Práctica 6 de Estadística

El problema de la dependencia entre variables medibles.

6.1 Variables bidimensionales

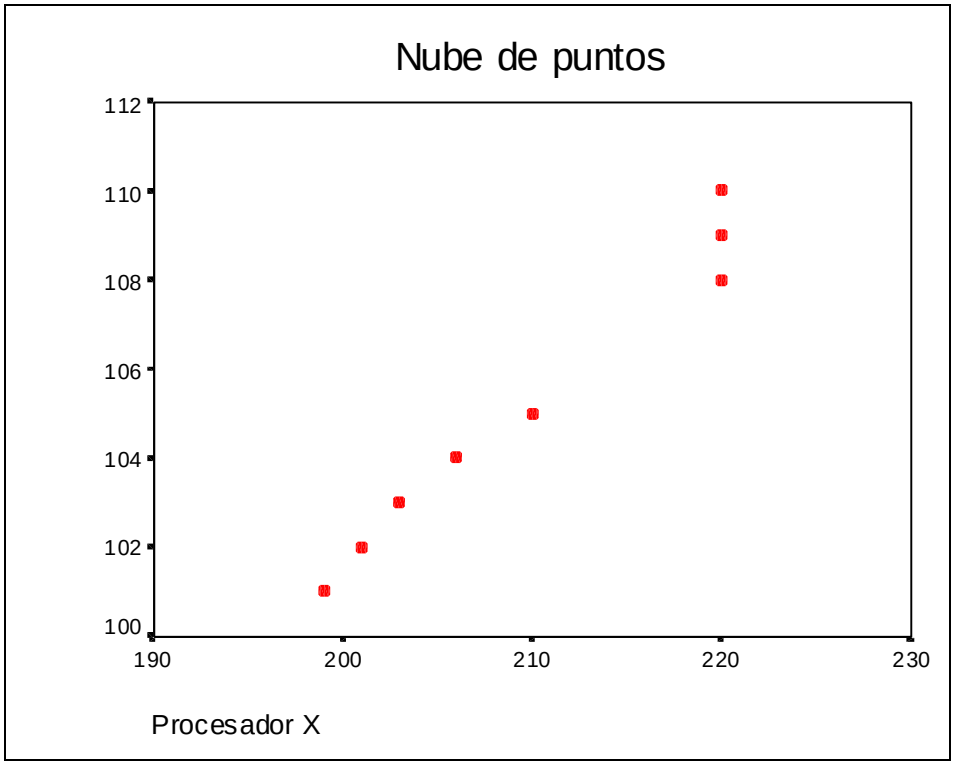
El problema de la dependencia consiste en el estudio de la relación existente entre dos o más características de los elementos de una población. Nos vamos a ocupar sólo del estudio de la relación entre dos características que puedan ser representadas mediante variables medibles, suponiendo que de cada elemento de la población podemos obtener un valor para cada una de las variables, es decir consideraremos *distribuciones estadísticas bidimensionales.*

- **Distribución bidimensional**: Intervienen dos conjuntos diferentes de datos, variando ambos a la vez.
  Por ejemplo supongamos que queremos comparar la velocidad de ejecución de dos procesadores. Para ello ejecutamos, con cada uno, 8 programas y medimos el tiempo empleado en cada ejecución. Consideremos las variables X e Y como el tiempo (en milisegundos) empleado. Estamos entonces ante una distribución bidimensional.

Al par de variables (X, Y), en donde para cada elemento de la población observamos conjuntamente el valor de las dos variables, lo llamamos *variable estadística bidimensional*. La variable la podemos representar mediante una tabla de dos filas (o dos columnas):

X	201	210	220	199	206	203	220	220
Y	102	105	108	101	104	103	109	110

También podemos representar la variable gráficamente, obteniendo su *diagrama de dispersión* o *nube de puntos*.



SPSS: Nube de puntos

Con el comando **Gráficos/Cuadro de diálogos antiguos/Dispersión...** obtenemos el cuadro de diálogo correspondiente en el que marcamos la opción **Simple** y pulsamos **Definir**. En el cuadro de diálogo obtenido seleccionamos las variables para los ejes, el tipo de gráfico y pulsamos **Aceptar**.

