Estadística y Probabilidad I
Estadística Interactiva en la Red.
Laboratorio Virtual de Estadística.

Contenidos Teóricos Unidad Temática 2.
Estadística Descriptiva Bidimensional.

A. Gámez, L.M. Marín, R. Huertas y S. Fandiño

Noviembre - 2005

Índice general

2. Estadística Descriptiva Bidimensional **2**

2.1. Introducción: Tablas Bidimensionales, Diagramas, Distribuciones marginales y condicionadas. 2

2.1.1. Introducción: distribución conjunta y tablas de doble entrada 2

2.1.2. Representaciones gráficas 4

2.1.3. Distribuciones marginales. Distribuciones condicionadas 5

2.2. Regresión y Correlación 7

2.2.1. Independencia de variables estadísticas. Dependencia funcional y dependencia estadística 7

2.2.2. Medias, varianzas y covarianzas 8

2.2.3. Ajustes. Método de mínimos cuadrados 9

2.2.4. Regresión lineal mínimo cuadrática 11

2.2.5. Coeficiente de determinación. Coeficiente de correlación lineal 14

Capítulo 2

Estadística Descriptiva Bidimensional

2.1. Introducción: Tablas Bidimensionales, Diagramas, Distribuciones marginales y condicionadas.

2.1.1. Introducción: distribución conjunta y tablas de doble entrada

En muchas ocasiones deseamos estudiar más de un carácter de una población determinada y nos interesa comprobar si existe relación entre dichos caracteres. Por ejemplo, podemos realizar un estudio de la relación entre la edad de los niños y su altura, para hacer una tabla de *alturas segun edad* de utilidad para los profesionales de la medicina. También podríamos preguntarnos si existe relación entre el número de habitantes de un país y su consumo energético, o entre el peso y la altura de sus habitantes, o entre los niveles de colesterol, glucosa, transaminasas y bilirrubina en la sangre. Centraremos nuestro estudio en el caso de dos variables.

Al igual que en el caso de una sola variable, los datos pueden venir presentados de diversas maneras. En el caso de tener pocos datos, pueden presentarse mediante una relación exhaustiva de todas las ocurrencias de las dos variables. Por ejemplo, si estamos estudiando el número de hijos e hijas que tienen los empleados de una empresa, se nos podrían presentar los siguientes datos:

$$(1,0),(2,1),(0,1),(1,1),(0,0),(0,2),(3,1),(1,0),(2,1),(2,0)$$

Este manera de presentar los datos solo es factible cuando se tiene un número muy pequeño de observaciones. Si el número de observaciones es grande, lo que se hace es agrupar los datos, indicando a continuación el número de ocurrencias de cada uno. Esto normalmente se realiza mediante una tabla de doble entrada, indicando en la intersección de cada fila y columna el número de ocurrencias.

En el ejemplo anterior, la tabla de doble entrada correspondiente sería:

	0	1	2
0	1	1	1
1	2	1	0
2	1	2	0
3	0	1	0

	30-50	50-70	70-90	90-110	Total
1.30-1.50	6	0	0	0	6
1.50-1.70	0	15	3	0	18
1.70-1.90	0	1	12	3	16
1.90-2.10	0	0	1	9	10
Total	6	16	16	12	50

Cuadro 2.1: Tabla de doble entrada por intervalos.

También es posible dar los datos con una tabla lineal. La que damos a continuación se refiere también al caso de los hijos e hijas de los empleados de la empresa. Los valores se han tomado de la tabla de doble entrada.

$X = hijos$	0	0	0	1	1	2	2	3
$Y = hijas$	0	1	2	0	1	0	1	1
$frecuencias$	1	1	1	2	1	1	2	1

Si llamamos X a la primera variable, que puede tomar los valores $x_1, x_2, \ldots, x_r$ y llamamos Y a la segunda variable, pudiendo tomar los valores $y_1, y_2, \ldots, y_s$, la tabla de doble entrada sería de la siguiente forma:

	y_1	y_2	$\ldots$	y_j	$\ldots$	y_s	
x_1	n_{11}	n_{12}	$\ldots$	n_{1j}	$\ldots$	n_{1s}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\ldots$	n_{2j}	$\ldots$	n_{2s}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\ldots$	n_{ij}	$\ldots$	n_{is}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	n_{r2}	$\ldots$	n_{rj}	$\ldots$	n_{rs}	$n_{r\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\ldots$	$n_{\bullet j}$	$\ldots$	$n_{\bullet s}$	$N = n$

Donde n_{ij} representa el número de veces que se presenta la observación (x_i, y_j). La última fila se obtiene sumando los elementos de la columna correspondiente y la última columna sumando los elementos de su misma fila:

$$n_{i\bullet} = \sum_j n_{ij} \qquad n_{\bullet j} = \sum_i n_{ij}$$

La suma de los $n_{i\bullet}$ coincide con la suma de los $n_{\bullet j}$ y vale N, el número total de pares de elementos de la muestra.

