TRABAJO DE ANÁLISIS EXPLORATORIO Y DESCRIPTIVO DE DATOS

Yosef Shmuel Guevara Salamanca

Killiam Ferney Mogollón Gómez

Juan David Ramírez Ávila

Juliana Lucia Rodríguez Castillo

Laura Sofía Rodríguez

Trabajo de Métodos Estadísticos Aplicados

Docente

Carlos Arturo Panza Ospino

UNIVERSIDAD NACIONAL DE COLOMBIA

FACULTAD DE CIENCIAS

ESPECIALIZACIÓN EN ESTADÍSTICA

NOVIEMBRE DEL 2020

1.Introducción

El presente documento presenta la solución de los cinco ejercicios propuestos, para aplicar los conceptos y técnicas vistas en el capítulo de análisis exploratorio y descriptivo, durante las primeras semanas del curso de Métodos Estadísticos aplicados.

2. Primer Ejercicio

Supóngase que los valores de la variable Z se obtienen de la suma de otras dos variables X y Y. Mostrar que:

a) La media aritmética de la variable Z es la suma de las medias aritméticas de X y Y.

b) La varianza de Z puede ser mayor, menor o igual que la suma de las varianzas de X y Y.

2.1 Solución Literal a

Se sabe que $z_i = x_i + y_i$ (1) , donde el subíndice i va de uno hasta n (valores), y este hace referencia a las observaciones de cada variable. Entonces, se desea verificar que, $\bar{Z} = \bar{X} + \bar{Y}$ (2) , es decir, que la suma de las medias aritméticas de las variables X e Y, es igual a la media aritmética de la variable Z.

Aplicando tanto la definición de media aritmética como factor común, en la ecuación (2), a los términos $\bar{X}$ y $\bar{Y}$, se obtiene la expresión (3), como se aprecia a continuación.

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\bar{Z} = \frac{1}{n}\left[\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i\right] \quad (3)$$

Luego, aplicando propiedades de sumatorias y la ecuación (1) se puede expresar la ecuación (3), como se observa en el resultado (4).

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} (x_i + y_i)$$

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} z_i \quad (4)$$

Y la ecuación (4), es la definición de media aritmética aplicada a la variable Z. Luego, con el proceso descrito en esta sección se verifica la igualdad de la ecuación (2).

2.2 Solución al literal b

Usando la definición de varianza en la variable Z, se obtiene la ecuación (5).

$$s_Z^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{Z})^2 \ (5)$$

Sustituyendo las ecuaciones (1) y (2) en (5), se obtiene el resultado (6), como se muestra seguidamente.

$$s_Z^2 = \frac{1}{n}\sum_{i=1}^{n}[(x_i + y_i) - (\bar{X} + \bar{Y})]^2 \ (6)$$

Ordenando, realizando un cambio de variables y desarrollando la expresión (6), se consigue el resultado (7)

$$s_Z^2 = \frac{1}{n}\sum_{i=1}^{n}[(x_i - \bar{X}) + (y_i - \bar{Y})]^2$$

Se define $a = x_i - \bar{X}$ y $b = y_i - \bar{Y}$, para obtener,

$$s_Z^2 = \frac{1}{n}\sum_{i=1}^{n}(a + b)^2$$

$$s_Z^2 = \frac{1}{n}\sum_{i=1}^{n}(a^2 + 2ab + b^2)$$

Ahora, remplazando nuevamente tanto a como b, en el resultado inmediatamente anterior se obtiene,

$$s_Z^2 = \frac{1}{n}\sum_{i=1}^{n}((x_i - \bar{X})^2 + 2(x_i - \bar{X})(y_i - \bar{Y}) + (y_i - \bar{Y})^2)$$

Luego, aplicando propiedades de sumatorias, la ecuación quedaría,

$$s_Z^2 = \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{X})^2 + 2\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}) + \sum_{i=1}^{n}(y_i - \bar{Y})^2\right]$$

Ahora, multiplicando toda la expresión por $1/n$, el resultado se transforma en:

$$s_Z^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2 + \frac{2}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}) + \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{Y})^2 \quad (7)$$

Al apreciar, la ecuación (7), se observa, que, en la parte de la derecha, los términos de los extremos son respectivamente la varianza de la variable X y de la variable Y. Entonces, se puede reescribir la expresión (7), como se aprecia en (8).

$$s_Z^2 = s_X^2 + \frac{2}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}) + s_Y^2 \quad (8)$$

Ahora, del resultado (8), se desarrolla, el término del centro, mediante la aplicación de algunas propiedades de las sumatorias y la definición de media aritmética, con el fin de obtener la ecuación (9).

$$s_Z^2 = s_X^2 + \frac{2}{n}\sum_{i=1}^{n}(x_i y_i - x_i\bar{Y} - \bar{X}y_i + \bar{X}\bar{Y}) + s_Y^2$$

$$s_Z^2 = s_X^2 + \frac{2}{n}\sum_{i=1}^{n}(x_i y_i - x_i\bar{Y} - \bar{X}y_i + \bar{X}\bar{Y}) + s_Y^2$$

$$s_Z^2 = s_X^2 + \frac{2}{n}\sum_{i=1}^{n}x_i y_i - \frac{2}{n}\sum_{i=1}^{n}x_i\bar{Y} - \frac{2}{n}\sum_{i=1}^{n}\bar{X}y_i + \frac{2}{n}\sum_{i=1}^{n}\bar{X}\bar{Y} + s_Y^2$$

$$s_Z^2 = s_X^2 + \frac{2}{n}\sum_{i=1}^{n}x_i y_i - \frac{2}{n}\sum_{i=1}^{n}x_i\bar{Y} - \frac{2}{n}\sum_{i=1}^{n}\bar{X}y_i + \frac{2}{n}\sum_{i=1}^{n}\bar{X}\bar{Y} + s_Y^2$$

$$s_Z^2 = s_X^2 + \frac{2}{n}\sum_{i=1}^{n}x_i y_i - \bar{Y}\,\frac{2}{n}\sum_{i=1}^{n}x_i - \bar{X}\,\frac{2}{n}\sum_{i=1}^{n}y_i + 2\bar{X}\bar{Y} + s_Y^2$$

$$s_Z^2 = s_X^2 + \frac{2}{n}\sum_{i=1}^{n}x_i y_i - \bar{Y}\,\frac{2}{n}\sum_{i=1}^{n}x_i - \bar{X}\,\frac{2}{n}\sum_{i=1}^{n} + 2\bar{X}\bar{Y} + s_Y^2$$

$$s_Z^2 = s_X^2 + \frac{2}{n}\sum_{i=1}^{n}x_i y_i - 2\bar{X}\bar{Y} - 2\bar{X}\bar{Y} + 2\bar{X}\bar{Y} + s_Y^2$$

$$s_Z^2 = s_X^2 + \frac{2}{n}\sum_{i=1}^{n}x_i y_i - 2\bar{X}\bar{Y} + s_Y^2$$

$$s_Z^2 = s_X^2 + 2\left[\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{X}\bar{Y}\right] + s_Y^2 \ (9)$$

Al apreciar la ecuación (9), se observa que el término del centro es la covarianza multiplicada por dos. Entonces, se puede reescribir como se muestra en el resultado (10).

$$s_Z^2 = s_X^2 + 2S_{XY} + s_Y^2 \ (10)$$

2.2.1 Varianza de *Z* igual a la suma de las varianzas de *X* e *Y*

Cuando la covarianza de las variables X e Y es cero, indica la ausencia de dependencia lineal entre las variables. Entonces, bajo la situación citada anteriormente se cumple que, $s_Z^2 = s_X^2 + s_Y^2$, es decir, la varianza de la variable Z, es igual a la suma de las varianzas de las variables X e Y.

2.2.2 Varianza de *Z* mayor a la suma de las varianzas de *X* e *Y*

Cuando la covarianza de las variables X e Y es menor que cero, indica que existe algún grado de correlación lineal negativa (Y decrece cuando X crece). Luego, bajo la situación descrita anteriormente, se satisface que $s_Z^2 > s_X^2 + s_Y^2$, es decir, la varianza de la variable Z, es mayor a la suma de las varianzas de las variables X e Y.

2.2.3 Varianza de *Z* menor a la suma de las varianzas de *X* e *Y*

Cuando la covarianza de las variables X e Y es mayor que cero, indica que existe algún grado de correlación lineal positiva (Y crece cuando X crece). Luego, bajo la situación mencionada anteriormente, se satisface que $s_Z^2 < s_X^2 + s_Y^2$, es decir, la varianza de la variable Z, es menor a la suma de las varianzas de las variables X e Y.

3. Segundo Ejercicio

Demostrar que el coeficiente de correlación lineal de Pearson es:

a) Invariante a cambios de escala.

b) En módulo igual a la unidad, si entre dos variables existe una relación lineal exacta.

3.1 Solución literal a

Para argumentar la solución del literal a, primero se presentan las pruebas de tres propiedades, que se necesitan para la demostración solicitada.

3.1.1 Media sometida a cambios de escala

Se tiene un conjunto de datos, que va desde $x_1, x_2, \ldots, x_n$, entonces, lo multiplicamos por una cantidad a, que puede ser positiva, negativa o cero. Luego, se puede definir que, $y_i = ax_i$, para $i = 1, \ldots, n$. Lo que significa, que la media del conjunto de datos transformado $y_1, \ldots, y_n$ es equivalente a $a\bar{X}$. A continuación, se presenta, la prueba de lo descrito en el presente párrafo.

$$\bar{y} = \frac{1}{n}(y_1 + \cdots + y_n)$$

$$= \frac{1}{n}(ax_1, + \cdots + ax_n)$$

$$= \frac{1}{n}a(x_1, + \cdots + x_n)$$

$$= a\bar{X} \ (11)$$

3.1.2 Varianza sometida a cambios de escala

Se tiene un conjunto de datos, que va desde $x_1, x_2, \ldots, x_n$, entonces, lo multiplicamos por una cantidad a, que puede ser positiva, negativa o cero. Luego, se puede definir que, $y_i = ax_i$, para $i = 1, \ldots, n$. Lo que significa, que la media del conjunto de datos transformado $y_1, \ldots, y_n$ es equivalente al producto $a^2 s_X^2$. Entonces, recordando que $\bar{y} = a\bar{X}$, se tiene que,

$$s_y^2 = \frac{1}{n}((y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2)$$

$$s_y^2 = \frac{1}{n}((ax_1 - a\bar{x})^2 + \cdots + (ax_n - a\bar{x})^2)$$

$$s_y^2 = \frac{1}{n}(a^2(x_1 - \bar{x})^2 + \cdots + a^2(x_n - \bar{x})^2)$$

$$s_y^2 = a^2\frac{1}{n}((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2)$$

$$s_y^2 = a^2 \, s_x^2 \ (12)$$

Entonces, con el proceso inmediatamente anterior, se prueba que la varianza de un conjunto de datos multiplicados por una constante es igual a la constante al cuadrado multiplicada por la varianza de los datos originales.

