Introducción a la Probabilidad y Estadística - Probabilidad y Estadística

Dra. Grisel Britos

2023

ESTADÍSTICA DESCRIPTIVA

La estadística descriptiva es el área de la estadística que se encarga de organizar, clasificar y resumir la información presente en los datos.

Poblaciones y Muestras

- Población: es la colección de elementos o sujetos de interés. Puede ser finita o infinita.
- Muestra: es un subconjunto de la población, elegido al azar, que será estudiado en detalle.

  Observación: para que la muestra proporcione información sobre la población deberá ser, en algún sentido, representativa de dicha población. Esto quiere decir que la muestra ha sido elegida de forma que todos los elementos de la población tengan la misma probabilidad de pertenecer a la muestra.

Ejemplo: supongamos que estamos interesados en aprender sobre la distribución de edades de los residentes de una ciudad y que obtenemos las edades de las 100 primeras personas que entran en una determinada biblioteca de la ciudad. Si la edad media de esas 100 personas es de 46.2 años, ¿podemos concluir justificadamente que este valor coincide aproximadamente con la edad media de toda la población? Posiblemente no, porque seguro que se podría argüir que la muestra elegida no es en este caso representativa de la población total, ya que generalmente son los estudiantes jóvenes y los ciudadanos mayores quienes frecuentan la citada biblioteca, en mayor medida que las personas que están en edad laboral.

Tipos de datos

Los datos que se presenten pueden clasificarse como **datos numéricos** o **datos categóricos**.

Los datos numéricos pueden clasificarse a su vez como **datos discretos** o **datos continuos**. Veamos algunos ejemplos:.
Ej. de datos discretos: número de hijos en una familia, número de materias aprobadas, número de accidentes de tránsito en una ciudad en determinado mes, etc.
Ej. de datos continuos: temperatura corporal, tiempo de exposición a determinado patógeno, ph del agua, etc.

Por otro lado, los datos categóricos, es decir, aquellos que expresan un atributo que puede pertencer a diferentes categorías, se clasifican como **datos ordinales** o **datos nominales**. Algunos ejemplos son:
Ej. de datos ordinales: estado de una enfermedad (severo, moderado, leve), grupo etario, nivel educativo alcanzado, etc.
Ej. de datos nominales: grupo sanguíneo (O, A, B, AB), estado civil, nacionalidad, etc.

La estadística descriptiva...

- Provee de métodos que permiten organizar y resumir la información de datos.
- De acuerdo al conjunto de datos se seleccionará el método más adecuado.
- Cuando los datos son numéricos continuos algunas opciones son:
  - Tablas de distribución de frecuencias (*)
  - Histogramas
  - Medidas de posición
  - Medidas de dispersión
  - Gráficos

(*) Se puede realizar para cualquier tipo de datos (numéricos o categóricos)

A continuación veamos en detalle estos métodos para organizar y resumir la información de los datos.

Tablas de distribución de frecuencias

Para construir una tabla de frecuencias debemos seguir estos pasos:

1) Tomar un intervalo que contenga al conjunto de datos.
2) Dividir el intervalo en k intervalos de clase (IC) tal que sean adyacentes y disjuntos.
3) Contar el número de observaciones en cada intervalo. Esto se llama Frecuencia Absoluta (FA).
4) Calcular la Frecuencia Relativa (FR) como la FA dividida por n para cada intervalo. (n es la cantidad de datos de la muestra).

Observaciones:

- ¿Cómo elegir k? No hay reglas. Entre 5 a 20 intervalos es lo usual. Se puede tomar $k \sim \sqrt{n}$ donde n es el tamaño de la muestra.
- Los intervalos no tienen por que tener igual longitud.
- Se tiene que cumplir:

$$\sum_{i=1}^{k} FA_i = n \quad \text{y} \quad \sum_{i=1}^{k} FR_i = 1.$$

Histogramas

Es el 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?

$$\text{Altura} = \frac{FR}{\text{longitud del intervalo de clase}}$$

De esta forma resultará que la suma de las áreas de los k intervalos de clase es igual a 1.

Observaciones:

- La suma de las k áreas es igual a 1.
- Si los intervalos de clase son de igual longitud, al tomar las alturas de los rectángulos como mencionamos recién (FR/long(IC)) o tomar como alturas las FA o las FR, el histograma nos mostrará la misma imagen visual pues estas alturas son proporcionales. 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 debemos observar las áreas de los rectángulos y no sus alturas.

