Estadística Descriptiva e Introducción a la Probabilidad

Juan Antonio Maldonado Jurado

Departamento de Estadística e Investigación Operativa
Universidad de Granada

UNIVERSIDAD DE GRANADA

Doble Grado en Ingeniería Informática y Matemáticas

Tema 2. Estadística descriptiva bidimensional. Regresión y correlación

Introducción

Cuando se estudia una población estadística, lo habitual es que se estudien al mismo tiempo varias características sobre cada uno de sus elementos.

Consideremos, por ejemplo, que nuestro objetivo es estudiar las causas que originan los distintos niveles de los gastos de los individuos de cierta población. Además de la mencionada variable, que normalmente se medirá en una muestra representativa de la población estadística, nos interesará medir otras características que pensamos que están relacionadas con ella: ingresos del individuo, estado civil, número de habitantes del municipio donde reside, forma de locomoción que emplea con más frecuencia, aficiones que tiene, edad, etc.

Todas estas características influirán en distinto grado en los niveles de gastos y nos podrán explicar su comportamiento. En general, a mayores ingresos existirá un mayor gasto, los tramos de edad más bajos gastarán más, ya que tendrán más movilidad y mayores aficiones lúdicas que comportan un mayor dispendio. Como es lógico, podrá estudiarse separadamente cada característica construyendo su distribución unidimensional y calculando sus medidas de posición, dispersión, forma, etc.; pero lo normal es presentar conjuntamente más de una característica con el propósito de estudiar sus posibles relaciones y responder a cuestiones como las siguientes, por ejemplo: ¿en qué medida el nivel de ingresos determina el nivel de gastos?, ¿existe relación entre el nivel de gastos y la edad?, ¿y el estado civil?, ¿y el tamaño del municipio?, ..
Con el propósito de dar tratamiento estadístico a estas cuestiones, surgen las **distribuciones conjuntas**, que vamos a estudiar a continuación.

Distribución conjunta de dos caracteres estadísticos

Supongamos una población formada por n individuos en la que se desea estudiar simultáneamente dos caracteres, que representaremos por X e Y. Dichos caracteres podrán ser ambos cualitativos, uno cualitativo y el otro cuantitativo (discreto o continuo), o bien ambos cuantitativos (los dos discretos, uno discreto y otro continuo, o los dos continuos).
Si designamos por $x_1, x_2, \ldots, x_i, \ldots, x_k$ las k modalidades posibles del carácter X, y por $y_1, y_2, \ldots, y_j, \ldots, y_p$ las p modalidades posibles para el carácter Y, las observaciones correspondientes a cada individuo serán de la forma (x_i, y_j), par ordenado que representa las modalidades tomadas por dicho individuo en los caracteres X e Y respectivamente.

Llamemos:

- n_{ij}: número total de individuos en la población que presentan simultáneamente la modalidad x_i del carácter X y la modalidad y_j del carácter Y, denominado **frecuencia absoluta** del par (x_i, y_j).
- f_{ij}: proporción de individuos en la población que presentan simultáneamente la modalidad x_i del carácter X y la modalidad y_j del carácter Y, denominado **frecuencia relativa** del par (x_i, y_j). Por definición de proporción sobre el total, obviamente se tiene que:

$$f_{ij} = \frac{n_{ij}}{n}; \quad i = 1, 2, \ldots, k; \quad j = 1, 2, \ldots, p.$$

Por el principio de incompatibilidad y exhaustividad de las modalidades, se tiene:

$$\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij} = n; \quad \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{p} \frac{n_{ij}}{n} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{p} n_{ij} = \frac{1}{n}n = 1$$

La distribución representada por $\{(x_i, y_j; n_{ij})\}_{i=1,\ldots,k;j=1,\ldots,p}$ recibe el nombre de **distribución conjunta** de los caracteres X e Y.

Distribución conjunta de dos caracteres estadísticos

De toda distribución conjunta es posible extraer, por cuanto a frecuencias absolutas y relativas se refiere, la información numérica que representaremos por:

- $n_{i.}$: número total de individuos que presentan la modalidad x_i del carácter X sin tener en cuenta las modalidades que puedan tomar para el carácter Y. Así:

$$n_{i.} = \sum_{j=1}^{p} n_{ij} \; ; \quad i = 1, 2, \ldots, k.$$

- $f_{i.}$: proporción de individuos que presentan la modalidad x_i del carácter X, sin tener en cuenta el carácter Y. Así:

$$f_{i.} = \sum_{j=1}^{p} f_{ij} = \frac{n_{i.}}{n} \; ; \quad i = 1, 2, \ldots, k.$$

- $n_{.j}$: número total de individuos que presentan la modalidad y_j del carácter Y sin tener en cuenta las modalidades que puedan tomar para el carácter X. Así:

$$n_{.j} = \sum_{i=1}^{k} n_{ij} \; ; \quad j = 1, 2, \ldots, p.$$

- $f_{.j}$: proporción de individuos que presentan la modalidad y_j del carácter Y, sin tener en cuenta el carácter X. Así:

$$f_{.j} = \sum_{i=1}^{k} f_{ij} = \frac{n_{.j}}{n} \; ; \quad j = 1, 2, \ldots, p.$$

Para estos parámetros así definidos se tiene que:

$$n = \sum_{i=1}^{k} n_{i.} = \sum_{j=1}^{p} n_{.j}; \quad 1 = \sum_{i=1}^{k} f_{i.} = \sum_{j=1}^{p} f_{.j}$$

Tablas estadísticas bidimensionales

La tabla estadística más habitual para describir a los n individuos de la población por los caracteres X e Y simultáneamente es la denominada **tabla de doble entrada**, en la que se disponen las modalidades de un carácter en la primera fila y las del otro en la primera columna. En la confluencia de la i-ésima fila con la j-ésima columna se colocará la frecuencia absoluta n_{ij} correspondiente al par (x_i, y_j), $\quad i = 1, \ldots, k;\ j = 1, \ldots, p$.

