Trabajo 1

Ivan Santiago Rojas Martinez

Estudiante de Pregrado en Estadística

Docente

Rene Iral Palomino

Asignatura

Introducción al Análisis Multivariado

Sede Medellín
Septiembre 2 de 2023

Índice

Índice de figuras

Índice de cuadros

Trabajo 1

Selección de la muestra de datos

Se incluye el código propuesto por el docente, con la intención de validar la extracción de la muestra

```
library(splitstackshape)
uno <- read.table("Data/data.txt", header=T, sep=",")

genera <- function(cedula){
set.seed(cedula)
aux <- stratified(uno, "CAT_IMC", 200/2100, bothSets=T)
mue <- aux$SAMP1
mue
}

data <- genera(1020479466)
```

1 Primer Punto

Para todas sus variables realice un análisis exploratorio gráfico e identifique posibles valores atípicos u otro tipo de anomalías. (Para las variables Categóricas diagramas de barras, para las continuas o discretas, use Histogramas y/o Boxplot). Comente brevemente.

1.1 Análisis Descriptivos

Breve descripción de la base de datos: La base de datos corresponde a las medidas antropométricas de la población laboral colombiana (ACOPLA). Esta base de datos cuenta con 200 observaciones y 9 variables de interés, las cuales son:

-**Sexo:** Variable categórica (Hom, Muj)

-**P1: Masa Corporal** Variable continua (kg)

-**P7: Perímetro muslo mayor** Variable continua (cm)

-**P16: Perímetro abdominal cintura** Variable continua (cm)

-**P22: Anchura de las caderas** Variable continua (cm)

-**P27: Longitud promedio de los pies** Variable continua (cm)

-**P29: Longitud promedio de las manos** Variable continua (cm)

-**P38: Estatura** Variable continua (cm)

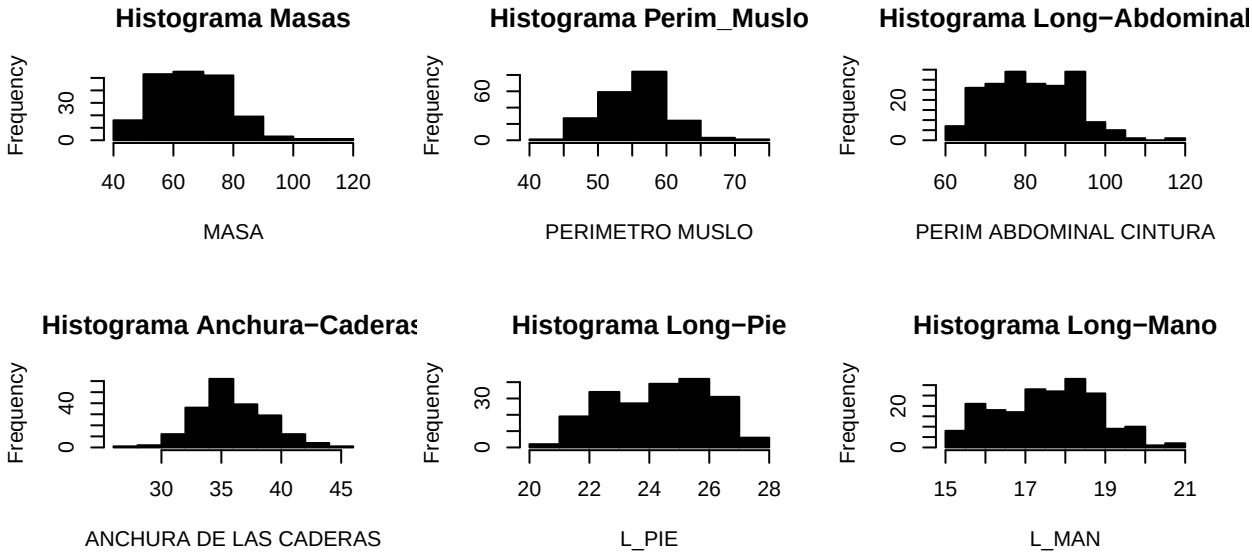
-**CAT_IMC: Categoría del indice de masa corporal** Variable categórica (DELGADO, NORMAL Y OBESO)

