# GUÍA DE CÓDIGO EN R PARA EL EXAMEN FINAL

# Esta guía contiene ejemplos de código para realizar cada parte del examen

# PARTE I: PREPARACIÓN Y ANÁLISIS EXPLORATORIO

# 1. Importación y exploración de datos

# Importar los archivos CSV

datos\_sin <- read.csv("sin\_estudios.csv")

datos\_con <- read.csv("con\_estudios.csv")

# Mostrar estructura y resumen

str(datos\_sin)

summary(datos\_sin)

str(datos\_con)

summary(datos\_con)

# 2. Estadísticas descriptivas básicas

# Total de datos

n\_con <- length(datos\_con$sueldo)

n\_sin <- length(datos\_sin$sueldo)

# Datos mayor y menor

dm\_con <- max(datos\_con$sueldo)

dn\_con <- min(datos\_con$sueldo)

dm\_sin <- max(datos\_sin$sueldo)

dn\_sin <- min(datos\_sin$sueldo)

# Rango

rango\_con <- dm\_con - dn\_con

rango\_sin <- dm\_sin - dn\_sin

# Número de clases

k\_con <- 1 + 3.322 \* log10(n\_con)

k\_con <- ceiling(k\_con) # Redondea hacia arriba

k\_sin <- 1 + 3.322 \* log10(n\_sin)

k\_sin <- ceiling(k\_sin) # Redondea hacia arriba

# Intervalo de clase

i\_con <- rango\_con / k\_con

i\_sin <- rango\_sin / k\_sin

# PARTE II: DISTRIBUCIÓN DE FRECUENCIAS

# 3. Tabla de frecuencias para persona con estudios

# Crear límites de clase

limites\_con <- seq(from = dn\_con, to = dm\_con, by = i\_con)

limites\_inf\_con <- limites\_con[-length(limites\_con)]

limites\_sup\_con <- limites\_con[-1]

# Límites aparentes

lim\_aparentes\_con <- data.frame(

Li = limites\_inf\_con,

Ls = limites\_sup\_con

)

# Límites reales

lim\_reales\_con <- data.frame(

Li\_real = limites\_inf\_con - 0.5,

Ls\_real = limites\_sup\_con + 0.5

)

# Marca de clase

marca\_clase\_con <- (limites\_inf\_con + limites\_sup\_con) / 2

# Frecuencia absoluta

# Usamos el paquete dplyr para facilitar la tarea

library(dplyr)

# Construir la tabla de frecuencias completa

tabla\_freq\_con <- data.frame(

Li = limites\_inf\_con,

Ls = limites\_sup\_con,

Li\_real = limites\_inf\_con - 0.5,

Ls\_real = limites\_sup\_con + 0.5,

Xi = marca\_clase\_con

)

# Calcular frecuencias

for (i in 1:nrow(tabla\_freq\_con)) {

tabla\_freq\_con$f[i] <- sum(datos\_con$sueldo >= tabla\_freq\_con$Li[i] &

datos\_con$sueldo < tabla\_freq\_con$Ls[i])

}

# Ajustar la última clase para incluir el valor máximo

tabla\_freq\_con$f[nrow(tabla\_freq\_con)] <- sum(datos\_con$sueldo >= tabla\_freq\_con$Li[nrow(tabla\_freq\_con)] &

datos\_con$sueldo <= tabla\_freq\_con$Ls[nrow(tabla\_freq\_con)])

# Frecuencia relativa

tabla\_freq\_con$fr <- tabla\_freq\_con$f / n\_con

# Frecuencia acumulada

tabla\_freq\_con$fa <- cumsum(tabla\_freq\_con$f)

# Porcentaje

tabla\_freq\_con$porcentaje <- tabla\_freq\_con$fr \* 100

# Repetir el proceso para persona sin estudios

# Crear límites de clase

limites\_sin <- seq(from = dn\_sin, to = dm\_sin, by = i\_sin)

limites\_inf\_sin <- limites\_sin[-length(limites\_sin)]

limites\_sup\_sin <- limites\_sin[-1]

# Construir la tabla de frecuencias completa

tabla\_freq\_sin <- data.frame(

Li = limites\_inf\_sin,

Ls = limites\_sup\_sin,

Li\_real = limites\_inf\_sin - 0.5,

Ls\_real = limites\_sup\_sin + 0.5,

Xi = (limites\_inf\_sin + limites\_sup\_sin) / 2

)

# Calcular frecuencias

for (i in 1:nrow(tabla\_freq\_sin)) {

tabla\_freq\_sin$f[i] <- sum(datos\_sin$sueldo >= tabla\_freq\_sin$Li[i] &

datos\_sin$sueldo < tabla\_freq\_sin$Ls[i])

