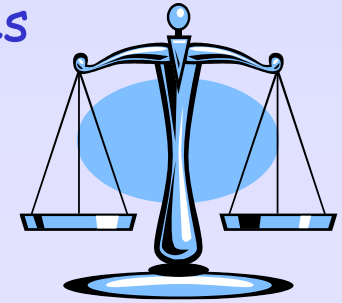
En todo proceso de investigación se generan datos y es la **Estadística** la disciplina encargada de :

Organizarlos y resumir la información

→ **Estadística Descriptiva**

Extraer conclusiones acerca de hipótesis planteadas



→ **Estadística Inferencial**

POBLACIÓN Y MUESTRA

POBLACIÓN:

- colección de elementos o sujetos de interés.
- puede ser finita o infinita.

MUESTRA:

- subconjunto elegido al azar y que sea representivo de la población.


MUESTRA
Estimar características
Inferir sobre hipótesis
POBLACIÓN

Tipos de datos de la muestra

Numéricos:

- discretos (determinados valores),

Ej: nº de hermanos, nº accidentes.

- continuos (valores en un intervalo),

Ej: concentración de glucosa en sangre.

Categóricos:

- ordinal (orden),

Ej: estado de una enfermedad (severo, moderado, suave).

- nominal (no orden),

Ej: grupo sanguineo.

Estadística Descriptiva

- Provee de métodos que permitan organizar y resumir la información de los datos.
- De acuerdo al conjunto de datos se seleccionará el método más adecuado.
- Cuando los datos son numéricos continuo algunas opciones son:

  Tablas de Distribución de frecuencias(*).

  Medidas de posición.

  Medidas de dispersión.

  Gráficos.

(*) Se puede realizar para cualquier tipo de conjunto de datos (numérico o categórico).

Tabla de distribución de frecuencia

- Tomar un intervalo que contenga al conjunto de datos.
- Dividir el intervalo en k intervalos de clase (IC) tal que sean adyacentes y disjuntos.
- Contar el número de observaciones en cada intervalo, que se llama Frecuencia Absoluta (FA).
- Calcular la Frecuencia Relativa (FR) como la FA dividida n para cada intervalo.

Observaciones:

- ¿Cómo elegir k?

  No hay reglas generales.

  Entre 5 a 20 intervalos.

  Tomar k ~ $\sqrt{n}$

- Los intervalos no tienen por que tener igual longitud.
- Además se tiene que:

$$\sum_{i=1}^{k} FA_i = n \qquad \sum_{i=1}^{k} FR_i = 1$$

Histograma

- Gráfico de mayor difusión y es la representación gráfica de la distribución de frecuencia.
- ¿Cómo hacerlo?

  - En una recta horizontal marcar los k intervalos.

  - Sobre cada intervalo trazar un rectángulo cuya área sea proporcional al número de observaciones en el mismo.

¿Cómo elegir la altura de los rectángulos?

Altura = FR / longitud del intervalo de clase

De esta forma resultará que la suma de las áreas de los k intervalos de clase es igual a

Observación:

- La suma de las k áreas es igual a 1.
- Si los intervalos de clase son de igual longitud tomar las alturas de los rectángulos como mencionamos recién o trazar la FA o FR nos mostrará la misma imagen visual. Además para la comparación de dos intervalos de clase bastará con comparar sus alturas.
- Si los intervalos de clase son de diferentes longitudes para compararlos se deben ver sus áreas y no sus alturas.

Ejemplo: Mediciones Uranio 238 en suelo

Es de gran importancia para los habitantes de cierta localidad controlar la cantidad de uranio 238 (238U) en suelo de áreas recuperadas de minas de fosfato. Las mediciones de 238U (medidas en pico-curies por gramo) obtenidas en 25 muestras de suelo y ordenadas de menor a mayor fueron:

0,32	**0,37**	**0,54**	**0,70**	**0,74**
0,75	**0,76**	**1,09**	**1,66**	**1,77**
1,90	**1,96**	**2,03**	**2,41**	**2,42**
2,69	**3,36**	**3,59**	**4,06**	**4,55**
5,07	**6,47**	**8,32**	**9,99**	**12,48**

Tabla de distribución de frecuencia

- El número de intervalos de clase tomados aquí es $k = 5$.
- Longitud de los intervalos de clase (IC)???

