4.1Introducción

En capítulos anteriores se empezó con el estudio de la *inferencia estadística.* El interés principal era inferir algo acerca de un parámetro poblacional, como la media poblacional, basándose en una muestra. Probamos si una media poblacional o una proporción poblacional eran razonables, la diferencia entre dos medias poblacionales, o si varias medias poblacionales eran iguales. En todas estas pruebas teníamos sólo *una* variable, como el promedio de velocidad de automovilistas, registros de gastos de alimentación, cantidad de bacterias en una planta de tratamiento, cantidad de desechos de una población, etc...

En este capitulo estudiaremos la *relación entre dos o más variables y desarrollaremos una ecuación que nos permite estimar una variable con base en otra.* Por ejemplo, podemos saber si *h*ay alguna relación entre la cantidad que una empresa gasta en publicidad y sus ventas, estimar el índice de alcoholismo basándonos en el índice de analfabetismo, estimar la producción de un empleado, en base al número de años que lleva laborando, etc. Observe que en cada uno de estos ejemplos hay dos variables, por ejemplo, el número de años trabajados y el número de unidades producidas. Empezaremos este capitulo estudiando **análisis de correlación.** Después vemos una gráfica, llamada **diagrama de dispersión,** diseñada para representar la relación entre dos variables. Continuamos nuestro estudio desarrollando un modelo matemático que nos permitirá estimar el valor de una variable basándonos en el valor de otra. A esto se le llama **análisis de regresión.** 1) Determinaremos la ecuación de la línea que mejor se ajuste a los datos, 2) estimaremos el valor de una variable basándonos en otra, 3) mediremos el error de nuestra estimación, y 4) estableceremos intervalos de confianza y de predicción para nuestra estimación.

4.2 ¿Qué es análisis de correlación?

Un ejemplo ilustrará mejor lo que es el análisis de correlación. Suponga que el gerente de ventas de Copier Sales of America, que tiene una gran fuerza de venta en todo Estados Unidos y Canadá, quiere determinar si hay alguna relación entre el número de

llamadas de venta hechas a posibles compradores en un mes y el número de copiadoras vendidas en ese mes. La gerente selecciona una muestra aleatoria de 10 representantes y determina el número de llamadas de venta hechas por cada representante el mes pasado y el número de copiadoras que vendió. La información muestral se da en la ilustración 1.

Ilustración 1. Llamadas de venta y copiadoras vendidas por diez vendedores

Representante de venta	Llamadas de venta	Copiadoras vendidas
Tom Keller	20	30
Jeff Hall	40	60
Briant Virost	20	40
Gregg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rick Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

Parece haber una relación entre el número de llamadas de venta hechas y el número de copiadoras vendidas. Es decir, los vendedores que hicieron más llamadas de venta vendieron más unidades. Sin embargo, la relación no es "perfecta" o exacta. Por ejemplo Soni Jones hizo menos llamadas que Jeff Hall, pero ella vendió más unidades. En lugar de estar hablando en general como hemos estado haciendo hasta aquí, vamos a desarrollar algunas mediciones estadísticas que reflejen más precisamente la relación entre las dos variables, llamadas de ventas y copiadoras vendidas. A este grupo de de técnicas estadísticas se le llama **análisis de correlación**.

Análisis de correlación: Un grupo de técnicas para medir la magnitud de la relación entre dos variables.

La idea básica del análisis de correlación es dar la magnitud de la relación entre dos variables. El primer paso es normalmente graficar los datos en un diagrama de dispersión.

Diagrama de dispersión: Un diagrama que refleja la relación entre dos variables.

Ejemplo:
Un ejemplo nos mostrará como se usa un diagrama de dispersión.

Copier Sales of America, Inc., vende copiadoras a negocios de todos tamaños en todo Estados Unidos y Canadá. Marcy Bancer acaba de ser promovida a gerente nacional de ventas. A la próxima reunión de vendedores vendrán los representantes de ventas de todo el país. A la señorita Bancer le gustaría subrayar la importancia de hacer diariamente las llamadas necesarias. Decide reunir información sobrela relación entre el número de llamadas de venta y el número de copiadoras vendidas. Selecciona una muestra aleatoria de 10 representantes de ventas y determina el número de llamadas de venta que hicieron el mes pasado y el número de copiadoras que vendieron. La información obtenida de esta muestra se da en la ilustración 1 de la tabla anterior. Represente esta información en un diagrama de dispersión. ¿Cuáles son sus observaciones acerca de la relación entre el número de llamadas de venta y el número de copiadoras vendidas?.