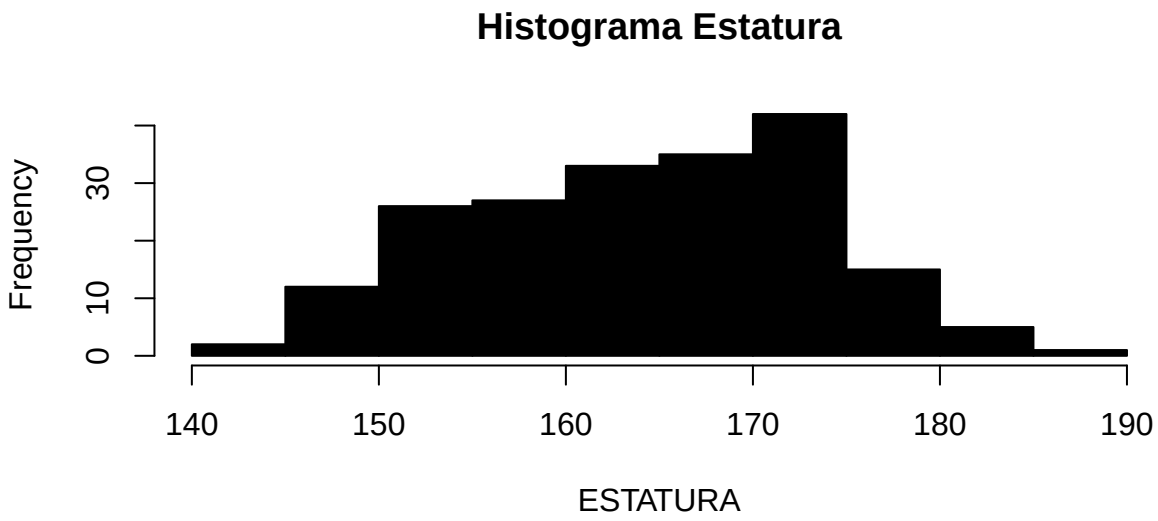
1.2 Resumen Numerico

Cuadro 1: Tabla de resúmenes estadísticos de variables continuas

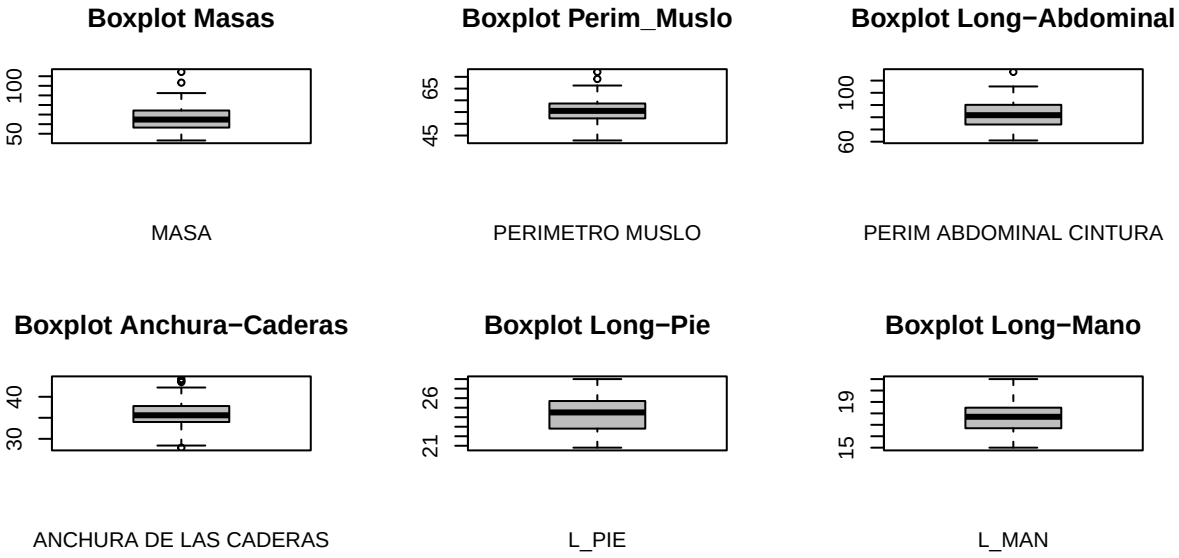
Variable	Media	Mediana	SD	Q1	Q2	Q3	Rango.intercuartil	Rango
P1	66.11600	64.80	12.140225	56.475	64.80	74.225	17.75	71.5
P7	55.47688	55.50	4.741592	52.300	55.50	58.650	6.35	29.2
P16	81.65200	81.70	10.163656	74.125	81.70	90.075	15.95	56.1
P22	35.87626	35.60	2.963242	34.000	35.60	37.800	3.80	16.3
P27	24.33600	24.50	1.681380	22.800	24.50	25.700	2.90	7.2
P29	17.63400	17.70	1.246480	16.700	17.70	18.500	1.80	6.0
P38	164.14949	164.95	9.178836	156.700	164.95	171.450	14.75	43.6

