ESTADÍSTICA DESCRIPTIVA BÁSICA

Argemiro Domínguez Villafañe[1]

INTRODUCCIÓN

La Estadística se ha considerado como el estudio de un conjunto de datos con el fin de transformarlos en información que pueda ser utilizada para tomar decisiones; sin embargo dar una definición formal de Estadística es menos importante que entender cómo puede ayudar en la solución de problemas en diferentes áreas; en particular la estadística es un fuerte soporte para la calidad de las conclusiones y recomendaciones que se desprenden de un estudio. El rápido desarrollo de paquetes computacionales de análisis estadístico ha facilitado la aplicación de técnicas estadísticas para el análisis de grandes volúmenes de datos; sin embargo es importante antes de hacer cualquier análisis estadístico, cuestionarnos un poco acerca de la validez estadística de los datos como: su representatividad, confiabilidad en instrumentos de medición utilizados, identificación de efectos de interacción y de factores de confusión.

La Estadística descriptiva constituye el primer paso en un análisis estadístico, ya que permite observar patrones en los datos, detectar errores en la información, conocer indicadores de tendencia central, de dispersión, etc.; esta información nos permite tener un conocimiento general de los datos y determinar si éstos cumplen los supuestos teóricos necesarios para análisis más complejos.

1. VARIABLES

Proposiciones tales como: El lote No 1 produjo 12 toneladas de uva en el primer ciclo del año, la dosis aplicada de Dormex como promotor de brotación fue efectiva, la intensidad de la adopción de la tecnología fue del 30%, son comunes e informativas dentro de nuestro trabajo; estas frases se refieren a características que no son constantes, sino que varían de un individuo a otro y que permiten describirlos y distinguirlos.

Lo anterior permite introducir el termino VARIABLE, el cual no es mas que una característica observada en dos o más individuos. Dado que este concepto es importante en análisis estadístico para un conjunto de datos, a continuación se da una descripción breve de las clases de variables.

[1] Economista Agrario, Esp. Biometría. CORPOICA, Centro de Investigación Palmira. Programa de Biometría. E-mail:ardominguez@telesat.com.co

1.1 Clasificación de variables

Las variables se pueden clasificar de acuerdo con su naturaleza, es decir lo que ella representa, y con el tipo de escala que permite hacer su medición.

1.1.1 Según su naturaleza

Variables Cualitativas: son variables cuyos valores corresponden a atributos o categorías que describen o identifican un individuo.

- Variables Binarias: Son variables que toman solo dos valores, como por ejemplo: masculino / femenino, positivo / negativo.
- Variables Multinomiales: Corresponde a variables que pueden tomar mas de dos posibles valores los cuales corresponden a niveles prefijados de una categoría, por ejemplo: color de la hoja (verde, amarilla, café, negra).

Variables Cuantitativas: Son variables que permiten identificar y describir sujetos con diferencia en cantidad o distancia.

- Variable continua: es aquella que puede asumir cualquier valor numérico en un intervalo dado.
- Variable discreta: es aquella que toma valores numéricos enteros únicamente.

1.1.2 Según las escalas de medida

La estadística tiene como primer objetivo el estudio cuantitativo de un conjunto de variables, las cuales son medidas utilizando diferentes escalas de medición.

Medición: Asignación de un número a la característica observada

Escalas de Medición

Nominal: Es el tipo de medición más limitado; los elementos se clasifican en categorías según un atributo.

Ordinal: Se establece una relación de orden (<, >, =); el orden en las categorías se debe especificar antes de hacer la medición.

Intervalo: Utiliza un cero arbitrario; no se pueden establecer razones por cociente.

Razón: Utiliza un cero real, por lo tanto se pueden utilizar todas las operaciones matemáticas.