*Ejemplo:*Entradas vendidas para un curso de RCP

Una médica influencer en redes sociales vende entradas para hacer un curso online de RCP. Todos los días sube historias a su cuenta; algunos días más y otros días menos pero siempre recuerda a sus seguidores de la disponibilidad del curso en cuestión. Para conocer el impacto de esas historias en las ventas decidió registrar en una tabla las entradas vendidas durante 25 días. En la siguiente tabla podemos observar los resultados ordenados de menor a mayor:

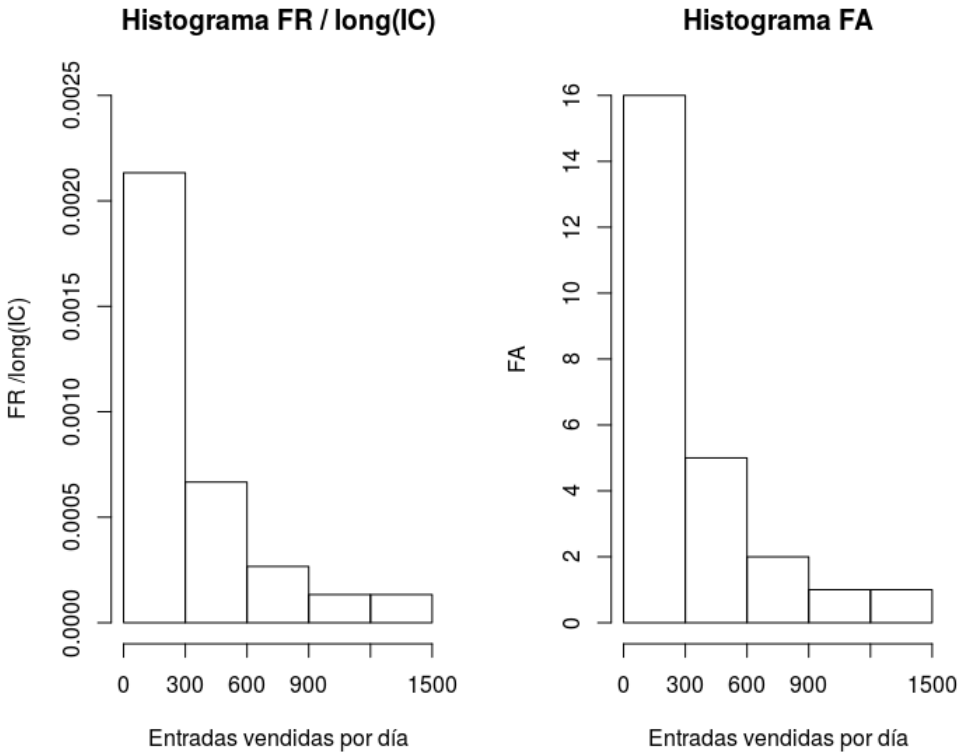
32	37	57	70	74
75	76	109	166	177
190	193	203	241	242
269	336	359	406	455
507	647	832	999	1248

Construyamos la tabla de distribución de frecuencia y el histograma para este caso:

- El número de intervalos de clase tomados aquí es $k = 5$.
- Si queremos una partición disjunta del intervalo [0 ; 1500] en $k = 5$ intervalos de igual longitud (L), entonces esta debe ser igual $L = (1500–0)/5 = 300$.

IC	FA	FR	FAA	FRA
[0, 300]	16	16/25	16	1600/2500=0.64
(300, 600]	5	5/25	21	21/25= 0.84
(600, 900]	2	2/25	23	23/25=0.92
(900, 1200]	1	1/25	24	24/25=0.96
(1200, 1500]	1	1/25	25	1

FAA: “Frecuencias Absolutas Acumuladas”.
FRA: “Frecuencias Relativas Acumuladas”.



Medidas resúmenes para describir conjuntos de datos

A continuación estudiaremos medidas que describen la tendencia central del conjunto de datos; es decir, que describen el centro del conjunto de valores de datos. Dentro de estas medidas veremos a la media muestral y los percentiles muestrales (mediana, primer cuartil, tercer cuartil). También estudiaremos medidas que describen la variabilidad de los datos respecto del centro. Esto responde a la pregunta de si la mayor parte de los valores están próximos al centro, o, por el contrario, varían mucho alrededor de éste. Dentro de estas medidas, analizaremos el rango muestral, la varianza y desvío muestral, el rango intercuartil y el coeficiente de variación.