3.1.3 Covarianza sometida a cambios de escala

Si a es una constante y se tiene el conjunto de datos, constituido por, $ax_1, \dots, ax_n$, se ha realizado un cambio de escala. Luego, recordando la propiedad del numeral 3.1.1 y que la media aritmética, para el conjunto de datos definido en el presente párrafo, seria igual a $a\bar{X}$. Entonces,

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(ax_i - a\bar{X})(y_i - \bar{Y})$$

$$= \frac{1}{n}\sum_{i=1}^{n}a(x_i - \bar{X})(y_i - \bar{Y})$$

$$= a\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})$$

$$= a\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})$$

$$= a\frac{1}{n}\sum_{i=1}^{n}x_i y_i - \bar{X}\bar{Y}$$

$$= as_{XY} \quad (13)$$

3.1.4 Coeficiente de correlación de Pearson invariante a cambios de escala

En esta sección se describe la prueba, con la que se verifica que la magnitud del coeficiente de correlación de Pearson es invariante a cambios de escala. Si a es una constante distinta de cero y se tiene el conjunto de datos transformados $ax_1, \dots, ax_n$, lo que corresponde a llevar un cambio de escala en la variable y recordando que la media de este conjunto de datos es $a\bar{X}$ con base en el numeral 3.1.1.

Entonces, aplicando la definición del coeficiente de correlación de Pearson y un cambio de escala, al multiplicar la variable X, por una constante a, se tiene la ecuación (14)

$$r_{XY} = \frac{cov(aX, Y)}{\sqrt{var(aX)}\sqrt{var(Y)}} \quad (14)$$

Utilizando en el numerador, la propiedad demostrada en el numeral 3.1.3 y en el denominador, la propiedad del numeral 3.1.2, la expresión (14), se puede reescribir como se muestra en la expresión (15).

$$r_{XY} = \frac{acov(X,Y)}{\sqrt{a^2 var(X)}\sqrt{var(Y)}} \quad (15)$$

Luego, utilizando la propiedad de que $\sqrt{x^2} = |x|$ y la definición de coeficiente de correlación, el resultado (15), queda como se ve en la ecuación (16).

$$r_{XY} = \frac{acov(X,Y)}{|a|\sqrt{var(X)}\sqrt{var(Y)}}$$

$$= \frac{a}{|a|} r_{XY} \quad (16)$$

$$= \begin{cases} +r_{XY}, & si \;\; a > 0 \\ -r_{XY}, & si \;\; a < 0 \end{cases}$$

Entonces, con base en el proceso de este numeral, se demuestra, que la magnitud, del coeficiente de correlación, se conserva, frente a cambios de escala, pero este adquiere el signo de la constante a.

3.2 Solución al literal b

Cuando hay una relación lineal exacta entre las variables, X e Y, se define que $Y = aX + b$, con $a \neq 0$ y b constante. Entonces, haciendo uso de las propiedades tanto de la varianza como de la covarianza, se puede hacer la demostración que se presenta a continuación.

Se parte de la definición, de coeficiente de correlación de Pearson (17), en esta la variable Y, se sustituye por $ax + b$ y se obtiene el resultado (18).

$$r_{XY} = \frac{cov(X,Y)}{\sqrt{var(X)}\sqrt{var(Y)}} \quad (17)$$

$$r_{XY} = \frac{cov(X, aX+b)}{\sqrt{var(X)}\sqrt{var(aX+b)}} \quad (18)$$

Luego, aplicando en el numerador tanto la propiedad de que $cov(ax, y) = acov(x, y)$ como la de que $cov(x + c, y) = cov(x, y)$, donde a y c son constantes y también usando en el denominador la propiedad de la varianza de que $var(aX + b) = a^2 var(x)$. Entonces, se obtiene la expresión (19).

$$r_{XY} = \frac{acov(X,X)}{\sqrt{var(X)}\sqrt{a^2 var(X)}} \quad (19)$$

Ahora, en la ecuación (19), utilizando en el numerador la propiedad de la covarianza, de que $cov(x,x) = var(x)$ y empleando en el denominador, que $\sqrt{x^2} = |x|$, se obtiene el resultado (20).

$$r_{XY} = \frac{avar(X)}{|a|\sqrt{var(X)}\sqrt{var(X)}} \quad (20)$$

En la ecuación (20), se observa que el denominador se convierte en la $var(X)$, y este término se encuentra en el numerador, entonces, se consigue la expresión (21).

$$r_{XY} = \frac{avar(X)}{|a|var(X)}$$

$$= \frac{a}{|a|} \quad (21)$$

$$= \begin{cases} +1, & si\ \ a > 0 \\ -1, & si\ \ a < 0 \end{cases}$$

Entonces, con el proceso descrito en este numeral, queda demostrado que el coeficiente de correlación de Pearson en una relación lineal exacta puede tomar el valor de -1 o de 1, es de decir, en este tipo de situaciones su módulo es uno ($|r_{XY}| = 1$).

4. Tercer Ejercicio

En los últimos años, el porcentaje de grasa corporal se ha convertido en un indicador del estado de salud y de acondicionamiento físico de una persona. Sin embargo, los métodos existentes para obtener valoraciones suficientemente confiables del porcentaje de grasa corporal resultan a la larga costosos y tampoco garantizan un nivel de precisión adecuado. El archivo **bodyfat** del sitio web https://dasl.datadescription.com/datafiles/ contiene 250 mediciones precisas del porcentaje de grasa corporal (**Pct.BF**) de 250 hombres adultos de varias edades junto con mediciones de distintas variables antropométricas. Se pide:

a) Quitar de la tabla de datos la variable **Density** y trabajar con el resto de la tabla.

b) Realizar la descripción de alguna (sólo una) de las variables (con excepción de la variable **Age**) de la tabla.

c) Dividir a los participantes en el estudio en tres grupos de acuerdo con su edad en menores de 30, entre 31 y 50 y mayores de 50. Realice la descripción de la variable escogida en el numeral anterior de acuerdo con esta categorización.

d) A su juicio, ¿cuál o cuáles de las variables medidas en el estudio serviría como mejor indicador del indicador del porcentaje de grasa corporal en hombres adulto.

4.1 Solución literal a

En este numeral se muestra la carga del "dataframe" a R, y como se retiró la variable "Dens ity. Con la línea de código `head(bodyfat)`, se carga el encabezado de la base de datos y e n la **Figura 1**, se aprecia, el resultado que arroja el software.

Figura 1. Encabezado de la base de datos en R

```
> head(bodyfat)
  Density Pct.BF Age Weight Height Neck Chest Abdomen    Waist   Hip Thigh Knee
1  1.0708   12.3  23 154.25  67.75 36.2  93.1    85.2 33.54331  94.5  59.0 37.3
2  1.0853    6.1  22 173.25  72.25 38.5  93.6    83.0 32.67717  98.7  58.7 37.3
3  1.0414   25.3  22 154.00  66.25 34.0  95.8    87.9 34.60630  99.2  59.6 38.9
4  1.0751   10.4  26 184.75  72.25 37.4 101.8    86.4 34.01575 101.2  60.1 37.3
5  1.0340   28.7  24 184.25  71.25 34.4  97.3   100.0 39.37008 101.9  63.2 42.2
6  1.0502   20.9  24 210.25  74.75 39.0 104.5    94.4 37.16535 107.8  66.0 42.0
  Ankle Bicep Forearm Wrist
1  21.9  32.0    27.4  17.1
2  23.4  30.5    28.9  18.2
3  24.0  28.8    25.2  16.6
4  22.8  32.4    29.4  18.2
5  24.0  32.2    27.7  17.7
6  25.6  35.7    30.6  18.8
```

Y en la **Figura 2**, se muestra, la evidencia de que se quitó la variable "Density", de la base de datos "bodyfat, con la línea de código:

```
bodyfat_sin_density <- bodyfat[, setdiff(colnames(bodyfat), "Density")]
```

Figura 2. Encabezado de la base de datos sin la variable "Density"

```
> bodyfat_sin_density <- bodyfat[, setdiff(colnames(bodyfat), "Density")]
> head(bodyfat_sin_density)
  Pct.BF Age Weight Height Neck Chest Abdomen    Waist   Hip Thigh Knee Ankle Bicep
1   12.3  23 154.25  67.75 36.2  93.1    85.2 33.54331  94.5  59.0 37.3  21.9  32.0
2    6.1  22 173.25  72.25 38.5  93.6    83.0 32.67717  98.7  58.7 37.3  23.4  30.5
3   25.3  22 154.00  66.25 34.0  95.8    87.9 34.60630  99.2  59.6 38.9  24.0  28.8
4   10.4  26 184.75  72.25 37.4 101.8    86.4 34.01575 101.2  60.1 37.3  22.8  32.4
5   28.7  24 184.25  71.25 34.4  97.3   100.0 39.37008 101.9  63.2 42.2  24.0  32.2
6   20.9  24 210.25  74.75 39.0 104.5    94.4 37.16535 107.8  66.0 42.0  25.6  35.7
  Forearm Wrist
1    27.4  17.1
2    28.9  18.2
3    25.2  16.6
4    29.4  18.2
5    27.7  17.7
```

4.2 Solución literal b

En este apartado se presenta tanto una descripción gráfica como una numérica de la variable biceps, del dataframe bodyfat. Lo primero que se hizo, fue extraer la variable mencionada, de la base de datos, con la línea de código `Bicep <- bodyfat_sin_density[, "Bicep"]` como se aprecia a continuación en la **Figura 3**.

Figura 3. Encabezado de la extracción de la variable Bicep

```
> Bicep <- bodyfat_sin_density[, "Bicep"]
>
> head(Bicep)
[1] 32.0 30.5 28.8 32.4 32.2 35.7
>
```

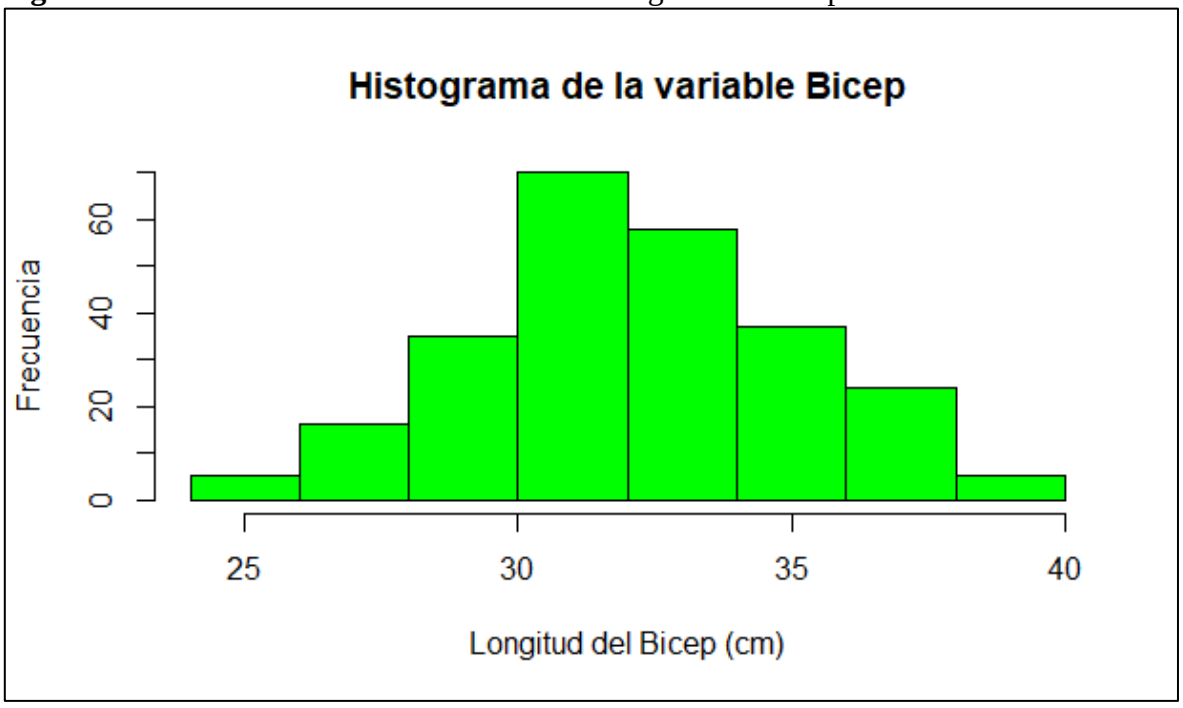
4.2.1 Descripción gráfica

En esta sección, se presenta tanto un histograma como un diagrama de caja y bigotes, para realizar un primer acercamiento descriptivo del comportamiento de los datos.

