COLEGIO DE BACHILLERES

ESTADÍSTICA DESCRIPTIVA E INFERENCIAL I

FASCÍCULO 3. CORRELACIÓN Y REGRESIÓN LINEALES

Autores: Alejandro Rosas Snell
Juan Zúñiga Contreras

COLEGIO DE
BACHILLERES

Colaboradores:

Asesoría Pedagógica:
Irma Cruz Santillán

Revisión de Contenido
Armando Martínez Cruz

Diseño Editorial
Leonel Bello Cuevas
Javier Darío Cruz Ortiz

ÍNDICE

P R O P Ó S I T O

En los fascículos anteriores de esta asignatura, has aprendido a utilizar eficazmente los métodos más usuales para organizar, analizar y cuantificar los datos aportados por observaciones estadísticas, todo ello dentro del contexto de la estadística descriptiva. De esa manera, tienes ya un panorama general de los elementos básicos de esta rama importante de la estadística paramétrica.

En este fascículo, efectuaremos una breve introducción a los temas de Correlación y Regresión lineales de datos bivariados, donde aprenderás a calcular, por un lado, en qué medida se relacionan dos variables estadísticas, a través del coeficiente de correlación de Pearson y por otro desarrollarás un método general para calcular la ecuación de regresión lineal que nos llevará a la recta de mejor ajuste, misma que nos permitirá realizar ciertas predicciones estadísticas, a partir de los datos registrados en una tabulación.

Cabe dentro del propósito de este fascículo, el que comprendas la diferencia entre los objetivos que se buscan con el análisis la correlación lineal y los del análisis de regresión.

Es necesario que recuerdes que el análisis de correlación y regresión lineales es un punto de partida para abordar los temas de la inferencia estadística, que serán abordados y analizados en el siguiente curso de Estadística.

INTRODUCCIÓN

Al iniciar el estudio de la correlación y la regresión lineales, te darás cuenta que en el campo de la estadística existen situaciones que requieren el análisis de más de una variable estadística. Por ejemplo, te has preguntado si alguna vez ¿existe una relación entre la estatura y el peso?, ¿están relacionadas la edad y la resistencia física?, ¿influye la temperatura en el índice de criminalidad?, ¿tienden a tener mayor escolaridad las personas con altos ingresos en comparación con las de bajos ingresos? Así también, un profesor puede estar interesado en conocer de qué manera se puede predecir el rendimiento en álgebra basándose en el puntaje obtenido en una prueba de aptitud en dicha asignatura. Así mismo, el psicólogo deseará saber si existe alguna relación entre el concepto que un alumno tiene de sí mismo y su promedio en las asignaturas. También, el sociólogo puede estar interesado en saber qué clase de relación existe entre la tasa de delincuencia juvenil que hay en una comunidad y el grado de hacinamiento de los hogares que ahí se encuentran. Como observarás son muchas situaciones cotidianas que necesitan analizarse estadísticamente utilizando por lo menos dos variables estadísticas.

En todos los ejemplos anteriores, deberás analizar los datos valiéndote de la correlación y la regresión lineales para obtener información acerca de los problemas planteados. Este análisis lo realizarás apoyándote en diagramas de dispersión, el cálculo del coeficiente de correlación de Pearson y la ecuación de mejor ajuste.

Cabe destacar un punto esencial en el análisis, las variables involucradas no necesariamente tienen una relación causa-efecto por lo que deberá tomarse la información obtenida mediante esta herramienta con una óptica estrictamente estadística.

Todas estas actividades te permitirán resolver problemas donde aplicarás la correlación y regresión lineales como instrumentos preliminares en la inferencia estadística.

CUESTIONAMIENTO GUÍA

Frecuentemente se presentan situaciones en las que es de gran interés estudiar la relación entre dos variables. Por ejemplo: el Profr. Gómez está interesado en conocer si el aprovechamiento de nivel superior y el aprovechamiento de nivel medio superior están relacionados. Él piensa que es razonable esperar que los alumnos en el nivel superior tenderán a obtener aproximadamente las mismas calificaciones que en el nivel bachillerato. Al realizar esta investigación se apoyó en quince estudiantes de nivel superior de los últimos años seleccionados al azar, cuyos promedios fueron los siguientes:

No.	ESTUDIANTE	PROMEDIO NIVEL MEDIO SUPERIOR	PROMEDIO NIVEL SUPERIOR
1	Jaime	80	1.0
2	Eduardo	82	1.0
3	Carolina	84	2.1
4	Marcia	85	1.45
5	Pedro	87	2.1
6	José	88	1.7
7	Tomás	88	2.0
8	Irene	89	3.5
9	Claudia	90	3.1
10	María	91	2.4
11	Antonio	91	2.7
12	Ana	92	3.0
13	Javier	94	3.9
14	Erika	96	3.6
15	Linda	98	4.0

¿Podrás ayudar al Profr. Gómez a solucionar este problema? ¿Existe alguna relación entre los promedios de nivel medio superior y de nivel superior?

Quizás al principio no tengas la menor idea de cómo ayudarlo, pero conforme estudies este fascículo, irás adquiriendo los conocimientos necesarios para llegar a la respuesta y, así poder resolverlo por ti mismo.

CORRELACIÓN LINEAL

CONCEPTO DE CORRELACIÓN

En las diferentes áreas del conocimiento existen problemas que requieren el análisis de más de una variable, como por ejemplo; un sociólogo puede estar interesado en saber qué clase de relación existe entre la tasa de delincuencia juvenil que hay en la comunidad y el grado de hacinamiento de los hogares que allí se encuentran; un profesor puede estar interesado en conocer de qué manera se puede predecir el rendimiento en álgebra de un estudiante con base en el puntaje obtenido en una prueba de aptitud en dicha asignatura; un psicólogo desea saber si existe alguna relación entre el concepto que tiene un alumno de sí mismo y su promedio en el estudio; un agrónomo desea conocer si existe relación entre la cantidad de lluvia caída y el rendimiento de ciertos productos agrícolas, es decir, si es afectado desfavorablemente tanto por la excesiva lluvia (humedad), como por la excesiva sequía del suelo.

Como te habrás dado cuenta, estas relaciones y muchas otras se pueden investigar por medio del análisis de correlación y/o regresión, simples o lineales, si la relación está limitada a dos variables (si fueran más de dos variables, este análisis de correlación y regresión sería múltiple). En esta sección del fascículo hablaremos de la correlación lineal cuyo objetivo principal es medir la intensidad de una relación lineal entre dos variables; la correlación lineal sirven para medir la relación entre dos variables.

Después de leer lo anterior, te preguntarás, ¿cómo es que una medida puede representar una relación? En realidad el término medida de correlación lineal implica encontrar un valor numérico que exprese el grado de correspondencia o dependencia que existe entre dos variables. Por ejemplo:

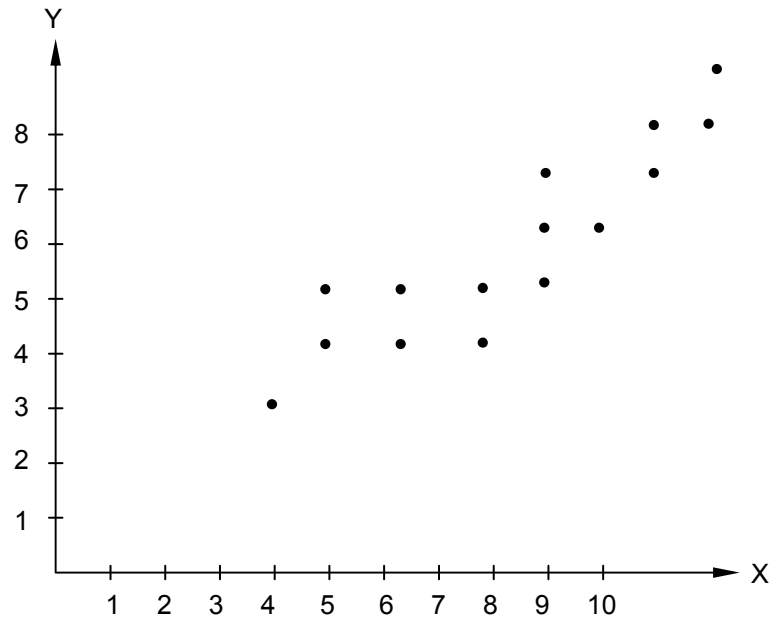
La siguiente tabla muestra las cantidades vendidas (y) por 15 vendedores de una compañía en un periodo dado. La tabla también muestra el número de periodos (x) de experiencia que cada vendedor tiene.

Tabla:

VENDEDOR	NÚMERO DE PERIODOS (x)	VENTAS (y)
1	3	2
2	4	3
3	4	4
4	5	3
5	5	4
6	6	3
7	6	4
8	7	4
9	7	5
10	7	6
11	8	5
12	9	6
13	9	7
14	10	7
15	10	8

Mostraremos la relación entre estas dos variables, gráficamente, para que te des cuenta de cómo están relacionadas estas variables. Más adelante, introduciremos el coeficiente de Pearson, y una fórmula para calcularlo, que nos indicará el grado de relación de estas variables.

Grafiquemos los puntos para observar la relación entre estas variables.



Gráfica No. 1

Este diagrama sugiere que a medida que los valores X aumentan, también los valores Y aumentan. Además, aparece que los puntos se agrupan a lo largo de una línea recta. Por lo mismo decimos que hay una relación lineal entre los variables X y Y.

Al hablar de la correlación lineal de dos variables es necesario distinguir dos casos: Correlación Positiva y Correlación Negativa.

Correlación Positiva. Ocurre cuando al crecer (o decrecer) una de las variables, la otra también crece (o decrece). Por ejemplo: a medida que se eleva el nivel de vida de una población, tiende a aumentar el consumo de artículos que no son de primera necesidad.

Correlación Negativa. Ocurre cuando al crecer alguna de las variables, la otra decrece o viceversa. Por ejemplo: a medida que se amplían los sistemas de salubridad y medicina preventiva, decrece el índice de mortalidad de las enfermedades infecto-contagiosas.

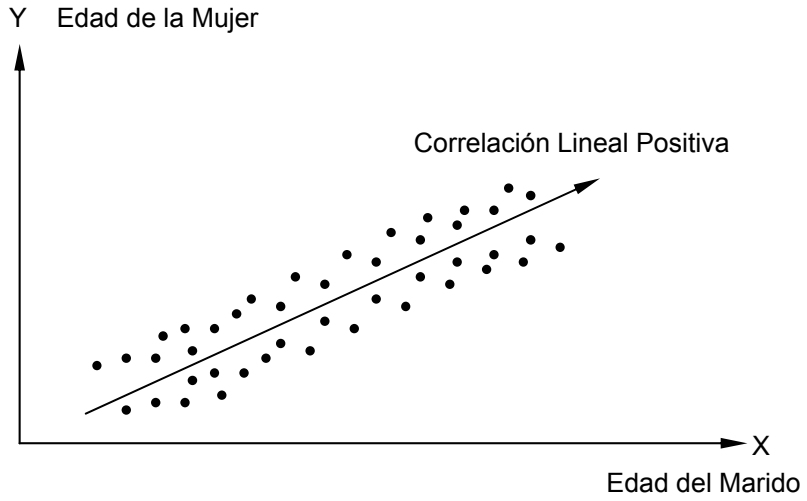
En el ejemplo anterior (las ventas) tenemos una correlación positiva. Estas dos correlaciones y otras más, se pueden mostrar utilizando los Diagramas de Dispersión, de los que nos ocuparemos enseguida.