$X \quad Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_p	$n_{i.}$
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1p}	$n_{1.}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2p}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ip}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	n_{kp}	$n_{k.}$
$n_{.j}$	$n_{,1}$	$n_{,2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.p}$	n

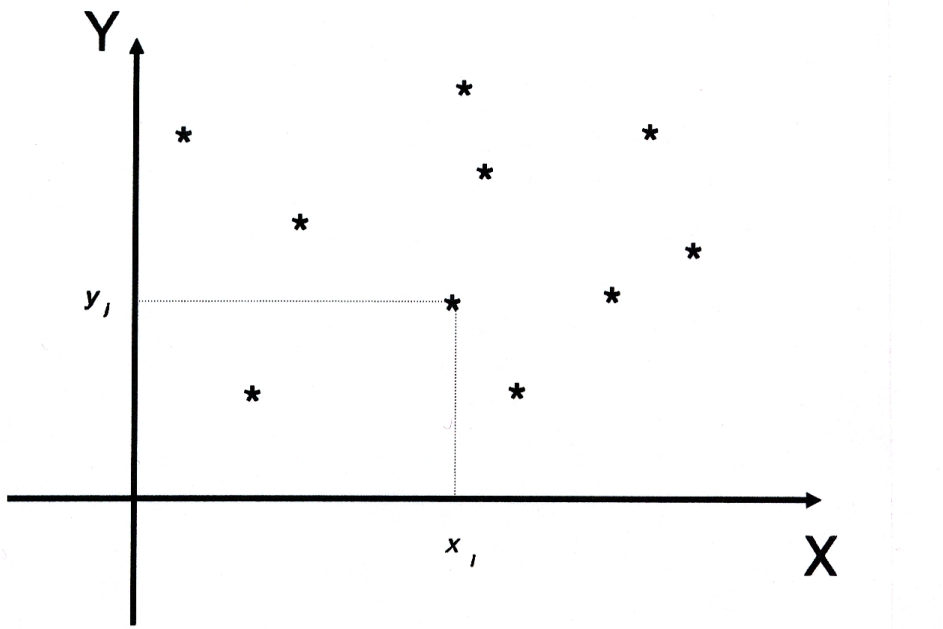
Así, por ejemplo, n_{21} indica el número de veces que se ha observado en la población el par (x_2, y_1), correspondiente a la segunda fila y primera columna de la tabla.

Si la distribución bidimensional corresponde a caracteres cualitativos, o atributos, la tabla de doble entrada recibe el nombre de **tabla de contingencia**. En ella, los valores de la primera fila y primera columna corresponderán a las modalidades para los caracteres X e Y respectivamente (o bien Y y X respectivamente, ya que la colocación de las variables en la tabla es indiferente).

A todo carácter cuantitativo discreto le corresponderá en la tabla una primera fila o columna formada por los valores numéricos que constituyen las modalidades para dicho carácter; mientras que a todo carácter cuantitativo continuo corresponderá una primera fila o columna formada por los intervalos de clase para dicho carácter, que constituyen las modalidades para el mismo.

Representaciones gráficas

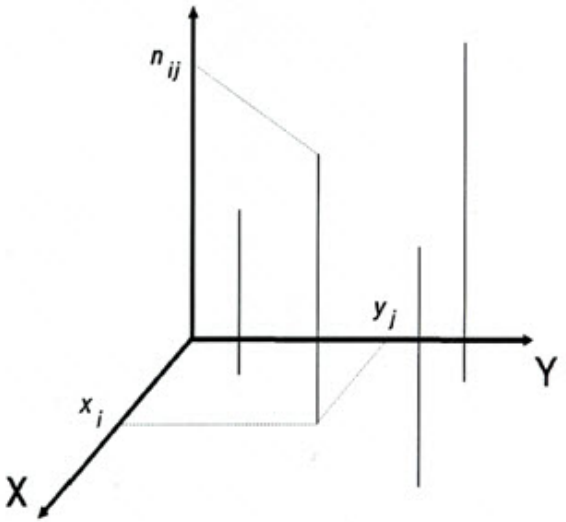
Las dos representaciones gráficas más utilizadas en el caso bidimensional son el **diagrama de dispersión o nube de puntos** y el **estereograma**.
La primera consiste en representar cada par de observaciones (x_i, y_j) por un punto o tantos puntos como indique su frecuencia, en la confluencia de dichos valores dentro de un espacio euclídeo bidimensional.
Así, la distribución vendrá representada por un conjunto de puntos en el plano euclídeo, de donde procede el nombre de nube de puntos. En ocasiones, cuando el par de valores o modalidades está repetido, junto a la representación del punto correspondiente se suele indicar el valor de la frecuencia absoluta de dicho par.