4.2.1.1 Histograma

En este apartado con la línea de código, `hist(Bicep, main="Histograma de la variable Bicep", col="green", xlab="Longitud del Bicep (cm)", ylab="Frecuencia"),` se dibujó el histograma de la **Figura 4**, con base en este, se puede apreciar, que los datos de la variable "Bicep", tienen una distribución aproximadamente simétrica, ya que no hay una acumulación de la frecuencia de los datos a izquierda o a derecha de la distribución.

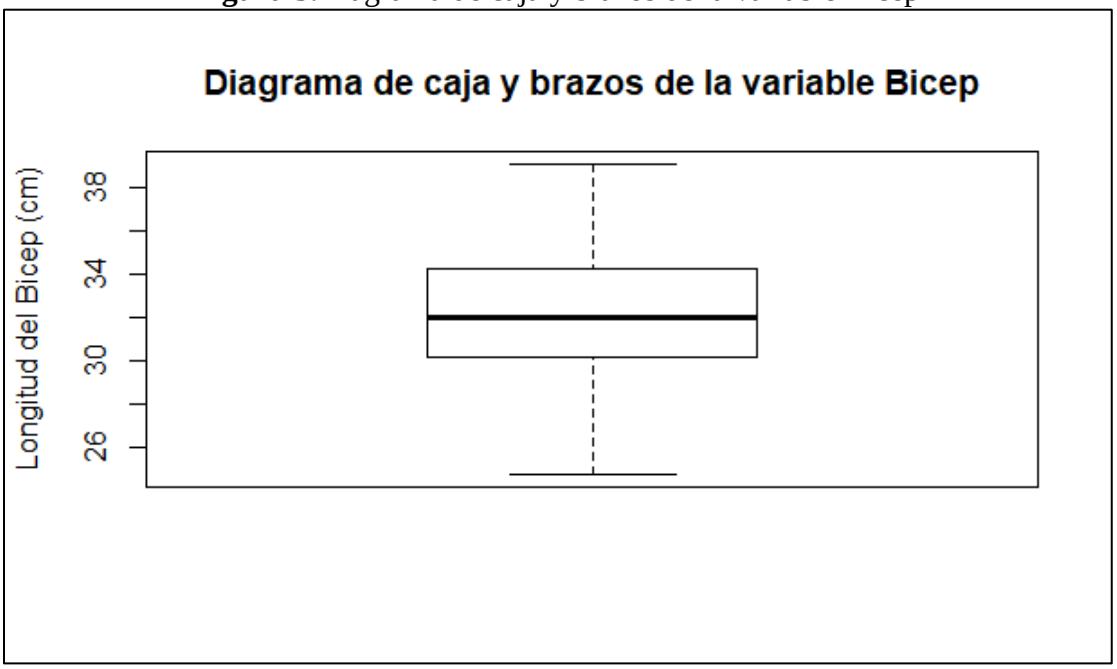
Figura 4. Distribución de las mediciones de la longitud del Bicep de los 250 hombres



4.2.1.2 Diagrama de caja y brazos

En este numeral, con la línea de código, `boxplot(Bicep, main="Diagrama de caja y brazos de la variable Bicep", ylab="Longitud del Bicep (cm)")`, se dibujó, el diagrama de la **Figura 5**, con este se aprecia, que no hay datos atípicos, ya que no se aprecian, observaciones debajo del brazo inferior o por encima del brazo superior y también se confirma que la distribución es aproximadamente simétrica, ya que la línea gruesa horizontal, que representa a la mediana, se encuentra aproximadamente en la mitad de los brazos.

Figura 5. Diagrama de caja y brazos de la variable Bicep



4.2.2 Descripción numérica

En esta sección se presentan una diversidad de medidas numéricas, que sirven para describir y dar una primera idea, tanto de la forma como de la distribución de los datos, de la variable Bicep.

4.2.2.1 Medidas de localización

Acá se presenta el cálculo y la interpretación de la moda, la media y la moda. Con las siguientes líneas de código:

```
#Cáculo de la media
mean(Bicep)
#Cálculo de la moda
names(sort(-table(Bicep)))
#Cálculo de mediana
median(Bicep)
```

Los datos obtenidos con las líneas de código anteriores se consignan en la **Tabla 1**.

Tabla 1. Medidas de localización para la variable

Media	Moda	Mediana
32.2172	30.5	32

Como se aprecia, en la **Tabla 1**, los valores para la media, la moda y la mediana son muy cercanos, entonces, la proximidad de estos valores, indica que la distribución es aproximadamente simétrica.

4.2.2.2 Medidas de dispersión

Acá se presenta el cálculo y la interpretación de varianza, desviación estándar, desviación media, rango y coeficiente de variación. Con las líneas de código que se presentan a continuación se hizo el cálculo de las medidas de dispersión citas en el presente párrafo.

```
> #Cálculo de la varianza
> sum((Bicep - mean(Bicep))^2) / length(Bicep)
[1] 8.506544
> #Cálculo de la desviación estándar
> sqrt(sum((Bicep - mean(Bicep))^2) / length(Bicep))
[1] 2.916598
> #Cálcuo de la desviación media
> sum(abs(Bicep-mean(Bicep)))/length(Bicep)
[1] 2.375901
> #Cálculo del rango
> range(Bicep)
[1] 24.8 39.1
> #Cálculo del rango
> max(Bicep)-min(Bicep)
[1] 14.3
> #Coeficiente de variación
> sqrt(sum((Bicep - mean(Bicep))^2) / length(Bicep))/mean(Bicep)
[1] 0.09052922
```

Tabla 2. Resultados de las medidas de dispersión de la variable Bicep

Medida de dispersión	**Valor calculado**
Varianza	8.506544
Desviación estándar	2.916598
Desviación media	2.375901
Rango	14.3
Coeficiente de variación	0.09052922

Las medidas de la **Tabla 2,** sirven para cualificar la variabilidad o el grado de dispersión de un conjunto de datos numéricos. En medidas tales, como la varianza, la desviación estándar, la desviación media, el rango y el coeficiente de variación, se debe considerar un valor central como punto de referencia, generalmente se toma la media como punto de partida.

Las medidas de dispersión de la **Tabla 2**, indican que la dispersión del conjunto de datos de la variable Bicep, con respecto a la media, lo que también se aprecia en el histograma de la

Figura 4, ya que este no presenta huecos y con el diagrama de cajas y brazos de la **Figura 5**, porque en este no se aprecian valores extremos o atípicos.

Por ejemplo, la desigualdad de Chebyshov, establece que más del 75% de las observaciones, se encuentran a dos desviaciones estándar de la media aritmética, entonces, con esta premisa, se sabe que en el intervalo comprendido entre 26.384004 y 38.050396, se encuentran más del 75% de los datos de la variable Bicep, lo que también viene siendo un indicador de que la dispersión en los datos no es muy alta.

4.2.2.3 Medidas de posición relativa (Cuantiles)

Un cuantil es un número o punto de referencia que divide a los datos en dos partes, un cierto porcentaje de las observaciones son menores o iguales al cuantil y el porcentaje complementario corresponde a datos que son mayores o iguales al cuantil. Dentro de los cuantiles, se tienen los deciles, percentiles y los cuartiles, en el presente numeral se hace la descripción de los cuartiles para la variable Bicep.

Los cuartiles son los valores que dividen una distribución ordenada en cuatro grupos que contiene cada uno el 25% de los datos. A continuación, se presentan la línea de código empleada para calcular los cuartiles, y en la **Tabla 3,** se consigan los resultados obtenidos.

```
> #Cálculo de los cuartiles en R
> quantile(Bicep)
  0%  25%  50%  75% 100%
24.8 30.2 32.0 34.3 39.1
```

Tabla 3. Cuartiles de la variable Bicep

Cuartil	Valor
Q_L	30.2
Q_C	32.0
Q_U	34.3
IQR	4.1

El cuartil Q_L indica que el 25% de las observaciones se encuentran por debajo, de 30.2, que el 50% de los datos están por debajo de 32, y que el 75% están por debajo de 34.3.

4.2.2.4 Coeficiente de asimetría

El coeficiente de asimetría es una medida descriptiva, que indica el grado de simetría de una distribución. Con las líneas de código que se muestran a continuación, se hizo el cálculo del coeficiente de asimetría de la variable Bicep.

```
> #Cálculo del coeficiente de asimetría
> library(moments)
> skewness(Bicep)
[1] 0.04125329
```

El coeficiente de asimetría para el conjunto de datos, de la variable Bicep, reporto un valor de 0.04125329, este valor es muy cercano a cero, lo que indica que la distribución es muy simétrica y confirma los análisis realizados a lo largo del presente documento.

4.2.2.5 Coeficiente de curtosis

Con las líneas de código que se presentan a continuación se hizo el cálculo del coeficiente de Curtosis para la variable Bicep.

```
> #Cálculo del coeficiente de apuntamiento o curtosis
> library(moments)
> kurtosis(Bicep)
[1] 2.540527
```

Para la variable Biceps, se determinó un coeficiente de curtosis de 2.540427, este valor es cercano a tres, lo que indica que existe un grado de dispersión moderado con respecto a la media y confirma nuevamente, lo que se ha intuido con otras estrategias descriptivas como los diagramas de las **Figuras 4** y **5** y las medidas de dispersión calculadas en la sección 4.2.2.2 del presente escrito.

4.3 Solución literal c

En este numeral se presenta el código empleado para categorizar los datos de la variable Bicep, en tres grupos de edad, menores a 30, entre 30 y 50 y los mayores a 50.

```
#Agrupando los datos por edades

#Grupo de menores de 30 años

bicep_menores_30 <- bodyfat_sin_density[bodyfat_sin_density $Age < 30,
c("Bicep")]
categoria <- rep("Menores de 30 años",length(bicep_menores_30))
bicep_menores_30 <- cbind(bicep_menores_30, categoria)

#Grupo de edad entre 30 y 50 años

bicep_entre_30_y_50 <- bodyfat_sin_density[bodyfat_sin_density$Age > 30 &
bodyfat_sin_density$Age < 50, c("Bicep")]
categoria <- rep("Entre 31 y 50 años", length(bicep_entre_30_y_50))
bicep_entre_30_y_50 <- cbind(bicep_entre_30_y_50, categoria)

#Grupo de edad mayores de 50 años
bicep_mayores_de_50 <- bodyfat_sin_density [bodyfat_sin_density $Age > 50
, c("Bicep")]
categoria <- rep("Mayores de 50", length(bicep_mayores_de_50))
bicep_mayores_de_50 <- cbind(bicep_mayores_de_50, categoria)


#Consolidar los tres grupos de edad en un solo data frame
bicep_por_edades <-
as.data.frame(rbind(bicep_menores_30,bicep_entre_30_y_50,bicep_mayores_de
_50))

#Nombrar las columnas
colnames(bicep_por_edades) <- c("Longitud del bicep en cm", "categoria")
```

```
bicep_por_edades[,"Longitud del bicep en cm"]<-
as.numeric(as.character(bicep_por_edades[,"Longitud del bicep en cm"]))
head(bicep_por_edades)
```

En la **Figura 6,** se muestra como quedo el encabezado del dataframe, que se armó de la variable Bicep, categorizada por grupos de edades.

