

## PRÁCTICA CALIFICADA N°1

**Curso:** Estadística para Ingeniería (EST218)

**Horario:** 0504

**Profesor:** Silvestre Valer, Jim Roland

**Nota: 18**

Nombres y Apellidos:	Correo Electrónico:	Código:
Axel Cárdenas Avellaneda	a20201941@pucp.edu.pe	20201941
Iván Aráoz Andrade	i.araoz@pucp.edu.pe	20201216
Carlos Camilo Vásquez Morales	a20202583@pucp.edu.pe	20202583

### Pregunta 1

Primero, se abre la librería DescTools para visualizar las tablas de frecuencias de fumadores y no fumadores en los géneros Masculino y Femenino. Entonces, procedemos a utilizar la función Freq.

```
library(DescTools)
Freq(d$smoke)

Freq(d$smoke[d$gender=="M"]) #Tabla de Frecuencias de fumadores
                               #y no fumadores Masculinos

Freq(d$smoke[d$gender=="F"]) #Tabla de Frecuencias de fumadores
                               #y no fumadores Femeninos
```

Obteniendo los siguientes resultados:

```
> library(DescTools)
> Freq(d$smoke)
  level  freq  perc cumfreq cumperc
1   No 62'325 91.2% 62'325  91.2%
2   Yes  6'012  8.8%  68'337 100.0%

> Freq(d$smoke[d$gender=="M"]) #Tabla de Frecuencias de fumadores y no fumadores Masculinos
  level  freq  perc cumfreq cumperc
1   No 18'618 78.1% 18'618  78.1%
2   Yes  5'220 21.9% 23'838 100.0%

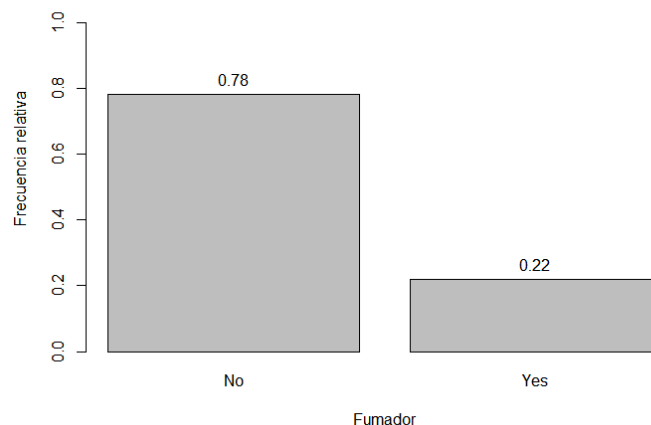
> Freq(d$smoke[d$gender=="F"]) #Tabla de Frecuencias de fumadores y no fumadores Femeninos
  level  freq  perc cumfreq cumperc
1   No 43'707 98.2% 43'707  98.2%
2   Yes   792  1.8% 44'499 100.0%
```

Con los resultados obtenidos podemos observar que hay menos encuestados hombres que mujeres.

Segundo, creamos los diagrama de barras para las dos tablas de frecuencias, porque el atributo de fumador es cualitativo.

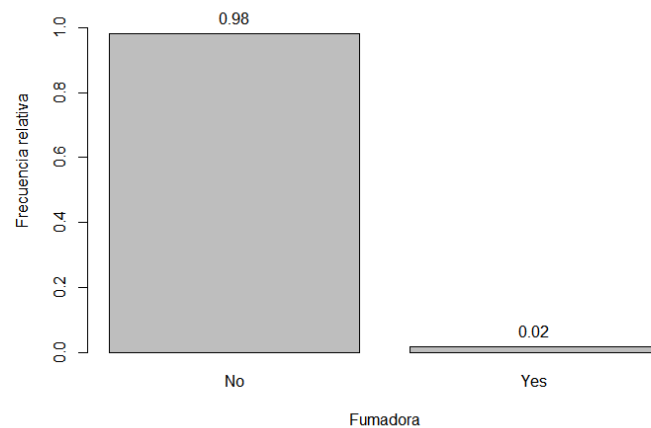
```
#Diagrama de barras para Hombres
f.j = prop.table(table(d$smoke[d$gender=="M"]))
#Gráfico de barras
barplot(f.j,
        main = "Distribución de frecuencias de fumadores y no fumadores Hombres",
        xlab = "Fumador",
        ylab = "Frecuencia relativa",
        ylim = c(0,max(f.j)*1.5))
#Adición de frecuencias
pos.j = barplot(f.j, plot = FALSE)
text(pos.j, f.j,
     labels = round(f.j,2),
     pos = 3)
```

Distribución de frecuencias de fumadores y no fumadores Hombres



```
#Diagrama de barras para Mujeres
f.j = prop.table(table(d$smoke[d$gender=="F"]))
#Gráfico de barras
barplot(f.j,
        main = "Distribución de frecuencias de fumadores y no fumadores Mujeres",
        xlab = "Fumadora",
        ylab = "Frecuencia relativa",
        ylim = c(0,max(f.j)*1.2))
#Adición de frecuencias
pos.j = barplot(f.j, plot = FALSE)
text(pos.j, f.j,
      labels = round(f.j,2),
      pos = 3)
```

Distribución de frecuencias de fumadores y no fumadores Mujeres



Podemos inferir que los hombres tienden a fumar más que las mujeres, y en general la mayoría de las personas no son fumadoras, ya que representan más del 70% en ambas gráficas. Además, al comparar el número de encuestados hombres y mujeres, se puede concluir que las mujeres tuvieron una mayor disposición para compartir sus datos en comparación con los hombres.

