Apuntes de Bioestadística: Tercer Ciclo en Ciencias de la Salud y Medicina

Francisco Javier Barón López
Dpto. Medicina Preventiva y Salud Pública (Bioestadística)
baron@uma.es
http://www.bioestadistica.uma.es/baron/

Francisco Téllez Montiel
Dpto. Matemática Aplicada

.

Diciembre de 2004

UNIVERSIDAD
DE MÁLAGA

Tabla de contenidos

Capítulo 1: Exploración de los datos ... 3
1.1 Datos Univariantes. ... 3
1.1.1 Datos Categóricos. ... 3
1.1.2 Datos Numéricos. ... 5
1.2 Datos Bivariantes ... 9
1.2.1 Categórica-categórica ... 10
1.2.2 Categórica-Numérica ... 11
1.2.3 Numérica-Numérica. ... 11
Capítulo 2: Intervalos de confianza ... 13
2.1 Error típico o estándar ... 13
2.2 Intervalo de confianza para una proporción ... 14
2.3 Intervalo de confianza para una media ... 15
2.4 Intervalos de confianza para otros parámetros ... 15
2.5 Contrastes de hipótesis basados en intervalos de confianza ... 16
Capítulo 3: Contrastes de hipótesis ... 17
3.1 ¿Qué es una hipótesis? ... 17
3.2 Tipos de error, significación, nivel de significación y potencia ... 18
3.2.1 Nivel de significación ... 18
3.2.2 Significación ... 19
3.2.3 Potencia ... 19
Capítulo 4: Diferencias que presenta una variable numérica entre dos grupos ... 21
4.1 Muestras apareadas o relacionadas ... 21
4.2 Muestras independientes ... 23
Capítulo 5: Diferencias que presenta una variable numérica entre varios grupos ... 26
5.1 Anova de un factor o una vía ... 26
5.1.1 En qué se basa el contraste ANOVA ... 27
5.1.2 Cómo se interpreta ANOVA ... 27
5.1.3 Contrastes no planeados o post-hoc ... 27
5.1.4 Comparaciones planeadas ... 29
5.1.5 ¿Qué hacer si no se verifican las premisas del modelo ANOVA? ... 29
5.2 Contraste no paramétrico de Kruskal-Wallis ... 33
Capítulo 6: Regresión múltiple ... 35
6.1 Aplicaciones de la regresión múltiple ... 35
6.2 Requisitos y limitaciones ... 36
6.3 Variables numéricas e indicadoras (dummy) ... 37
6.4 Interpretación de los resultados ... 38
6.5 Variables confusoras ... 39
Capítulo 7: Independencia de variables categóricas ... 44
7.1 Limitaciones de la prueba de independencia ... 45
Capítulo 8: Identificación de factores de riesgo ... 50
8.1 Riesgo, Oportunidad, Riesgo Relativo y Odds Ratio ... 50
8.2 Regresión logística ... 52
8.2.1 Codificación de las variables ... 52
8.2.2 Requisitos y limitaciones ... 53
8.2.3 Interpretación del modelo ... 54

Capítulo 1: Exploración de los datos

Cuando abordamos el estudio de un conjunto de datos, antes de introducirnos en cuestiones más detalladas, es necesario hacer una exploración inicial de los mismos. Así podemos tener una idea más clara de las características principales de los datos que hemos recogido, y de las posibles asociaciones.

En primer lugar daremos unas ideas sobre la manera de presentar ordenadamente y resumir variables consideradas aisladamente de las demás, para después explorar conjuntamente grupos de variables.

1.1 Datos Univariantes.

Los métodos para visualizar y resumir los datos dependen de sus tipos, que básicamente diferenciamos en dos: ***categóricos*** **y** ***numéricos***.

Los datos **categóricos** (o *factores*) son aquellos que registran *categorías* o *cualidades*. Si hacemos una base de datos de pacientes, ejemplos de variables categóricas son el sexo, el estado civil, fumar. Dentro de las categóricas podemos a su vez distinguir entre variable nominal y ordinal. En esta última hay un orden entre las distintas categorías. Por ejemplo, en la variable *Intensidad del dolor* tenemos las categorías: no perceptible, dolor tenue, doloroso y muy doloroso.