Figura 6. Encabezado de la variable Bicep categorizada por grupos de edad

```
> head(bicep_por_edades)
  Longitud del bicep en cm          categoria
1                     32.0 Menores de 30 años
2                     30.5 Menores de 30 años
3                     28.8 Menores de 30 años
4                     32.4 Menores de 30 años
5                     32.2 Menores de 30 años
6                     35.7 Menores de 30 años
```

4.3.1 Descripción del grupo de edad menores de 30 años

En este apartado se presenta una descripción tanto gráfica como numérica para la variable Bicep, restringida a las personas que tienen menos de 30 años.

4.3.1.1 Descripción numérica

En este literal se presentan las medidas de localización, dispersión, los cuantiles, el coeficiente de curtosis y de asimetría para los datos de la variable Bicep, que corresponden a las personas que tienen menos de 30 años.

El código que se presenta a continuación se utilizó para crear el vector de logitudes de Biceps, para las personas con menos de 30 años.

```
> # Se construye el vector de longitudes de Biceps para los que tienen me
nos de 30 años
> bicep_menores_de_30_años <-bicep_por_edades[bicep_por_edades$categoria=
="Menores de 30 años",c("Longitud del bicep en cm")]
> bicep_menores_de_30_años
 [1] 32.0 30.5 28.8 32.4 32.2 35.7 31.9 30.5 35.9 35.6 32.8 37.2 37.2 32.
5 33.0 31.1
[17] 29.9 30.5 30.1 30.1 29.0 30.5 31.8 33.5 36.1 33.3 25.8 36.0 31.6 38.
5 27.8 30.6
[33] 33.7 32.2 35.2 31.6
```

Y con base en el vector `bicep_menores_de_30_años` y el código que se presenta a continuación, se construyó la **Tabla 4**.

```
> # Se calculan las medidas de localización para bicep_menores_de_30_años
> #Media
> mean(bicep_menores_de_30_años)
[1] 32.41944
> #Moda
> names(sort(-table(bicep_menores_de_30_años)))[1]
[1] "30.5"
> #Mediana
> median(bicep_menores_de_30_años)
```

```
[1] 32.1
> # Se calculan las medidas de dispersión para bicep_menores_de_30_años
> #Cálculo de la varianza
> sum((bicep_menores_de_30_años - mean(bicep_menores_de_30_años))^2) / le
ngth(bicep_menores_de_30_años)
[1] 7.8599
> #Cálculo de la desviación estándar
> sqrt(sum((bicep_menores_de_30_años - mean(bicep_menores_de_30_años))^2)
/ length(bicep_menores_de_30_años))
[1] 2.803551
> #Cálcuo de la desviación media
> sum(abs(bicep_menores_de_30_años-mean(bicep_menores_de_30_años)))/lengt
h(bicep_menores_de_30_años)
[1] 2.21713
> #Cálculo del rango
> max(bicep_menores_de_30_años)-min(bicep_menores_de_30_años)
[1] 12.7
> #Coeficiente de variación
> sqrt(sum((bicep_menores_de_30_años - mean(bicep_menores_de_30_años))^2)
/ length(bicep_menores_de_30_años))/mean(bicep_menores_de_30_años)
[1] 0.08647746
> # Se calculan los cuartiles
> quantile(bicep_menores_de_30_años)
    0%    25%    50%    75%   100%
25.800 30.500 32.100 34.075 38.500
> #Cálculo del coeficiente de asimetría
> library(moments)
> skewness(bicep_menores_de_30_años)
[1] 0.1504847
> #Cálculo del coeficiente de de apuntamiento o curtosis
> library(moments)
> kurtosis(bicep_menores_de_30_años)
[1] 2.71590
```

Tabla 4. Medidas descriptivas númericas de la variable Bicep para el grupo de edad de menores de 30 años

Medida	**Valor**
Media	32.41944
Mediana	32.1
Moda	30.5
Varianza	7.8599
Desviación media	2.21713
Desviación estándar	2.803551
Coeficiente de variación	0.08647746
Rango	12.7
Q_L	30.5
Q_C	32.1
Q_U	34.075
IQR	3.575
Coeficiente de asimetría	0.1504847
Coeficiente de Curtosis	2.71590

Con base en los valores de la media, la mediana y la moda de la **Tabla 4**, se intuye, que la distribución de los datos de la variable "Bicep", para personas menores a 30 años, es aproximadamente simétrica, ya que estos tres valores son muy cercanos.

Observando los valores de, varianza, desviación media, desviación estándar y coeficiente de variación, se aprecia, que son relativamente bajos, lo que invita a pensar que la distribución de los datos de la longitud del Bicep, de los 250 hombres, no presenta una variabilidad o dispersión muy alta con respecto a la media.

Con base, en los cuantiles, se sabe que para la distribución de datos de la variable Bicep, para personas con menos de 30 años, el cuartil Q_L indica que el 25% de las observaciones se encuentran por debajo, de 30.5, que el 50% de los datos están por debajo de 32.1, y que el 75% están por debajo de 34.075.

El coeficiente de asimetría para el conjunto de datos, de la variable Bicep acotada ala grupo de edad de menores de 30 años, reporto un valor de 0.1504847, este valor es cercano a cero, esto indica que la distribución tiende a ser simétrica.

Por último, para la variable Biceps, delimitada a los hombres menores de 30 años, reportó un coeficiente de curtosis de 2.71590, este valor es cercano a tres, lo que indica que existe un grado de dispersión moderado con respecto a la media, es decir, gran cantidad de observaciones se acumulan cerca a media de la distribución.

4.3.1.2 Descripción Gráfica

En este literal, se presenta la **Figura 7** que hace referencia al histograma para los datos de la variable Bicep, acotados al grupo de edad, de los hombres que tienen menos de 30 años y en la **Figura 8,** se presenta un diagrama de caja y brazos.

Figura 7. Histograma de la variable Bicep para los menores de 30 años

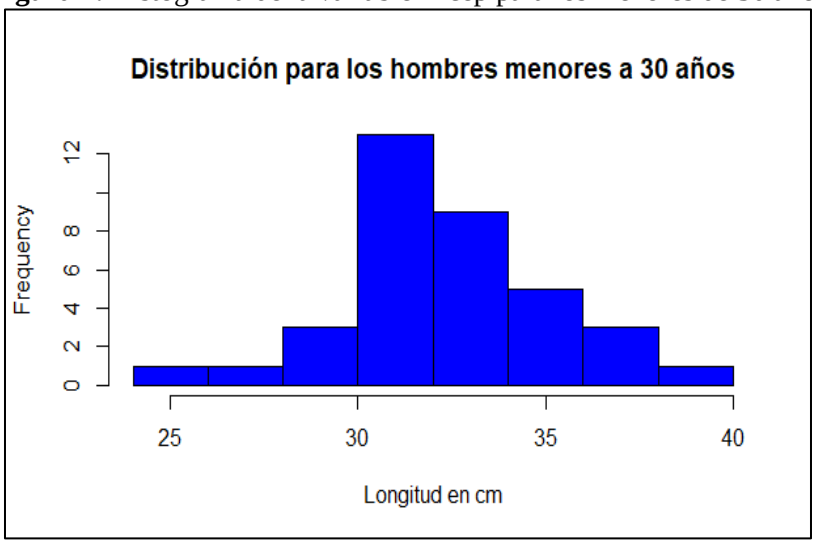
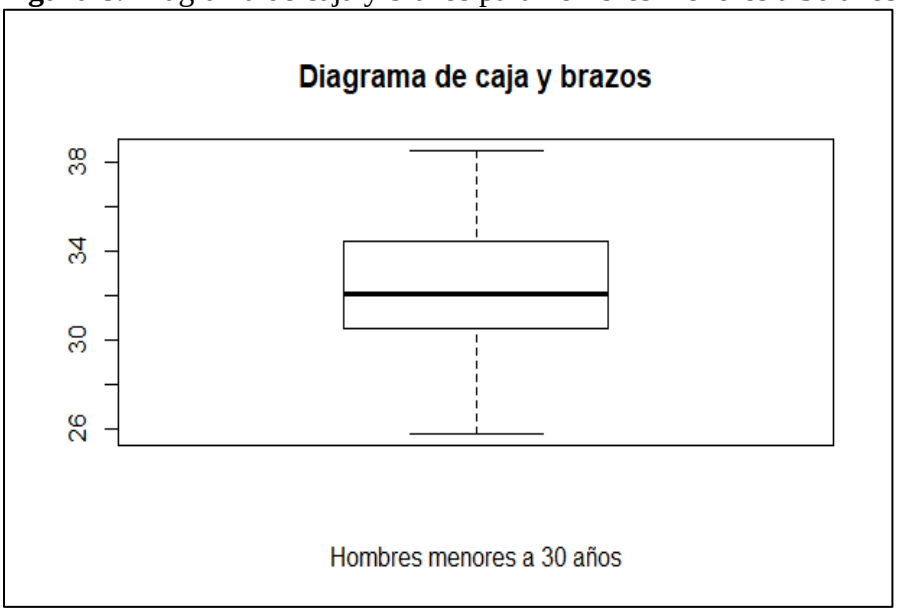


Figura 8. Diagrama de caja y brazos para hombres menores a 30 años



Con base en el histograma de la **Figura 7**, se aprecia, que la distribución es aproximadamente simétrica y observando la **Figura 8**, se identifica que no hay datos atípicos en el conjunto de datos restringido a los hombres menores de 30 años.

Con las líneas de código que se presentan a continuación se dibujaron las **Figuras 7** y **8**.

```
> hist(bicep_menores_de_30_años, main = "Distribución para los hombres me
nores a 30 años", xlab = "Longitud en cm", col="blue")

boxplot(bicep_menores_de_30_años, main = "Diagrama de caja y brazos", xla
b = "Hombres menores a 30 años")
```

4.3.2 Descripción para el grupo de edad entre 30 y 50 años

En este apartado se presenta una descripción tanto gráfica como numérica para la variable Bicep, restringida a las personas que tienen entre 30 y 50 años.

4.3.2.1 Descripción numérica

En este numeral se presentan las medidas de localización, dispersión, los cuantiles, el coeficiente de curtosis y de asimetría para los datos de la variable Bicep, que corresponden a las personas que tienen entre 30 y 50 años.