Solución

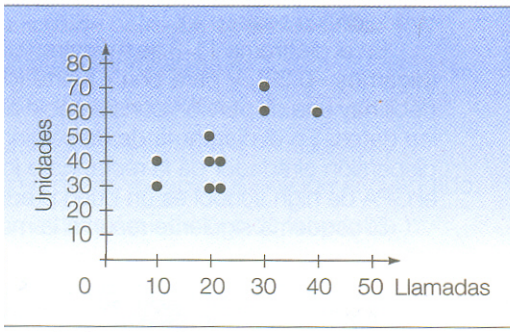
Basándose en la información de la ilustración 1, la señora Bancer sospecha que si hay relación entre el número de llamadas de venta hechas en un mes y el número de copiadoras vendidas. Soni Jones es quien vendió más copiadoras el mes pasado, y ella fue una de los tres representantes que hicieron 30 llamadas de venta o más. Por otro lado Susan Welch y Carlos Ramírez hicieron sólo 10 llamadas de venta el mes pasado. La señorita Welch es, de los representantes en la muestra, la que menos copiadoras vendió.
La conclusión es que el número de copiadoras vendidas se relaciona con el número de llamadas de venta hechas. Conforme aumenta el número de llamadas de venta, también aumenta el número de copiadoras vendidas. Al número de llamadas de venta le llamamos la *variable independiente, y* al número de copiadoras vendidas le llamamos la *variable dependiente.*

Variable dependiente: La variable que se va a predecir, o estimar

Variable independiente: Una variable que da la base para la estimación. Es la variable predictora.

Lo que se suele hacer es colocar la escala de la variable dependiente (copiadoras vendidas) en el eje vertical o eje *Y* y la escala de la variable independiente (número de llamadas de ventas) en el eje horizontal o eje X. Para trazar el diagrama de dispersión correspondiente a la información de las ventas de Copier Sales of America, empezamos con el primer representante de ventas, Tom Keller. Tom hizo 20 llamadas el mes pasado y vendió 30 copiadoras, por lo que X = 20 y *Y*= 30. Para localizar este punto, muévase a lo largo del eje horizontal hasta X = 20, después en forma vertical hasta *Y*= 30 y coloque un punto en la intersección. Así se sigue hasta que se han localizado todos los pares de datos, como se muestra en el diagrama 1. El diagrama de dispersión muestra gráficamente que los representantes de ventas que hicieron más llamadas tienden a vender más copiadoras. Para la señora Bancer, gerente nacional de ventas de Copier Sales of America, es razonable decir a sus vendedores que entre más llamadas de venta hagan, más copiadoras podrán vender. Observe que aunque parece haber una relación positiva entre las dos variables, no todos los puntos caen en una recta. En la siguiente sección usted podrá medir numéricamente la magnitud y dirección de esta relación entre dos variables, determinando el coeficiente de correlación.

Diagrama 1. Diagrama de dispersión que muestra llamadas de venta y copiadoras vendidas