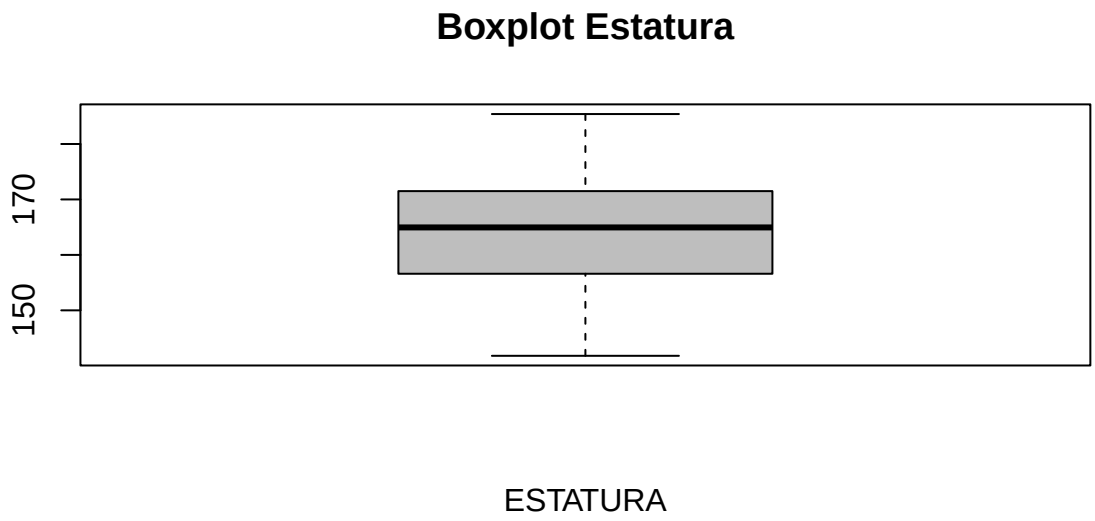
1.3 Histogramas



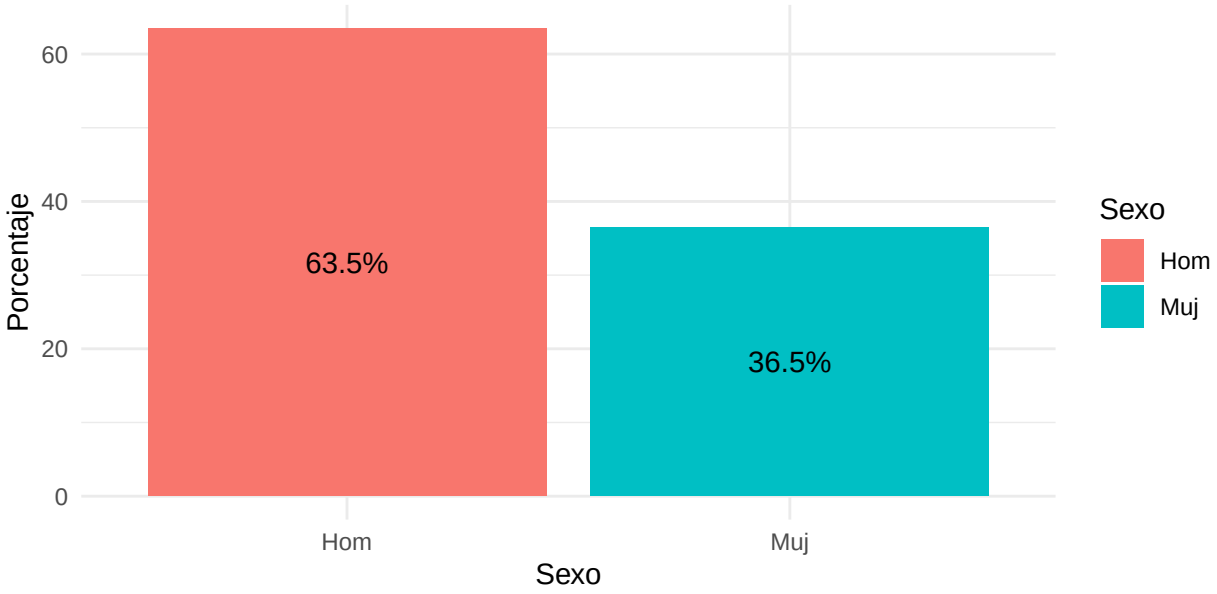

Histograma Estatura
Frequency
0
10
30
140
150
160
170
180
190
ESTATURA

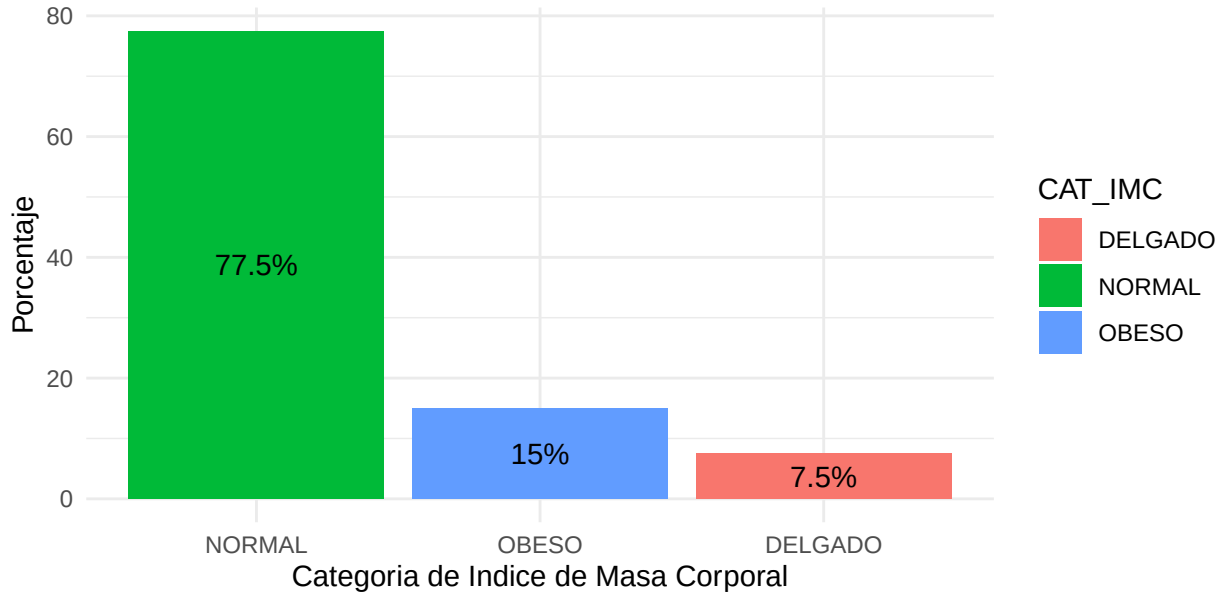
1.4 Boxplots


Boxplot Masas
50
100
MASA
Boxplot Perim_Muslo
45
65
PERIMETRO MUSLO
Boxplot Long−Abdominal
60
100
PERIM ABDOMINAL CINTURA
Boxplot Anchura−Caderas
