CAPÍTULO 11
ESTADÍSTICA DESCRIPTIVA BIDIMENSIONAL

11.1 DISTRIBUCIONES MARGINALES Y CONDICIONALES

Cuando sobre cada individuo de una población se observan **dos** características aleatorias de naturaleza cuantitativa se tiene una variable aleatoria bidimensional cuantitativa.

Ejemplo 1: en la población constituida por los estudiantes universitarios españoles se observa la ESTATURA (cms) y el PESO (kgs) de cada estudiante. Una muestra de esta variable bidimensional puede estar constituida por los 130 pares de valores constatados en los 130 alumnos que respondieron a la encuesta (archivo *curs8990.sf3)*

Ejemplo 2: para el control del consumo de energía en calefacción en una factoría durante los meses de invierno se anota diariamente el CONSUMO (en una determinada unidad de medida, como m^3 de gas o termias) y la temperatura diaria (1C a las 12). Una muestra de esta variable bidimensional puede estar constituida por los 57 pares de valores constatados en 57 días laborables del invierno de 1985 (archivo *gas.sf3*[1]

En el Capítulo 7 se ha expuesto cómo podía describirse, mediante una Tabla de Contingencia, la relación entre las dos componentes de una variable bidimensional en el caso de que ambas fueran de tipo **cualitativa**.

Cuando las dos variables[2] sean de tipo **cuantitativo**, y especialmente cuando se trate de variables **continuas** (como sucede en los dos ejemplos anteriores) es posible utilizar técnicas más adecuadas para describir y analizar la relación existente entre ambas.

Por supuesto es posible, en primer lugar, construir una **tabla de frecuencias cruzada** entre las dos variables, aunque será necesario previamente agruparlas en intervalos.

La siguiente tabla (construida mediante Statgraphics previa una recodificación de las variables) refleja las frecuencias observadas para cada combinación de tramos de ESTATURA y PESO.

[1] Todos los archivos de datos que se mencionan en este texto pueden bajarse libremente de la URL http://personales.upv.es/rromero/descargas. Los alumnos de la Universidad Politécnica de Valencia pueden también bajárselos del Poliformat de la asignatura.

[2] Utilizaremos frecuentemente la expresión "dos variables aleatorias" por ser más cómoda que "dos componentes de la variable aleatoria bidimensional", aunque ésta última es más correcta

ESTATURA PESO	145 155	155 165	165 175	175 185	185 195	Row Total
40 55	9 75.0	17 44.7	0 .0	0 .0	0 .0	26 20.0
55 70	3 25.0	18 47.4	31 53.4	5 29.4	0 .0	57 43.8
70 85	0 .0	3 7.9	24 41.4	12 70.6	3 60.0	42 32.3
85 99	0 .0	0 .0	3 5.2	0 .0	2 40.0	5 3.8
Column Total	12 9.2	38 29.2	58 44.6	17 13.1	5 3.8	130 100

En el margen derecho se recogen las frecuencias (absolutas y relativas, estas últimas expresadas como porcentaje) de los 4 tramos considerados para PESO. Estas frecuencias, que están obtenidas sumando para todos los valores posibles de ESTATURA se denominan **marginales**. A la pauta de variabilidad que sigue en la población la variable PESO considerada aisladamente, o sea prescindiendo de los posibles valores que tome la ESTATURA, se le denomina **distribución marginal** del PESO.

De forma análoga en el margen inferior de la tabla se reflejan las frecuencias (absolutas y relativas) observadas en la muestra para la **distribución marginal** de la ESTATURA.

Dentro de cada columna se recogen las frecuencias observadas para los diferentes tramos de PESO en los individuos cuya ESTATURA se halla en el tramo considerado. Las frecuencias relativas están calculadas respecto a la frecuencia total de la columna considerada y se denominan **frecuencias relativas condicionales**. Así de los individuos cuya ESTATURA está en el tramo 145-155, el 75% pesan entre 40 y 55 kgs y el 25% entre 55 y 70 kgs, mientras que de los que miden entre 175 y 185 cms el 29.4% pesan entre 55 y 70 kgs y el 70.6% pesan entre 70 y 85 kgs.

La pauta de variabilidad que sigue en la población la variable PESO, si nos limitamos a considerar sólo aquellos individuos cuya ESTATURA pertenece a un determinado tramo, se denomina **distribución condicional** del PESO, y en general será diferente según el tramo considerado para la ESTATURA. En la tabla siguiente se recogen los valores de la media, desviación típica, mínimo y máximo, para las 4 distribuciones condicionales del PESO asociadas a distintos tramos de la variable ESTATURA.