Medidas de posición o tendencia central

* Media muestral o promedio muestral

Dados los datos $x_1, x_2, ..., x_n$ la media muestral, denotada por $\overline{x}$ o $\overline{x}_n$, se define como

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

La media muestral es la medida más utilizada para aproximar la media poblacional (μ), sin embargo, es muy sensible a la presencia de datos extremos en la muestra.

Ejemplo. Consideremos los conjuntos de datos:
A) Datos: 37, 40, 46, 50, 57; $\overline{x}_A = 46$
B) Datos: 37, 40, 46, 57, 200; $\overline{x}_B = 76$
En este ejemplo, la media muestral del conjunto de datos B se ve fuertemente afectada por el dato extremo "200".

* Percentiles muestrales

El percentil i es aquel valor que acumula a su izquierda el i % de los datos.

- **Mediana muestral**

  La mediana muestral, también llamada percentil 50, se denota con $\tilde{x}$ y es el valor que deja el 50 % de las observaciones por debajo y por encima de él. Puede o no ser un valor de la muestra. Dados los datos $x_1, x_2, ..., x_n$, la mediana muestral se define como:

$$\tilde{x} = \begin{cases} \dfrac{x_{(\frac{n}{2})} + x_{(\frac{n}{2})+1}}{2}, & \text{si } \ \text{es par} \\ x_{(\frac{n+1}{2})}, & \text{si } \ n \text{ es impar} \end{cases}$$

  Cabe aclarar que $x_{(1)}, x_{(2)}, ..., x_{(n)}$ denota la muestra ordenada de menor a mayor. Es decir, $x_{(1)}$ es el menor valor de la muestra, $x_{(2)}$ es el segundo menor valor de la muestra y $x_{(n)}$ es el mayor valor de la muestra.

- **Primer cuartil**

  El primer cuartil o cuartil inferior, denotado Q_1, se calcula como la mediana de las $(n/2)$ o las $(n+1)/2$ observaciones más pequeñas dependiendo de que n sea par o impar, respectivamente.

$$Q_1 = \begin{cases} mediana\{x_{(i)} : 1 \leq i \leq \dfrac{n}{2}\}, & \text{si } \ n \text{ es par} \\ mediana\{x_{(i)} : 1 \leq i \leq \dfrac{n+1}{2}\}, & \text{si } \ \text{es impar} \end{cases}$$

- **Tercer cuartil**

  El tercer cuartil o cuartil superior, denotado Q_3, se calcula como la mediana de las $(n/2)$ o las $(n+1)/2$ observaciones más grandes dependiendo de que n sea par o impar respectivamente.

$$Q_3 = \begin{cases} mediana\{x_{(i)} : \dfrac{n}{2} + 1 \leq i \leq n\}, & \text{si } \ n \text{ es par} \\ mediana\{x_{(i)} : \dfrac{n+1}{2} \leq i \leq n\}, & \text{si } \ \text{es impar} \end{cases}$$

Nótese que si n es impar, la mediana $\tilde{x}$ está incluida en ambas mitades.

En el ejemplo de las entradas vendidas para un curso de RCP:

$$\tilde{x} = x_{(13)} = 203 \quad , \quad Q_1 = x_{(7)} = 76 \quad , \quad Q_3 = x_{(19)} = 406.$$

Medidas de dispersión o variabilidad

* Rango muestral

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. Consideremos los conjuntos de datos:
A) Datos: 0, 5, 5, 5, 10; Rango(A)=10
B) Datos: 0, 4, 5, 6, 10; Rango(B)=10
En este ejemplo, a pesar de que las muestras son muy distintas, el rango muestral es el mismo.

* Varianza muestral

Se define como:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Tiene la desventaja de que no tiene la misma unidad de medida que los datos.

* Varianza muestral corregida

Se define como:

$$s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

* Desvío estándar muestral

Se define como:

$$s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

* Desvío estándar muestral corregido

Se define como:

$$s_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Nótese que las varianzas y los desvíos muestrales utilizan el valor de la media muestral, esto hace que sean medidas sensibles a la presencia de datos extremos.

