UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

COLEGIO DE CIENCIAS Y HUMANIDADES
PLANTEL VALLEJO
ÁREA DE MATEMÁTICAS
ESTADÍSTICA Y PROBABILIDAD I

UNIDAD II
DATOS BIVARIADOS

Introducción:
Sabemos que la 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

Por ejemplo: En el periódico La Jornada del mes de abril de 2004 qp43dió esta tabla de información:

MATRIZ DE REASIGNACIÓN DE MUNICIPIOS 2000 Y 2003

		GANADOR 2003				
		PRI	PAN	PRD	OTROS	TOTALES
GANADOR 2000	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?

Siendo 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.
Por ejemplo los siguientes datos de los censos de población de México.

Año	Población
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. Será posible estimar la población en los años en los que no se realizaron censos durante el periodo 1950 – 2000? (interpolación).
3. ¿Será posible estimar la población del año 2005, y en algunos años futuros, a partir de estos datos? (predicción).

Las situaciones planteadas en la primera unidad de han referido a la observación de una sola variable. Con el propósito de conocer los métodos más usuales empleados en la organización, análisis y medición de los datos aportados por dichas observaciones. Aunado a lo anterior, se presentan frecuentemente en la investigación casos que se refieren a la observación de dos o más variables, relacionadas o ligadas por algún tipo de relación que es importante medir.

Frecuentemente deseamos analizar más de una característica de un elemento de una muestra o población dando por resultado un análisis llamado análisis multivariado, restringiéndonos en esta unidad a desarrollar el análisis de dos atributos o variables asociadas a dos características de un mismo elemento de la muestro o población.

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.
Ala consideración conjunta de dos variables X, Y o dos atributos A, B se les llama variable bidimensional (X, Y) con valores (x_i, y_i) o atributo bidimensional (A, B) con valores (a_i, b_i), junto con sus frecuencias constituyen una distribución bidimensional.

Un subconjunto de datos bivariados consiste en una colección de observaciones simultaneas de dos variables X, Y. Los datos bivariados relacionados con un elemento se representan mediante pares ordenados (x_i, y_i).
De la misma forma que se hace en el caso unidimensional, debemos buscar una forma organizada de presentar las observaciones. Esta organización la conseguimos al utilizar una tabla de doble entrada. Una tabla de doble entrada tiene el siguiente aspecto:

	X/Y	Variable Y: y_1	y_2	....	y_t	Totales
Variable X	x_1	f_{11}	f_{12}	....	f_{1t}	$\sum_{j=1}^{t} f_{1j} = f_{1\bullet}$
	x_2	f_{21}	f_{22}	....	f_{2t}	$\sum_{j=1}^{t} f_{2j} = f_{2\bullet}$
	....	....	....	....	....	
	x_k	f_{k1}	f_{k2}	....	f_{kt}	$\sum_{j=1}^{t} f_{kj} = f_{k\bullet}$
	Totales	$\sum_{i=1}^{k} f_{i1} = f_{\bullet 1}$	$\sum_{i=1}^{k} f_{i2} = f_{\bullet 2}$	....	$\sum_{i=1}^{k} f_{it} = f_{\bullet t}$	$\sum_{i=1}^{k}\sum_{j=1}^{t} f_{ij} = f_{\bullet\bullet} = n$

Donde f_{ij} representa la frecuencia absoluta de la observación conjunta (x_i, y_i), $\sum_{j=1}^{t} f_{ij} = f_{i\bullet}$ representa el numero de elementos de la muestra que poseen el nivel i de la variable X, y $\sum_{i=1}^{k} f_{ij} = f_{\bullet j}$ es la frecuencia de elementos que tienen el nivel j de la variable o atributo Y.

Las frecuencias relativas se calculan al dividir la correspondiente frecuencia absoluta de cada dato ordenado, por el total de observaciones, ***n*** es decir;

$$fr_{ij} = \frac{f_{ij}}{n}$$

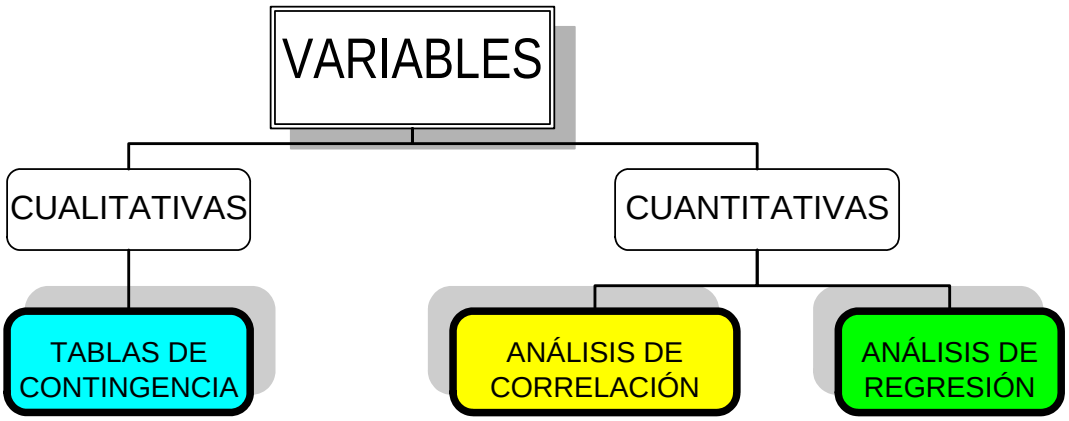
Donde n es el número total de pares observados.

El análisis para datos bivariados lo haremos considerando dos aspectos:

1) Que al menos una de las variables es cualitativa.
2) Ambas variables con cuantitativas.

Estos aspectos nos plantean dividir el estudio de datos bivariados en dos secciones:

a) Tablas de contingencia.
b) Análisis de correlación y análisis de correlación.



Tablas de Contingencia

Sea (a_1,b_1), (a_2,b_2), (a_3,b_3),…, (a_n,b_n), una muestra de n observaciones de un atributo estadístico bidimensional (A, B), de manera que A representa k modalidades $(a_1,a_2,a_2,...,a_k)$ y B representa t modalidades $(b_1,b_2,b_2,...,b_t)$ (El número de k modalidades distintas que adopta A no tiene por qué ser el mismo que adopta B).
Llamaremos tabla de contingencia de dos atributos (A; B) a una tabla de doble entrada que representa los valores observados de ambos atributos y las frecuencias (absolutas o relativas) de aparición de cada par de valores, dicha tabla tiene la siguiente estructura:

		Variable B				
	A/B	b_1	b_2	….	b_t	**Totales**
Variable A	a_1	f_{11}	f_{12}	….	f_{1t}	$\sum_{j=1}^{t} f_{1j} = f_{1\bullet}$
	a_2	f_{21}	f_{22}	….	f_{2t}	$\sum_{j=1}^{t} f_{2j} = f_{2\bullet}$
	….	….	….	….	….	
	a_k	f_{k1}	f_{k2}	….	f_{kt}	$\sum_{j=1}^{t} f_{kj} = f_{k\bullet}$
	Totales	$\sum_{i=1}^{k} f_{i1} = f_{\bullet 1}$	$\sum_{i=1}^{k} f_{i2} = f_{\bullet 2}$	….	$\sum_{i=1}^{k} f_{it} = f_{\bullet t}$	$\sum_{i=1}^{k}\sum_{j=1}^{t} f_{ij} = f_{\bullet\bullet} = n$

Donde f_{ij} es el número de veces que aparece repetido el par (a_i,b_j), y que llamaremos frecuencia absoluta del par (a_i,b_j). Denotamos por fr_{ij} la frecuencia relativa de dicho par, y que vendrá dada por la expresión $fr_{ij} = \frac{f_{ij}}{n}$, donde n es el número total de pares observados.