Las variables cualitativas según su escala se consideran:

Variables Nominales: es la escala de medida con menor precisión, los datos corresponden al número de individuos en cada clase o categoría. Ej.: Tipo de hongo que esta afectando el cultivo (roya, mildeo, oidium), Forma de aplicar las recomendaciones (bien o mal), tipo de productor (veterano, maduro o iniciado). En esta clase de variable solo se habla de igualdad o diferencia entre los valores.

Variables Ordinales: Poseen un nivel mayor de precisión respecto a las anteriores; los valores se pueden ordenar con base en la cantidad del atributo poseído. Los valores se comparar en términos de mayo que o menor que. Al calificar el estado de un cultivo (bueno, regular, malo), el nivel de ingestación de por una plaga (alta, media baja), efecto de la aplicación del dormex para inducir brotación (brotación pareja, no broto, broto irregularmente).

Dentro de las variables cuantitativas están:

Variables de Intervalo: Aquí se tienen el concepto de distancia de un punto a otro y además se puede saber cual es el mejor, se puede medir que tanto es mejor; un ejemplo clásico de variable de intervalo es la temperatura: 20 °C es menor que 27°C, exactamente en 7°C.

En esta escala existe un cero arbitrario ya que 0°C no es precisamente la ausencia de temperatura y tampoco se puede decir que 30°C es 2 veces más caliente que 15°C.

Variables de Razón: Aquí se consideran dos aspectos adicionales. Existe un cero absoluto que indica un punto de partida real y existe el concepto de razón; permite todas las operaciones matemáticas. Consideremos el peso de un racimo de uva, cero (0) gramos indica ausencia de peso, pero también podemos afirmar que si un racimo peso 50 gramos y otro 100 gramos el segundo tiene el doble del peso del primero.

Rendimiento en peso de uva, número de racimos con perlilla, cantidad de plantas en descanso, número de visitas que hace el TAE al viticultor en el año, altura de planta a los seis meses de sembrada.

Para el análisis estadístico es importante conocer tanto la forma como ha sido medida cada variable, como los valores (distintos estados de cada variable) que toma con el fin de utilizar las estadísticas apropiadas. Por ejemplo la variable producción puede considerarse:

- Nominal: Si se habla del fin de la producción: uva para embandejado, consumo fresco, transformación industrial.
- Ordinal: Si se habla de la calidad de la producción. Primeras, segundas o terceras.
- Razón: Si se considera la cantidad en kilos cosechados en cada lote.

2. ANÁLISIS DESCRIPTIVO DE DATOS

Los métodos utilizados para describir conjuntos de datos numéricos se agrupan en métodos gráficos y métodos numéricos. No existe una regla única y discriminante para cada conjunto de datos, se deben utilizar ambos métodos en forma complementaria para el análisis.

Para introducir estos métodos, se hará uso de conjuntos de datos tomados de la experiencia en la explotación de la uva Isabela, en la zona de influencia de CorpoGinebra, municipios de El Cerrito, Ginebra y Guacarí, en el Valle del Cauca.

2.1 Distribución de Frecuencias

Tablas de frecuencia: método numérico que permite arreglar y ordenar los datos en intervalos.

- Ejemplo 1

En un ensayo de agroquímicos se observa el número de racimos picados por pájaros cuando se ha aplicado un tratamiento repelente específico; en el ensayo se consideraron 40 muestras de 100 racimos cada una.

Datos : 4, 2, 3, 5, 4, 2, 1, 6, 3, 6, 4, 5, 6, 6, 5, 4, 3, 5, 5, 3, 1, 6, 4, 3, 2, 3, 4, 6, 2, 3, 2, 5, 4, 3, 5, 1, 2, 1, 4, 3.

