UNIVERSIDAD TECNOLÓGICA NACIONAL

Facultad Regional San Nicolás

PROBABILIDAD y ESTADÍSTICA II

UNIDAD Nº2

Licenciatura en Enseñanza de la Matemática

Año 2011

Mg. Lucía C. Sacco

UNIDAD Nº2

Variables aleatorias bidimensionales discretas. Correlación y regresión.

Correlación. Diagramas de dispersión.
Parámetros de una distribución bidimensional. Esperanza matemática y varianza. Covarianza.
Correlación lineal. Coeficiente de correlación lineal. Coeficientes de determinación.
Regresión lineal. Recta de regresión mínimo cuadrática. Fiabilidad de la recta de regresión.

Propósitos:

Brindar oportunidades para la construcción de herramientas que permitan:

- Utilizar los diagramas de dispersión para representar conjuntos de datos de dos variables.
- Aprender el significado de correlación estadística.
- Medir la dependencia estadística con ayuda del coeficiente de correlación lineal.
- Calcular la recta de regresión y emplearla para hacer estimaciones.

Bibliografía sugerida:

- Canavos, George. *Probabilidad y Estadística. Aplicaciones y Métodos*. México. McGraw Hill. 1988.
- Meyer Paul L. *Probabilidad y Aplicaciones Estadísticas*. México. Addison Wesley Iberoamericana .1993.
- Walpole Ronald, Myers Raymond. *Probabilidad y Estadística*. México. Pearson Educación. 1999.
- Zylberberg, Alejandro. *Probabilidad y Estadística P(X).* Nueva Librería.

1. Correlación

Al estudiar distribuciones bidimensionales, el objetivo perseguido es determinar si existe relación estadística entre las dos variables consideradas. Es decir, ver si los cambios en una de las variables influyen en los cambios de la otra. Cuando sucede esto, diremos que ambas variables están correlacionadas o que hay **correlación** entre ellas.

Si las variables crecen conjuntamente, la **correlación es directa.** Si, por el contrario, al aumentar una de ellas disminuye la otra, la **correlación será inversa.**

La correlación puede clasificarse como *fuerte* cuando el grado de dependencia es alto; y como *débil* en caso contrario.

Si la correlación es fuerte, a partir de una variable puede estimarse la otra con una probabilidad alta. Si la correlación es débil, la estimación de una variable a partir de otras es poco fiable.

Ejemplos

a) La correlación entre el número de zapato y la estatura de las personas es directa y fuerte. (Las fábricas de zapatos hacen tallas de zapatos más grandes en Suecia que en Japón, pues, en general, los suecos son más altos que los japoneses. No obstante, nadie se compra los zapatos por su estatura, todo el mundo se los prueba!!).

b) Las variables temperatura y el número de enfermos de gripe están inversamente correlacionadas: a menor temperatura más enfermos de gripe. Quizás se trate, también, de una correlación fuerte.

c) Las variables altura y cociente intelectual de las personas no están correlacionadas.

d) La correlación entre el número de errores cometidos y tiempo empleado en realizar una tarea por un grupo de personas no sabemos cómo es; para determinarla habría que tener datos concretos.

1.1. Diagrama de dispersión

El primer paso para determinar el sentido y el grado de la correlación entre dos variables consiste en representar gráficamente, en el plano cartesiano, los pares de valores conocidos. Estos gráficos, que reciben el nombre de **diagramas de dispersión**, permiten visualizar la posición de los datos en el plano. La forma de la **nube de puntos** asociada a cada diagrama nos permitirá establecer conjeturas sobre la correlación existente entre las variables estudiadas.

Ejemplo 1

Los siguientes datos corresponden a la vida útil y a la velocidad de corte de una herramienta:

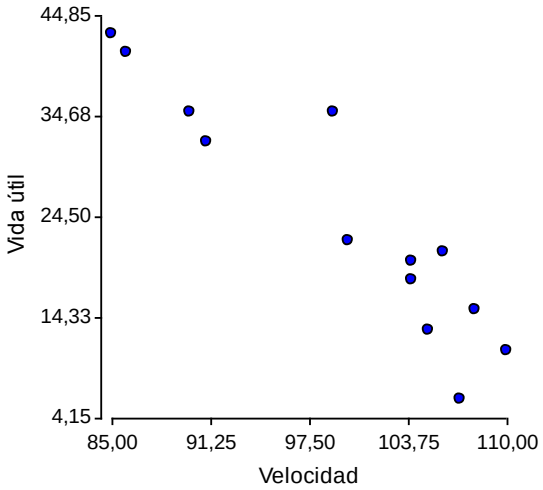
Velocidad de corte:	86	104	100	85	107	104	106	99	90	110	108	105	91
Vida útil:	41	18	22	43	6	20	21	35	35	11	15	13	32