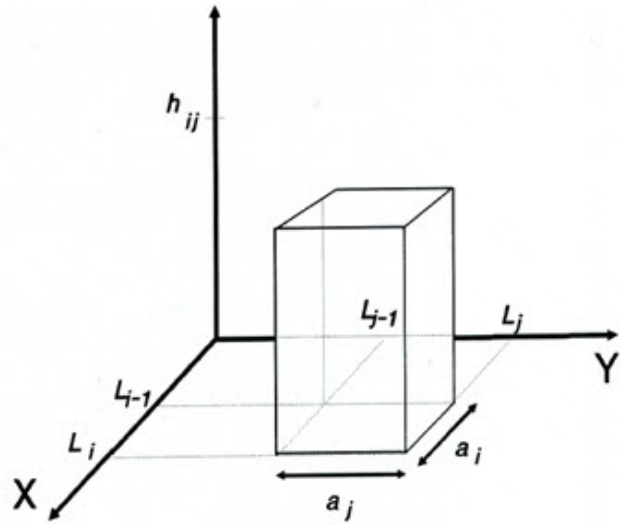


Esta representación gráfica se utilizará para el caso de caracteres cuantitativos fundamentalmente, teniendo en cuenta que si son continuos se utilizarán las marcas de clase de los intervalos como valores a situar en los ejes.

Representaciones gráficas

El estereograma es un gráfico tridimensional, formado por barras (en el caso de dos caracteres cuantitativos discretos), prismas con base común (en el caso de caracteres cualitativos), o prismas con base común o distinta (en el caso de caracteres cuantitativos continuos).
En el primer caso, la altura de cada barra, situada en el punto de confluencia de cada par de observaciones, será igual a la frecuencia absoluta del par.





Para caracteres cuantitativos continuos, los prismas (que no tienen por qué tener igual base) son tales que su volumen es igual a la frecuencia absoluta de cada pareja de intervalos de clase, teniendo por altura el cociente h_{ij} entre la frecuencia absoluta n_{ij} y el producto de las amplitudes de cada intervalo de clase.

$$h_{ij} = \frac{n_{ij}}{(L_i - L_{i-1})(L_j - L_{j-1})}$$

Distribuciones marginales

Las modalidades del carácter X, junto con las frecuencias $n_{i.}$ de la columna marginal de la tabla forman la denominada **distribución marginal del carácter** X, $\{x_i; n_{i.}\}_{i=1,2,\ldots,k}$, mientras que las modalidades del carácter Y, junto con las frecuencias de la fila marginal de la tabla forman la **distribución marginal del carácter** Y, $\{y_j; n_{.j}\}_{j=1,2,\ldots,p}$.

Ambas son distribuciones unidimensionales, a las que es posible dar el tratamiento oportuno, ya visto en anteriores temas. Como ya expusimos, al tratarse de este tipo de distribuciones, su presentación tabular adopta la forma

X	$n_{i.}$	$f_{i.}$
x_1	$n_{1.}$	$f_{1.}$
x_2	$n_{2.}$	$f_{2.}$
$\vdots$	$\vdots$	$\vdots$
x_i	$n_{i.}$	$f_{i.}$
$\vdots$	$\vdots$	$\vdots$
x_k	$n_{k.}$	$f_{k.}$
	n	1

Y	$n_{.j}$	$f_{.j}$
y_1	$n_{,1}$	$f_{,1}$
y_2	$n_{,2}$	$f_{,2}$
$\vdots$	$\vdots$	$\vdots$
y_j	$n_{.j}$	$f_{.j}$
$\vdots$	$\vdots$	$\vdots$
y_p	$n_{.p}$	$f_{.p}$
	n	1

para el caso de la distribución marginal de X, o de Y, respectivamente.
Hay que decir, no obstante, que estos datos pueden irse añadiendo a la propia tabla de doble entrada, en columnas sucesivas o filas sucesivas a las de los propios recuentos, de manera que sea posible calcular lo que en cada momento necesitemos.

Distribuciones condicionadas

En ocasiones, puede resultar de interés el estudio de un carácter sólo sobre los individuos que presentan una modalidad (o algunas modalidades) del otro carácter; así, por ejemplo, podría ser interesante estudiar el carácter X sólo en la subpoblación formada por aquellos individuos que presentan la modalidad y_j del carácter Y. Dicho estudio corresponde al de un carácter unidimensional (en este caso X), descrito por las frecuencias de la j-ésima columna de la tabla de doble entrada. En ella, n_{ij}; $i = 1, 2, \ldots, k$ individuos en $n_{\cdot j}$ presentan la modalidad x_i del carácter x. Se dice también que la frecuencia relativa de la modalidad x_i de X en aquellos individuos que presentan la modalidad y_j de Y es

$$f_{i/j} \equiv f_i^j = \frac{n_{ij}}{n_{\cdot j}}; \quad i = 1, 2, \ldots, k.$$

Análogamente se define la distribución condicionada del carácter Y sobre aquellos individuos que presentan la modalidad x_i del carácter X, y que vendrá determinada por la i-ésima fila de la tabla de doble entrada. Las frecuencias relativas condicionadas serán ahora:

$$f_{j/i} \equiv f_j^i = \frac{n_{ij}}{n_{i\cdot}}; \quad j = 1, 2, \ldots, p,$$

Así obtenemos las denominadas **distribuciones condicionadas**.

Existen p distribuciones condicionadas según el carácter X para una sola modalidad de Y: cada una corresponde a una modalidad y_j del carácter Y; si bien éstas no son las únicas, ya que puede resultar igualmente de interés condicionar a más de una modalidad simultáneamente (entre dos niveles de renta, dos valores de estatura, etc.) Existen k distribuciones condicionadas según elcarácter Y para una sola modalidad de X: cada una corresponde a una modalidad x_i de X; si bien éstas tampoco son las únicas, por la misma razón de antes.