si queremos una partición disjunta del intervalo [0 ; 15] en k=5 intervalos de igual longitud (L), entonces esta debe ser igual

$$L = (15 - 0) / 5 = 3$$

Tabla de distribución de frecuencia

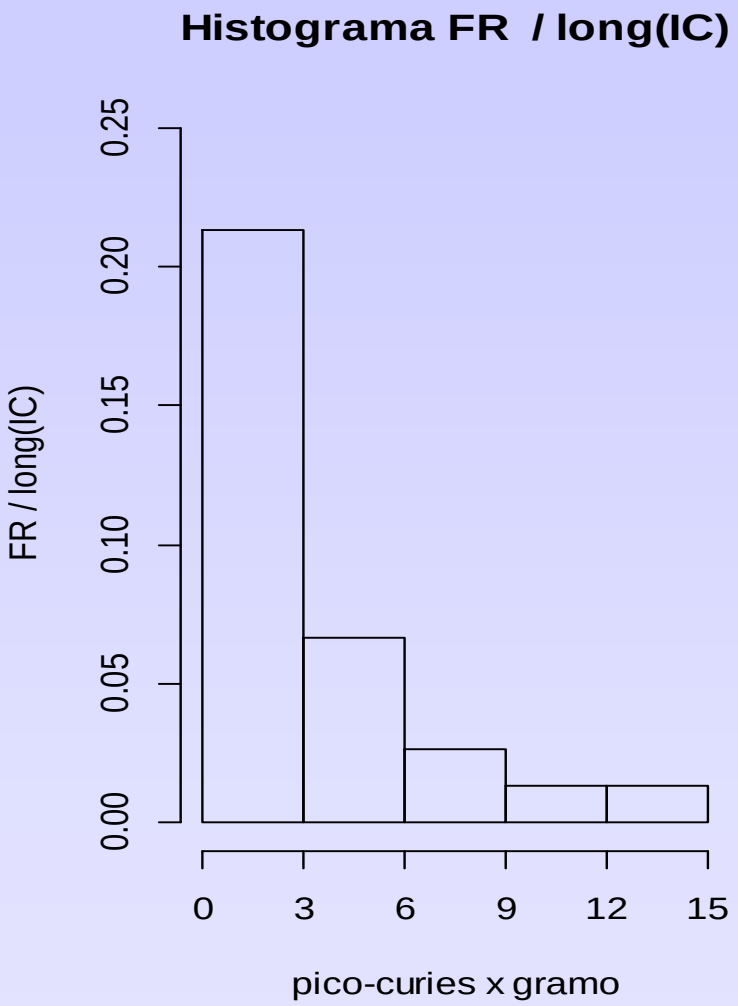
IC	MC	FA	FR	FAA	FRA
[0,00 ; 3,00]	1,50	16	0,64	16	0,64
(3,00 ; 6,00]	4,50	5	0,20	21	0,84
(6,00 ; 9,00]	7,50	2	0,08	23	0,92
(9,00 ; 12,00]	10,50	1	0,04	24	0,96
(12,00 ; 15,00]	13,50	1	0,04	25	1,00