DIAGRAMAS DE DISPERSIÓN

La forma más sencilla que tienen para predeterminar si existe o no correlación entre dos variables es construir un diagrama de dispersión.

Para construir un diagrama de dispersión tienes que utilizar un sistema de coordenadas rectangulares, el cual aprendiste en los fascículos de Matemáticas I, II y IV, ¿lo recuerdas?, bien. El sistema de coordenadas rectangulares, en el eje X (abscisas), es donde se marca una escala adecuada para registrar los valores de una de las variables y sobre el eje Y (ordenadas), se marca otra escala adecuada para representar o registrar los valores de la otra variable. Los dos valores de las variables forman pares ordenados (X, Y) dispersos en dicho sistema de coordenadas rectangulares. Esta dispersión de los pares ordenados deben de sugerir una línea recta, (de aquí el nombre de correlación lineal) como lo muestra el diagrama de dispersión del ejemplo anterior. La dispersión de estos puntos tienen las siguientes formas generales:

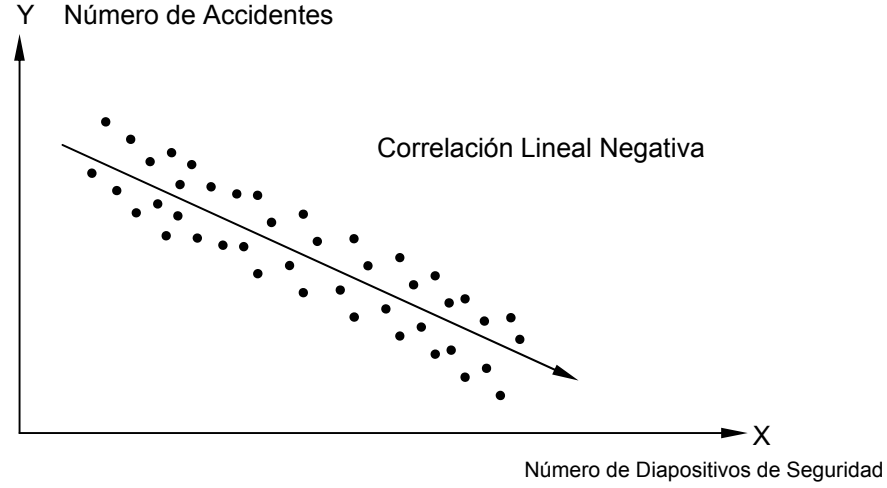
a) Cuando los puntos se van localizando en los ejes coordenados de manera que veas que si los valores de la variable X aumentan y los valores de la variable Y también aumentan, entonces existe una Correlación Lineal Positiva. Un ejemplo así ocurre al correlacionar las edades del marido y de la mujer en las parejas conyugales. En este caso a mayor edad del marido, mayor edad de la mujer.



Gráfica No. 2

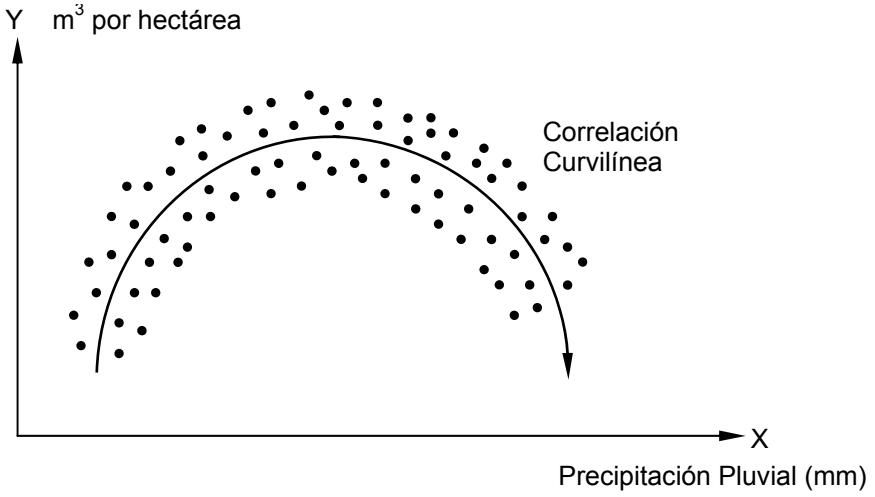
Como vemos en el diagrama de dispersión anterior, conforme la edad del marido (X) aumenta, aumenta la edad de la mujer (Y), por lo que tendremos una correlación lineal positiva.

b) Si los puntos se localizan en los ejes coordenados y observas que los valores de la variable X aumentan mientras que los valores de la variable Y decrecen, entonces existe una Correlación lineal negativa. Un ejemplo así ocurre al correlacionar el número de accidentes de trabajo acaecidos en un periodo de tiempo, con el número de dispositivos de seguridad operantes en las plantas de una industria. En este caso a mayor número de dispositivos de seguridad, menor número de accidentes de trabajo.



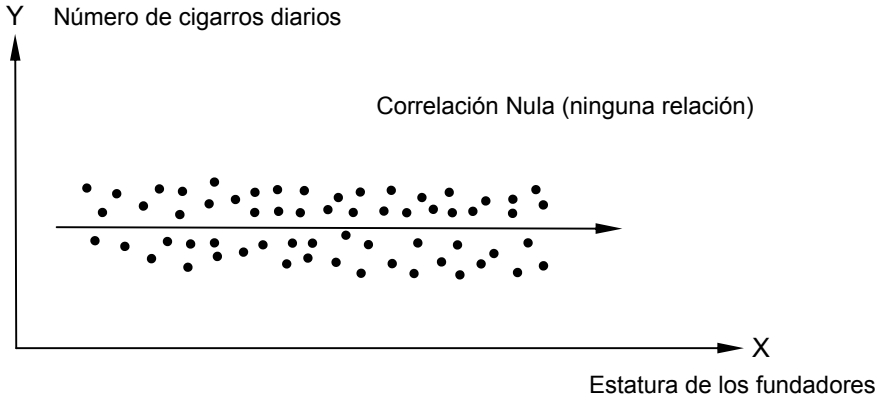
Gráfica No. 3

c) Cuando los puntos se localizan en el eje de coordenadas y observes que su relación no es lineal, es decir, aunque su patrón de dispersión está definido, estas variables presentan una relación no lineal. Por ejemplo: al correlacionar la cantidad de lluvia caida y el rendimiento de ciertos productos agrícolas, que es afectado desfavorablemente tanto por la excesiva sequía, como por la humedad excesiva del suelo, se tiene una correlación que se denomina Correlación Curvilínea.



Gráfica No. 4

d) Cuando los valores de X tienen la misma probabilidad de aparecer aparcadas con valores de Y o con valores pequeños de Y, decimos que no hay relación entre X y Y. Por ejemplo: ¿habrá alguna relación entre la estatura de los que fuman cigarros, con el número de cigarros que fuman a diario? No. entre estas dos variables (estatura de fumadores y números de cigarros que fuman diariamente) no existe relación.



Gráfica No. 5

Los diagramas de dispersión que acabas de ver te muestran las diferentes relaciones entre la variable independiente (X) y la variable dependiente (Y), por lo que podemos señalar que si tanto los valores de X como los valores de Y tienden a seguir un patrón recto, entonces existe una correlación lineal.

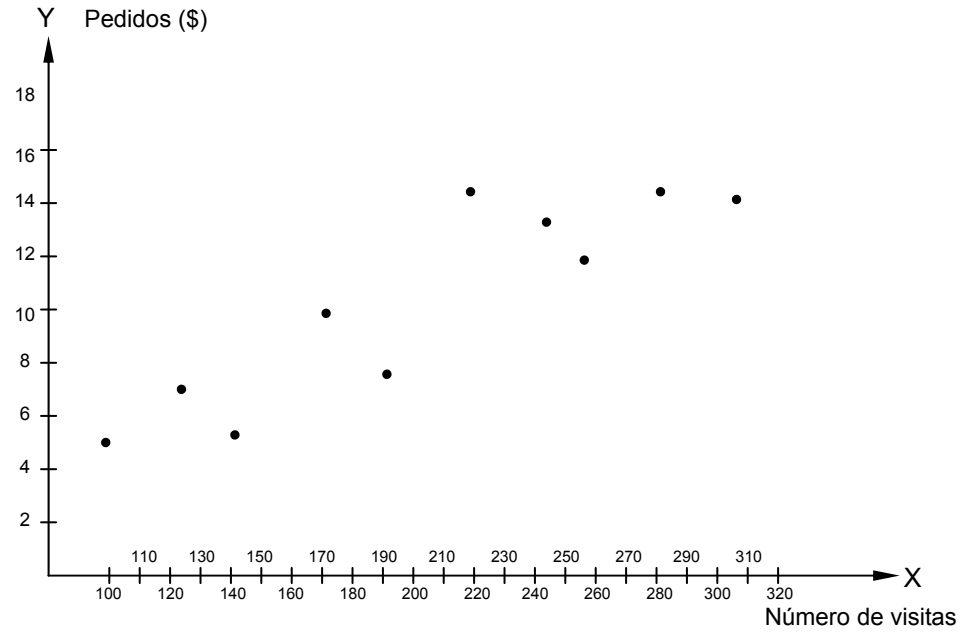
Para mostrar estos tipos de diagramas de dispersión y recordando cómo se localizan los puntos o parejas ordenadas en los ejes cartesianos, te invito a que resuelvas gráficamente los problemas que a continuación mencionamos e infieras algún tipo de correlación.

Ejemplo: El Departamento de Ventas de una empresa realiza un análisis comparativo entre el volumen de pedidos levantados y el número de visitas efectuadas por sus 10 vendedores en un cierto periodo de tiempo. Todos los vendedores trabajan en zonas similares, en lo referente al número de clientes y al potencial de compra de dichos clientes. Los resultados de la comparación se muestran a continuación:

Considera el número de visitas como la variable (X) y el monto de los pedidos como la variable (Y), construye el diagrama de dispersión correspondiente e infiere si existe algún tipo de correlación.