6.2 Dependencia funcional y dependencia estadística.

Decimos que una variable Y depende funcionalmente de otra X cuando podemos establecer una aplicación f que nos transforme los elementos de X en elementos de Y, es decir $Y = f(X)$. Es el caso, por ejemplo, de la relación existente entre el espacio, s, recorrido por un móvil y su velocidad v para un tiempo dado t_0:

$$s = v \cdot t_0$$

Pero hay muchas otras variables como estatura y peso, consumo y renta, temperatura y consumo de agua, etc, en las que es evidente que existe relación pero ésta no puede ser expresada mediante una función. Este tipo de relación no expresable de forma funcional se conoce mediante el nombre de *dependencia estadística,* y admite distintos grados, según que la relación entre las variables sea más fuerte o más débil.

- **Dependencia funcional**: Al determinar una variable, la otra queda determinada de forma unívoca.

- **Dependencia estadística**: Al determinar una variable, la otra queda determinada sólo en términos *probabilísticos*. En este caso se habla de ***correlación*** entre las variables.

La teoría de la ***regresión*** se ocupa de ajustar una función *f* que represente la relación estadística entre dos variables. Una de las dos se considera coma variable independiente, *X*, y la otra coma variable dependiente *Y*, de forma que mediante la función ajustada, para cada valor de *X* se obtenga un valor *f* (*X*) que pueda servir como predicción del valor de la variable *Y*.

Nos vamos a ocupar sólo del caso en que la función ajustada es una recta. La bondad o calidad del ajuste realizado será medido por un estadígrafo muestral.

6.3 Covarianza. Correlación lineal.

Consideremos una muestra bidimensional de tamaño *n*

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

La medida de relación lineal más simple entre las dos variables es la *covarianza muestral*, definida mediante la expresión

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

donde $\bar{x}$ e $\bar{y}$ son las respectivas medias de las variables *X* e *Y*:

(2) $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Podemos obtener una expresión más simple de la covarianza desarrollando la suma del segundo miembro:

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + n\bar{x}\bar{y}$$

y teniendo en cuenta (2) obtenemos

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}$$

Sustituyendo en (1) resulta

(3) $$\boxed{\mathrm{Cov}(X,Y)=\frac{1}{n-1}\left[\sum_{i=1}^{n}x_i y_i - n\,\bar{x}\,\bar{y}\right]}$$

Esta medida de la relación lineal existente entre las dos variables tiene dos inconvenientes:

1. Los puntos atípicos influyen considerablemente en el resultado
2. Depende de las unidades de medida de las variables

Para evitar estos inconvenientes, sobre todo para obtener una medida adimensional de la relación lineal existente, se define el *coeficiente de correlación lineal* (coeficiente de correlación de Pearson):

(4) $$\boxed{R=\frac{\mathrm{Cov}(X,Y)}{S_x\,S_y}}$$

donde S_x y S_y son las desviaciones típicas muestrales:

(5) $$S_x=\sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - n\,\bar{x}^2\right]}\,,\quad S_y=\sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n}y_i^2 - n\,\bar{y}^2\right]}$$

Como propiedades importantes del coeficiente de correlación lineal destacamos:

1. $|R|\le 1$
2. Si $|R|$ es próximo a 1, la dependencia lineal es fuerte.
3. Si $|R| = 1$, la dependencia entre las variables es funcional. Una variable es función lineal de la otra.
4. Si R es próximo a 0, la dependencia lineal es muy pobre.
5. $100R^2$ se interpreta normalmente como el porcentaje de la varianza común entre las dos variables.

SPSS: Covarianza y correlación

Con el comando **Analizar/Correlaciones/Bivariadas...** obtenemos el cuadro de diálogo correspondiente en el que trasladamos las dos variables a la ventana de **Variables** y señalamos el coeficiente de correlación de **Pearson**. En el cuadro de diálogo **Opciones** señalamos **Productos cruzados y covarianzas** (este paso no es necesario si sólo necesitamos el coeficiente de correlación)**.** Obtenemos los resultados al pulsar **Continuar**/**Aceptar**.