Distribuciones condicionadas

X	n_{ij}	f_i^j
x_1	n_{1j}	f_1^j
x_2	n_{2j}	f_2^j
$\vdots$	$\vdots$	$\vdots$
x_i	n_{ij}	f_i^j
$\vdots$	$\vdots$	$\vdots$
x_k	n_{kj}	f_k^j
	$n_{\cdot j}$	1

Y	n_{ij}	f_j^i
y_1	n_{i1}	f_1^i
y_2	n_{i2}	f_2^i
$\vdots$	$\vdots$	$\vdots$
y_j	n_{ij}	f_j^i
$\vdots$	$\vdots$	$\vdots$
y_p	n_{ip}	f_p^i
	$n_{i\cdot}$	1

En el primer caso tenemos la distribución de X condicionada a la modalidad y_j de Y (lo que denotaremos $X/Y = y_j$), y en el segundo la distribución de Y condicionada a la modalidad x_i de X (lo que denotaremos $Y/X = x_i$).

De las definiciones anteriores se tienen las siguientes relaciones:

$$f_{ij} = \frac{n_{ij}}{n} = \frac{n_{i\cdot}}{n}\,\frac{n_{ij}}{n_{i\cdot}} = \frac{n_{\cdot j}}{n}\,\frac{n_{ij}}{n_{\cdot j}},$$

o lo que es lo mismo:

$$f_{ij} = f_{i\cdot}\, f_j^i = f_{\cdot j}\, f_i^j,$$

o bien

$$f_{ij} = f_{i\cdot}\, f_{j/i} = f_{\cdot j}\, f_{i/j}.$$

Dependencia e independencia estadística

Dos caracteres X e Y se dirán **estadísticamente dependientes** cuando la variación en uno de ellos influya en la distribución del otro.

Se dice que el carácter X es **independiente estadísticamente**, (o, simplemente, independiente) del carácter Y si las distribuciones de X condicionadas a cada valor y_j de Y $(X/Y = y_j)$ son todas idénticas para cualquier valor de $j = 1, 2, \ldots, p$; es decir,

$$f_i^j \equiv f_{i/j} \text{ no depende de } j.$$

En este caso, las distribuciones de $X/Y = y_j$ son idénticas, e idénticas además a la distribución marginal de X. En efecto: la independencia de X respecto de Y se escribe

$$\frac{n_{i1}}{n_{\cdot 1}} = \frac{n_{i2}}{n_{\cdot 2}} = \ldots = \frac{n_{ij}}{n_{\cdot j}} = \ldots = \frac{n_{ip}}{n_{\cdot p}} \quad \forall i = 1, 2, \ldots, k.$$

Si estas fracciones son iguales, lo serán también a la fracción obtenida sumando numeradores y denominadores; así:

$$f_{i/j} = f_i^j = \frac{n_{ij}}{n_{\cdot j}} = \frac{n_{i1} + n_{i2} + \ldots + n_{ip}}{n_{\cdot 1} + n_{\cdot 2} + \ldots + n_{\cdot p}} = \frac{n_{i\cdot}}{n} = f_{i\cdot}; \quad \forall i = 1, 2, \ldots, k.$$

Por lo tanto,

$$f_i^j = f_{i/j} = f_{i\cdot} \quad \forall i = 1, 2, \ldots, k$$

denotará que el carácter X es independiente del carácter Y.

Dependencia e independencia estadística

Se dice que el carácter Y es **independiente estadísticamente**, (o, simplemente, independiente) del carácter X si las distribuciones de Y condicionadas a cada valor x_i de X $(Y/X = x_i)$ son todas idénticas para cualquier valor de $i = 1, 2, \ldots, k$; es decir,

$$f_j^i \equiv f_{j/i} \text{ no depende de } i.$$

En este caso, las distribuciones de $Y/X = x_i$ son idénticas, e idénticas además a la distribución marginal de Y. En efecto: la independencia de Y respecto de X se escribe

$$\frac{n_{1j}}{n_{1\cdot}} = \frac{n_{2j}}{n_{2\cdot}} = \ldots = \frac{n_{ij}}{n_{i\cdot}} = \ldots = \frac{n_{kj}}{n_{k\cdot}} \quad \forall j = 1, 2, \ldots, p.$$

Si estas fracciones son iguales, lo serán también a la fracción obtenida sumando numeradores y denominadores; así:

$$f_{j/i} = f_j^i = \frac{n_{ij}}{n_{i\cdot}} = \frac{n_{1j} + n_{2j} + \ldots + n_{kj}}{n_{1\cdot} + n_{2\cdot} + \ldots + n_{k\cdot}} = \frac{n_{\cdot j}}{n} = f_{\cdot j}; \quad \forall j = 1, 2, \ldots, p.$$

Por lo tanto,

$$f_j^i = f_{j/i} = f_{\cdot j} \quad \forall j = 1, 2, \ldots, p$$

denotará que el carácter Y es independiente del carácter X.

Proposición:
Si el carácter X es independiente del carácter Y, entonces, Y es independiente de X (la independencia es una propiedad recíproca)

Hablamos entonces de **independencia** de dos caracteres estadísticos, sin referirnos al sentido de la misma.

Dependencia funcional

Se dice que el carácter X **depende funcionalmente** del carácter Y si a cada modalidad y_j de Y corresponde una única modalidad posible de X con frecuencia no nula; esto es, para cualquier $j = 1, 2, \ldots, p$, la frecuencia absoluta n_{ij} es nula, excepto para un valor $i = \varphi(j)$ donde n_{ij} es igual a $n_{\cdot j}$. Así, en cada columna de la tabla de doble entrada, un término y sólo un término es diferente de cero, aunque puede haber varios términos no nulos en una misma fila. Por ejemplo, en la siguiente tabla

X Y	y_1	y_2	y_3	y_4	y_5	
x_1	3	0	6	0	0	9
x_2	0	4	0	0	2	6
x_3	0	0	0	5	0	5
	3	4	6	5	5	20

el carácter X depende funcionalmente del carácter Y, ya que a cada modalidad de Y le corresponde una sola modalidad de X con frecuencia no nula (en cada columna hay sólo una frecuencia distinta de cero); sin embargo, Y no depende funcionalmente de X, ya que en cada fila no existe una única frecuencia no nula.