Ejemplo: Se realizó una encuesta entre un grupo de trabajadores para determinar su nivel de estudios y su satisfacción respecto al trabajo que desempeñan, obteniéndose los siguientes resultados:

		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

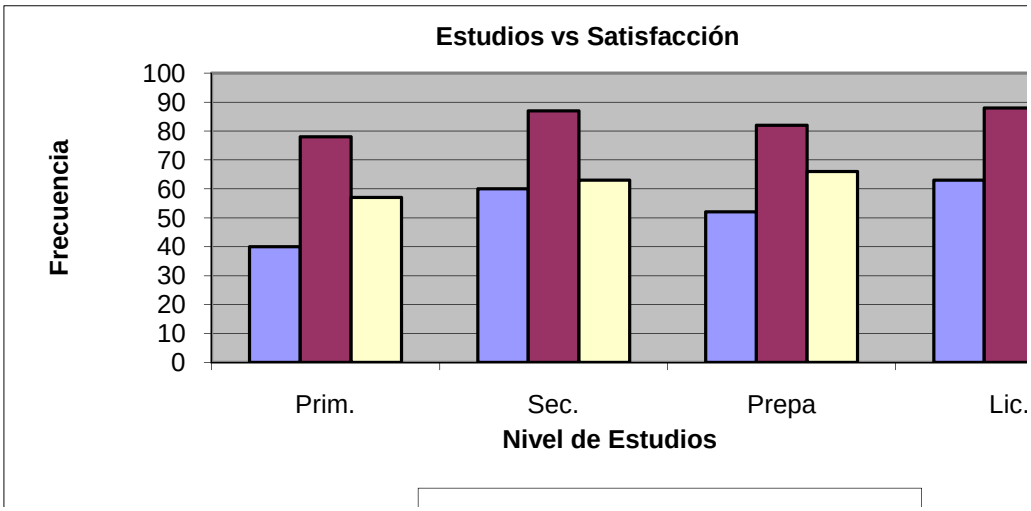
Para este ejemplo tenemos dos atributos:
A) Satisfacción en el trabajo, el cual tiene 3 modalidades: Mucha, Regular, Poca.
B) Nivel de estudios, que tiene 4 modalidades: Primaria, Secundaria, Preparatoria y Licenciatura.

Cada uno de los valores de ésta tabla representa el número de trabajadores con un cierto nivel de satisfacción en el trabajo y con establecido nivel de estudios. Por ejemplo: la intersección de mucha satisfacción con primaria encontramos el valor de 40 ($f_{11} = 40$), el cual nos indica que hay 40 empleados que están muy satisfechos con su trabajo y que al mismo tiempo tienen un nivel de estudios de primaria.

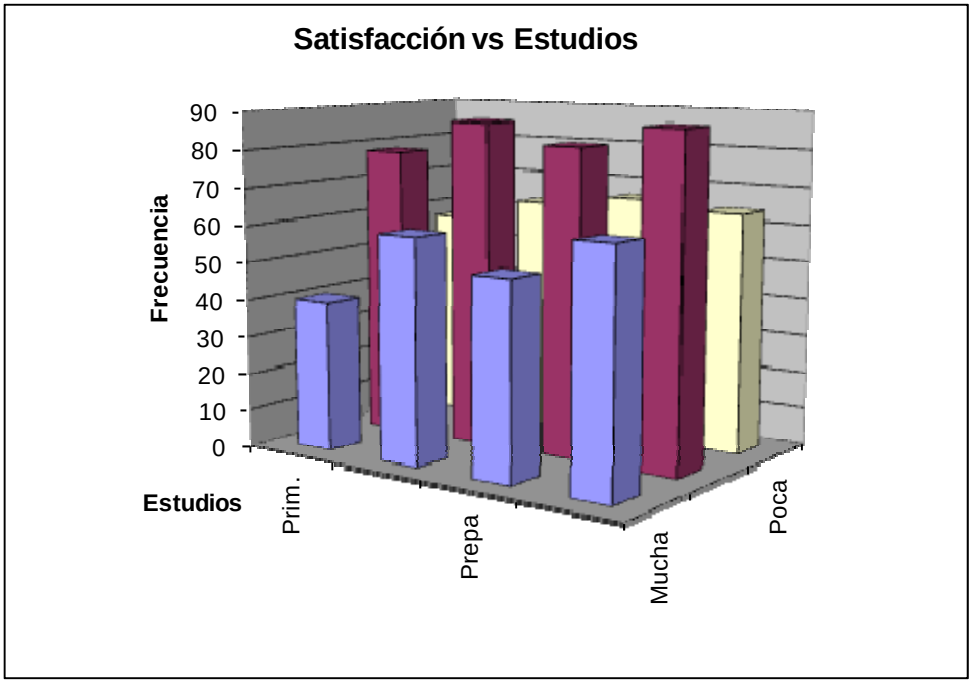
Analizando algunos resultados más podemos observar que 87 trabajadores que tienen una satisfacción regular, cuando su nivel de estudios es de secundaria; que hay 64 empleados con licenciatura que se sienten poco satisfechos con su trabajo, etc.

Representaciones graficas

La representación grafica más útil de dos atributos agrupados es el diagrama de barras que se obtiene representando cada celda (a_i, b_j), como una barra de altura f_{ij} en el plano cartesiano o en el plano tridimensional (a_i, b_j, f_{ij}).



O en la forma:

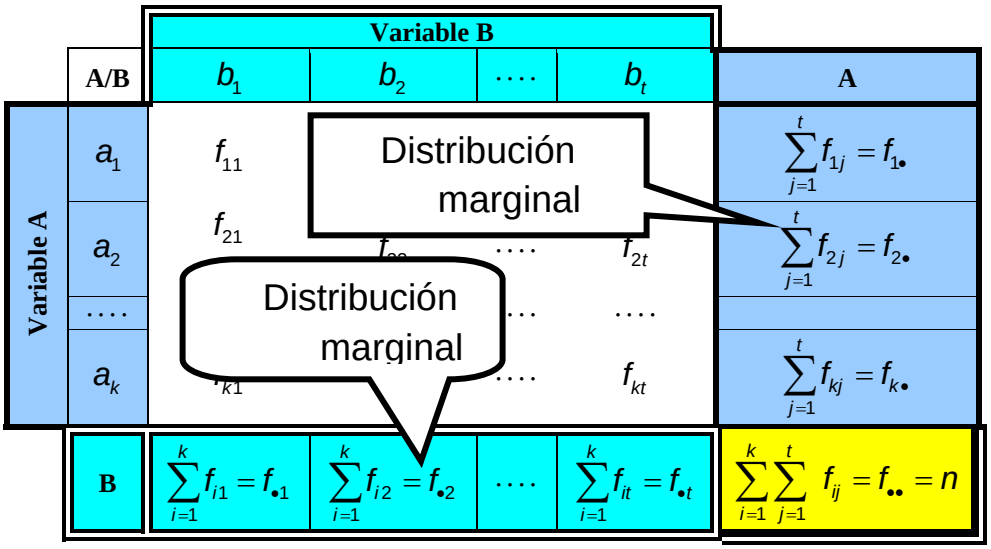


Podemos observar que el número total de empleados entrevistados fue de 800, que 215 de ellos tenían mucha satisfacción en su trabajo, que 355 desempeñan su trabajo don una satisfacción regular y que 250 tiene poca satisfacción al realizar su trabajo; también podemos notar que tenemos 175 empleados con nivel de primaria, que 210 terminaron la secundaria, que 200 tienen estudios de preparatoria y que 215 alcanzaron a tener estudios de licenciatura.