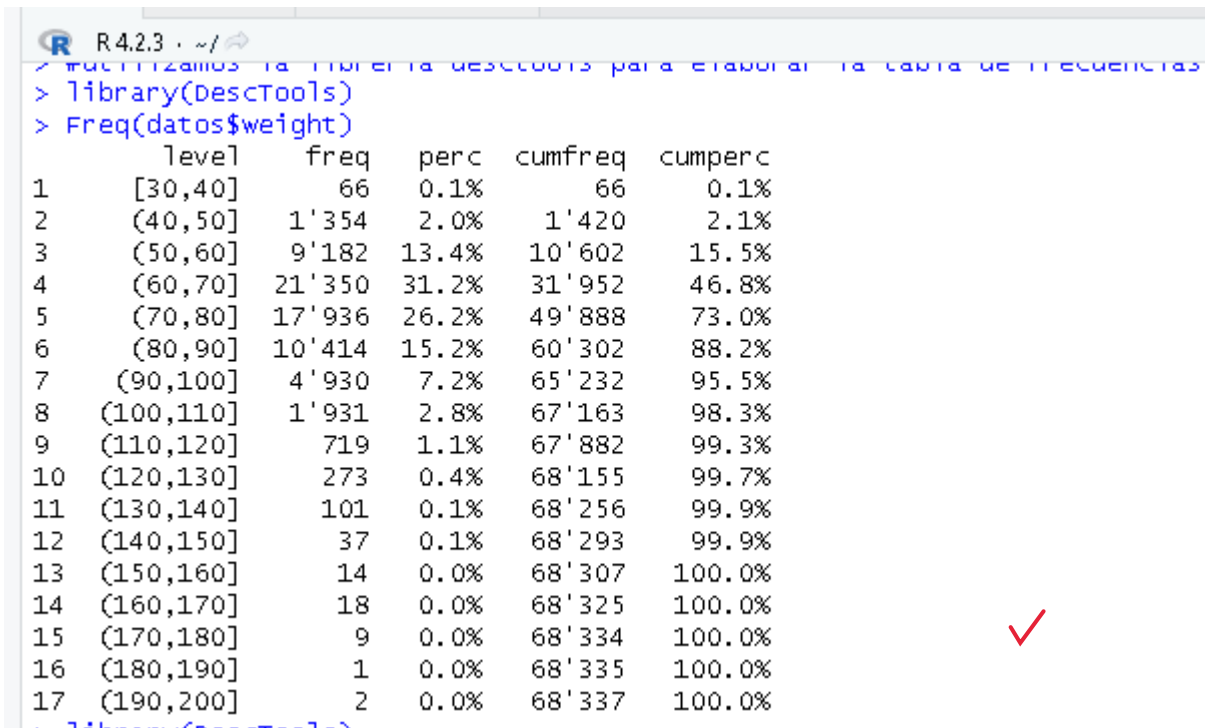
## Pregunta 2:

2) 2.5

Debido a que la variable weight es una variable de tipo cuantitativo continuo.  
Por tanto, el gráfico que se utilizará será el histograma.

```
7 #utilizamos la librería descTools para elaborar la tabla de frecuencias
8 library(DescTools)
9 Freq(datos$weight)
10 |
```

El ejercicio indica explícitamente que usen 14 intervalos.



R 4.2.3 ~/

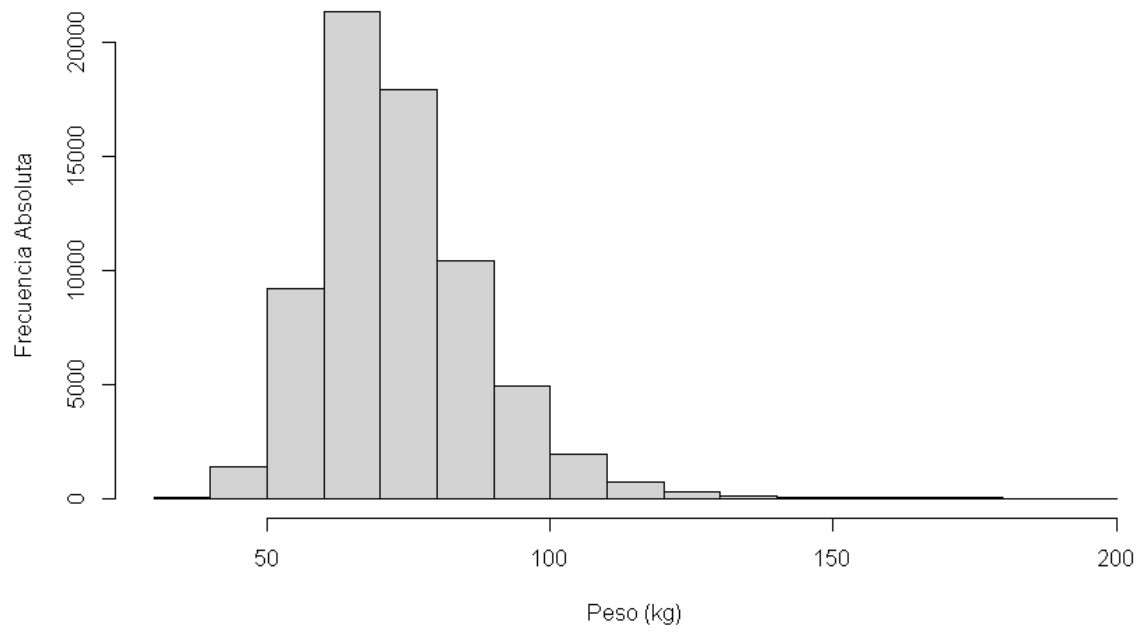
```
> #utilizamos la librería descTools para elaborar la tabla de frecuencias
> library(DescTools)
> Freq(datos$weight)
```