La dependencia funcional, así, no es recíproca. Cuando la correspondencia $i = \varphi(j)$ es biunívoca; es decir, cuando en cada fila y en cada columna de la tabla de doble entrada existe una única casilla con frecuencia no nula, se dice que la dependencia funcional es recíproca. Evidentemente, para que esto ocurra es necesario que, en particular, la tabla sea cuadrada; esto es, ambos caracteres presenten el mismo número de modalidades ($k = p$). Por ejemplo, en la siguiente tabla

X Y	y_1	y_2	y_3	
x_1	3	0	0	3
x_2	0	0	2	2
x_3	0	4	0	4
	3	4	2	9

la dependencia funcional es recíproca.

Momentos bidimensionales

Definición:

Dada una variable estadística bidimensional (X, Y), con distribución conjunta $\{(x_i, y_i); f_{ij}\}_{i=1,\dots,k;\ j=1,\dots,p}$, se define el *momento conjunto respecto al origen de órdenes r y s*, con r y s enteros no negativos, como

$$m_{rs} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij} x_i^r\ y_j^s.$$

Definición:

Dada una variable estadística bidimensional (X, Y), con distribución conjunta $\{(x_i, y_i); f_{ij}\}_{i=1,\dots,k;\ j=1,\dots,p}$, se define el *momento conjunto central de órdenes r y s*, con r y s enteros no negativos, como

$$\mu_{rs} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij} (x_i - \bar{x})^r\ (y_j - \bar{y})^s.$$

Los momentos centrales más utilizados son las varianzas marginales, dadas por los momentos conjuntos centrales $\mu_{20} = \sigma_X^2$ y $\mu_{02} = \sigma_Y^2$, y el momento μ_{11}, denominado **covarianza** de las variables X e Y, y que también se suele representar por $\sigma_{XY} = \mathrm{Cov}(X, Y)$.

Como resultados importantes relativos a estos momentos tenemos:

- $\sigma_{XY} = \mu_{11} = m_{11} - m_{10} m_{01}$.
- Si X e Y son independientes estadísticamente, entonces $m_{rs} = m_{r0} m_{0s}$ y $\mu_{rs} = \mu_{r0} \mu_{0s}$.
- Si X e Y son independientes estadísticamente, entonces la covarianza entre ambas variables es cero.

Regresión

Uno de los objetivos de toda ciencia es el de encontrar relaciones entre los sucesos que se le presentan en su campo de investigación. El científico intenta traducir estas relaciones en estructuras manejables, para lo cual hace uso de un lenguaje ya conocido, fundamentalmente el *estadístico-matemático*, a través del establecimiento de relaciones funcionales en donde un número finito de magnitudes (variables o atributos) $X_1, \ldots, X_n$ se suponen que están relacionadas con una variable Y, a través de la expresión

$$Y = f(X_1, \ldots, X_n).$$

El tratamiento que necesita esta modelización no es estrictamente matemático, sino que requiere un análisis fundamentalmente estadístico.
Desde esta perspectiva, son dos los enfoques con que simultáneamente se puede abordar el problema. A saber:

- La determinación de la estructura de dependencia que mejor expresa el tipo de relación de la variable Y con las demás o **regresión**.
- El estudio del grado de dependencia existente entre las variables, que será el propósito de la **correlación**.

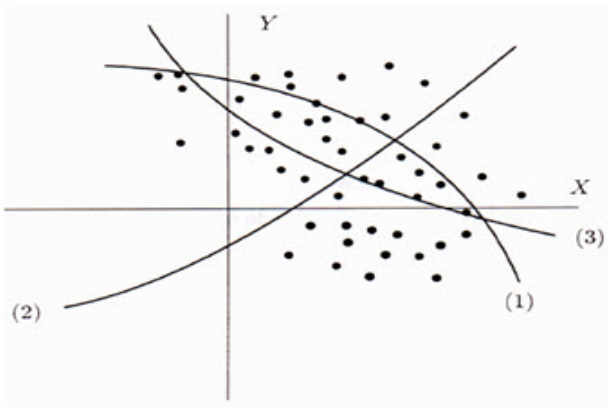
Existen, pues, tres motivos fundamentales por los que una variable que vamos a denominar **dependiente** o **endógena** está influida por otra que actúa como **independiente, explicativa** o **exógena**. A saber:

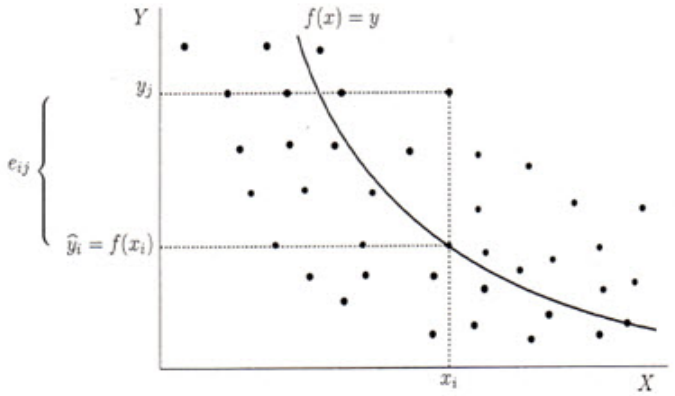
- la casualidad o el azar ha hecho que ambas variables estén relacionadas estadísticamente (por ejemplo, podría existir relación entre el número de accidentes de automóviles y la producción de queso manchego en una determinada región);
- una tercera variable está determinando a las que estamos estudiando (por ejemplo, el consumo de caviar y la compra de yates de recreo están determinadas por la renta disponible de las personas);
- y, por último, puede existir una *relación causa-efecto* como en el ejemplo de que los niveles de consumo estén determinados fundamentalmente por la renta disponible.