}

# Ajustar la última clase para incluir el valor máximo

tabla\_freq\_sin$f[nrow(tabla\_freq\_sin)] <- sum(datos\_sin$sueldo >= tabla\_freq\_sin$Li[nrow(tabla\_freq\_sin)] &

datos\_sin$sueldo <= tabla\_freq\_sin$Ls[nrow(tabla\_freq\_sin)])

# Frecuencia relativa

tabla\_freq\_sin$fr <- tabla\_freq\_sin$f / n\_sin

# Frecuencia acumulada

tabla\_freq\_sin$fa <- cumsum(tabla\_freq\_sin$f)

# Porcentaje

tabla\_freq\_sin$porcentaje <- tabla\_freq\_sin$fr \* 100

# Repetir el proceso para persona sin estudios

# [código similar al anterior pero con la variable sin\_estudios]

# PARTE III: MEDIDAS DE TENDENCIA CENTRAL

# 5. Media para ambas variables

media\_con <- sum(tabla\_freq\_con$Xi \* tabla\_freq\_con$f) / n\_con

media\_sin <- sum(tabla\_freq\_sin$Xi \* tabla\_freq\_sin$f) / n\_sin

# 6. Moda para con estudios

# Identificar la clase modal

clase\_modal\_con <- which.max(tabla\_freq\_con$f)

Li\_modal\_con <- tabla\_freq\_con$Li[clase\_modal\_con]

i\_modal\_con <- i\_con

# Frecuencias necesarias

if (clase\_modal\_con > 1) {

f\_anterior\_con <- tabla\_freq\_con$f[clase\_modal\_con - 1]

} else {

f\_anterior\_con <- 0

}

if (clase\_modal\_con < nrow(tabla\_freq\_con)) {

f\_posterior\_con <- tabla\_freq\_con$f[clase\_modal\_con + 1]

} else {

f\_posterior\_con <- 0

}

f\_modal\_con <- tabla\_freq\_con$f[clase\_modal\_con]

# Calcular delta1 y delta2

delta1\_con <- f\_modal\_con - f\_anterior\_con

delta2\_con <- f\_modal\_con - f\_posterior\_con

# Calcular la moda

moda\_con <- Li\_modal\_con + ((delta1\_con / (delta1\_con + delta2\_con)) \* i\_modal\_con)

# Repetir para sin estudios

# [código similar al anterior]

# 7. Mediana para con estudios

# Posición de la mediana

pos\_mediana\_con <- n\_con / 2

# Encontrar la clase mediana

for (i in 1:nrow(tabla\_freq\_con)) {

if (tabla\_freq\_con$fa[i] >= pos\_mediana\_con) {

clase\_mediana\_con <- i

break

}

}

# Parámetros para el cálculo

Li\_mediana\_con <- tabla\_freq\_con$Li[clase\_mediana\_con]

fa\_anterior\_con <- if (clase\_mediana\_con > 1) tabla\_freq\_con$fa[clase\_mediana\_con - 1] else 0

f\_mediana\_con <- tabla\_freq\_con$f[clase\_mediana\_con]

# Calcular la mediana

mediana\_con <- Li\_mediana\_con + ((pos\_mediana\_con - fa\_anterior\_con) / f\_mediana\_con) \* i\_con

# Repetir para sin estudios

# [código similar al anterior]

# PARTE IV: MEDIDAS DE DISPERSIÓN

# 8. Desviación respecto a la media para con estudios

tabla\_freq\_con$xi\_menos\_media <- tabla\_freq\_con$Xi - media\_con

tabla\_freq\_con$abs\_xi\_menos\_media <- abs(tabla\_freq\_con$xi\_menos\_media)

tabla\_freq\_con$f\_abs\_xi\_menos\_media <- tabla\_freq\_con$f \* tabla\_freq\_con$abs\_xi\_menos\_media

suma\_f\_abs\_xi\_menos\_media\_con <- sum(tabla\_freq\_con$f\_abs\_xi\_menos\_media)

# Repetir para sin estudios

# [código similar al anterior]

# 9. Varianza y desviación estándar para con estudios

tabla\_freq\_con$xi\_menos\_media\_cuadrado <- tabla\_freq\_con$xi\_menos\_media^2

tabla\_freq\_con$f\_xi\_menos\_media\_cuadrado <- tabla\_freq\_con$f \* tabla\_freq\_con$xi\_menos\_media\_cuadrado

varianza\_con <- sum(tabla\_freq\_con$f\_xi\_menos\_media\_cuadrado) / n\_con

desv\_estandar\_con <- sqrt(varianza\_con)