Tabla: Parámetros de las distribuciones condicionales de PESO en función de la Estatura

ESTATURA	Número de casos
145 155	12
155 165	38
165 175	53
175 185	17
185 195	5

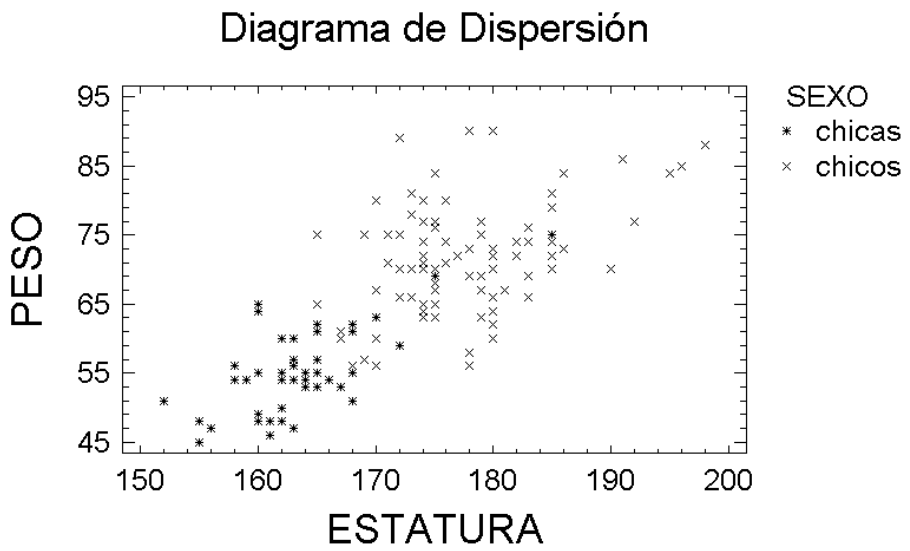
Autoevaluación: En la tabla anterior las medias de las distribuciones condicionales aumentan al aumentar los valores considerados para la variable condicionante ESTATURA. Justificar la lógica de este resultado

Autoevaluación: La desviación típica en la muestra de la distribución marginal del PESO es 10.7, sensiblemente superior a las desviaciones típicas constatadas para las distribuciones condicionales. Justificar lógicamente este resultado.

11.2 DIAGRAMAS DE DISPERSIÓN

Una forma sencilla de describir gráficamente las relaciones constatadas entre dos variables consiste en representar cada observación por un punto en un plano, cuya abscisa sea el valor de la primera variable y cuya ordenada sea el de la segunda. A este tipo de gráfico se le denomina **diagrama de dispersión**.

La siguiente figura refleja el diagrama de dispersión de la variable PESO frente a ESTATURA. Para mayor información los puntos correspondientes a chicas se han codificado como un $*$ y los correspondientes a chicos con una $\times$.



El diagrama pone claramente de manifiesto una relación positiva entre las dos variables estudiadas, que se refleja en una nube de puntos en forma de elipse cuyo eje principal tiene un sentido creciente, como consecuencia del hecho de que, en términos generales, los individuos más altos pesan más que los más bajos. El diagrama también pone de manifiesto que las chicas tienen en general valores menores de ambas variables que los chicos, pero que la relación entre PESO y ESTATURA es bastante similar en ambos sexos.

Autoevaluación: Para estudiar un ejemplo en el que el Diagrama de Dispersión pone claramente en evidencia una relación negativa entre dos variables obtener el diagrama para las variables TEMPER y CONSUMO del fichero gas.sf3.

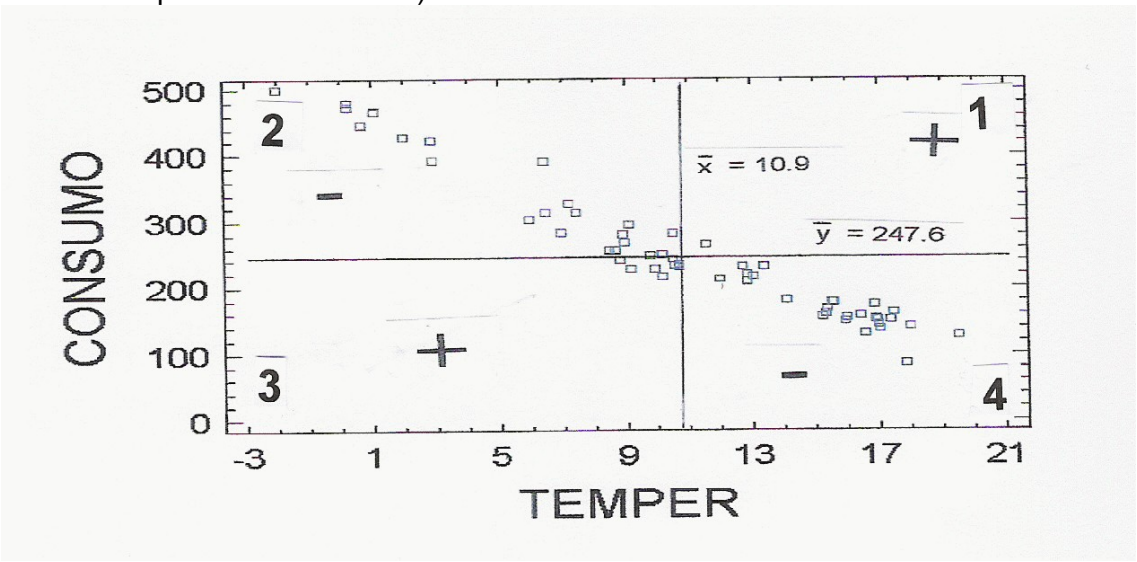
En general cuanto más estrechamente se agrupen los puntos del diagrama de dispersión alrededor de una recta más fuerte es el grado de relación lineal existente entre las dos variables consideradas.