El código que se presenta a continuación se utilizó para crear el vector de logitudes de Biceps, para las personas que tienen entre 30 y 50 años.

```
> # Se construye el vector de longitudes de Biceps para los que tienen me
nos de 30 años
> bicep_entre_30_y_50 <-bicep_por_edades[bicep_por_edades$categoria=="Ent
re 31 y 50 años",c("Longitud del bicep en cm")]
```

Y con base en el vector `bicep_entre_30_y_50` y el código que se presenta a continuación, se construyó la **Tabla 5**.

```
> # Se calculan las medidas de localización para bicep_menores_de_30_años
>
> #Media
> mean(bicep_entre_30_y_50)
[1] 32.27984
> #Moda
> names(sort(-table(bicep_entre_30_y_50)))[1]
[1] "31"
> #Mediana
> median(bicep_entre_30_y_50)
[1] 32
>
> # Se calculan las medidas de dispersión para bicep_menores_de_30_años
> #Cálculo de la varianza
> sum((bicep_entre_30_y_50 - mean(bicep_entre_30_y_50))^2) / length(bicep
_entre_30_y_50)
[1] 9.126726
>
> #Cálculo de la desviación estándar
> sqrt(sum((bicep_entre_30_y_50 - mean(bicep_entre_30_y_50))^2) / length(
bicep_entre_30_y_50))
[1] 3.021047
>
> #Cálcuo de la desviación media
> sum(abs(bicep_entre_30_y_50-mean(bicep_entre_30_y_50)))/length(bicep_en
tre_30_y_50)
[1] 2.477351
>
> #Cálculo del rango
> max(bicep_entre_30_y_50)-min(bicep_entre_30_y_50)
[1] 14.3
>
> #Coeficiente de variación
> sqrt(sum((bicep_entre_30_y_50 - mean(bicep_entre_30_y_50))^2) / length(
bicep_entre_30_y_50))/mean(bicep_entre_30_y_50)
[1] 0.09358927
>
> # Se calculan los cuartiles
> quantile(bicep_entre_30_y_50)
  0%  25%  50%  75% 100%
24.8 30.2 32.0 34.4 39.1
>
> #Cálculo del coeficiente de asimetría
> library(moments)
> skewness(bicep_entre_30_y_50)
[1] 0.05713119
>
> #Cálculo del coeficiente de de apuntamiento o curtosis
> library(moments)
> kurtosis(bicep_entre_30_y_50)
[1] 2.450999
```

Con base en los valores de la media, la mediana y la moda de la **Tabla 5**, se intuye, que la distribución de los datos de la variable "Bicep", para personas entre 30 y 50 años, es aproximadamente simétrica, ya que estos tres valores son muy similares entre sí.

Observando los valores de, varianza, desviación media, desviación estándar y coeficiente de variación, de la **Tabla 5** se aprecia, que son relativamente bajos, lo que incita a pensar que la

distribución de los datos de la longitud del Bicep, de los 250 hombres, no presenta una variabilidad o dispersión muy alta con respecto a la media, para las personas que tienen entre 30 y 50 años.

Con base, en los cuantiles, de la **Tabla 5** se sabe que para la distribución de datos de la variable Bicep, para personas entre 30 y 50 años, el cuartil Q_L indica que el 25% de las observaciones se encuentran por debajo, de 30.2, que el 50% de los datos están por debajo de 32, y que el 75% están por debajo de 34.4.

El coeficiente de asimetría para el conjunto de datos, de la variable Bicep acotada ala grupo de edad entre 30 y 50 años, reporto un valor de0.05713119, este valor es muy cercano a cero, esto indica que la distribución tiende a ser simétrica.

Por último, para la variable Biceps, delimitada a los hombres menores de 30 años, reportó un coeficiente de curtosis de 2.450999, este valor es próximo a tres, lo que muestra que existe un grado de dispersión moderado con respecto a la media, es decir, gran cantidad de observaciones se acumulan cerca a la media de la distribución.

Tabla 5. Medidas descriptivas numéricas de la variable Bicep para el grupo de edad entre 30 y 50 años

Medida	**Valor**
Media	32.27984
Mediana	31
Moda	32
Varianza	9.126726
Desviación media	2.477351
Desviación estándar	3.021047
Coeficiente de variación	0.09358927
Rango	14.3
$\boldsymbol{Q_L}$	30.2
$\boldsymbol{Q_C}$	32.0
$\boldsymbol{Q_U}$	34.4
$\boldsymbol{IQR}$	4.2
Coeficiente de asimetría	0.05713119
Coeficiente de Curtosis	2.450999

4.3.2.2 Descripción Gráfica

En este literal, se presenta la **Figura 9** que hace referencia al histograma para los datos de la variable Bicep, acotados al grupo de edad, de los hombres que tienen entre 30 y 50 años y en la **Figura 10,** se presenta un diagrama de caja y brazos.

Figura 9. Histograma para el grupo de edad entre 30 y 50 años

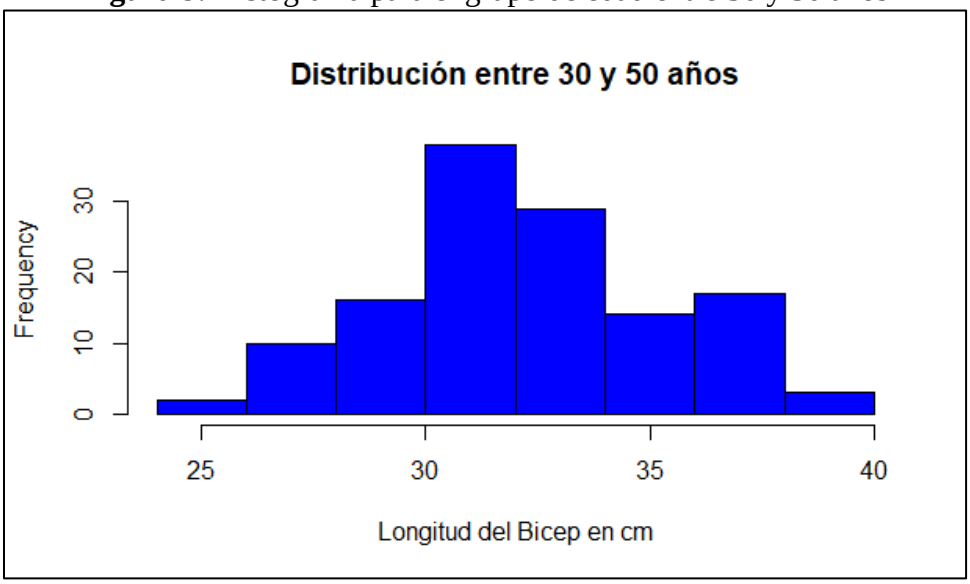
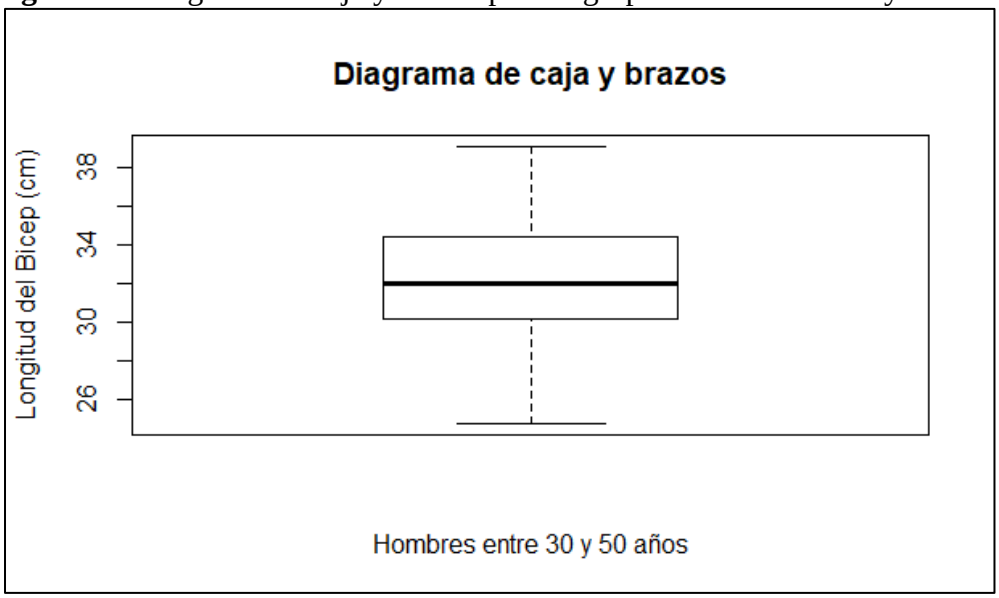


Figura 10. Diagrama de caja y brazos para el grupo de edad entre 30 y 50 años



Con base en el histograma de la **Figura 9**, se aprecia, que la distribución es aproximadamente simétrica y observando la **Figura 10**, se identifica que no hay datos atípicos en el conjunto de datos restringido a los hombres entre 30 y 50 años.

A continuación, se presentan las líneas de código usadas para construir las **Figuras 9** y **10.**

```
> hist(bicep_entre_30_y_50, main = "Distribución entre 30 y 50 años", xla
b = "Longitud del Bicep en cm", col="blue")
> boxplot(bicep_entre_30_y_50, main = "Diagrama de caja y brazos", xlab =
"Hombres entre 30 y 50 años",ylab="Longitud del Bicep (cm)")
```

4.3.3 Descripción para el grupo de edad de mayores de 50 años

En este apartado se presenta una descripción tanto gráfica como numérica para la variable Bicep, restringida a las personas que tienen más de 50 años.

4.3.3.1 Descripción numérica

En este literal se presentan las medidas de localización, dispersión, los cuartiles, el coeficiente de curtosis y de asimetría para los datos de la variable Bicep, que corresponden a las personas que tienen más de 50 años.

El código que se presenta a continuación se utilizó para crear el vector de logitudes de Biceps, para las personas con más de 50 años.

```
 > # Se construye el vector de longitudes de Biceps para los que tienen m
ás de 50 años
> bicep_mayores_de_50 <-bicep_por_edades[bicep_por_edades$categoria=="May
ores de 50",c("Longitud del bicep en cm")]
```

Y con base en el vector `bicep_mayores_de_50` y el código que se presenta a continuación, se construyó la **Tabla 6**.

```
> #Media
> mean(bicep_mayores_de_50)
[1] 32
> #Moda
> names(sort(-table(bicep_mayores_de_50)))[1]
[1] "29.4"
> #Mediana
> median(bicep_mayores_de_50)
[1] 31.75
> # Se calculan las medidas de dispersión para bicep_menores_de_30_años
> #Cálculo de la varianza
> sum((bicep_mayores_de_50 - mean(bicep_mayores_de_50))^2) / length(bicep
_mayores_de_50)
[1] 7.690789
> #Cálculo de la desviación estándar
> sum(abs(bicep_mayores_de_50-mean(bicep_mayores_de_50)))/length(bicep_ma
yores_de_50)
[1] 2.281579
> #Cálculo del rango
> max(bicep_mayores_de_50)-min(bicep_mayores_de_50)
[1] 13.1
> #Coeficiente de variación
> sqrt(sum((bicep_mayores_de_50 - mean(bicep_mayores_de_50))^2) / length(
bicep_mayores_de_50))/mean(bicep_mayores_de_50)
[1] 0.08666335
> # Se calculan los cuartiles
> quantile(bicep_mayores_de_50)
    0%    25%    50%    75%   100%
```

```
25.300 29.775 31.750 34.150 38.400
> #Cálculo del coeficiente de asimetría
> library(moments)
> skewness(bicep_mayores_de_50)
[1] -0.005085902
> #Cálculo del coeficiente de de apuntamiento o curtosis
> library(moments)
> kurtosis(bicep_mayores_de_50)
[1] 2.608249
>
> sum(abs(bicep_mayores_de_50-mean(bicep_mayores_de_50)))/length(bicep_ma
yores_de_50)
[1] 2.281579
```

Tabla 6. Medidas descriptivas numéricas de la variable Bicep para el grupo de edad mayor a 50 años

Medida	Valor
Media	32
Mediana	31.75
Moda	29.4
Varianza	7.690789
Desviación media	2.281579
Desviación estándar	2.281579
Coeficiente de variación	0.08666335
Rango	13.1
Q_L	29.775
Q_C	31.750
Q_U	34.150
IQR	4.375
Coeficiente de asimetría	-0.005085902
Coeficiente de Curtosis	2.608249

Con base en los valores de la media, la mediana y la moda de la **Tabla 6**, se vislumbra, que la distribución de los datos de la variable "Bicep", para personas mayores de 50 años, es aproximadamente simétrica, ya que estos tres valores son muy similares entre sí.