	level	freq	perc	cumfreq	cumperc
1	[30,40]	66	0.1%	66	0.1%
2	(40,50]	1'354	2.0%	1'420	2.1%
3	(50,60]	9'182	13.4%	10'602	15.5%
4	(60,70]	21'350	31.2%	31'952	46.8%
5	(70,80]	17'936	26.2%	49'888	73.0%
6	(80,90]	10'414	15.2%	60'302	88.2%
7	(90,100]	4'930	7.2%	65'232	95.5%
8	(100,110]	1'931	2.8%	67'163	98.3%
9	(110,120]	719	1.1%	67'882	99.3%
10	(120,130]	273	0.4%	68'155	99.7%
11	(130,140]	101	0.1%	68'256	99.9%
12	(140,150]	37	0.1%	68'293	99.9%
13	(150,160]	14	0.0%	68'307	100.0%
14	(160,170]	18	0.0%	68'325	100.0%
15	(170,180]	9	0.0%	68'334	100.0%
16	(180,190]	1	0.0%	68'335	100.0%
17	(190,200]	2	0.0%	68'337	100.0%

✓

```
10
11 #elaboramos el gráfico para la variable weight
12 hist(datos$weight,
13       xlab = "Peso (kg)",
14       ylab = "Frecuencia Absoluta",
15       main = "Distribución de frecuencias de pacientes según su peso" )
```

**Distribución de frecuencias de pacientes según su peso**



```

16
17 media = mean(datos$weight)
18 media
19 mediana = median(datos$weight)
20 mediana

```

20:8 (Top Level) ⌵

Console Terminal × Background Jobs ×

R 4.2.3 ~ /

2	(40,50]	1'334	2.0%	1'420	2.1%
3	(50,60]	9'182	13.4%	10'602	15.5%
4	(60,70]	21'350	31.2%	31'952	46.8%
5	(70,80]	17'936	26.2%	49'888	73.0%
6	(80,90]	10'414	15.2%	60'302	88.2%
7	(90,100]	4'930	7.2%	65'232	95.5%
8	(100,110]	1'931	2.8%	67'163	98.3%
9	(110,120]	719	1.1%	67'882	99.3%
10	(120,130]	273	0.4%	68'155	99.7%
11	(130,140]	101	0.1%	68'256	99.9%
12	(140,150]	37	0.1%	68'293	99.9%
13	(150,160]	14	0.0%	68'307	100.0%
14	(160,170]	18	0.0%	68'325	100.0%
15	(170,180]	9	0.0%	68'334	100.0%
16	(180,190]	1	0.0%	68'335	100.0%
17	(190,200]	2	0.0%	68'337	100.0%

```

> #elaboramos el gráfico para la variable weight
> hist(datos$weight,
+       xlab = "Peso (kg)",
+       ylab = "Frecuencia Absoluta",
+       main = "Distribución de frecuencias de pacientes según su peso" )
> media = mean(datos$weight)
> media
[1] 74.13485
> mediana = median(datos$weight)
> mediana
Error: object 'mediana' not found
> mediana = median(datos$weight)
> mediana
[1] 72

```

**No justificó la asimetría positiva con la medida estadística apropiada: Fisher o Pearson**

Calculamos la media y la mediana de la distribución para determinar la asimetría. Como la media es mayor que la mediana y del histograma observamos que la moda está en el intervalo de [60-70] (menor que la media y la mediana) concluimos que la asimetría es positiva.

**Pregunta 3:**

**3) 3.0**

Ejecutamos el código a continuación:

```

quantile(data$bmi[data$gender == "F"],0.85)
quantile(data$bmi[data$gender == "M"],0.90)

```

```
> quantile(data$bmi[data$gender == "... [TRUNCATED]
85%
33.5

> quantile(data$bmi[data$gender == "M"],0.90)
90%
32.3
```