Dados los pares de valores (x, y) entre los cuales se desea estudiar la relación, se representan a los mismos en un sistema de ejes cartesianos ortogonales, seleccionando una escala de modo que la lectura del diagrama resulte más fácil

(generalmente se considera la diferencia entre el máximo y mínimo de cada variable y a esa diferencia se le asigna la misma longitud en cada eje).

Si una de las variables se puede considerar como la variable que causa, o explica los cambios observados en la otra, a esa variable se la denomina **explicativa** y se la representa sobre el eje x. En este caso, a la otra variable se la denomina **variable respuesta** y se la representa sobre el eje y. Si no se quiere distinguir entre variable explicativa y variable respuesta, cualquiera de las dos puede representarse en el eje de las abscisas.

El siguiente diagrama de dispersión corresponde a los datos sobre la velocidad de corte y vida útil de una herramienta.



Para interpretar un diagrama de dispersión es necesario reconocer primero su aspecto general que debe revelar la dirección, la forma e intensidad de la relación entre las variables.

A partir del diagrama de dispersión se percibe una relación con tendencia lineal y pendiente negativa. Valores superiores al promedio de la velocidad de corte están en correspondencia con valores que son inferiores al promedio de la vida útil y valores inferiores al promedio de velocidad de corte están en correspondencia con valores que superan el promedio de la vida útil. En este caso decimos que las variables están **relacionadas negativamente**.

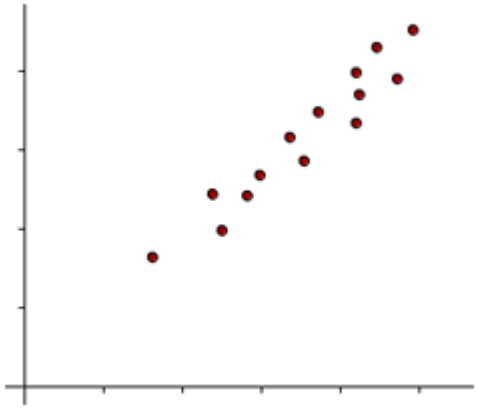
Asimismo decimos que dos variables están **relacionadas positivamente** cuando los valores que se representan sobre el eje x y que superan su promedio tienden a estar en correspondencia con los valores que se representan sobre el eje y superan su promedio, y los valores inferiores al promedio de cada variable también tienden a ocurrir conjuntamente.

Si existe una "fuerte" relación entre dos variables, el conocimiento de una de ellas permite predecir el comportamiento de la otra, pero cuando la relación es débil, la información de una de las variables no ayuda demasiado a extraer conclusiones sobre la otra.

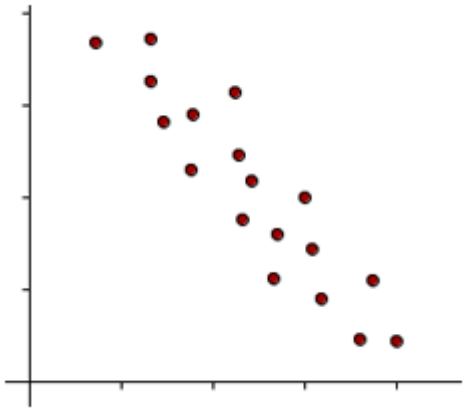
En general, dependiendo de la forma de la nube de puntos, puede asegurarse:

- Una nube de puntos alargada indica **correlación lineal**: los puntos se distribuyen en torno a una línea recta. La estrechez de la nube expresa que la correlación es fuerte.
- Si la recta que se ajusta a la nube tiene pendiente positiva, **la correlación será directa**; al crecer la variable X, lo hace también la variable Y.
- Una recta con pendiente negativa, indica que la **correlación es inversa**, al crecer X, disminuye Y.