Autoevaluación: ¿Sigue creciendo la gente a partir de los 19 años? Para ver si hay alguna evidencia la respecto en los datos de la encuesta, construir el Diagrama de Dispersión de ESTATURA frente a EDAD. ¿Sugieren los datos algún tipo de relación?

11.3 COVARIANZA. COEFICIENTE DE CORRELACIÓN

Con el fin de cuantificar en un índice numérico el grado de relación lineal existente entre dos variables se utilizan en Estadística dos parámetros: la **covarianza** y el **coeficiente de correlación**.

Con el fin de dar una idea intuitiva del concepto de covarianza vamos a razonar sobre el siguiente diagrama de dispersión, correspondiente a las variables TEMPERATURA diaria y CONSUMO de energía, en el que hemos trazado una línea horizontal a la altura del valor medio $\bar{y}$ de la segunda variable (247.6 es el consumo diario medio) y una línea vertical situada sobre el valor medio $\bar{x}$ de la primera variable (10.9 1C es la temperatura media en el período considerado)



En este caso, en el que existe claramente una fuerte relación negativa, la mayor parte de los puntos caen en los cuadrantes 2 y 4. Por el contrario cuando la relación existente sea positiva la mayoría de los puntos caerán en los cuadrantes 1 y 3.

Si consideramos el signo que para cada punto (x_i,y_i) del diagrama tiene el producto $(x_i-\bar{x})(y_i-\bar{y})$, vemos que éste resulta positivo en los cuadrantes 1 y 3 y negativo en los cuadrantes 2 y 4. Por lo tanto el producto anterior será **en promedio** positivo si existe una relación creciente, (o sea positiva), entre las dos variables (es decir si la Y tiende a crecer cuando lo hace la X) y negativo si la relación existente es decreciente (o sea negativa)

Por definición la **covarianza** entre dos variables en una muestra no es más que el promedio de los productos de las desviaciones de ambas variables respecto a sus medias respectivas. (Por consideraciones que no son del caso, y de forma similar a como se procedió al definir la varianza, el promedio se calcula dividiendo por N-1 en vez de por N)

$$\mathrm{cov}(X,Y)=\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{N-1}$$

Nota: a nivel poblacional, la covarianza entre dos variables aleatorias **X** e **Y** se define como la esperanza matemática del producto de las desviaciones de ambas variables respecto a sus medias respectivas $\mathrm{cov}(\mathbf{X},\mathbf{Y}) = E(\mathbf{X}-m_x)(\mathbf{y}-m_y)$

La covarianza presenta el inconveniente de que depende de las dimensiones en que se expresan las variables. Así la covarianza entre ESTATURA y PESO será 100 veces mayor si la primera variable se mide en centímetros que si se mide en metros. Para obviar este problema se utiliza universalmente en Estadística, como medida del grado de relación lineal existente entre dos variables, el **coeficiente de correlación lineal** que no es más que la covarianza dividida por el producto de las desviaciones típicas de las dos variables

$$r(X,Y) = \frac{\text{cov}(X,Y)}{s_x s_y}$$

Nota: a nivel poblacional, el coeficiente de correlación lineal entre dos variables aleatorias **X** e **Y** se define como cov(**X,Y**)/($\sigma_x\sigma_y$)

Se demuestra fácilmente que el coeficiente de correlación entre dos variables se mantiene inalterable si cualquiera de ellas sufre una transformación lineal. Así, por ejemplo, el coeficiente de correlación entre CONSUMO y TEMPERATURA no se modifica por el hecho de que esta última variable se exprese en grados Fahrenheit en vez de en grados centígrados.