Tabla de distribución de frecuencias:

Clase	n_i	f_i	N_i	F_i
1	4	4/40	4	0.100
2	6	6/40	10	0.250
3	9	9/40	19	0.475
4	8	8/40	27	0.675
5	7	7/40	34	0.850
6	6	6/40	40	1.000
Total	40	1.00		

Interpretación:

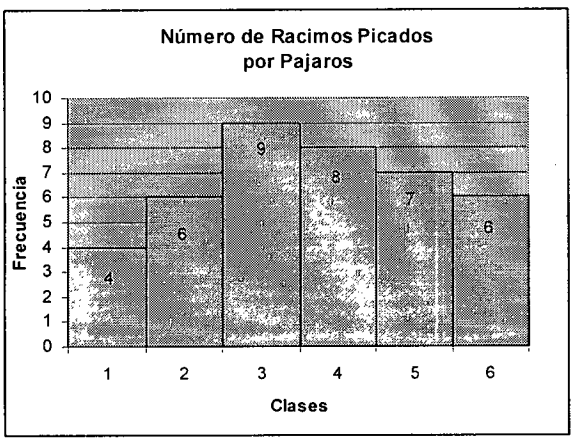
$n_2 = 6$: 6 de las 40 muestras presentan 2 racimos picados

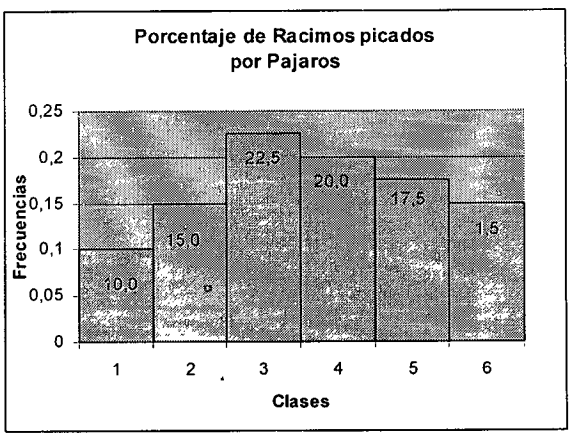
$f_4 = 0.2$: El 20% de las muestras presentan 4 racimos picados

$N_3 = 19$: De las 40 muestras, 19 tienen 3 o menos racimos con picaduras

$F_5 = 0.85$: El 85% de las muestras tienen 5 o menos racimos picados

Histograma de Frecuencias: Método gráfico que permite presentar la distribución de los datos organizados en clases o intervalos.





- Ejemplo 2

En un ensayo de fungicidas, con el fin de observar el efecto curativo de un producto X, sobre la Roya en uva, se registró el porcentaje de incidencia (presencia/ausencia de la enfermedad en la planta).

Datos:19.5, 20.3, 18.8, 19.4, 22.3, 22.4, 21.9, 26.0, 22.1, 22.5, 24.7, 23.5, 24.2, 23.9, 21.8, 19.5, 18.2, 21.4, 22.6, 21.6, 21.7, 18.0, 24.3, 21.8, 19.0, 22.2, 24.8, 23.6, 24.0, 20.9, 23.1, 21.9, 25.8, 19.6, 25.3, 26.0, 23.0, 23.4, 25.3, 22.8, 24.0, 19.8