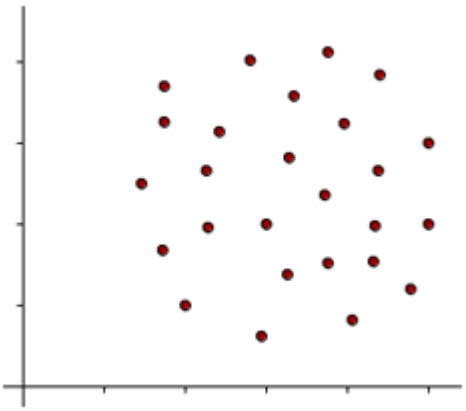
Correlación directa fuerte



Correlación directa débil



Correlación nula



Actividades:

1. Ocho personas, con similar destreza en mecanografía, teclearon 40 líneas de texto en un ordenador. El tiempo empleado, en minutos, y el número de errores cometidos, fueron:

Tiempo (X)	6	7	8	9	9	10	12	12
Errores (Y)	22	15	12	17	21	13	9	6

¿Qué tipo de correlación se da entre las variables estudiadas?

2. Los siguientes datos corresponden al consumo de combustible de un auto a medida que aumenta su velocidad.

Velocidad	10	20	30	40	50	60	70	80	90	100	120	130	140	150
Consumo	21	13	10	8	7	5.9	6.3	6.95	7.57	8.27	9.87	10.79	11.77	12.83

a) Dibuja un diagrama de dispersión. ¿Cuál consideras que es la variable explicativa?
b) Describe la forma de la relación.

2. Parámetros de una distribución bidimensional

Los datos de una distribución bidimensional suelen darse en forma de tabla. Los datos correspondientes a cada una de las variables se llaman **datos marginales.** En el caso de tablas de doble entrada puede hablarse de **frecuencias marginales.** Estos datos permiten el cálculo de los **parámetros marginales** de cada una de las variables.

2.1. Medias

Las medias marginales para cada una de las variables X e Y valen respectivamente:

$$\bar{x} = \frac{\sum x_i}{n} \qquad \bar{y} = \frac{\sum y_i}{n}$$

El punto $(\bar{x}, \bar{y})$ se llama **centro medio** de la distribución. Es el centro de gravedad (o centro de masas) de la nube de puntos.

2.2. Varianzas y desviaciones típicas

Las varianzas marginales, que denotamos s_x^2 y s_y^2, valen:

Para la variable X: $s_x^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\sum x_i^2}{n} - \bar{x}^2$

Para la variable Y: $s_y^2 = \frac{\sum (y_i - \bar{y})^2}{n} = \frac{\sum y_i^2}{n} - \bar{y}^2$

Las desviaciones típicas marginales, s_x y s_y son la raíz cuadrada de las respectivas varianzas.

2.3. La covarianza

La covarianza es un parámetro estadístico conjunto, pues, en su cálculo intervienen las dos variables a la vez. Se define como la media aritmética de los productos de las diferencias de los valores de cada variable respecto de su media marginal. Por tanto, vale:

$$s_{xy} = \frac{\sum (x_i - \bar{x})\ (y_i - \bar{y})}{n} \quad \Leftrightarrow \quad \frac{\sum x_i\ y_i}{n} - \bar{x}\,\bar{y}$$

La covarianza es un número tal que:

1. Su signo indica el sentido de la correlación entre las variables.
   - Si $S_{xy} > 0$, la correlación es directa.
   - Si $S_{xy} < 0$, la correlación es inversa.
2. Un valor grande de S_{xy} advierte que la correlación entre las variables puede ser fuerte. Pero a la covarianza le pasa lo mismo que a la varianza: se ve afectada por la unidad de medida en que vengan los datos. Por otra parte, la distinta naturaleza de los fenómenos estudiados hace que la comparación entre covarianzas carezca de significado. En definitiva, la covarianza sólo nos da el

sentido de la correlación: directa si es positiva e inversa si su valor en negativo.

Actividad:

3. Si consideramos la distribución que mide la altura de niños en cm de diferentes años:

Años (X)	1	2	3	6
Estatura (Y)	63	80	88	101

a. Realizar el diagrama de dispersión

b. Hallar su covarianza.

c. ¿Qué tipo de correlación presentan las variables años y estatura? ¿qué implica esto?