En nuestro estudio, sólo nos ocuparemos de estudiar las relaciones de causa-efecto, que son las que tienen una base teórica.

Regresión
Ajuste de funciones y datos por el método de mínimos cuadrados

Si dos variables presentan una dependencia estadística; es decir, no funcional, no es posible encontrar una ecuación tal que los valores que puedan presentar dichas variables la satisfagan. Gráficamente equivale al hecho de que no es posible encontrar una función tal que su gráfica pase por todos los puntos correspondientes al diagrama de dispersión asociado a las variables observadas.





Ante la imposibilidad de encontrar una gráfica que pase por todos los puntos de la nube, hay que aceptar el razonamiento lógico de que la función cuya gráfica más se aproxime a los datos observados expresará mejor la relación entre los mismos. Y resulta claro en esta figura que la función con gráfica (1) expresa mejor que la función con gráfica (2) la relación entre X e Y. Pero si tuviésemos que decidir entre (1) y (3), la elección no sería tan clara. Esto indica que el método gráfico, aunque es de gran ayuda, no es suficiente para encontrar la función que mejor expresa la dependencia estadística entre dos variables. Tendremos que recurrir a métodos matemáticos que no dependan de opiniones subjetivas.

Hacer regresión consiste, pues, en ajustar lo mejor posible una función a una serie de valores observados, encontrando una curva que, aunque no pase por todos los puntos de la nube, al menos esté lo más próxima posible a ellos. El siguiente razonamiento nos conducirá al conocido **ajuste por mínimos cuadrados**.

Regresión
Ajuste de funciones y datos por el método de mínimos cuadrados

El método de mínimos cuadrados consiste en encontrar dicha función f tal que minimice la media de los cuadrados de los residuos

$$\psi(a_0, a_1, \ldots, a_n) = \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} \ (y_j - f(x_i, a_0, a_1, \ldots, a_n))^2.$$

Esta función, ψ, se denomina **error cuadrático medio de la función** f y la denotaremos $ECM(f(x_i, a_0, a_1, \ldots, a_n)$. Dado que los parámetros $(a_0, a_1, \ldots, a_n)$ sólo están sometidos a sumas, productos y cuadrados dentro de la función ψ, ésta es derivable respecto a cada a_i, $i = 0, \ldots, n$. Además puede asegurarse que el punto $(\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_n)$ donde se anulan las derivadas parciales primeras respecto de cada a_i corresponde a un mínimo de la función ψ.

El cálculo de los parámetros de la función de ajuste óptima según el método de los mínimos cuadrados consiste en resolver el sistema siguiente, denominado **sistema de ecuaciones normales**:

$$\frac{\partial \psi}{\partial a_r} = 0 \Rightarrow \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} \ (y_j - f(x_i, a_0, a_1, \ldots, a_n)) \ \frac{\partial f}{\partial a_r} = 0, \quad r = 0, 1, \ldots, n.$$

Todo lo expuesto aquí es igualmente válido para el ajuste por el método de mínimos cuadrados de una función de X sobre Y, donde ahora la predicción para cada valor y_j es

$$\hat{x}_j = g(y_j, c_0, c_1, \ldots, c_m),$$

y el correspondiente sistema de ecuaciones normales es

$$\frac{\partial \psi}{\partial c_s} = 0 \Rightarrow \sum_{i=1}^{k} \sum_{j=1}^{p} f_{ij} \ (x_i - f(y_j, c_0, c_1, \ldots, c_m)) \ \frac{\partial g}{\partial c_s} = 0, \quad s = 0, 1, \ldots, m.$$

Regresión

Aproximación polinomial mínimo cuadrática. Sistemas ortonormales

Una de las funciones de regresión más utilizadas para expresar el comportamiento de una variable en función de la otra es un polinomio de grado n. Comenzaremos estudiando el caso de una recta ($n = 1$).

Supongamos que queremos ajustar por el método de mínimos cuadrados una recta que exprese Y en función de X. La función sería $y = f(x; a, b) = a + bx$, con lo que habrá que calcular el mínimo en a y b de la función

$$\psi(a, b) = ECM(a, b) = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}[y_j - (a + bx_i)]^2.$$

La función $\psi(a, b)$ es una función continua y derivable respecto a las variables a y b, por ser suma de funciones polinómicas de estas variables. Obtenemos, entonces, el *sistema de ecuaciones normales*:

$$\frac{\partial \psi}{\partial a} = 0 \Rightarrow \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\ [y_j - (a + bx_i)] = 0$$

$$\frac{\partial \psi}{\partial b} = 0 \Rightarrow \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\ [y_j - (a + bx_i)]x_i = 0$$

Si tenemos en cuenta los momentos conjuntos definidos anteriormente, el sistema adopta ahora la forma

$$\begin{aligned} m_{01} &= a + bm_{10} \\ m_{11} &= am_{10} + bm_{20} \end{aligned}$$

cuya resolución proporciona los coeficientes buscados:

$$\hat{a} = m_{01} - \frac{m_{11} - m_{10}m_{01}}{m_{02} - m_{01}^2} m_{10} = \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2}\bar{x}$$

$$\hat{b} = \frac{m_{11} - m_{10}m_{01}}{m_{02} - m_{01}^2} = \frac{\sigma_{xy}}{\sigma_x^2}$$

Regresión

Aproximación polinomial mínimo cuadrática. Sistemas ortonormales

La *recta de regresión de Y sobre X* tiene por expresión

$$y = \frac{\sigma_{xy}}{\sigma_x^2}x + \bar{y} - \frac{\sigma_{xy}}{\sigma_x^2}\bar{x}, \qquad o\ bien \qquad y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x}).$$

Al número $\frac{\sigma_{xy}}{\sigma_x^2}$ se le denomina *coeficiente de regresión lineal de Y sobre X*.