Observando los valores de, varianza, desviación media, desviación estándar y coeficiente de variación, de la **Tabla 6** se aprecia, que son relativamente bajos, lo que indica que la distribución de los datos de la longitud del Bicep, de los 250 hombres, no presenta una variabilidad o dispersión muy alta con respecto a la media, para las personas que tienen más de 50 años.

Con base, en los cuantiles, de la **Tabla 6** se sabe que para la distribución de datos de la variable Bicep, para personas mayores a 50 años, el cuartil Q_L indica que el 25% de las observaciones se encuentran por debajo, de 29.775, que el 50% de los datos están por debajo de 31.75, y que el 75% están por debajo de 34.150.

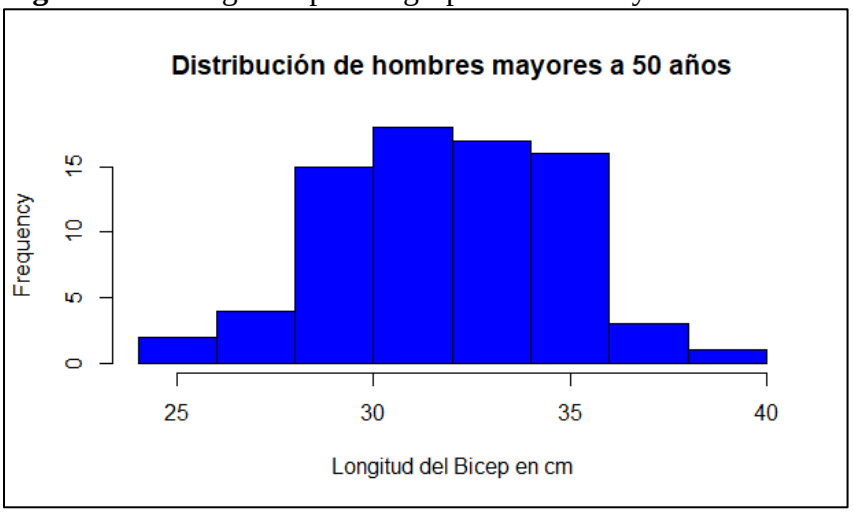
El coeficiente de asimetría para el conjunto de datos, de la variable Bicep acotada al grupo de edad entre 30 y 50 años, reporto un valor de -0.005085902, este valor es muy cercano a cero, esto indica que la distribución tiende a ser simétrica.

Por último, para la variable Biceps, delimitada a los hombres menores de 30 años, reportó u n coeficiente de curtosis de 2.608249, este valor es próximo a tres, lo que muestra que exist e un grado de dispersión moderado con respecto a la media, es decir, gran cantidad de obser vaciones se acumulan cerca a la media de la distribución.

4.3.3.2 Descripción gráfica

En este literal, se presenta la **Figura 11** que hace referencia al histograma para los datos de la variable Bicep, acotados al grupo de edad, de los hombres que tienen más de 50 años y en la **Figura 12,** se presenta un diagrama de caja y brazos.

Figura 11. Histograma para el grupo de edad mayores de 50 años



Con base en el histograma de la **Figura11**, se aprecia, que la distribución es aproximadamente simétrica y observando la **Figura 12**, se identifica que no hay datos atípicos en el conjunto de datos restringido al grupo de edad de mayores de 50 años.

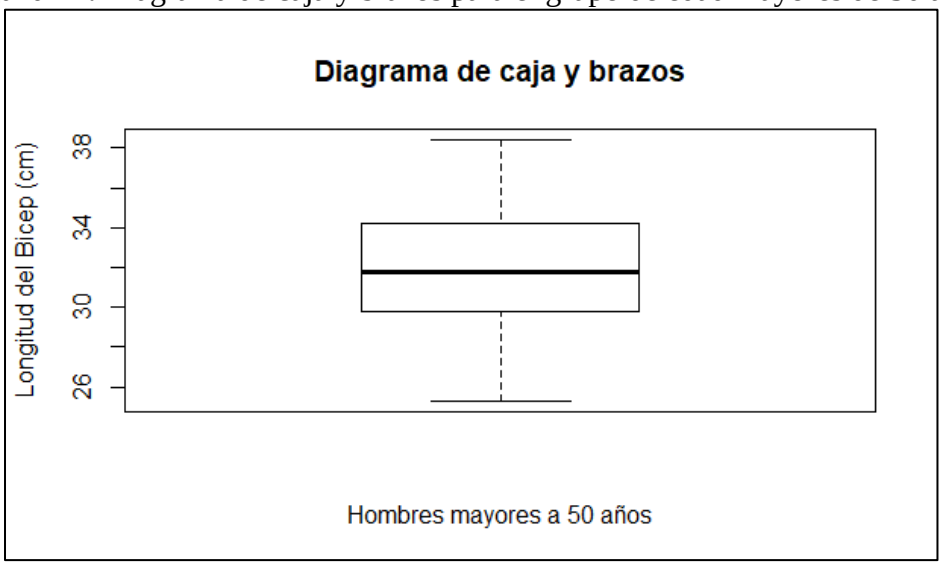
Las **Figuras 11** y **12**, se realizaron con las líneas de código que se muestran a continuación:

```
> hist(bicep_mayores_de_50, main = "Distribución de hombres mayores a 50
años", xlab = "Longitud del Bicep en cm", col="blue")
> boxplot(bicep_mayores_de_50, main = "Diagrama de caja y brazos", xlab =
"Hombres mayores a 50 años",ylab="Longitud del Bicep (cm)")
```

4.3.4 Análisis y conclusiones de los resultados obtenidos en los tres grupos de edad

Al observar la **Figura 13**, se aprecia que, en ninguno de los tres grupos de edad, existen datos atípicos y que, en los tres casos, la distribución de los datos es similares. Al observar las Tablas **4**, **5** y **6**, se aprecia que tanto las medidas de localización, dispersión, posición como los coeficientes de curtosis y asimetría, dieron muy cercanos, lo que es también un indicador de la semejanza en la distribución de los datos para los grupos de edad trabajados.

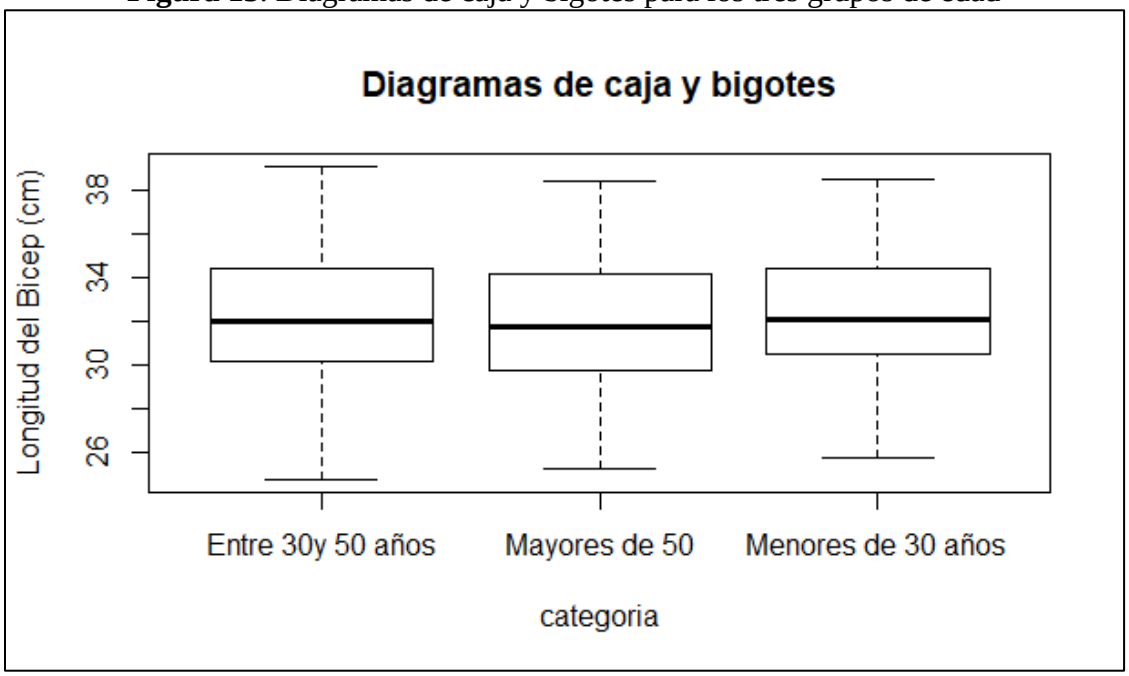
Figura 12. Diagrama de caja y brazos para el grupo de edad mayores de 50 años



Y al observar los coeficientes de variación de las tres distribuciones, aunque todos son muy cercanos, el del grupo de edad entre 30 y 50 años, arrojo un valor de 0.09358927, que es superior al de los otros dos subconjuntos de datos, entonces se aprecia, que este seria el subconjunto de datos, que más variabilidad o dispersión en los datos, tiene con respecto a la media.

Por último, con base la comparación realizada entre grupos de edad, y la similitud obtenida en las distribuciones, por grupo de edad, se aprecia que la longitud en el Bicep de un hombre no sufre grandes cambios a medida que aumenta su edad.

Figura 13. Diagramas de caja y bigotes para los tres grupos de edad



4.4 Solución literal d

Para solucionar este literal, se construyó una matriz de correlación, para medir el grado de dependencia lineal entre todas las variables de la base de datos y el porcentaje de grasa corporal. La matriz se construyó con base en el siguiente código,

```
> matriz_correlacion <- cor(bodyfat_sin_density)
> correlaciones_vs_bf <- matriz_correlacion[, "Pct.BF"]
> correlaciones_vs_bf <- Filter(function(x) x != 1, correlaciones_vs_bf)
> corrplot(matriz_correlacion,method="square")
```

Y se obtuvieron los datos de la **Tabla 7**, donde se aprecia que las variables Abdomen y Waist (cintura), tienen una fuerte relación lineal, con el porcentaje de masa corporal de una persona. Lo descrito en el presente párrafo, es bastante lógico, ya que biológicamente los hombres, tienden a acumular la mayor cantidad de grasa en estas dos zonas. Entonces, a mayores valores en las variables Abdomen o Waist en un hombre se espera un mayor porcentaje de grasa corporal, ya que se identificó una relación lineal positiva tanto entre porcentaje de masa corporal y Abdomen, y como entre porcentaje de grasa corporal y Waist.

Tabla 7. Coeficientes de correlación entre el porcentaje de grasa corporal y las demás variables

Variable	Coeficiente de correlación
Age	0.29
Weight	0.61
Height	-0.02
Neck	0.48
Chest	0.70
Abdomen	0.82
Waist	0.82
Hip	0.63
Thigh	0.54
Knee	0.49
Ankle	0.24
Bicep	0.48
Forearm	0.36
Wrist	0.33

5. Ejercicio Cuatro

El archivo tattoos del sitio web https://dasl.datadescription.com/datafiles/ contiene los conteos de 626 individuos categorizados de acuerdo con su calidad de tatuado (tattoo status) y su estado de infectado de hepatitis C ("hepatitis status"). Evalúe y cuantifique, de existir, el nivel de asociación entre las variables involucradas en el estudio.