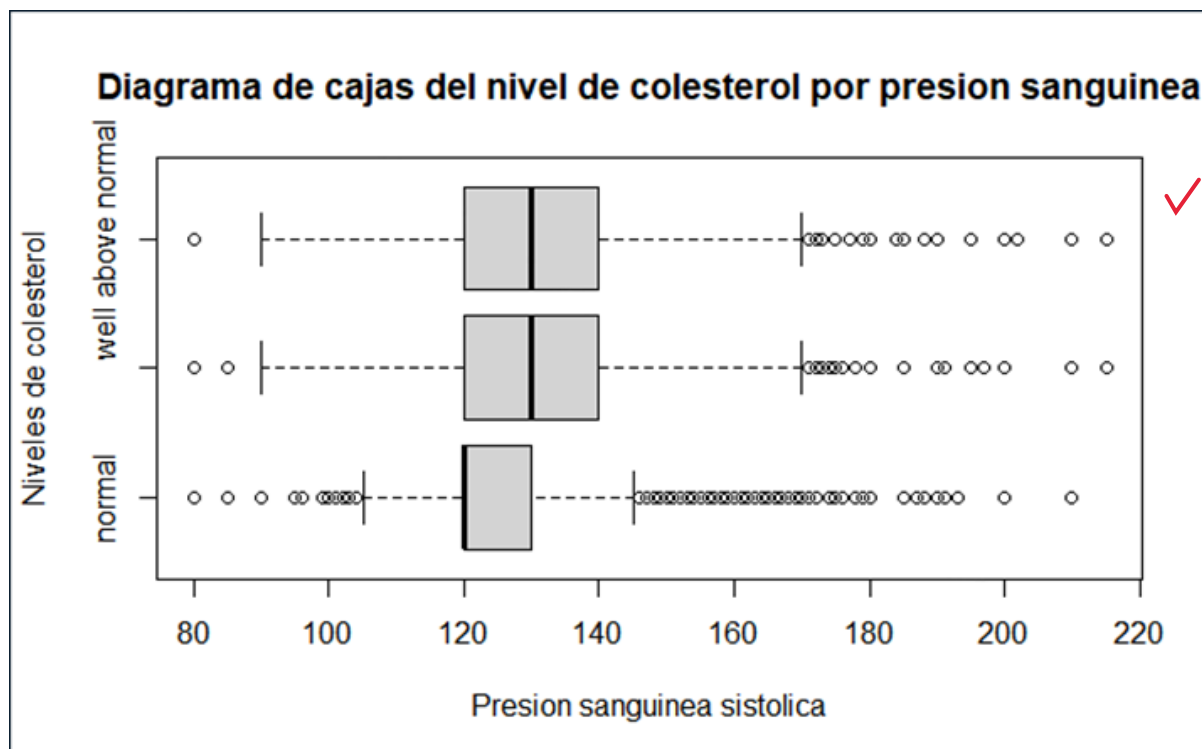
Como se puede observar, los resultados obtenidos son los valores que se deben de superar, respectivamente, si es que se desea incluir una mujer u hombre en el estudio que se está considerando.

PREGUNTA 4:

4) 2.75

Creamos el diagrama de cajas, que involucra la variable colesterol, y la presión sanguínea sistólica.

```
boxplot(formula = ap_hi ~ cholesterol, data = data, horizontal = TRUE,
        ylab = "Niveles de colesterol",
        xlab = "Presion sanguinea sistolica",
        main = "Diagrama de cajas del nivel de colesterol por presion sanguinea")
```



Teniendo en cuenta la mediana, podemos afirmar que el nivel de colesterol normal presenta una presión sanguínea menor, a comparación del resto de los niveles.

Con respecto a la medida de dispersión *RIC*, el gráfico nos da a entender que, para el nivel normal de colesterol, presenta una menor variabilidad.

Finalmente, el diagrama de cajas presenta una asimetría positiva para los niveles de colesterol presentados, como también una cantidad considerable de valores atípicos mayores al RIC (más para el caso del nivel normal). Esto se puede interpretar que existen datos los cuales presentan una presión sanguínea inusualmente elevada a comparación del resto de los pacientes.

Pero los que tienen nivel de colesterol normal tienen una "mayor" asimetría positiva que los otros 2 grupos

#### PREGUNTA 5:

Primero trabajamos sobre los fumadores y luego sobre los no fumadores para compararlos.

```

21
22 #seleccionamos a los fumadores primero
23 fumadores = datos[datos$smoke == "Yes" ,]
24 Freq(fumadores$ap_lo)
25 |

```

5) 2.75

25:1
(Top Level) ▾

Console
Terminal x
Background Jobs x

R 4.2.3 · ~/

> Freq(fumadores)

Error in Freq(fumadores) : object 'fumadores' not found

> #seleccionamos a los fumadores primero

> fumadores = datos[datos\$smoke == "Yes" ]

Error in `[.data.frame'](datos, datos\$smoke == "Yes") :  
undefined columns selected

> #seleccionamos a los fumadores primero

> fumadores = datos[datos\$smoke == "Yes" ,]

> Freq(fumadores)

Error in table(x, useNA = useNA) :  
attempt to make a table with >= 2^31 elements

> Freq(fumadores)

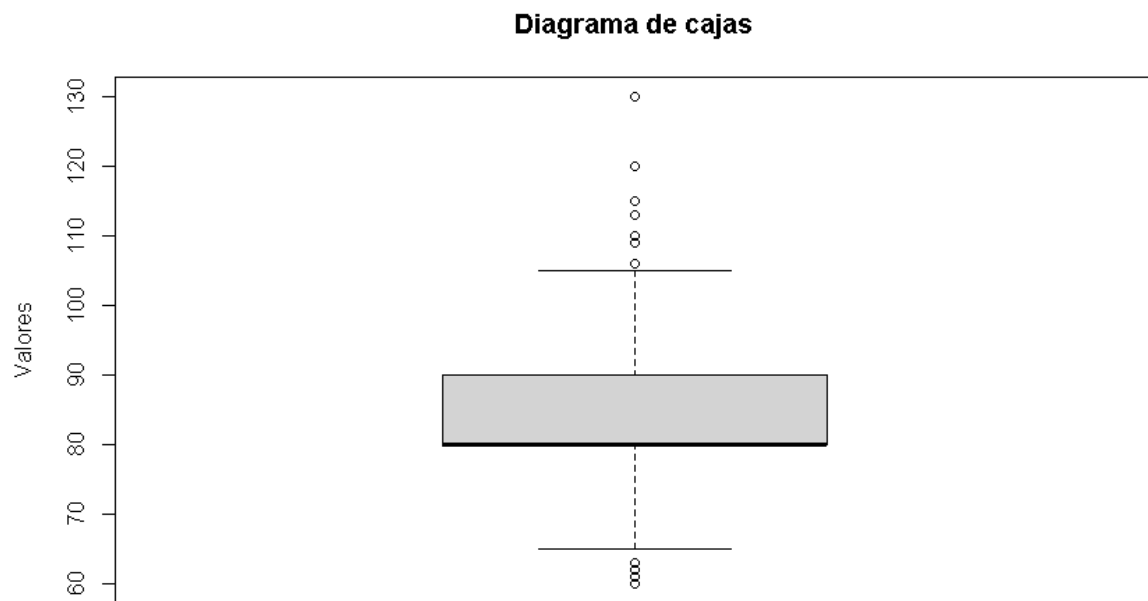
Error in table(x, useNA = useNA) :  
attempt to make a table with >= 2^31 elements