De forma análoga, la recta de regresión mínimo cuadrática de X sobre Y es la recta $x = h(y; c, d) = c + dy$ que hace mínima la función

$$\phi(c, d) = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(x_i - \hat{x}_j)^2,$$

donde $\hat{x}_j = c + dy_j$. Siguiendo un procedimiento análogo al anterior se llega a la recta de regresión

$$x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y}).$$

Al número $\frac{\sigma_{xy}}{\sigma_y^2}$ se le denomina *coeficiente de regresión lineal de X sobre Y*.

Los coeficientes de regresión son las pendientes de las rectas de regresión. Los signos de dichos coeficientes son los mismos para las dos rectas ajustadas a la nube de puntos, y, además igual al de la covarianza. Una covarianza positiva dará dos coeficientes de regresión positivos y sus correspondientes rectas de regresión crecientes. Si la covarianza es negativa, las dos rectas de regresión serán decrecientes. En el caso de que la covarianza sea nula, la rectas de regresión serán paralelas a los ejes, y perpendiculares entre sí. Cuando exista dependencia funcional lineal (todos los puntos de la nube están sobre una recta), las dos rectas de regresión coinciden con la recta de dependencia.

Propiedades de las rectas de regresión:

1. Pasan las dos por el centro de gravedad de la distribución $(\bar{x}, \bar{y})$.
2. La media de los valores ajustados coincide con la de los valores observados de la variable.
3. La media de los residuos vale cero.
4. La media de los productos de los residuos por los valores de la variable explicativa vale cero.
5. La media de los productos de los residuos por los valores ajustados vale cero.

Regresión

Aproximación polinomial mínimo cuadrática. Sistemas ortonormales

Si se quiere ajustar un polinomio de grado superior a dos, de expresión general

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n,$$

el método de mínimos cuadrados, nos conduce al sistema de ecuaciones normales

$$\begin{aligned} m_{01} &= a_0 + a_1 m_{10} + a_2 m_{20} + \ldots + a_n m_{n0} \\ m_{11} &= a_0 m_{10} + a_1 m_{20} + a_2 m_{30} + \ldots + a_n m_{n+1,0} \\ m_{21} &= a_0 m_{20} + a_1 m_{30} + a_2 m_{40} + \ldots + a_n m_{n+2,0} \\ &\vdots \quad \vdots \\ m_{n1} &= a_0 m_{n0} + a_1 m_{n+1,0} + a_2 m_{n+2,0} + \ldots + a_n m_{n+n,0} \end{aligned}$$

cuya resolución nos proporciona los valores de los parámetros buscados $a_0, a_1, \ldots, a_n$.

Para el ajuste a la nube de puntos de otro tipo de función, intentaremos pasar a un ajuste polinómico (lineal, a ser posible), mediante transformaciones adecuadas sobre el conjunto de los datos.
Así, para ajustar una hipérbola equilátera, de expresión

$$y = a + b\frac{1}{x},$$

efectuaremos la transformación $z = 1/x$, con lo que el ajuste por mínimos cuadrados de la hipérbola se reduce a ajustar una recta de la forma $y = a + bz$ sobre las variables (Z, Y) como se ha descrito anteriormente.

Por su parte, para el modelo potencial (de difícil resolución analítica directa por mínimos cuadrados), linealizaremos igualmente antes de aplicar el criterio, tomando logaritmos, con lo que de la función $y = ax^b$ pasaremos a $\log y = \log a + b \log x$. Llamando a las nuevas variables $\log Y = V, \log X = U$ y a la constante $\log a = A$, queda $v = A + b\ U$, de forma que hemos transformado el problema inicial en el ajuste de una recta. De forma similar se realizaría el ajuste de una función exponencial.

Regresión

Curvas generales de regresión. Regresión de tipo I

Uno de los objetivos más importantes del análisis de la regresión es predecir el valor de una variable dependiente Y dado el valor x_i de una variable independiente asociada X; es decir, predecir el comportamiento de la variable condicionada $Y/X = x_i$.
Son bien conocidas las cualidades de la media aritmética como valor representativo del comportamiento de una variable. Teniendo en cuenta este hecho y el comentario anterior, se define la *curva de regresión de tipo I de* Y/X como la curva que pasa por los puntos $(x_i, \bar{y}_i)$; $i = 1, \ldots, k$.. Análogamente se define la curva de regresión de tipo I de X/Y como la curva que pasa por los puntos $(\bar{x}_j, y_j); j = 1, \ldots, p$.

Las curvas así definidas tienen la importante propiedad de ser entre todas las funciones las que mejor se ajustan a los datos observados (en cada caso), según el criterio de mínimos cuadrados.

A pesar de esta importante propiedad, esta curva de regresión no es de gran utilidad práctica, pues el hecho de conocerla solamente en puntos aislados la hace inútil para la predicción en la mayoría de los casos.

Correlación

Tan importante es conocer la forma en que se relacionan las variables como su grado de asociación. De esto último se ocupa la *correlación*.
El grado de asociación entre las variables nos indicará en qué medida la expresión encontrada mediante la regresión explica una variable en función de la otra. El estudio de la correlación también equivale al estudio de la bondad del ajuste de una curva a una nube de puntos.

