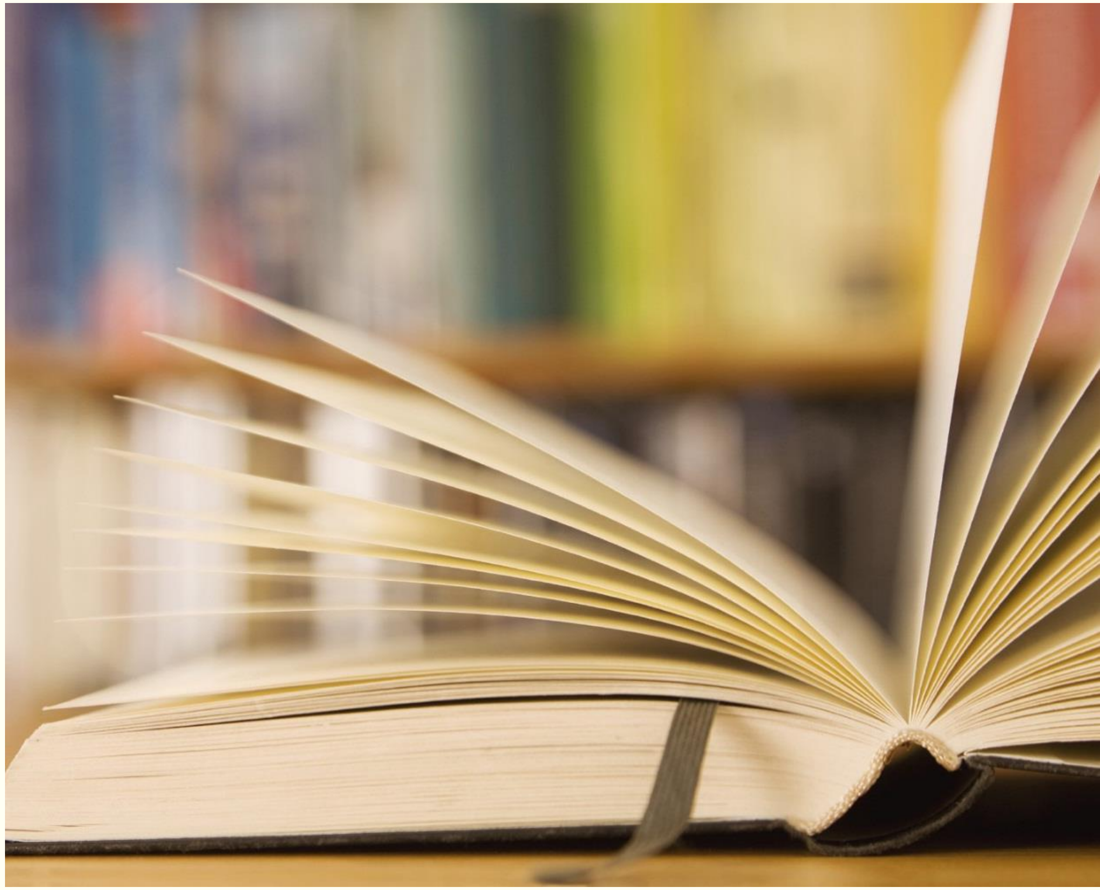
UNIVERSIDAD NACIONAL DE CHIMBORAZO

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA CIVIL



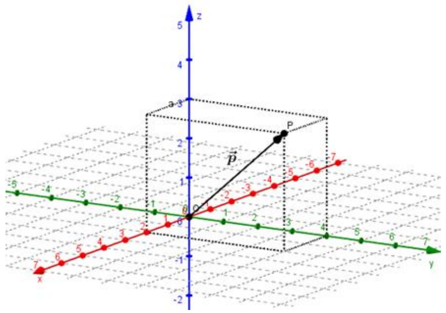
ASIGNATURA: Estadística

DOCENTE: Ing. Lidia Castro M.Sc

UNIDAD 1 → ESTADÍSTICA DESCRIPTIVA

OBJETIVO DE LA UNIDAD

Conocer y comprender el análisis de datos que intervienen en la estadística descriptiva.