> Freq(fumadores\$ap\_lo)

	level	freq	perc	cumfreq	cumperc
1	[60,65]	219	3.6%	219	3.6%
2	(65,70]	853	14.2%	1'072	17.8%
3	(70,75]	31	0.5%	1'103	18.3%
4	(75,80]	2'937	48.9%	4'040	67.2%
5	(80,85]	44	0.7%	4'084	67.9%
6	(85,90]	1'407	23.4%	5'491	91.3%
7	(90,95]	18	0.3%	5'509	91.6%
8	(95,100]	423	7.0%	5'932	98.7%
9	(100,105]	3	0.0%	5'935	98.7%
10	(105,110]	54	0.9%	5'989	99.6%
11	(110,115]	4	0.1%	5'993	99.7%
12	(115,120]	16	0.3%	6'009	100.0%
13	(120,125]	0	0.0%	6'009	100.0%
14	(125,130]	3	0.0%	6'012	100.0%

> |





Al realizar el diagrama de cajas vemos que hay gran presencia de valores atípicos. Por tanto, el valor que nos ayudará a compararlos con más exactitud es la mediana o la moda (porque no se ven tan afectadas por valores extremos como la media).

Calculamos las medidas de tendencia central para ambas distribuciones.

```

30
31 #Calculamos las medidas de tendencia central de fumadores y no fumadores
32 mean(fumadores$ap_lo)
33 mean(noFumadores$ap_lo)
34 median(fumadores$ap_lo)
35 median(noFumadores$ap_lo)
36 |
37
36:1 | (Top Level) ⌵

```

---

Console   Terminal x   Background Jobs x

R 4.2.3 · ~/

```

> noFumadores = datos[datos$smoke == "no",]
> Freq(noFumadores$ap_lo)

```

	level	freq	perc	cumfreq	cumperc
1	[60,65]	2'591	4.2%	2'591	4.2%
2	(65,70]	9'471	15.2%	12'062	19.4%
3	(70,75]	263	0.4%	12'325	19.8%
4	(75,80]	32'146	51.6%	44'471	71.4%
5	(80,85]	362	0.6%	44'833	71.9%
6	(85,90]	13'021	20.9%	57'854	92.8%
7	(90,95]	212	0.3%	58'066	93.2%
8	(95,100]	3'699	5.9%	61'765	99.1%
9	(100,105]	36	0.1%	61'801	99.2%
10	(105,110]	343	0.6%	62'144	99.7%
11	(110,115]	12	0.0%	62'156	99.7%
12	(115,120]	153	0.2%	62'309	100.0%
13	(120,125]	3	0.0%	62'312	100.0%
14	(125,130]	13	0.0%	62'325	100.0%

```

>
> #Calculamos las medidas de tendencia central de fumadores y no fumadores
> mean(fumadores$ap_lo)
[1] 82.06637
> mean(noFumadores$ap_lo)
[1] 81.25842
> median(fumadores$ap_lo)
[1] 80
> median(noFumadores$ap_lo)
[1] 80

```

Calculamos las medidas de dispersión y concluimos que la asimetría en los fumadores y no fumadores es positiva.

En conclusión, las medidas adecuadas de tendencia para comparación son la moda y la mediana. Vemos que en ambas distribuciones presentan datos similares lo que significa que la variable `ap_lo` no se ve influida en gran medida por la variable `smoke`. ✓

En el caso de las medidas de dispersión, el rango intercuartil (RIC) es la más adecuada para realizar una comparación. ✓

Vemos que el RIC en ambos casos es 10, ya que  $1.000e+01$  es igual a 10. Como ambos son iguales, esto indica que los datos están dispersos de manera similar alrededor de sus respectivas medianas. Por tanto, ambos conjuntos de datos son similares. ✓

Los fumadores presentan una asimetría positiva un poco mayor a los no fumadores

```

35 median(noFumadores$ap_lo)
36
37 #calculamos las medidas de dispersión para ambos casos
38 library(EnvStats)
39 summaryFull(fumadores$ap_lo)
40 summaryFull(noFumadores$ap_lo)
41 |