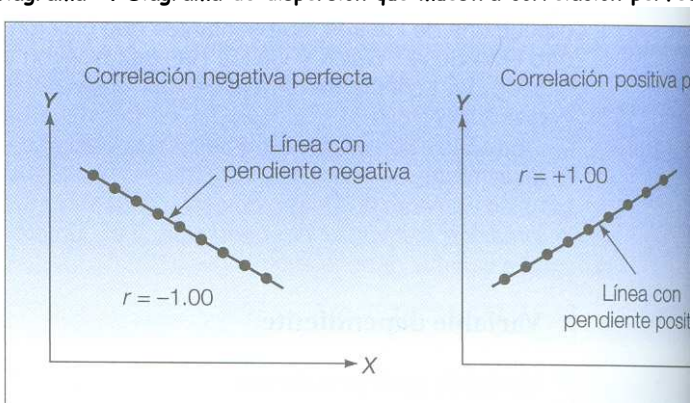


4.3 El coeficiente de correlación

Encontrado por Karl Pearson alrededor de 1900, el **coeficiente de correlación** describe la magnitud de la relación entre dos conjuntos de variables de intervalo o de razón. Se designa como *r* y con frecuencia se le llama *r de Pearson* o *coeficiente de correlación de Pearson.* Puede tomar cualquier valor desde -1.00 hasta + 1.00 inclusive. Un coeficiente de correlación de -1.00 o de +1.00 indica correlación perfecta. Por ejemplo, si en el ejemplo anterior se obtuviera un coeficiente de correlación de +1.00, esto indicaría que el número de llamadas de venta sería un predictor perfecto del número de copiadoras vendidas. Es decir el número de llamadas de venta y el número de copiadoras vendidas están relacionadas perfectamente en un sentido lineal positivo. Un valor calculado de-1.00 indicaría que la variable independiente X y la variable

dependiente Y están perfectamente relacionadas de una manera lineal negativa. En el diagrama 2 se muestra cómo se vería un diagrama de dispersión si la relación entre los dos conjuntos de datos fuera lineal y perfecta.

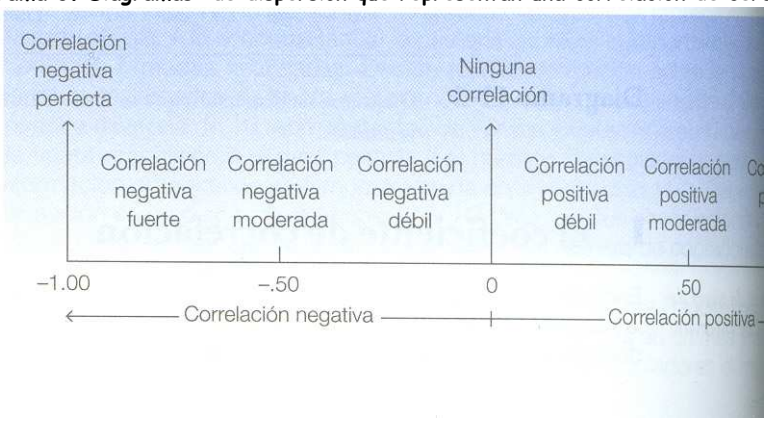
Diagrama 2. Diagrama de dispersión que muestra correlación perfecta negativa y positiva.

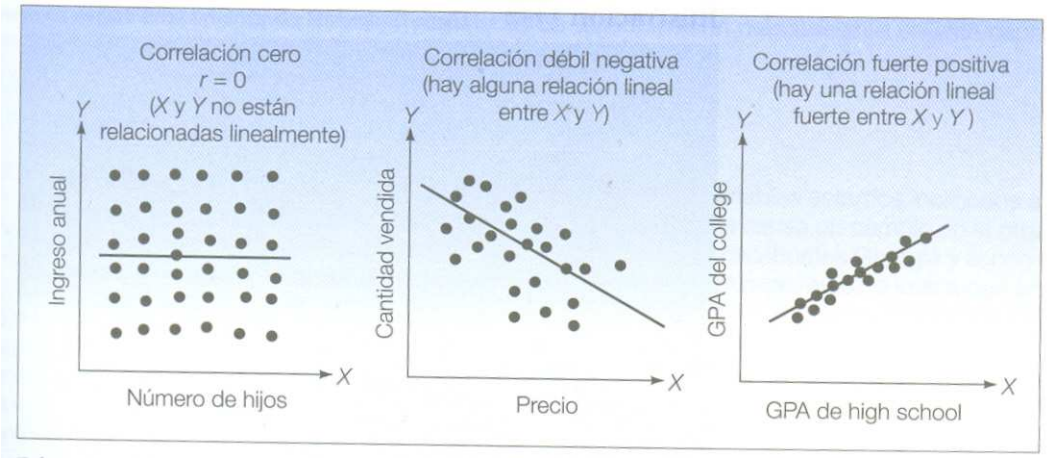


Si no hay absolutamente ninguna relación entre los dos conjuntos de variables, la r de *Pearson* será cero. Un coeficiente de correlación *r* cercano a cero (digamos 0.08) indica que la relación es muy débil. A la misma conclusión se llega si r = -0.08. Coeficientes de -0.91 y +0.91 tienen igual magnitud; ambos indican una correlación muy fuerte entre los dos conjuntos de variables. *La magnitud del coeficiente de dirección (ya sea – o +)* .En el diagrama 3 se muestran diagramas de dispersión para r = 0, para una r débil (digamos -0.23), y para una *r* fuerte (digamos +0.87). *Observe que si la relación es débil hay una dispersión considerable alrededor de la recta trazada a través del centro de los datos*. En el diagrama de dispersión que muestra una fuerte relación, hay muy poca dispersión alrededor de la recta. Esto indica en el ejemplo mostrado en el diagrama, que el GPA de High School es un buen predictor para el desempeño en el College

El esquema siguiente resume la magnitud y dirección del coeficiente de correlación.