30
40
ANCHURA DE LAS CADERAS
Boxplot Long−Pie
21
26
L_PIE
Boxplot Long−Mano
15
19
L_MAN



1.5 Diagrama de Barras



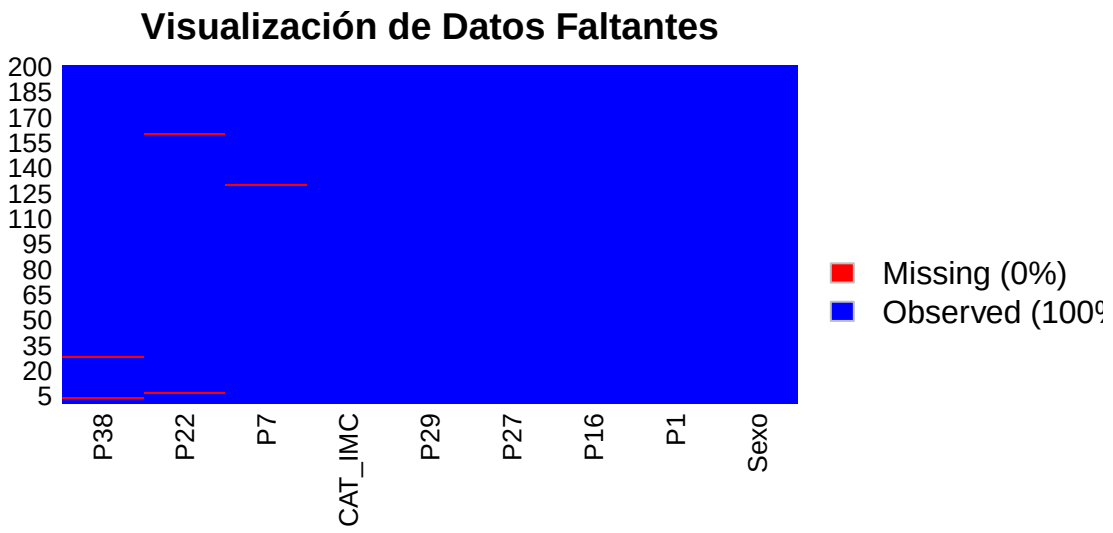


2 Segundo Punto

Realice el respectivo proceso de imputación para los datos faltantes en su base de datos. Explique cómo realiza dicha imputación, cuál criterio utiliza y muestre un par de ejemplos ilustrativos.