Distribuciones marginales

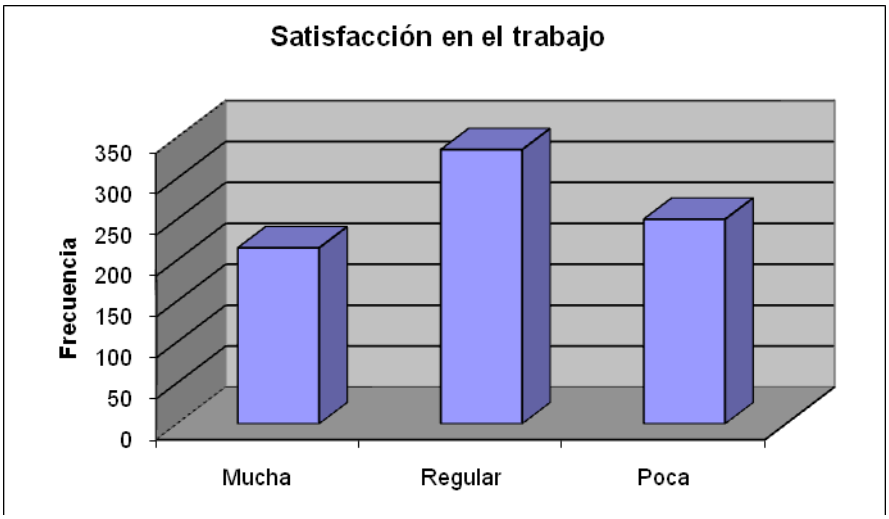
Muchas veces, a pesar de que observamos pares de datos (a_i, b_j), podemos estar interesados en estudiar el comportamiento de una sola de las variables, por ejemplo B, independientemente de la otra. A partir de las frecuencias conjuntas (a_i, b_j), f_{ij} podemos obtener la frecuencia de observación a_i (con independencia de los valores de B). Esto es lo que se llama ***distribución marginal del atributo A***.

Se denomina distribución marginal a la distribución de frecuencias obtenida al estudiar la variable (a o B), aisladamente, es decir con independencia del resto de atributos.



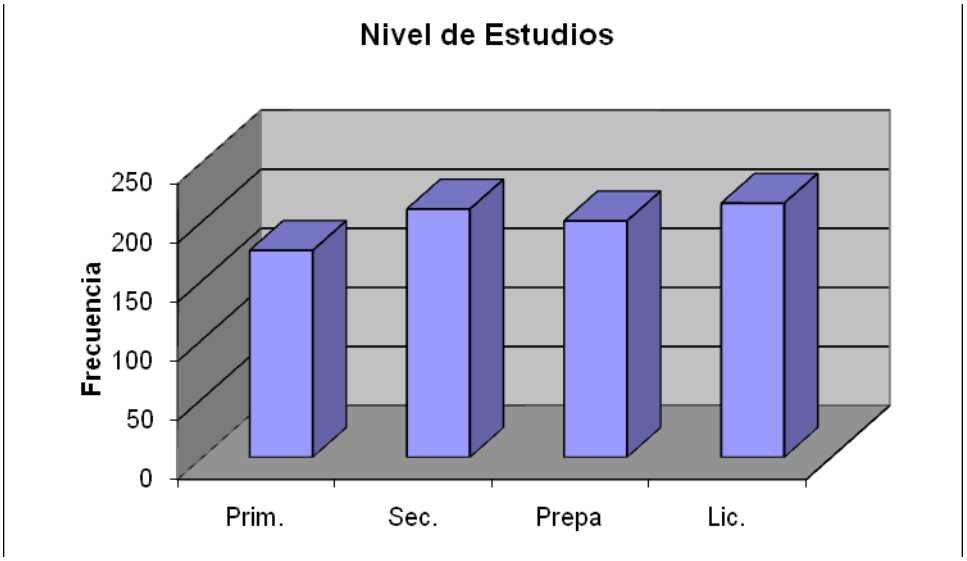
Para nuestro ejemplo la distribución marginal para el atributo ***satisfacción en el trabajo*** es de la forma:

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



La distribución marginal para el atributo ***nivel de estudios*** es de la forma:

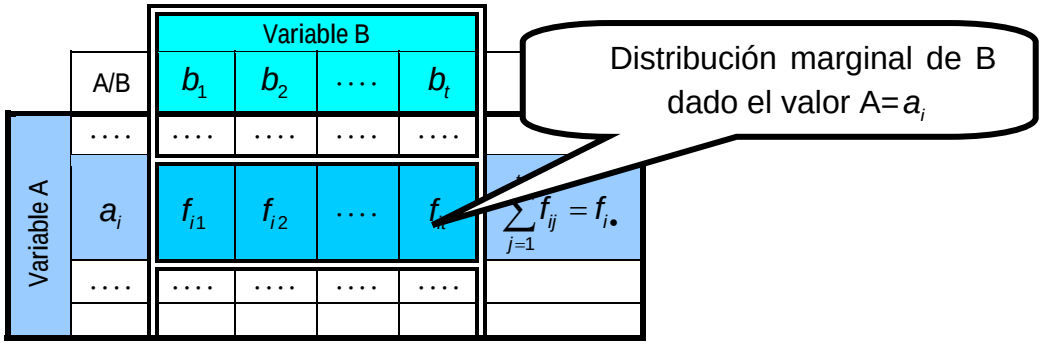
	Nivel de estudios				
	Primaria	Secundaria	Preparatoria	Licenciatura	Totales
Totales	175	210	200	215	800



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 A cuando la variable B toma el valor fijo b_j? Esto es lo que se conoce como **distribuciones condicionadas**.