Diagrama 3. Diagramas de dispersión que representan una correlación de cero, una débil y una fuerte.





Coeficiente de correlación: Una medida de la magnitud de la relación lineal entre dos variables.

Para determinar el coeficiente de correlación usamos la fórmula siguiente:

COEFICIENTE DE CORRELACIÓN: $$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{\left[n(\sum X^2) - (\sum X)^2\right]\left[n(\sum Y^2) - (\sum Y)^2\right]}}$$

donde :

n es el número de pares de observaciones
ΣX es la suma de las variables X
ΣY es la suma de las variables Y
(ΣX^2) es la suma de los cuadrados de la variable X
$(\Sigma X)^2$ es la suma de las variables X elevadas al cuadrado
(ΣY^2) es la suma de los cuadrados de la variable Y
$(\Sigma Y)^2$ es la suma de las variables Y elevadas al cuadrado
ΣXY es la suma de los productos X y Y

Ejemplo 1: Vea el ejemplo anterior en el que trazamos un diagrama de dispersión que representa la relación entre el número de llamadas de venta y el número de copiadoras vendidas. Determine el coeficiente de correlación e interprete su valor.

Solución:

Representante de venta	Llamadas de venta (X)	Copiadoras vendida (Y)	X^2	Y^2	XY
Tom Keller	20	30	400	900	600
Jeff Hall	40	60	1600	3600	2400
Briant Virost	20	40	400	1600	800
Gregg Fish	30	60	900	3600	1800
Susan Welch	10	30	100	900	300
Carlos Ramirez	10	40	100	1600	400
Rick Niles	20	40	400	1600	800
Mike Kiel	20	50	400	2500	1000
Mark Reynolds	20	30	400	900	600
Soni Jones	30	70	900	4900	2100
TOTAL	220	450	5600	22100	10800

$$r = \frac{10(10800) - (220)(450)}{\sqrt{\left[10(5600) - (220)^2\right]\left[10(22100) - (450)^2\right]}} = 0.759$$

¿Cómo interpretamos una correlación de 0.759?. Primero, es Positiva, por lo que vemos que hay una relación directa entre el número de llamadas hechas y el número de copiadoras vendidas. Esto confirma nuestro razonamiento basado en el diagrama de dispersión 1 . Concluimos que la relación es fuerte, pues el valor 0.759 esta cercano a +1.00. Dicho de otra manera, u n aumento de un 25% de las llamadas, posiblemente llevaría a un 25% de aumento en las ventas.

4.4 El coeficiente de determinación

En el ejemplo anterior, interpretamos el coeficiente de correlación de 0.759 con respecto de la relación entre las llamadas de venta y el número de copiadoras vendidas, como "fuerte". Términos como débil, moderado y fuerte, no tienen un significado preciso. Una medición que tiene una mejor interpretación es el coeficiente de determinación. Se calcula elevando al cuadrado el coeficiente de correlación, en el ejemplo, el coeficiente de determinación r^2, es 0.576, que es $(0.759)^2$. Este es una proporción, o un porcentaje, podemos decir que 57.6 por ciento de la variación en el número de copiadoras vendidas se explica, o se debe a la variación en el número de llamadas de venta.

Coeficiente de determinación: La proporción ó porcentaje de la variación total de la variable dependiente Y que se explica por, o se debe a la variación en la variable independiente X.

Advertencia: Cuando hay una relación muy fuerte (digamos 0.91) entre dos variables, estamos obligados a pensar que un incremento o una disminución en una variable causa un cambio en la otra variable. Por ejemplo, se puede mostrar que el consumo de cerveza, está fuertemente relacionado con el consumo de aspirinas, sin embargo esto no indica que un aumento en el consumo de cerveza, este causando un aumento en el consumo de aspirinas. De la misma manera, una disminución en el número de reprobados, no implica un aumento en la calidad de la enseñanza de los profesores. A correlaciones como éstas se les llama **correlaciones espurias**. Lo que podemos concluir es que cuando encontramos dos variables fuertemente correlacionadas, es que hay una relación entre las dos variables, pero no que un cambio en una, causa un cambio en la otra. Entonces la señora Bancer, gerente de ventas de *Copier Sales of America*, puede concluir que hay una relación positiva entre las llamadas de venta y las copiadoras vendidas. No puede concluir que mas llamadas de venta cause que se vendan más copiadoras.