No. de clases = $1 + 3.3 \text{Log}_{10}(n)$

m = $1 + 3.3 \text{Log}_{10}(42)$ = 6.35 Entonces $m \approx 6$

Longitud del Intervalo: C_i = (Valor Máximo - Valor mínimo)/ m

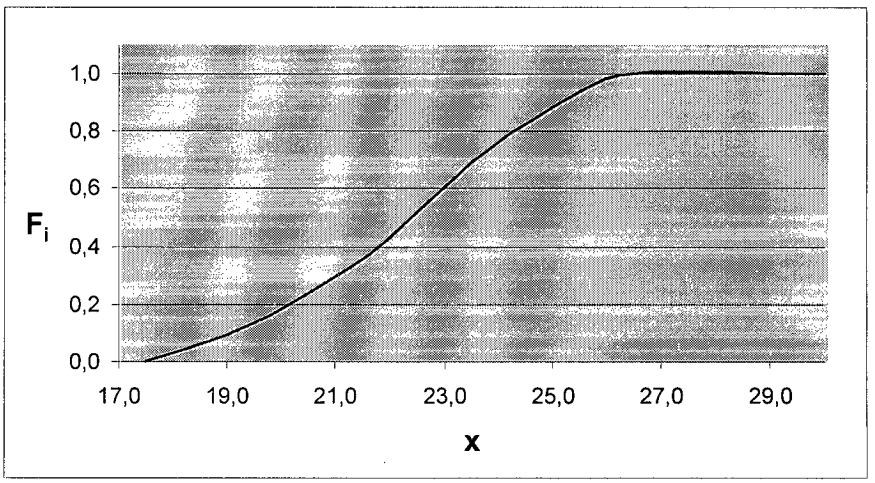
$C_i = (26 - 18) / 6 = 1.33$ Luego $C_i \approx 1.5$

Punto de inicio = Valor mínimo – (m * Ci – Rango) / 2 = 18 – (6*1.5 – 8) / 2 = 17.5

Tabla de distribución de frecuencias:

Intervalo de Clase	Marca de Clase	n_i	f_i	N_i	F_i
(17.5-19.0]	18.25	4	0.095	4	0.095
(19.0-20.5]	19.75	6	0.143	10	0.238
(20.5-22.0]	21.25	8	0.190	18	0.428
(22.0-23.5]	22.75	11	0.262	29	0.690
(23.5-25.0]	24.25	8	0.190	37	0.880
(25.0-26.5]	25.75	5	0.119	42	1.000
Total		42	1,000		

Función de distribución acumulativa: Método grafico que permite representar la forma como se van acumulando los datos en cada clase en un orden ascendente



- Ejemplo 3.

En el estudio de la Pre-adopción de la tecnología entregada por CorpoGinebra desde 2001 hasta 2003 para el control de la Perlilla, requirió encuestar a los viticultores de los municipios, El Cerrito, Ginebra y Guacarí. Se requiere conocer:

i. La distribución de los encuestados por las UBA que le corresponden a Ginebra
ii. La distribución de los encuestados de acuerdo con la zona y municipio a donde están ubicados los lotes
iii. La distribución de los encuestados de acuerdo con el municipio que pertenecen

i. Distribución por UBA. Para el municipio de Ginebra.

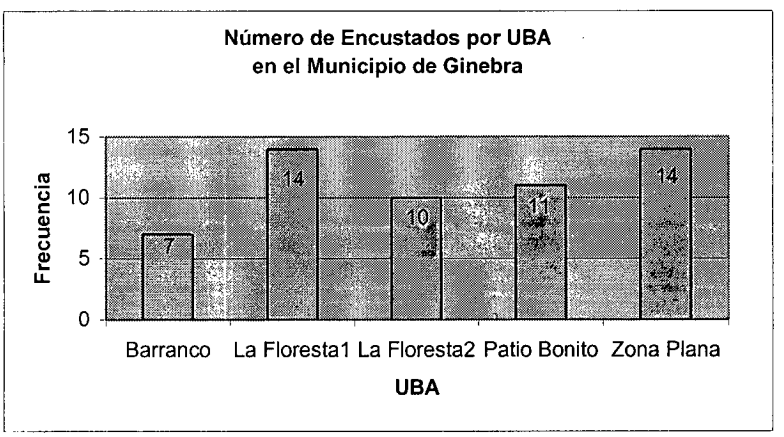
Conteo:

UBA	Encuestas
Barranco	/////// = 7
La Floresta1	////////////// = 14
La Floresta2	////////// = 10
Patio Bonito	////////// = 11
Zona Plana	////////// //// = 14

Tabla de distribución de frecuencias:

Clase	n_i	f_i	N_i	F_i
Barranco	7	0,13	7	0,13
La Floresta1	14	0,25	21	0,38
La Floresta2	10	0,18	31	0,55
Patio Bonito	11	0,20	42	0,75
Zona Plana	14	0,25	56	1,00
Total	56	1.00		