* **Rango intercuartil**

Se define como:

$$RIC = Q_3 - Q_1$$

* **Coeficiente de variación**

Se define como:

$$CV = \frac{s_n}{\overline{x}} 100\,\%.$$

Esta medida es adimensional y permite comparar la variabilidad de características medidas en distintas escalas, luego el que tenga menor CV será el de menor variabilidad.

Proposición: Sea $x_1, x_2, ..., x_n$ una muestra, a y b números reales con $a \neq 0$. Si $y_i = ax_i + b$, $i = 1, ..., n$ entonces

$$\overline{y} = a.\overline{x} + b \quad \text{y} \quad s_y^2 = a^2.s_x^2$$

donde $\overline{y}$ y s_y^2 son el promedio y la varianza muestral para la muestra y, respectivamente.

Consecuencia: Si $w_i = \frac{x_i - \overline{x}}{s_x}$ con $i = 1, ..., n$, entonces $\overline{w} = 0$ y $s_w^2 = 1$.

Gráfico de caja o Boxplot

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 Boxplot.

¿Para qué sirve un gráfico de caja? Las características incorporadas por este gráfico son:

a) Medidas de tendencia central o posición

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 Boxplot:

Paso 1. Ordenar los datos de menor a mayor.

Paso 2. Calcular la mediana, el cuartil superior (Q_3), el cuartil inferior (Q_1) 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.

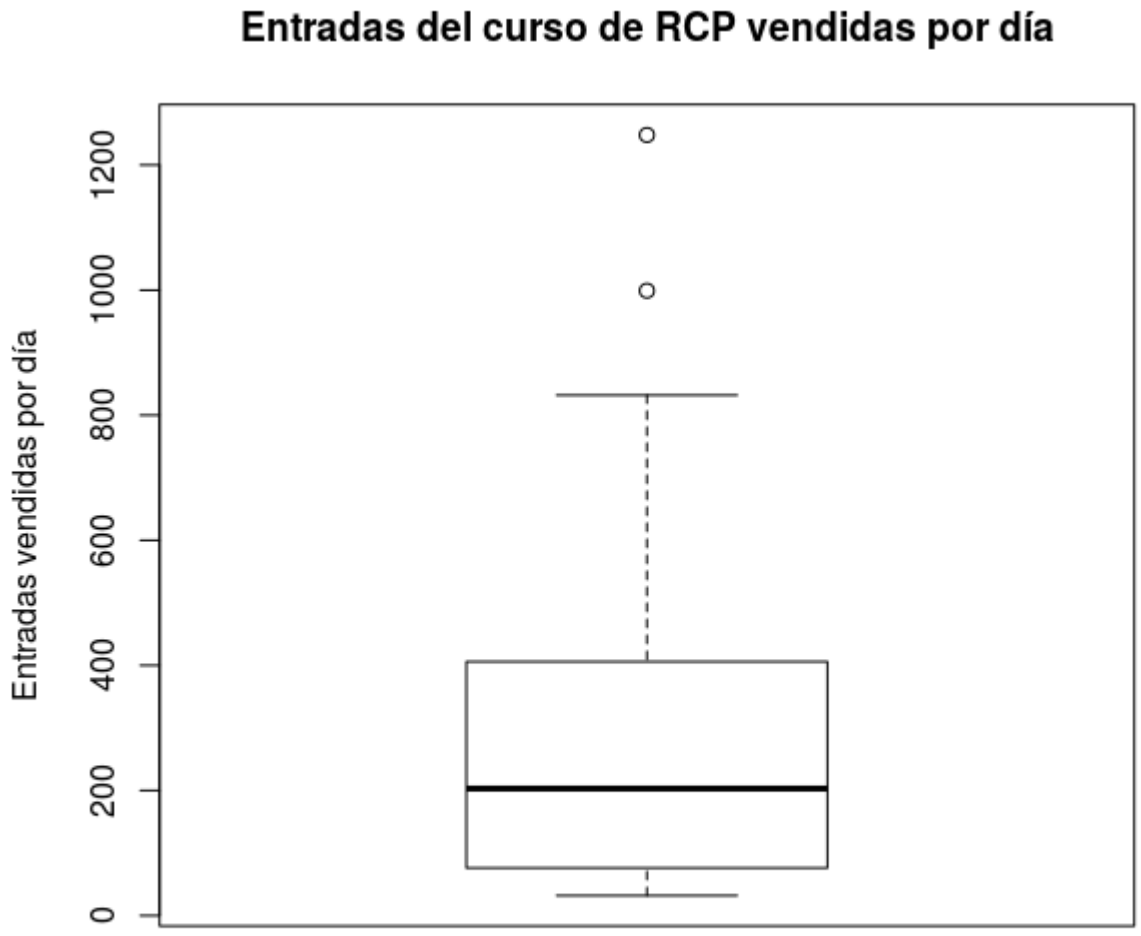
¿Cómo obtener descriptivas con R? (Ejemplo de las entradas al curso de RCP)

```
>sum(x) /25
[1] 320
>mean(x)
[1] 320
>S2=sum((x-mean(x))^2)/25
>S2
[1] 95184.32
>S = sqrt(S2)
>S
[1] 308.52
>summary(x)
Min.   1st Qu.   Median   Mean   3rd Qu.   Max.
32     76        203      320     406      1248
>Q1=summary(x)[[2]]
>Q3=summary(x)[[5]]
>RIC = 406-76
>RIC
[1] 330
>f=1.5*330
>f
[1] 495
>c(Q1-f, Q3+f)
[1] (419, 901)
```