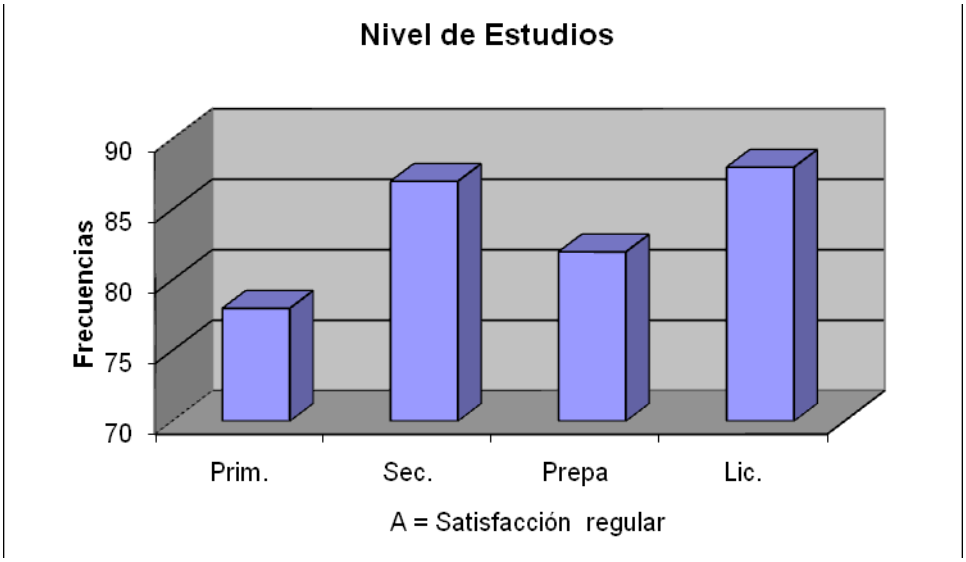
La distribución de atributo B condicionado a que A toma el valor a_i (A=a_i) es la distribución de B que se obtiene considerando sólo los elementos que tienen para el atributo A el valor a_i



Por ejemplo deseamos determinar la distribución condicionada del nivel de estudios dada una satisfacción regula en el trabajo. Esta distribución está indicada por el renglón que indica una satisfacción regular en el trabajo, la cual incluye todos los niveles de estudios.

		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

Para la cual podemos construir el siguiente diagrama de barras;



En forma análoga, la distribución del atributo A condicionado a que el atributo B toma el valor b_j (B = b_j), es la distribución de A que se obtiene considerado sólo los elementos que tienen para el atributo B el valor b_j.

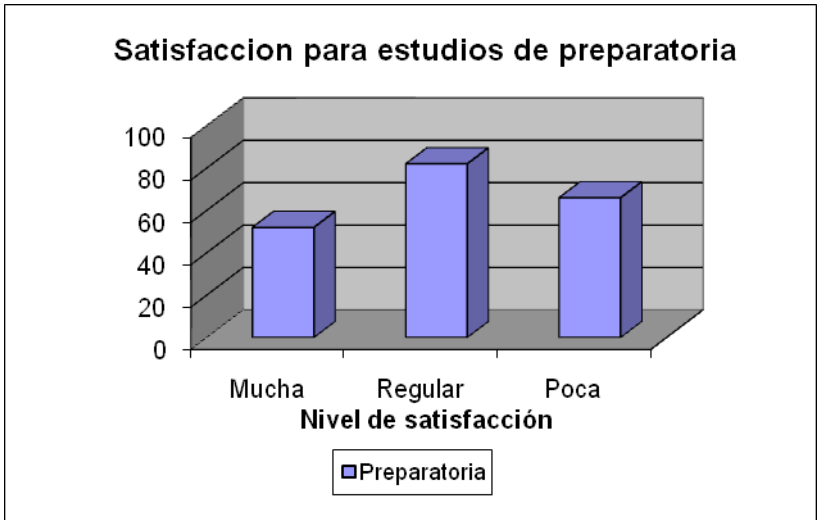
	A/B	....	Variable B: b_j	....	Tot
Variable A	a_1	....	f_{1j}	....	
	a_2	....	f_{2j}	....	
	....	....	....	....	
	a_k	....	f_{kj}	....	
	Totales		$\sum_{i=1}^{k} f_{ij} = f_{\bullet j}$	....	

Distribución condicional de A dado el valor de B = b_j

Siguiendo con nuestro ejemplo deseamos encontrar la 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

Que al presentarla mediante un diagrama de barras obtenemos:



Ejercicio: Tenemos hambre pero no tenemos mucho tiempo para comes 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					

a. Determinar los totales marginales.
b. Construir la tabla de frecuencias relativas respecto al total de la muestra.
c. Elaborar la gráfica de barras para la tabla de contingencia.
d. Construir la tabla marginal y construir el diagrama de barras con respecto a la bebida adquirida.
e. Encontrar la tabla marginal y construir el diagrama de barras con respecto al tipo de alimento comprado.
f. Determinar la distribución condicional de bebida dado que el alimento que se compra es de frituras.
g. Encontrar la distribución condicional de los alimentos, dado que adquirieron como bebida yogurt.

CORRELACIÓN Y REGRESIÓN LINEAL

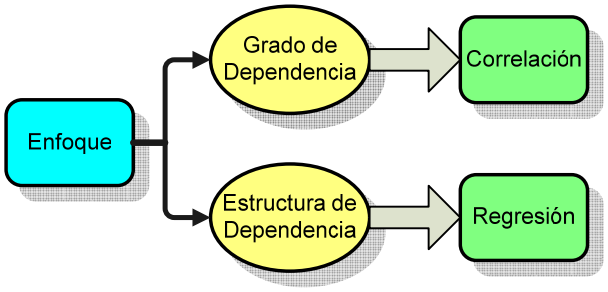
Consideremos ahora que nuestras variable bidimensional (X, Y) son del tipo cuantitativo y queremos determinar si existe una relación entre las dos variables, y de existir, identificar qué tipo de relación es. Si existe tal relación, sería bueno expresarla con una sencilla ecuación que nos permita predecir el valor de una de las variables si conocemos el valor de la otra.

Lo primero que se hará es considerar el concepto de correlación, que sirve para determinar si existe una relación estadísticamente significativa entre dos variables.

Posteriormente en el análisis de regresión, expresamos la relación existente entre las dos variables con una ecuación, es decir encontraremos una formula matemática que relaciona dichas variables y se aprenderá a usar dicha ecuación para predecir los valores de una de ellas.

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

a) La estructura verdadera de la relación se desconoce.
b) No existe una dependencia funcional exacta entre las variables consideradas.



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

1) El análisis de correlación que tiene como objetivo determinar el grado de relación entre variables.
2) El análisis de regresión que trata de establecer la “naturaleza de la relación” entre variables, es decir, estudiar la relación funcional entre las variables y por lo tanto proporcionar un mecanismo de predicción o pronostico.

El análisis de asociación puede dividirse en simple y múltiple, el primero se aplica solamente cuando dos variables y el segundo cuando la asociación es entre tres o más variables. Además existe también la diferencia entre asociación lineal y no lineal, según el tipo de relación que existe entre las variables.

Primer aspecto ¿Existe dependencia entre las variables?

Ejemplo 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?

Ejemplo 2: Como estudiantes nos dedicamos a estudiar y a resolver exámenes, ¿Será cierto que cuanto más se estudie tanto mayor es la calificación obtenida?

Ejemplo 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?