2.1 Grafico de datos faltantes

Procederemos realizando un gráfico de datos faltantes, el cual nos permitirá determinar el porcentaje de datos ausentes en cada variable.



Se puede observar que el porcentaje de datos faltantes es aproximadamente 0. Esto nos indica que la base de datos no presenta muchos problemas con valores faltantes (missing o NA). Sin embargo, las variables P38, P22 y P7 contienen datos faltantes, los cuales son:

- **P38:** observación 173 y 197.
- **P22:** observación 41 y 194.
- **P7:** observación 71.

2.2 Imputación de datos faltantes

Como se observó en el análisis descriptivo previo, el género parece ser un factor discriminante en las variables de estatura (P38), anchura de las caderas (P22) y el perímetro del muslo mayor (P7), indicando medidas promedio mayores o menores dependiendo del genero. Por lo tanto, el criterio de imputación de datos se basara en el promedio de la variable respecto al genero de la observación que cuenta con un dato faltante en alguna de las anterios variables.

Dos ejemplos de como se realizo la imputación de los datos para la variable P38

- Para la observación **173** que tiene un valor faltante en la variable *P38* (Estatura) se procede a calcular el promedio para hombres y mujeres los cuales son: 169.3024 cm y 155.1319 cm respectivamente. Como la observación **173** es una **mujer** el valor a imputar es **155.1319**
- Para la observación **197** que tiene un valor faltante en la variable *P38* (Estatura) se procede a calcular el promedio para hombres y mujeres los cuales son: 169.3024 cm y 155.1319 cm respectivamente. Como la observación **197** es un **hombre** el valor a imputar es **169.3024**

De esta manera se imputan los datos faltantes en las demás variables.

Se anexa el código usado en R para hacer la imputación de datos.

```
data <- data %>%
  group_by(Sexo) %>%
  mutate(P38 = ifelse(is.na(P38), mean(P38, na.rm = TRUE), P38),
         P22 = ifelse(is.na(P22), mean(P22, na.rm = TRUE), P22),
         P7 = ifelse(is.na(P7), mean(P7, na.rm = TRUE), P7))
```

3 Tercer Punto

Considere las variables P1, P29 y P38. ¿Se puede afirmar que cada variable por separado permitiría discriminar entre Hombres y Mujeres? Elabore los resúmenes numéricos y gráficos que considere pertinentes para responder la pregunta.