```

41:1 (Top Level) ↕

Console Terminal × Background Jobs ×

R 4.2.3 · ~/

```

[1] TRUE
attr(,"drop0trailing")
[1] TRUE
> noFumadores = datos[datos$smoke == "No" ,]
> summaryFull(noFumadores$ap_lo)
noFumadores$ap_lo
N 6.232e+04
Mean 8.126e+01
Median 8.000e+01
10% Trimmed Mean 8.107e+01
Geometric Mean 8.073e+01
Skew 2.963e-01
Kurtosis 1.137e+00
Min 6.000e+01
Max 1.300e+02
Range 7.000e+01
1st Quartile 8.000e+01
3rd Quartile 9.000e+01
Standard Deviation 9.245e+00
Geometric Standard Deviation 1.121e+00
Interquartile Range 1.000e+01
Median Absolute Deviation 0.000e+00
Coefficient of Variation 1.138e-01
attr(,"class")
[1] "summaryStats"
attr(,"stats.in.rows")
[1] TRUE
attr(,"drop0trailing")
[1] TRUE
> |

```

## Problema 6

- a) Creamos un diagrama de caja para la variable peso

```

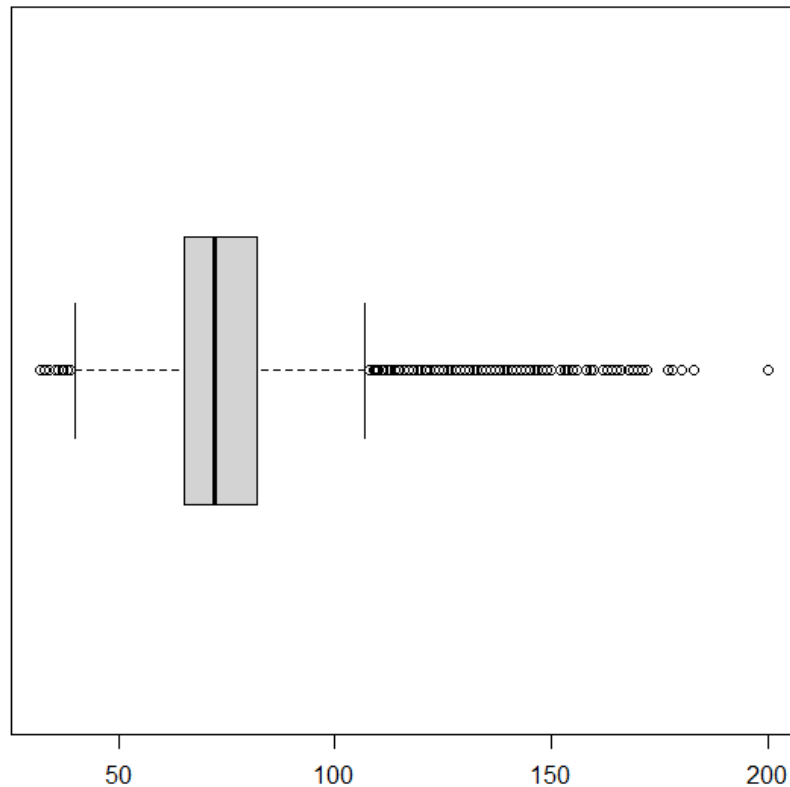
7 boxplot(d$weight,
8         main = "Diagrama de caja de la variable weight",
9         horizontal = 1)
0 |

```

Obteniendo el siguiente gráfico:

a) 1.0

Diagrama de caja de la variable weight



Donde se puede observar que, la variable weight presenta una <sup>✓</sup>asimetría positiva. Entonces, es más recomendable utilizar la mediana, debido a que, la variable <sup>✓</sup>presenta valores atípicos y no es simétrica

- b) Creamos la tabla de frecuencias y un histograma para la variable de cholesterol

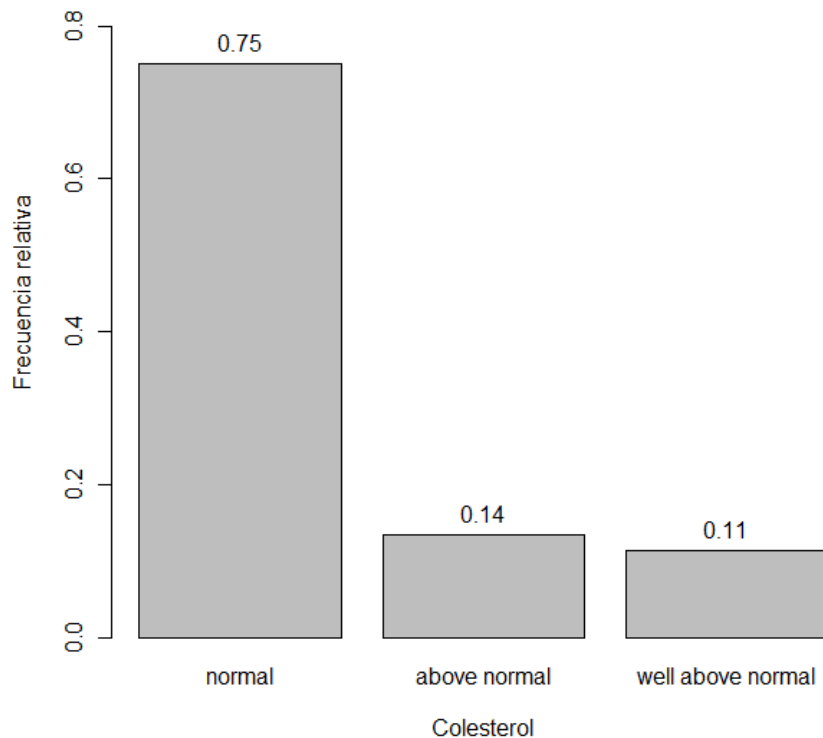
b) 1.0

```
11 library(DescTools)
12 #Tabla de Frecuencias del colesterol
13 Freq(d$cholesterol)
14 #Histograma de Frecuencias
15 f.j = prop.table(table(d$cholesterol))
16 barplot(f.j,
17         main = "Distribución de frecuencias",
18         xlab = "Colesterol",
19         ylab = "Frecuencia relativa",
20         ylim = c(0,max(f.j)*1.2))
21 pos.j = barplot(f.j, plot = FALSE)
22 text(pos.j, f.j,
23      labels = round(f.j,2),
24      pos = 3)
```