En el caso de que sean muchos los posibles valores que pueda tomar la variable, agrupamos los datos en intervalos. Obtendremos entonces una tabla de doble entrada equivalente a las tablas de tipo III para la variable unidimensional.

La tabla de doble entrada que puede verse en la tabla 2.1 corresponde a la distribución conjunta de las variables Talla y Peso de los alumnos del instituto cuyos datos están en la tabla **??** de la página **??**. Los datos se han clasificado en cuatro intervalos para el peso y otros cuatro para la talla.

También podemos construir tablas de frecuencias relativas sin más que dividir todos los elementos de la tabla por el número total de datos N. Así pues, denominamos frecuencia relativa de la pareja (x_i, y_j) a

$$fr(x_i, y_j) = f_{ij} = \frac{n_{ij}}{N}$$

Es evidente comprobar que la suma de todas las frecuencias relativas es 1. Análogamente se pueden definir las cantidades $f_{i\bullet}$ y $f_{\bullet j}$.

$$f_{i\bullet} = \sum_j f_{ij} = \frac{n_{i\bullet}}{N} \qquad f_{\bullet j} = \sum_i f_{ij} = \frac{n_{\bullet j}}{N}$$

La tabla de frecuencias relativas correspondiente a las variables peso y talla es la siguiente

	$30-50$	$50-70$	$70-90$	$90-110$	$Total$
$1{,}30-1{,}50$	0,12	0	0	0	$f_{1\bullet}=0{,}12$
$1{,}50-1{,}70$	0	0,30	0,06	0	$f_{2\bullet}=0{,}36$
$1{,}70-1{,}90$	0	0,02	0,24	0,06	$f_{3\bullet}=0{,}32$
$1{,}90-2{,}10$	0	0	0,02	0,18	$f_{4\bullet}=0{,}2$
$Total$	$f_{\bullet 1}=0{,}12$	$f_{\bullet 2}=0{,}32$	$f_{\bullet 3}=0{,}32$	$f_{\bullet 4}=0{,}24$	1

2.1.2. Representaciones gráficas

La representación gráfica más usual es la llamada **Nube de Puntos o Diagrama de Dispersión**. En el plano delimitado por dos ejes que sirvan para representar las variables X e Y se dibuja un punto (x, y) por cada vez que las variables tomen este par de valores. Si coinciden varias observaciones en un mismo punto puede optarse por dibujar un pequeño círculo de radio proporcional a su frecuencia o indicar en la gráfica esta frecuencia al lado del punto.

Otra manera de representar los datos es mediante un **diagrama de barras tridimensional**. Sobre cada punto del plano se levanta una barra de altura proporcional a su frecuencia. Queda por tanto un gráfico tridimensional.

En el caso de que los datos vengan agrupados en intervalos se dibuja un histograma tridimensional, también llamado **estereograma**. Sobre cada uno de los rectángulos determinados por un intervalo de X y otro de Y se levanta un paralepípedo rectángulo. En este caso, su volumen ha de ser proporcional a la frecuencia con que aparecen los puntos contenidos en dicho rectángulo. A continuación aparece un estereograma para las variables Talla y Peso correspondientes a la tabla 2.1 de la página 3.