1. Introducción
2. Definiciones fundamentales
3. Población y muestra
4. Frecuencia y rango
5. Distribución de frecuencias
6. Medidas de tendencia central
7. Medidas de dispersion

MEDIDAS DE POSICIÓN

Asociados a la mediana se tiene otras mediadas que se fundamentan en las divisiones proporcionales que pueden hacerse en datos agrupados o sin agrupar y se denominan **cuantiles.**

Al generalizar este concepto se define los cuartiles, deciles y percentiles, para ello si se divide los datos en:

4 partes iguales, se los llama cuartiles.

10 partes iguales, a estos valores se los llama deciles.

100 partes iguales, se los llama percentiles.

CUARTILES

- Los intervalos dentro de los cuales quedan proporcionalmente repartidos los datos sin agrupar o agrupados en una distribución formada por cuatro partes.

- Para los cuartiles, al dividir por 4, se tiene un "dato" que deja por debajo de él la primera cuarta parte de los datos (25%), un valor que deja la 2/4 parte de los datos, y otro para la 3/4 parte de los datos.

- Estos valores se denotan por Q_1, Q_2 y Q_3.

- Esta sería la idea intuitiva, pero al igual que para la mediana, no siempre existe tal dato, y entonces definimos el cuartil, como también el decil o el percentil, como un promedio.

Para localizar Q_1 se divide $\frac{N}{4}$

Para localizar Q_2 se realiza: $\frac{2 \times N}{4} = \frac{N}{2}$ que coincide con la mediana

Para localizar Q_3 $\frac{3 \times N}{4}$

CUARTILES – DATOS AGRUPADOS

El cálculo de los cuartiles de una distribución de frecuencias se determina por:

$$Q_k = L_i + \left[\frac{\frac{KN}{4} - (\sum f)_i}{f_k}\right] C$$

Donde:

f_k = frecuencia de la clase del $\mathrm{k}_{\text{ésimo}}$ cuartil, la cual no debe ser mayor que $KN/10$.

$K = k_{ésimo}\ cuartil, se\ refiere\ al\ primero, segundo\ y\ tercer\ cuartil.$

$L_i = límite\ real\ inferior\ de\ la\ clase\ del\ k_{ésimo}\ cuartil, es\ decir\ la\ clase\ que\ contiene\ el\ valor.$

$N = número\ total\ de\ datos\ (frecuencia\ total).$

$(\sum f)_i$ = suma de las frecuencias de todas las clases por debajo del $\mathrm{k}_{\text{ésimo}}$ cuartil.

$C = tamaño\ del\ intervalo\ de\ la\ clase\ del\ k_{ésimo}\ cuartil.$

DECILES

Con la misma idea se calcularán los deciles, al dividir los datos entre 10.

Se obtendría D1, D2,...,D9 donde Dk se calcula a partir del valor

$$\frac{K \times N}{10}$$

PERCENTILES

Los percentiles P1, P2, P3,…,P99 se calculan usando:

$$\frac{K \times N}{100}$$

DECILES – DATOS AGRUPADOS

$$D_k = L_i + \left[\frac{\frac{KN}{10} - (\sum f)_i}{f_k}\right] C$$

Donde:

f_k = frecuencia de la clase del $\mathrm{k}_{\text{ésimo}}$ decil, la cual no debe ser mayor que $KN/10$.

$K = k_{\text{ésimo}}\ decil = se\ refiere\ al\ primero, segundo, tercero, \ldots, noveno\ decil.$

$L_i = límite\ real\ inferior\ de\ la\ clase\ del\ k_{\text{ésimo}}\ decil, es\ decir\ la\ clase\ que\ contiene\ el\ valor.$

$N = número\ total\ de\ datos\ (frecuencia\ total).$

$(\sum f)_i$ = suma de las frecuencias de todas las clases por debajo del $\mathrm{k}_{\text{ésimo}}$ decil.

$C = tamañano\ del\ intervalo\ de\ la\ clase\ del\ k_{\text{ésimo}}\ decil.$

PERCENTILES – DATOS AGRUPADOS

$$P_k = L_i + \left[\frac{\frac{KN}{100} - (\sum f)_i}{f_k}\right] C$$

Donde:

$f_k =$ frecuencia de la clase del $\mathrm{k_{ésimo}}$ percentil, la cual no debe ser mayor que $KN/10$.