Obteniendo los siguientes resultados:

```
> Freq(d$cholesterol)
      level      freq      perc  cumfreq  cumperc
1    normal 51'261  75.0%   51'261   75.0%
2  above normal   9'240  13.5%   60'501   88.5%
3 well above normal   7'836  11.5%   68'337  100.0%
```

Distribución de frecuencias de la variable cholesterol



En la tabla podemos ver que hay un total de 51261 personas con niveles de colesterol normales. Entonces, la mediana de la distribución de niveles de colesterol se ubicaría en el nivel normal, ya que es el punto central que divide los datos en dos partes iguales, concluyendo que la mediana no es above normal. ✓

c) Escribimos el código requerido

```
7 library(EnvStats)
8 summary(d$height)
```

c) 0.25

Obteniendo el siguiente resultado:

```
> summary(d$height)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
140.0  159.0   165.0   164.5  170.0   198.0
```

Observamos que la mediana es 165.0 cm y la media aritmética es 164.5. Entonces, como la mediana, el valor central que separa los datos de manera simétrica, es mayor a la media aritmética, el valor que se puede esperar que tome la variable. Se puede concluir que, aproximadamente el 50% de las personas tiene una altura mayor a 164.5. ✗

Falso, aproximadamente el 50% de las personas tiene una altura mayor a 165cm

d) Escribimos el código requerido

d) 1.0

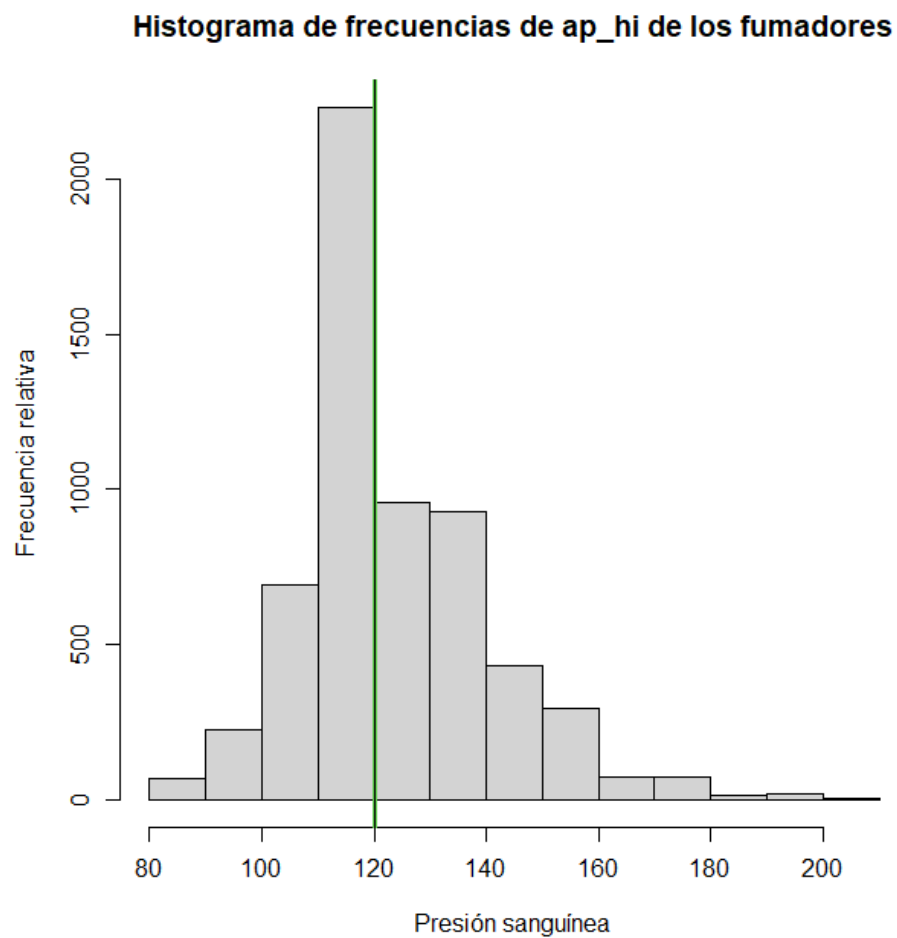
```
#Función para los datos de la ap_hi de los fumadores
summary(d$ap_hi[d$smoke=="Yes"])
#Función para los datos de la ap_hi de los no fumadores
summary(d$ap_hi[d$smoke=="No"])
```

Los resultados son los siguientes:

```
> #Función para los datos de la ap_hi de los fumadores
> summary(d$ap_hi[d$smoke=="Yes"])
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  80.0  120.0   120.0   128.1  140.0   210.0
> #Función para los datos de la ap_hi de los no fumadores
> summary(d$ap_hi[d$smoke=="No"])
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  80.0  120.0   120.0   126.5  140.0   215.0
```