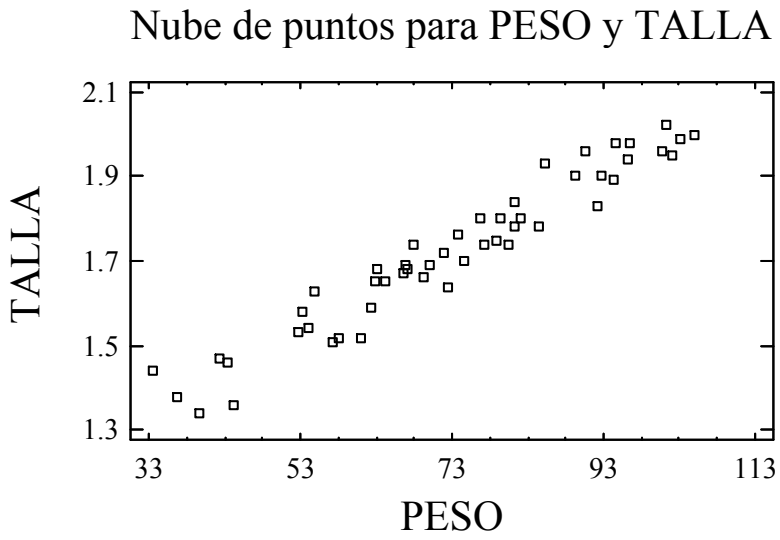
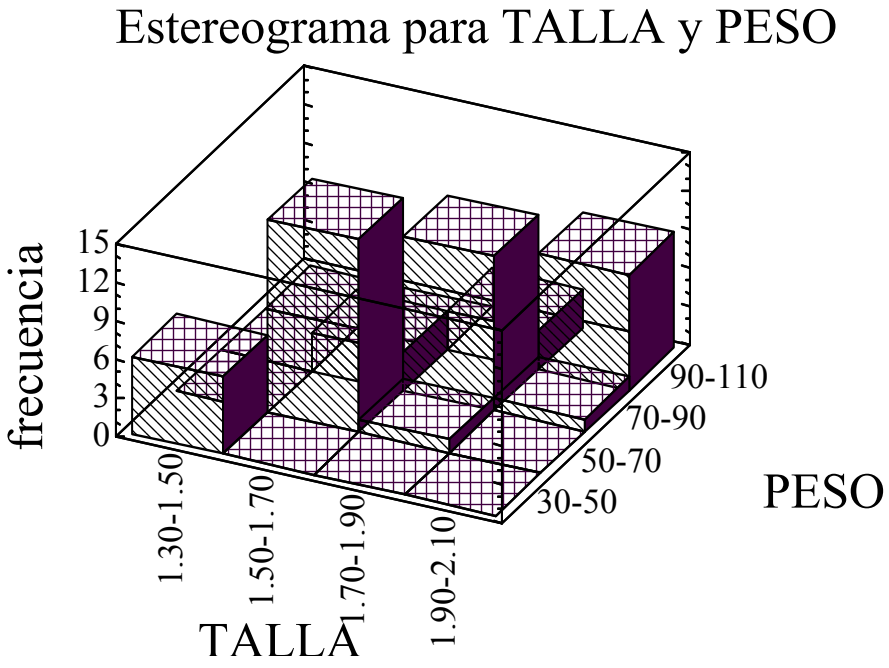


Figura 2.1: Diagrama de Dispersión para Peso y Talla.



2.1.3. Distribuciones marginales. Distribuciones condicionadas

En las distribuciones de frecuencia bidimensionales cabe estudiar por separado cada una de las variables unidimensionales que la componen, haciendo caso omiso de la otra. Estas distribuciones reciben el nombre de *distribuciones marginales.* Son obviamente dos: la distribución marginal de X y la distribución marginal de Y. Las frecuencias absolutas

asociadas a los distintos valores x_i de la variable X son las $n_{i\bullet}$ y las de los y_j son las $n_{\bullet j}$.

Así pues, la distribución marginal de X se obtiene tomando, en la tabla de doble entrada, la primera y última columnas

x_1	$n_{1\bullet}$
x_2	$n_{2\bullet}$
$\vdots$	$\vdots$
x_i	$n_{i\bullet}$
$\vdots$	$\vdots$
x_r	$n_{r\bullet}$
	N

y la marginal de Y tomando la primera y última fila.

y_1	y_2	$\dots$	y_j	$\dots$	y_s	
$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet s}$	N

Las frecuencias relativas de las distribuciones marginales se obtendrán dividiendo las frecuencias absolutas entre el número total de observaciones N. Es decir, la frecuencia relativa de x_i será:

$$fr(x_i) = \frac{n_{i\bullet}}{N} = f_{i\bullet}$$

y la frecuencia relativa de y_j será:

$$fr(y_j) = \frac{n_{\bullet j}}{N} = f_{\bullet j}$$

Las tablas de frecuencia relativa para la marginal correspondiente a la variable talla sería

Talla	frec. relativa
$1{,}30 - 1{,}50$	$f_{1\bullet} = 0{,}12$
$1{,}50 - 1{,}70$	$f_{2\bullet} = 0{,}36$
$1{,}70 - 1{,}90$	$f_{3\bullet} = 0{,}32$
$1{,}90 - 2{,}10$	$f_{4\bullet} = 0{,}2$
Total	1

y la de la variable peso:

Peso	frec. relativa
$30 - 50$	$f_{\bullet 1} = 0{,}12$
$50 - 70$	$f_{\bullet 2} = 0{,}32$
$70 - 90$	$f_{\bullet 3} = 0{,}32$
$90 - 110$	$f_{\bullet 4} = 0{,}24$
Total	1

En otras ocasiones nos interesará analizar los datos obtenidos por una de las variables cuando se presenta exactamente un determinado valor de la otra variable. Esta idea da lugar a las llamadas *distribuciones condicionadas de frecuencias.*