3.1 Resumen Numerico

Cuadro 2: Resumen Numerico para P1

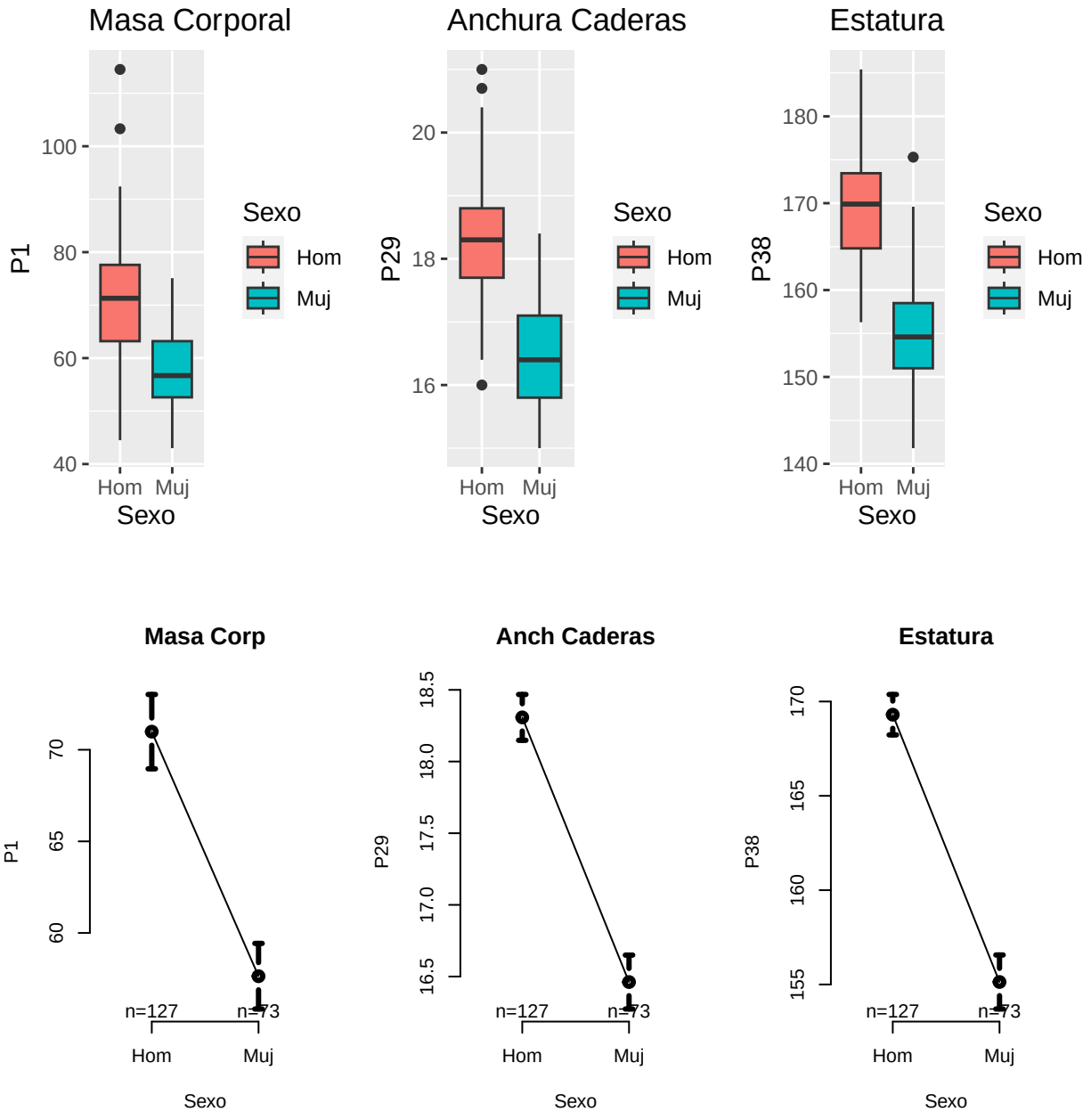
Sexo	Promedio	DesviacionEstandar	Mediana
Hom	70.98898	11.557182	71.3
Muj	57.63836	7.671857	56.7

Cuadro 3: Resumen Numerico para P29

Sexo	Promedio	DesviacionEstandar	Mediana
Hom	18.30787	0.9103990	18.3
Muj	16.46164	0.8058174	16.4

Cuadro 4: Resumen Numerico para P38

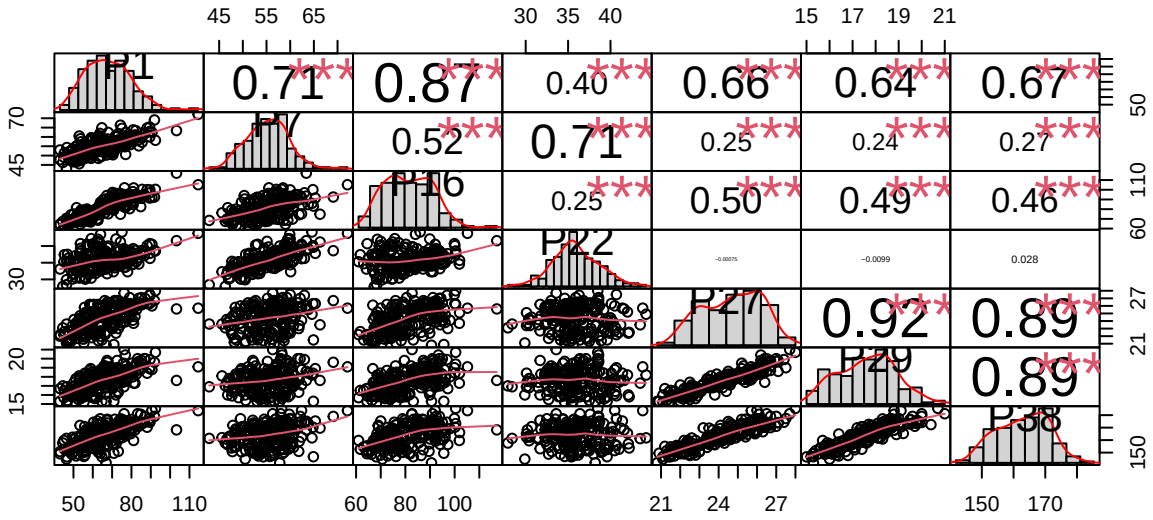
Sexo	Promedio	DesviacionEstandar	Mediana
Hom	169.3024	6.098132	169.9
Muj	155.1319	6.136770	154.6