Diagramas de Barras: Método grafico que permite representar las clases cuando corresponden a una variable cualitativa

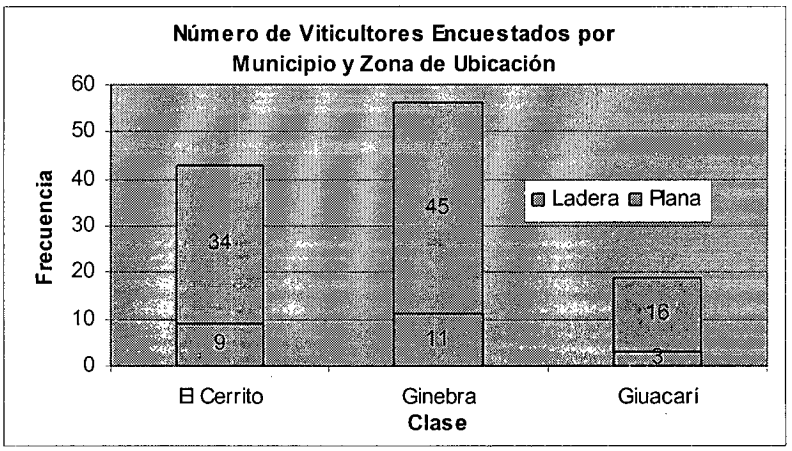


ii. La distribución de acuerdo con la zona y municipio

Tabla de distribución de frecuencias: doble entrada

Clase		Zona		n_i	f_i
		Plana	Ladera		
Municipio	El Cerrito	34	9	43	0.36
	Ginebra	45	11	56	0.48
	Guacarí	16	3	19	0.16
Total		95	23	118	1.00

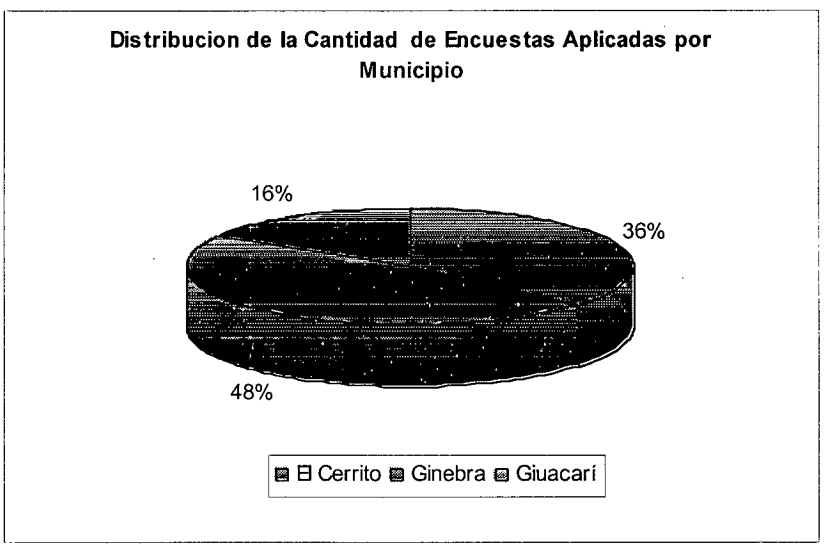
Diagramas de Barras



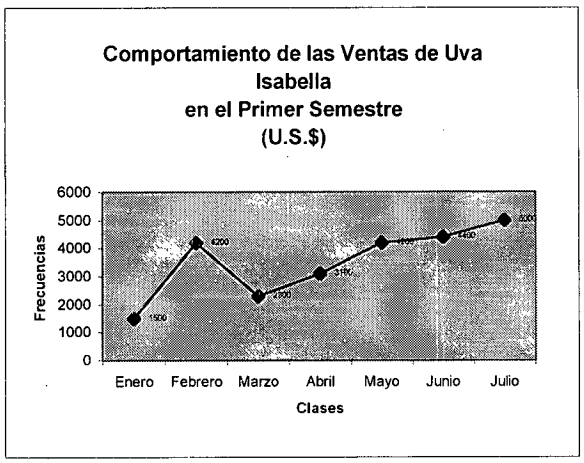
iii. La distribución de acuerdo con el municipio que pertenecen