Podemos estudiar la distribución de X condicionada a que la variable Y tome el valor y_j . A esta variable la denotaremos por X/y_j, obteniéndose a partir de la primera columna y la correspondiente al valor y_j .

x_1	n_{1j}
x_2	n_{2j}
$\vdots$	$\vdots$
x_i	n_{ij}
$\vdots$	$\vdots$
x_r	n_{rj}
$Total$	$n_{\bullet j}$

También podemos estudiar la distribución de Y condicionada a que la variable X tome el valor x_i . A esta variable la denotaremos por Y/x_i, obteniéndose a partir de la primera fila y la correspondiente al valor x_i .

y_1	y_2	$\ldots$	y_j	$\ldots$	y_s	$Total$
n_{i1}	n_{i2}	$\ldots$	n_{ij}	$\ldots$	n_{is}	$n_{i\bullet}$

Las frecuencias relativas de las distribuciones condicionadas se obtendrán dividiendo las frecuencias absolutas entre el número total de observaciones que cumplen la condición requerida, que en los casos anteriores son, respectivamente, $n_{\bullet j}$ y $n_{i\bullet}$.

Es decir, la frecuencia relativa de x_i/y_j será:

$$fr(x_i/y_j) = \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij}}{f_{\bullet j}}$$

y la frecuencia relativa de y_j/x_i será:

$$fr(y_j/x_i) = \frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}}$$

La distribución del peso, condicionada a que la talla esté en el intervalo 1.70-1.90 viene dada en la tabla siguiente, donde se muestran la frecuencias absolutas de cada uno de los pesos de los individuos de cuya talla está en el intervalo de 1.70 a 1.90.

Peso	$30-50$	$50-70$	$70-90$	$90-110$	*Total*
Frecuencia absoluta	0	1	12	3	16

Para obtener la tabla de frecuencias relativa hay que dividir estas frecuencias por el total de individuos de la talla considerada, que en este caso son 16.

La tabla de frecuencias relativa para la distribución condicionada resulta:

Peso	$30-50$	$50-70$	$70-90$	$90-110$	*Total*
Frecuencia relativa	0	$\frac{1}{16} = 0{,}0\,625$	$\frac{12}{16} = 0.75$	$\frac{3}{16} = 0.187\,5$	$\frac{16}{16} = 1$

2.2. Regresión y Correlación

2.2.1. Independencia de variables estadísticas. Dependencia funcional y dependencia estadística

Dos variables estadísticas se dicen dependientes cuando el conocimiento de que se ha presentado una determinada ocurrencia en una de ellas condiciona en algún sentido el valor que pueda tomar la otra. Así si observamos la nube de puntos de las variable peso

y talla de la figura 2.1 de la página 5, podemos ver que los valores bajos de talla dan valores bajos para el peso, y en cambio, a las tallas mayores corresponden pesos mayores. A la vista de los valores de los tallas sabemos que están comprendidas entre 1.34 y 2.02. Sin embargo conociendo el peso de una persona podemos dar una información más precisa sobre su talla. Así, mirando la nube de puntos observamos que las personas cuyo peso es aproximadamente 73, tienen una altura entre 1.60 y 1.80. En cambio las personas que pesan más o menos 99 tienen una talla comprendida entre 1.90 y 2. Por tanto, conociendo el peso de una persona podríamos obtener alguna información suplementaria sobre su peso.

Damos, a partir de esta idea, la siguiente definición: dos variables X e Y (que constituyen una variable bidimensional) son *independientes* si las distribuciones de frecuencias relativas de la variable X condicionada a cualquier valor y_j de la variable Y son todas idénticas, sin depender del valor de y_j. La distribución de X no depende del valor que tome la variable Y.

Es decir:

$$fr(x_i/y_1) = fr(x_i/y_2) = \ldots = fr(x_i/y_s)$$

De aquí se deduce que

$$fr(x_i/y_j) = f_{i\cdot} \qquad \forall i = 1, 2, \ldots, r$$

Es decir, las distribuciones condicionadas de X, coinciden exactamente con la distribución de frecuencias relativas marginal de X.

Debemos resaltar la diferencia entre dependencia funcional y dependencia estadística. Cuando la variable Y depende funcionalmente de la variable X eso significa que conociendo el valor que toma la variable X tenemos perfectamente determinado el valor que tomará la variable Y. Sin embargo, cuando se produce dependencia estadística eso significa que al conocer el valor que toma la variable X obtenemos "alguna" información sobre la distribución de frecuencias de los valores de la variable Y, pero no obtenemos un valor concreto para esta variable.