¿Hay datos en la muestra que no pertenecen a éste intervalo? SÍ. Hay dos datos atípicos o anómalos en la muestra que no pertenecen al intervalo (Q1-f ; Q3 +f) y ellos son: 999 y 1248.

Con la orden siguiente se puede hacer el boxplot de estos datos:

```
boxplot(x, main= Entradas del curso de RCP vendidas por día , ylab= Entradas vendidas por día)
```


Entradas del curso de RCP vendidas por día
Entradas vendidas por día
0
200
400
600
800
1000
1200

Ejemplo. Datos de octanaje en gasolina

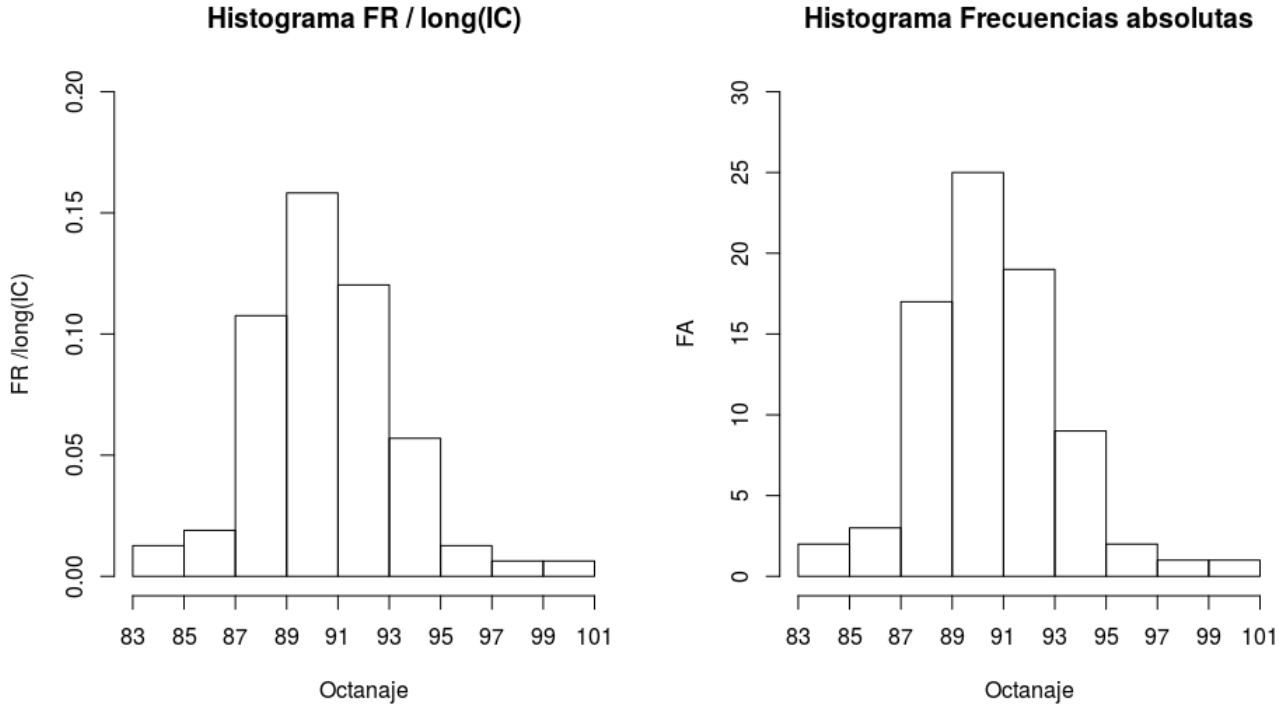
Los datos que se presentan a continuación corresponden al octanaje en muestras de gasolina, que han sido tomados de un artículo en la revista Technometrics (vol 19, pag. 425). Los 79 datos ya están ordenados de menor a mayor.