$K = k_{ésimo}\ percentil\ se\ refiere\ al\ primero, segundo, tercero, \ldots, nonogéismo\ noveno\ percentil.$

$L_i = límite\ real\ inferior\ de\ la\ clase\ del\ k_{ésimo}\ percentil, es\ decir\ la\ clase\ que\ contiene\ el\ valor.$

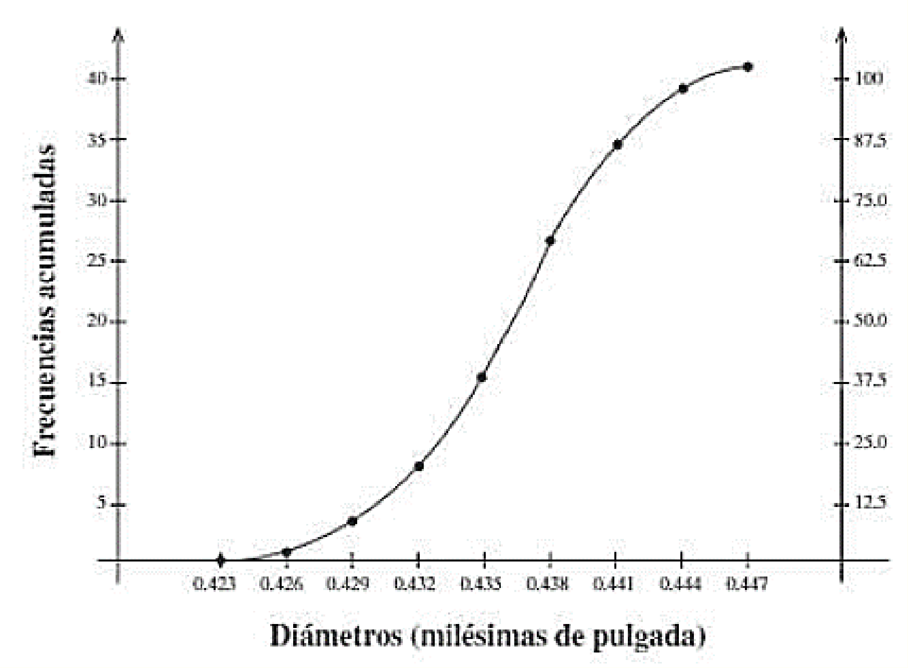
$N = número\ total\ de\ datos\ (frecuencia\ total).$

$(\sum f)_i =$ suma de las frecuencias de todas las clases por debajo del $\mathrm{k_{ésimo}}$ percentil.

$C\ = tamañano\ del\ intervalo\ de\ la\ clase\ del\ k_{ésimo}\ percentil.$

Para datos agrupados en tablas de frecuencia

Para el caso que no tener los datos completos y solo una tabla de frecuencias, se puede determinar los cuartiles, deciles y percentiles a partir del polígono de frecuencias acumuladas (ojiva).



RANGO INTERCUARTÍLICO

- Es la diferencia de sus cuartiles superior e inferior.
- Es una medida de cuán alejada está la porción media de los datos en valor.

$$RIQ = Q_3 - Q_1$$

Diagrama de Caja y bigotes (box and whisker plot o BoxPlot)

- Es un tipo de gráfico que muestra un resumen de una gran cantidad de datos en cinco medidas descriptivas (cuartiles y otras), además de intuir su morfología y simetría.
- Este tipo de gráficos permite identificar valores atípicos y comparar distribuciones.
- Además de conocer de una forma cómoda y rápida como el 50% de los valores centrales se distribuyen.
- Las dimensiones de la caja está determinada por la distancia del rango intercuartílico, que es la diferencia entre el primer y tercer cuartil.