2.2.2. Medias, varianzas y covarianzas

Tanto las distribuciones marginales como condicionadas que hemos visto son distribuciones unidimensionales, y por tanto podemos calcular para ellas todas las medidas descriptivas expuestas en el apartado correspondiente a variables unidimensionales, sin más que tener en cuenta las frecuencias relativas de cada caso. En particular, la media de la distribución marginal de X será:

$$\overline{X} = E[X] = \sum_{i=1}^{r} f_{i\bullet} x_i$$

Del mismo modo, la media de la distribución marginal de Y será:

$$\overline{Y} = E[Y] = \sum_{j=1}^{s} f_{\bullet j} y_j$$

Análogamente pueden calcularse las respectivas varianzas.

Para las distribuciones condicionadas tendremos los siguientes valores:

$$E[X/y_j] = \sum_{i=1}^{r} \frac{fij}{f_{\bullet j}} x_i$$

$$E[Y/x_i] = \sum_{j=1}^{s} \frac{fij}{f_{i\bullet}} y_j$$

Es inmediato deducir que si X e Y son independientes se verifica que

$$E[X/y_j] = E[X] \qquad E[Y/x_i] = E[Y]$$

Las medidas vistas hasta ahora corresponden a distribuciones unidimensionales. También existen parámetros conjuntos para ambas variables, característicos de la distribución bidimensional y que, como veremos más adelante, van a estar ligados a la dependencia de las variables. Una de estas medidas recibe el nombre de *covarianza* de las variables X e Y:

$$\text{cov}(X,Y) = S_{XY} = \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}(x_i - \overline{X})(y_j - \overline{Y})$$

Esta fórmula puede simplificarse hasta quedar:

$$\text{cov}(X,Y) = \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}x_i y_j - \overline{X}\,\overline{Y}$$

Si las dos variables son independientes, se verifica que

$$\text{cov}(X,Y) = 0$$

Nota: El recíproco no es cierto, pues si $\text{cov}(X,Y) = 0$, entonces no significa que X e Y sean independientes.

En muchas situaciones prácticas es frecuente encontrar que existe una cierta dependencia de tipo estadístico entre dos variables. Así, si estudiamos el peso de un coche y su gasto de combustible observaremos que guardan una cierta relación. Una relación de dependencia es de tipo funcional cuando podemos encontrar una función matemática de modo que para cada valor de X podamos encontrar el valor correspondiente de Y. En las dependencias de tipo estadístico, sin embargo, no es posible establecer tal función, y lo normal es que a un valor determinado de X le puedan corresponder distintos valores de Y.

Si se representa la nube de puntos correspondiente a los datos observados es posible establecer la relación de dependencia entre las variables. En los casos de dependencia funcional se podría encontrar una función cuya gráfica pasara por todos los puntos dibujados. En el caso de la dependencia estadística se podría encontrar una función de modo que la distancia entre la nube de puntos y su gráfica sean pequeños.

En la figura 2.2 se consideran ejemplos de nubes de puntos entre los que existe dependencia estadística de tipo lineal entre variables. En la figura 2.3 hay un primer diagrama en el que no existe dicha dependencia estadística y otro ejemplo en el que existiendo dependencia estadística no es de tipo lineal.

2.2.3. Ajustes. Método de mínimos cuadrados

Consideramos N observaciones que son pares de valores del tipo (x_i, y_i). Si tomamos como variable independiente a X y como variable dependiente a Y, debemos de hallar