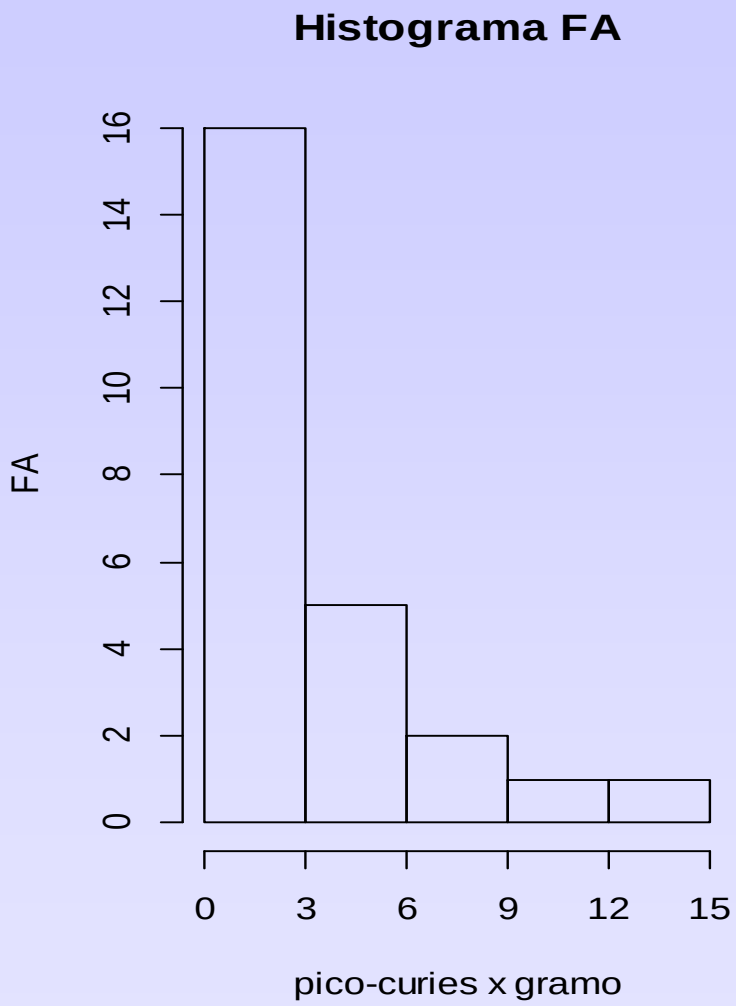
Histogramas para los datos de Uranio 238

Comandos en R para realizar los histogramas

```
par(mfrow= c( 1 , 2 ))
hist(x , breaks = w , freq = FALSE , main= "Histograma FR  / long(IC)" ,
xlab= "pico-curies x gramo", ylim = c( 0 , 0.25 ) ,  axes = FALSE, ylab = "FR /
long(IC)"  )
axis(1,at= seq(0, 15 ,by=3))
axis(2,at=seq(0,0.25,by=0.05))
hist(x , breaks = w , freq = TRUE , main= "Histograma FA" , xlab= "pico-curies x
gramo",  ylim = c( 0 , 16 ) ,  axes = FALSE , ylab = "FA" )
 axis(1,at= seq(0 , 15 , by=3))
 axis(2,at=seq(0,16,by=2))
```

Histogramas





Medidas resúmenes para describir conjuntos de datos

Medidas de posición o tendencia central:

Promedio muestral

Percentiles muestrales (mediana, primer cuartil y tercer cuartil muestrales)

Medidas de dispersión o variabilidad:

Rango muestral

Varianza y desvío estándar muestrales

Rango intercuartil

Coeficiente de variación

Medidas de posición

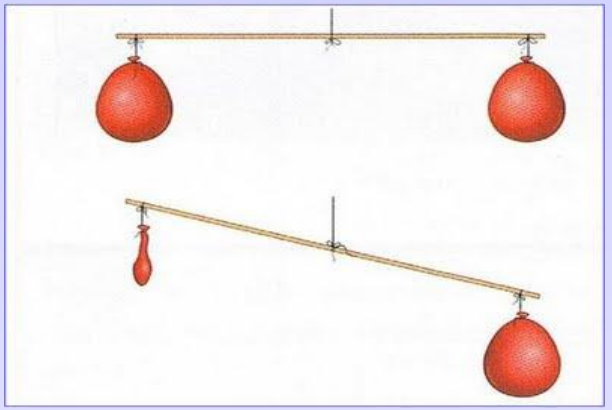
*Media o promedio muestral:

- Media muestral o Promedio $\bar{x} = (x_1 + x_2 + \ldots + x_n) / n$.

-Mejor estimador para la media poblacional (μ).

- Propiedad de centro de masa:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$



- Muy sensible a la presencia de datos extremos.

Ejemplo:

A) 37, 40, 46, 50, 57

B) 37, 40, 46, 57, 200

Percentiles muestrales

- El percentil i es aquel valor que acumula a su izquierda el i% de los datos.
- El percentil 50 también es llamado **mediana muestral**.
- Los percentiles 25 y 75 también son conocidos como **primer y tercer cuartil** y denotaremos con Q_1 y Q_3 respectivamente.

• ¿Cómo calcular la mediana, el Q_1 y Q_3 para un conjunto de datos?

*Mediana muestral:

- $\tilde{x}$ es un valor que deja el 50% de observaciones por debajo y por encima de el.
- Puede o no ser un valor de la muestra.
- Es el valor central o el promedio de los dos valores centrales si n es impar o par respectivamente.