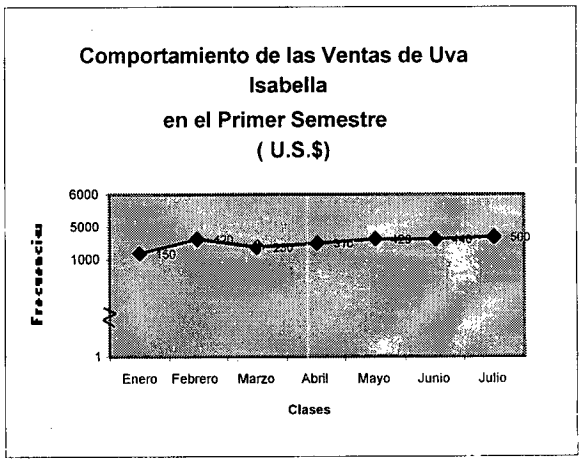
A partir de la tabla de frecuencias de doble entrada anterior, se tiene.

Diagramas Circulares: Método grafico que permite mostrar como se reparte un total entre las clases.



Graficas que engañan: Aun cuando los métodos gráficos para describir datos son muy útiles, deben ser examinados con mucho cuidado. Es posible en algunos casos construir graficas que inducen al lector inexperto a conclusiones inadecuadas. El método más común para confundir al lector es modificar la escala de los ejes de la gráfica.





La segunda grafica deja ver rápidamente un comportamiento casi uniforme en las ventas a través del semestre; mientras que la primera deja ver el incremento que se presenta en febrero y luego su descenso en marzo, además se observa una tendencia a aumentar las ventas. Otra forma de aumentar o encoger la pendiente de la grafica es alargar o encoger los ejes.

2.2 Medidas de tendencia central

Media aritmética:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$ Datos simples

$$\bar{x} = \sum_{i=1}^{m} \frac{n_i x_i}{n}$$ Datos agrupados

- Ejemplo 1

Datos: 12, 14, 15, 18, 21

$\bar{x}$ = [12+14+15+18+21] / 5 = 16

Para el ensayo de agroquímicos:

$\bar{x}$ = [4*1+6*2+9*3+8*4+7*5+6*6]/40 = 3.65

Características:

- La suma de las desviaciones con respecto al promedio es cero.

x_i	$x_i - \bar{x}$
12	-4
14	-2
15	-1
18	2
21	5
Σ	0

- Sensible a valores extremos

A: 1, 3, 5, 7, 10 $\longrightarrow$ $\bar{x}$ = 5.2

B: 1, 3, 5, 7, 10, 28 $\longrightarrow$ $\bar{x}$ = 9.0

- Si Y = X + K entonces $\bar{y} = \bar{x} + k$
- Si Y = k*X entonces $\bar{y} = k^* \bar{x}$;

Donde X y Y son variables y k es una constante

La mediana:

Valor por debajo del cual se encuentra el 50% de los datos

Datos simples:

Si n es par $p = n/2$ y $q = n/2+1$

$$Me = (x_p + x_q) / 2$$

Si n es impar $r = (n + 1) / 2$

$$Me = x_r$$

- Ejemplo 1

n par : 16, 14, 10, 12

x_1= 10, x_2=12, x_3= 14, x_4=16
p = 4/2 = 2 , q = 3
Me = $(x_2+x_3)/2$ = (12+14)/2 = 13

n impar : 14, 10, 12

$x_1 = 10,\ x_2 = 12,\ x_3 = 14$
$r = (3+1)/2 = 2$
$Me = x_2 = 12$

Datos agrupados:

$$M_e = L_{i-1} + \frac{\left[0.5 - F_{L_{i-1}}\right]}{f_i} * C_i$$

L_{i-1}: Límite inferior del intervalo de clase que contiene la mediana.