4 Cuarto Punto

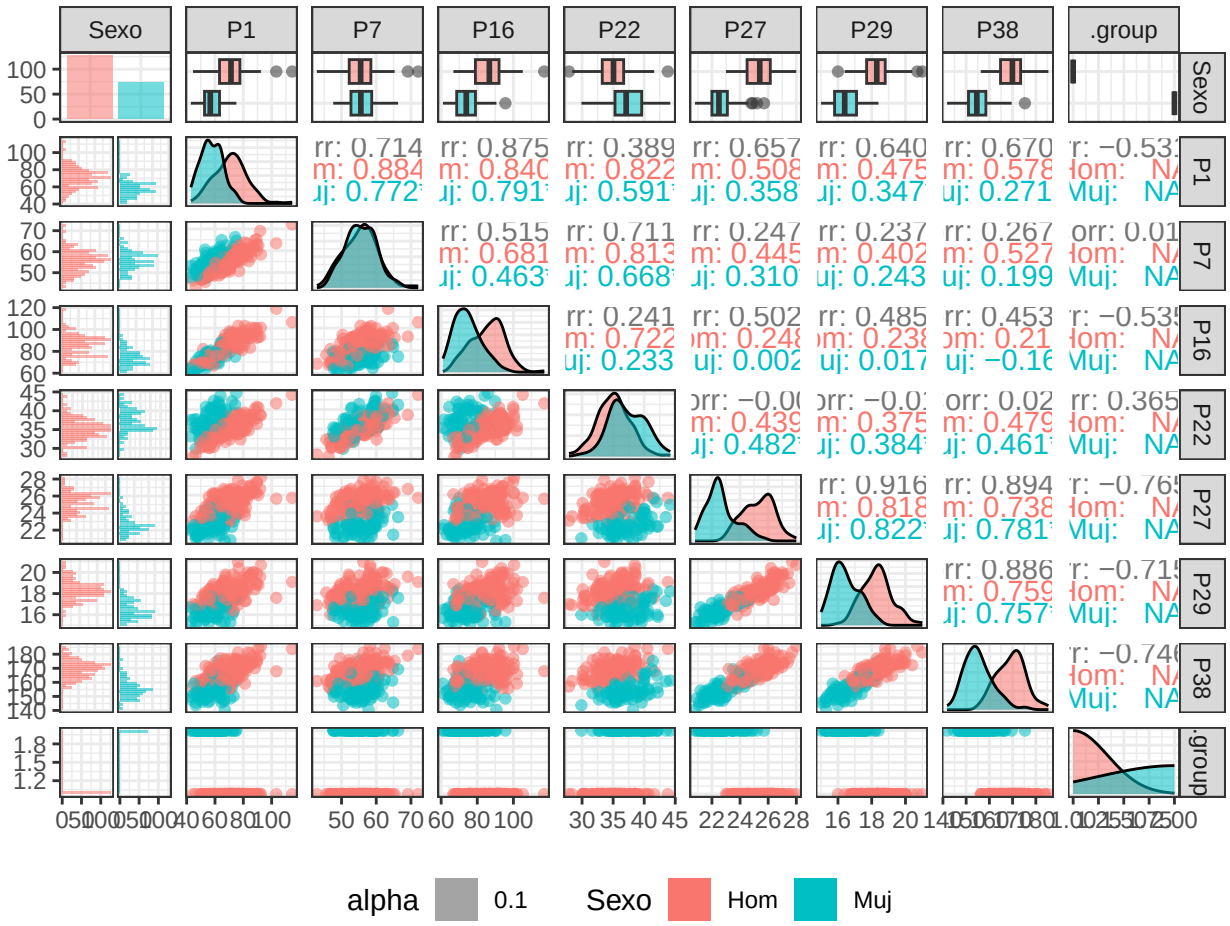
Usando las variables continuas, realice un gráfico de dispersión para identificar posibles relaciones entre sus variables. Explique si lo que se observa gráficamente tiene sentido o es coherente a la luz de sus datos. Corrobore lo observado con el cálculo de la matriz de correlaciones. Comente. Repita el proceso discriminando por SEXO. ¿Hay cambios en las estructuras de Covarianzas para ambos grupos? Comente