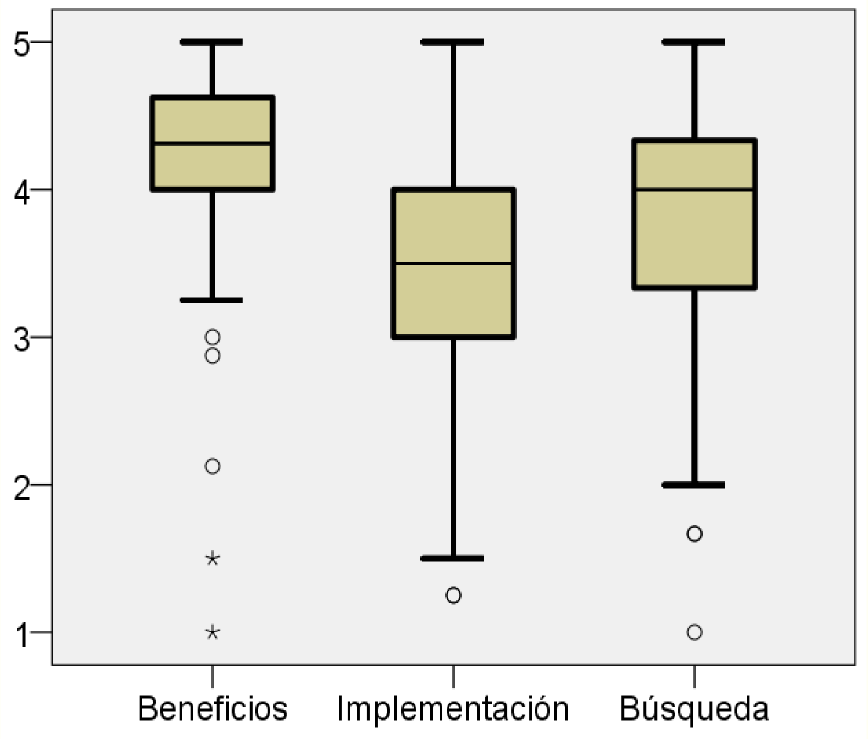


Diagrama de Caja y bigotes (box and whisker plot o BoxPlot)

- El segmento que divide la caja en *dos partes es la mediana*, que facilitará la comprensión de si la distribución es simétrica o asimétrica.

- Si la mediana se sitúa en el centro de la caja entonces la distribución es simétrica y tanto la media, mediana y moda coinciden.

- Si la mediana corta la caja en dos lados desiguales se tiene:

  a) Asimetría positiva o segada a la derecha si la parte más larga de la caja es la parte superior a la mediana. Los datos se concentran en la parte inferior de la distribución. La media suele ser mayor que la mediana.

  b) Asimetría negativa o sesgada a la izquierda si la parte más larga es la inferior a la mediana. Los datos se concentran en la parte superior de la distribución. La media suele ser menor que la mediana.

Diagrama de Caja y bigotes (box and whisker plot o BoxPlot)

- Que un lado de la caja sea más largo que otro, no quiere decir que ese lado contenga más datos. Indica un rango más amplio, por lo que los datos estarán mas dispersos. Un rango menos amplio, indica que los datos están más próximos.

- La continuación de dos segmentos en la caja se denominan bigotes (whisker) que determina el límite para la detección de valores atípicos. Los bigotes deben tener una longitud máxima. Dicha longitud no debe ser superior al 150% del rango intercuartílico.

- Habrá un límite superior, que no podrá superar el 1.5 veces el RIC, si el máximo no supera ese valor, la longitud del bigote será desde el tercer cuartil hasta el máximo.

- Habrá un límite inferior, que no podrá superar el 1.5 veces el RIC, si el mínimo no supera ese valor, la longitud del bigote será desde el primer cuartil hasta el mínimo.

- Los valores atípicos (outilers) son aquellos puntos que están mas allá del límite inferior o superior.