Vendedor Número	Visitas Realizadas (X)	Pedidos en Millones (N$) (Y)
1	245	13.4
2	172	10.3
3	291	15.1
4	124	6.9
5	191	7.3
6	218	14.2
7	101	5.2
8	259	11.28
9	307	14.3
10	142	5.5

Solución: La tabla de valores nos proporciona los pares para localizarlos en los ejes, como se muestra en la siguiente gráfica. Verifica estas localizaciones.



Gráfica No. 6

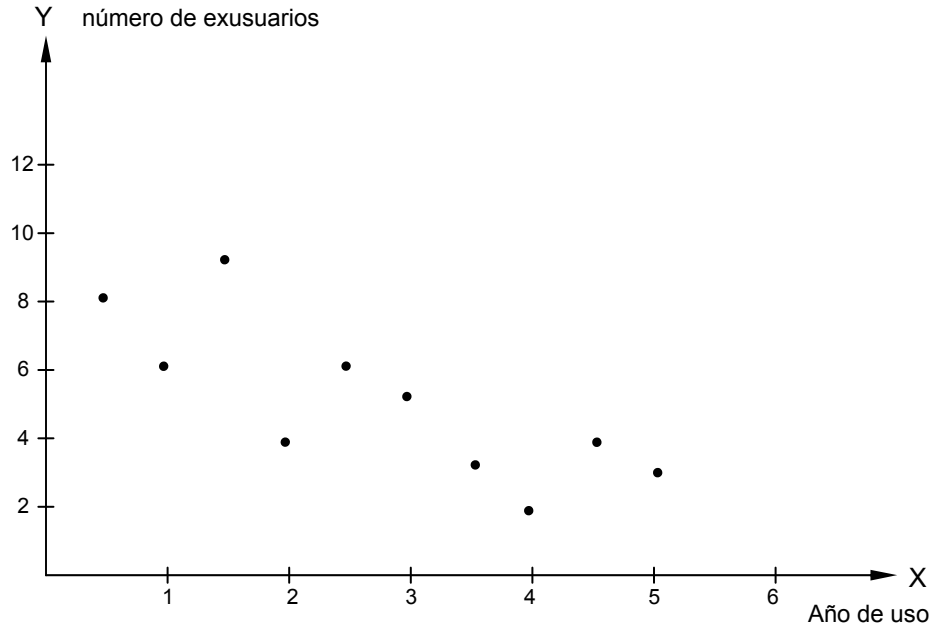
El diagrama de dispersión indica que existe una correlación lineal positiva, ¿sabes por qué?

La construcción de diagramas de dispersión es sencilla, si consideras que tienes antecedentes de este conocimiento desde Matemáticas I. Ahora, el siguiente ejemplo te brinda la oportunidad para que tú construyas la gráfica correspondiente e indiques qué tipo de correlación tiene.

Ejemplo: Al efectuarse un estudio sobre la marca de cierto producto se encontró que 50 personas habían usado anteriormente dicha marca y la habían cambiado. La relación entre el tiempo que habían usado la marca, antes de sustituirla por otra, y el número de exusuarios en cada caso, fue:

Años de Uso (X)	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
Número de Exusuarios (Y)	8	6	9	4	6	5	3	2	4	3

Solución:



Gráfica No. 7

La tabla del ejemplo te facilitó la localización de los puntos en los ejes y confirmaste que existe una correlación lineal negativa. A estas alturas te puedes dar cuenta de la facilidad con que se construye este tipo de diagramas y se reconoce el tipo de correlación que existe entre las variables.

Te recomiendo realices tú solo el siguiente ejemplo, inténtalo y estoy seguro que lo lograrás. Si tienes alguna duda, acude con tu profesor o asesor.

Ejemplo: Para apoyar la venta de un producto de consumo masivo en un mercado altamente competitivo, una empresa inició a comienzos de año una intensa campaña publicitaria y promocional. La comparación entre la inversión publicitaria y las ventas del producto en 12 meses se indican en la siguiente tabulación:

Mes	Publicidad (X) (miles de N$)	Ventas (Y) (miles de N$)
Enero	200	350
Febrero	250	300
Marzo	300	630
Abril	250	840
Mayo	330	930
Junio	180	1060
Julio	150	1280
Agosto	350	850
Septiembre	240	700
Octubre	250	1160
Noviembre	230	910
Diciembre	170	1500

Construye el diagrama de dispersión e indica si existe alguna correlación entre las variables. ¿De qué tipo es la correlación?

Para que reafirmes cómo se construye un diagrama de dispersión y los tipos de correlación que puedes inferir, es aconsejable que realices los siguientes ejercicios y si acaso tuvieses dudas, acude con tu profesor o asesor para que te puedan orientar.

Ejercicios: Para cada uno de los siguientes enunciados, dibuja un diagrama de dispersión e infiere qué tipo de correlación existe.

1) La siguiente tabla muestra los puntajes obtenidos en satisfacción en el trabajo y los puntajes que obtuviste en una prueba de aptitud al iniciar sus estudios universitarios en medicina algunos estudiantes.

Puntaje de satisfacción (Y)	58	54	67	64	66	73	70	85	74	85
Puntaje de aptitud (X)	50	55	60	65	70	75	80	85	90	95

2) La siguiente tabla muestra el peso de 11 ovejas y el peso de sus madres a la misma edad.

Puntaje de la Oveja (Y)	68	63	70	66	81	74	82	76	81	92	85
Paso de la Madre (X)	60	64	68	72	76	80	84	88	92	96	100

3) La siguiente tabla muestra el número de horas por semana que estudiaron diez universitarios y su promedio de calificaciones acumulativas.

Promedio de Calificaciones (Y)	2.1	2.7	2.6	2.5	3.5	3.0	3.5	3.7	2.9	4.0
Horas de Estudio (X)	5	6	7	8	9	10	11	12	13	14

4) La siguiente tabla muestra los siguientes datos de 11 trabajadores de una empresa, el tiempo en minutos requeridos para completar una tarea y el número de minutos invertido en aprender la tarea.

Tiempo gastado en aprender (X)	30	30	40	40	50	50	60	60	60	70	70
Tiempo para hacer la tarea (Y)	45	35	20	38	17	26	28	22	12	12	5

5) La siguiente tabla muestra los resultados de una prueba para medir el nivel de seguridad en sí mismo y de otra prueba para medir el nivel de madurez social de 15 estudiantes de preparatoria.

Puntaje de seguridad en sí mismo (Y)	5	10	15	15	20	20	25	25	25	32	40	37	45	35	50
Puntaje de madurez social (X)	5	5	8	20	15	25	20	35	30	30	30	35	35	40	40

Recordemos que la obtención de datos para un análisis estadístico es un proceso integral que incluye las siguientes etapas:

a) Definición de los objetivos del estudio del experimento.

b) Definición de la variable y la población de interés.

c) Definición de los métodos para la obtención y la medición de los datos.

d) Determinación de las técnicas descriptivas o inferenciales que sean apropiadas para el análisis de datos.

Se sugiere para la recopilación de un conjunto de datos, se emplee técnicas que uno mismo utilice.

La descripción gráfica se realiza mediante el diagrama de dispersión, el cual se construye localizando los pares ordenados en el plano cartesiano. No olvides que la disposición de los puntos en el plano X Y sugiere también el tipo de correlación entre las variables de estudio. Con este tipo de diagramas y con el cálculo del coeficiente de correlación r de Pearson, podemos decidir si la correlación es positiva ($r > 0$) y negativa ($r < 0$) o nula ($r = 0$).

COEFICIENTE DE CORRELACIÓN

Ahora que has aprendido a construir los diagramas de dispersión y a identificar cuándo hay correlación (positiva y negativa), y cuándo no hay, podemos empezar a estudiar cómo se calcula el Coeficiente de Correlación de Pearson.

De los diversos coeficientes de correlación que existen, el más popular y utilizado es el Coeficiente de Correlación de Pearson. Para su aplicación es indispensable que la correlación sea lineal.

El coeficiente de correlación de Pearson, que se simboliza con la letra minúscula r, se calcula dividiendo la suma de los productos de las desviaciones de cada variante de X e Y, con respecto a sus medias (suma que se denomina covarianza de X e Y), por el producto de las desviaciones estándar de ambas variables. En forma práctica, el coeficiente de correlación de Pearson es:

$$r = \frac{N \sum_{i=1}^{N} (XY) - \left(\sum_{i=1}^{N} X\right)\left(\sum_{i=1}^{N} Y\right)}{\sqrt{\left[N \sum_{i=1}^{N} X^2 - \left(\sum_{i=1}^{N} X\right)^2\right]\left[N \sum_{i=1}^{N} Y^2 - \left(\sum_{i=1}^{N} Y\right)^2\right]}}$$

donde N es el número de datos.

Por medio de ejemplos, veremos cómo se utiliza esta fórmula, para que puedas hacer interpretaciones de este valor.

Ejemplo: La siguiente tabla muestra los datos registrados en una muestra aleatoria de 10 escuelas para niños superdotados. La razón alumno/maestro es (X) y los estudiantes que se salen antes de completar el curso es (Y).

X	20	18	16	15	14	12	12	10	8	5
Y	12	16	10	14	12	10	9	8	7	2

Solución: Se recomienda para hacer el cálculo directo del coeficiente r de Pearson, realizar una tabla como la siguiente:

(1) X	(2) Y	(3) X^2	(4) Y^2	(5) XY
20	12	400	144	240
18	16	324	256	288
16	10	256	100	160
15	14	225	196	210
14	12	196	144	168
12	10	144	100	120
12	9	144	81	108
10	8	100	64	80
8	7	64	49	56
5	2	25	4	10

$\sum X = 130 \quad \sum Y = 100 \quad \sum X^2 = 1878 \quad \sum Y^2 = 1138 \quad \sum XY = 144$

De la tabla, ves que en las columnas (1) y (2) se han escrito las puntuaciones originales. En la columna (3) se obtuvieron los cuadrados de las puntuaciones X y en la columna (4) los cuadrados de las puntuaciones Y. La columna (5) se forma con el producto de cada X por cada Y, finalmente se suman los valores de las cinco columnas y se sustituyen en la fórmula que ya conoces, obteniendo el siguiente resultado.

$$r = \frac{N \sum_{i=1}^{N} (XY) - \left(\sum_{i=1}^{N} X\right)\left(\sum_{i=1}^{N} Y\right)}{\sqrt{\left[N \sum_{i=1}^{N} X^2 - \left(\sum_{i=1}^{N} X\right)^2\right]\left[N \sum_{i=1}^{N} Y^2 - \left(\sum_{i=1}^{N} Y\right)^2\right]}}$$

$$r = \frac{10\,(1440) - (130)(100)}{\sqrt{\left[10\,(1878) - (130)^2\right]\left[10\,(1138) - (100)^2\right]}}$$

$$r = \frac{14400 - 13000}{\sqrt{(18780 - 16900)(11380 - 10000)}} = \frac{1400}{\sqrt{(1880)(1380)}}$$

$$r = \frac{1400}{\sqrt{2594400}}$$

$$r = \frac{1400}{1610.7141} = 0.869180$$

Ahora interpretaremos este valor. Para ello es necesario conocer las siguientes características del coeficiente de correlación lineal.