Siguiendo con la misma base de datos de pacientes, si recogemos, el peso de una persona es una *cantidad numérica*. En particular **continua** (los valores dentro de cualquier intervalo son posibles); Esto no ocurre cuando recogemos el número de hijos; Esta variable es discreta.

Variables	*Categóricas*	*Nominales*
		Ordinales
	Numéricas	*Discretas*
		Continuas

1.1.1 Datos Categóricos.

Los datos categóricos los examinamos bien con tablas de frecuencias o con representaciones gráficas como diagramas de barras o de sectores.

Frecuencias y porcentajes

Las **frecuencias** pueden obtenerse en términos absolutos (*frecuencias absolutas*), mostrando las repeticiones de cada categoría, o bien en términos relativos (*porcentajes*), mostrando los participación de cada categoría en relación con el total. Las frecuencias absolutas se utilizan con muestras de tamaño pequeño, y las relativas tienen más sentido con muestras de tamaño grande.

Si las variables son categóricas ordinales (o numéricas) pueden sernos de interés los *porcentajes acumulados*. Nos indican para cada valor de la variable, en qué porcentaje de ocasiones se presentó un valor inferior o igual.

Diagrama de barras
El diagrama de barras se representa asignándole a cada modalidad de la variable una barra de una altura proporcional a su frecuencia absoluta o a su porcentaje. En ambos casos el gráfico es el mismo, sólo se modifica la escala.

Diagramas de sectores
En este diagrama se le asigna a cada valor un sector cuyo ángulo sea proporcional a su frecuencia. Se suele utilizar en datos categóricos nominales y no en los ordinales.

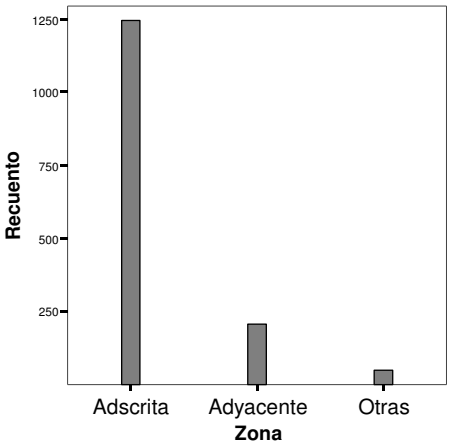
Ejemplo: En un estudio sobre el grado de satisfacción de los servicios ofrecidos en determinado Centro de Salud, se han tomado, entre otras, las variables *Zona* (categórica nominal) que recoge la zona a la que pertenece el enfermo y la variable *Grado de satisfacción* (categórica ordinal), que recoge el nivel de satisfacción que tiene con los servicios recibidos. La variable Zona tiene tres modalidades: Zona Adscrita al Centro, Zonas Adyacentes y Otras, codificadas para la recogida de datos con los valores 1, 2 y 3, respectivamente. La variable Grado de Satisfacción tiene las modalidades: No satisfecho, Satisfecho, Muy satisfecho, codificadas con 1, 2 y 3 respectivamente.

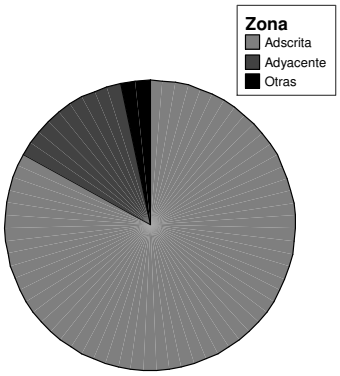
Si usamos SPSS, tanto tablas de frecuencias como los gráficos mencionados los encontramos en la opción de menú "Analizar – Estadísticos Descriptivos – Frecuencias".
Para la variable "*zona"* hemos obtenido las frecuencias, eliminando la columna de frecuencias acumuladas, puesto que no tiene sentido en las variables nominales.

Zona

		Frecuencia	Porcentaje
Válidos	Adscrita	1247	83,1
	Adyacente	204	13,6
	Otras	49	3,3
	Total	1500	100,0

El diagrama de barras y el de sectores son dos presentaciones diferentes de la misma información que hay en la tabla.





Para la variable "*grado de satisfacción"* sí tiene sentido la columna de porcentajes acumulados.