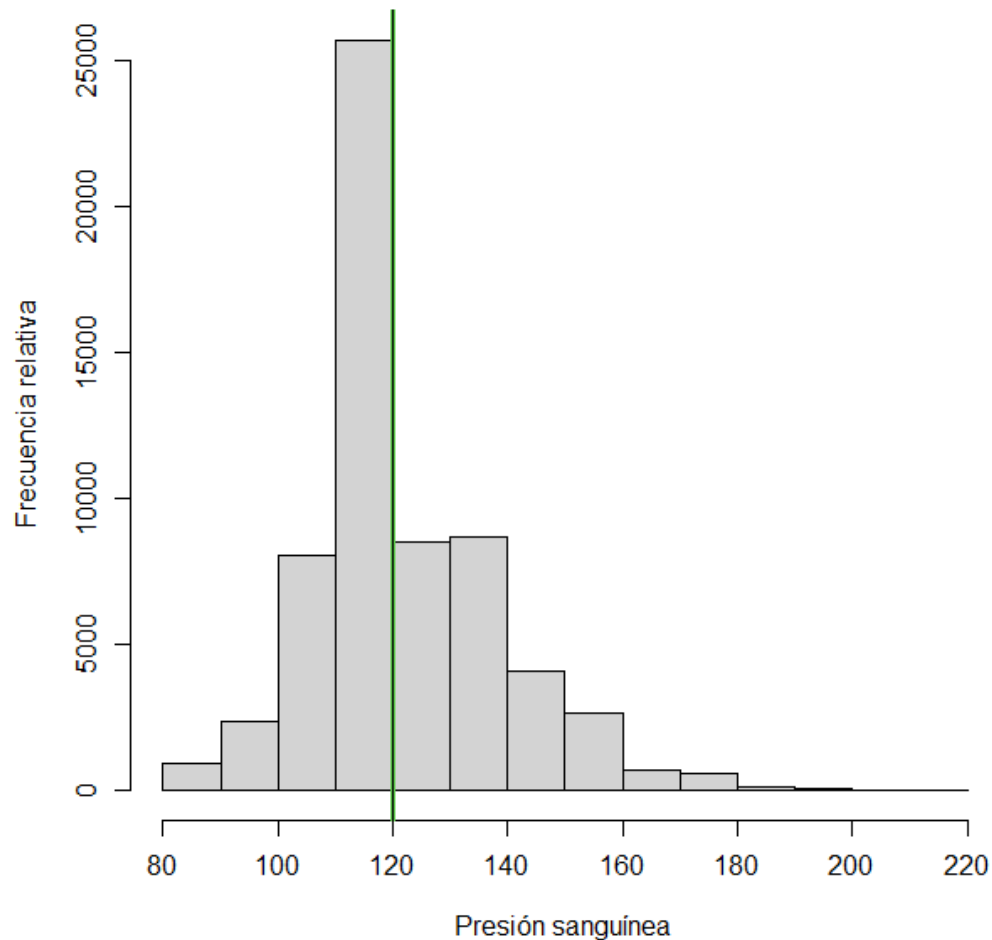
Además, también podemos realizar los histogramas respectivos e incluir dos líneas que representan las medianas.

```
#Histograma de la ap_hi de los fumadores
hist(d$ap_hi[d$smoke == "Yes"],
     main = "Histograma de frecuencias de ap_hi de los fumadores",
     xlab = "Presión sanguínea",
     ylab = "Frecuencia relativa")
#Línea de la mediana de la ap_hi de los fumadores (Verde)
abline(v = median(d$ap_hi[d$smoke == "Yes"]), col=3, lwd=3)
#Línea de la mediana de la ap_hi de los no fumadores (Negro)
abline(v = median(d$ap_hi[d$smoke == "No"]), col=1, lwd=1.8)
```



```
#Histograma de la ap_hi de los no fumadores
hist(d$ap_hi[d$smoke == "No"],
     main = "Histograma de frecuencias de ap_hi de los no fumadores",
     xlab = "Presión sanguínea",
     ylab = "Frecuencia relativa")
#Linea de la mediana de la ap_hi de los fumadores (Verde)
abline(v = median(d$ap_hi[d$smoke == "Yes"]), col=3, lwd=3)
#Linea de la mediana de la ap_hi de los no fumadores (Negro)
abline(v = median(d$ap_hi[d$smoke == "No"]), col=1, lwd=1.8)
```

### Histograma de frecuencias de ap\_hi de los no fumadores



En conclusión, podemos decir que la mediana de la presión sanguínea sistólica es la misma para fumadores y no fumadores, comprobándolo gráficamente y con la función summary que nos permitió saber que numéricamente son iguales. ✓

- e) Para calcular la dispersión de las variables bmi y age, utilizaremos la función sd() que nos permite obtener la desviación estándar.

```
#Desviación estandar de la variable Bmi  
sd(d$bmi)  
#Desviación estandar de la variable Age  
sd(d$age)
```

e) 0.5

Obteniendo el siguiente resultado:

```
> sd(d$bmi)  
[1] 5.181394  
> sd(d$age)  
[1] 6.759342
```

✗

Como podemos observar, la desviación estándar de la variable bmi es menor que el de age. Sabiendo que, la desviación estándar pequeña nos indica una poca dispersión.

Entonces, la variable bmi presenta una menor dispersión que la de age. ✗

Cuando son "diferentes unidades" se debe usar el coeficiente de variabilidad