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.

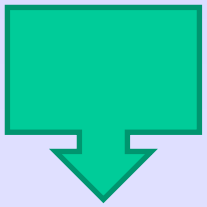
MUESTRA:

- subconjunto elegido al azar de la población.
- tamaño muestral n.

MUESTRA

Estimar características

Inferir acerca de hipótesis



POBLACIÓN

Tipos de datos

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.
- ¿cómo hacerlo?

Realizando: Tablas de Distribución de frecuencias.
Medidas de posición.
Medidas de dispersión.
Gráficos.

Tabla de distribución de frecuencia

- Tomar un intervalo que contenga al conjunto de datos.
- Dividir el intervalo en k intervalos de clase tal que sean adyacentes y disjuntos.
- Contar el número de observaciones en cada intervalo (FA).
- Calcular las FR como la FA dividida n en 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 sea igual a 1.
- Luego 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 comparar dos se deben ver sus áreas y no sus alturas.

Ejemplo

Para decidir el número de cajeras necesaria para una cadena de supermercado, se requiere tener información sobre el tiempo (en minutos) requerido para atender a los clientes. Para tal fin, se tomó una muestra aleatoria de n=60 clientes y se midió el tiempo que se demora en atenderlos.

Los datos previamente ordenados de menor a mayor fueron:

0.20	**0.20**	**0.30**	**0.30**	**0.30**	**0.40**	**0.40**	**0.40**	**0.50**	**0.50**
0.60	**0.60**	**0.60**	**0.60**	**0.70**	**0.70**	**0.70**	**0.80**	**0.80**	**0.80**
0.80	**0.90**	**0.90**	**1.00**	**1.00**	**1.10**	**1.10**	**1.10**	**1.10**	**1.10**
1.10	**1.10**	**1.20**	**1.20**	**1.20**	**1.30**	**1.30**	**1.30**	**1.40**	**1.40**
1.60	**1.60**	**1.70**	**1.70**	**1.80**	**1.80**	**1.80**	**1.80**	**1.90**	**1.90**
2.10	**2.20**	**2.30**	**2.50**	**2.80**	**3.10**	**3.10**	**3.60**	**4.50**	**5.20**

Tabla de distribución de frecuencia

- Elección del número de intervalos de clase

$k \cong \sqrt{60} \cong 7.75$ entonces tomar k = 8.

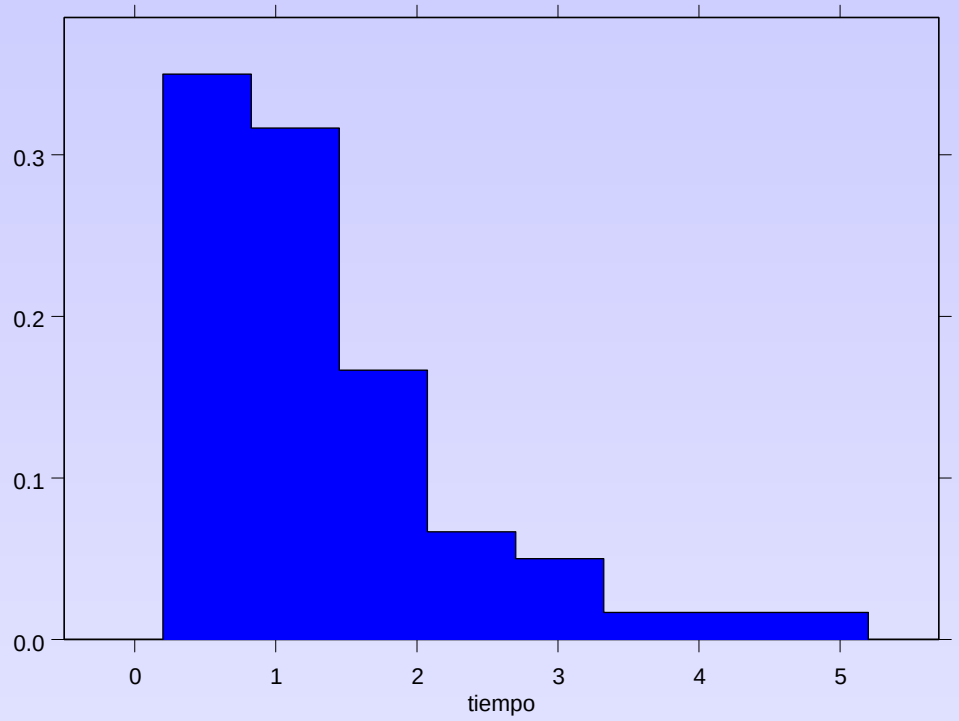
- Longitud de los intervalos de clase (IC)??? si queremos una partición disjunta del intervalo [0.2 , 5.2] en k=8 intervalos de igual longitud (l), entonces esta debe ser igual

$$l = (5.2 - 0.2) / 8 = 0.625$$

Tabla de distribución de frecuencia

IC	FA	FR
[0.2, 0.825]	21	21/60 ≅ 0.35
(0.825, 1.45]	19	19/60 ≅ 0.32
(1.45, 2.075]	10	10/60 ≅ 0.17
(2.075, 2.7]	4	4/60 ≅ 0.07
(2.7, 3.325]	3	3/60 ≅ 0.05
(3.325, 3.95]	1	1/60 ≅ 0.02
(3.95, 4.575]	1	1/60 ≅ 0.02
(4.575, 5.2]	1	1/60 ≅ 0.02
	n=60	1

Histograma de frecuencias relativas



Ejemplo

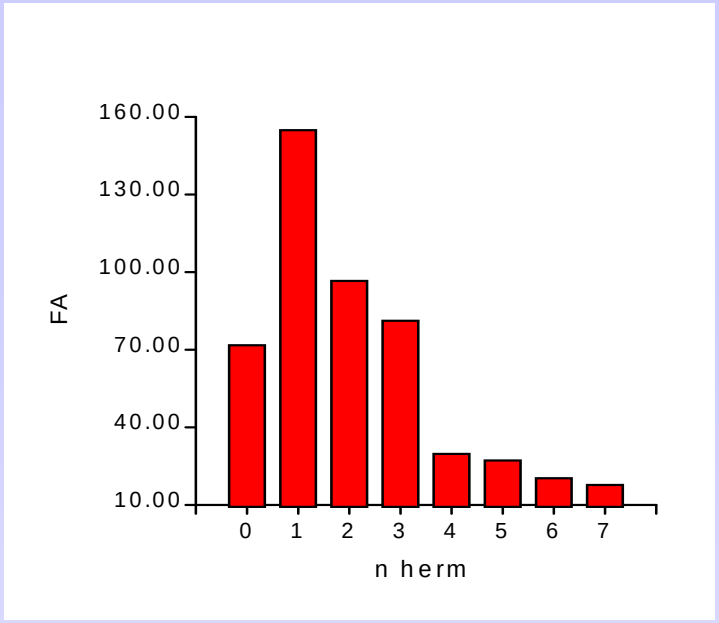
Distribución del número de hermanos (excluido él mismo) de una muestra de 500 alumnos varones de una Universidad

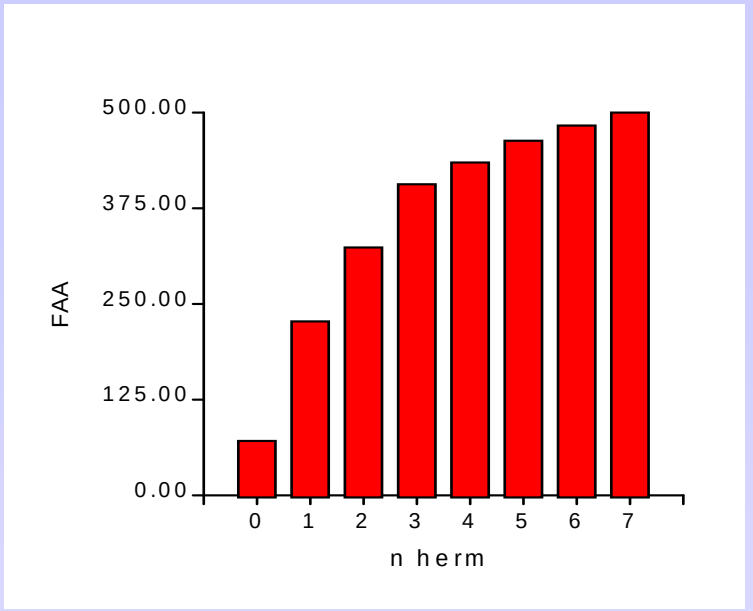
Número de hermanos	FA	FAA	FR	FRA	Porcentaje
0	72	72	0.144	0.144	14.4
1	155	227	0.310	0.454	31.0
2	97	324	0.194	0.648	19.4
3	81	405	0.162	0.810	16.2
4	30	435	0.060	0.870	6.0
5	27	462	0.054	0.924	5.4
6	20	482	0.040	0.964	4.0
Más de 6	18	500	0.036	1.000	3.6
total	500	500	1.000	1.000	100

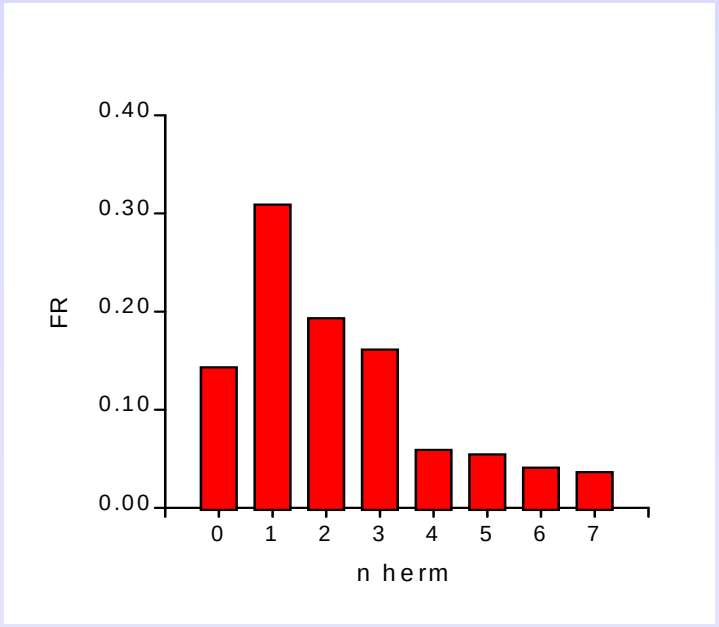
Distribución del peso (x) en Kg de una muestra de 500 alumnos varones de una Universidad

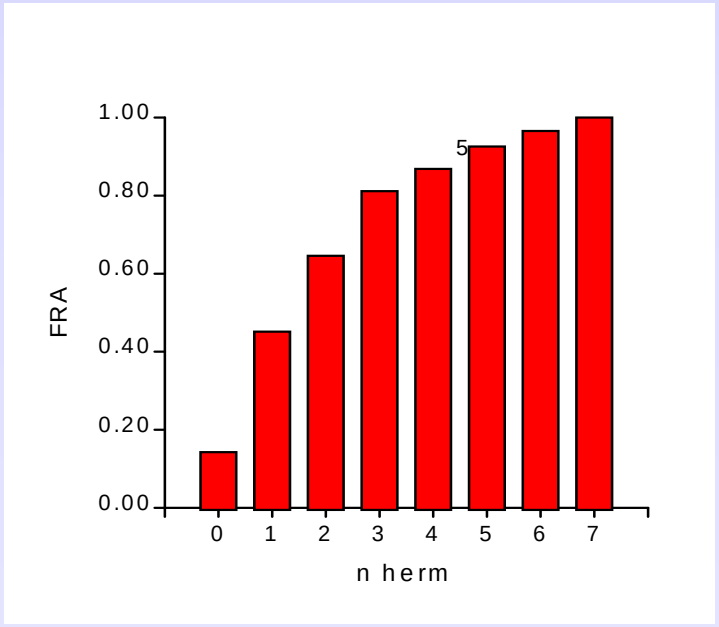
Intervalo de clase	FA	FAA	FR	FRA	Porcentaje	Marca de clase
$40 < x \leq 45$	1	1	0.002	0.002	0.2	42.5
$45 < x \leq 50$	3	4	0.006	0.008	0.6	47.5
$50 < x \leq 55$	12	16	0.024	0.032	2.4	52.5
$55 < x \leq 60$	75	91	0.150	0.182	15.0	57.5
$60 < x \leq 65$	103	194	0.206	0.388	20.6	62.5
$65 < x \leq 70$	155	349	0.310	0.698	31.0	67.5
$70 < x \leq 75$	101	450	0.202	0.900	20.2	72.5
$75 < x \leq 80$	29	479	0.058	0.958	5.8	77.5
$80 < x \leq 85$	11	490	0.022	0.980	2.2	82.5
$85 < x \leq 90$	8	498	0.016	0.996	1.6	87.5
$90 < x \leq 95$	2	500	0.004	1.000	0.4	92.5
total	500	500	1.000	1.000	100.0	-

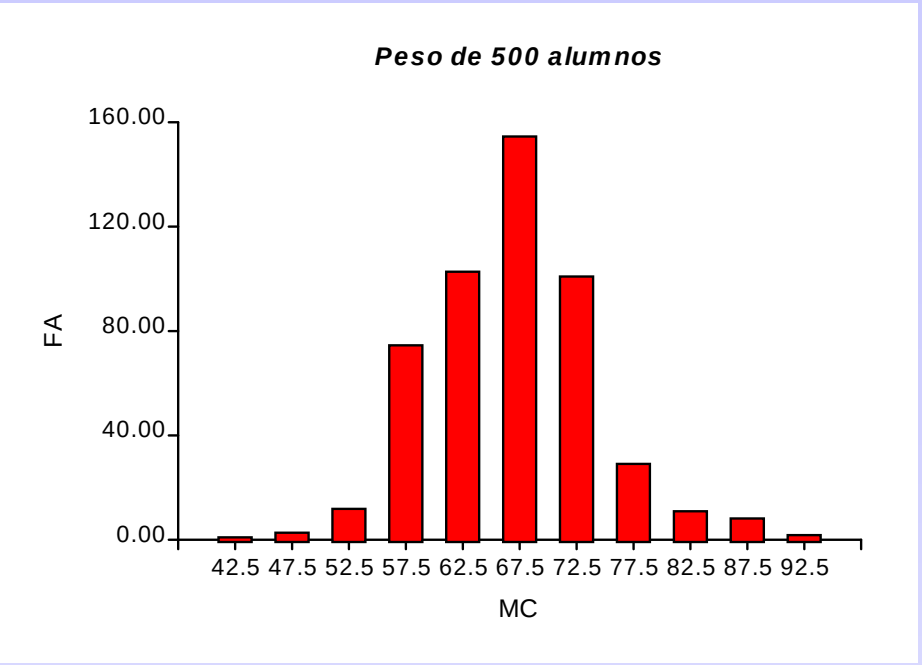
FAA= Frecuencias absolutas acumuladas
FRA= Frecuencias relativas acumuladas