F (L_{i-1}): Frecuencia relativa acumulada hasta el intervalo de clase anterior al que se encuentra el 50% de los datos.

f_i : Frecuencia relativa asociada al Intervalo donde está el 50% de los datos.

C_i: Longitud del I. de Clase donde se encuentra el 50% de los datos.

Características:

- No es sensible a valores extremos
- Su cálculo no involucra todos los datos

- Ejemplo 1

A partir de los datos del ensayo de sobre dosis de funguicidas para el control de la roya en uva, y de su tabla de distribución de frecuencias se tiene:

L_{i-1}: 22.0

F (L_{i-1}): 0.428

f_i : 0.262

C_i: 1.5

$$M_e = 22.0 + \frac{\left[0.5 - 0.428\right]}{0.262} * 1.5 = 24.41$$

La Moda:

Corresponde al valor o valores que más se repiten.

- Ejemplo 1

A: 4, 5, 5, 5, 8,3 ⟶ Mo = 5

B: 3, 6, 6, 6, 4, 3, 5, 6 ⟶ Mo = 3, 6

C: 4, 8, 6, 5, 2 ⟶ Mo: no hay

Características:

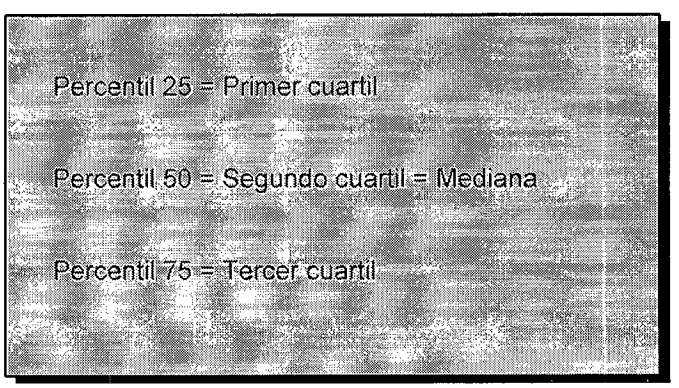
- No siempre se puede calcular.
- Con datos agrupados su valor depende de los límites de clase considerados.

2.3 Medidas de Posición

Indican la posición relativa de una calificación.

Rango percentil: porcentaje de los casos que alcanzan valores menores que el citado

Percentiles utilizados con más frecuencia:

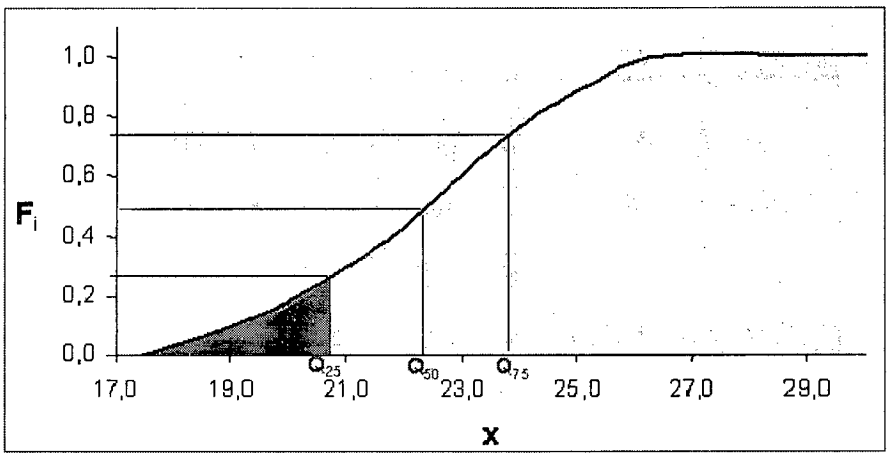


Cuartiles

Q25: El 25% de los datos están por debajo de este valor

Q50: El 50% de los datos están por debajo de este valor

Q75: El 75% de los datos están por debajo de este valor



Forma de Cálculo:

$$Q_k = F_{L_{i-1}} + \left[\frac{x - l_{i-1}}{c_i}\right] * f_i$$

- Ejemplo 1

Para el ensayo de fungicidas, calcular el cuartil 25:

$$0.25 = 0.238 + \left[\frac{x-20.5}{1.5}\right] * 0.190$$

x = 20.59 ≈ 21

2.4 Medidas de dispersión

Describen variabilidad de los datos, cuantificando su dispersión con respecto a una medida de tendencia central.

Desviación Media:

$$DM_e = \sum_{i=1}^{n} \frac{|x_i - M_e|}{n}$$ Datos simples

$$DM_e = \sum_{i=1}^{m} \frac{n_i |x_i - M_e|}{n}$$ Datos agrupados

- Ejemplo 1

A partir de los datos del ensayo de sobre dosis de funguicida para el control de la roya en uva se tiene:

Intervalo de Clase	Marca de Clase x_i	n_i	M_e	$\lvert x_i - M_e \rvert$	$\frac{n_i \lvert x_i - M_e \rvert}{n}$
(17.5-19.0]	18.25	4	22.41	4,16	0,396
(19.0-20.5]	19.75	6	22.41	2.66	0,380
(20.5-22.0]	21.25	8	22.41	1.16	0,221
(22.0-23.5]	22.75	11	22.41	0.34	0,088
(23.5-25.0]	24.25	8	22.41	1.84	0,350
(25.0-26.5]	25.75	5	22.41	3.34	0.397
		n = 42			DM= 1.832

Varianza y Desviación Estándar

Miden la dispersión de los datos con respecto a la media aritmética.

Varianza:

$$S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$ Datos simples

$$S^2 = \sum_{i=1}^{m} \frac{n_i (x_i - \bar{x})^2}{n-1}$$ Datos agrupados

Desviación Estándar:

$$S = \sqrt[2]{S^2}$$

Coeficiente de variación:

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

Permite comparar la variabilidad para dos conjuntos de datos ya que éste es un indicador adimensional.

- Ejemplo 1

Para los datos de Agroquímicos, en el control del número de racimos picados por pájaro se tiene:

Clase	n_i	$\bar{x}$	$(x_i - \bar{x})^2$	$\frac{n_i (x_i - \bar{x})^2}{n-1}$
1	4	3.65	7,02	0,720
2	6	3.65	2,72	0,419
3	9	3.65	0,42	0,098
4	8	3.65	0,12	0,025
5	7	3.65	1,82	0,327
6	6	3.65	5,52	0,850
Total	40			2,438

$$\bar{X} = [4*1+6*2+9*3+8*4+7*5+6*6]/40 = 3.65$$

$$S^2 = 2.438$$

$$S = \sqrt{2.438} = 1.561$$

$$CV = \left(\frac{2.438}{3.65}\right)*100 = 42.7$$

BIBLIOGRAFÍA

Dallas e. Jonson. 2000. Métodos Multivariados Aplicados al Análisis de Datos. Internacional Thomnson editores, México. 572 p.

Guilford, J. P. y Fruchter, Benjamín. 1994. Estadística Aplicada a la Psicología y la Educación. McGraw-Hill, Colombia. 234 p.

Mendenhall, William y Reinmuth, James E. 1981. Estadística para Administración y Economía. Grupo editorial Iberoamericana, México. 260 p.

Mesa, F. Eloina. 1997. Curso de Estadística para Investigadores. Documento de trabajo, Colombia. 139p.