- El valor de r es un número que satisface la desigualdad $-1 \leq r \leq 1$.

- Cuando 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 (unidad positiva).

- Cuando 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 (unidad positiva).

- Cuando no existe relación entre las 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 lineal es cero.

Lo anterior significa que, entre 0 y +1 cabe toda una gama de correlaciones positivas, que serán tanto más directamente proporcionales, cuanto más se acerquen a +1. Similarmente entre –1 y 0 cabe toda una gama de correlaciones negativas, que serán tanto más inversamente proporcionales, cuanto más se acerquen a –1. Los coeficientes de correlación, cuanto más cerca de cero, indican menor correlación.

Con todas estas características, podemos interpretar el resultado que calculamos del coeficiente r de Pearson. Como r = 0.869180 podemos concluir que la correlación es fuerte y positiva.

Con base a las características del coeficiente de correlación lineal (r) de Pearson, se muestra a continuación una tabla que indica cuándo una correlación lineal es débil, fuerte, positiva o negativa.

Tabla Significado de

	Tendencia del agrupamiento con respecto a la línea de regresión.
R = 0	Correlación nula
0 < r << 1,	Correlación baja positiva
1 – r << 1	Correlación alta positiva
0 < r << 1, < 0	Correlación baja negativa
1 + r << 1	Correlación alta negativa

Como puedes observar, lo único tedioso es la tabla, pero ésta concentra los resultados para obtenerlos con cierta facilidad. Te invito a que resuelvas el siguiente ejemplo sin ver los resultados, salvo te aparezcan dudas, ¡inténtalo!

Ejemplo: Retomemos los valores utilizados del ejemplo de las visitas realizadas y los pedidos hechos por diez vendedores de un Departamento de Ventas, ¿lo recuerdas?, te mostraré la tabla de valores que utilizamos; calcula el coeficiente r de Pearson.

Vendedor Número	Visitas realizadas (X)	Pedidos en millones (N$) (Y)
1	245	13.4
2	172	10.3
3	291	15.1
4	124	6.9
5	191	7.3
6	218	14.2
7	101	5.2
8	259	11.8
9	307	14.3
10	142	5.5

X – Y	X^2	Y^2
3283.00	60025	179.56
1771.60	29584	106.09
4394.10	84681	228.01
855.60	15376	47.61
1394.30	36481	53.29
3095.60	47524	201.64
525.20	10201	27.04
3056.20	67081	139.24
4390.10	94249	204.49
781.00	20164	30.25

Solución: Recuerda que para facilitar este cálculo, se puede elaborar una tabla para mostrar los totales, la cual está a continuación de la tabla de datos, como observas.

La suma de las visitas realizadas es: $\sum X = 2050$

La suma de los pedidos hechos es: $\sum Y = 104$

La suma del producto de (X) por (Y) es: $\sum X\,Y = 23546.70$

La suma de los cuadrados de (X) es: $\sum X^2 = 465366$

La suma de los cuadrados de (Y) es: $\sum Y^2 = 1217.22$

Ahora procedemos a sustituir en la fórmula del coeficiente de correlación de Pearson, r:

$$r = \frac{10\,(23546.6) - 213200}{\sqrt{\left[10\,(465366) - 4202500\right]\,\left[10\,(1217.22) - 10816\right]}} = 0.9$$

Si te apoyas en la tabla del significado de r, ves que existe un grado apreciable de correlación entre las visitas y los pedidos, y ésta resulta ser positiva.

En el cálculo de r se omitieron algunos procedimientos para crear la necesidad en ti de hacerlo completo e ir aclarando posibles dudas que pudieran surgir. Si no lo entendiste después de haberlo hecho de nuevo, revisa el primer ejemplo del cálculo del coeficiente de correlación r de Pearson.

Ejercicios: Calcula el coeficiente de correlación r de Pearson para los siguientes problemas.

6) Para poder medir los resultados de un curso de capacitación realizado con 12 técnicos de una empresa, se tomó un examen teórico antes de comenzar el curso y se realizó una prueba teórica-práctica al final del curso. La calificación máxima de cada una de dichas pruebas fue de 10 puntos. El grupo estuvo compuesto por 6 técnicos recientemente ingresados a la empresa (No. 1 al 6) y 6 técnicos con mayor antigüedad (No. 7 al 12). Los resultados de ambas pruebas fueron:

No.	Prueba Previa	Prueba Final
1	6.0	6.5
2	4.0	5.5
3	3.0	7.0
4	5.0	5.0
5	6.0	7.0
6	4.0	6.5
7	7.0	10.0
8	4.0	5.0
9	6.5	9.0
10	5.5	7.0
11	6.0	8.5
12	5.0	6.0

Con estos resultados calcula los coeficientes de correlación r de:

a) Todo el grupo.

b) El grupo de recién ingresados.

c) El grupo de mayor antigüedad.

¿Qué conclusiones obtienes de los incisos anteriores?

7) En dos tests, diez alumnos obtuvieron las siguientes puntuaciones:

Alumno	Test 1	Test 2
1	15	12
2	14	14
3	10	9
4	9	10
5	8	8
6	8	7
7	7	8
8	6	4
9	4	6
10	2	4

¿Cuál es el coeficiente de correlación r de Pearson? Interprétalo.

8) A veinte estudiantes se les aplica un test de capacidad mental y otro sobre conocimientos de francés. Se obtuvieron las siguientes puntuaciones.

Capacidad Mental	Francés
54	203
53	196
51	202
50	186
48	204
47	184
47	196
46	182
45	170
45	178
44	181
44	175
44	168
43	174
40	162
38	158
37	170
36	144
34	141

Calcula el coeficiente de correlación r de Pearson e interprétalo.

9) La siguiente tabla muestra los valores obtenidos en asistencia a juntas tanto para hombres como para mujeres.

Asistencia a juntas

Hombres (X)	Mujeres (Y)
10	8
10	7
9	7
9	6
8	5
7	6
7	5
7	4
6	4
6	3
5	4
5	3
4	4
4	3
3	2

¿Cuál es el coeficiente de correlación r de Pearson? Interprétalo.

10) Los siguientes pares de valores representan las dimensiones en cms. de las hojas del árbol del fresno:

(2,1), (3,2), (2,3), (3,3), (4,3), (3,5), (5,5) y (6,7)

Si se conoce la anchura (X) y la longitud (Y) de las hojas, ¿habrá alguna relación entre estas dos variables?, y si la hay, ¿ésta es fuerte o débil? Realiza los cálculos adecuados para que contestes estas preguntas.

REGRESIÓN LINEAL

Ahora que has analizado el grado de relación que existe entre dos variables estadísticas (datos bivariados), a través del cálculo del coeficiente de correlación de Pearson, es importante dar un contexto adecuado al tema de **Regresión Lineal**, con el objeto de ubicar correctamente algunos de los conceptos que se utilizarán en el proceso de predicción estadística. Es probable que hayas escuchado una expresión tan popular como "para muestra basta un botón", que ilustra muy bien lo que sucede en la inferencia estadística. El proceso inferencial consiste en obtener información acerca de una **Población** de objetos cuantitativos (datos), a partir de información contenido en una parte de esta población llamada **Muestra**. Cabe preguntarnos ¿por qué no utilizar todos los datos de una Población? Pongamos por ejemplo que un especialista desea información acerca de las dimensiones de las alas de la mariposa Monarca que anualmente hace una emigración desde Canadá hasta México. ¿Será posible estudiar todas y cada una de las mariposas monarcas que llegan cada año a nuestro país? Desde luego que no, pues ello implica un enorme gasto de recursos humanos y materiales entre otros, cosa que haría prácticamente imposible el estudio. Para llevar adelante su investigación el especialista tomaría una muestra de la población, mediría y analizaría estadísticamente los datos que le interesan y apoyándose en un modelo matemático adecuado trataría de deducir las características esenciales de toda la población de mariposas. Este modo de proceder del especialista lo realizamos todos cotidianamente, aunque no de manera tan rigurosa. Por ejemplo, una ama de casa en el supermercado quiere comprar naranjas y sabe por experiencia que no siempre las más grandes son las más jugosas, escoge unas cuantas para observar su peso, consistencia, madurez y si es posible prueba una de ellas, sólo después de hacer estas operaciones toma una decisión. Al hacerlo no fue necesario que probara todas las naranjas que había en el aparador o en la bodega o en la huerta del productor que provee al supermercado, sólo le bastó una muestra.

En los ejemplos siguientes, se hará referencia a muestras de datos, esperamos que con la explicación anterior logres observar que éstas forman parte de poblaciones más grandes.

Hablemos ahora de la **Regresión Lineal**. En primer lugar nos surgen interrogantes como:

- ¿Qué es la regresión lineal?
- ¿En qué consiste el análisis de represión?
- ¿Qué técnicas se utilizan en el análisis?
- ¿Qué relación existe entre el análisis de regresión y la predicción estadística?
- ¿Existe una diferencia cuantitativa entre lo observado y lo predicho?

Estas y otras preguntas tratarán de ser contestadas en los siguientes párrafos, a fin de que puedas usar el modelo estadístico de Regresión Lineal para hacer deducciones o predicciones estadísticas. Las respuestas a cada una de ellas si bien no serán definitivas sí serán válidas para nuestro análisis, mismo que deberá ser ampliado y profundizado en estudios posteriores.

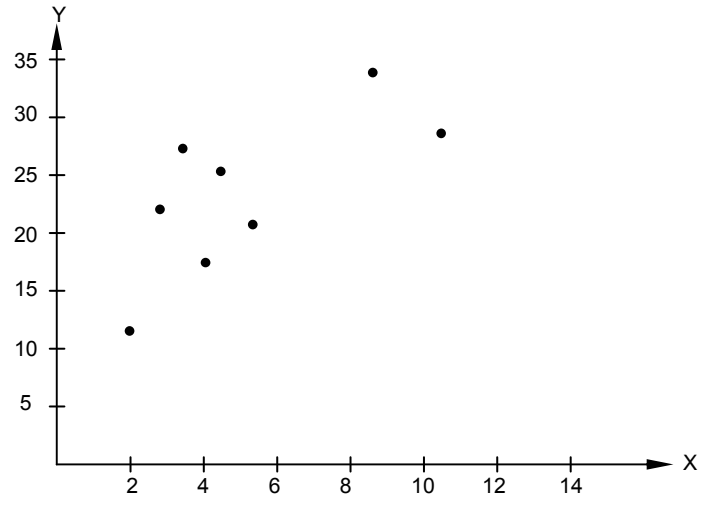
Dentro de las aplicaciones de la estadística, podemos encontrar problemas que tienen que ver con procesos de planeación en la administración de recursos materiales y humanos, tal es el caso del ejemplo que a continuación te presentamos.