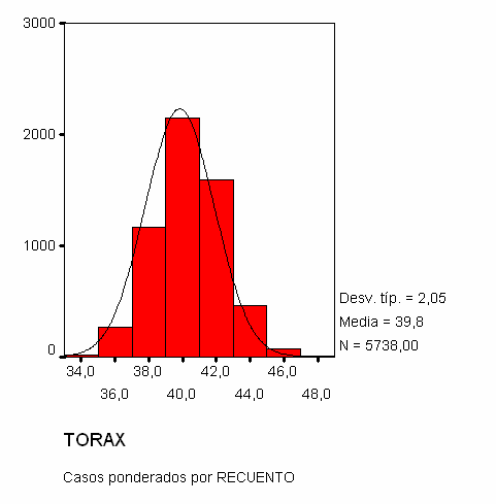
Grado de Satisfacción

		Frecuencia	Porcentaje	Porcentaje acumulado
Válidos	No satisfecho	464	30,9	30,9
	Satisfecho	867	57,8	88,7
	Muy satisfecho	169	11,3	100,0
	Total	1500	100,0	

1.1.2 Datos Numéricos.

Los datos numéricos son mucho más ricos en información que los datos categóricos. Por tanto además de las tablas, tenemos otras medidas que sirven para resumir la información que contienen. Dependiendo de cómo se distribuyan los datos, usaremos grupos de medidas de resumen diferentes.

Cuando se tiene una variable numérica, lo primero que nos puede interesar es alrededor de qué valor se agrupan los datos, y cómo se dispersan con respecto a él.



En múltiples ocasiones los datos presentan cierta distribución acampanada como la de la figura adjunta, denominada *distribución normal*. En estos casos con sólo dos medidas como son la *media* y la *desviación típica* tenemos resumida prácticamente toda la información contenida en las observaciones.

La **media:** es el promedio de todos los valores de la variable, es decir, la suma de todos los datos dividido por el número de ellos.

La **desviación típica** nos da una medida de la dispersión que tienen los datos con respecto a la media:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

La media y la desviación muestral no tienen tanto interés cuando los datos presentan largas colas u observaciones anómalas[1] (*outliers*), es decir, son muy influenciables por las *asimetrías* y los *valores extremos*. En estos casos, debemos considerar medidas más resistentes a estas influencias.

Como medidas de centralización resistentes podemos utilizar en sustitución de la media:

- La **mediana**, que es aquel valor que deja la mitad de los datos por debajo de él.
- La **media recortada** (*trimmed mean*), muy utilizada en datos preferentemente simétricos, con muchas observaciones anómalas y, que se obtiene eliminando un determinado porcentaje de los datos menores y mayores; Así calculamos la media sin contar con ese porcentaje de datos *extremos,* haciendo desaparecer su influencia.

En cuanto a las medidas de dispersión más resistentes podemos utilizar el **rango intercuartílico** (IQR), que es la diferencia entre el tercer cuartil y el primero. El **primer cuartil** (Q_1) deja al 25% de los datos por debajo de él y el **tercer cuartil** (Q_3) deja al 75%, por tanto sabemos que entre ambos valores se encuentra el 50% central de las observaciones.

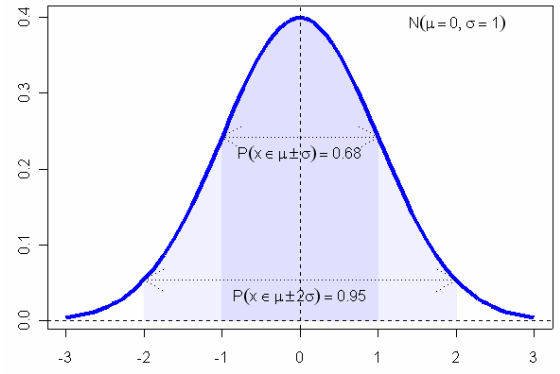
Cuando la muestra no parece tener una distribución normal, para obtener una idea aproximada de la distribución de los datos se acostumbra a mostrar **un resumen en cinco números**, que son el *valor mínimo*, el *primer cuartil*, la *mediana*, el *tercer cuartil*, y el *valor máximo*.