En nuestro ejemplo obtenemos Cov(X,Y) = 29,893 y R = 0,986. Deducimos, por tanto que la dependencia lineal es muy fuerte. Podemos decir, además, que el 97,2%

$(100 \cdot 0{,}986^2)$ de la variabilidad de ambas variables es común. Los valores los hemos obtenido de la tabla de resultados del programa SPSS:

Correlaciones

		Procesador X	Procesador Y
Procesador X	Correlación de Pearson	1,000	,986
	Sig. (bilateral)	,	,000
	Suma de cuadrados y productos cruzados	566,875	209,250
	Covarianza	80,982	29,893
	N	8	8
Procesador Y	Correlación de Pearson	,986	1,000
	Sig. (bilateral)	,000	,
	Suma de cuadrados y productos cruzados	209,250	79,500
	Covarianza	29,893	11,357
	N	8	8

6.4 Regresión lineal simple.

Dada una muestra bidimensional de tamaño *n*

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

para dos variables medibles *X* e *Y* entre las que suponemos que existe dependencia estadística. Vamos a obtener la recta de la forma general

$$y = ax + b$$

que mejor represente la relación de dependencia entre las dos variables basándonos en el criterio de mínimos cuadrados, es decir, la recta que minimice la suma de las diferencias cuadráticas entre los valores observados y los ajustados.

Esto se puede expresar así:

- $y_1, y_2, \ldots y_n$ son los valores observados
- $ax_1 + b, ax_2 + b, \ldots, ax_n + b$ son los valores ajustados.

Buscamos *a* y *b* para que sea mínima la expresión

(6) $$D = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Como *D* depende de dos variables *a* y *b*, hay que derivar respecto a cada una de ellas e igualar a cero:

$$\frac{\partial D}{\partial a} = -2\sum_{i=1}^{n}(y_i - ax_i - b)x_i = 0 \Rightarrow a\sum_{i=1}^{n}x_i^2 + b\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}x_i\, y_i$$

$$\frac{\partial D}{\partial b} = -2\sum_{i=1}^{n}(y_i - ax_i - b) = 0 \Rightarrow a\sum_{i=1}^{n}x_i + nb = \sum_{i=1}^{n}y_i$$

Despejando *b* de la segunda ecuación, y teniendo en cuenta (2), resulta

$$b = \frac{1}{n}\sum_{i=1}^{n}y_i - a\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) = \bar{y} - a\bar{x}$$

Sustituyendo ahora este resultado en la primera ecuación y volviendo a tener en cuenta (2) obtenemos

$$a\left(\sum_{i=1}^{n}x_i^2 - \bar{x}\sum_{i=1}^{n}x_i\right) = \sum_{i=1}^{n}x_i y_i - \bar{y}\sum_{i=1}^{n}x_i \Rightarrow a\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right) = \sum_{i=1}^{n}x_i y_i - n\bar{x}\,\bar{y}$$

de donde

(7) $$\boxed{a = \frac{\sum_{i=1}^{n}x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}}$$

Una vez calculado el coeficiente de regresión *a* obtenemos *b* mediante

(8) $$\boxed{b = \bar{y} - a\bar{x}}$$

La recta

$$y = ax + b$$

recibe el nombre de *recta de regresión de Y sobre X*, que se puede expresar sólo en función de *a*

$$y - \bar{y} = a(x - \bar{x})$$

probando que el punto $(\bar{x}, \bar{y})$ siempre pertenece a la recta.

Multiplicando el numerador y el denominador de (7) por el factor $\frac{1}{n-1}$ obtenemos la expresión más simple

(9) $$a = \frac{\mathrm{Cov}(X,Y)}{S_x^2}$$

siendo S_x^2 la varianza de la muestra $x_1, x_2,...,x_n$.

Resulta, pues, la ecuación de la *recta de regresión de Y sobre X*:

$$y - \bar{y} = \frac{\mathrm{Cov}(X,Y)}{S_x^2}(x - \bar{x})$$

Cambiando los papeles de las variables y razonando de forma similar obtendríamos la *recta de regresión de X sobre Y*:

$$x - \bar{x} = \frac{\mathrm{Cov}(X,Y)}{S_y^2}(y - \bar{y})$$

SPSS: Recta de regresión lineal (ecuación)

La secuencia de comandos **Analizar/Regresión/Estimación curvilínea...** abre el cuadro de diálogo, en él trasladamos la variable dependiente e independiente a las casillas correspondientes y marcamos **Incluir constante en la ecuación** y **Lineal** (si queremos la representación gráfica marcamos también **Representar los modelos**). Al **Aceptar** se obtiene el resultado en forma de texto donde **R cuadrado** representa el cuadrado del coeficiente de correlación lineal, **Constante** corresponde al termino independiente de la ecuación (en el desarrollo le hemos llamado *b*) y **b1** es la pendiente de la recta (término *a* del desarrollo)

En nuestro ejemplo el resultado obtenido es

Resumen del modelo y estimaciones de los parámetros

Variable dependiente:ProcesadorY

Ecuación	Resumen del modelo					Estimaciones de los parámetros	
	R cuadrado	F	gl1	gl2	Sig.	Constante	b1
Lineal	,972	205,085	1	6	,000	27,779	,369

La variable independiente esProcesadorX.