El coeficiente de correlación **r**(X,Y) tiene una serie de propiedades que lo hacen especialmente adecuado para medir el grado de relación (lineal) entre dos variables:

- **r**(X,Y) está siempre comprendido entre -1 y +1

- Los valores extremos, -1 y +1, sólo se alcanzan si existe una relación lineal exacta entre la Y y la X, o sea, si los puntos del diagrama de dispersión están exactamente alineados en una recta. (El valor +1 se tiene si la recta es creciente y el -1 si es decreciente)

- Cuando dos variables aleatorias **X** e **Y** son independientes r(**X,Y**) es igual a cero. (En la práctica en una muestra de dos variables independientes **r**(X,Y) será "cercano" a cero, pues debido al azar del muestreo **r** fluctuará algo alrededor de su verdadero valor poblacional)

- Por tanto, contra más estrecho es el grado de relación lineal existente entre dos variables más cercano a 1 es el valor de **r** (o a -1 si la relación es decreciente). Por el contrario un valor de **r** nulo o cercano a cero indicará una relación lineal inexistente o muy débil.

- El cuadrado del coeficiente de correlación mide la proporción (o porcentaje si se multiplica por 100) de la varianza de Y que está asociada linealmente a la variabilidad de X

Autoevaluación: Calcular los coeficientes de correlación entre ESTATURA y PESO, entre EDAD y ESTATURA y entre TEMPER y CONSUMO y contrastarlos con el aspecto de los diagramas de dispersión correspondientes. ¿Hasta qué punto las diferencias de peso entre los alumnos están asociadas a las diferencias de estatura entre ellos?

Es importante resaltar que tanto la covarianza como el coeficiente de correlación miden sólo el grado de relación lineal existente entre dos variables. Dos variables pueden tener una relación estrecha y sin embargo resultar **r** cercano a cero por ser dicha relación no lineal.

Autoevaluación: Introducir en Statgraphics dos variables: una X de valores -3,-2,-1, 0,1,2,3 y otra Y de valores 9,4,1,0,1,4,9. Dibujar el diagrama de dispersión y hallar el coeficiente de correlación entre ambas. ¿Están relacionadas las variables?) Lo están linealmente?

11.4 INTERPRETACION DE RELACIONES

Es importante señalar que la existencia de una relación estadística entre dos variables, constatada por ejemplo a partir de su coeficiente de correlación en una muestra, no significa necesariamente que haya una relación de causalidad entre las mismas. Una correlación constatada entre dos variables puede presentarse fundamentalmente en dos contextos diferentes:

A - Existe una dependencia causal unidireccional. La relación entre TEMPER y CONSUMO es de este tipo, pues está claro (no por los datos estadísticos sino por el conocimiento previo existente sobre el tema) que la disminución de la temperatura ambiental influye en el consumo de energía, por utilizarse ésta en la climatización de las naves de la factoría. El valor de **r** y el cálculo de la recta de regresión que se expone en el siguiente apartado, permiten cuantificar la magnitud de esta relación, lo que resulta imprescindible si se desea controlar el consumo.

B - Las dos variables dependen parcialmente de otra u otras variables que no se están a lo mejor considerando. La correlación entre ESTATURA y PESO se debe posiblemente a que ambas variables vienen condicionadas por las características genéticas del individuo así como por las condiciones en que se ha desarrollado.

Autoevaluación: ¿Crees que tras alargar unos centímetros a una persona en el potro de tormentos ésta habrá aumentado algo de peso, como consecuencia de la relación existente entre ESTATURA y PESO.

Autoevaluación: En una encuesta sobre hábitos de consumo en hogares españoles se constató una correlación positiva entre consumo de zapatos y consumo de libros. (Los hogares que compraban más zapatos eran también los que compraban más libros).) A qué crees que se debe esta relación?) Qué te parece la idea de fomentar el hábito de lectura subvencionando el precio del calzado para que los hogares compren más zapatos? (Dado que "está demostrado" que cuantos más zapatos se compran más libros se compran).

11.5 MODELOS DE REGRESIÓN

En muchas ocasiones resulta necesario cuantificar la relación existente entre dos variables con el fin de predecir el valor de una de las variables a partir del valor constatado de la otra.

Por ejemplo, el responsable del control de consumo de energía de la factoría desea saber si el consumo de 290 termias constatado el día anterior puede considerarse "normal", sabiendo que la temperatura fue de 10ºC.

Para responder a preguntas como la anterior se utiliza en Estadística la recta de regresión . Mediante esta recta se pretende predecir el valor que en promedio corresponde a una variable Y, cuando otra variable X tiene un valor determinado.
La recta de regresión constituye un caso particular de los Modelos de Regresión que serán estudiados a continuación, como último capítulo del libro.

Estos modelos permiten analizar, no sólo la relación entre dos variables aleatorias, sino cuestiones mucho más generales, como el estudio de la relación que existe entre una determinada variable aleatoria y una o más variables explicativas, aleatorias o no, de las que la primera puede depender.