Ahora bien, ¿qué criterios aproximados podemos utilizar para clasificar unos datos como normales o no? Para ello destacamos varias características de la distribución normal. El alejamiento de las mismas es indicación de falta de normalidad:

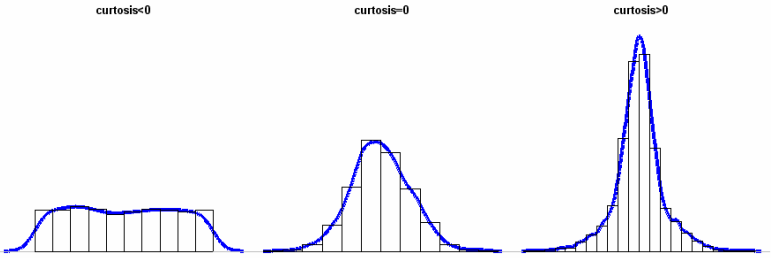
- Es simétrica (el coeficiente de asimetría vale cero)

[1] Observaciones demasiado grandes o pequeñas con respecto al resto de valores.

- Tiene forma de campana (el apuntamiento o curtosis vale cero).
- Coinciden la media y la mediana
- Aproximadamente el 95% de las observaciones se encuentran en el intervalo de centro la media y radio 2 veces la desviación típica.



Los indicadores que miden la simetría y la forma de la campana son el coeficiente de asimetría (negativo en distribuciones con cola a la izquierda, positivo en distribuciones con cola a la derecha) y la curtosis (negativa para las aplanadas y positiva para las apuntadas).



En la siguiente gráfica mostramos cómo SPSS nos ofrece realizar los cálculos de algunas de las medidas mencionadas anteriormente. La ventana la encontramos al pulsar el botón "Estadísticos…" al realizar una tabla de frecuencias como las de la sección anterior.

Histogramas

Como representación gráfica de una variable numérica, la más usada es el **histograma**. Si la variable es discreta y presenta pocas modalidades, deberíamos considerar en su lugar el diagrama de barras mencionado con anterioridad.

Para dibujar el histograma dividimos el rango de la variable en *intervalos*; A cada intervalo le asignamos como frecuencia el número de datos que contiene. El histograma se forma construyendo rectángulos con *base* el intervalo y *altura* tal que el área del rectángulo sea proporcional a la frecuencia.

Diagrama de caja.

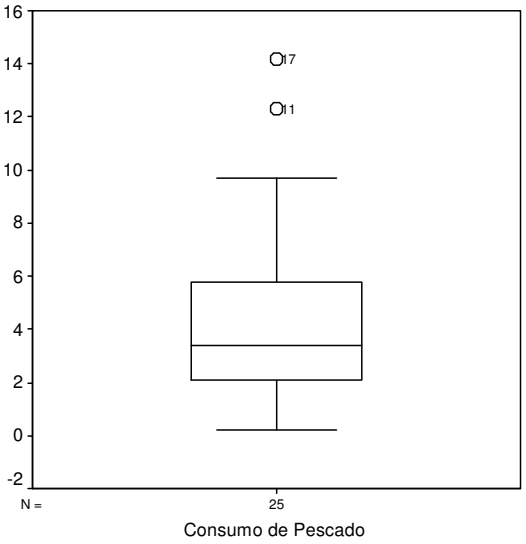
El **diagrama de caja** (*boxplot*) no es otra cosa que un resumen de la variable que se usa para ver rápidamente si los datos son *simétricos* o si incluyen observaciones anómalas. Está basado en el *resumen de los 5 números*. Su composición se basa en una **caja** cuyos extremos son el primer y tercer cuartil (aproximadamente), con una marca interior para la mediana, y dos **bigotes**, cuya misión es delimitar hasta donde podemos considerar los datos de las colas como no anómalos. El bigote derecho se extiende desde el límite superior de la caja hasta el valor más pequeño de los dos siguientes: el máximo valor de los datos o 1,5 veces el rango intercuartílico (anchura de la caja). Análogamente, el bigote izquierdo se extiende desde el límite inferior de la caja hasta el mayor de los dos siguientes: el valor mínimo de los datos o 1,5 veces el rango intercuartílico. Cualquier valor que quede fuera de los bigotes es marcado como anómalo.

Ejemplo: En un estudio sobre los hábitos de alimentación, realizado sobre 25 familias de igual tamaño, se estudiaron, entre otras, la variable *Consumo medio semanal de pescado (en Kg.)*. Una exploración (en SPSS, "Analizar – Estadísticos descriptivos – Explorar") arroja los siguientes resultados.