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & \text{si n es impar} \\ [x_{(n/2)} + x_{(n/2)+1}]/2 & \text{si n es par.} \end{cases}$$

¿Cómo calcular Q_1 y Q_3 para un conjunto de datos?

Q_1 se calcula como la mediana de las (n/2) o las a (n+1)/2 observaciones más pequeñas dependiendo que n sea par o impar respectivamente.

Q_3 se calcula como la mediana de las (n/2) o las a (n+1)/2observaciones más grandes dependiendo que n sea par o impar respectivamente.

Calcular la mediana, Q1 y Q3 para los datos del Uranio 238

Medidas de dispersión o variabilidad

Rango muestral

Varianza y Desviación Estandar muestral

Rango intercuartil

Coeficiente de Variación

Rango

Se define como la diferencia entre la máxima y mínima observación, o sea (x (n) – x (1)).

Ventajas

- Fácil de calcular
- Iguales unidades que los datos de origen

Desventajas

- Considera solo dos valores de la muestra
- Ejemplo:

  Muestra 1: 0, 5, 5, 5, 10

  Muestra 2: 0, 4, 5, 6, 10

Varianza muestral → $$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Desventaja: No tiene la misma unidad de medida de los datos.

Desviación estándar muestral → $$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

NOTAR: Ambas utilizan el valor de la media muestral, luego son sensibles a la presencia de datos extremos.

Rango intercuartil → $RIC = Q_3 - Q_1$

Coeficiente de Variación

$$CV = \frac{S}{\bar{x}}\ 100\%$$

Notar:

-Adimensional.

-Permite comparar la variabilidad de características medidas en distintas escalas, luego el que tenga menor CV será el de menor variabilidad.

Ejemplo:

Medidas de alturas de:

Personas

$\bar{x}$ =1.70m

S=0.02m

CV=1,18%

Edificios

$\bar{x}$ =20m

S=0.1m

CV=0,50%

Proposición:

Sea $\{x_i\}_{i=1}^n$ una muestra, a y b números reales con $a \neq 0$. Si $y_i = a\,x_i + b \;;\; i = 1, \ldots, n$ entonces:

$$\bar{y} = a\,\bar{x} + b \qquad \text{y} \qquad S_y^2 = a^2\,S_x^2$$

donde $\bar{y}$ y S_y^2 son el promedio y varianzas muestrales para al muestra y .

Consecuencia: Sea $w_i = \dfrac{x_i - \bar{x}}{S_x} \;;\; i = 1, \ldots, n$

Entonces:

$$\bar{w} = 0 \text{ y } S_w^2 = 1$$

Gráfico de Caja o Box Plot

Tukey (1977) presentó un simple método gráfico que resume varias de las características más destacadas de un conjunto de datos, conocido con el nombre de **Gráfico de caja o Box-plot**.

¿Para qué sirve un gráfico de caja?

Las características incorporadas por este gráfico son:

a) medidas de tendencia central o posición, o sea valor representativo de la muestra,

b) medida de dispersión,

c) naturaleza y magnitud de cualquier desviación de la simetría,

d) identificación de los puntos no usuales o atípicos, o sea puntos marcadamente alejados de la masa principal.

Construcción del Gráfico de caja o *Box plot*

Paso 1: Ordenar los datos de menor a mayor.

Paso 2: Calcular la mediana, el cuartil superior (Q3), el cuartil inferior (Q1) y el RIC.

Paso 3 Sobre un eje horizontal, dibujar una caja cuyo borde izquierdo sea el cuartil inferior y el borde derecho el cuartil superior.

Paso 4: Dentro de la caja trazar un segmento perpendicular cuya posición corresponde al valor de la mediana y marcar con un punto el valor promedio muestral.

Paso 5: Trazar segmentos desde cada extremo de la caja hasta las observaciones más alejadas, que no superen (1,5 RIC) de los bordes correspondientes.