3. Cálculo de la correlación lineal

En las distribuciones de una sola variable, el conocimiento de la varianza (o la desviación típica) no basta para precisar la dispersión de la distribución, sino que es necesario definir una medida relativa de la variabilidad: **el coeficiente de variación.**

Algo parecido hay que hacer con las distribuciones bidimensionales, pues tampoco la covarianza da una medida objetiva (comparable) de la correlación de las variables.

3.1. Coeficiente de correlación lineal

El valor del **coeficiente de correlación lineal** es el criterio que se utiliza para medir la fuerza de la correlación entre dos variables.

Este coeficiente, denotado ρ, se define así $\rho = \frac{s_{xy}}{s_x \; s_y}$

Esto es, la razón entre la covarianza de las variables X e Y y el producto de sus desviaciones típicas marginales.

Las propiedades fundamentales del coeficiente de correlación son:

1. El valor de ρ no cambia al hacerlo la escala de medición, pues la covarianza y el producto de las desviaciones típicas varían en la misma proporción.
2. El signo de ρ es el mismo que el de la covarianza, pues las desviaciones típicas siempre son positivas. Luego:
   - Si $\rho > 0$, la correlación es directa.
   - Si $\rho < 0$, la correlación es inversa.

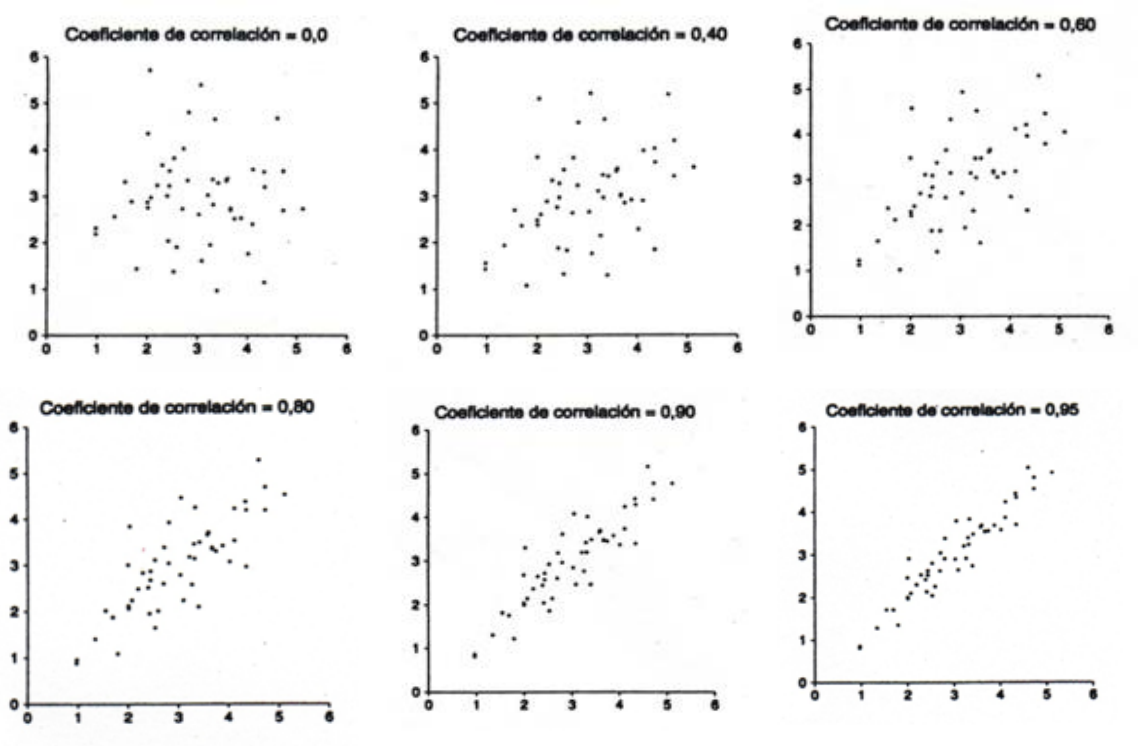
   Este coeficiente mide exclusivamente la correlación lineal entre variables. Por tanto, puede haber otro tipo de correlación no detectada por ρ. Por ejemplo,

ρ no detectaría si la correlación exponencial perfecta que hay entre los puntos (-1, 0,5), (0,1), (1,2) y (4,16), que pertenecen todos a la gráfica de $y = 2^x$

3. El valor de ρ está entre -1 y +1: $-1 \leq \rho \leq 1$
4. Si ρ toma valores cercanos a 0 por izquierda, la correlación es débil (e inversa).
5. Si ρ toma valores cercanos a +1, la correlación es fuerte (y directa).