Por lo que la ecuación de la recta de regresión de *Y* sobre *X* es

$$y = 0{,}3691x + 27{,}7791$$

Podemos utilizar la recta de regresión para predecir el valor de la variable dependiente conocido el valor de la variable independiente, para ello sustituimos dicho valor en la ecuación de la recta.

Por ejemplo, si deseamos estimar el tiempo que tardará en ejecutarse un programa en el procesador Y, sabiendo que dicho programa ha consumido 215 milisegundos con el procesador X, tendríamos:

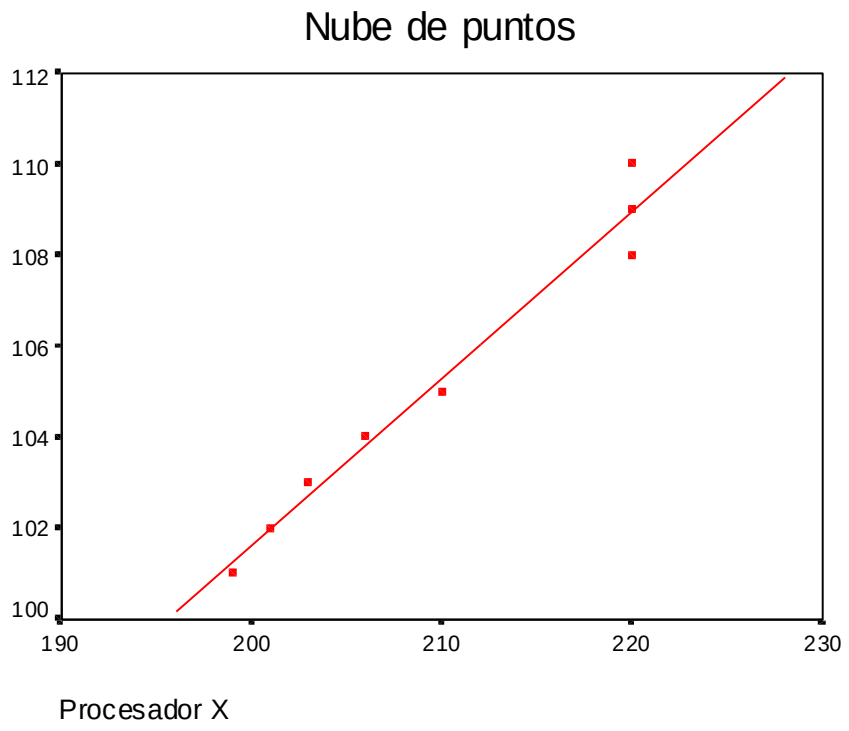
$$y = 0{,}3691 \cdot 215 + 27{,}7791 = 107{,}14 \text{ milisegundos}$$

SPSS: Recta de regresión lineal (gráfico)

Para obtener la gráfica de la recta de regresión podemos considerar dos métodos:

1. Activando la opción **Representar los modelos** en el cuadro de diálogo correspondiente a la ecuación de la recta de regresión (**Analizar/ Regresión/Estimación curvilínea...**)
2. Editando el gráfico de dispersión (doble clic sobre él), ejecutando los comandos **Elementos/Línea de ajuste total** y marcando la opción **Lineal**.

Para obtener el gráfico de la recta de regresión de nuestro ejemplo empleamos el método 2:



6.5 Estadígrafos para medir la bondad de ajuste.

La recta de regresión obtenida es la que mejor representa la relación de dependencia entre las variables según el criterio de mínimos cuadrados, es decir, minimiza la expresión

(10) $$D = \sum_{i=1}^{n} (y_i - ax_i - b)^2 = \sum_{i=1}^{n} d_i^2$$

donde

$$d_i = y_i - ax_i - b, \quad i = 1 \ldots n$$

se conocen como *residuos* del ajuste.