Ejemplo 2

Una tienda de ropa tiene sucursales en varias áreas metropolitanas grandes. La gerente general de ventas planea transmitir un comercial de televisión a través de las estaciones locales por lo menos dos veces, antes de una venta gigante que empezará en sábado y terminará el domingo. Quiere comparar las ventas de sábado y domingo en las diversas sucursales y el número de veces que apareció la publicidad en la estación local de televisión. El propósito principal de la investigación es ver si hay alguna relación entre el número de veces que se transmitió al aire la publicidad y las ventas. Las comparaciones son:

Ubicación de la estación de televisión	Número de emisiones	Ventas en sábado y domingo (miles de dólares)
Guadalajara	4	15
Monterrey	2	8
Puebla	5	21
D.F.	6	24
Tijuana	3	17

a) ¿Cuál es la variable dependiente?
b) Dibuje un diagrama de dispersión
c) Determine el coeficiente de correlación
d) Determine el coeficiente de determinación.
e) Interprete éstas mediciones estadísticas.

Solución:

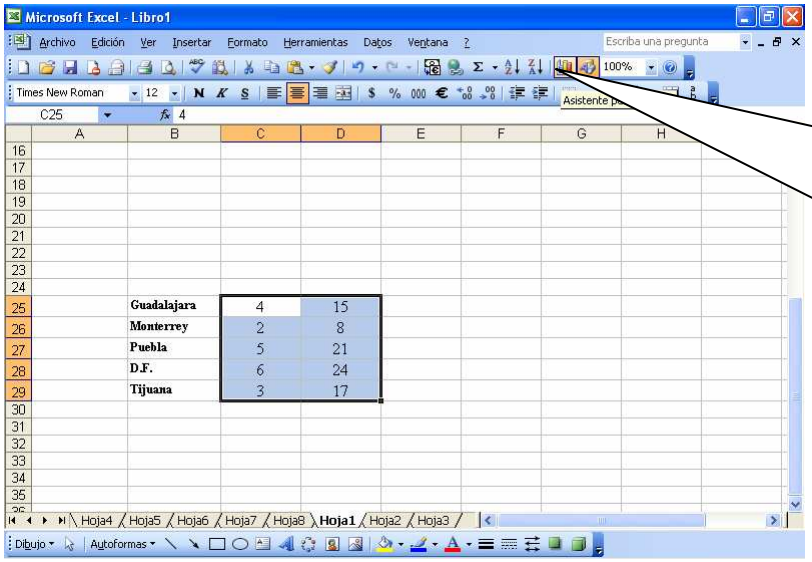
	X	Y	X^2	Y^2	XY
Guadalajara	4	15	16	225	60
Monterrey	2	8	4	64	16
Puebla	5	21	25	441	105
D.F.	6	24	36	576	144
Tijuana	3	17	9	289	51
$\sum$	**20**	**85**	**90**	**1595**	**376**

$$r = \frac{5(376) - (20)(85)}{\sqrt{\left[5(90) - (20)^2\right]\left[5(1595) - (85)^2\right]}} = \frac{180}{\sqrt{50 * 750}} 0.9295$$