Una compañía comercializadora desea contratar vendedores, para lo cual se ha seleccionado una muestra de ocho aspirantes, tomando en cuenta dos parámetros de selección que pueden servir de referencia para tomar una decisión sobre otros aspirantes. Dichas variables son: los años de experiencia (X) y el monto de ventas promedio (Y). Los datos se incluyen en la tabla de valores siguiente:

VENDEDOR	AÑOS (x)	MONTO EN MILES N% (Y)
1	2	12
2	4	18
3	5	25
4	3	23
5	4	27
6	6	19
7	20	32
8	12	26

Tabla

El conjunto de datos que incluye la tabulación, los llevaremos al plano cartesiano para obtener la gráfica siguiente:



Gráfica No. 8

Los valores de las variables X y Y forman parejas ordenadas (x,y) susceptibles de ser graficadas en el plano cartesiano. Al exhibir gráficamente los datos de la tabla No. 1 obtenemos el **Diagrama de Dispersión**. De la tabulación se puede considerar que al haber pares ordenados (x,y), teóricamente puede existir una relación Funcional entre las variables X a la que llamaremos variable independiente y Y a la que llamaremos variable dependiente suponiendo que el problema es saber ¿cómo varía Y en función de X? Para hacer esto más claro, te pedimos que apoyándote en la tabulación y en la gráfica escribas en el siguiente cuadro ¿cuánto esperarías que vendiera un aspirante con tres años de experiencia?, ¿cuánto si tiene siete u ocho años en ventas?

Vendedor con tres años en ventas: _____

Vendedor con siete años en ventas: _____

Vendedor con once años en ventas: _____

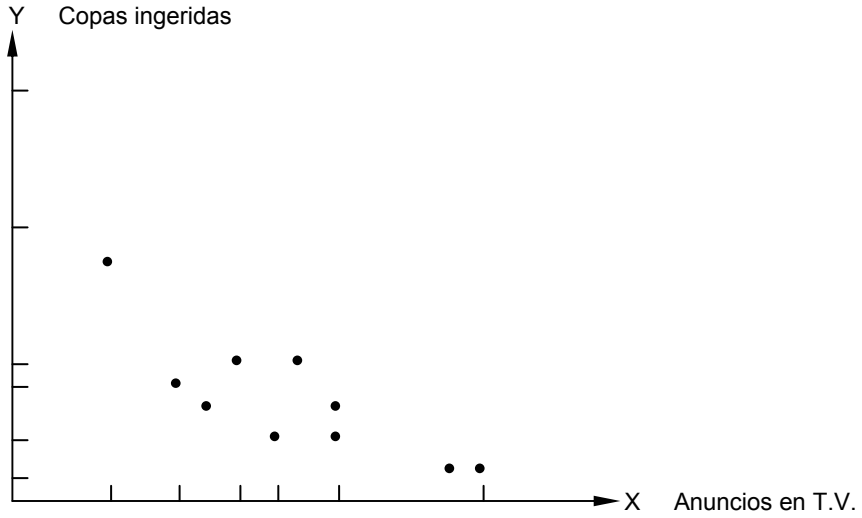
Como te habrás dado cuenta, lo que hiciste para contestar las preguntas anteriores fue apoyarte en la observación de datos conocidos y en tu experiencia, es decir, has hecho una estimación empírica a partir de cierta información estadística. Esta forma de proceder ha sido la base del desarrollo de la estadística moderna, pues de esa manera, los procesos prospectivos o de planeación a futuro tienen una fundamentación teórica basada en observaciones hechas con anterioridad. Volveremos a este ejemplo para proponer un método general de análisis, que nos permita hacer predicciones estadísticas consistentes. Pero ahora te pedimos que analices el siguiente caso donde encontrarás nuevas interrogantes.

Una Empresa de publicidad, ha sido contratada para llevar a cabo una campaña para disminuir el consumo de bebidas alcohólicas entre la juventud. Los planificadores de la empresa estiman que el consumo disminuirá si incrementan el número de anuncios televisivos con el eslogoan “sin alcohol la vida es más placentera”. Para verificar esta hipótesis toman una muestra de diez personas al azar y hacen una encuesta que arroja los siguientes resultados:

NOMBRE	EDAD (AÑOS)	No. ANUNCIOS VISTOS EN T.V.	No. COPAS INGERIDAS EN UNA FIESTA
Jorge	18	3	8
Andrés	19	5	4
Carlos	21	7	5
Sandra	16	10	3
Martha	22	6	3
Ruth	18	10	2
Juan	17	14	1
Pedro	23	9	5
Raúl	19	8	2
Claudia	22	15	1

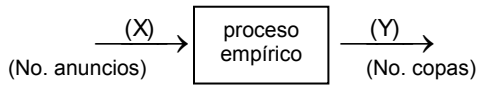
Tabla No. 9

Construye el diagrama de dispersión correspondiente a los valores tabulares tomando a "x" (variable independiente) como el número de anuncios de T.V. y a "y" (variable dependiente) como el número de copas ingeridas por persona y compáralo con el que a continuación te mostramos.



Gráfica No. 10

Por el texto del problema, nos percatamos de que los planificadores de esta Empresa desean analizar teóricamente, la variación entre el consumo de alcohol y el número de anuncios vistos por el público, tomando como variable independiente o de entrada este número de anuncios (X) y como variable dependiente o de salida el número de copas de bebida ingeridas en una fiesta (Y). Ilustramos esto mediante el siguiente esquema:



Esquema No. 1

Ahora contesta las preguntas siguientes apoyándote tanto en la tabulación como en el diagrama de dispersión:

1. ¿Estás de acuerdo con los planificadores, de que la campaña publicitaria influirá para que el público joven disminuya su consumo de alcohol? Explica.

   _____

2. ¿Se puede aumentar indefinidamente el número de anuncios para garantizar que una mayor población consuma menos alcohol? Explica.

   _____

3. ¿Teóricamente es posible alcanzar el objetivo de eliminar absolutamente el consumo de alcohol entre la juventud que ha visto el anuncio publicitario? Explica.

_____

4. ¿Qué método propones para comprobar el impacto del anuncio publicitario con relación al consumo de bebidas alcohólicas? Explica.

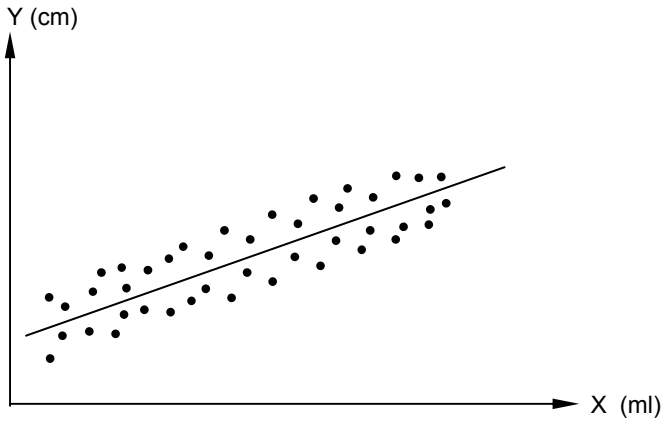
_____

Al contestar las preguntas anteriores, habrás observado que lo que teóricamente es posible, en la práctica no es tan inmediato, es decir, es probable que estadísticamente exista una relación entre las variables, pero, eso no quiere decir que existe necesariamente una relación causa-efecto entre ellas, por lo que, se sugiere interpretar prudentemente las observaciones derivadas del análisis estadístico.

Este ejemplo nos coloca en el centro de la discusión acerca de cómo predecir un evento, en este caso el número de copas ingeridas (Y) en términos del número de anuncios vistos por una persona. Surge la necesidad de encontrar un modelo teórico para realizar predicciones estadísticas, que nos permita a la vez comparar nuestras observaciones empíricas con respecto a dicho modelo. Para que sea útil, el modelo en cuestión, deberá poseer ciertas características entre las cuales se deben contar su sencillez en el manejo y su eficacia para hacer predicciones estadísticas. El comentario anterior nos pone en evidencia un punto medular en el análisis, que consiste en colocar nuestras observaciones empíricas a la luz de un modelo estadístico teórico al que llamaremos **CURVA DE REGRESIÓN o CURVA DE PREDICCIÓN o también CURVA DE MEJOR AJUSTE**.

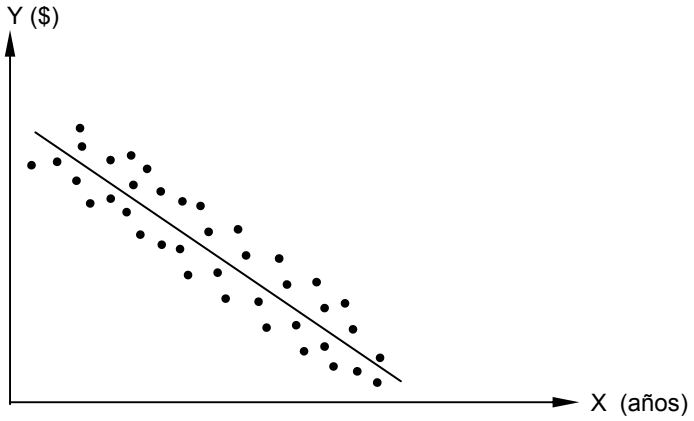
El párrafo anterior, nos indica que el objetivo primordial en el análisis de **Regresión** es encontrar la **Curva de regresión** para que realicemos con ella predicciones y observemos que para cada valor (Y) registrado en la tabulación existe un valor de predicción, $\hat{y}$, que pertenece a la curva. La sola presencia en el diagrama de dispersión de una **Curva de regresión** nos conduce a preguntarnos entre otras cosas ¿cuál es la ecuación algebráica o trascendente que define a esta curva? ¿Cómo saber si esta curva es la que ofrece las mejores predicciones estadísticas? Observa en seguida algunas gráficas de dispersión que incluyen diferentes **Curvas de Regresión**.

1. **BIOLOGÍA**. El crecimiento de una cierta especie de alga marina al aplicarle cierta dosis de líquido proteínico.



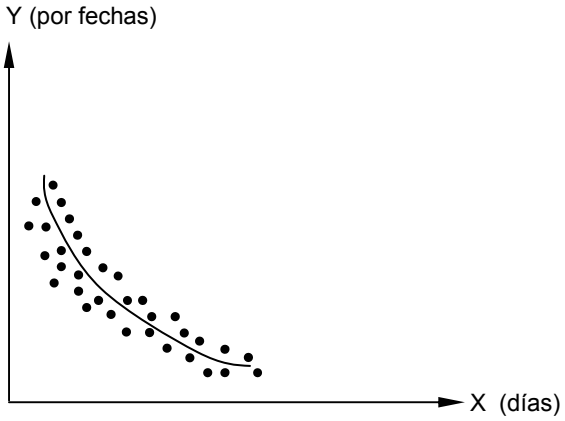
Gráfica No. 11

2. **ECONOMÍA**. Los años de antigüedad de un automóvil y su valor de reventa.