Paso 6: Si existen observaciones que superen (1,5 RIC) respecto de cada uno de los bordes marcar los valores con círculos, y a estos puntos le llamaremos ***puntos anómalos o atípicos.***

¿Cómo obtener descriptivas con R?

```
datos<-read.table("uranio.txt",header=T)
x=datos$Uranio238
sum(x) /25
[1] 3.2
mean(x)
[1] 3.2
sum( (x-mean(x) )^2) /24
[1] 9.915033
var(x)
[1] 9.915033
S = sqrt(var(x))
S
[1] 3.148815

summary(x)
   Min.   1st Qu.  Median    Mean   3rd Qu.    Max.
   0.32      0.76      2.03       3.20      4.06       12.48
```

RIC = 4.06 – 0.76

3.30

f=1.5 * 3.3

f

4.95

(Q1-f ; Q3 +f) =(-4.19 ; 9.01)

¿Hay datos en la muestra que no pertenecen a éste intervalo?

SI hay dos datos atípicos o anómalos en la muestra que no pertenecen al intervalo

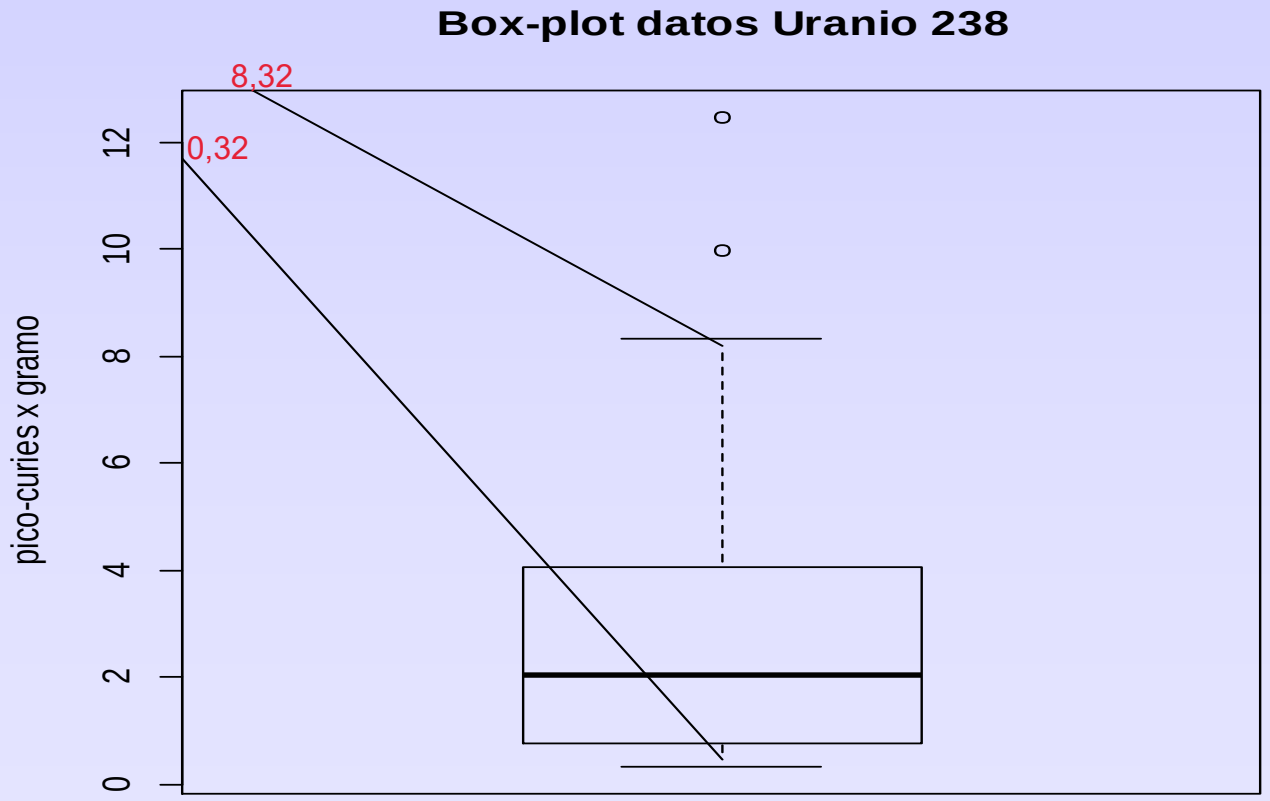
(Q1-f ; Q3 +f) y que ellos son:

9,99 y 12,48.

Hacer el gráfico de caja para estos datos del Uranio 238.