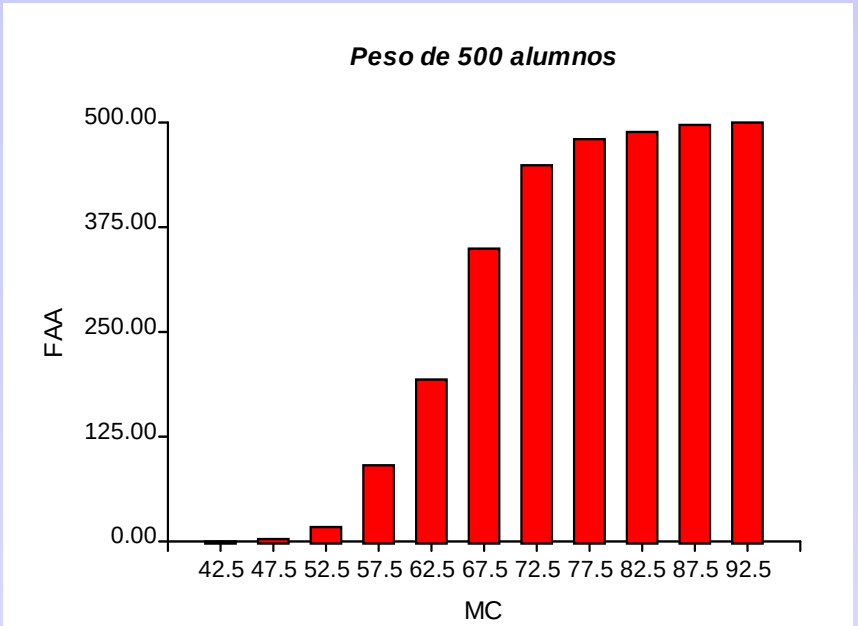

160.00
130.00
100.00
70.00
40.00
10.00
FA
0 1 2 3 4 5 6 7
n herm

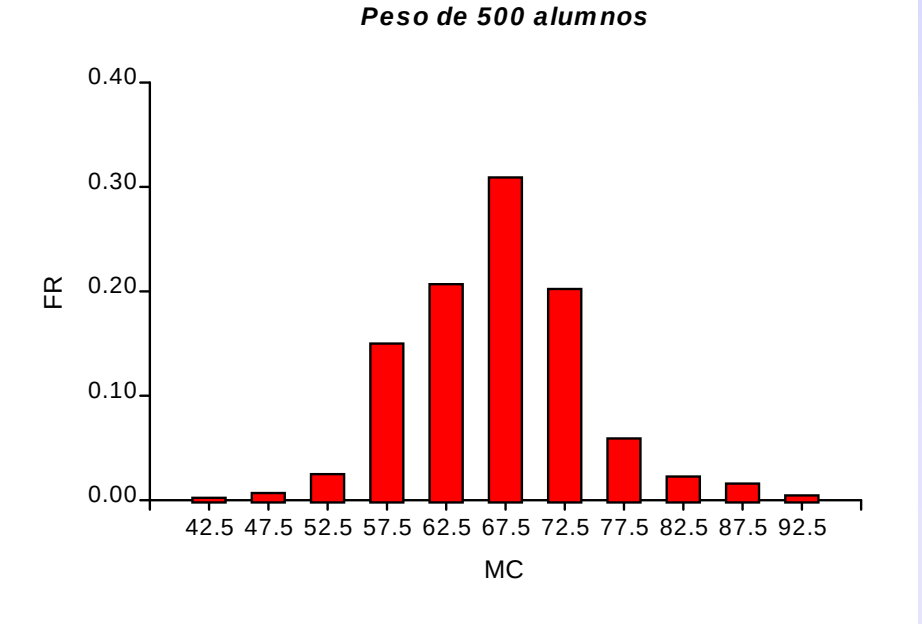

500.00
375.00
250.00
125.00
0.00
FAA
0 1 2 3 4 5 6 7
n herm

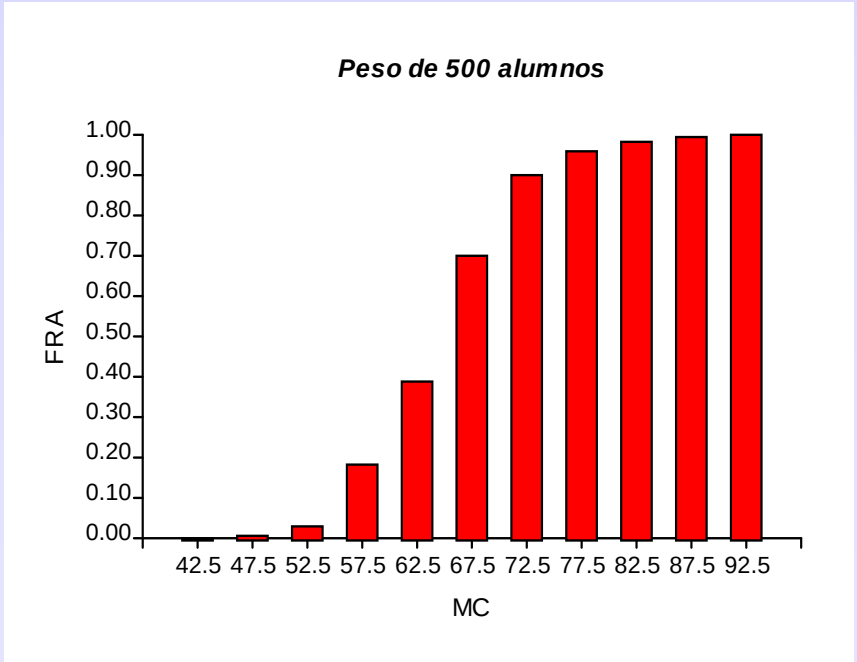

0.40
0.30
0.20
0.10
0.00
FR
0 1 2 3 4 5 6 7
n herm


1.00
0.80
0.60
0.40
0.20
0.00
FRA
5
0 1 2 3 4 5 6 7
n herm


Peso de 500 alumnos
160.00
120.00
80.00
40.00
0.00
FA
42.5 47.5 52.5 57.5 62.5 67.5 72.5 77.5 82.5 87.5 92.5
MC


Peso de 500 alumnos
500.00
375.00
250.00
125.00
0.00
FAA
42.5 47.5 52.5 57.5 62.5 67.5 72.5 77.5 82.5 87.5 92.5
MC


Peso de 500 alumnos
0.40
0.30
0.20
0.10
0.00
FR
42.5 47.5 52.5 57.5 62.5 67.5 72.5 77.5 82.5 87.5 92.5
MC


Peso de 500 alumnos
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
FRA
42.5 47.5 52.5 57.5 62.5 67.5 72.5 77.5 82.5 87.5 92.5
MC

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).
- Muy sensible a la presencia de datos extremos.
- Propiedad de centro de masa: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

Ejemplo:

A)37, 40, 46, 50, 57

B) 37, 40, 46, 57, 200

Percentiles muestrales

- El percentil i% (p_i) es aquel valor que acumula a su izquierda el i% de los datos.
- Al percentil 50% es también llamado como **mediana muestral**.
- Otros percentiles de interés son el 25% y 75%, también llamados **primer y tercer cuantil**, que 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.

Medidas de dispersión o variabilidad

Rango

Varianza y Desviación Estandar

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

La muestra 2 es más variable que la 1!

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 → $f = Q_3 - Q_1$

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%

Algunos tipos de gráficos

1) Gráfico de barras o histograma.

2) Gráfico de caja.

3)Diagrama de dispersión.

4) Gráfico de densidad de puntos.

5) Q Q plot.

¿Cómo construir un gráfico de caja?

Los histogramas nos dan una información cualitativa sobre el comportamiento de los datos, dado que presentan de forma resumida cómo se distribuyen los datos respecto de la media y permiten visualizar cuál es el valor (o valores) más frecuentes. En 1977, Tukey presentó un simple método gráfico-cuantitativo que resume varias de las características más destacadas de un conjunto de datos. Tal método se conoce con el nombre de **diagrama de caja o box-plot**.

Las características de los datos incorporadas por este diagrama son:

a) centro o posición del valor más representativo,

b) 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 de datos.

Construcción del *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 RIQ.

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 RIQ) de los bordes correspondientes.

Paso 6: Si existen observaciones que superen (1,5 RIQ) entonces marcarlos con circunferencias aquellos puntos comprendidos entre (1,5 RIQ) y (3 RIQ) respecto del borde más cercano, estos puntos se llaman ***puntos anómalos suaves***, y con asteriscos aquellos puntos que superen los (3 RIQ) respecto de los bordes más cercanos, estos puntos se llaman ***puntos anómalos extremos***.

Ejemplo

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.