# Repetir para sin estudios

# [código similar al anterior]

# 10. Coeficiente de variación

cv\_con <- (desv\_estandar\_con / media\_con) \* 100

cv\_sin <- (desv\_estandar\_sin / media\_sin) \* 100

# PARTE V: ASIMETRÍA Y CURTOSIS

# 11. Coeficiente de asimetría para con estudios

tabla\_freq\_con$xi\_menos\_media\_cubo <- tabla\_freq\_con$xi\_menos\_media^3

tabla\_freq\_con$f\_xi\_menos\_media\_cubo <- tabla\_freq\_con$f \* tabla\_freq\_con$xi\_menos\_media\_cubo

sumatoria\_f\_xi\_menos\_media\_cubo\_con <- sum(tabla\_freq\_con$f\_xi\_menos\_media\_cubo)

asimetria\_con <- sumatoria\_f\_xi\_menos\_media\_cubo\_con / (n\_con \* desv\_estandar\_con^3)

# Repetir para sin estudios

# [código similar al anterior]

# 12. Coeficiente de curtosis para con estudios

tabla\_freq\_con$xi\_menos\_media\_cuarta <- tabla\_freq\_con$xi\_menos\_media^4

tabla\_freq\_con$f\_xi\_menos\_media\_cuarta <- tabla\_freq\_con$f \* tabla\_freq\_con$xi\_menos\_media\_cuarta

sumatoria\_f\_xi\_menos\_media\_cuarta\_con <- sum(tabla\_freq\_con$f\_xi\_menos\_media\_cuarta)

curtosis\_con <- sumatoria\_f\_xi\_menos\_media\_cuarta\_con / (n\_con \* desv\_estandar\_con^4)

# Interpretación

if (curtosis\_con == 3) {

interpretacion\_curtosis\_con <- "mesocúrtica"

} else if (curtosis\_con < 3) {

interpretacion\_curtosis\_con <- "platicúrtica"

} else {

interpretacion\_curtosis\_con <- "leptocúrtica"

}

# Repetir para sin estudios

# [código similar al anterior]

# PARTE VI: REPRESENTACIÓN GRÁFICA

# 13. Gráficos

# Histograma para ambas variables

par(mfrow=c(1,2)) # Dividir el área de gráficos en 1 fila y 2 columnas

# Histograma para con estudios

hist(datos\_con$sueldo,

breaks = k\_con,

main = "Histograma de sueldos con estudios",

xlab = "Sueldo",

ylab = "Frecuencia",

col = "lightblue",

border = "black")

# Histograma para sin estudios

hist(datos\_sin$sueldo,

breaks = k\_sin,

main = "Histograma de sueldos sin estudios",

xlab = "Sueldo",

ylab = "Frecuencia",

col = "lightgreen",

border = "black")

# Polígono de frecuencias

par(mfrow=c(1,2))

# Polígono para con estudios

plot(tabla\_freq\_con$Xi, tabla\_freq\_con$f,

type = "o",

main = "Polígono de frecuencias - Con estudios",

xlab = "Marca de clase",

ylab = "Frecuencia",

col = "blue",

pch = 19)

# Polígono para sin estudios

plot(tabla\_freq\_sin$Xi, tabla\_freq\_sin$f,

type = "o",

main = "Polígono de frecuencias - Sin estudios",

xlab = "Marca de clase",

ylab = "Frecuencia",

col = "red",

pch = 19)

# Ojiva (frecuencia acumulada)

par(mfrow=c(1,2))

# Ojiva para con estudios

plot(tabla\_freq\_con$Xi, tabla\_freq\_con$fa,

type = "o",

main = "Ojiva - Con estudios",

xlab = "Marca de clase",

ylab = "Frecuencia acumulada",

col = "purple",

pch = 19)

# Ojiva para sin estudios

plot(tabla\_freq\_sin$Xi, tabla\_freq\_sin$fa,

type = "o",

main = "Ojiva - Sin estudios",

xlab = "Marca de clase",

ylab = "Frecuencia acumulada",

col = "orange",

pch = 19)

# Gráfico circular

par(mfrow=c(1,2))

# Circular para con estudios

pie(tabla\_freq\_con$f,

labels = paste(round(tabla\_freq\_con$porcentaje, 1), "%"),

main = "Porcentajes de sueldos - Con estudios",

col = rainbow(nrow(tabla\_freq\_con)))

# Circular para sin estudios

pie(tabla\_freq\_sin$f,

labels = paste(round(tabla\_freq\_sin$porcentaje, 1), "%"),

main = "Porcentajes de sueldos - Sin estudios",

col = rainbow(nrow(tabla\_freq\_sin)))

# Diagrama de cajas y bigotes comparativo

par(mfrow=c(1,1)) # Restablecer el área de gráficos

boxplot(datos\_con$sueldo, datos\_sin$sueldo,

names = c("Con estudios", "Sin estudios"),

main = "Comparación de sueldos",

ylab = "Sueldo ($)",

col = c("lightblue", "lightgreen"),

notch = TRUE) # El notch permite visualizar si hay diferencias significativas