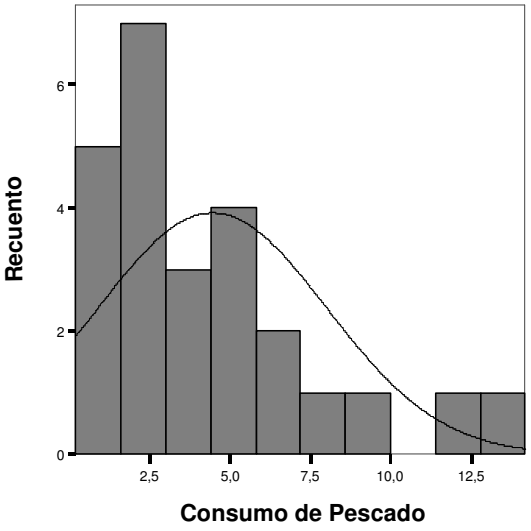
Descriptivos - Consumo de Pescado

		Estadístico	Error típ.
Media		4,444	,7124
Intervalo de confianza para la media al 95%	Límite inferior	2,974	
	Límite superior	5,914	
Media recortada al 5%		4,154	
Mediana		3,400	
Varianza		12,689	
Desv. típ.		3,5622	
Mínimo		,2	
Máximo		14,2	
Rango		14,0	
Amplitud intercuartil		3,800	
Asimetría		1,330	,464
Curtosis		1,516	,902

El valor de asimetría positivo indica la presencia de una cola a la derecha. En el diagrama de caja observamos la existencia de dos observaciones anómalas.



En el histograma se aprecia un cierto alejamiento de la distribución de la variable con respecto a la normal. En este tipo de situaciones se suele preferir describir la variable con el resumen de los cinco números que con la media y la desviación típica.



1.2 Datos Bivariantes

Si resumir la información de una variable es de por sí interesante, en investigación lo es mucho más el poner de manifiesto la posible relación entre dos de ellas. ¿Hay relación entre el tabaco y el cáncer de pulmón? ¿Aumentando la dosis de un medicamento, mejoramos la respuesta?

Para ello realizamos estudios donde intervienen ambas variables simultáneamente. Según sean los tipos de cada una de ellas usaremos técnicas diferentes.

1.2.1 Categórica-categórica

Cuando ambas variables son categóricas (o discretas con pocas modalidades), se suele presentar las observaciones en una *tabla de contingencia.* Esta una tabla de doble entrada donde se presentan la distribución de frecuencias conjunta de las dos variables.

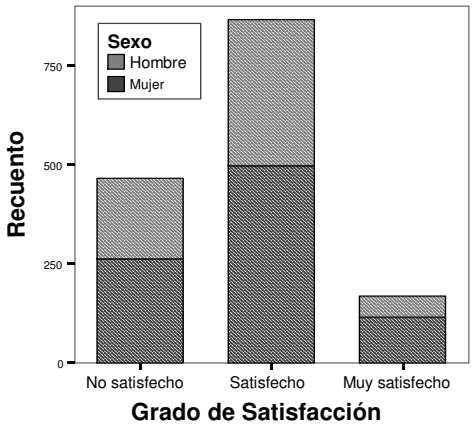
Ejemplo: Siguiendo el ejemplo anterior sobre el grado de satisfacción de los servicios prestados en determinado centro de salud se recogió además de la variable "*grado de satisfacción*", la variable "*sexo*". La tabla de contingencia es:

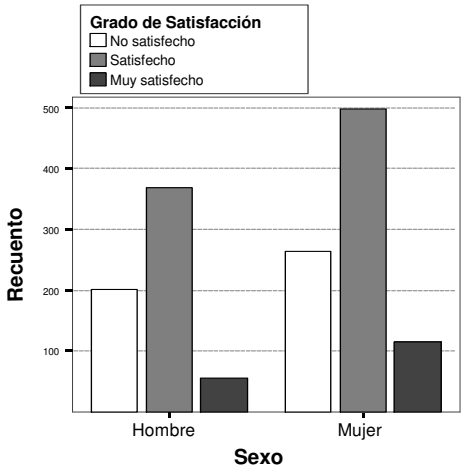
Tabla de contingencia Sexo * Grado de Satisfacción

Recuento

		Grado de Satisfacción			
		No satisfecho	Satisfecho	Muy satisfecho	Total
Sexo	Hombre	201	369	55	625
	Mujer	263	498	114	875
Total		464	867	169	1500

En cuanto a la representación, podemos utilizar el diagrama de barras apiladas o agrupadas.