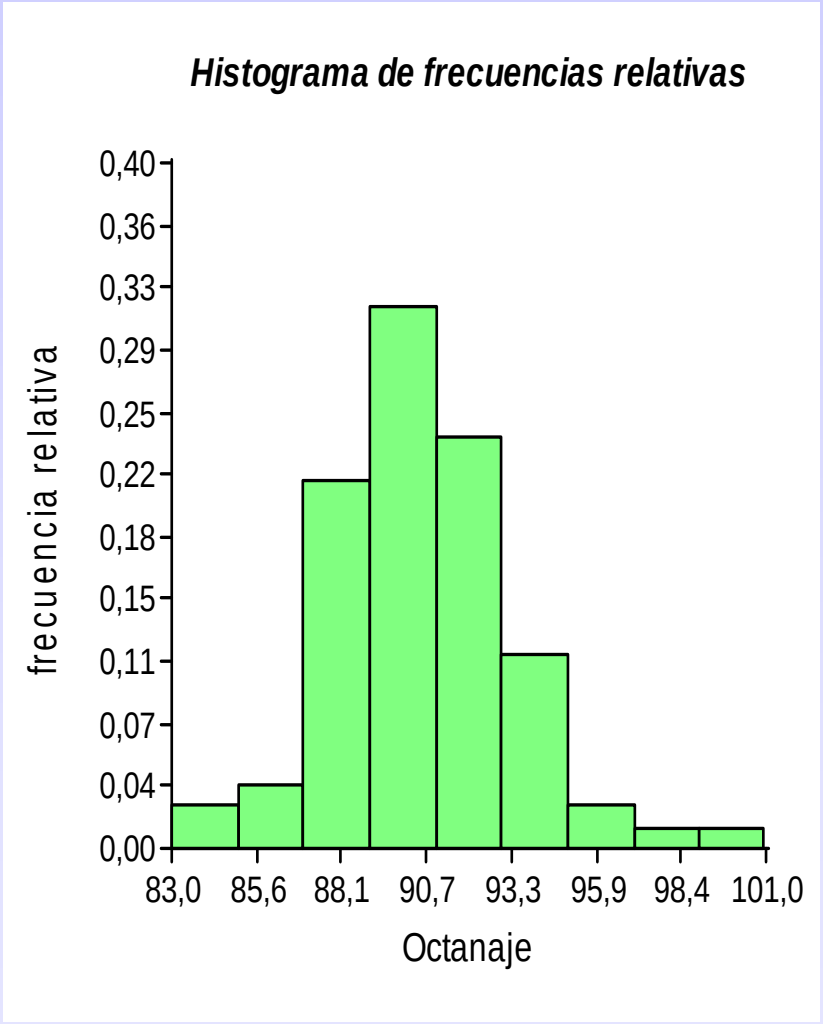
Datos del octanaje

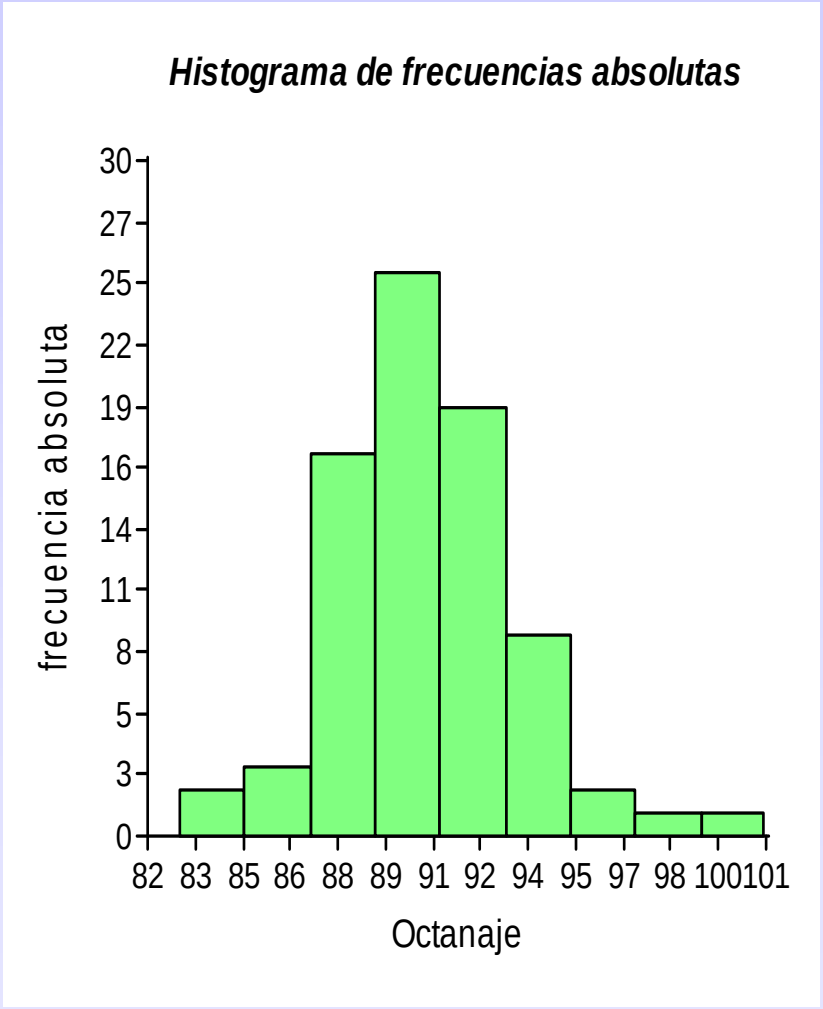
83,4	**87,9**	**88,9**	**89,9**	**90,6**	**91,2**	**92,3**	**93,7**
84,3	88,2	89	90	90,7	91,2	92,6	94,2
85,3	88,3	89,2	90,1	90,8	91,5	92,7	94,2
86,7	88,3	89,3	90,1	90,9	91,6	92,7	94,4
86,7	88,3	89,3	90,3	91	91,6	92,7	94,7
87,4	88,5	89,6	90,3	91	91,8	93	95,6
87,5	88,5	89,7	90,4	91	91,8	93,2	96,1
87,5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