Figura 2.2: Dependencia estadistica de tipo lineal entre variables

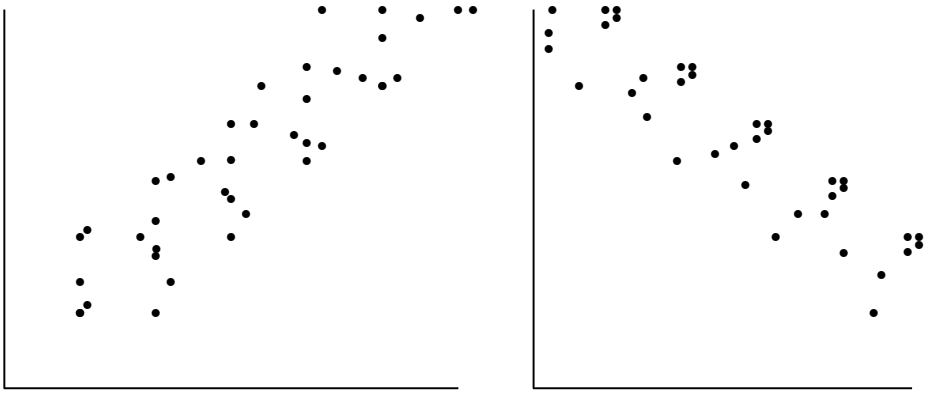
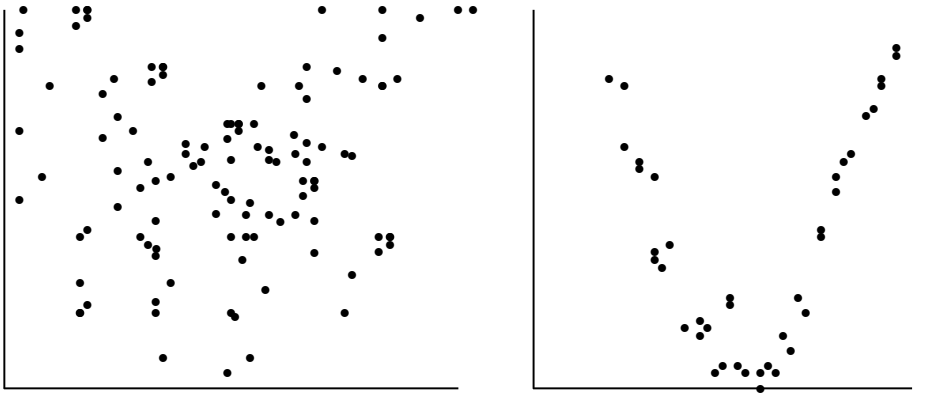


Figura 2.3: No existe dependencia estadistica entre variables o bien no es lineal



una expresión de Y como función de X. En este caso diremos que estamos realizando una regresión de Y sobre X.

La función de regresión, $y = f(x)$, puede adoptar distintas formas: línea recta, parábola, polinomio de grado n, función exponencial, etcétera. El modo usual de proceder es prefijar el tipo de función que se va a considerar. Todas las posibles funciones de ese tipo tendrán una formulación general que dependerá de unos parámetros. La determinación de dichos parámetros se hará de modo que los valores observados estén "próximos" a los puntos de la función de regresión.

Consideremos como variable independiente a X y como variable dependiente a Y. Si hemos observado un punto (x_i, y_i), llamamos y_i^* al valor previsto por la función de regresión para x_i, es decir $y_i^* = f(x_i)$. Denominaremos *error o residuo* y lo denotaremos por e_i a la diferencia entre el valor observado y el valor previsto, es decir:

$$e_i = y_i - y_i^*$$

Lógicamente se desea que los residuos o errores sean lo más pequeños posible. El método que más se utiliza para obtener los parámetros es el de *ajuste por mínimos cuadrados*, que consiste en obtener el valor de los parámetros que hagan mínima la suma de los cuadrados de los residuos. Es decir, habría que minimizar la expresión

$$\sum_{i=1}^{N} e_i^2$$

2.2.4. Regresión lineal mínimo cuadrática

Estudiaremos el caso más sencillo y de mayor importancia, que es aquel en que la función de regresión es una línea recta. La expresión general de una línea recta será del tipo

$$y = a + bx$$

con lo que tendremos que los valores previstos o predichos por la regresión serían

$$y_i^* = a + bx_i$$

de donde deducimos que

$$e_i = y_i - y_i^* = y_i - a - bx_i$$

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - y_i^*)^2 = \sum_{i=1}^{N} (y_i - a - bx_i)^2$$

Para determinar los valores de los parámetros a y b que minimizan dicha suma igualamos a cero las derivadas parciales respecto a los parámetros a y b.

$$\begin{aligned} -2\sum_{i=1}^{N} (y_i - a - bx_i) &= 0 \\ -2\sum_{i=1}^{N} (y_i - a - bx_i)x_i &= 0 \end{aligned}$$

Operando obtenemos que las ecuaciones se convierten en:

$$\begin{cases} \sum_{i=1}^{N} y_i &= b\sum_{i=1}^{N} x_i + Na \\ \sum_{i=1}^{N} y_i x_i &= b\sum_{i=1}^{N} x_i^2 + a\sum_{i=1}^{N} x_i \end{cases}$$

que recibe el nombre de *sistema de ecuaciones normales.*

Dividiendo ambas ecuaciones por N obtenemos:

$$\begin{array}{rcl} \overline{Y} &=& b\,\overline{X} + a \\ \frac{\sum_{i=1}^{N} y_i x_i}{N} &=& b\ \frac{\sum_{i=1}^{N} x_i^2}{N} + a\overline{X} \end{array}$$

Para calcular los valores de a y b que son la únicas incognitas de este sistema hay que resolverlo. La primera ecuación del sistema

$$\overline{Y} = a + b\overline{X} \tag{2.1}$$

nos indica que la recta de regresión de Y sobre X pasa por el punto $(\overline{X}, \overline{Y})$, que es el centro de gravedad de la nube de puntos.

