CAPITULO VI.

ESTADISTICA APLICADA

Luis Alfredo Valdés Hernández

I. Estadística

- Es una rama de las matemáticas que se especializa en datos de enumeración y en su relación con los datos métricos.
- La información cuantitativa apropiada para análisis estadístico deber ser un conjunto (o conjuntos) de números que muestren relaciones significativas.
- Los datos estadísticos son números que pueden ser comparados, analizados e interpretados. Un número aislado que no se compara o que no muestra relación significativa con otro número no es dato estadístico.

1.1 El Proceso de la Investigación científica.

Para avanzar en este tema es necesario hacer referencia al proceso de investigación científica, que servirá para ubicar a la estadística, así como su importancia, dentro del citado proceso.

- La investigación científica es rigurosa y cuidadosamente realizada, y se define como un tipo de investigación "Sistemática, controlada, empírica, crítica, de proposiciones hipotéticas sobre las presumidas relaciones entre fenómenos naturales "[5].

- El considerarla sistemática y controlada implica que hay una disciplina constante para hacer investigación científica y que no se dejan los hechos a la casualidad, empírica significa que se basa en fenómenos observables de la realidad. Y crítica quiere decir que juzga constantemente de manera objetiva, eliminando las preferencias personales y los juicios de valor.

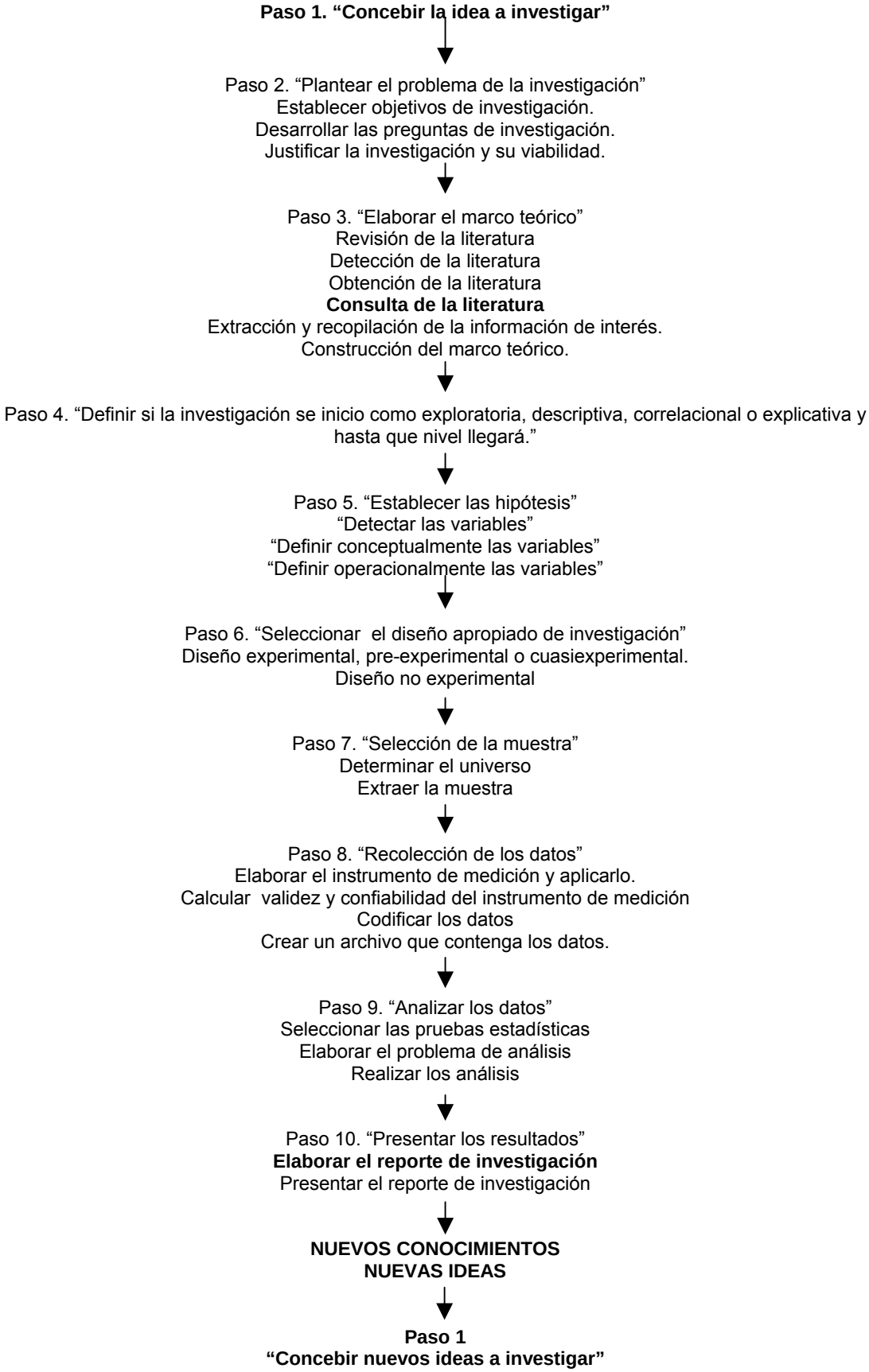
La investigación científica es un proceso, dinámico, cambiante y continuo. De acuerdo a Hernández este proceso esta compuesto por una serie de etapas (figura 1), las cuales son continuas e indivisibles. Por ello, al llevar a cabo un estudio o investigación, ni se puede, ni se debe omitir ninguna etapa, o alterar su orden. Aquellos que violaran este principio de la investigación científica, terminarán con investigaciones no válidas o no confiables, o que no cumplió con el propósito para los que fue realizada y por lo tanto dejó de ser científica.[6]

---

[5] Kerlinger, F.N.; "Investigación del comportamiento: técnicas y metodología"; Nueva Editorial Interamericana; México D.F. 1975.

[6] Hernández Sampieri R., Fernández collado, C., Baptista Lucio P.,; "Metodología de la Investigación"; Edit. Mc Graw Hill, México D.F. 1991.

FIGURA 1. Etapas del proceso de investigación científica.



La principal característica de la investigación científica es que debemos seguir ordenadamente y de manera rigurosa el proceso.

La investigación puede cumplir dos propósitos fundamentales:
a) Producir conocimiento y teorías (investigación básica).
b) Resolver problemas prácticos (investigación aplicada).

Es por estos dos tipos de investigación que la humanidad ha evolucionado.

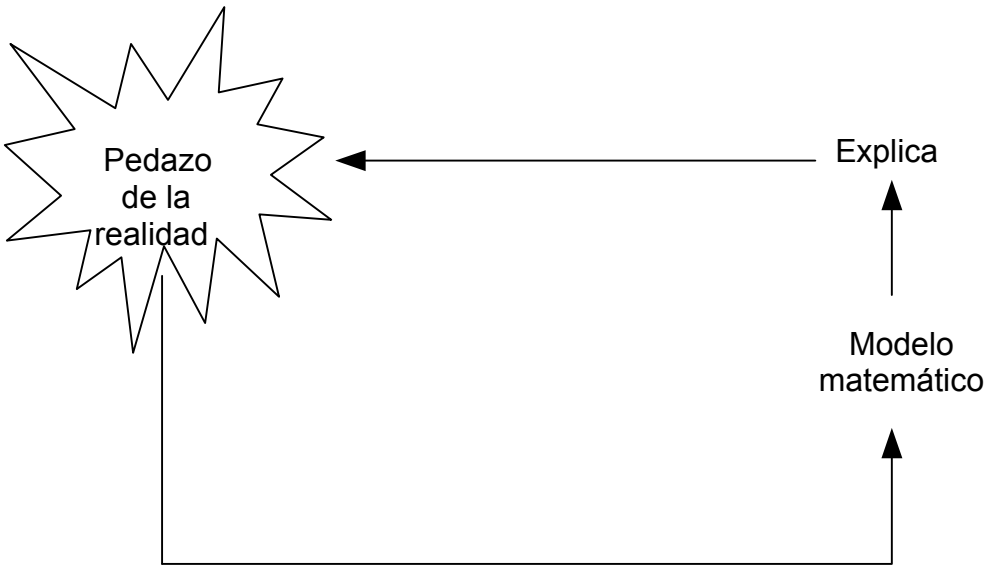
Con la aplicación del proceso de investigación científica se generan nuevos conocimientos, los cuales a su vez producen nuevas ideas e interrogantes para investigar, y es así como avanzan las ciencias y el conocimiento en general.

En la gráfica 1 se puede observar claramente la importancia tanto del conocimiento como de la correcta aplicación de la Estadística y sus métodos, dentro del proceso de la investigación científica.

1.2 Aplicación de la estadística y sus métodos.

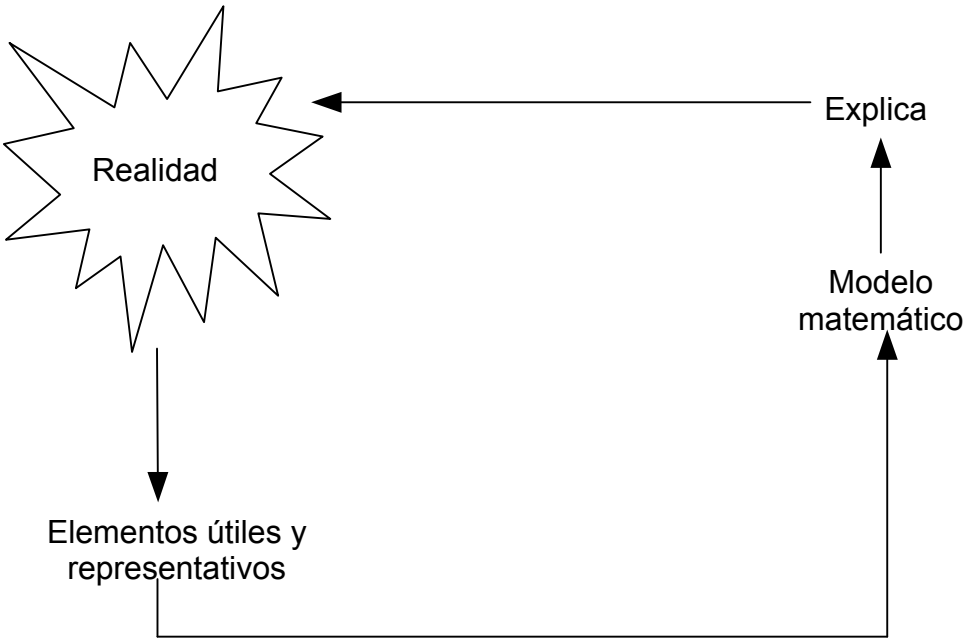
En las ciencias sociales la estadística es una herramienta que al aplicarla nos auxilia para llegar a conocer un pedazo de la realidad, mediante la representación matemática de esa realidad, al que llamaremos modelo matemático.
La utilidad de éste modelo es nos permite tomar decisiones para modificar esa realidad disminuyendo las condiciones de incertidumbre. En otras palabras, podemos tomar decisiones con mayor seguridad.

Figura 2. La gráfica como objeto de estudio y su relación con el modelo matemático que la explicará.



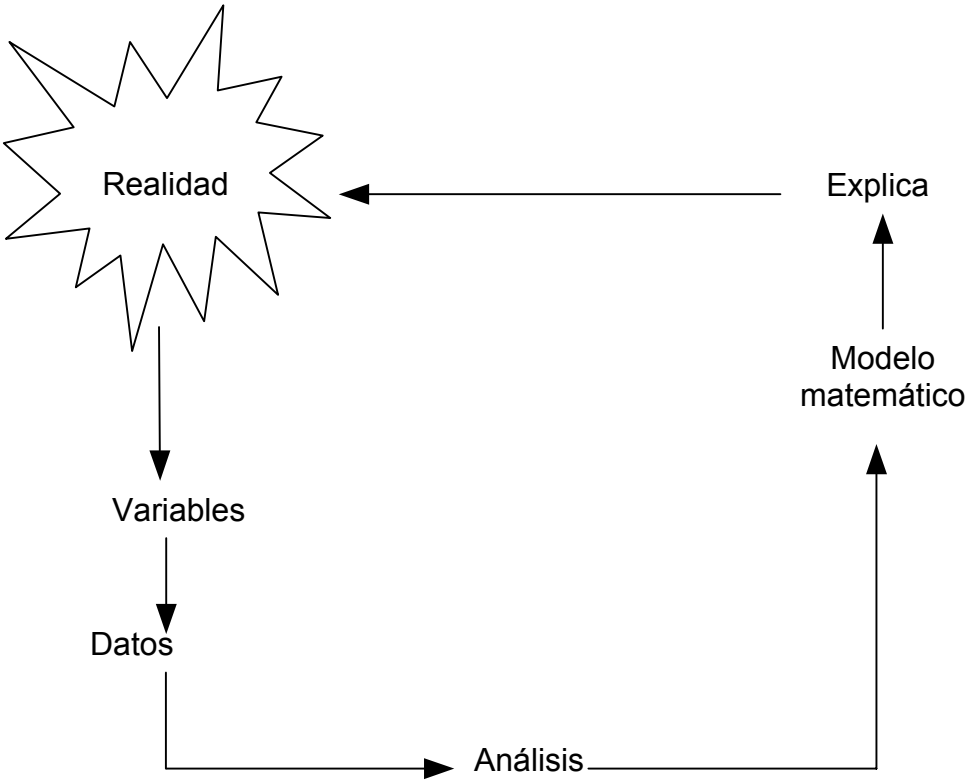
El modelo matemático explica la realidad en función de los datos con que se construye. Esto quiere decir que el modelo es adecuado si los datos son correctos, ya que las relaciones matemáticas son independientes de la representatividad de los datos. Por lo tanto para desarrollar un modelo que sea útil primero deberemos conocer lo más posible acerca de esa realidad con el fin de identificar sus elementos más útiles y representativos de la misma. Con estos últimos se construye el modelo como concepto.

Figura 3. La elaboración del modelo se hace con elementos representativos de la realidad



A esos elementos útiles y representativos los llamaremos variables, las cuales al definirles sus unidades de medida nos proporcionaran datos numéricos (también llamados índices, indicadores, etcétera.) mismos que al analizarlos nos permitirán establecer por análisis el tipo de relación matemática que guardan las variables representativas de la realidad bajo estudio.

Figura 4. Los elementos representativos se transforman en variables y se vuelven útiles cuando se les define escalas y unidades de medida para que nos proporciones datos numéricos



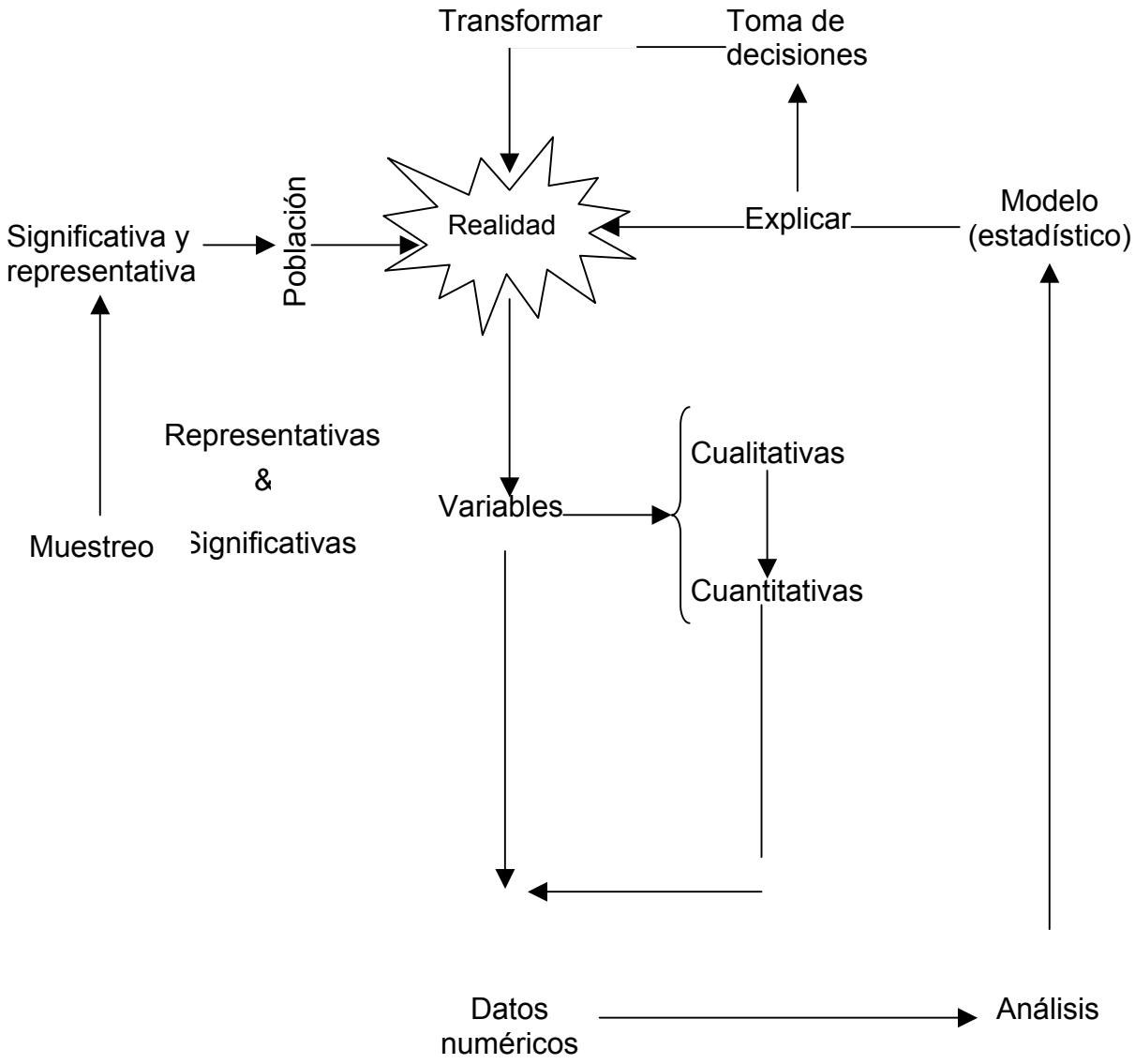
Para que las variables seleccionadas sean útiles deben de poseer dos características básicas: representatividad y significancia.

La primera característica, representatividad es el grado en que esa variable puede explicar la realidad bajo estudio. Por ejemplo si lo que deseamos estudiar es la salud de los paciente, algunas de las variables que la pueden explicar son los signos vitales, frecuencia cardiaca, temperatura corporal, respiración y presión arterial, etc. Estas variables se miden en latidos por minuto, grados celsius o Fahrenheit, inhalaciones-exhalaciones por minuto y mm de Hg.

La segunda característica se considera el grado de seguridad en su ocurrencia, consideremos por ejemplo que medimos la presión arterial de un paciente, después de que éste ha estado haciendo ejercicio, y el dato es de 200/160, este único dato, aun cuando es de una variable representativa, no se puede considerar significativo para los fines de explicar el estado de salud del paciente, por lo que deberemos tomar mas datos en un numero suficientes

como para que podamos considerar que son realmente significativo para explicar ese pedazo de realidad que estamos estudiando.

Figura 5. Relaciones entre la realidad y los elementos del modelo



Una ultima consideración para este proceso es el modelo estadístico obtenido forzosamente deberá servir para tomar decisiones que transforman esa realidad. De no ser así tan solo habremos desarrollado un mero ejercicio académico.

1.3 Métodos Estadísticos.

La palabra “estadística” ha sido frecuentemente referida a la información cuantitativa o numérica. También ha sido referida ampliamente a los métodos que tratan con la información. Sin embargo, esto debería aclararse y llamar a la información *datos estadísticos* y a los métodos *métodos estadísticos*.

La información cuantitativa o numérica puede encontrarse casi dondequiera, sin embargo no toda la información cuantitativa es considerada como dato estadístico. La información cuantitativa apropiada para análisis estadístico debe ser un conjunto (o conjuntos) de números que muestra relaciones significativas. En otras palabras, los datos estadísticos son números que pueden ser comparados, analizados e interpretados. Un número aislado que no se compara o que no muestra relación significativa con otro número no es dato estadístico.

El área en la cual los datos estadísticos son recopilados es generalmente referida como la población o universo. Una población puede ser finita o infinita. Una población finita tiene un numero limitado de individuos u objetos, mientras que una población infinita tiene un número ilimitado. La tarea de recopilar un conjunto completo de datos de una población finita pequeña es relativamente simple; Sin embargo, recopilar tales datos de una población finita pero grande, es algunas veces imposible o impráctico; la recopilación de datos completos de una población infinita es definitivamente imposible. A fin de evitar las tareas imposible o impráctica, usualmente se extrae una muestra de elementos representativos de la población; la muestra es entonces, utilizada para su estudio estadístico y los resultados de la muestra son usados como las bases para describir, estimar o predecir las características de la población.

De acuerdo con el orden de aplicaciones en un estudio estadístico, los métodos estadísticos son divididos en cinco pasos básicos:

1.3.1 Recopilación – Recopilación de datos estadísticos.

La información cuantitativa suministra hechos para resolver problemas. Después de que el problema ha sido definido y entendido, ciertos hechos relevantes que pueden ser presentados cuantitativamente, sí los hay, deberán ser recopilados. De acuerdo a la localización de la información, los datos estadísticos pueden ser clasificados en: 1) datos internos y 2) datos externos. Cuando la información cuantitativa es obtenida dentro de la organización que hace el estudio estadístico, la información es llamada datos internos; y cuando es obtenida fuera de la organización, es llamada datos externos. Los datos externos son usualmente obtenidos de dos maneras: a) datos publicados y b) encuesta de datos originales.

1.3.2 Organización – Organización de datos recopilados.

Los datos recopilados de fuentes publicadas están usualmente en forma organizada. Sin embargo, una gran masa de cifras que son recopiladas en una

encuesta frecuentemente necesitan de una organización. El primer paso al organizar un grupo de datos es la corrección. Los datos recopilados deben ser corregidos muy cuidadosamente, de tal manera que las omisiones, inconsistencias, respuestas irrelevantes y cálculos equivocados en los resultados de la encuesta puedan ser corregidos o ajustados. El siguiente paso es la clasificación. El propósito de este es decidir las clasificaciones adecuadas en las cuales los datos corregidos serán agrupados. Las clasificaciones adecuadas son muy importantes en los estudios estadísticos, puesto que los pasos sucesivos (es decir, presentación, análisis e interpretación de los datos) son afectados por las clasificaciones dadas. El último paso es la tabulación. Los elementos semejantes son numerados y registrados en esta etapa de acuerdo con las clasificaciones adecuadas.

1.3.3 Presentación de datos organizados en una forma fácil de leer.

Después de que los datos recopilados son organizados de acuerdo con clasificaciones adecuadas, están listos para su presentación. Los datos presentados en una forma fácil de leer pueden facilitar el análisis estadístico.

Existen tres formas básicas de presentar los datos recopilados y son:

1) mediante enunciados,
2) tablas estadísticas y
3) gráficas estadísticas.

El enunciado mediante palabras es conveniente para presentar un conjunto pequeño de datos. Cuando el conjunto es grande, la presentación mediante palabras se vuelve ineficiente y pesada, puesto que las clasificaciones, unidades de medidas y otras explicaciones detalladas tendrán que ser repetidas muchas veces en el enunciado. Las tablas estadísticas de doble entrada (matriciales) sí son bien construidas son de gran ayuda.

Una gráfica o diagrama estadístico es un medio para presentar datos estadísticos, que bien utilizados dan una buena aproximación del conjunto de datos en un instante. Sí se desea un valor exacto es preferible la tabla estadística.

1.3.4 Análisis – Análisis de los datos presentados.

Los métodos empleados en analizar datos estadísticos son numerosos y van desde la simple observación de los datos hasta los métodos complicados, sofisticados y de investigación matemática.

1.3.5 Interpretación – Interpretación de los resultados del análisis.

Después de que el análisis de los datos estadísticos está completo, los resultados del análisis deben ser interpretados. La interpretación correcta guiará a una conclusión válida del estudio y así poder auxiliar en la construcción de las conclusiones.

Los métodos estadísticos se pueden describir como métodos utilizados para obtener conclusiones respecto a poblaciones, por medio de muestras[7]. En vez de referirse a los métodos estadísticos como tales, normalmente se les llamará Estadística simplemente.

3. Estadística descriptiva e Inferencial.

A los métodos estadísticos concernientes a la recopilación y resumen de los datos se les llama *Estadística Descriptiva*. Y los concernientes a la obtención de las conclusiones a partir de la fuente de información de los datos significativos se les llama *Estadística Inferencial.*

Esta división de la Estadística la aplicamos en nuestras observaciones como cuando, por ejemplo, utilizando los datos de la Primera Convención Bancaria[8] . tenemos la Figura 6 que representa una tabla descriptiva.

Figura 6: Tablas descriptivas para robos bancarios reportados.

Año	1988	1989	1990	1991	1992	1993	Ier. Cuatrimestre 1994
Robos	272	207	163	215	194	136	54
bancarios reportados	22%	16.7%	13.1%	17.3%	15.6%	11%	4.3%

Fuente: El Financiero, 9 de mayo de 1994

A estos datos organizados los consideraremos Estadística Descriptiva, pero sí indicamos que los robos bancarios cometidos en 1991 aumentaron un 31.9% con respecto a 1990, podríamos inferir que en el intervalo de un año se incrementaron los robos bancarios en 31.9%. Se usaran los datos descriptivos para hacer una inferencia estadística.

La estadística Inferencial y la Descriptiva son básicas para el proceso de la investigación científica. El razonamiento lógico conlleva un método inductivo, así como la Estadística Inferencial siempre va del conocimiento de los datos muestrales al conocimiento de la población; no así la Lógica Deductiva y la Probabilidad, que se orientan del universo a los hechos particulares o de la población a la muestra.

[7] Hoel, Paul G.; "Estadística Elemental; CECSA, México 1986.
[8] El Financiero, 9 de mayo de 1994.

En el proceso de la investigación científica los métodos más importantes son los inferenciales, más no por ello podemos soslayar la importancia intrínseca de los métodos descriptivos.

Señalando que un estadístico es una medida usada para describir alguna característica de la muestra y los parámetros son las medidas descriptivas de la población[9].

El manejo adecuado de los métodos estadísticos implica el conocimiento estructurado de los siguientes conceptos básicos:

a) Medidas de Tendencia Central.
b) Medidas de dispersión.
c) Medidas de relación.

Los cuales serán tratados con amplitud en aportados capitulos subsecuentes.

[9] Phillips, John L.; "la lógica del pensamiento estadístico", Editorial El Manual Moderno S.A.; México 1980.

Capitulo VI.

MEDIDA Y CONTEO

Dos tipos de datos numéricos

Los datos numéricos son generalmente de dos categorías principales.

Objetivo	Obtiene	Se llaman
Se cuentan cosas	Frecuencias	Datos de enumeración
Se miden cosas	Valores métricos o valores de escala	Medidas o datos métricos

Los procedimientos estadísticos tratan ambos tipos de datos.

Los datos y la estadística.

Estadística: Es una rama de las matemáticas que se especializa en datos de enumeración y en su relación con datos métricos.

Datos: Detalles en registros o informes numéricos.

Datos en categoría.

Probablemente la mayoría de los datos sociales están en forma de frecuencias categóricas, o sea los números de casos en clases o categorías definidas.

Clasificación

Antes de contar, con el fin de acumular información útil, tenemos que saber qué es lo que estamos contando. No se cuenta indiscriminadamente. La frecuencia que se registra se refiere a una clase particular de objetos o sucesos, y ello supone un proceso de clasificación anterior.

La clasificación es un proceso muy útil y necesario en la ciencia como en la vida práctica. Es el procedimiento mediante el cual los objetos quedan repartidos en categorías para su recuento.

Al progresar la ciencia, es probable lograr abstraer *variables* de sus datos. Las variables son variaciones continuas en direcciones separadas.

Muchas categorías que se usan en investigaciones no son estáticas sino que cambian a medida que se aclara más el campo de estudio. Ciertas categorías se inventan para una labor temporal como estructura provisional sobre la cual se disponen los datos ara su mejor examen.

Al elegir o construir categorías útiles, éstas deberán ser bien definidas, mutuamente exclusivas y exhaustivas

Bien definidas: la asignación apropiada de los casos a las clases depende de lo adecuado de su definición.
Mutua exclusividad: significa que hay una y solo una base de clasificación, su ausencia indica interdependencia.
Exhaustivas: para ser exhaustivas las categorías deberán ofrecer lugar para "todos" los casos.
Categoría cualitativa: por sus atributos o cualidades.
Categoría cuantitativa: se ordenan de acuerdo a la cantidad.

Ejemplos:

1) Clasificación de la vida
- Vegetal
  - Especie
  - Género
  - Ordenes
- Animal

2) Categorías absolutas:
- Vivos y Muertos.
- Casados y Solteros.
- Aprobados y No-aprobados.
- Criminales y No-criminales.}

ESTADISTICA DESCRIPTIVA FUNDAMENTAL

Datos:

Los datos fueron inventados por los hombres como un sistema simbólico de ideas internamente coherente que pueden utilizarse eficazmente para describir el mundo tal como lo conocía, logrando así el control del mismo.

Los datos numéricos son de 2 categorías

Uso	Obtiene	Denominan
Se cuentan cosas	Frecuencias	Datos de enumeración
Se miden cosas	Valores métricos Valores de escala	Medidas Datos métricos

CAPITULO VI.
MEDIDAS DE TENDENCIA CENTRAL

MEDIA ARITMÉTICA

Un promedio puede representar a la muestra o a la población de la cual salió dicha muestra. Una población contiene a todos los elementos que comparten ciertas características, podemos hablar de promedio cuando indicamos que en la Zona Metropolitana de la Ciudad de México (ZMCM) cada 30 segundos se comete un delito, referido a salto, robo a personas, a casas, de vehículos, de accesorios de autos, tentativa de robo y otros no especificados; obteniendo este de la muestra obtenida por INEGI, en estudio particular.

La media aritmética o media de un conjunto de "n" números X_1, X_2, X_3,…, X_n se representa por y se calcula de acuerdo:

$$\text{Media} = \frac{\text{Suma de todos los valores}}{\text{Número de valores}}$$

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{n}$$

$$\bar{X} = \frac{\Sigma X}{n} \qquad \text{Para datos no agrupados}$$

$$\bar{X} = \frac{\Sigma fX}{\Sigma f} = \frac{\Sigma fX}{n} \qquad \text{Para datos agrupados}$$

Ejemplos:

1.- Las calificaciones de un estudiante en seis pruebas fueron 84, 91, 72, 68, 87 y 78.

Hallar la media aritmética de las calificaciones

$$\overline{X} = \frac{84+91+72+68+87+78}{6} = \frac{480}{6} = 80$$

2.- Los salarios anuales de cuatro hombres fueron $5,000; $6,000; $6,500 y $30,000.

a) Hallar la media aritmética de sus salarios.
b) Se diría que este promedio es representativo de los salarios?

$$\overline{X} = \frac{5000+6000+6500+30000}{4} : \frac{47500}{4} = \$11,875$$

3.- Los kilómetros recorridos por 5 estudiantes para ir a la escuela desde su casa son:

Estudiante	Kilómetros recorridos (Variable x)
1	1
2	4
3	10
4	8
5	10
5	33

Calcular la media aritmética de los kilómetros recorridos por los cinco estudiantes.

$$\overline{X} = \frac{1+4+10+8+10}{5} = \frac{33}{5} = 6.6 \text{ Kms.}$$

4.- Los tiempos de reacción de un individuo a determinados estímulos fueron 0.53, 0.46, 0.50, 0.49, 0.52, 0.53, 0.44 y 0.55 seguidos, respectivamente. Determinar el tiempo medio de reacción del individuo a los estímulos.

$$\overline{X} = \frac{0.53+0.46+0.50+0.49+0.52+0.53+0.44+0.55}{8} = \frac{4.02}{8} = 0.5 \text{ segundos}$$

5.- De un total de 100 números, 20 eran 4, 40 eran 5, 30 eran 6 y el resto eran 7.

Hallar la media aritmética de los números.

$$\overline{X} = \frac{\Sigma fX}{\Sigma f} = \frac{\Sigma fX}{n} = \frac{(20)(4)+(40)(5)+(30)(6)+(10)(7)}{100} = \frac{530}{100} = 5.3$$

6.- Cuatro grupos de estudiantes, formados por 15, 20, 10 y 18 individuos registran una media de pesos de 73, 67, 69 y 64 Kilogramos, respectivamente.

Hallar el peso medio de todos los estudiantes.

$$\overline{X} = \frac{\Sigma fX}{\Sigma f} \qquad \frac{(15)(73)+(20)(67)+(10)(69)+(18)(64)}{100} \qquad \frac{4277}{63} \qquad = 67.89 \text{ Kgs.}$$

7.- Se tienen la distribución de frecuencias de las alturas de 100 estudiantes. Hallar la altura media de los 100 estudiantes.

Altura cms. (Clase)	Frecuencias (f)	Marca de Clase (X)	(f)(X)
150-54	5	152	760
55-59	18	157	2826
60-64	42	162	6804
65-69	27	167	4509
70-74	8	162	1296
	$\Sigma f=100=n$		$\Sigma fX=16195$

$$\overline{X} = \frac{\Sigma fX}{\Sigma f} = \frac{\Sigma fX}{n} = \frac{16195}{100} = 161.95 \text{ cms.}$$

8.- Las millas recorridas por 20 estudiantes para ir a la escuela desde sus casas son:

0.8	1.2	2.6	2.8	3.3
3.4	3.7	4.0	4.5	5.3
5.8	6.1	6.2	6.5	7.1
7.3	7.4	7.6	7.8	9.2
				Σ = 102.6 millas

a) La media aritmética para los datos no-agrupados será:

$$\overline{X} = \frac{102.6}{20} = 5.13$$

b) Clasificando los datos en cinco clases

Millas recorridas (Intervalo de clase) Clase	Millas promedio (punto medio) X	No. De estudiantes (frecuencia de clase) f	Total de millas recorridas fX
0 y menos de 2	1	2	2
2 y menos de 4	3	5	15
4 y menos de 6	5	4	20
6 y menos de 8	7	8	56
8 y menos de 10	9	1	9
		$\Sigma f=n=20$	$\Sigma(fX)=102$

$$\overline{X} = \frac{\Sigma fX}{\Sigma f} = \frac{\Sigma fX}{n} : \frac{102}{20} : 5.1 \text{ millas}$$

Nota: La diferencia de las dos medias (0.03) es la información que se pierde, cuando se agrupan los datos.

MEDIANA

La mediana de un conjunto de valores es el valor del elemento central del conjunto.

La mediana de una colección de datos ordenados en orden de magnitud, es el valor medios o la media aritmética de los dos valores medios.

Ejemplos:

1.- Sean los números 3, 4, 4, 5, 6, 8, 8, 8, 10

Su mediana será: 6 Md=6

2.- Sean los números 5, 5, 7, 9, 11, 12, 15, 18

Su mediana será: $\frac{9+11}{2} = 10$; Md = 10

Nota: Si el número de valores en un conjunto de datos no agrupados es PAR, no hay mediana verdadera. El valor de la mediana se supone, por lo tanto, que es igual a la mitad entre los dos elementos centrados en el arreglo.

3.- Los salarios percibidos por un grupo de 25 empleados, en un periodo dado, están dados en la primera y segunda columnas de la tabla.
Calcular la mediana.

Salarios (Intervalo de clase)	Número de empleados (frecuencia)	Frecuencia acumulada
1-3	1	1
4-6	4	4+1=5
7-9	9	5+9=14 Clase mediana
10-12	6	14+6=20
13-15	2	20+2=22
16-18	3	22+3=25
	n=25	

$$Md = Li + \left(\frac{\frac{n}{2} - C}{f_{Md}} \right) i_{Md}$$

Donde:

Md = Mediana.
Li = Límite inferior verdadero de la clase mediana.
n = Número total de datos (Σf).
c= frecuencia acumulada hasta la clase anterior hasta la clase mediana
fmd = Frecuencia de la clase mediana.
imd = Amplitud o tamaño de la clase mediana.

En nuestro ejemplo:

La mediana es:

$$\frac{n}{2} = \frac{25}{2} = 12.5$$

La mediana es el elemento 12.5 mismo que se ubica en la tercera clase de la distribución; por lo tanto la clase mediana es la perteneciente 7-9.

Li = Límite inferior verdadero de la clase mediana es:

$$Li = \frac{6+7}{2} \qquad 6.5$$

=

$\Sigma f = n = 50$ datos

C; Frecuencia acumulada hasta la clase anterior a la clase Md es 5.
fmd; Frecuencia de la clase mediana = 9
imd; Amplitud o intervalo de la clase mediana es 9.5 - 6.5 = 3.0

$$Md = 6.5 + \left[\frac{\frac{25}{2} - 5}{9} \right] (3)$$

$$Md = 6.5 + \left[\frac{12.5 - 5}{9} \right] (3)$$

$$Md = 6.5 + \left[\frac{7.5}{9} \right] (3) = 6.5 + \frac{7.5}{3} = 9: Md =$$

4.- Hallar la mediana de los pesos de 40 estudiantes

Clase Pesos (lbs)	F Frecuencias	Fa Frecuencia acumulada
118-126	3	3
127-135	5	8
136-144	9	17
145-153	12	29 Clase mediana
154-162	5	34
163-171	4	38
172-180	2	40
	Σ40	

$$Md = Li + \left[\frac{\frac{N}{2} - C}{f_{Md}} \right] i_{Md}$$

$$Md = 144.5 + \left[\frac{\frac{40}{2} - 17}{12} \right] 9$$

$$Md = 144.5 + \left[\frac{20-17}{12} \right] 9$$

$$Md = 144.5 + \left[\frac{3}{12} \right] 9$$

$$Md = 144.5 + \left[\frac{1}{4} \right] (9) = 144.5 + \frac{2.2}{5} = Md =$$

5.- Los tiempos de reacción de un individuo a determinados estímulos fueron 0.53, 0.46, 0.50, 0.49, 0.52, 0.53, 0.44 y 0.55 segundos, respectivamente. Determine la mediana del tiempo de reacción del individuo a los estímulos.

Número del momento o caso (ordinal)	Valor del Elemento	
1	0.44	
2	0.46	
3	0.49	
4	0.50	Posición de la Mediana
5	0.52	
6	0.53	
7	0.53	
8	0.55	

6.- Una serie de números esté formada por 6, 7, 8, 9 y 10. Cuál es su mediana?

Secuencia	Valores	
1	6	
2	7	
3	8	Posición de la Mediana
4	9	Md=8
5	10	

7.- Según datos del CONACYT los egresados de los programas de Posgrado a nivel especialidad en Ciencias Sociales y Humanidades desde 1984 hasta 1990, han sido.
975, 872, 1163, 940, 1012, 1378, 717, respectivamente.

Determine la media y la mediana.

$$\bar{X} = \frac{\Sigma X}{\Sigma f} = \frac{\Sigma X}{n} = \frac{7057}{7} = 1{,}008 \text{ egresados}$$

X
717
872
940
975
1012
1163
1378
Σ7057

$\bar{X}$= 1,008 egresados

Posición de la mediana

Md = 975 egresados

8. Según datos de SECOFI el número de patentes concedidas en México a Mexicanos y ciudadanos de E.U.A., desde 1980 a 1991, han sido:

Mexicanos 165; 188; 197; 162; 138; 100; 41; 67; 256; 194; 132; 129.
E.U.A. 1140; 1225; 1524; 1222; 981; 646; 605; 625: 1697; 1237; 1237; 957; 801.

Determine medidas de tendencia central.

Mexicanos	E.U.A.
41	605
67	625
100	646
129	801
132	957
138	981
162	1,140
165	1,222
188	1,225
194	1,237
197	1,524
256	1,697
Σ 1,769	Σ 12,660

Mexicanos

$$\bar{X} = \frac{1769}{12} = 147.42 \text{ patentes}$$

$$Md = \frac{138+162}{2} = 150 \text{ patentes}$$

E.U.A.

$$\bar{X} = \frac{12{,}660}{12} = 1{,}055 \text{ patentes}$$

$$Md = \frac{981+1140}{2} = 1{,}060.5 \text{ patentes}$$

CAPITULO VI
MEDIDAS DE DISPERSION
MEDIDAS DE RELACION
REGRESION
CORRELACION

Probabilidad de aparición de un suceso (ocurrencia).
Por ejemplo:

a) Probabilidad de que aparezca sol en un "volado"

$$p = P\{s\} = \frac{1}{2} \begin{matrix} \longrightarrow \text{Un sol} \\ \longrightarrow \text{Dos caras (águila y sol)} \end{matrix}$$

b) Probabilidad de que al tirar un dado, la cara con el número 6 quede hacia arriba.

$$p = P\{6\} = \frac{1}{6} \begin{matrix} \longrightarrow \text{Una cara con 6} \\ \longrightarrow \text{Seis caras del dado} \end{matrix}$$

Probabilidad de la no aparición

$$q = P\{\text{no } E\} = \frac{N - h}{n} = 1 - \frac{h}{n} = 1 - p = 1 - P\{E\}$$

El suceso {no E} $\Longleftrightarrow$ Ē, Ẽ, ˜E.

Valores de la probabilidad

———————— 0 ——————————————— 1

Suceso			
	Imposible	**Probables**	Cierto
	Imposibilidad		Certeza

PROBABILIDAD EMPÍRICA

La probabilidad estimada o empírica de un suceso se toma como la frecuencia relativa de la aparición del suceso, cuando el número de observaciones es muy grande.

Dados

Evento	Frecuencia (aparición real)	Frecuencia relativa	
2	1	1/66	1/36
3	3	3/66	2/36
4	6	6/66	3/36
5	7	7/66	4/36
6	8	8/66	5/36
7	12	12/66	6/36
8	9	9/66	5/36
9	7	7/66	4/36
10	5	5/66	3/36
11	4	4/66	2/36
12	2	2/66	1/36
	66		

La probabilidad es el límite de la frecuencia relativa.

ANÁLISIS DE REGRESIÓN Y CORRELACIÓN

Luis Alfredo Valdés Hernández

Ningún procedimiento estadístico ha abierto tantos caminos en las Ciencias Sociales, como el de la *correlación,* ya que una gran cantidad del conocimiento que desarrollaremos se basa en averiguar cuáles elementos están relacionados y cuáles no.

Luego entonces, es necesario conocer el comportamiento que siguen esos elementos (también llamados variables) para poder medir la relación que guardan entre sí y como se verá esto no es más que la tendencia que manifiestan los fenómenos bajo observación. Esto condiciona a que se desarrolle un modelo en forma de ecuación matemática (*regresión)*, que describa en forma aproximada dicho comportamiento por la relación existente y una vez determinada la ecuación, se tendrá que analizar si guardan una relación estrecha o independiente (*correlación*), a través de medir la dispersión que presentan los valores de cada fenómeno.

Los procedimientos nos servirán para poder analizar los fenómenos y podemos pronunciar sobre ellos en forma precisa. Sin embargo, suele ocurrir que se caiga en error por emplearlos sin precaución, como por ejemplo en las situaciones de

a) Se les emplee de una manera inadecuada por una mala selección en el procedimiento
b) Se aplican a fenómenos sin ninguna relación razonable

c) Se les de una interpretación abusiva limitándose a ciertos procedimientos, sin utilizar otros que pueden arrojar resultados diferentes
d) Se puede llegar a conclusiones absurdas.

Los problemas anteriores son determinantes en nuestros estudios ya que podemos llegar a considerar relaciones, donde realmente hay poca o nula aproximación, en nuestro afán de representar la realidad mediante una relación matemática,

Relación entre variables

Las relaciones entre dos o más fenómenos pueden ser descritas como la vinculación que guarda una serie de valores que adquiere un fenómeno, con respecto a otra serie de valores que toma el otro fenómeno; por ejemplo, los alumnos de secundaria que tienen una comprensión verbal (tabla 1) pueden tener valores del razonamiento, relacionados con los valores de la comprensión verbal. Por lo que, si consideramos al conjunto total de los 10 alumnos, tendremos un espacio dado por sus valores, que nos indicará la posible relación, tan solo por la posición que ocupan en dicho espacio.
Regresando a nuestra suposición original, de que podemos representar por una ecuación matemática a ese segmento de la realidad, que está acotado por la relación existente entre los dos fenómenos, y a los que les asignaremos de manera arbitraria los papeles de las *variables dependiente* e *independiente* a la *comprensión verbal* y al *razonamiento* respectivamente.
Es decir, que si consideramos a la *comprensión verbal* como la *variable dependiente (Y)* y al *razonamiento* como la *variable independiente* (X) en nuestro modelo matemático llamado ecuación, esto quiere decir que los valores que tome la Y *(comprensión verbal),* dependerán de los valores de la X *(razonamiento).*

Diagrama de Dispersión

Si graficamos los valores de X y Y, podremos observar una serie de puntos que no son representaciones exactas de rectas o cualquier tipo de curva sino más bien tienden a ser de manera aproximada algún tipo de ellas y a esta representación gráfica le llamaremos diagrama de dispersión (Gráficas 1 y 2).
El coeficiente de correlación es un numero que nos dice hasta donde dos fenómenos están relacionados dicho de otra manera hasta donde las variaciones en los valores de uno de ellos acompañan a las variaciones en los valores del otro.
Este coeficiente puede tomar valores desde +1.00 pasando por el 0, hasta el -1.00; que nos indican una correlación positiva perfecta, independencia total o sea ninguna correlación, hasta la correlación negativa perfecta respectivamente.

Regresión

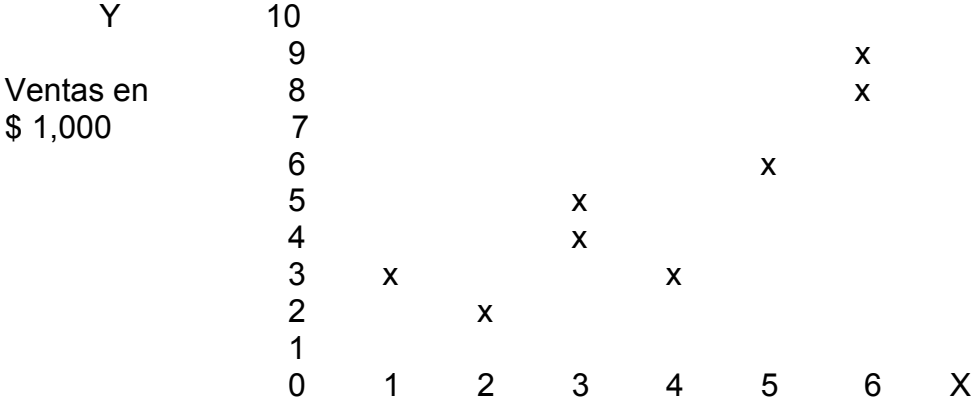
Los datos de dos o más fenómenos a través de una ecuación matemática, que suele llamársele modelo matemático de regresión o de relación entre los

fenómenos. El tipo de la curva de regresión dependerá de la tendencia que muestren los datos en el diagrama de dispersión.
Actualmente, el término regresión es sinónimo de estimación; el vocablo se empleó originalmente cuando al estudiar la relación entre la estatura de los padres y de sus hijos se encontró que los hijos de las personas más altas del grupo tendían a tener estaturas en promedio inferior a la de los padres, y los hijos de padres con estaturas menores al promedio tendían a tener estaturas superiores, esto es, se observó una tendencia de regresión en torno al promedio.
Cuando la ecuación de regresión da la relación entre dos variables, donde una será la variable independiente y la otra la variable dependiente, nos encontramos con el caso de regresión simple.
El fenómeno más común es el de que los datos describan en el diagrama de dispersión un comportamiento rectilíneo.

1.- Las primeras tres columnas de la tabla, muestran las cantidades de ventas (Y) hechas por un grupo de 8 vendedores en una compañía durante un período dado y los años de experiencia en ventas (X) de cada vendedor.
1.a) Elabore un diagrama de dispersión
1.b) Calcular la ecuación de la regresión lineal mediante mínimos cuadrados
1.c) Dibujar en la gráfica la línea de regresión
1.d) Estimar la cantidad de ventas para un vendedor que tiene 4 años de experiencia

Vendedor	Cantidad de Ventas (en 1,000)	Años de Experiencia en Ventas			
	Y	**X**	**XY**	**X^2**	**Y^2**
A	9	6	54	36	81
B	6	5	30	25	36
C	4	3	12	9	16
D	3	1	3	1	9
E	3	4	12	16	9
F	5	3	15	9	25
G	8	6	48	36	64
H	2	2	4	4	4
Suma	40	30	178	136	244

1 a) Diagrama de dispersión



I.b) $Y_C = a + bX$ Ecuación de una recta

$$b = \frac{n \sum (XY) - \sum X * \sum Y}{n \sum X^2 - (\sum X)^2} = \frac{8(178) - 30(40)}{8(136) - 30^2} = \frac{224}{118} = 1.19$$

$$a = \frac{\sum X^2 * \sum Y - \sum X * \sum (XY)}{n \sum X^2 - (\sum X)^2} = \frac{136(40) - 30(178)}{8(136) - 30^2} = \frac{100}{188} = 0.53$$

Otra forma de calcular "a" es:

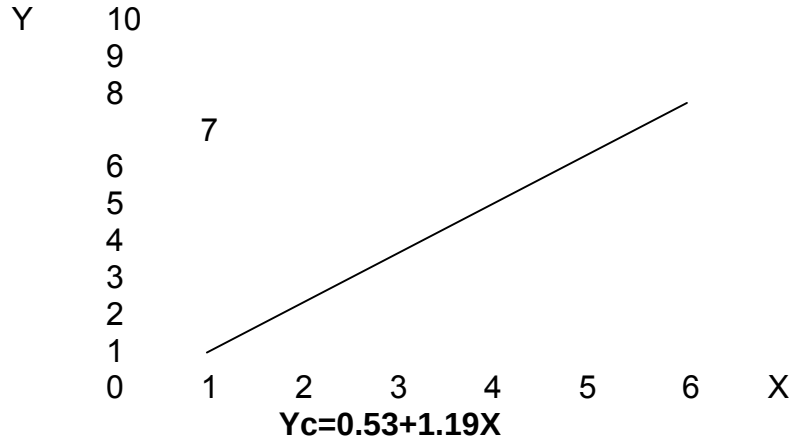
$$a = \frac{\Sigma Y}{n} - \left(b \quad \frac{\Sigma X,}{n} \right) = \frac{40}{8} - \left(\frac{56}{47} * \frac{30}{8} \right) = \frac{25}{47} = 0.53$$

$$a = \bar{Y} - b\bar{X}$$

1.c)

X	Yc = 0.53 + 1.19
1	0.53 + (1.19) (1) = 1.72
4	0.53 + (1.19) (4) = 5.29
6	0.53 + (1.19) (6) = 7.67

1. d)



l.c)

X	Yc = 0.53+1.19X	
1	0.53+(1.19)(1)	= 1.72
4	0.53+(1.19)(4)	= 5.29
6	0.53+(1.19)(6)	= 7.67

2.- Calcular la desviación estándar de regresión de los valores de Y para los datos dados en

SYX =~ X(Y-YcY 1 n

a

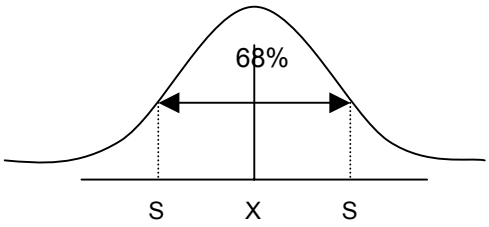
SYX =XY^2-aTY-bYXfl

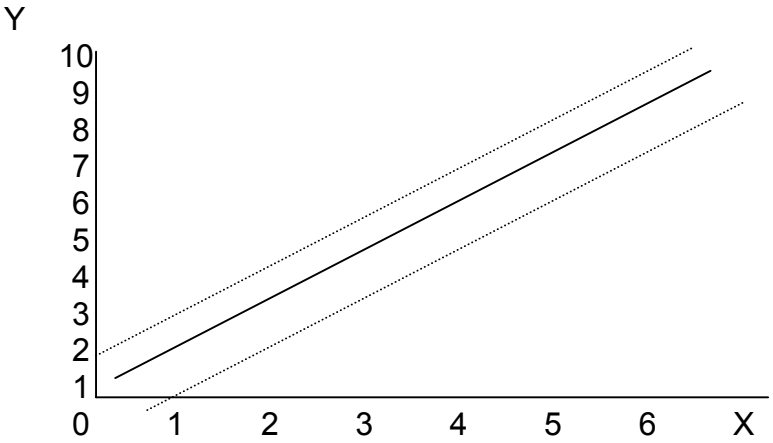
X	Yc = 0.53 + 1.19X	Y - Yc	$(Y - Yc)^2$
6	Yc =0.53 + (1.19) (6) = 7.67	1.33	1.77
5	Yc =0.53 + (1.19) (5) = 6.48	-0.48	0.23
3	Yc =0.53 + (1.19) (3) = 4.10	-.010	0.01
1	Yc =0.53 + (1.19) (1) = 1.72	1.28	1.64
4	Yc =0.53 + (1.19) (4) = 5.29	-2.29	5.24
3	Yc =0.53 + (1.19) (3) = 4.10	0.90	0.81
6	Yc =0.53 + (1.19) (6) = 7.67	0.33	0.11
2	Yc =0.53 + (1.19) (2) = 2.91	-0.91	0.83
30	39.94	0.06	10.64

Sxy = $\sqrt{\frac{10.64}{8}}$ = $\sqrt{1.33}$ = 1.15
|

$$Sxy = \sqrt{\frac{244 - 25/47(40) - 56/47(178)}{8}} = \sqrt{1.33} = 1.15$$

La interpretación de SYX con respecto a Yc (línea) es similar a la de Sy con respecto a calcular la desviación estándar de regresión de los valores de Y , para los datos dados
la media.

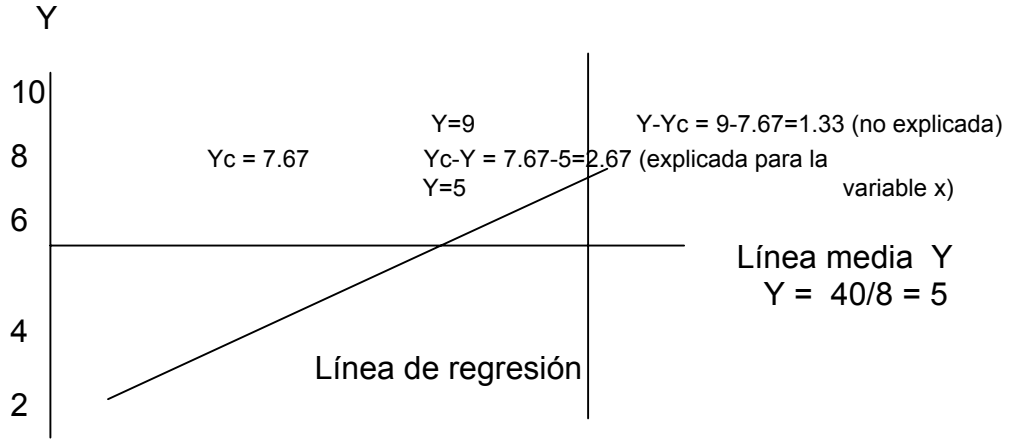


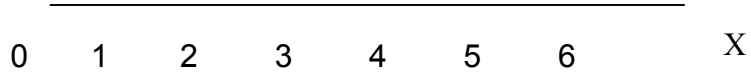


5/8 = dentro = 62.5 % (real)
3/8 = fuera = 68.0 % (teórica)

3.- Calcular el coeficiente de determinación r² y el coeficiente de correlación r

Coeficiente de determinación = Variación explicada
Variación total
(variación explicada + variación no explicada)





$$r^2 = \frac{X(Yc-Y)^2}{X(Y-Y)^2}$$

Vendedor	Y	Y-Y	$(Y-Y)^2$	Yc-Y	$(Yc-Y)^2$
A	9	4	16	7.67-5=2.67	7.013
B	6	1	1	6.48-5=1.48	2.09
C	4	-1	1	4.10-5=0.90	0.81
D	3	-2	4	1.72-5=-3.28	10.76
E	3	-2	4	5.29-5=0.29	0.08
F	5	0	0	4.10-5=-1.90	0.81
G	8	3	9	7.67-5=2.67	7.13
H	2	-3	9	2.91-5=-2.09	4.37
Σ	40	0	44		33.28

Y= 40/8 = 5

$$r^2 = \frac{33.36}{44} = 0.7582$$

Cuando r^2 es cercano a 1 los valores de Y están muy cercanos a la línea de regresión. Por lo tanto, la variación total de los valores de Y es más explicada por la línea y la variable Y está estrechamente relacionada a la variable X.
Redondeando: r^2 =0.7582 = 75.82 ~76%

76 % de la variación en las cantidades de ventas (Y) está linealmente relacionada con la variación en los años de experiencia en ventas (X) del vendedor.

$r = \sqrt{0.7582} = 0.87$ Correlación positiva.

DATOS AGRUPADOS

La exposición de las selecciones previas trató con datos no agrupados. Cuando las técnicas usadas en análisis de regresión y correlación son extendidas a datos agrupados, deberá tenerse cuidado del factor frecuencia *f* en cada intervalo de clase de ambas variables X e Y. Las formulas para

calcular las diferentes medidas a partir de datos agrupados, arreglados en orden para conveniencia de los cálculos están dadas en seguida.

Sea:

$$① = n \Sigma f d_x f_y - (\Sigma f_x d_x)(\Sigma f_y d_y),$$
$$② = n \Sigma f d_x d_x^2 - (\Sigma f_x d_x)^2,$$
$$③ = n \Sigma f d_y d_y^2 - (\Sigma f_y d_y)^2,$$

donde f = la frecuencia de las clases conjuntas de las variables X e Y,

f_y = la frecuencia en una clase de la variable X,

f_y = la frecuencia en una clase de la variable Y,

n = la frecuencia total $= \Sigma f = \Sigma f_x = \Sigma f_y$,

d_x = la desviación del punto medio de una clase de la variable X con respecto a la media supuesta en unidades de intervalo de clase,

d_y = la desviación en unidades de intervalo de clase para la variable Y.

La formula para la constante *b* de la ecuación de regresión, es ahora escrita:

$$b = \frac{1}{2} \cdot \frac{i_y}{i_x}$$

donde i_y e i_x representan el tamaño del intervalo de clase de la variable Y y X respectivamente.

La formula para la constante *a* de la ecuación de regresión, es ahora escrita:

$$a = \bar{Y} - b\bar{X},$$

donde, basados en la formula:

$$Y = \bar{A}_y + \frac{(\Sigma f_x d_x) i_y}{n},$$

y

$$X = \bar{A}_x + \frac{(\Sigma f_x d_x) i_x}{n},$$

A es la media supuesta.

La fórmula de producto-momento para el coeficiente de correlación, fórmula es ahora escrita:

$$r = \frac{1}{\sqrt{2.3}}$$

La fórmula par la desviación estándar de regresión, fórmula es ahora escrita:

$$S_{yx} = i_y \frac{3 \ (1-\sqrt{r^2})}{n}$$

Cuando se aplican las fórmulas de arriba para datos agrupados, usualmente se prepara una *tabla de correlación* para facilitar los cálculos. Una tabla de correlación es también llamada una *tabla de frecuencia bivariante, bidimensional, o de doble entrada.* Muestra la distribución de frecuencias de dos variables.

Ejemplo:Las marcas en la tabla muestran las cantidades de vetas (Y) hechas por un grupo de 40 vendedores durante un período dado y los años de experiencia en ventas (X) correspondientes a los vendedores individuales. Las marcas en la última celda de la primera columna, por ejemplo, indican que dos vendedores con “0 y o menos de 2” años de experiencia en ventas, vendieron de $500 a $3,500 (los límites reales 0.5 y 3.5 de la clase “1-3” en unidades de $1,000).

TABLA DE CLASIFICACIÓN CRUZADA
X= Años de experiencia en ventas

Y cantidades de ventas $1000

Intervalo de clase	0 y menos de 2	2 y menos de 4	4 y menos de 6	6 y menos de 8	8 y menos de 10	Número total de vendedores
16-18					///	1
13-15			///	///	///	5
10-12			////	//		7
7-9		////	////	////		21
4-6		///		/		4
1-3	//					2
Número total de vendedores, f_x	2	8	16	10	4	40

La fórmula de la desviación estándar, formula, puede ser escrita:

$$s_y = i_y \sqrt{\frac{\sum f_y d_y^2}{n} - \frac{(\sum f_y d_y)^2}{n}} = i_y \sqrt{\frac{n.\sum f_y d_y^2 - (\sum f_y d_y)^2}{n}} = i_y \sqrt{\frac{3}{n}}$$

$$s_{yx} = s_y \sqrt{1 - r^2} = i_y \sqrt{\frac{3}{n}} \sqrt{1 - r^2} = i_y \sqrt{\frac{3 \quad (1-r^{2)}}{n}}$$

Solución: 1. La tabla de correlación es dispuesta de acuerdo con los datos en la tabla cruzada. Se cuenta el número de marcas en cada celda y se pone en la celda correspondiente de la tabla de correlación en la segunda línea, denotado por el símbolo *f*.

TABLA DE CORRELACIÓN

x- Años de experiencia en ventas (i_x = 2 años)

Y Cantidad de ventas en $1000

Intervalo de clase	0 @*2	2 @ 4	4 @ 6 A_x=5	6 @ 8	8 @ 10	f_y	d_y	$f_y d_y$	$f_y d_y 2$	$fd_x d_y$
16-18					2.3 = 6 x1 6	1	3	3	9	6
13-15			0.2=0 x1 0	1.2= 2 x1 2	~~2.2 = 4~~ x3 12	~~5~~	2	10	20	14
10-12			0.1= 0 x5 0	1.1= 1 x2 2		7	1	7	7	2
7-9 A_y=8		-1.0 = 0 x5 0	0.0 = 0 x10 0	1.0 = 0 x6 0		21	0	0	0	0
4-6		-1.-1= 1 x3 3		1.-1= -1 x1 -1		4	-1	-4	4	2
1-3	-2.-2=4 x2=*f* 8					2	-2	-4	8	8
f_x	2	8	16	10	4	40	-	12	48	32
d_x	-2	-1	0	1	2	-				
$f_x d_x$	-4	-8	0	10	8	6				
$f_x d_x^2$	8	8	0	10	16	42				

*@ representa " y menos de "
$\sum f_x d_x = 6$, $\sum f_x dx^2 = 42$, n= $\sum f_x = \sum f_y = \sum f = 40$,

$\sum f_y d_y = 12$, $\sum f_y dy^2 = 48$, $\sum f_x d_y = 32$.

2. Las medidas supuestas son seleccionadas como sigue:

A_x, medida supuesta para la variable X= 5 años, el puento medio de la clase "4 y menos de 6"

A_y, medida supuesta para la variable Y= $8,000, el punto medio de la clase "7-9"

Los números en la hilera d_x y en la columna d_y son, por lo tanto, determinados - +1, +2, y así sucesivamente para representar los números en unidades de

intervalo de clase de las clases individuales arriba (o mayores que) la clase de la media supuesta (donde d_x o d_y = 0) ; -1, -2, y así sucesivamente para representar los números abajo (o menores que) la clase correspondiente a la medida supuesta.

Los valores de fd_xd_y se calculan para las celdas individuales. El producto de d_x y d_y es puesto en la primera línea de cada celda. El producto de f, d_x y d_y está en la tercera línea. Por ejemplo, los números en la última celda de la primera columna representan:

1ª línea: -2 (d_x en la clase X "0 @ 2"). –2 (d_y en la clase Y "1-3")= 4, o d_x. d_y = 4

2ª línea: 2= la frecuencia conjunta de la clase X "0 @ 2" y la clase Y "1-3", o f =2.

3ª línea: 8= 4 x 2, o fd_xd_y= 8. La suma de los productos de cada hilera aparece en la última columna de la derecha con el encabezado fd_xd_y. Por lo tanto, el segundo número en la columna fd_xd_y, o 14= 0 + 2+12.

Nótese que le valor de fd_xd_y en cada celda en la hilera o en la columna de la clase correspondiente a la medida supuesta, es siempre igual a cero, puesto que d_x y d_y = 0 en las clases. El cálculo para los valores en la hilera o columna no es, por lo tanto, necesario, excepto para propósitos ilustrativos.
Calcular y sumar los valores requeridos como se muestra en la tabla de correlación son resumidos y colocados directamente debajo de la tabla. Los valores resumidos se sustituyen ahora en las fórmulas, como sigue:

$$1 = n\sum fd_xd_y - (\sum f_xd_x)(\sum f_yd_y) = 40(32) - 6(12) = 1{,}208$$

$$2 = n\sum f_xd_x^2 - (\sum f_xd_x)^2 = 40\,(42) - 6^2 = 1{,}664,$$

$$3 = n\sum f_yd_y^2 - (\sum f_yd_y)^2 = 40\,(48) - 12^2 = 1{,}776.$$

a. La ecuación de regresión: usar la fórmula para b.

$$b = \frac{1}{2}\cdot\frac{i_y}{i_x} = \frac{1{,}208}{1{,}664}\cdot\frac{3}{2} = \frac{151}{137} = 1.1.$$

$$a = \bar{Y} - b\bar{X} = 8.9 - 1.1\,(5.3) = 3.1.$$

Los valores de $\bar{Y}$ y $\bar{X}$ se calculan usando la formula :

$$\bar{Y} = A_y + \left(\frac{\Sigma\, f_xd_x)i_y,}{n}\right. = 8 + \frac{(12)\,3}{40} = 8.9$$

$$\bar{X} = A_x + \left(\frac{\Sigma\, f_xd_x)i_x,}{n}\right. = 5 + \frac{(6)\,2}{40} = 5.3$$

La ecuación de regresión es:

$$Y_c = 3.1+1.1X.$$

Los siguientes dos puntos se calculan de la ecuación para dibujar la línea de regresión sobre la tabla de clasificación cruzada.

Cuando X = 0, limite inferior de la primera clase de X “0 @ 2”

$Y_c = 3.1 + 1.1\ (0) = 3.1;$

Cuando X = 10, el limite superior de la última clase de X “8 @ 10”,

$Y_c = 3.1 + 1.1\ (10) = 14.1$

b. El coeficiente de correlación:

$$r = \frac{1}{\sqrt{2.3}} = \frac{1208}{\sqrt{(1644)\ (1766)}} = \frac{1208}{\sqrt{2,919,744}} = \frac{1208}{1709}$$

$$= 0.71$$

$$r^2 = 0.71^2 = 0.5041 \text{ o } 50.41\%$$

Por lo tanto, 50.41% de la variación de las cantidades de ventas es explicada por la variación de los años de experiencia en ventas del grupo de 40 vendedores.

c. La desviación estándar de regresión:

$$S_{yx} = i_y \frac{3\sqrt{(1- r^2)}}{n} = \frac{3}{\sqrt{1776}}\sqrt{(1-0.5041)} = \frac{3(29.68)}{40} = 2.226 \text{ o } \$\ 2,226$$