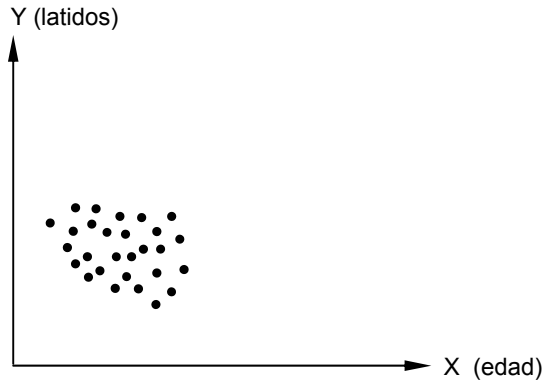
Gráfica No. 12

3. **PSICOLOGÍA**. La cantidad de fechas memorizadas-recordadas por un sujeto y el número de días transcurridos.



Gráfica No. 13

4. **MEDICINA**. El ritmo cardiaco de un espectador de basquetbol y la edad de los jugadores de su equipo favorito.



Gráfica No. 14

Como ya se mencionó, las curvas trazadas sobre el diagrama de dispersión son llamadas Curvas de ajuste y como se puede notar tienen distintas formas geométricas dependiendo del tipo de modelo que la define. Así por ejemplo, tenemos que si $\hat{y} = f(x)$ es la ecuación de predicción, entonces:

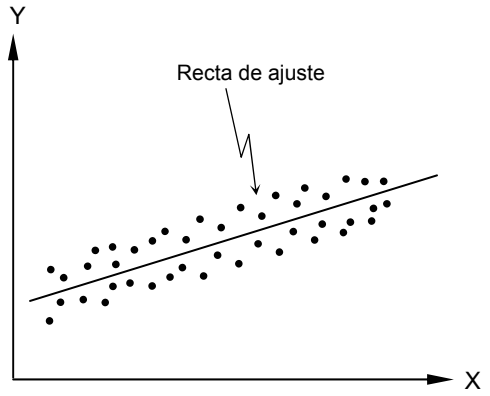
$f(x) = a + bx$ es **lineal**.

$f(x) = ax^2 + bx + c$ se llama **cuadrática**.

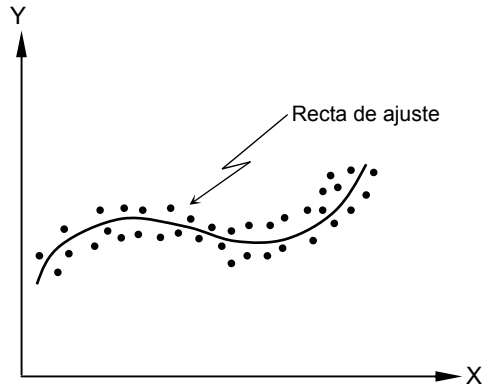
$f(x) = a(b^x)$ se llama **exponencial**.

$f(x) = a \log_b x$ es **logarítmica**.

Si se trata del modelo lineal, entonces la gráfica es una recta a la que llamaremos: **Recta de ajuste** o **Recta de regresión**. En todo caso, los puntos registrados en el diagrama de dispersión sugieren el tipo de función de regresión que se debe utilizar. Ver las siguientes figuras:



Gráfica No. 15



Gráfica No. 16

Desde luego que encontrar la expresión de esta función, no siempre es sencillo, por lo que, se propone el modelo de la ecuación lineal:

$$y = a + bx$$

como una buena alternativa de solución al problema de la predicción estadística. Por cierto, recuerdas ¿cuáles son los parámetros que determinan la función lineal, en este caso a y b? Si no es así coméntalo con tu profesor o asesor.

Es tiempo de contestar las preguntas básicas, ¿cómo encontrar las rectas de ajuste para un problema en particular?, ¿qué criterio se debe utilizar para asegurar la recta de mejor ajuste?

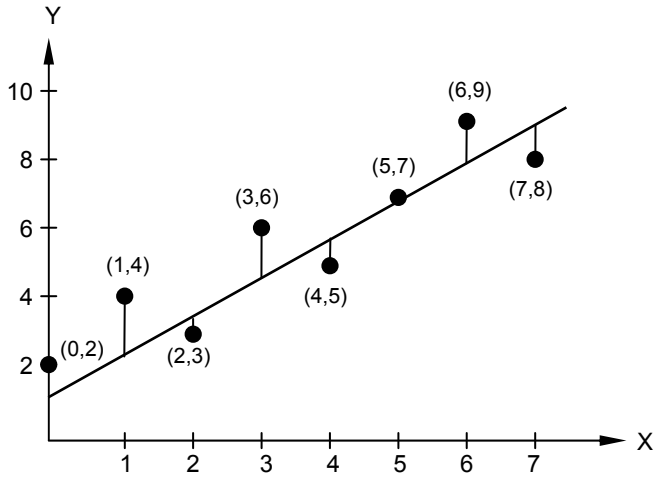
El ejemplo siguiente, nos muestra un método de trazo rápido (“mano alzada”) de la recta de ajuste sobre el diagrama de dispersión. No olvidemos que al trazar la recta, ésta coincidirá con algunos puntos pero en general habrá puntos que se encuentren arriba o debajo de la recta. Observa la gráfica del siguiente ejemplo.

Ejemplo No. 3

Dibuja en el plano cartesiano un diagrama de dispersión con los datos x,y de la tabulación dada. Sobre el diagrama de dispersión traza una recta que incluya los datos si es posible, si no es así, trata de minimizar las distancias entre la recta y los puntos tabulados. Mide la distancia entre cada punto (x,y) de la tabulación y su correspondiente punto de predicción (x,y) que pertenece a la recta. Observa la figura.

X	Y
0	2
1	4
2	3
3	6
4	5
5	7
6	9
7	8

Tabla

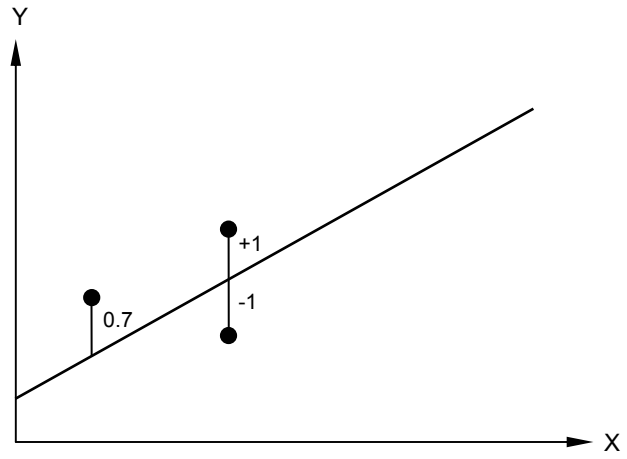


Gráfica No. 17

Es adecuado hacer las convenciones prácticas siguientes:

- La recta de ajuste tiene como ecuación $\hat{y} = a + bx$
- Si el punto se encuentra arriba de la recta la distancia será positiva.
- Si el punto se encuentra debajo de la recta la distancia será negativa.

Esto se ilustra a continuación.



Gráfica No. 18

¿Cuánto resultó la suma de las distancias que mediste? _____

¿Puede disminuirse la suma de las distancias que hay entre los puntos y la recta de ajuste? Explica. _____

_____

Cabe mencionar, que la recta trazada puede no ser la de mejor ajuste, entonces ¿cómo encontrar la de mejor ajuste? Analicemos qué pasa si las distancias $(y - \hat{y})$ son tan pequeñas como sea posible, es decir, que estas distancias estén cerca de cero. ¿Cómo varía el cuadrado de la diferencia cuando ésta tiende a cero? Observa los siguientes ensayos hipotéticos.

Si $(y - \hat{y}) = 0.25$ entonces $(y - \hat{y})^2 = (0.25)^2 = 0.0625$

Si $(y - \hat{y}) = 0.12$ entonces $(y - \hat{y})^2 = (0.12)^2 = 0.0144$

Si $(y - \hat{y}) = 0.6$ entonces $(y - \hat{y})^2 = (0.06)^2 = 0.0036$

Como te habrás dado cuenta, cuando las diferencias $(y - \hat{y})$ son cada vez más cercanas a cero, el valor del cuadrado de la diferencia también tiende a cero. Esto es muy importante, ya que si esta diferencia al cuadrado la asociamos a un cierto valor de **ERROR** en la predicción entonces decimos que la **Curva de mejor ajuste** es aquella en donde la **suma de los errores cuadráticos es mínima**. Es decir:

Si al valor $(y_i - \hat{y}_i)^2$ lo llamamos ERROR (el error es la diferencia al cuadrado entre un valor tabular (y_i) y su respectiva predicción (y) tabular (y_i) y su respectiva predicción ($\hat{y}$) entonces la curva de regresión óptima será la que cumpla con un:

$$\text{ERROR} = D = d_1^2 + d_2^2 + d_3^2 + \ldots d_n^2 \qquad \text{(Mínimo)}$$

donde: $d_i = (y_i - \hat{y}_i)^2$

Los resultados anteriores nos inducen a pensar por un lado, que existe una recta que minimiza las distancias que hay entre ésta y los puntos del diagrama de dispersión y por otro, que la diferencia entre los puntos registrados y la recta nos ofrece una medida de la "bondad" de la recta de regresión como instrumento de predicción estadística. En otras palabras, si la diferencia ($y - \hat{y}$) entre la recta y cada uno de los puntos de la tabulación es mínima entonces se tendrá un mejor modelo de predicción. Para determinar este párrafo, diremos que, a cada valor de la tabulación le corresponderá un valor de predicción obtenido por la ecuación de regresión:

$$\hat{y} = a + bx \qquad \longrightarrow \quad (1)$$

De lo anterior, tendremos los siguientes valores:

Tabulado		Predicho	Diferencia
y_1	$\longrightarrow \hat{y}_1$	$y_1 - \hat{y}_1$	$(y_1 - \hat{y}_1)^2$
Y_2	$\longrightarrow \hat{y}_2$	$y_2 - \hat{y}_2$	$(y_2 - \hat{y}_2)^2$
Y_3	$\longrightarrow \hat{y}_3$	$y_3 - \hat{y}_3$	$(y_3 - \hat{y}_3)^2$
•	$\longrightarrow$ •	•	•
		Diferencia al cuadrado	
•	$\longrightarrow$ •	•	•
y_n	$\longrightarrow \hat{y}_n$	$y_n - \hat{y}_n$	$(y_n - \hat{y}_n)^2$

Si ahora tomamos la suma de las diferencias al cuadrado para encontrar la expresión del error (D) tendremos:

$$D = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \left(y_1 - \hat{y}_1 \right)^2 + \left(y_2 - \hat{y}_2 \right)^2 + \ldots + \left(y_n - \hat{y}_n \right)^2 \longrightarrow (2)$$

Si sustituimos la ecuación de predicción $\hat{y} = a + bx$ (1) en la ecuación de error (2) tenemos:

$$D = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\left[y_i - (a + bx_i)\right]$$

$$= \sum_{i=1}^{n}\left(y_i{}^2 - a - bx_i\right) \longrightarrow (3)$$

Como te darás cuenta, los valores x_i y y_i son valores incluidos en la tabulación, por lo tanto, el error mínimo (D) sólo depende de los valores que tomen los parámetros a y b que determinan la **recta de regresión** o **predicción**. Esto nos conduce a una conclusión sorprendente, pues el problema de calcular la recta de regresión o predicción se reduce a calcular los valores de a y b para los cuales el valor del error (D) es mínimo.