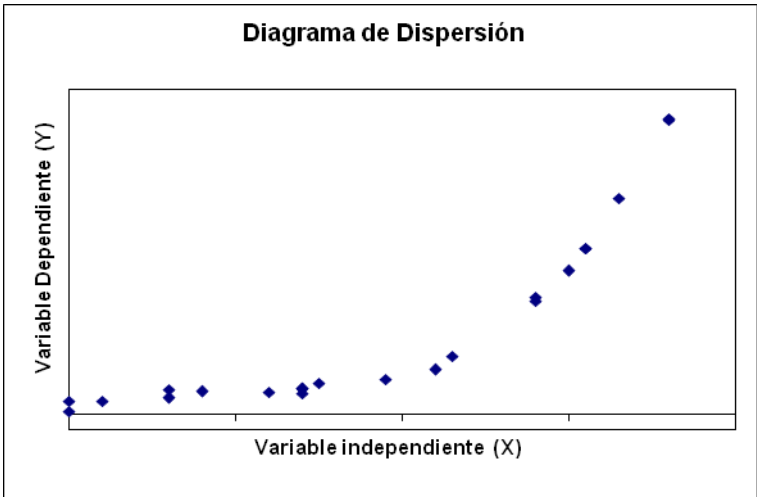
Es decir, en todos los casos queremos saber si existe una cierta variación conjunta entre las dos variables, y si es asó determinar el grado de dependencia que existe entre ellas y por supuesto verla reflejada mediante una regla o ecuación.

El estudio de la relación entre dos variables inicia con el caso más sencillo, el de la asociación existente entre las variables, supongamos que se toman 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.

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 A, se utiliza para el eje horizontal, y la variable Y para el eje vertical.

Para analizar el diagrama de dispersión es necesario utilizar nuestra intuición para determinar si la relación es lineal (una línea recta), o una curva. Si la relación es lineal se deseará saber si la relación es positiva o negativa y cuál es la pendiente de la línea que se ajusta a los puntos dados y por último se necesita saber el grado de la relación, esto es, qué tan cerca están los puntos de la curva que mejor los ajusta.

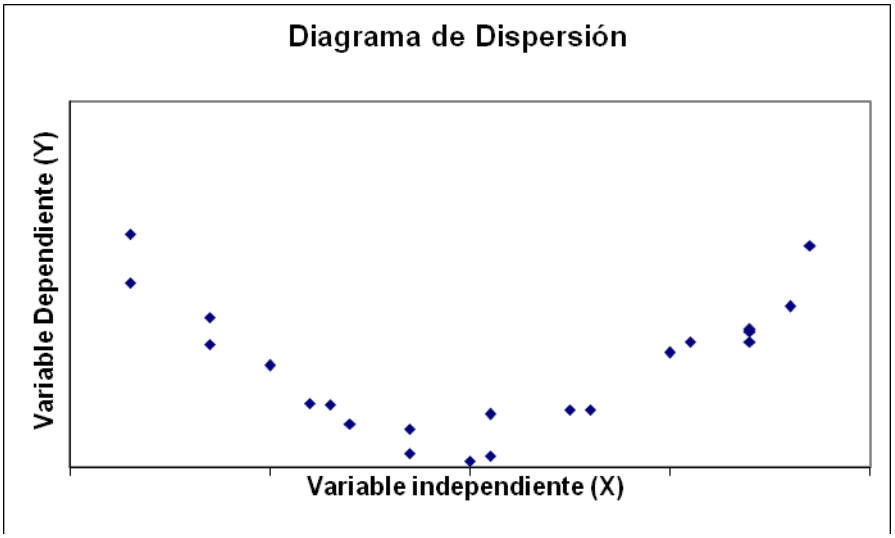
Ejemplo 1: Observemos el siguiente diagrama de dispersión ¿A qué conclusiones podemos llegar?



En este diagrama de dispersión podemos observar que a medida que el valor de la variable independiente (regularmente indicada por X) aumenta, el valor de la variable

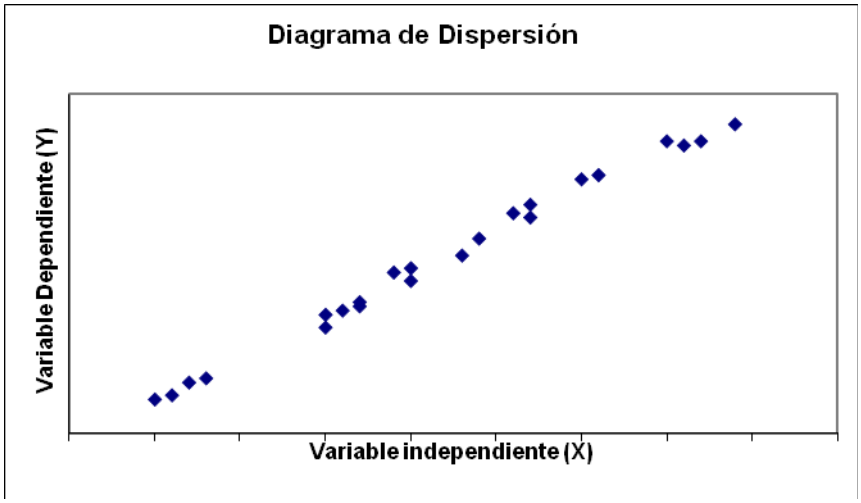
dependiente (usualmente indicada por Y) aumenta, esto nos indica que existe una evolución positiva, y que evidentemente los puntos se agrupan alrededor de una curva (que no es una recta) de la cual no conocemos su forma funcional.

Ejemplo 2: Revisemos el siguiente diagrama de dispersión, ¿Cuáles son ahora nuestras conclusiones?



Podemos observar que no hay un comportamiento uniforme ya que primero conforme aumenta el valor de la variable independiente el valor de la variable dependiente disminuye y posteriormente comienza a aumentar, además observar que los puntos siguen un patrón no lineal, por lo que evidentemente la relación que pudiéramos utilizar para expresar la relación entre ellas no sería una línea recta.

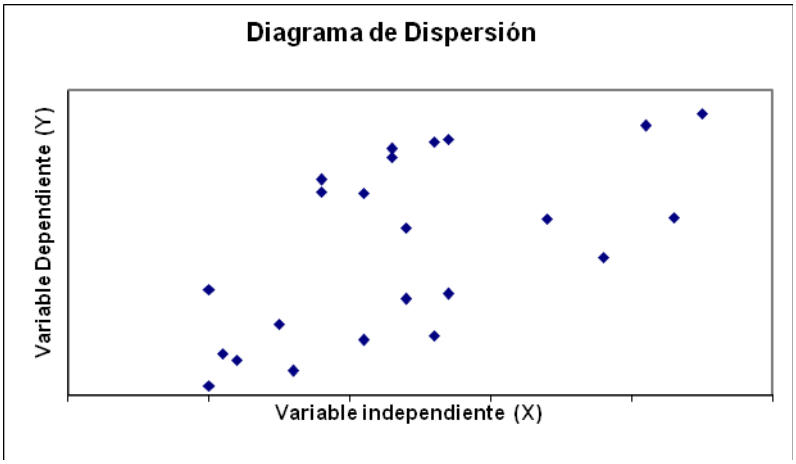
Ejemplo 3: Comentemos el siguiente diagrama de dispersión, ¿Qué conclusiones podemos obtener?



En este diagrama de dispersión podemos observar dos propiedades; primera cuando una de las variables aumenta su valor, el valor de la otra variable aumenta, es decir,

tiene una variación positiva, segundo; evidentemente la relación que pudiéramos utilizar para relacionarlas pudiera ser una línea recta, ya que el conjunto de puntos presenta un patrón de comportamiento cercano a una recta con pendiente positiva.