por lo que existe una relación bastante fuerte entre las dos variables

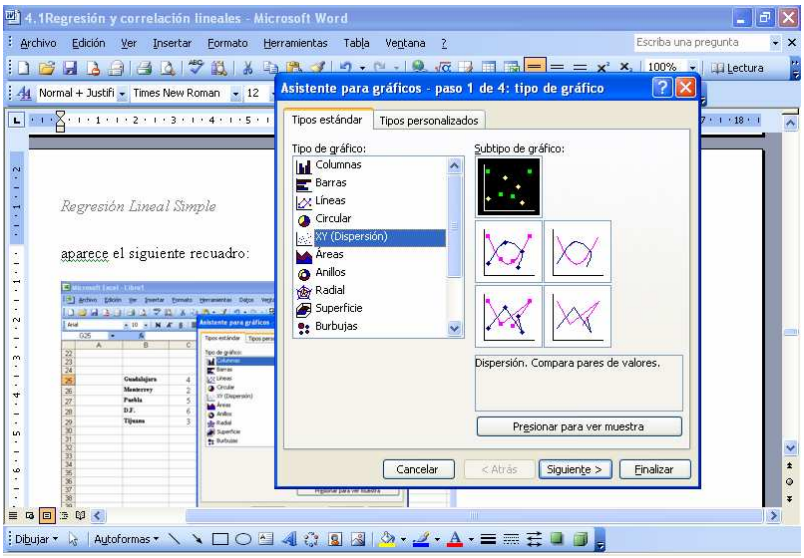
r^2 = 0.8639 por lo que la T.V. resultó efectiva en un 86.4%

El diagrama de dispersión también puedes graficarlo en Excel de la siguiente manera: Introducimos la información del problema. Después seleccionamos los datos con el ratón. A continuación seleccionamos el asistente para gráficos.

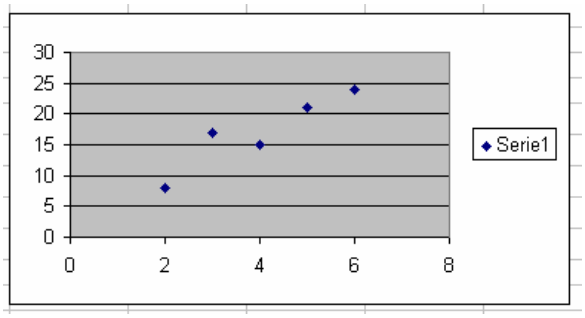


Ya introducidos los datos, seleccionamos los valores con el Mouse. tecleamos en el *asistente para gráficos* o también en el menú: ***Insertar- gráficos***

aparece el siguiente recuadro: donde seleccionaremos la forma de gráfico ***XY dispersión***



Damos click en ***Siguiente*** hasta ver la opción finalizar, deberás especificar donde quieres que aparezca el gráfico, se obtiene el siguiente diagrama



4.5Análisis de regresión

A continuación desarrollaremos un modelo para expresar la relación entre dos variables y estimar el valor de la variable dependiente Y, basándonos en el valor de la variable independiente X. El modelo lo formularemos a partir de la ecuación de una línea recta que se ajuste a los datos graficados en el diagrama de dispersión(Ver diagrama 4).

A esta ecuación de la línea recta que se usa para estimar el valor de Y basándose en X, se le llama **ecuación de regresión**.

Sin embargo, la línea recta trazada con una regla tiene una desventaja: su posición se basa, en parte en el juicio de la persona que traza la línea. Las líneas trazadas en el diagrama 5 representan los juicios de cuatro personas. Todas las líneas, excepto la línea A, parecen ser razonables. Sin embargo, con cada una se tendría una estimación diferente de las unidades vendidas.

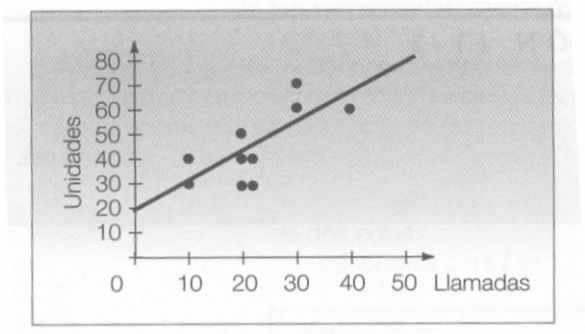


Diagrama 4. Llamadas de venta y copiadoras vendidas por 10 representantes de venta.

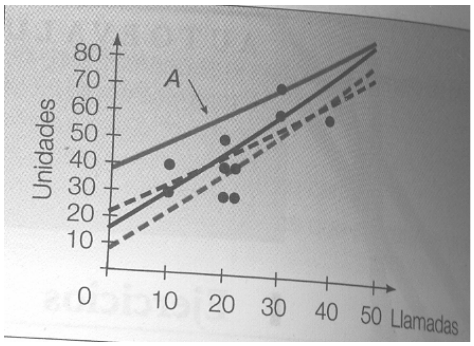


Diagrama 5. cuatro líneas trazadas cada una por diferentes personas