5.1 Carga y visualización de los datos

En este apartado se explica como se hizo la carga de datos a R y se describen los datos de la base de datos utilizada, en esta práctica.

5.1.1 Carga de los datos

Para realizar la carga de los datos lo primero que debemos hacer es descargar el archivo tattos.txt, usar el comando read.delim para subir los datos a R y asignarlos a la variable tattoos,

```
tattoos <- read.delim("tattoos.txt")
```

5.1.2 Visualización de los datos

Para fines exploratorios solo visualizaremos los primeros 6 datos de la data frame usando el comando head, para darnos una idea general de la composición de los datos.

```
head(tattoos)
```

	Location <chr>	Has.hepatitis.C <chr>
1	Commercial Parlor	Yes
2	Commercial Parlor	Yes
3	Commercial Parlor	Yes
4	Commercial Parlor	Yes
5	Commercial Parlor	Yes
6	Commercial Parlor	Yes

5.2 Creación de la tabla de contingencia

En este numeral se muestran los códigos y metodología utilizada para establecer la tabla de contingencia.

5.2.1 Frecuencias en entre variables

La tabla de contingencia muestra las frecuencias absolutas observadas para la combinación de cada categoría.

```
> tcon  <- table(tattoos);tcon
                  Has.hepatitis.C
Location            No Yes
  Commercial Parlor  35  17
  Elsewhere          53   8
  No Tattoo         491  22
```

Gracias a las frecuencias podemos ver que casi el 50% de las personas que se tatuaron en establecimientos comerciarles tienen hepatitis C.

5.2.2 Tabla de contingencia

Para la creación de la tabla de contingencia completa debemos añadir los marginales es decir que debemos añadir una columna y una fila con la cantidad de observaciones para “x” y para “y”.

```
> tcon <- addmargins(tcon);tcon
                  Has.hepatitis.C
Location            No Yes Sum
  Commercial Parlor  35  17  52
  Elsewhere          53   8  61
  No Tattoo         491  22 513
  Sum               579  47 626
```

Fácilmente se evidencia que la mayoría de los individuos categorizados no tiene tatuajes variables “x” y no tienen hepatitis C variable “y”.

5.3. Distribuciones de frecuencias relativas

```
> addmargins(prop.table(table(tattoos)))*100
                  Has.hepatitis.C
Location                    No       Yes       Sum
  Commercial Parlor   5.591054  2.715655  8.306709
  Elsewhere           8.466454  1.277955  9.744409
  No Tattoo          78.434505  3.514377 81.948882
  Sum                92.492013  7.507987 100.000000
```

Dadas las frecuencias relativas marginales es posible evidencia que apenas el 7.51% de los individuos categorizados tienen hepatitis C, de los cuales el 2.71% se ha tatuado en un local comercial y apenas el 1.27% lo ha hecho en otro lugar.

5.3.1 Distribuciones de frecuencias relativas por filas

```
> prop.table(table(tattoos),1)*100
                  Has.hepatitis.C
Location                    No       Yes
  Commercial Parlor  67.307692 32.692308
  Elsewhere          86.885246 13.114754
  No Tattoo          95.711501  4.288499
```

Usando esta distribución podemos ver que la mayoría de los individuos no está infectada, sin embargo, es evidente que existe una tasa mayor de individuos infectados de Hepatitis C que son tatuados en un establecimiento comercial llegando a ser esta del 32.7% del total de los tatuados en estos lugares

5.3.2 Distribuciones de frecuencias relativas por columnas

```
> prop.table(table(tattoos),2)*100
                    Has.hepatitis.C
Location                        No       Yes
  Commercial Parlor       6.044905 36.170213
  Elsewhere               9.153713 17.021277
  No Tattoo              84.801382 46.808511
```

Usando la distinción por columnas es claro que la mayoría de las personas no infectadas con Hepatitis C no poseen ningún tatuaje siendo estas el 85% de la población no infectada, sin embargo, entre los individuos infectados los porcentajes son muy similares independientemente si están tatuados o no.

5.4 Diagramas de barra

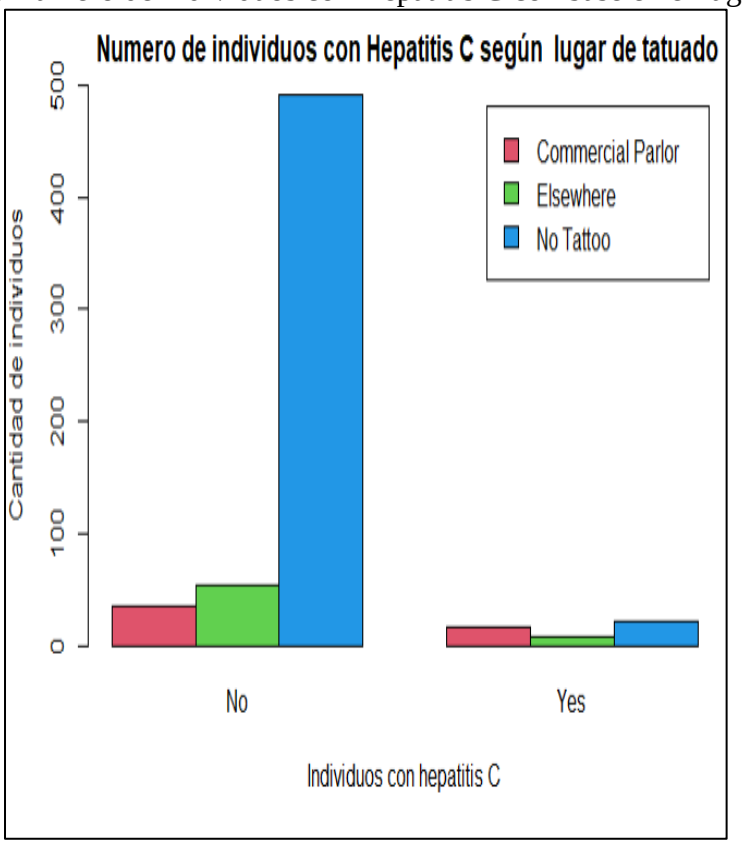
Podemos observar fácilmente las afirmaciones anteriores mediante un diagrama de barras en el cual se muestra que la mayoría de los individuos no están infectados con Hepatitis C y no poseen ningún tatuaje, y que la cantidad de individuos infectados es muy similar independientemente si el individuo esta tatuado o no o de en donde se realizó su tatuaje.

```
> barplot(table(tattoos), beside=T, legend=T, col=c(2,3,4), xlab= "Individuos
con hepatitis C", ylab= "Cantidad de individuos", main=" Numero de individuos
con Hepatitis C según  lugar de tatuado")
```

5.5 Medidas de asociación

Para verificar el nivel de asociación entre 2 variables es necesario construir una tabla de contingencia 2x2, para esto simplificaremos nuestra tabla inicial de tal manera que solo contenga dos categorías para “x” y dos categorías para “y”. Para este caso en particular una de la variable “x” (Location), se categorizará entre los individuos que tienen o no tatuajes y su correspondiente distribución marginal.

Figura 14. Número de individuos con Hepatitis C con base en el lugar tatuado



5.5.1 Llamado a la tabla de contingencia.

Para realizar este proceso no es necesario eliminar la columna y la fila de la distribución marginal, de tal manera que solo se tenga en cuenta las frecuencias relativas.

```
> tcon <- table(tattoos);tcon
                  Has.hepatitis.C
Location            No Yes
  Commercial Parlor  35  17
  Elsewhere          53   8
  No Tattoo         491  22
```

5.5.2 Creación de nuevas categorías para la variable x

Para la creación de las nuevas categorías que definan a la variable “x”, tal como se había indicado anteriormente se sumarán las filas de las personas que se tatuaron en “Commercial Parlor” y las que se tatuaron en “Elsewhere”, tal que:

```
> tatuados <- colSums(tcon[1:2,]);tatuados
 No Yes
 88  25
```

La categoría de no tatuados para la variable “x”, se dejará tal como esta:

```
> noTatuados <- tcon[3,];noTatuados
 No Yes
491  22
```

Luego de esto se creará una nueva tabla de contingencia con sus correspondientes marginales

```
> tcon <- rbind(tatuados,noTatuados);tcon
            No Yes
tatuados    88  25
noTatuados 491  22
> tcon <- addmargins(tcon);tcon
            No Yes Sum
tatuados    88  25 113
noTatuados 491  22 513
Sum        579  47 626
```

- $\%\ Individuos\ tatuados = \frac{113}{626} = 0.18$
- $\%\ Individuos\ tatuados\ infectados = \frac{25}{113} = 0.22$

Por medio de esta nueva tabla de contingencia podemos ver que el 18% de los individuos del grupo esta tatuado y que de estos el 22% está infectado con Hepatitis C.

5.6 Verificación de la independencia

Para que ambas variables sean independientes se debe cumplir que:

$$f_{1|1} = \frac{(f_{1|1} + f_{1|2})(f_{1|1} + f_{2|1})}{n} = tcon[1,1] == \frac{tcon[1,3] * tcon[3,1]}{tcon[3,3]}$$

Ejecutando el código en R

```
> tcon[1,1]==(tcon[1,3]*tcon[3,1])/tcon[3,3]
[1] FALSE
```

Pues $tcon[1,1]$ es igual a 88

```
> tcon[1,1]
[1] 88
```

Y, $\frac{tcon[1,3]*tcon[3,1]}{tcon[3,3]}$ es igual a 104.516

```
> (tcon[1,3]*tcon[3,1])/tcon[3,3]
[1] 104.516
```

Este resultado nos permite concluir que las variables no son independientes entre sí, es decir existe una asociación entre tener un tatuaje y estar infectado de hepatitis C, en otras palabras, no haberse tatuado no es independiente de no tener hepatitis C, sin embargo, el hecho de no tener tatuaje está disminuyendo la posibilidad de no tener hepatitis C, dentro del grupo de individuos categorizados.

5.6.1 Análisis de Riesgos relativos.

Calcularemos los riesgos relativos para establecer la relación entre nuestras variables "x" y "y", nuevamente con nuestra tabla de 2x2 categorías, para esto compararemos las distribuciones condicionales.

5.6.1.1. Comportamiento de los individuos tatuados respecto a los que tienen o no hepatitis C

$$\frac{f_{1|1}}{f_{1|2}} = \frac{n_{11}/n_{+1}}{n_{12}/n_{+2}} = \frac{No\ infectados}{Infectados} = \frac{88 * 47}{25 * 579} = 0.2857 \cong 0.3$$

Este resultado nos permite decir que entre los individuos tatuados el riesgo de estar infectado es del 70% aproximadamente.

5.6.1.2. Comportamiento de los individuos **NO tatuados** respecto a los que tienen o no hepatitis C:

$$\frac{f_{2|1}}{f_{2|2}} = \frac{n_{11}/n_{+1}}{n_{12}/n_{+2}} = \frac{No\ infectados}{Infectados} = \frac{491 * 47}{22 * 579} = 1.8116 \cong 2$$

Este resultado nos permite decir que entre las personas no tatuadas el riesgo de no estar infectado con hepatitis C es 2 veces mayor.