   El signo de ρ no determina la fuerza de correlación: sólo el sentido. Tampoco indica la mayor o menor pendiente de la recta asociada a la nube de puntos.
6. Si $|\rho| = 1$, la correlación es perfecta. Hay dependencia lineal entre las variables X e Y.
7. Si ρ toma valores cercanos a 0, la correlación es débil.

Los diagramas siguientes aclaran estas propiedades.



Actividades:

4. Hallar el coeficiente de correlación entre el tiempo empleado y el número de errores cometidos por ocho personas al realizar un trabajo de mecanografía (Actividad 1).

5. Hallar el coeficiente de correlación de la distribución dada por la siguiente tabla:

X	4	7	3	9
Y	3	6	7	5

3.2. Coeficiente de determinación

Cuando ρ está próximo a 1 (o a -1), la correlación lineal es fuerte. Esto significa que los cambios en la variable Y se explican, en gran medida, por los cambios de la variable X, en consecuencia, se pueden hacer estimaciones fiables de Y a partir de X.

Una media de esta fiabilidad es ρ^2, siendo ρ el coeficiente de correlación, pues su valor indica la proporción de la variación en la variable Y que puede ser explicada por los cambios de la variable X. A ρ^2 se le llama **coeficiente de determinación.**

Si multiplicamos ρ^2 por 100 se obtiene el porcentaje de cambio de Y explicado por X. Así, si $\rho = 0$, los cambios en la variable explican el 0 % de los cambios en la variable Y, o sea, nada: las variables X e Y son linealmente independientes. Y si $\rho = 1$ (o $\rho = -1$), la variación de Y se explica totalmente, al 100% por la variación de la X; en este caso, las variables X e Y son linealmente dependientes. Fuera de estos límites, el porcentaje explicado es $100\ \rho^2$.

Ejemplos

a) El coeficiente de correlación entre la edad y la altura de niños (Actividad 3), vale $\rho = 0{,}94$. Por tanto, el coeficiente de determinación será $\rho^2 = 0{,}94^2 = 0{,}88$. Esto significa que, en los niños de nuestro ejemplo, el 88% de su altura se explica por la edad; el resto, hasta el 100%, será debido a otras causas: altura de sus padres, dieta, etc.

b) El coeficiente de determinación entre el tiempo empleado y el número de errores cometidos por ocho personas (Actividad 1), vale $\rho^2 = (-0{,}764)^2 = 0{,}584$. O sea, el tiempo empleado explica el 58,4% de las diferencias de errores. El 41,6% de la variación restante se debe a causas desconocidas por nosotros.

4. Regresión lineal

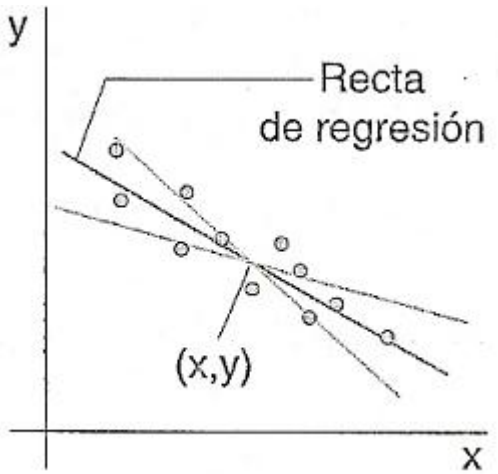
Hasta ahora nos hemos limitado, a decir cuándo dos variables están correlacionadas. Primero, representábamos el diagrama de dispersión y decíamos que, cuando una recta se ajustaba bien a la nube de puntos, entonces la correlación era fuerte. Después, con el coeficiente ρ, dábamos una medida del sentido y fuerza de la corrección. En esta ocasión, sacaremos el máximo provecho a la correlación, hallando qué recta es la que mejor se ajusta a la nube de puntos.

Esta recta nos permitirá calcular qué valor de Y es el que *cabe esperar* para un valor conocido de X. O sea, podremos hacer estimaciones de una variable a partir de la otra.