Variable	n	Media	D.E.	Var(n-1)	Mín	Máx
Octanaje	79	90,6696	2,8081	7,8855	83,4000	100,3000

Mediana = 90,5

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

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

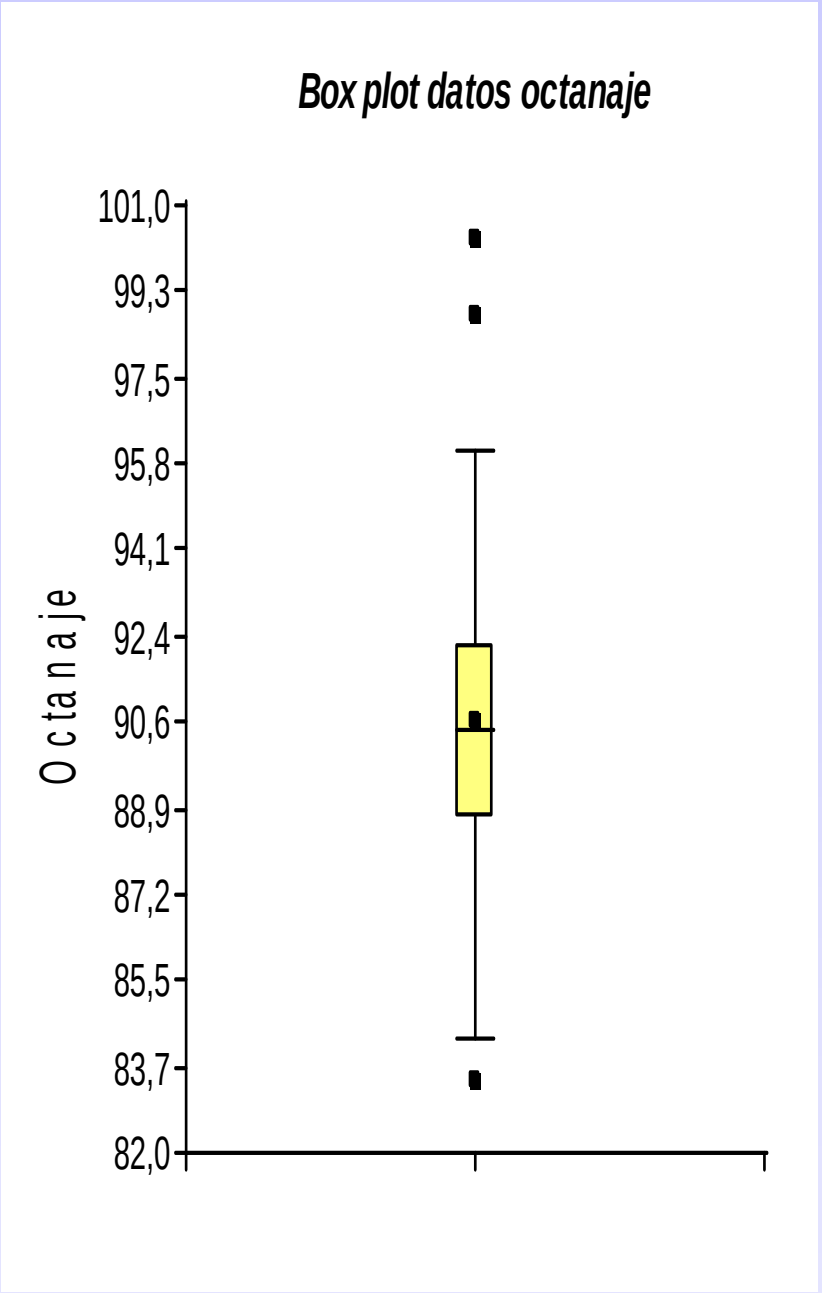
Cálculos para hacer el gráfico de caja

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

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

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

(98,8 y 100,3).



Ejemplo :

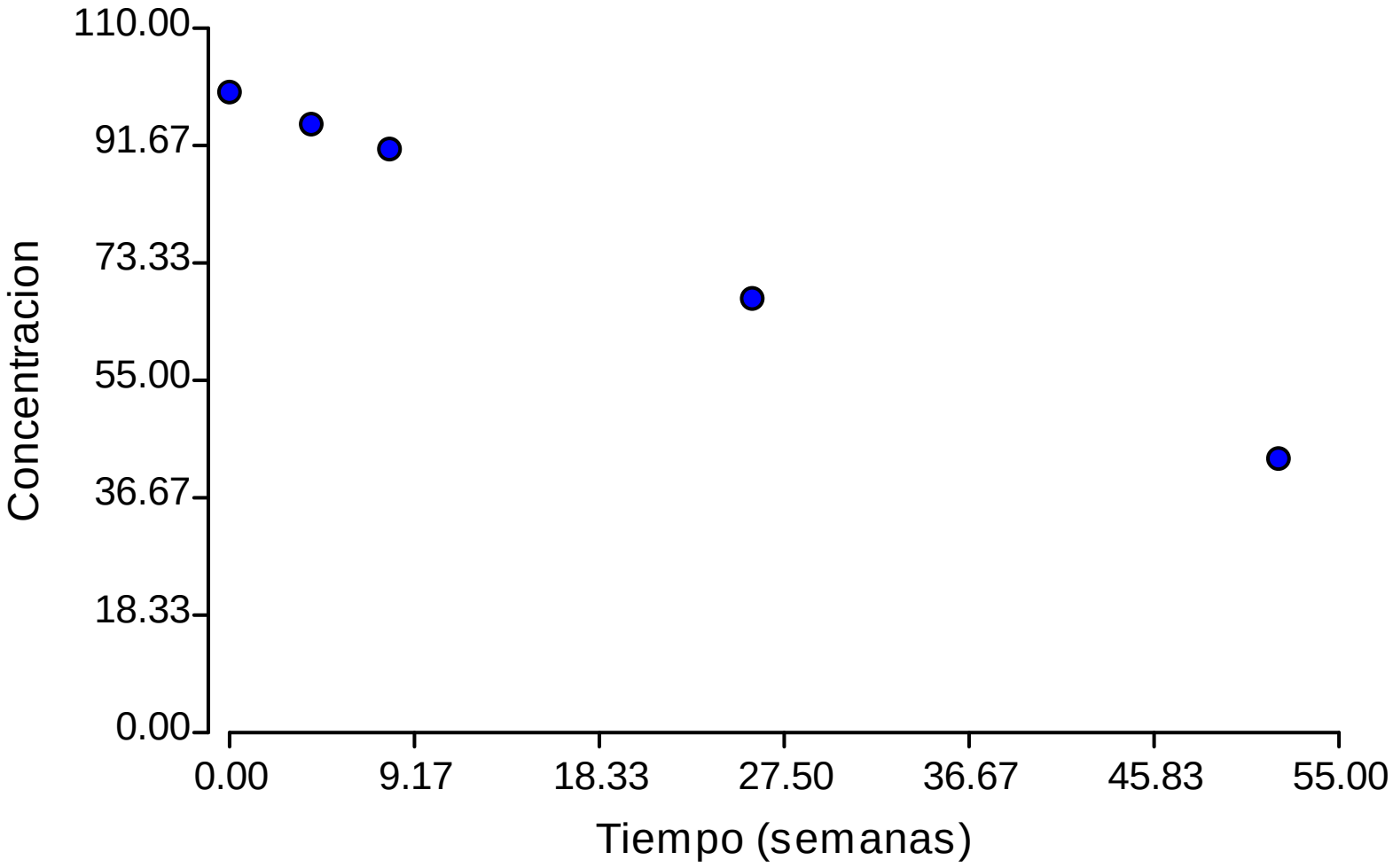
Se midió la concentración de una droga en solución como función del tiempo

Tiempo (semanas)	Concentración
0	100
4	95
8	91
26	68
52	43

Represente la concentración en función del tiempo.

¿Cómo se llama este tipo de gráfico?

Diagrama de dispersión

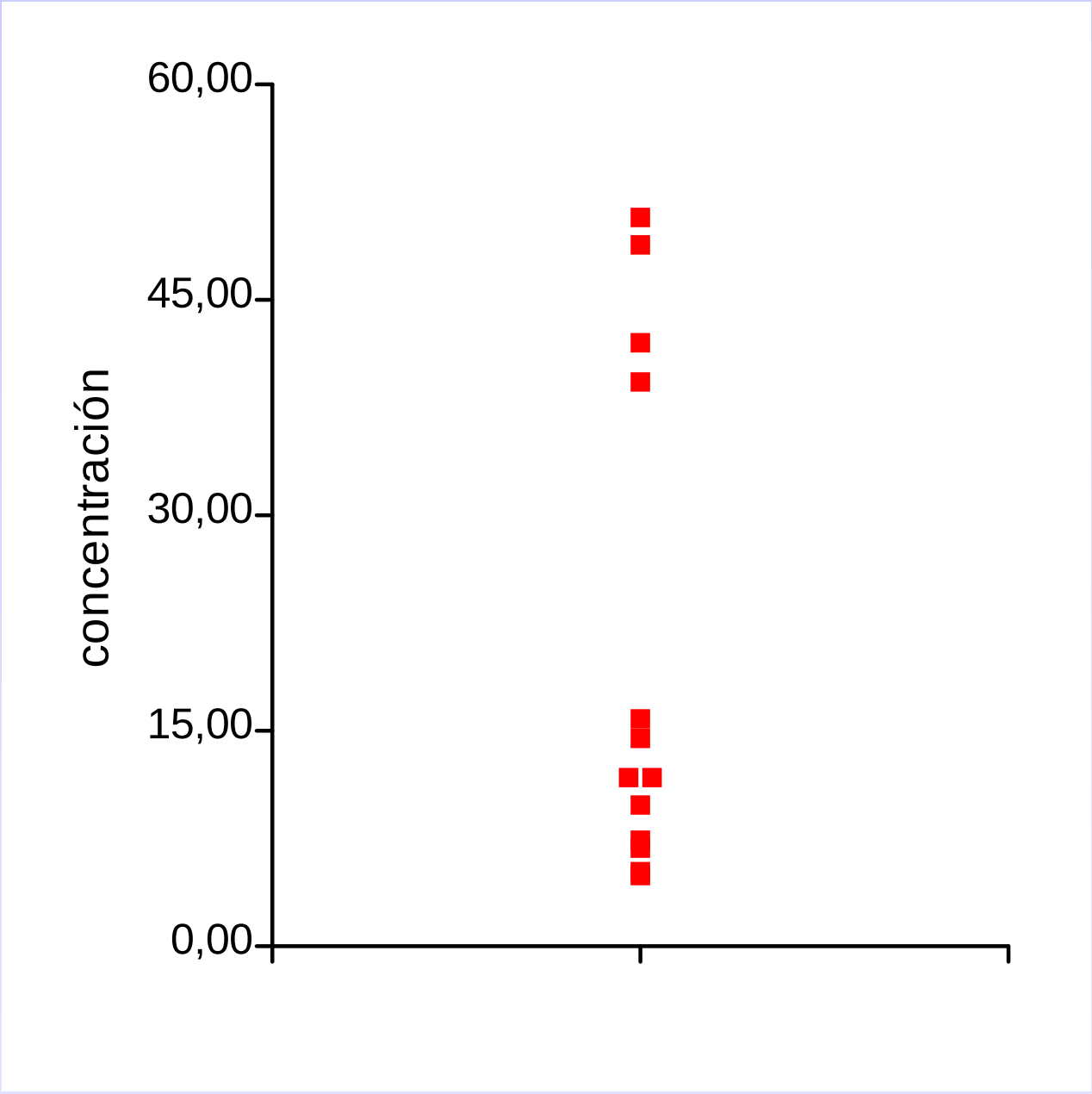

110.00
91.67
73.33
55.00
36.67
18.33
0.00
Concentracion
0.00
9.17
18.33
27.50
36.67
45.83
55.00
Tiempo (semanas)

Ejemplo:

Los siguientes valores de contenido de un metabolito en la sangre de un paciente en 13 extracciones diferentes :

11,6	39,2	4,9	7,3	50,6	9,8	11,6	6,7	42,1	14,4	5,1	48,8	15,9

Los datos están informados en $mg.L^{-1}$.
Haga un gráfico de densidad de puntos y analice los resultados.