Ejemplo 4; ¿Qué conclusiones son apropiadas para el siguiente diagrama de dispersión?



Para este diagrama de dispersión los puntos no sigue patrón alguno ya que no se vislumbra ningún comportamiento uniforme que describa a los datos por lo que resultaría difícil encontrar alguna relación que permitiera describir a este conjunto de datos.

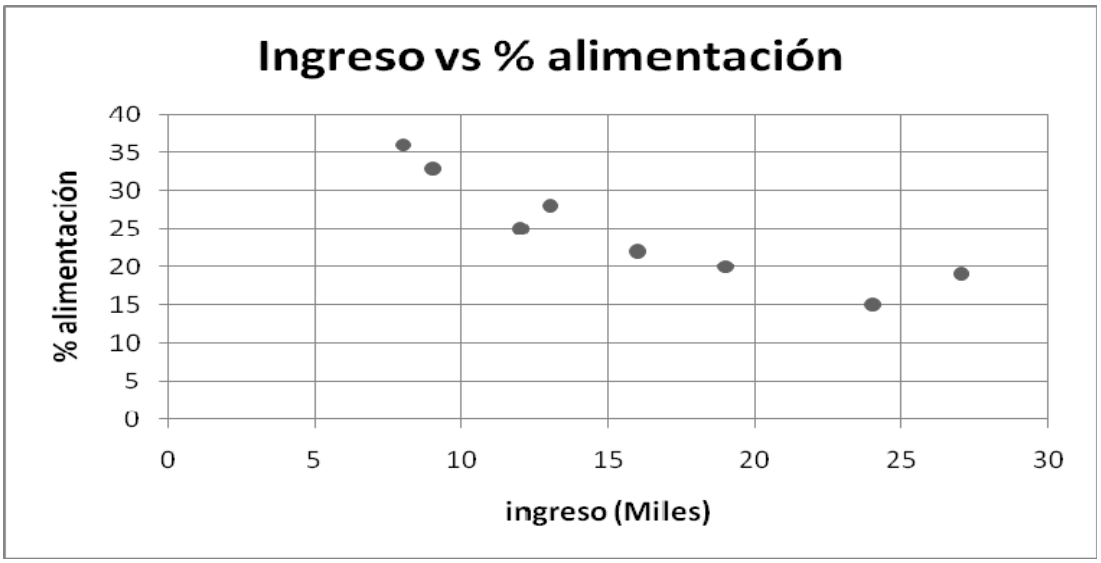
En resumen analizar el diagrama de dispersión nos es útil para:
a) Que visualmente busquemos patrones que nos indiquen que las variables están relacionadas.
b) Y si esto sucede en él se esboza el tipo de curva (recta, parábola, etc.) que puede describir esta relación.

Apliquemos estos conceptos al siguiente ejercicio:

Ejemplo 1: Un economista, está interesado en determinar si existe alguna relación entre el ingreso familiar (X) y el porcentaje de ingreso gastado en alimentación (Y).
La tabla, abajo mostrada, indica los resultados de un estudio de 8 familias seleccionadas al azar.

.	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

Primero construyamos el diagrama de dispersión para este conjunto de datos, para analizarlo posteriormente y determinar si las variables están relacionadas y de qué forma están relacionadas, obtenemos el siguiente diagrama:



Parece ser que los puntos del diagrama de dispersión se aglutinan alrededor de un patrón de tipo lineal, por lo que podríamos intuir que existe una relación de tipo lineal entre el ingreso y el porcentaje de ingreso gastado en alimentación. Aún más podemos observar que a medida que el ingreso aumenta el porcentaje de gasto en alimentación disminuye (es decir tienen un comportamiento inverso).

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.

CORRELACIÓN

Ahora el problema consiste en determinar si hay alguna relación aparente entre dos variables, a una relación de este tipo se le llama ***correlación lineal***.

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

Una vez que se ha detectado que existe una correlación lineal entre nuestra dos variables, nuestro objetivo es medir el grado de asociación entre estas variables.

Para nuestro ejemplo, en el primer análisis determinamos que la relación es posiblemente de tipo lineal, utilizaremos el coeficiente de correlación lineal (el cual sirve para detectar patrones de línea recta).

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

Es decir elcoeficiente de correlación mide el grado al cual se relaciona en forma lineal dos varibles 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.

El coeficiente de correlación se calcula mediante la ecuación:

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

donde:

NOTACIÓN PARA EL COEFICIENTE DE CORRELACIÓN	
n	Número de pares de datos
$\sum x$	Suma de los valores de ***x***
$\sum y$	Suma de los valores de ***y***
$\sum x^2$	Indica que cada valor de ***x*** se debe elevar al cuadrado y luego sumar todos los cuadrados
$\sum y^2$	Indica que cada valor de ***y*** se debe elevar al cuadrado y luego sumar todos los cuadrados
$\left(\sum x\right)^2$	Indica el cuadrado de la suma de los valores de ***x*** (primero se suman los valores de ***x*** y la suma se eleva al cuadrado)
$\left(\sum y\right)^2$	Indica el cuadrado de la suma de los valores de ***y*** (primero se suman los valores de ***y*** y la suma se eleva al cuadrado)
$\sum xy$	Indica la suma de los productos de ***x*** y ***y*** (se multiplica cada valor de ***x*** por su correspondiente ***y*** y luego se suman los productos)

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, pero si r está cerca de –1 o +1, concluimos que existe una correlación lineal significativa entre x, y.

En la correlación de dos variables se distinguen dos casos básicos: los casos de ***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). Por otra parte existen también los casos de ***correlación negativa*** que ocurre cuando al crecer una de las variables, la otra decrece (es decir tienen un comportamiento inverso).

Los casos extremos ocurren cuando:

1.- La relación de dos variables es perfectamente ***positiva***, o sea cuando al variar la primera, la segunda varía ***en las mismas proporciones y en la misma dirección***, el coeficiente de correlación es +1.

2.- La relación de dos variables es perfectamente ***negativa***, o sea cuando al variar la primera, la segunda varía ***en las mismas proporciones pero en dirección contraria***, el coeficiente de correlación es -1.

3.- ***No existe relación entre dos variables***, o sea cuando al variar la primera, las variaciones de la segunda no reflejan dependencia o conexión alguna con las variaciones de la primera, el coeficiente de correlación es 0.

