UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

MAESTRÍA EN DOCENCIA PARA LA EDUCACIÓN MEDIA SUPERIOR

ESTADÍSTICA Y PROBABILIDAD I

Eleazar Gómez Lara

DATOS BIVARIADOS

MADEMS Maestría en Docencia para la Educación Media Superior MATEMÁTICAS

Estadística es un medio de comunicación científica que suministra un lenguaje claro y conciso.

Por medio de:

- **Una grafica**
- **Una tabla**
- **Una formula**
- **Un enunciado**

En el periódico La Jornada del mes de abril de 2004 apareció ésta tabla de información:

GANADOR 2000	GANADOR 2003 PRI	PAN	PRD	OTROS	TOTALES
PRI	40	11	12	6	69
PAN	16	12	1	1	30
PRD	8	1	10	2	21
OTROS	2	0	0	0	2
Nuevos Mpios.	1	0	1	0	2
TOTALES	67	24	9	24	122

a) ¿Podemos extraer información de ella?
b) ¿Qué tipo de información?
c) ¿Cómo podemos representarla?

¿Cómo la usamos?

- **La estadística es un medio de comunicación efectivo para la predicción, logrando esto a través de los modelos matemáticos o de la "matematización" de situaciones reales, los cuales; permiten explicar el comportamiento de estas situaciones y predecir con cierta aproximación, cuestiones desconocidas.**

Los siguientes datos son resultado de diversos censos de la población de México.

Año	Población Total
1950	25 71 017
1960	34 923 129
1970	48 225 238
1990	81 249 645
1995	91 158 290
2000	97 483 412

1. **¿Podemos encontrar alguna expresión matemática que responda a esos datos? (explicación).**
2. **¿Es posible estimar la población en los años en los que no se realizaron censos durante el periodo 1950 – 2000? (interpolación).**
3. **¿Es factible estimar la población del año 2005, y en algunos años futuros, a partir de estos datos? (predicción).**

DATOS BIVARIADOS

- **Consideremos que de un elemento tenemos dos características útiles para ciertos estudios, dichas características podrían ser analizadas cada una por separado, más sin embargo nuestro interés está centrado en analizarlas en forma conjuntas es decir cuando ellas interactúan sobre el elemento en consideración.**

VARIABLES CON POSIBLE RELACIÓN

- Distancia - Combustible
- Edad – Peso
- Estatura – Peso
- Escolaridad – Ingreso
- Calidad – Precio

Análisis del conjunto de datos

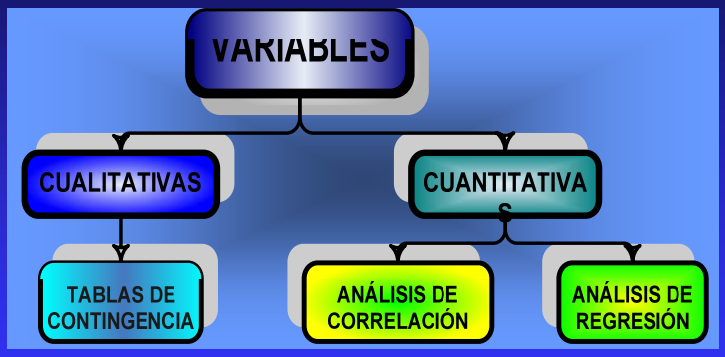
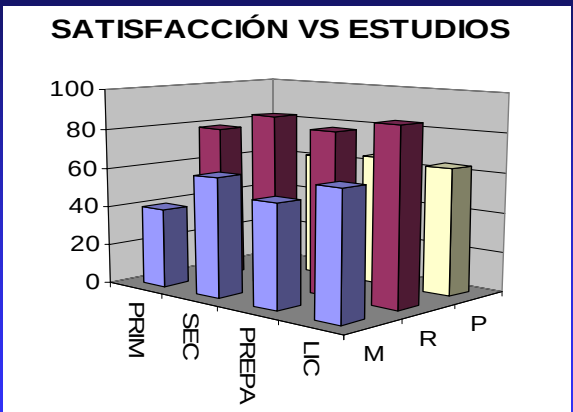


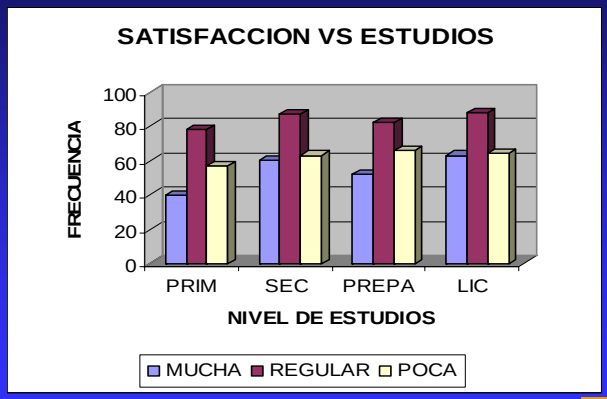
TABLA DE CONTINGENCIA

Satisfacción en el trabajo	Nivel de estudios: Primaria	Secundaria	Preparatoria	Licenciatura	Totales
Mucha	40	60	52	63	215
Regular	78	87	82	88	335
Poca	57	63	66	64	250
Totales	175	210	200	215	800

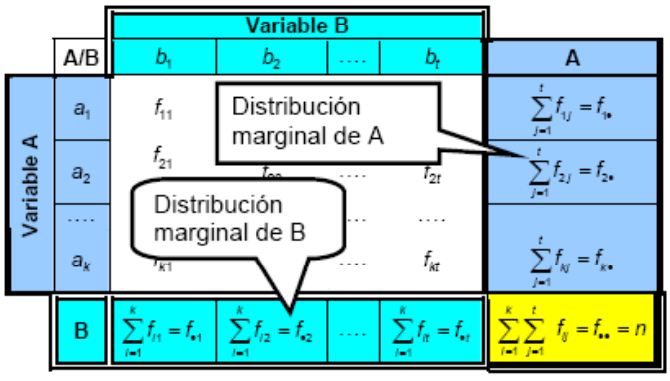
REPRESENTACION GRAFICA



REPRESENTACIÓN GRÁFICA



DISTRIBUCIONES MARGINALES



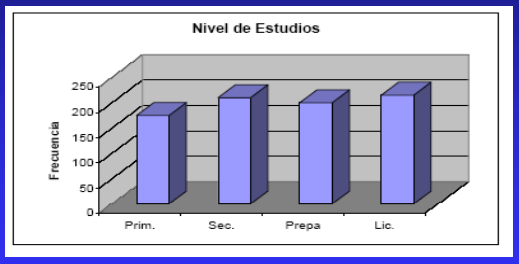
DISTRIBUCIÓN MARGINAL

- **SATISFACCION EN EL TRABAJO**

Satisfacción en el trabajo	Totales
Mucha	215
Regular	335
Poca	250
Totales	800

DISTRIBUCIONES MARGINALES

- **NIVEL DE ESTUDIOS**



	Nivel de estudios				Totales
	Primaria	Secundaria	Preparatoria	Licenciatura	
Totales	175	210	200	215	800

DISTRIBUCIONES CONDICIONADAS

- **En otras ocasiones estamos interesados en la distribución de una de las variables para un valor fijo de la otra, es decir tratamos de responder a la pregunta ¿Cómo se comporta la variable B cuando la variable A toma un valor fijo ? Esto es lo que se conoce como distribuciones condicionada.**
- **La distribución de atributo B condicionado a que A toma un valor $(A= a_j)$; es la distribución de B que se obtiene considerando sólo los elementos que tienen para el atributo A el valor a_j**

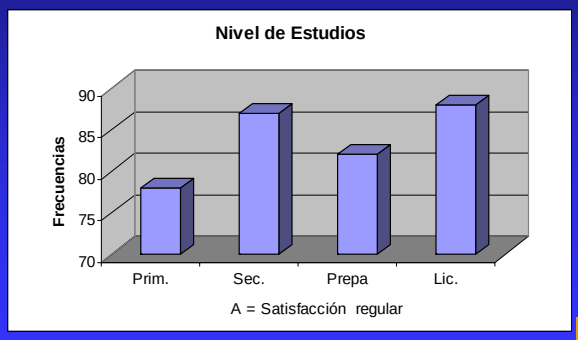
DISTRIBUCION CONDICIONADA

- **Distribución condicionada del nivel de estudios dada una satisfacción regular en el trabajo.**

		Nivel de estudios				
		Primaria	Secundaria	Preparatoria	Licenciatura	Totales
Satisfacción en el trabajo	Mucha	40	60	52	63	215
	Regular	78	87	82	88	335
	Poca	57	63	66	64	250
	Totales	175	210	200	215	800

REPRESENTACIÓN GRÁFICA

- **Distribución condicionada del nivel de estudios dada una satisfacción regular en el trabajo.**



DISTRIBUCIONES CONDICIONADAS

- **Ahora la pregunta es ¿Cómo se comporta la variable A cuando la variable B toma un valor fijo ?**
- **En forma análoga, la distribución del atributo A condicionado a que el atributo B toma un valor $(B = b_k)$; es la distribución de A que se obtiene considerado sólo los elementos que tienen para el atributo B el valor b_k.**

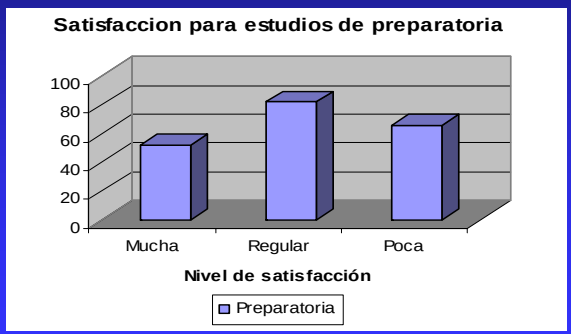
DISTRIBUCION CONDICIONADA

- **Distribución condicionada de la satisfacción en el trabajo dado el nivel de estudios de preparatoria.**

		Nivel de estudios				
		Primaria	Secundaria	Preparatoria	Licenciatura	Totales
Satisfacción en el trabajo	Mucha	40	60	52	63	215
	Regular	78	87	82	88	335
	Poca	57	63	66	64	250
	Totales	175	210	200	215	800

REPRESENTACIÓN GRÁFICA

- **Distribución condicionada de la satisfacción en el trabajo dado el nivel de estudios de preparatoria.**



EJERCICIO

- **ESTADO CIVIL Y NIVEL DE EMPLEO**
- **En algunas ocasiones se ha escuchado que el estar casado es bueno para una carrera dentro de nuestro empleo.**
- **La siguiente tabla presenta un estudio para revisar esta afirmación.**

ESTADO CIVIL Y NIVEL DE EMPLEO

		ESTADO CIVIL				
		SOLTERO	CASADO	DIVORCIADO	VIUDO	TOTAL
NIVEL DE EMPLEO	A	58	874	15	8	
	B	222	3927	70	20	
	C	50	2396	34	10	
	D	7	533	7	4	
	TOTAL					

ESTADO CIVIL Y NIVEL DE EMPLEO

- **Determina los distribuciones marginales de estado civil y nivel de empleo.**
- **Elaborar la grafica de barras de las distribuciones marginales para el estado civil y el nivel de empleo.**
- **¿Qué porcentaje de hombres tienen el nivel más bajo de empleo? ¿Qué porcentaje de hombres solteros tienen el empleo más bajo?**
- **¿Cuál es la distribución marginal en porcentaje de hombres solteros?**

ESTADO CIVIL Y NIVEL DE EMPLEO

- **Compara las distribuciones condicionales del nivel 1 de empleo con el nivel 4 de empleo. Describe brevemente las principales diferencias entre estos dos grupos de hombres apoyándote en los valores porcentuales . Muestra estas distribuciones mediante un diagrama de barras.**
- **Se desea publicar una revista dirigida a hombres casados. Determina la distribución condicional de los niveles de empleo entre los hombres casados, represéntala utilizando un diagrama de barras ¿A que nivel o niveles de empleo se debería de dirigir ésta revista?**

EJERCICIO

- Tenemos hambre pero no tenemos mucho tiempo para comer y decidimos pasar a una tienda para comprar “algo” para calmar nuestra hambre. ¿Qué tipo de “alimento” y “bebida” compraríamos?

		Bebidas				
	A/B	Agua	Refresco	Yogurt	Jugo	Totales
Alimento	Papas					
	Frituras					
	Pan					
	Galletas					
	Totales					

CORRELACIÓN Y REGRESIÓN LINEAL

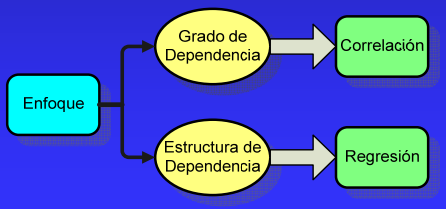
Consideremos ahora que nuestras variables son del tipo cuantitativo y queremos determinar

- **Si existe una relación entre las dos variables, y**
- **De existir, identificar qué tipo de relación es.**
- **Si existe tal relación, expresarla con una ecuación que permita predecir el valor de una de las variables si conocemos el valor de la otra.**

CORRELACIÓN Y REGRESIÓN LINEAL

El análisis de las relaciones existentes entre dos o más variables requiere del tratamiento estadístico cuando:

1. **La estructura verdadera de la relación se desconoce.**
2. **No existe una dependencia funcional exacta entre las variables consideradas.**



CORRELACIÓN Y REGRESIÓN LINEAL

Por lo que en el estudio de la asociación entre variables existen dos aspectos relacionados:

- **El análisis de correlación que tiene como objetivo determinar el grado de relación entre variables.**
- **El análisis de regresión que trata de establecer la "naturaleza de la relación" entre variables, es decir, propone la relación funcional entre las variables.**

POSIBLES RELACIONES

1. **A medida que una persona aumenta de estatura, se espera que gane peso, se podrá preguntar en este caso ¿Existe una relación entre la estatura y el peso?**
2. **Como estudiantes nos dedicamos a estudiar y a resolver exámenes, ¿Será cierto que cuanto más se estudie tanto mayor es la calificación obtenida?**
3. **Como profesores deseamos saber si el desempeño de los estudiantes en secundaria tiene un efecto en las calificaciones obtenidas en matemáticas, ¿habrá una relación entre el promedio de calificaciones de secundaria y el promedio de calificación en matemáticas?**

CORRELACIÓN Y REGRESIÓN LINEAL

En los casos anteriores queremos saber si existe una cierta variación conjunta entre las dos variables, y si es así determinar el grado de dependencia que existe entre ellas y por supuesto verla reflejada mediante una regla o ecuación.

El estudio de la relación entre dos variables inicia realizando dos mediciones a cada uno de varios objetos. Deseamos determinar cuál de estas variables medibles denominada Y, tiende a aumentar o disminuir mientras la otra variable, llamada X, varía.

CORRELACIÓN Y REGRESIÓN LINEAL

El primer paso en la determinación de sí existe o no una relación entre dos variables es examinar la gráfica de los datos observados, a la cual se le da el nombre de diagrama de dispersión. El diagrama de dispersión consiste en el trazo de todos los pares ordenados de datos bivariados sobre un sistema de ejes coordenados. La variable de entrada X, se utiliza para el eje horizontal, y la variable Y para el eje vertical.

DIAGRAMA DE DISPERSIÓN

Observemos el siguiente diagrama de dispersión ¿A qué conclusiones podemos llegar?

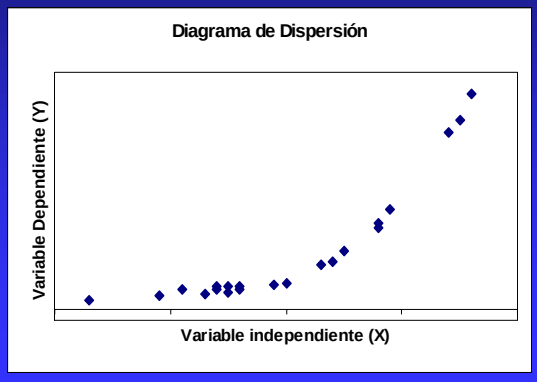


DIAGRAMA DE DISPERSIÓN

Revisemos el siguiente diagrama de dispersión, ¿Cuáles son ahora nuestras conclusiones?

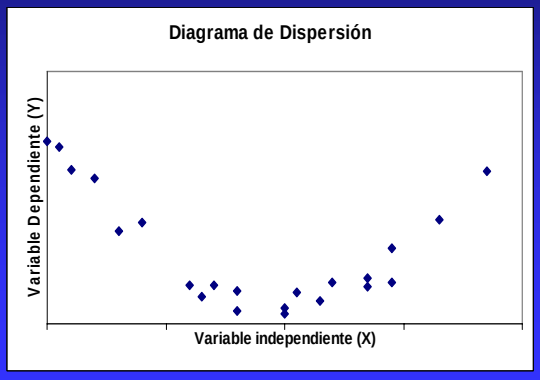


DIAGRAMA DE DISPERSIÓN

Para éste diagrama de dispersión, ¿Qué conclusiones podemos obtener?

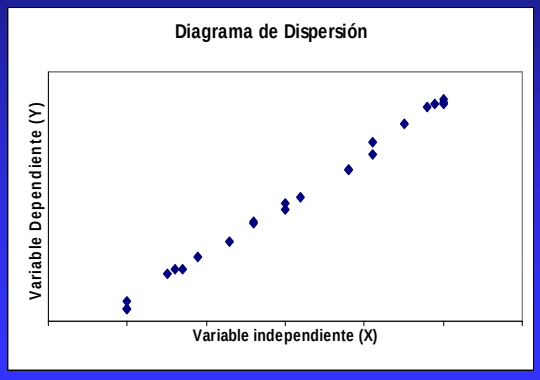


DIAGRAMA DE DISPERSIÓN

¿Qué conclusiones son apropiadas para el siguiente diagrama de dispersión?

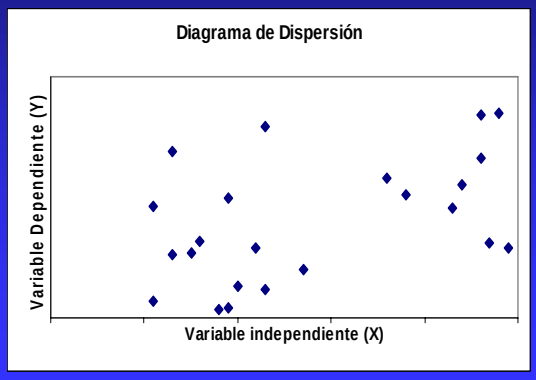


DIAGRAMA DE DISPERSIÓN

En resumen analizar el diagrama de dispersión nos es útil para:

- **Visualmente buscar patrones que nos indiquen que las variables están relacionadas.**
- **Y si esto sucede en él se esboza el tipo de curva (recta, parábola, etc.) que puede describir esta relación.**

CORRELACIÓN Y REGRESIÓN LINEAL

Un economista, está interesado en determinar si existe alguna relación entre el ingreso familiar (X) y el porcentaje de ingreso gastado en alimentación (Y).

-	X ($ 1000)	Y (%)
1	8	36
2	9	33
3	12	25
4	13	28
5	16	22
6	19	20
7	24	15
8	27	19

DIAGRAMA DE DISPERSIÓN

¿Cuáles son nuestras conclusiones?

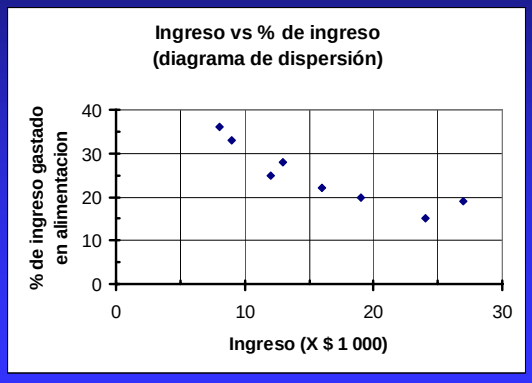


DIAGRAMA DE DISPERSIÓN

Debido a que las conclusiones que podemos sacar de los diagramas de dispersión tienden a ser subjetivas, se necesitan métodos precisos y objetivos para confirmar nuestras conclusiones alcanzadas al analizar el diagrama de dispersión.

Definición: Existe una correlación entre dos variables si una de ellas está relacionada o ligada con la otra de alguna manera.

CORRELACIÓN Y REGRESIÓN LINEAL

Definición: Se llama coeficiente de correlación a un índice numérico abstracto, que indica el grado de relación entre dos variables.

- **El coeficiente de correlación mide el grado al cual se relaciona en forma lineal dos variables entre sí.**
- **El más popular y utilizado de los coeficientes de correlación es el de Pearson, que para su aplicación es requisito indispensable que la correlación sea de tipo lineal.**

COEFICIENTE DE CORRELACIÓN

El coeficiente de correlación de Pearson se calcula mediante la expresión:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

COEFICIENTE DE CORRELACIÓN

El valor de *r* siempre debe quedar entre –1 y +1 inclusive.

- **Si r es cercano a 0, concluimos que no existe una correlación lineal significativa entre x, y,**
- **Si r está cerca de –1 o +1, concluimos que existe una correlación lineal significativa entre x, y.**
- ***Correlación positiva*, que ocurre cuando al crecer o decrecer una de las variables la otra crece o decrece paralelamente. (es decir, ambas tienen el mismo comportamiento, ambas crecen o ambas decrecen).**
- ***Correlación negativa* que ocurre cuando al crecer una de las variables, la otra decrece (y viceversa, es decir tienen un comportamiento inverso).**

COEFICIENTE DE CORRELACIÓN

- **De acuerdo al valor del coeficiente de correlación de Pearson, podemos describir el tipo de relación existente entre dos variables de acuerdo a la siguiente tabla:**

	CORRELACION					
Tipo de correlación	Negativa o inversa			Positiva o directa		
	Fuerte	Moderada	Débil	Débil	Moderada	Fuerte
Valor de R	-1 a -0.8	-0.8 a -0.5	-0.5 a 0	0 a 0.5	0.5 a 0.8	0.8 a 1

COEFICIENTE DE CORRELACIÓN

- **Para calcular el valor del coeficiente de correlación lineal r, realizamos las operaciones indicadas en la siguiente tabla.**

No.	x	y	xy	x^2	y^2
1	8	36	288	64	1296
2	9	33	297	81	1089
3	12	25	300	144	625
4	13	28	364	169	784
5	16	22	352	256	484
6	19	20	380	361	400
7	24	15	360	576	225
8	27	19	513	729	361
	128	198	2854	2380	5264
	$\sum x$	$\sum y$	$\sum xy$	$\sum x^2$	$\sum y^2$

COEFICIENTE DE CORRELACIÓN

Sustituyendo los valores obtenidos en la tabla en la formula para el coeficiente de correlación

$$r = \frac{(8)(2854) - (128)(198)}{\sqrt{\left[(8)(2380) - (128)^2\right]\left[(8)(5264) - (198)^2\right]}} = -0.9038$$

De acuerdo a este valor podemos concluir que existe una correlación lineal inversa (r es negativo) significativamente fuerte entre el ingreso y el porcentaje en gastos de alimentación, es decir a medida que el ingreso aumenta, el porcentaje de ingreso gastado en alimentación disminuye.

REGRESIÓN LINEAL

Ahora como ya sabemos que existe una correlación lineal significativa entre dos variables, queremos describir la relación encontrando la ecuación de la línea recta que la representa y posteriormente trazar la gráfica de la misma sobre el diagrama de dispersión.

Definición: Dada una colección de datos de muestras apareados, la ecuación de regresión

$$\hat{y} = m \cdot x + b$$

describe la relación entre las dos variables.

REGRESIÓN LINEAL

Los valores de b y m se calculan mediante las expresiones:

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} \qquad b = \frac{\sum y - m\sum x}{n}$$

Una vez evaluados m y b podemos identificar la ecuación de regresión estimada, que tiene la siguiente propiedad: *La línea de regresión es la que mejor ajusta a los puntos de la muestra.*

REGRESIÓN LINEAL

Calculemos ahora los coeficientes de la recta de regresión para nuestro ejemplo, no es necesario realizar más cálculos ya que los necesarios están en la tabla que utilizamos para calcular el coeficiente de correlación:

$$m = \frac{8(285) - (128)(198)}{8(2380) - (128)^2} = \frac{22832 - 25344}{19040 - 16384} = \frac{-2512}{2656} = -0.9457$$

$$b = \frac{198 - (-0.9457)(128)}{8} = \frac{198 + 121.0602}{8} = 39.8825$$

REGRESIÓN LINEAL

La recta de regresión tiene como ecuación:

$$\hat{y} = -0.9457 \cdot x + 39.8825$$

La pendiente se interpreta como: que por cada $1000 que aumenta el ingreso, el porcentaje de ingreso gastado en alimentación disminuye 0.94%.

La ordenada al origen es el valor que se esperaría tendría la variable dependiente cuando la variable independiente es cero.

En este caso su interpretación no tiene sentido práctico ya que nos indicaría que cuando el ingreso es cero, el porcentaje del mismo gastado en alimentación es del 39.88%.

REGRESIÓN LINEAL

Usando la ecuación de regresión, para cada valor de *x* calculamos los valores de estimación mediante la recta de regresión, para esto sustituimos cada valor del ingreso en la ecuación de regresión, realizando las operaciones indicadas.

Por ejemplo: Para *x = 8*

$\hat{y}$ **= -0.9457*(8) + 39.8825 = 32.3169**

Esto significa que para un ingreso de $ 8 000 el porcentaje de ingreso gastado en alimentación se estima en 32.31 %

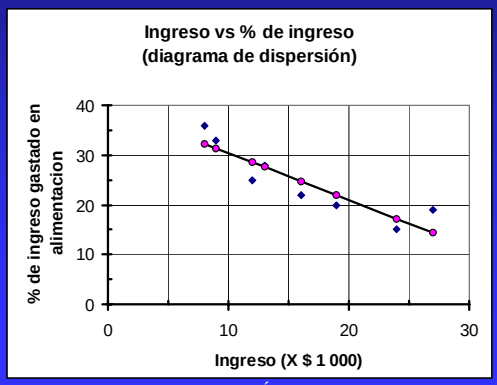
REGRESIÓN LINEAL

- **Los resultados de estas evaluaciones se expresan en la siguiente tabla:**

No.	X	y	$\hat{y}$
1	8	36	32.3169
2	9	33	31.3712
3	12	25	28.5341
4	13	28	27.5884
5	16	22	24.7513
6	19	20	21.9142
7	24	15	17.1857
8	27	19	14.3486

REGRESIÓN LINEAL

Graficando en forma conjunta tanto los datos pareados como la línea de regresión obtenemos la siguiente gráfica.



REGRESIÓN LINEAL

- **Si la recta de regresión se utiliza para determinar valores de la variable Y considerando valores dentro del rango de la variable independiente, decimos que estamos *interpolando valores*.**
- **Si la recta de regresión se utiliza datos cercanos a los que conocemos para la variable independiente, pero que quedan fuera de su rango se dice que estamos *extrapolando valores* o *realizando un pronóstico*.**
- **<u>Observación.</u> Esta ecuación es una estimación de la verdadera recta de regresión, que se basa en un conjunto especifico de datos muestra, y cualquier otra muestra extraída de la misma población muy probablemente proporcionará una recta de regresión un poco diferente.**

EJERCICIO

Se quiere explicar la participación en el mercado de una marca de pasta en función del precio de venta. Los datos aparecen en la siguiente tabla para los últimos 12 meses.

% mercado	3.63	4.2	3.33	4.54	2.89	4.87	4.9	5.29	6.18	7.2	7.25	6.09
Precio ($)	9.7	9.5	9.9	9.1	9.8	9.0	8.9	8.6	8.5	8.2	7.9	8.3

a. **Construya el diagrama de dispersión.**
b. **Calcule el valor del coeficiente de correlación entre el precio y la participación en el mercado, y establezca el tipo de relación.**
c. **Desarrolle la ecuación de regresión para determinar la participación en el mercado en base a su precio.**
d. **Determine la participación del mercado cuando la pasta cuesta $ 10**

BIBLIOGRAFÍA

BÁSICA

- Johnson, Robert, Estadística Elemental, Grupo Editorial Iberoamérica, México, 1990.
- Daniel, Wayne W. Estadística con aplicaciones a las ciencias sociales y a la educación, ED. Mc Graw HiII, México 1988.
- Portilla Chimal, Enrique. Estadística, Primer curso, Ed. Mc Graw HiII, Primera edición, México, 1992.
- Berenson, Mark, Estadística para la administración y economía. Editorial Interamericana, México, 1988.
- Levine, Richard I. Estadística para administradores, Ed. Harla, México. 1985

COMPLEMENTARIA

- Proaño, Humberto. Estadística aplicada a la mercadotecnia, Ed. Diana, México 1975.
- Chou, Ya-Iun, Análisis estadístico, Ed. Interamericana, México, 1997.
- Hoel, Paul G. Estadística Elemental, Ed. CECSA, 3a. edición, México, 1982.