4.1 Relaciones entre variables



Cuadro 5: Matrix de correlaciones

	P1	P7	P16	P22	P27	P29	P38
P1	1.0000000	NA	0.8748881	NA	0.6570800	0.6397675	NA
P7	NA	1	NA	NA	NA	NA	NA
P16	0.8748881	NA	1.0000000	NA	0.5020976	0.4852569	NA
P22	NA	NA	NA	1	NA	NA	NA
P27	0.6570800	NA	0.5020976	NA	1.0000000	0.9155520	NA
P29	0.6397675	NA	0.4852569	NA	0.9155520	1.0000000	NA
P38	NA	NA	NA	NA	NA	NA	1



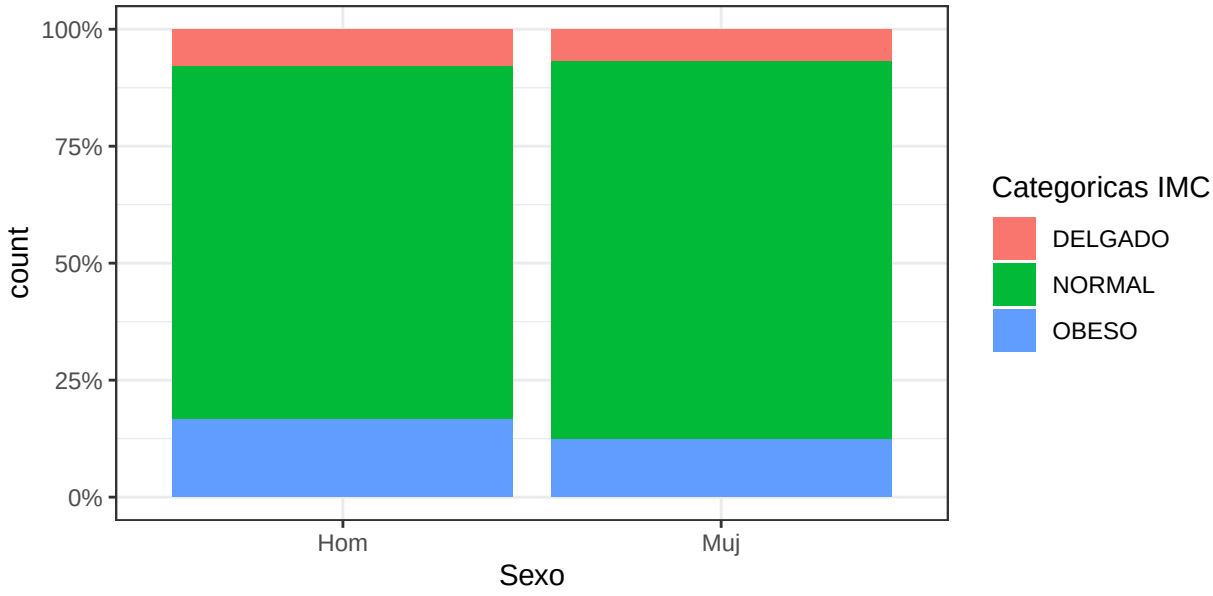
5 Quinto Punto

Elabore una tabla de porcentajes de doble entrada con las variables CAT_IMC y SEXO. Luego presente la información gráficamente. ¿Se puede afirmar que la distribución porcentual de la variable CAT_IMC es diferente para hombre y mujeres? Justifique su respuesta.

5.1 Distribución porcentual y tabla de contingencia.

Cuadro 6: Tabla de contingencia porcentual de doble entrada.

	Hom	Muj	Sum
DELGADO	5.0	2.5	7.5
NORMAL	48.0	29.5	77.5
OBESO	10.5	4.5	15.0
Sum	63.5	36.5	100.0



6 Sexto Punto

Se tienen los siguientes datos de 5 personas, de las cuales se desconoce su CAT_IMC.

P1	P7	P16	P22	P27	P29	P38	CAT_IMC
66.1	53.9	73.8	34.7	27.6	20.9	181.6	
55.8	50.1	76.9	39.5	24.7	17.3	154.5	
62.8	54.3	80.4	37.5	23.5	16.5	156.6	
63.9	50.6	75.6	31.5	24.9	18.6	173.1	
50.7	46.3	72.7	30.4	23.5	16.7	159.5	

Usando la distancia estadística, determine a cuál de las tres categorías pertenece cada sujeto. Explique claramente el proceso empleado para clasificar los sujetos.

Anexe el código empleado.

6.1 Clasificación usando la distancia estadística