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	89.7	90.4	91	91.8	93.2	96.1
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 $n = 79$ se puede elegir $k = 9$ 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]	17	0.215	0.278
(89,91]	25	0.316	0.594
(91,93]	18	0.228	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 de los datos de octanaje



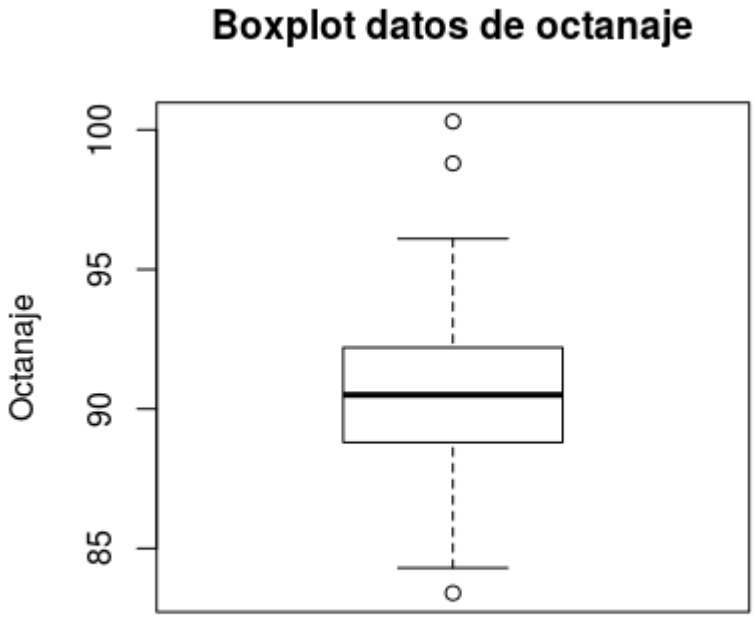
Las medidas de posición y dispersión para los datos del octanaje son:

$$n = 79 \qquad \overline{x} = 90{,}67 \qquad s_n = 2{,}79 \qquad s_n^2 = 7{,}78$$
$$x_{(1)} = 83{,}4 \qquad x_{(n)} = 100{,}3 \qquad \tilde{x} = 90{,}5 \qquad Q_1 = 88{,}8$$
$$Q_3 = 92{,}2 \qquad RIC = 3{,}4 \qquad 1{,}5 * RIC = 5{,}1$$

Los cálculos para hacer el gráfico de caja son:

$$Q_1 - 1{,}5 * RIC = 83{,}7$$
$$Q_3 + 1{,}5 * RIC = 97{,}3$$

El bigote inferior es 84.3 y el bigote superior es 96.1. Hay tres datos atípicos suaves, uno por debajo (83,4) y dos por arriba (98,8 y 100,3).


Boxplot datos de octanaje
Octanaje
85
90
95
100

Ejemplo. La siguiente tabla muestra la efectividad observada (en una escala del 1 al 10) en un grupo de mujeres y en un grupo de hombres de una nueva droga contra la migraña.

Grupo	Respuestas						
Mujeres	5	7	3	5	4	4	7
Hombres	5	2	9	9	3	1	3

¿Hay diferencias en los valores medios entre los dos grupos?
Medidas obtenidas:

$$\overline{x}_M = 5 \qquad \overline{x}_H = 4{,}57 \qquad \tilde{x}_M = 5 \qquad \tilde{x}_H = 3$$
$$s^2_M = 2 \qquad s^2_H = 9{,}1$$

Tanto el grupo de las mujeres como el de los hombres tienen valores medios de efectividad similares. Sin embargo, el grupo de los hombres presentó mayor dispersión, es decir que el efecto de la droga en ellos fue variada. No se observaron valores atípicos en las observaciones.

Boxplots ejemplos Nuevas Drogas: