Universidad Nacional Autónoma de México
Colegio de Ciencias y Humanidades
Plantel Vallejo

ÁREA DE MATEMÁTICAS
*ESTADÍSTICA Y PROBABILIDAD* I

# DIAGRAMA DE CAJA Y BIGOTES

En este escrito ilustraremos la construcción de una gráfica que suele formar parte de los informes de análisis estadístico y que es de gran utilidad como técnica de análisis al proporcionarnos una amplia visión sobre como se distribuyen los datos.

Dentro del análisis exploratorio en estadística, resultan útiles los *box plots* o diagrama de caja con bigotes.

Este tipo de gráfica fue desarrollado por J. Tukey (denominándolo *"box and whisker plot"*) para evaluar la forma de las distribuciones, ya que permite detectar problemas en las colas de la distribución (casos extremos, atípicos o errores), los cuales pueden distorsionar cualquier análisis de un conjunto de datos.

A diferencia de los otros gráficos, los diagramas de caja hacen énfasis en las medidas de posición.

El diagrama de caja-bigotes es un resumen gráfico que permite visualizar, para un conjunto de datos, la tendencia central, la dispersión y la presencia posible de datos atípicos. Para realizarlo se necesita calcular la mediana, el primer cuartil, y el tercer cuartil de los datos.

Este diagrama consiste en un rectángulo cuya longitud es el rango intercuartílico, el rectángulo está dividido por un segmento vertical que indica la posición de la mediana y esta complementado por dos líneas (llamadas bigotes) que parten de los extremos del rectángulo, cuya longitud puede llegar a ser el equivalente a 1.5 veces el rango intercuartílico y que intentan encerrar los valores mínimo y máximo observados.

La mayor utilidad de los diagramas caja-bigotes es para comparar dos o más conjuntos de datos.

## ¿CÓMO SE DIBUJA UN DIAGRAMA DE CAJA?

Un diagrama de caja se construye como sigue:

1) Se ordenar los datos de la muestra y se obtienen; el valor mínimo, el máximo, y los tres cuartiles Q1, Q2 y Q3.

2) Se dibuja un rectángulo (de anchura arbitraria) cuyos extremos son Q1 y Q3 y se indica en su interior la posición de la mediana, Q2, mediante una línea vertical.

3) Se calcula el rango intercuartílico del conjunto de datos: $Q = Q_3 - Q_1$

4) Se determinan los límites admisibles superior e inferior. Donde el límite inferior es igual al ***máximo*** entre el valor mínimo de los datos y el primer cuartil menos una vez y

media el rango intercuartílico. Y el límite superior es el ***mínimo*** entre el valor mayor de los datos y la suma del tercer cuartil con una vez y medio el rango intercuartílico.

$$L_i = \max(x_{\min},\ Q_1 - 1.5Q)$$
$$L_s = \min(x_{\max},\ Q_3 + 1.5Q)$$

Estos límites nos permitirán identificar los valores atípicos, que serán aquellos datos que queden fuera del intervalo $(L_i,\ \ L_s)$

5) Se dibuja una línea horizontal desde cada extremo del rectángulo central hasta el valor más alejado no atípico, es decir, que está dentro del intervalo $(L_i,\ \ L_s)$.

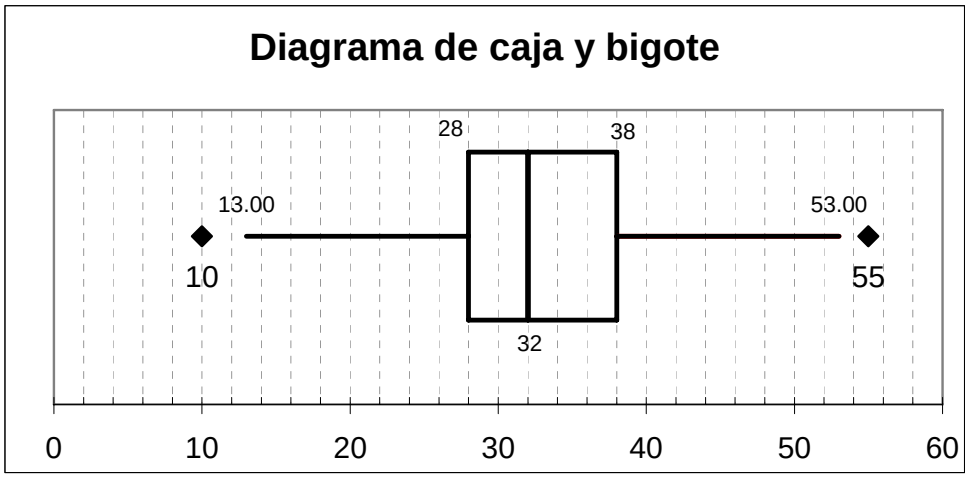
6) Identificar todos los datos que están fuera del intervalo $(L_i,\ \ L_s)$, marcándolos como atípicos.

Ejemplo: Construir el diagrama de caja para un conjunto de datos que tiene: valor mínimo 10, valor máximo 55, $Q_1 = 28$, $Q_2 = 32$, $Q_3 = 38$,
El rango intercuartílico es $Q = Q_3 - Q_1 = 38 - 28 = 10$, por lo que

$$L_i = \max(x_{\min},\ Q_1 - 1.5Q) = \max(10,\ \ 13) = 13$$

$$L_s = \min(x_{\max},\ Q_3 + 1.5Q) = \min(55,\ \ 53) = 53$$

Con todos estos elementos construimos el siguiente diagrama:




En el diagrama podemos observar que tenemos dos valores atípicos, (el 10 y el 55), puesto que ambos caen o están fuera del intervalo $(L_i,\ \ L_s) = (13,\ \ 53)$ (13, 53).

Además la distribución de datos es sesgada en forma positiva.

Recordemos que la caja central contiene el 50% de los datos. Cada línea o bigote puede contener hasta un 25% de los datos. Por último, es recomendable señalar con una marca especial los valores atípicos.

SIMETRÍA

Este diagrama nos muestra la posición relativa de la mediana, los cuartiles y extremos de la distribución. El diagrama de caja proporcionan una idea intuitiva de la simetría de la distribución de los datos: si la mediana no está en el centro del rectángulo esto significa que la distribución es asimétrica, indicando además el tipo de sesgo.

VALORES ATÍPICOS

Son aquellos valores que se alejan demasiado de los valores centrales de la distribución, es decir son valores que difieren bastante con respecto a la gran mayoría de ellos. El diagrama detecta la presencia de valores atípicos y sugiere o no la necesidad de utilizar estadísticos robustos.

RECONOCIMIENTO DE PATRONES

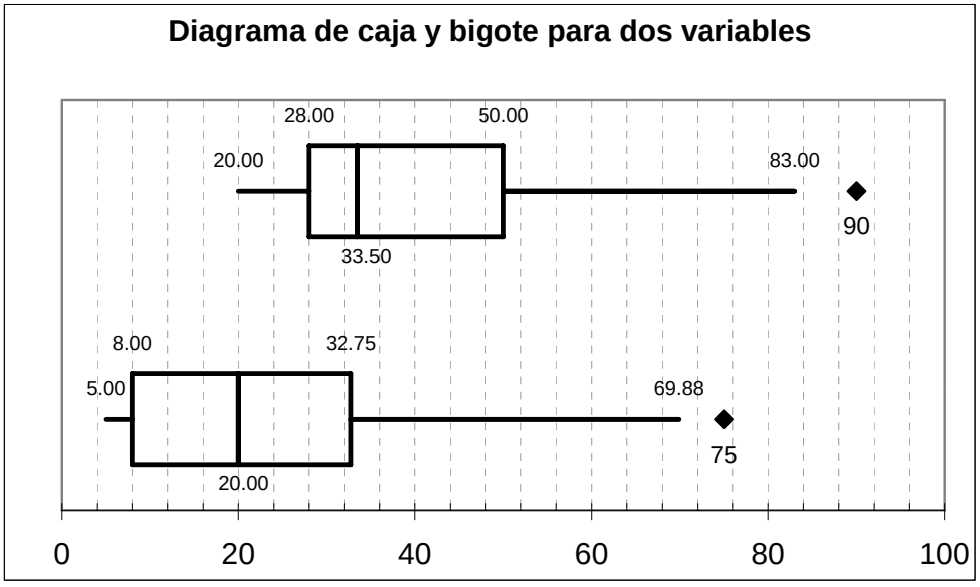
En general, pueden descubrirse diferentes patrones con este tipo de gráficos:
Si tenemos líneas (bigotes) largas implica una situación de desequilibrio, indicándonos que hay valores extremos.
Cajas grandes y líneas cortas indican que en dicha población presenta valores afines, por lo que tendríamos una baja dispersión.

COMPARACIÓN ENTRE POBLACIONES

La comparación entre poblaciones es sencilla de realizar ya que simplemente se construye un diagrama para cada población.

Ejemplo de dos poblaciones:

| | Mínimo | $Q_1$ | $Q_2 = \tilde{x}$ | $Q_3$ | Máximo |
|---|---|---|---|---|---|
| Población 1 | 5.00 | 8.00 | 20.00 | 32.75 | 75.00 |
| Población 2 | 20.00 | 28.00 | 33.50 | 50.00 | 90.00 |




¿Qué población tiene mayor sesgo?
¿Qué población tiene mayor dispersión?
¿Cuáles son los valores atípicos para cada población?
¿Qué conclusiones se pueden obtener al comparar estas dos poblaciones?
¿Cuál tiene de las poblaciones es más puntiaguda?

Para cada uno de los siguientes casos elabore el diagrama de caja – bigotes.

1.- La siguiente tabla muestra el resultado de una encuesta realizada en los hogares de la Ciudad de Guadalajara respecto al “numero de cuartos” en una casa habitación

| Número de cuartos por hogar | Frecuencia |
|---|---|
| 1 | 154 |
| 2 | 235 |
| 3 | 184 |
| 4 | 97 |
| 5 | 53 |

2.- La empresa automovilística “Autos de Excelencia” ha realizado un control de potencia sobre los 500 motores a gasolina que se han fabricado a lo largo del mes de noviembre del año 2001 obteniendo la siguiente tabla:

| Potencia en CV | Nº de motores |
|---|---|
| 0-50 | 30 |
| 50-60 | 100 |
| 60-65 | 200 |
| 65-70 | 150 |
| 70-80 | 20 |

3. –Una empresa interesada en determinar la edad de los empleados del departamento de producción, realizo un estudio, sobre este departamento obteniendo la siguiente distribución de frecuencias;

| Intervalo de edad | Frecuencia |
|---|---|
| 20-23 | 2 |
| 24-27 | 5 |
| 28-31 | 17 |
| 32-35 | 55 |
| 36-39 | 123 |
| 40-43 | 105 |
| 44-47 | 71 |
| 48-52 | 18 |

¿Cuál es la edad promedio de los empleados?
¿Cuál es la edad que más predomina?
¿Cuál es la edad para la cual el 50% de los empleados son menores que ella?