Lo anterior significa que, entre 0 y +1 cabe toda una gama de correlaciones positivas, que serán tanto más *directamente proporcionales,* cuando más se acerquen a +1,

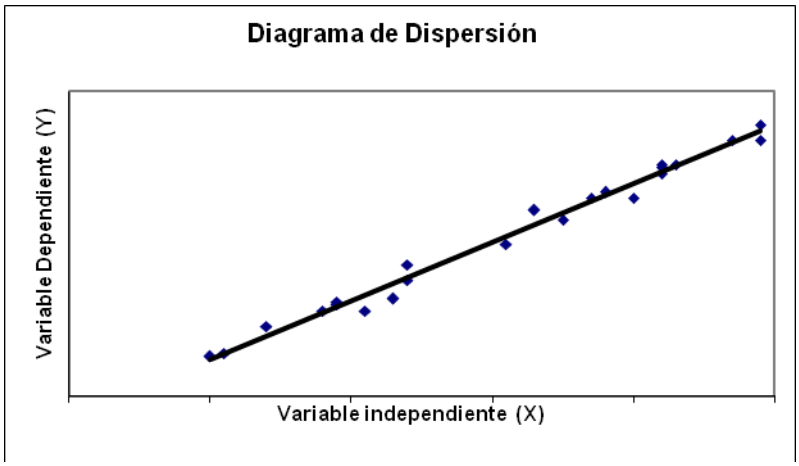


Diagrama de dispersión con una **correlación positiva fuerte**
(r cercana a +1)

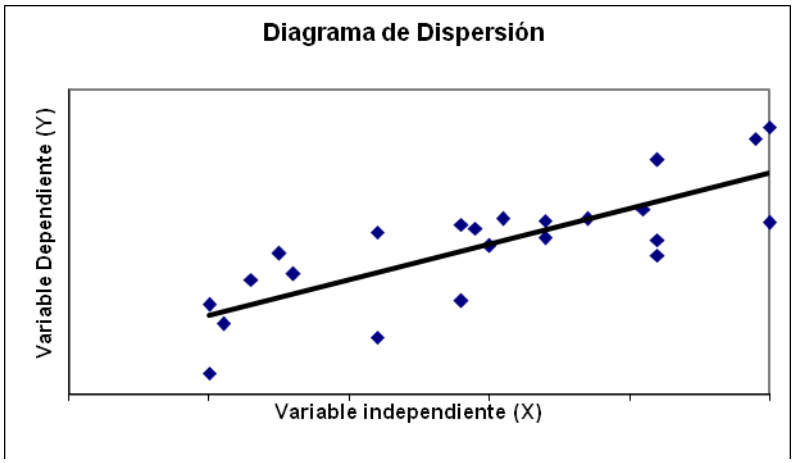


Diagrama de dispersión con una correlación positiva moderada
$(0.5 \leq r \leq 0.8)$

Y entre –1 y 0 cabe toda una gama de correlaciones negativas, que serán tanto más *inversamente proporcionales* cuanto más se acerquen a –1.

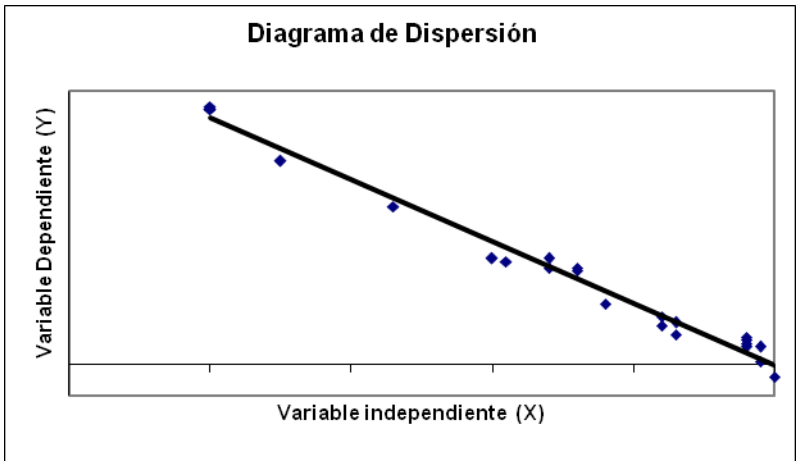


Diagrama de dispersión con una **correlación negativa fuerte**
(r cercana a -1)

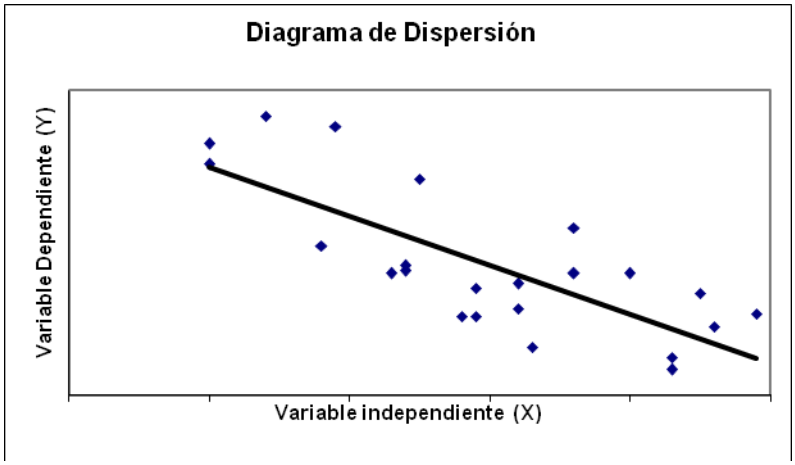


Diagrama de dispersión con una correlación positiva moderada
$(-0.8 \leq r \leq -0.5)$

Los coeficientes de correlación en las cercanías del 0 indicarán ausencia de correlación.

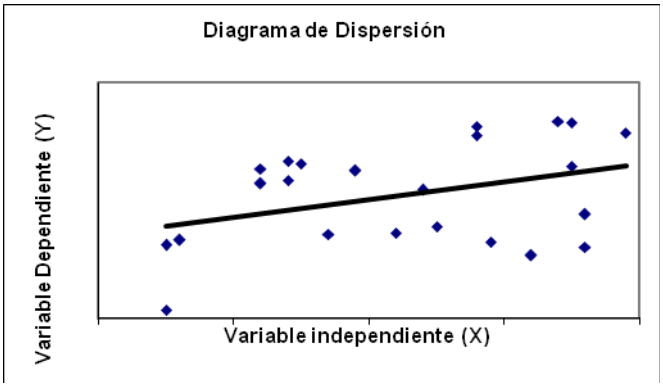


Diagrama de dispersión con una correlación positiva (o negativa) débil ($-0.5 \le r \le +0.5$)

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

	C O R R E L A C I O N					
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**

Continuando con nuestro ejemplo; calcularemos el valor del coeficiente de correlación lineal r para nuestro conjunto de datos, los términos que necesitamos se obtienen realizando los cálculos indicados en la siguiente tabla. Este tipo de tabla facilita los cálculos.

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
	$\sum x$	$\sum y$	$\sum xy$	$\sum x^2$	$\sum y^2$

$$r = \frac{(8)(2854) - (128)(198)}{\sqrt{\left[(8)(2380) - (128)^2\right]\left[(8)(5264) - (198)^2\right]}} = \frac{22832 - 25344}{\sqrt{[19040 - 16384][42112 - 39204]}}$$

$$r = \frac{-2512}{\sqrt{[2656][2908]}} = \frac{-2512}{\sqrt{7723648}} = \frac{-2512}{2779.1452} = -0.9038$$

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.

Ahora como sabemos ya, 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 trazando la gráfica de la misma sobre el diagrama de dispersión. Esta ecuación se denomina ***ecuación de regresión*** y su gráfica se denomina ***línea de regresión***.

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.

Donde : **m:** es la pendiente de la recta de regresión
b: es la ordenada al origen de la recta de regresión.
$\hat{y}$ **:** es el valor estimado mediante la ecuación de regresión.
x: es la variable independiente.

Esta definición expresa una relación entre ***x*** (llamada ***variable independiente o variable predictora)*** y ***y*** (llamada ***variable dependiente o variable de respuesta***).