Hasta aquí, hemos preparado el terreno para desarrollar el método general para encontrar la **Recta de regresión**, al que llamaremos **Método de Mínimos Cuadrados**. Retomaremos la tabulación del ejemplo No. 1, para observar cómo se calcula la recta de regresión, a la que también llamaremos: **Recta de mínimos Cuadrados**. En este cálculo utilizaremos los valores cuadráticos x^2, y^2 y xy, así como también las sumatorias correspondientes $\sum x_i$, $\sum y_i$ y $\sum x_i{}^2$ que ya habías utilizado para el cálculo del coeficiente de correlación (r).

Consideremos la tabulación donde se incluyen los datos correspondientes a los años de experiencia (X) y Monto en miles N$ de ventas (Y) de un grupo de vendedores. Se completa con los valores de X^2, y^2 y XY, además de las sumatorias ($\sum$) correspondientes.

	X	Y	X^2	Y^2	XY
	2	12	4	144	24
	4	18	16	324	72
	5	25	25	625	125
	3	23	9	529	69
	4	27	16	729	108
	6	19	36	361	114
	10	32	100	1024	320
	12	26	144	676	312
$\sum$	46	182	350	4412	1144

Número de parejas ordenadas n = 8

Promedio de X = $\bar{x} = \frac{\sum x}{n}$

Promedio de Y = $\bar{y} = \frac{\sum y}{n}$

Como sabemos la ecuación de la recta de mínimos cuadrados

$$\hat{y} = a + bx \longrightarrow (1)$$

está definida por su pendiente b y su ordenada al origen a. Cada uno de estos parámetros se calcularán a partir de los valores de la tabla, en donde se incluyen las sumatorias $\sum x$, $\sum y$, $\sum xy$ y $\sum x^2$. De hecho algunos de estos valores ya los utilizaste en el cálculo del coeficiente de correlación (r) de Pearson. Estos valores serán aplicados a las relaciones siguientes:

$$b = \frac{(1/n)\sum xy \; - \bar{x}\,\bar{y}}{(1/n)\sum n^2 \; - (\bar{x})^2} \qquad \text{(Pendiente de la recta)} \longrightarrow (2)$$

Si suponemos que el punto $(\bar{x}, \bar{y})$ satisface la ecuación de regresión $\hat{y} = a + bx$

entonces:

$$\bar{y} = a + b\,\bar{x}$$

de donde despejamos el parámetro a, y obtenemos:

$$a = \bar{y} - b\,\bar{x} \quad \text{(Ordenada al origen)} \longrightarrow (3)$$

SOLUCIÓN.

Calculando los promedios $\bar{x}$ y $\bar{y}$ tenemos:

$$\bar{x} = \frac{\sum n}{n} = \frac{46}{8} = 5.75 \qquad \bar{y} = \frac{\sum y}{n} = \frac{182}{8} = 22.75$$

Sustituyendo los valores anteriores y los de la tabulación en la ecuación de la pendiente (2) tenemos:

$$b = \frac{(1/8)(1144) - (5.75)(22.75)}{/1/8)(350) - (5.75)^2} = \frac{12.1875}{10.6875} = 1.1403 \approx 1.14$$

Ahora, calculamos la ordenada al origen, mediante la ecuación (3).

$$a = 22.75 - (1.1403)(5.75) = 16.1932 \approx 16.2$$

Por lo tanto la ecuación de predicción o regresión será:

$$\hat{y} = 16.2 + 1.14x$$ RECTA DE MÍNIMOS CUADRADOS

Otra forma de calcular la recta de mínimos cuadrados es mediante las relaciones:

$$b = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2} \longrightarrow (4)$$

y la ecuación de mínimos cuadrados:

$$y_p = \bar{y} + b\,(x - \bar{x}) \longrightarrow (5)$$

sustituyendo valores tenemos:

$$b = \frac{(8)(1144) - (46)(182)}{(8)(350) - (46)^2} = \frac{780}{684} = 1.1403$$

Para la ecuación de regresión sustituimos valores:

$$y_p = 22.75 + 1.1403(x-5.75)$$

$$y_p = 22.75 + 1.1403x - 6.5570$$

$$y_p = 16.192 + 1.1403x$$

la cual corresponde a la ecuación calculada anteriormente. El manejo de los números decimales y del redondeo cobra gran importancia en este punto, de ahí que se deben manejar adecuadamente durante los cálculos.

Para terminar el ejemplo, utilizaremos la ecuación de regresión encontrada para realizar las estimaciones solicitadas en el ejemplo No. 1.

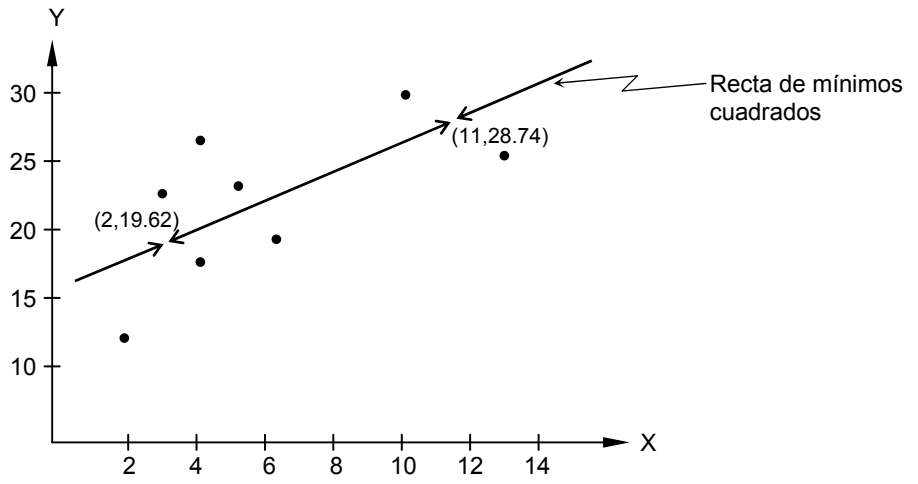
Ventas estimadas para un vendedor con tres años de experiencia.

$\hat{y}$ = 16.2 + 1.14(3) = 19.62 (miles de N$)

Ventas estimadas para un vendedor con once años de experiencia.

$\hat{y}$ = 16.2 + 1.14(11) = 28.74 (miles de N$)

Aún cuando no se mencionó al principio de este problema, nosotros esperaríamos que un vendedor con más experiencia vendería más que un vendedor con menos experiencia, los resultados anteriores corroboran esta suposición, ya que según nuestro modelo, un vendedor con 11 años de experiencia vende más que uno que tiene sólo 3 años en ventas. Por otro lado, si copiamos la gráfica de dispersión del ejemplo 1 y sobre ésta trazamos la recta que une los dos puntos estimados entonces tenemos el diagrama completo.



Gráfica No. 19

Un elemento de comprobación de la ecuación de mínimos cuadrados, lo podemos obtener al sustituir en ésta los valores de $\bar{x}$ y $\bar{y}$ con lo cual verificamos que esta pareja $(\bar{x}, \bar{y})$ pertenece a la recta de regresión.

En primer término, comprobemos que el punto $(\bar{x}, \bar{y})$, pertenece a la recta de regresión tal y como lo habíamos supuesto.

Si la ecuación de regresión es:

$\bar{y}$ = 16.2 + 1.14 x

al sustituir $\bar{x}$ obtenemos:

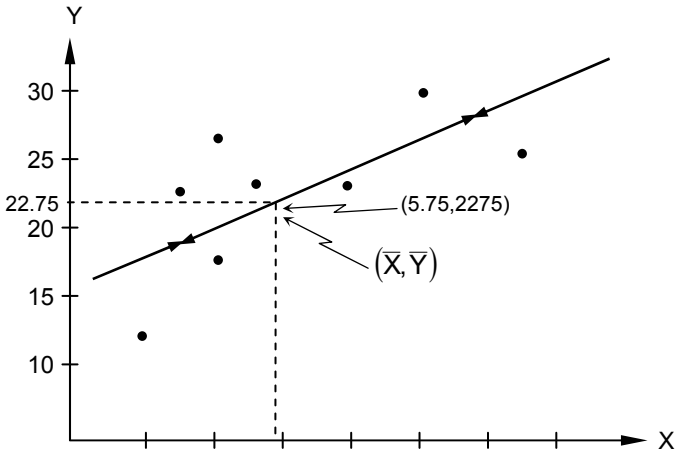
$$\bar{y} = 16.2 + 1.14\,\bar{x}$$

pero $\bar{x} = 5.75$ luego:

$$\bar{y} = 16.2 + 1.14(5.75) = 16.2 + 6.555 = 22.755 \approx 22.75$$

lo que es el valor de $\bar{y}$

Que se puede observar en la gráfica siguiente:

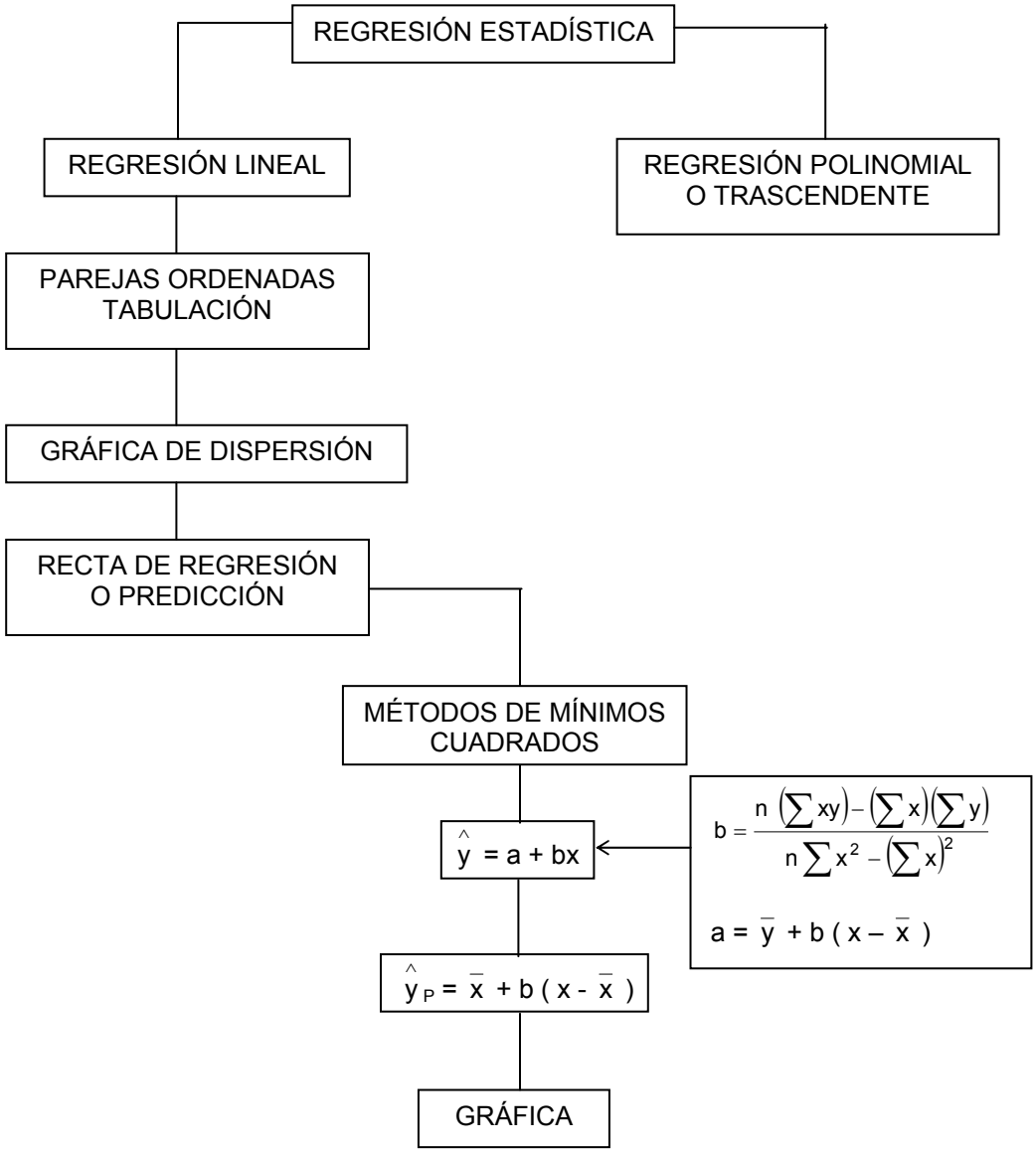


Gráfica No. 20

Una vez que has desarrollado estos conceptos, te recomendamos que calcules las ecuaciones de regresión de los ejemplos 2 y 3 de este tema con el fin de que practiques el desarrollo del método de mínimos cuadrados.

RECAPITULACIÓN

Un esquema de los temas de correlación y regresión lineales se te presenta a continuación, compleméntalo y agrega algún resumen de los puntos que consideres más relevantes de los mismos. Coméntalo con tu profesor o asesor.



RESUMEN DE CORRELACIÓN Y REGRESIÓN LINEALES

ACTIVIDADES DE CONSOLIDACIÓN

Para reafirmar los conocimientos que adquiriste sobre los temas de **Correlación** y **Regresión Lineales** al estudiar este fascículo, te sugerimos realizar las siguientes actividades:

1. Los siguientes datos muestran el número de horas (x) dedicadas a estudiar para un examen y la calificación (y) obtenida en dicha prueba. Observa en el diagrama de dispersión si existe alguna correlación lineal y en caso de que así sea, calcula el coeficiente de correlación de Pearson (r).

x (horas-estudio)	2	3	3	4	4	5	5	6	6	6	7	7	7	8	8
y (calificación)	5	5	7	5	7	7	8	6	9	8	7	9	10	8	9

2. Se realizó un estudio para investigar la relación que existe entre el peso (x) en libras (lb), la presión sanguínea (y), de adultos varones cuyas edades oscilan entre 19 y 30 años. Se obtuvieron los siguientes resultados.

x(lb)	173	178	145	146	157	175	173	137	199	131	152	172	163	170	135	159
y(lb/pul^2)	76	76	74	70	80	68	90	70	96	80	90	72	76	80	68	72

Calcula el coeficiente de correlación de Pearson (r) e interpreta tu resultado.

3. Se efectuó un experimento para investigar las variables que probablemente estuvieran relacionadas con el espíritu de iniciativa en las situaciones de resolución de problemas. Los sujetos formaban parte de una muestra aleatoria de 14 estudiantes de penúltimo año de una prestigiada universidad. Los resultados se muestran en la tabla. Calcula el coeficiente de Pearson (r) e interprétalo.

Puntaje de auto-concepto (y)	5	6	6	7	8	8	8	9	9	9	10	10	11	12
Puntaje iniciativa personal (x)	5	6	8	7	9	11	12	11	12	14	14	16	15	17

4. De acuerdo con lo que has desarrollado en este fascículo, contesta las preguntas que se encuentran al inicio del tema de regresión y coméntalas con tu profesor o asesor.

5. En una de las Secretarías del gobierno federal se ha implantado el sistema de retiro voluntario. Para analizar dicho proceso se toma una muestra aleatoria en los distintos departamentos, donde se relaciona el número de empleados que han renunciado y el número de años de servicio. Se pretende estimar cuántos trabajadores renunciarían en función de su antigüedad. Se obtuvieron los siguientes datos:

No. de años de servicio (X)	No. de empleados que Renunciaron (Y)
16	14
9	15
13	16
10	14
15	17
10	10
11	15
12	12

Calcula el coeficiente de correlación de Pearson (r) y obtén la ecuación de regresión. Estima cuántos empleados renunciarían si tuvieran 14 o 17 años de servicio. Construye la gráfica de dispersión junto con la recta de mejor ajuste.

6. Te sugerimos realices una lectura comentada de los capítulos siguientes: Relación entre correlación y regresión lineales páginas 485-491 del libro Estadística elemental por R. Johnson, de la bibliografía.

   El modelo bivariante, páginas 339-347 del libro Estadística con aplicaciones a las Ciencias Sociales y a la educación por W.W. Daniel, de la bibliografía.

AUTOEVALUACIÓN

A continuación te proporcionamos algunas de las respuestas de los problemas que están redondeadas a dos o tres cifras, de las Actividades de Consolidación. Complétalos y verifica tus respuestas.

SOLUCIONES:

1) El diagrama de dispersión lo dejamos para que los compares con tus compañeros y cambies impresiones. El cálculo de r redondeado a tres cifras, da como resultado 0.741.

2) El coeficiente r de Pearson redondeado a tres cifras tiene un valor de 0.453 y como recuerdas, el tipo de correlación que existe entre las variables se llama…. _____ Completa la respuesta, con base a los diferentes diagramas de dispersión e interpreta dicho resultado.

3) El coeficiente r de Pearson redondeado a tres cifras tiene un valor de 0.95.

4) El coeficiente de Pearson redondeado a tres cifras tiene un valor de 0.999.

Para el tema de Regresión Lineal, se sugiere elaborar un ensayo acerca de los puntos esenciales del tema, de manera que el profesor o asesor observe el manejo de éstos.

ACTIVIDADES DE GENERALIZACIÓN

El objetivo de las siguientes actividades es el que puedas realizar no sólo cálculos de correlación sino que también apliques e interpretes tus resultados.

1. En un grupo de observaciones de estaturas de padres e hijos, que obtengas de tu entorno social (familiares o amistades), comprueba la hipótesis de que si los padres son altos, entonces sus hijos serán altos también y si los padres son bajos entonces sus hijos serán bajos. ¿Cuál es el comportamiento de estaturas de los hijos con relación a la estatura promedio de los padres? Tiene esto que ver con los conceptos de **Correlación y Regresión lineales**? Si es así, explica.

(Sugerencia: Construye la gráfica de dispersión. Calcula el coeficiente de correlación y obtén la recta de mínimos cuadrados para que te sirva de base en el análisis).

2. Explica en forma completa la diferencia entre relaciones causales y relaciones estadísticas.

3. Explica ampliamente los conceptos de correlación y regresión.

4. Los siguientes resultados muestran las puntuaciones obtenidas por 6 estudiantes tomados al azar en las asignaturas de idiomas y matemáticas.

Idiomas (Y)	525	515	510	495	430	400
Matemáticas (X)	550	535	535	520	455	420

Construye la gráfica de dispersión. Calcula el coeficiente de correlación de Pearson (r) y encuentra la ecuación de regresión. Haz una conjetura acerca de ¡cuánto obtendría en matemáticas un estudiante que hubiera obtenido 480 puntos en idiomas? Si se considera el aprendizaje de las matemáticas como un problema de lenguaje ¿es razonable pensar que el buen manejo de otros idiomas facilitaría el manejo del lenguaje matemático? Explica.

5. Comprueba que la recta $\hat{y} = a + bx$ puede expresarse como $Y = \bar{y} + b(X - \bar{x})$.

   Te sugerimos encuentres la recta de regresión de los ejemplos desarrollados durante el fascículo, con esta relación).

BIBLIOGRAFÍA CONSULTADA

ARNOL Naiman, R. Rosenfeld, G. Zirkel. Introducción a la Estadística. México, D.F. Editorial Mc Graw Hill 1987.
Este texto cubre el 100% del programa, manejando el enfoque del mismo. Sobre el tema incluye una variedad de ejemplos prácticos que permiten una visión amplia en este terreno.

JOHNSON, Robert. Estadística elemental. México, D.F., grupo Editorial Iberoamérica 1990.
Este texto cubre el 90% del programa, siguiendo el enfoque del mismo. Con relación al tema su tratamiento es muy adecuado.

N. M. Downie, R. W. Heath. Métodos Estadísticos Aplicados. 3ª. Edición. México, D.F. Editorial Harla. 1973.

PORTILLA Chimal, E. Estadística (primer curso). México, D.F. Nueva Editorial Interamericana. 1980.
Este libro aborda el tema de manera muy adecuada, incluye ejemplos muy ilustrativos.

PROAÑO, Humberto. Estadística Aplicada a la Mercadotecnia. 4ta. Edición. México, D.F. Editorial Diana. 1983.
Este texto cubre el 80% del curso. El tratamiento de los temas es muy claro, además de que incluye ejemplos de aplicación práctica.

WAYNE W. Daniel. Estadística con Aplicaciones a las Ciencias Sociales y a la Educación. México, D.F. Editorial Mc Graw Hill / Interamericana de México. 1988.