Despejando en esta ecuación el valor de a y sustituyendo en la segunda ecuación del sistema obtenemos:

$$b = \frac{S_{XY}}{S_X^2} \tag{2.2}$$

que nos indica que el parámetro b de la recta de regresión puede calcularse como el cociente entre la covarianza y la varianza de la variable independiente. Este parámetro, llamado *coeficiente de regresión de Y sobre X*, representa la pendiente de la recta. Por tanto una expresión de la recta de regresión es

$$y - \overline{Y} = \frac{S_{XY}}{S_X^2}(x - \overline{X}) \tag{2.3}$$

que se obtiene usando la ecuación *punto-pendiente* de una recta.

Usando las expresiones 2.1 y 2.2, o también operando en la ecuación 2.3 obtenemos que

$$a = \overline{Y} - \frac{S_{XY}}{S_X^2}\,\overline{X}$$

Calculamos ahora la covarianza y la recta de regresión correspondiente a los datos de la siguiente tabla que se refieren a los hijos e hijas de los empleados de la empresa.

	0	1	2	*Marginal de X*
0	1	1	1	3
1	2	1	0	3
2	1	2	0	3
3	0	1	0	1
Marginal de Y	4	5	1	$total = 10$

Comenzamos hallando la media de X e Y, ya que son necesarias para evaluar la covarianza e igualmente los coeficientes de la recta de regresión.

$$\overline{X} = \frac{3\times0+3\times1+3\times2+1\times3}{10} = 1.2$$
$$\overline{Y} = \frac{4\times0+5\times1+1\times2}{10} = 0.7$$
$$S_{XY} = \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}(x_i - \overline{X})(y_j - \overline{Y}) =$$
$$= \tfrac{1}{10}(0-1{,}2)(0-0{,}7) + \tfrac{1}{10}(0-1{,}2)(1-0{,}7) + \tfrac{1}{10}(0-1{,}2)(2-0{,}7) +$$
$$+ \tfrac{2}{10}(1-1{,}2)(0-0{,}7) + \tfrac{1}{10}(1-1{,}2)(1-0{,}7) +$$
$$+ \tfrac{1}{10}(2-1{,}2)(0-0{,}7) + \tfrac{2}{10}(2-1{,}2)(1-0{,}7) +$$
$$+ \tfrac{1}{10}(3-1{,}2)(1-0{,}7) = -0{,}04$$

Para hallar la recta de regresión calculamos también la varianza de X.
Usando la expresión alternativa de la varianza:

$$\frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \quad \overline{x}^2 = \frac{1}{10}(3\times0^2 + 3\times1^2 + 3\times2^2 + 1\times3^2) - 1{,}2^2 = 0.96$$

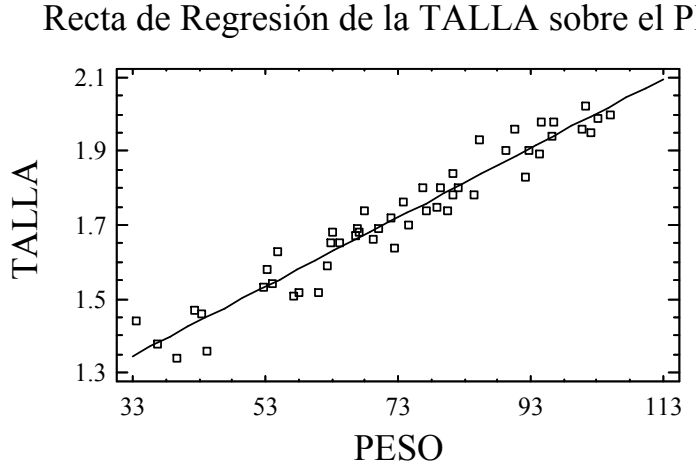
La recta de regresión es

$$(y - 0{,}7) = \tfrac{-0{,}04}{0{,}96}(x - 1{,}2)$$
$$y = -4.1667\times10^{-2}x + 0{,}75$$

Si calculamos la recta de regresión correspondiente a la nube de puntos representada en la figura 2.1 de la página 5, obtenemos

$$Talla = 0{,}0094 \times peso + 1{,}038$$

La representación gráfica de esta última recta de regresión puede verse en la siguiente figura:



2.2.5. Coeficiente de determinación. Coeficiente de correlación lineal

Si consideramos el caso de una regresión de Y sobre X, para medir el grado de dependencia estadística entre X e Y puede utilizarse el llamado *coeficiente de determinación*, que denotaremos por R^2, y que se calcula como

$$R^2 = \frac{S_{y^*}^2}{S_y^2}$$

El numerador es la varianza de los valores calculados para cada x_i mediante la función de regresión y el denominador es la varianza de los valores observados para Y. Se demuestra que R^2 sólo puede tomar valores en el intervalo [0, 1].