5.6.2. Análisis de la Razón de riesgos

$$OR = \frac{f_{1|1}/f_{1|2}}{f_{2|1}/f_{2|2}} = \frac{Tatuados\ no\ infectados}{No\ tatuados\ no\ \ infectados} = \frac{88 * 22}{25 * 491} = 0.1577 \cong 0.16$$

Es decir que por cada 16 personas que no se infectan con Hepatitis C siendo tatuadas, hay 100 que no se infectan sin ser tatuadas, es decir que hay 6.34 más posibilidades de no estar infectado sin ser tatuado que siendo tatuado.

5.7 Análisis para las personas tatuadas en diferentes lugares.

Para este punto analizaremos la asociación entre estar o no infectado de Hepatitis C y el lugar donde se realizó el tatuaje el individuo categorizado.

5.7.1 Construcción de la tabla de contingencia

```
> tcon <- addmargins(tcon[1:2,]);tcon
                    Has.hepatitis.C
Location              No Yes Sum
  Commercial Parlor   35  17  52
  Elsewhere           53   8  61
  Sum                 88  25 113
```

- $\%\ del\ total\ de\ Individuos\ tatuados\ en\ lugares\ comerciales = \frac{52}{113} = 0.46$
- $\%\ del\ total\ de\ \ Individuos\ tatuados\ en\ lugares\ comerciales = \frac{17}{113} = 0.15$

Con esta tabla de contingencia se observa que el 46% de los individuos tatuados lo hicieron en lugares comerciales, y dentro de esto el 15% está infectado con Hepatitis C.

5.7.2 Verificación de independencia

$$f_{1|1} = \frac{(f_{1|1} + f_{1|2})(f_{1|1} + f_{2|1})}{n} = tcon[1,1] == \frac{tcon[1,3] * tcon[3,1]}{tcon[3,3]}$$

```
> tcon[1,1]==(tcon[1,3]*tcon[3,1])/tcon[3,3]
[1] FALSE
> tcon[1,1]
[1] 35
> (tcon[1,3]*tcon[3,1])/tcon[3,3]
[1] 40.49558
```

Mediante este resultado validamos que las variables no son independientes entre sí, en otras palabras, haberse tatuado en los lugares comerciales no es independiente de no tener Hepatitis C, no obstante, si se puede notar un cierto nivel de asociación entre ambas variables debido a la cercanía del resultado de la prueba de independencia al valor $f_{1|1}$.

5.7.1 Análisis de Riesgos relativos.

5.7.1.1. Comportamiento de los individuos tatuados en "Establecimientos comerciales" respecto a los que tienen o no hepatitis C:

$$\frac{f_{1|1}}{f_{1|2}} = \frac{n_{11}/n_{+1}}{n_{12}/n_{+2}} = \frac{No\ infectados}{Infectados} = \frac{35 * 25}{17 * 88} = 0.5848 \cong 0.6$$

Es decir que para los individuos que se tatuaron en un establecimiento comercial hay un riesgo de un 40% de estar infectado de Hepatitis C, en otras palabras 6 de cada 10 individuos tatuados en un establecimiento comercial no estará infectado.

5.7.1.2. Comportamiento de los individuos tatuados en "otros lugares" respecto a los que tienen o no hepatitis C:

$$\frac{f_{2|1}}{f_{2|2}} = \frac{n_{21}/n_{+1}}{n_{22}/n_{+2}} = \frac{No\ infectados}{Infectados} = \frac{53 * 25}{8 * 88} = 1.882 \cong 2$$

Bajo este resultado podemos decir que el riesgo de no estar infectado con Hepatitis C es 2 veces mayor si el individuo se tatúa en un lugar diferente a un establecimiento comercial.

5.7.2. Análisis de la Razón de riesgos

$$OR = \frac{f_{1|1}/f_{1|2}}{f_{2|1}/f_{2|2}} = \frac{No\ infec.\ tatuados\ en\ loc.comerciales}{No\ infec.\ tatuados\ en\ otros\ lugares} = \frac{35 * 8}{17 * 53} = 0.317 \cong 0.3$$

Es decir que, por cada 3 personas no infectadas con Hepatitis C tatuadas en un lugar comercial, habrá 10 personas tatuadas en otros lugares que no estén infectadas. Es decir que es 3.3 veces menos probable infectarse si el individuo se tatúa en otro lugar que en un establecimiento comercial.

Conclusión

En términos generales se puede decir que hay un menor riesgo a contraer Hepatitis C, si el individuo no se encuentra tatuado a que, si lo está, siendo 6.34 más riesgoso infectarse estando tatuado que no estarlo. De igual manera se puede afirmar que existe una relación débil entre estar infectado con Hepatitis C y haber sido tatuado en un lugar comercial, a tal punto que hay 3.33 veces más riesgo de estar infectado de hepatitis C cuando el individuo fue tatuado en un lugar comercial a que si fuera tatuado en otro lugar.

6. Ejercicio cinco

La tabla adjunta indica la proporción de su renta, en euros, que una muestra de hogares se gasta en alimentación.

Tabla 8. Datos de proporción de renta gastada en alimentación

Renta	30	30	22	23	19	20	15	14
Proporción %	22	24	25	28	30	33	37	40

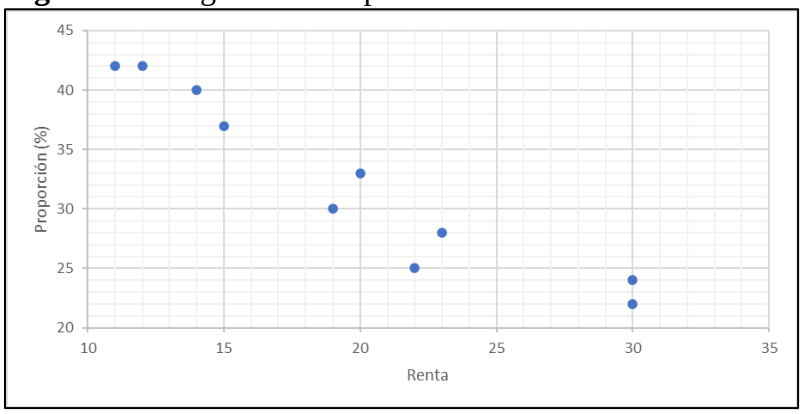
Evalúe y cuantifique, de existir, el nivel de asociación entre las variables.

6.1 Solución punto cinco

Entonces, para apreciar si existe algún grado de asociación entre las variables de la **Tabla 8**, se realizó el diagrama de dispersión de la **Figura 15**, al observar esta figura, se aprecia que al parecer si hay una dependencia entre los datos, ya que se ve como a medida que la renta aumenta la proporción que se gasta de esta en alimentación disminuye.

Gracias al diagrama de dispersión de la **Figura 15**, se identificó que existe una correlación entre las variables, ahora se desea cuantificar la intensidad de esa relación y averiguar si dicha dependencia es lineal o no. Luego, con las herramientas vistas durante el curso, podríamos emplear el coeficiente de Pearson o de Spearman.

Figura 15. Diagrama de dispersión de los datos de la **Tabla 8**



Se sabe que el coeficiente de Spearman, se recomienda principalmente cuando se tiene la presencia datos atípicos, mientras que si no se tienen observaciones extremas, entonces se puede usar el coeficiente de Pearson con seguridad, para verificar lo mencionado en el presente párrafo, se elaboraron los gráficos de caja y brazos de las **Figuras 16** y **17**, donde se aprecia que no hay presencia de datos atípicos en ninguna de las dos variables, entonces con base en eso, se decide, utilizar el coeficiente de Pearson, tanto para cuantificar el grado de asociación entre las variables como para confirmar si hay una dependencia lineal.

Figura 16. Diagrama de caja y bigotes de la variable renta

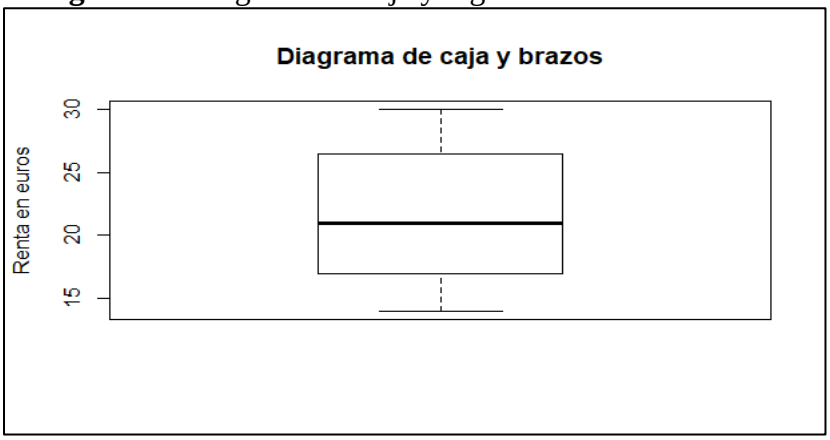
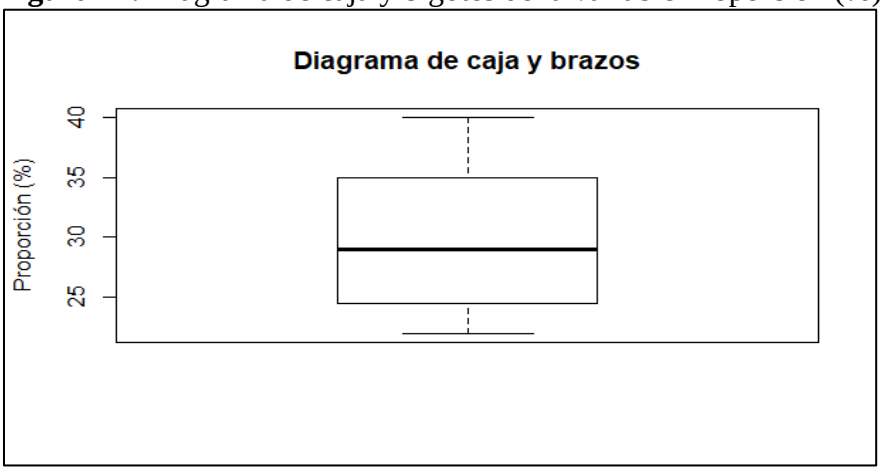


Figura 17. Diagrama de caja y bigotes de la variable Proporción (%)



Las **Figuras 16** y **17,** se elaboraron con el código que se muestra a continuación.

```
> Renta<-c(30,30,22,23,19,20,15,14,11,12)
> Proporcion<-c(22,24,25,28,30,33,37,40,42,42)
>
> boxplot(Renta, main = "Diagrama de caja y brazos",ylab="Renta en euros"
)
>
> boxplot(Proporcion, main = "Diagrama de caja y brazos",ylab="Proporción
(%)")
```

Y con el código que se muestra a continuación, se determinó el valor del coeficiente de corr elación de Pearson, obteniendo un valor de -0.9538221, este valor al ser muy cercano a -1, i ndica que las variables tienen un nivel de asociación lineal, y el signo negativo, refleja que a medida que la renta aumenta la proporción que se gasta en alimentación disminuye.

```
> Renta<-c(30,30,22,23,19,20,15,14,11,12)
> Proporcion<-c(22,24,25,28,30,33,37,40,42,42)
>
> # se calcula la covarianza entre Renta y Proporción
> sxy<-cov(Renta,Proporcion)*(9/10)
> # se calcula la desviación estándar de Renta
> sx<-sqrt(sum((Renta - mean(Renta))^2) / length(Renta))
>
> #se calcula la desviación estándar de Proporción
> sy<-sqrt(sum((Proporcion - mean(Proporcion))^2) / length(Proporcion))
>
> rxy=sxy/(sx*sy)
> rxy
[1] -0.9538221
```