SPSS: Residuos
Los residuos se pueden obtener en la ventana de datos de la siguiente forma:
En el cuadro de diálogo Estimación curvilínea (**Analizar/Regresión/Estimación Curvilínea...**) pulsamos **Guardar** y marcamos **Residuos**, después de pulsar **Continuar / Aceptar** aparece una nueva variable con el nombre **err_1** (el número depende de si hay otra variable cuyo nombre comience igual) que contiene a los residuos.
Para calcular la suma de los cuadrados de los residuos (*D*) creamos una nueva variable (**Transformar/Calcular**) que contenga los cuadrados y después obtenemos su suma por ejemplo con el cuadro de diálogo Estadísticos descriptivos (**Analizar/ Estadísticos/ Descriptivos**) marcando la opción suma.

A la expresión

$$S_{ry}^2 = \frac{D}{n-2}$$

se le llama varianza residual de *Y* que, al desarrollar *D*, se puede expresar

$$S_{ry}^2 = \frac{\sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} x_i y_i - b\sum_{i=1}^{n} y_i}{n-2}$$

Para muestras suficientemente grandes y no alejadas de la normalidad se puede comprobar gráficamente la calidad del ajuste lineal. Si trazamos dos rectas paralelas a la recta $y = ax + b$ a una distancia de $2S_{ry}$, aproximadamente el 95% de los valores están comprendidos entre las dos rectas.

En nuestro caso D = 2,26 y S_{ry} =0,614.

El estadígrafo más usado para medir la bondad de ajuste es el coeficiente de correlación lineal *R*, del que ya hemos hablado anteriormente. Ahora estamos en disposición de demostrar que $|R|\leq 1$:

Puesto que *D* minimiza la suma de cuadrados de la diferencia entre y_i y los valores obtenidos con cualquier recta se tiene la desigualdad

$$D \leq \sum_{i=1}^{n} (y_i - \bar{y})^2$$

por lo que si

$$r = \sqrt{1 - \frac{D}{\sum (y_i - \bar{y})^2}}$$

se tiene $r^2 \leq 1$.

Ahora bien, teniendo en cuenta (8), (10) se puede escribir

$$D = \sum_{i=1}^{n} \left[(y_i - \bar{y}) - a(x_i - \bar{x}) \right]^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2a \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + a^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

por lo que

$$r^2 = \frac{2a \sum (x_i - \bar{x})(y_i - \bar{y}) - a^2 \sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2}$$

Pero recordando que

$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad S_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

y teniendo en cuenta (1) resulta

$$r^2 = \frac{2a\mathrm{Cov}(X,Y) - a^2 S_x^2}{S_y^2}$$

por fin usando la igualdad (9) obtenemos

$$r^2 = \frac{\mathrm{Cov}^2(X,Y)}{S_x^2 S_y^2}$$

que, junto con (4), prueba que $R^2 = r^2 \leq 1$ y, por tanto que, $|R| \leq 1$.

Además hemos obtenido una expresión nueva para el cuadrado de la correlación:

$$R^2 = 1 - \frac{D}{\sum (y_i - \bar{y})^2}$$

6.6 Ejercicios.

1. En un estudio llevado a cabo en Italia, 10 pacientes con Hipertrigliceridemia se sometieron a una dieta baja en grasas y alta en carbohidratos para investigar si había alguna relación entre estas variables. La tabla muestra los valores antes de comenzar la dieta:

Nivel colesterol (mmol/l)	5,12	6,18	6,77	6,65	6,36	5,9	5,48	6,02	10,34	8,51
Nivel triglicéridos (mmol/l)	2,3	2,54	2,95	3,77	4,18	5,31	5,53	8,83	9,48	14,2

Se pide:

a) Construye un diagrama de dispersión para estos datos.
b) ¿Existe evidencia de relación lineal entre los niveles de colesterol y triglicéridos antes de la dieta?
c) Estimar el nivel de triglicéridos cuando el nivel de colesterol es de 6,10 mmol/l.

2. Las mediciones de peso y altura de una muestra de ocho estudiantes se en la siguiente tabla:

Peso (Kg)	75	60	72	80	65	68	73	55
Altura (cm)	180	158	170	169	170	176	175	160

Se pide:

a) Construye el gráfico de dispersión (nube de puntos) de los datos. ¿Parece plausible ajustar una recta de regresión? ¿Cómo debe salir el coeficiente de correlación? Razona la respuesta.
b) Calcular la covarianza existente entre ambas variables así como el coeficiente de correlación.
c) Calcular las dos rectas de regresión.
d) ¿Qué altura cabe esperar para un estudiante que pese 70 kg?