los cursos de KIrill Emerenko en Udemy:
    - Básico: https://www.udemy.com/course/r-programming/ (Enlaces a un sitio externo.)
    - Intermedio: https://www.udemy.com/course/r-analytics/

Lab: Inferencia Estadística 27/Ago/2021