Box plot para los datos del Uranio 238

```
boxplot(x, main= "Box-plot datos Uranio 238" , ylab=
    "pico-curies x gramo")
```



Ejemplo de datos de octanaje en gasolina

Los datos que se presentan a continuación corresponden al octanaje en muestras de gasolina, que han sido tomados de un artículo en la revista Technometrics (vol 19, pag. 425).

Los 79 datos ya están ordenados de menor a mayor.

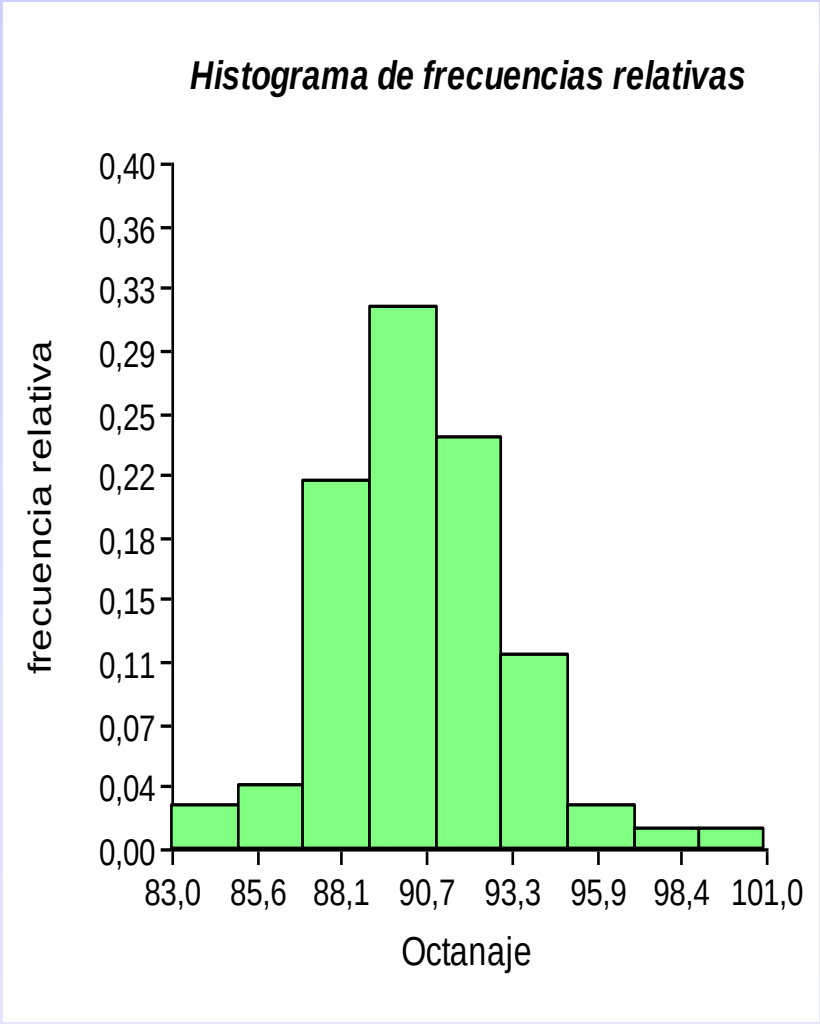
Datos de octanaje de gasolina

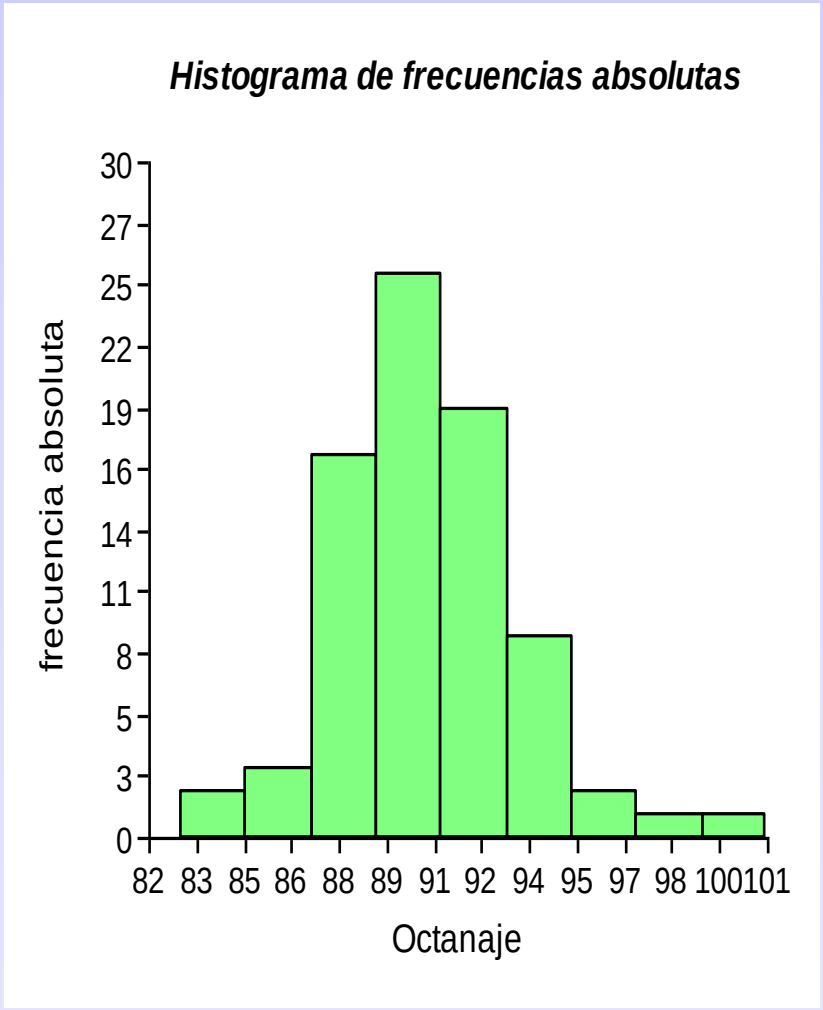
83,4	**87,9**	**88,9**	**89,9**	**90,6**	**91,2**	**92,3**	**93,7**
84,3	88,2	89	90	90,7	91,2	92,6	94,2
85,3	88,3	89,2	90,1	90,8	91,5	92,7	94,2
86,7	88,3	89,3	90,1	90,9	91,6	92,7	94,4
86,7	88,3	89,3	90,3	91	91,6	92,7	94,7
87,4	88,5	89,6	90,3	91	91,8	93	95,6
87,5	88,5	89,7	90,4	91	91,8	93,2	96,1
87,5	88,6	89,8	90,4	91,1	92,2	93,3	98,8
87,6	88,6	89,9	90,4	91,1	92,2	93,3	100,3
87,8	88,7	89,9	**90,5**	91,1	92,2	93,4	

Como el tamaño de la muestra es n=79 elijo k=9 Intervalos de Clase

Intervalo de clase	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Relativa Acumulada
[83,85]	2	0,025	0,025
(85,87]	3	0,038	0,063
(87,89]	16	0,202	0,265
(89,91]	23	0,291	0,556
(91,93]	21	0,266	0,822
(93,95]	10	0,126	0,948
(95,97]	2	0,025	0,973
(97,99]	1	0,013	0,986
(99,101]	1	0,013	0,999