60,00
45,00
30,00
15,00
0,00
concentración

Ejemplo:

La siguiente tabla muestra los resultados de un experimento de respuesta a una dosis, aplicado a tres grupos de 5 animales cada uno a los que se les aplicaron una determinada dosis.

Dosis (mg)	Respuesta
1	8, 12, 9, 14, 6
2	16, 20, 12, 15, 17
4	20, 17, 25, 27, 16

¿Qué gráfico haría ?

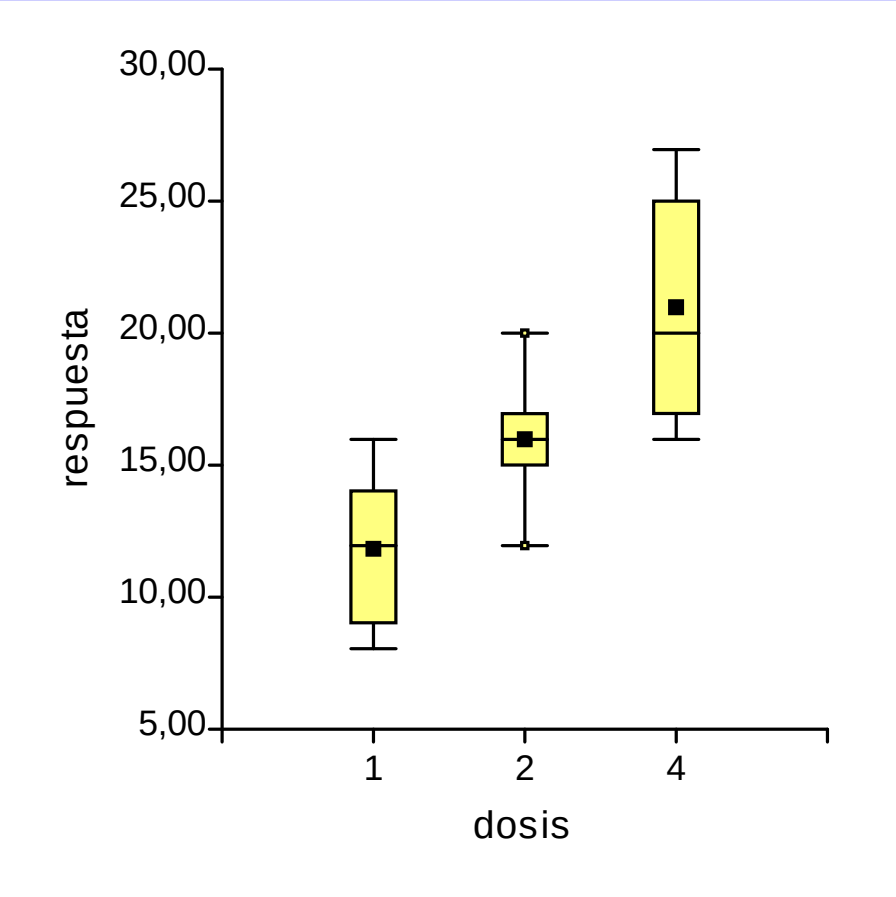


Grafico de cajas

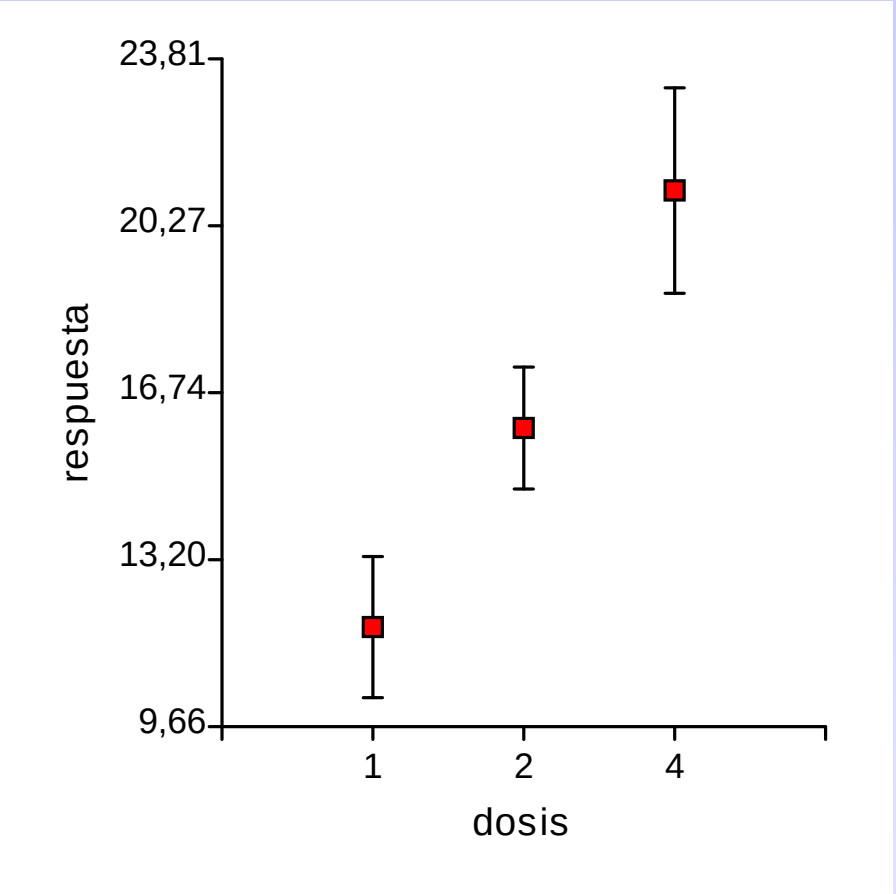


Grafico de puntos.

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

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

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

$f = Q_3 - Q_1 = 48,5$

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

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

NO HAY DATOS EXTREMOS.