4.1. Recta de regresión mínimo cuadrática

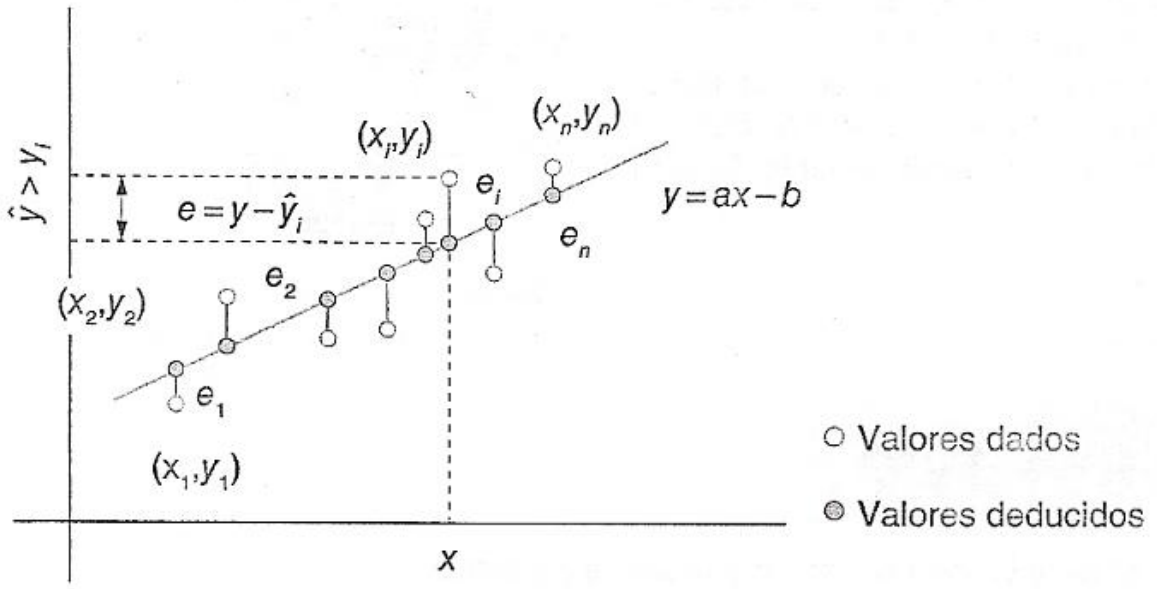
La recta de regresión es la recta que mejor se ajusta a la nube de puntos. Es una recta ideal que asignaría a cada valor x_i de la variable X el promedio de los y_i correspondientes a x_i. En consecuencia, debe pasar por el punto $(\bar{x}, \bar{y})$, centro de gravedad de la distribución bidimensional.

Al ser una recta ideal no tiene por que pasar por ninguno de los puntos dados, pero sí lo más cerca posible de todos ellos. De cualquier manera, siempre se cometerán errores en la estimación. Obviamente, lo que se pretende es que estos errores sean lo más pequeños posibles.



La recta que mejor se ajusta a estos propósitos es la **recta de regresión mínimo cuadrática,** que es aquella que minimiza la suma de los cuadrados de los errores. Esto es, si la ecuación de esta recta es $y = ax + b$, y los puntos dados son $(x_1, y_1), (x_2, y_2), \ldots$ y (x_n, y_n), debe cumplirse que la suma $e_1^{\,2} + e_2^{\,2} + \ldots + e_n^{\,2}$ sea mínima, siendo e_i la diferencia entre el valor dado, y_i, correspondiente a x_i, y el valor deducido $\hat{y}_i$ por la recta para ese mismo x_i:

$$\hat{y}_i = ax_i + b \qquad \text{y} \qquad e_i = |\hat{y}_i - y_i|$$



Con estas condiciones, los valores de la pendiente a y de la ordenada al origen b de esa recta valen:

$$a = \frac{s_{xy}}{s_x^{\,2}} \qquad \text{y} \qquad b = \bar{y} - \frac{s_{xy}}{s_x^{\,2}}\bar{x}$$

Luego, **la ecuación de la recta de regresión** es:

$$y - \bar{y} = \frac{s_{xy}}{s_x^{\,2}}(x - \bar{x})$$

Siendo $\bar{x}$ e $\bar{y}$ las medias marginales de las variables X e Y, $s_x^{\,2}$ la varianza de X y s_{xy} la covarianza. Esta recta de regresión se llama **Y sobre X**, pues se utiliza para

predecir (estimar) los valores de Y a partir de los de X. Si lo que se desea es estimar los valores de X partiendo de los de Y, se empleará la ecuación de la **recta de regresión de X sobre Y**, que es:

$$x - \bar{x} = \frac{s_{xy}}{s_x^{\,2}}(y - \bar{y})$$

Actividades:

6. Con los datos de la Actividad 1:

a) hallar la ecuación de la recta de regresión que permita estimar los errores a partir del tiempo.

b) representar la recta y los puntos dados.

c) estimar el número de errores de una persona que tardase 11 minutos en teclear las 40 líneas.

d) Ídem para otra persona que tardase 9 minutos. Compararlo con datos del problema.