Varianza residual. Coeficiente de determinación

El método de mínimos cuadrados toma como medida del error que se comete, cuando ajustamos una curva, la media de los residuos al cuadrado

$$\sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij} e_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(y_j - \hat{y}_i)^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}[y_j - f(x_i)]^2.$$

Esta cantidad se usa como medida de la bondad del ajuste. En funciones lineales en los parámetros (recta, parábola, hipérbola equilátera, etc.), la media de los residuos es cero, por lo que la expresión anterior es precisamente la varianza de los residuos, o **varianza residual**. En funciones no lineales en los parámetros, la media de los residuos no es nula; a pesar de ello, se mantiene la denominación de varianza residual para la media de los cuadrados de los residuos.

Tomaremos como medida del grado de ajuste la proporción de la varianza total de la variable Y explicada por la regresión; es decir, el cociente o razón entre la varianza explicada y la total. Esto será posible siempre que se pueda llevar a cabo una descomposición de la varianza en suma de varianza residual y varianza explicada por la regresión

$$\sigma_y^2 = \sigma_{ey}^2 + \sigma_{ry}^2, \qquad \textit{donde} \qquad \sigma_{ey}^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}(\hat{y}_j - \bar{y})^2,$$

denominada **varianza explicada por la regresión** nos indica en qué medida queda explicada la variable dependiente mediante el modelo estimado y la varianza residual σ_{ry}^2 se puede interpretar como una medida de lo que queda sin explicar después de haber efectuado la regresión.

Correlación

A esta razón se denomina *coeficiente de determinación o, en general, razón de correlación*. Dicho coeficiente podrá escribirse de las dos formas siguientes:

$$\eta^2_{Y/X} = \frac{\sigma^2_{ey}}{\sigma^2_y} = 1 - \frac{\sigma^2_{ry}}{\sigma^2_y}.$$

De su misma expresión se deduce que $0 \leq \eta^2_{Y/X} \leq 1$.

- $\eta^2_{Y/X} = 0 \Leftrightarrow \dfrac{\sigma^2_{ry}}{\sigma^2_y} = 1 \Leftrightarrow \sigma^2_{ry} = \sigma^2_y,\ \sigma^2_{ey} = 0$; es decir, el modelo no explica nada de Y a partir de X. El ajuste es el peor posible que se puede hacer por mínimos cuadrados.
- $\eta^2_{Y/X} = 1 \Leftrightarrow \dfrac{\sigma^2_{ry}}{\sigma^2_y} = 0$; es decir, todos los residuos son nulos. El ajuste es perfecto.
- Para valores intermedios entre 0 y 1, según estén más próximos a un extremo u otro nos indicarán un peor o mejor ajuste, respectivamente.

Correlación

Correlación en el caso lineal

Es fácil comprobar que la razón de correlación en el caso lineal vale

$$\eta^2_{Y/X} = \eta^2_{X/Y} = r^2 = \frac{\sigma^2_{xy}}{\sigma^2_x\ \sigma^2_y},$$

y que coincide con el producto de los coeficientes de regresión o pendientes de las dos rectas de regresión.

La raíz cuadrada del coeficiente de determinación lineal anterior (con el signo de la covarianza) recibe el nombre de **coeficiente de correlación lineal**

$$r = \pm\sqrt{r^2} = \frac{\sigma_{xy}}{\sigma_x\ \sigma_y}.$$

Dicho coeficiente se utiliza para determinar el grado de dependencia lineal de la variable endógena ante los valores de la exógena. Esta dependencia puede ser directa o positiva, ó indirecta o negativa, según sea el signo de la covarianza.

Si la covarianza es positiva, la correlación también lo es y su coeficiente tomará valores entre 0 y 1. Si $r = 1$ existirá una dependencia lineal exacta o funcional, ya que los valores estimados mediante la recta coincidirán con los observados. Si, por el contrario, $r = 0$, no existe ninguna dependencia o asociación entre las variables de tipo lineal, aunque sí puede haberla de otra naturaleza, convirtiéndose las rectas de regresión en dos paralelas a los ejes de coordenadas, ya que los valores estimados de las rectas coinciden con las medias aritméticas marginales.

Si la covarianza es negativa, la correlación también lo es y su coeficiente tomará valores entre -1 y 0. Si $r = -1$, la correlación es perfecta, existiendo una dependencia funcional lineal, pero negativa. Las rectas de regresión coincidirían en una sola, que sería decreciente, al tener una pendiente negativa.

Concluyendo, diremos que el campo de variación del coeficiente de correlación lineal es $-1 \leq r \leq 1$. Cuando varía de -1 a 0 estamos en una correlación negativa y la dependencia será mayor cuanto más se aproxime a -1. Si la variación está entre 0 y 1, la correlación es positiva y el grado de asociación o dependencia será mayor cuanto más se aproxime a 1.

Correlación

Predicción

Uno de los objetivos que persigue la regresión y correlación es hacer predicciones de la variable dependiente o endógena en función de los que toma la independiente o exógena. Las predicciones se efectúan utilizando la función estimada por mínimos cuadrados f. Con ella se obtienen los valores *teóricos* que ajustan a los observados. La predicción será tanto más fiable cuanto mayores sean los coeficientes de determinación correspondientes, o razones de correlación, ya que menor será la varianza de los residuos, que es la que nos indica la cuantía de la separación entre lo observado y lo estimado.

Hay que tener presente que la fiabilidad de las predicciones disminuye a medida que los valores de la variable exógena se alejan de su recorrido, pues se corre el riesgo adicional de que el modelo ajustado no sea válido para dichos valores en la medida dada por η^2.