Histogramas de Frecuencias Absolutas y Relativas
Datos del octanaje





Medidas de posición y dispersión para los datos del octanaje

Estadística descriptiva

n	Media	S	Varianza muestral	Mín	Máx
79	90,6696	2,8081	7,8855	83,4000	100,3000

Mediana = 90,5

$Q_1 = 88{,}8$ $Q_3 = 92{,}2$

RIC= 3,4 y (1,5 RIC)= 5,1

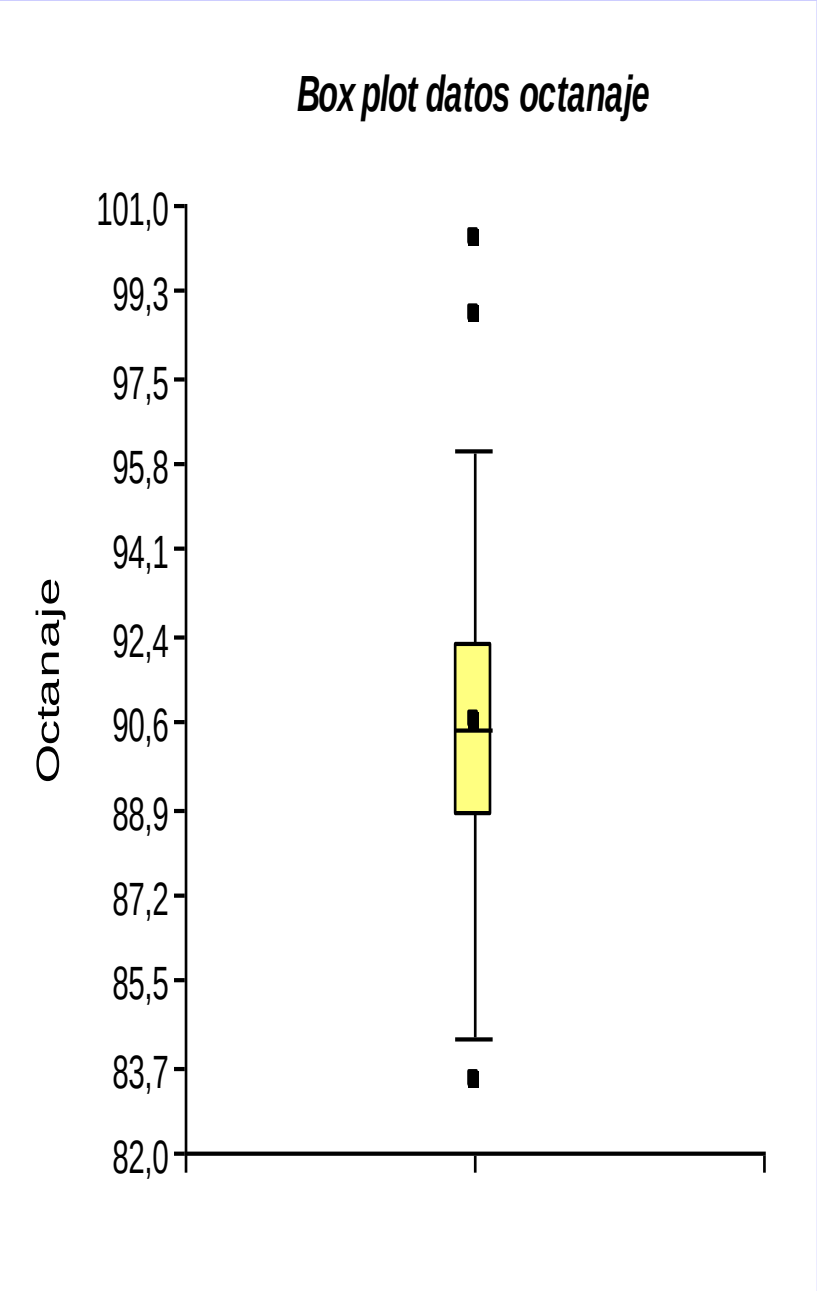
Cálculos para hacer el gráfico de caja

$$Q_1 - 1{,}5\,RIQ = 88{,}8 - 5{,}1 = 83{,}7$$

$$Q_3 + 1{,}5\,RIQ = 92{,}2 + 5{,}1 = 97{,}3$$

Hay tres datos atípicos suaves, uno por debajo (83,4) y dos por arriba

(98,8 y 100,3).



Ejemplo:

Los siguientes datos corresponden a los tiempos de oxidación-inducción (en minutos) para varios aceites comerciales.

87	87	93	99	103	105	119	129	130	132
138	145	145	152	153	160	180	195	211	

Realizar en gráfico de caja para los siguientes datos.

Se puede ver que

$Q_1 = (103+105)/2 = 104$

$Q_3 = (152+153)/2 = 152,5$

$RIC = Q_3 - Q_1 = 48,5$

$Q_1 - 1,5\ RIC = 31,25$

$Q_3 + 1,5\ RIC = 225,25$

NO HAY DATOS EXTREMOS.

Ejemplo:

La siguiente tabla muestra la respuesta observada en dos grupos de conejos, que fueron sometidos a una dosis de una determinada droga, de 1 mg uno de los grupos y el otro a una dosis de 4 mg.

Dosis (mg)	Respuesta
1	8, 12, 9, 14, 6
4	20, 17, 25, 27, 16

¿Hay diferencias en los valores medios entre los dos grupos?

Algunos gráficos posibles para visualizar el interés del problema.