Si R^2 vale 1 nos indica que existe una dependencia exactamente funcional, todos los puntos observados están sobre la gráfica de la función de regresión obtenida. En cambio, si R^2 vale 0, entonces el modelo de regresión seleccionado no explica nada sobre la variación de Y. Si R^2 está próximo a 1 se acepta que el modelo explica la relación de dependencia.

Todo lo anterior es válido para el caso de una función de regresión cualquiera, sin importar la forma que adopte. En el caso particular de regresión de tipo lineal podemos calcular el *coeficiente de correlación lineal* mediante la expresión:

$$r = \frac{S_{XY}}{S_X S_Y} \ ; \ \text{donde se puede comprobar que } -1 \leq r \leq 1,$$

es decir, como el cociente entre la covarianza y el producto de las desviaciones típicas de X e Y. Se verifica que $r^2 = R^2$, pero mientras que R^2 puede calcularse para cualquier tipo de regresión, r sólo tiene sentido en el caso de regresión lineal.

El valor obtenido para estos parámetros en el caso de los hijos y las hijas de los empleados es:

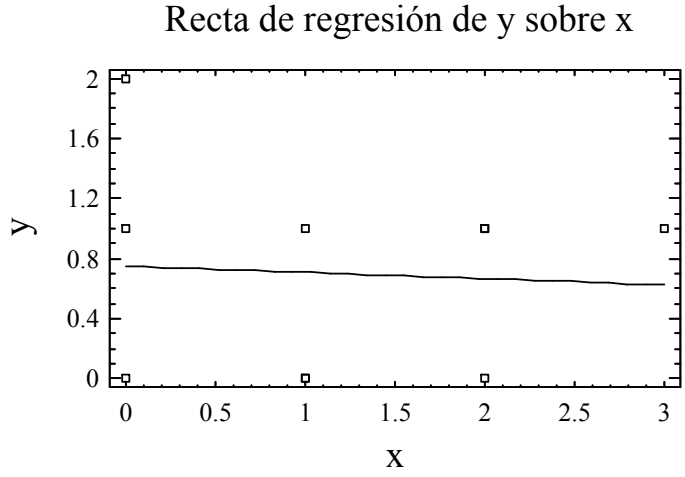
$$r = \frac{S_{XY}}{S_X S_Y} = \frac{-0{,}04}{\sqrt{0{,}96}\sqrt{0{,}41}} = -0{,}06375$$

donde la varianza de Y se ha obtenido de la expresión

$\frac{1}{n}\sum_{j=1}^{k} n_i y_i^2 - \quad \overline{y}^2 = \frac{1}{10}(4 \times 0^2 + 5 \times 1^2 + 1 \times 2^2) - 0{,}7^2 = 0.41$

Podemos obtener $R^2 = (-0{,}06375)^2 = 0{,}004\,064$.

Como estos valores son próximos a 0, concluimos que el ajuste es muy pobre. Es decir que los datos no están cerca de la recta de regresión. En efecto, esto puede apreciarse en la siguiente gráfica que nos da la representación gráfica de los puntos y de la recta de regresión.



Si cálculamos ambos parámetros en el caso de regresión lineal de la variable talla sobre la variable peso de los alumnos del instituto obtendremos:

$$r = \textit{coeficiente de Correlación} = 0.967641$$
$$R^2 = r^2 = \textit{coeficiente de Determinación} = 0.93633$$

Como estos valores son cercanos al valor 1, nos indican un buen ajuste de los puntos a la recta de regresión.

El signo de r coincide con el de S_{XY}. Si $r > 0$ la recta tiene pendiente positiva, es decir cuando una variable crece la otra también. Si $r < 0$ cuando una variable crece la otra decrece.

Si las variables son independientes, la covarianza es nula, y por tanto $r = 0$. El recíproco no tiene por qué ser cierto.

La teoría de regresión nos permite hacer predicciones del valor que tomará la variable dependiente conociendo el valor que toma la variable independiente, sustituyendo el valor de esta última en la función de regresión. Hay que tener en cuenta, sin embargo, que las predicciones tienen mayor validez si se consideran valores de la variable cercanos a su media. Conforme los valores van estado más alejados de la media más arriesgada será la predicción, y por tanto existen riesgos en las extrapolaciones.

En la parte correspondiente a los ejercicios y ejemplos de este tema, vamos a describir otros tipos de ajustes como los parabólicos, logarítmicos, exponenciales, hiperbólicos, etc.