7. a) Hallar la recta que mejor se ajuste a la distribución dada por la siguiente tabla:

X	1	3	4	5	6
Y	3	4	6	6	8

b) mediante la recta obtenida, estimar el valor para Y para x=2 y x=7.

4.2. Fiabilidad de la recta de regresión

La fiabilidad de las estimaciones hechas a partir de la recta de regresión depende fundamentalmente de:

1. **El valor del coeficiente de correlación** ρ

   Una correlación alta (ρ próximo a 1), asegura estimaciones fiables

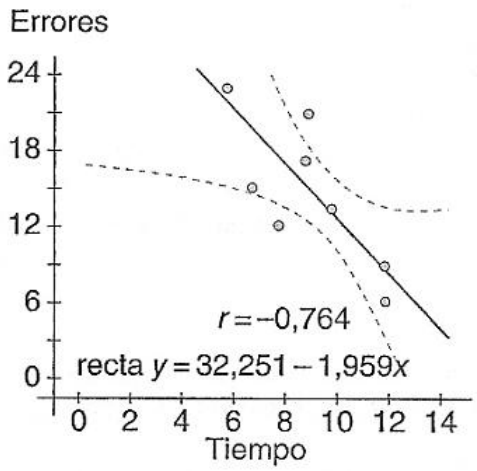
2. **El número de datos considerados**

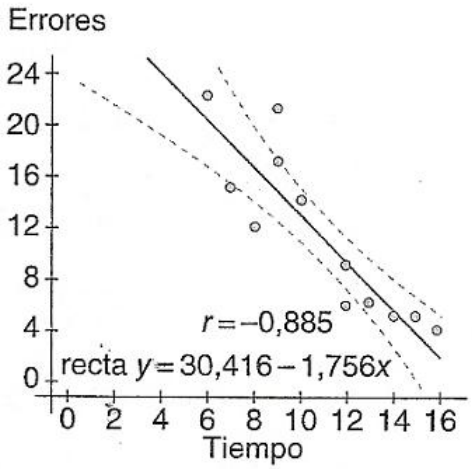
   La fiabilidad aumenta al aumentar los datos. Una recta obtenida a partir de pocos datos genera grandes riesgos, aunque ρ sea muy alto.

3. **La proximidad del valor x_0,** para el que se quiere hacer la estimación, **a la media** $\bar{x}$

   La estimación de y_0 para un x_0 dado es más fiable cuando x_0 está próximo a $\bar{x}$; a medida que x_0 se aleja de $\bar{x}$ la estimación se hace más arriesgada.

Se observa en las siguientes figuras como las líneas de puntos determinan una banda alrededor de la recta de regresión. Esta banda indica los márgenes del valor estimado $\hat{y}_0$, para cada x_0 dado. En este caso, la banda se ha generado para una probabilidad de acierto del 95%.





4.2.1. Observaciones sobre la recta de regresión

1. La banda se ensancha a medida que nos alejamos de la media $\bar{x}$. Esto indica que si la estimación desea hacerse para valores alejados de la media, la probabilidad de acierto es menor.
2. La banda se ensancha más rápidamente cuando el coeficiente de correlación es menor.
3. Para la misma distribución, la estimación será más fiable si aumentan los datos de la muestra considerada.

4.2.2. Limitaciones de la recta de regresión

La recta de regresión debe usarse para hacer estimaciones en valores próximos a los considerados. Pretender una estimación en puntos lejanos puede conducir a soluciones absurdas. Por ejemplo, si con la recta que obtuvimos para la Actividad 1 estimamos el número de errores que cometería una persona que tardase 30 minutos en teclear 40 líneas, se obtendría:

$$y = 32{,}251 - 1{,}959 \,.\, 30 = -26{,}519$$ ¿-27 errores? ¡Absurdo!