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

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

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.***

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:

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

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

tomando los valores de *m* y *b,* obtenemos que la recta de regresión tiene como ecuación:

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

La pendiente se puede interpretar como que por cada $1000 de 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. Pero hay que tener cuidado con su interpretación practica ya que en ocasiones no tiene sentido hablar de su valor.

Para este ejemplo su interpretación no tiene sentido practico ya que nos indicaría que cuando el ingreso es cero, el porcentaje del mismo gastado en alimentación es del 39.88%. No tiene interpretación práctica ya que al no haber ingreso no puede haber porcentaje de gasto.

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 = - 7.5656 + 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 %

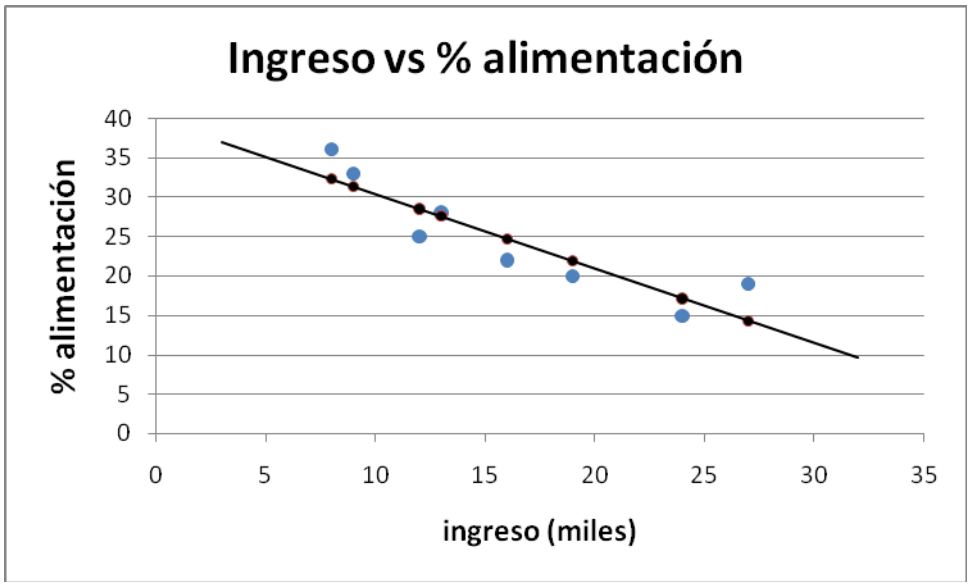
Para *x = 9* $\hat{y}$ = -0.9457*(9) + 39.8825 = - 8.5113 + 39.8825 = 31.3712
Esto nos indica que para un ingreso de $ 9 000 el porcentaje de ingreso gastado en alimentación se estima en 31.37 %

Para *x = 12* $\hat{y}$ = -0.9457*(12) + 39.8825 = - 11.4564 + 39.8825 = 28.5341
y así sucesivamente.

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

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



Cuando 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***.

Cuando 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***.

Observación. Debemos tener presente que esta ecuación es una estimación de la verdadera ecuación de regresión ya que esta se basa en un conjunto especifico de datos muestra, ya que cualquier otra muestra extraída de la misma población probablemente proporcione una ecuación de regresión un poco diferente.

ANÁLISIS DE REGRESIÓN LINEAL Y DE CORRELACIÓN SIMPLE ALGUNAS PRECAUCIONES

El empleo adecuado del análisis de regresión y de correlación proporciona herramientas útiles y poderosas para el análisis de datos. Sin embargo, se debe evitar su uso incorrecto o la mala interpretación de resultados al emplearla inapropiadamente.

El hecho de que los datos indiquen una alta relación lineal no debe interpretarse como un indicio de relación de causa y efecto. Un coeficiente de correlación muestral muy significativo, entre X e Y, puede ser un reflejo cualquiera de las siguientes situaciones:

1. X es causa de Y. Cuando las variaciones de la variable dependiente Y son efecto de las variaciones de X (Regresión causal).
2. Y es causa de X. Cuando las variaciones de la variable dependiente X, dependen causalmente de las variaciones de Y, (Relación causal)
3. Alguna tercera variable es causa de X y de Y, directa o indirectamente. Cuando las variaciones de Y no dependen de X sino que ambas varían en función de una causa común. (Relación concomitante)
4. Ha ocurrido un evento improbable y se ha obtenido una muestra con coeficiente de correlación significativo, como resultado de la casualidad solamente, en una población en la que X e Y no están correlacionadas (Relación fortuita)
5. La correlación es puramente espuria. Estas correlaciones son debido a suposiciones engañosas, y esta es uno de los principales peligras que hay que evitar, ya que el índice de correlación por sí sólo no constituye evidencia de que exista una relación plausible, o que el hecho de haber encontrado un alto coeficiente de correlación implique automáticamente que existe dependencia entre las dos variables.

En el análisis de regresión hay que proceder con cautela cuneado se está considerando la predicción de Y para valores de X fuera de los límites de aquella variable representada en la muestra. Esta práctica, que denominamos extrapolación, puede producir resultados erróneos. Cuando se emplea una ecuación de regresión lineal simple para predecir Y en un valor de X más allá del límite superior de los valores de X de la muestra, se supone que la relación entre X e Y continúa siendo lineal en esta región. Si la suposición no se hace, la predicción será errónea, El mismo razonamiento se aplica a la estimación de valores Y correspondientes a valores de X más pequeños que el valor mínimo de X en la muestra. Se debe usar la extrapolación solamente cuando de sabe que la relación es lineal en el área en que aquella tiene lugar.

RESUMEN DE LA UNIDAD

En esta unidad nos ocupamos del análisis de datos bivariados, es decir de información sobre elementos de una muestra que describen dos de sus características.

Cuando las variables a estudiar son atributos, la información la ordenamos formando una tabla de doble entrada denominada tabla de contingencia.

Con respecto a las tablas de contingencia, las hemos utilizado para obtener

- Diagramas de barras que representan la información.
- La distribución marginal para una de las variables.
- La construcción del diagrama de barras para una distribución marginal.
- La distribución condicionada de una de las variables con respecto a un valor especifico de otra variable.
- La construcción de la distribución condicionada de una de las variables.

También se estudio la relación entre dos variables cuantitativas mediante los procedimientos del análisis de regresión lineal y del análisis de correlación simple, y las hemos empleado para obtener interpolaciones de la variable dependiente conociendo algunos valores de la variable independiente.

Se ha sugerido el siguiente procedimiento para el empleo del análisis de regresión lineal:

1. Identificar el modelo mediante el diagrama de dispersión.
2. Determinar el valor del coeficiente de correlación de Pearson.
3. Obtener la ecuación de regresión, mediante el método de mínimos cuadrados.
4. Utilizar la ecuación de regresión.

Hemos estudiado el uso de la ecuación de regresión muestral para predecir el valor que probablemente tomará Y para una X dada.

El análisis de correlación se ha centrado en el coeficiente de correlación de Pearson como medida de la fuerza de la relación entre dos variables.

Por último hemos mencionado algunas de las precauciones que se deben tomar en cuenta cuando se emplea el análisis de regresión y correlación.