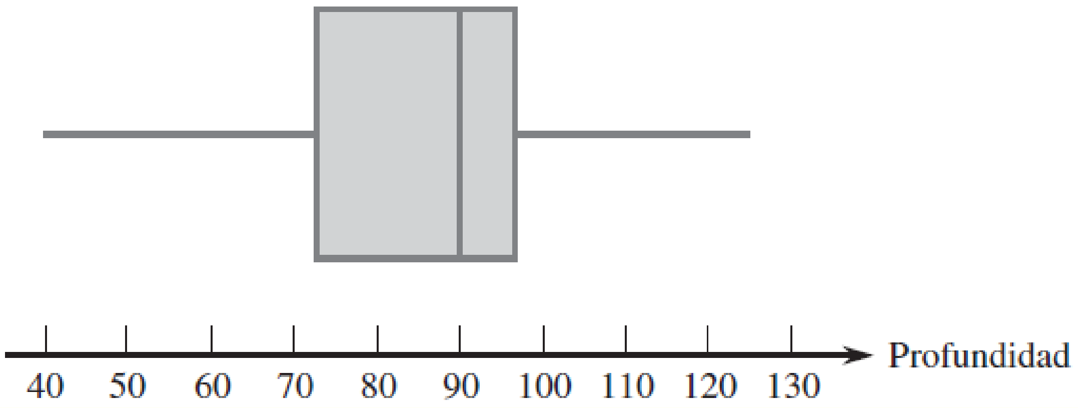
EJEMPLO

Se utilizó ultrasonido para reunir los datos de corrosión adjuntos de la placa de piso de un tanque elevado utilizado para almacenar petróleo crudo (“Statistical Analysis of UT Corrosion Data from Floor Plates of a Crude Oil Aboveground Storage Tank”, *Materials Eval.*, 1994: 846-849); cada observación es la profundidad de picadura más grande en la placa, expresada en milésimas de pulgada.

40 52 55 60 70 75 85 85 90 90 92 94 94 95 98 100 115 125 125

El resumen de cinco números es como sigue:

x_i más pequeña = 40 cuarto inferior = 72.5 $\tilde{x}$ = 90 cuarto superior = 96.5
x_i más grande = 125



EJERCICIO

1. En el semáforo de una ciudad se realiza una revisión vehicular aleatoria y los número de autos revisados fueron:

27, 22, 19, 16, 28, 35, 33, 39, 54, 60, 53, 48, 65, 76, 68, 83, 89, 92, 103, 85, 57, 43, 46, 41

Determine:

- Los cuartiles Q1, Q2, Q3
- Los deciles D1,D2,...,D9
- Los percentiles P1, P2, P3, P10, P35, P55, P95, P99

EJERCICIO

2. La siguiente tabla de distribución de frecuencias registra salarios semanales en cientos de dólares de 75 empleados. Determinar Q1, Q2, Q3, los deciles D1, D2, D3,…,D9 y los percentiles P1, P2, P3, P15, P40, P75, P95.

Intervalos (salarios)	Frecuencia (empleados)
110 - 119	5
120 - 129	9
130 - 139	19
140 - 149	15
150 - 159	11
160 - 169	7
170 - 179	4
180 - 189	3
190 - 199	2
	$N = 75$

Ejemplo

3. Realice el diagrama de cajas y bigotes del siguiente problema.

Los efectos de descargas parciales en la degradación de materiales para cavidades aislantes tienen implicaciones importantes en relación con las duraciones de componentes de alto voltaje. Considérese la siguiente muestra de $n = 25$ anchos de pulso de descargas lentas en una cavidad cilíndrica de polietileno. (Estos datos son consistentes con un histograma de 250 observaciones en el artículo "Assessment of Dielectric Degradation by Ultrawide-band PD Detection", *IEEE Trans. on Dielectrics and Elec. Insul.*, 1995: 744-760.) El autor del artículo señala el impacto de una amplia variedad de herramientas estadísticas en la interpretación de datos de descarga.

5.3	8.2	13.8	74.1	85.3	88.0	90.2	91.5	92.4	92.9	93.6	94.3	94.8
94.9	95.5	95.8	95.9	96.6	96.7	98.1	99.0	101.4	103.7	106.0	113.5	

Las cantidades pertinentes son

$\tilde{x} = 94.8$	cuarto inferior $= 90.2$	cuarto superior $= 96.7$
$f_s = 6.5$	$1.5f_s = 9.75$	$3f_s = 19.50$

¡GRACIAS!