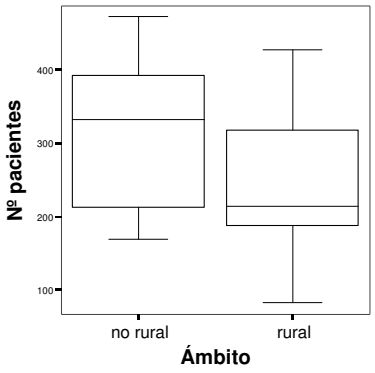


1.2.2 Categórica-Numérica

Supongamos que tenemos datos numéricos para varias categorías. Por ejemplo, en un experimento donde hacemos mediciones numéricas en dos grupos: uno al que se le aplica determinado tratamiento y otro de control. Podemos describir los resultados del experimento con sólo dos variables: Una variable categórica que representa el grupo de tratamiento, y otra que representa el resultado numérico

En estos casos, lo que se realiza es un estudio descriptivo de la variable numérica en cada una de las muestras y comparamos los resultados. Por ejemplo, se suelen enfrentar los diagramas de caja generados para cada categoría de la variable categórica.

Si usamos SPSS, tenemos a nuestra disposición la opción de menú "Analizar – Estadísticos descriptivos – Explorar…". En la casilla denominada "dependientes" situamos la variable numérica y en "factores" situamos la categórica.

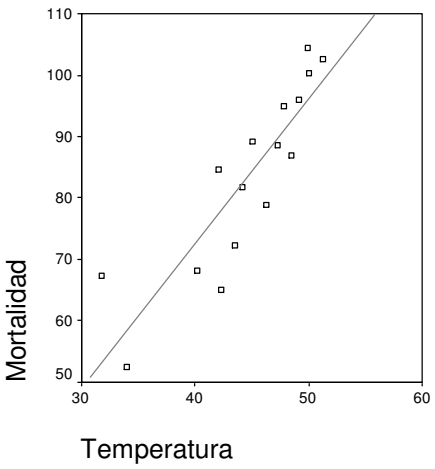


Ejemplo: Se realiza un estudio sobre el número de pacientes al mes que atiende un servicio de enfermería en dos ámbitos: rural y no rural. Para ello se toma una muestra de 51 centros. Representando los diagramas de caja para cada ámbito podemos comparar la distribución de la variable número de pacientes en cada ámbito.

1.2.3 Numérica-Numérica.

Cuando hablamos de comparar dos variables numéricas, pensamos en establecer la posible relación entre ellas. ¿Estarán relacionados la altura y el peso de los individuos? ¿Cuanto mayor es el tamaño del cerebro, mayor es el coeficiente intelectual?

La vía más directa para estudiar la posible asociación consiste en inspeccionar visualmente un **diagrama de dispersión** (*nube de puntos*). Si reconocemos una tendencia, es una indicación de que puede valer la pena explorar con más profundidad. Si es el caso, puede interesarnos proseguir con un análisis de regresión. En este tipo de análisis se pretende encontrar un modelo matemático (recta de regresión) que explique los valores de una de las variables (dependiente) en función de la otra (independiente). A ello le dedicamos un capítulo con posterioridad.

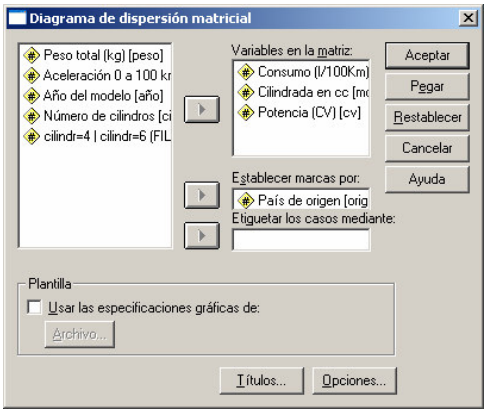


Ejemplo: Se discute la relación entre el índice de mortalidad en mujeres por neoplasma de mama y la temperatura media anual (medida en grados Fahrenheit). Se recogen datos de ambas variables de determinadas zonas de Gran Bretaña, Suecia y Noruega, y se construye el diagrama de dispersión y la recta de regresión, usando la opción de menú de SPSS "Gráficos – Dispersión… - Simple".

Se aprecia una relación directa entre las variables, que desde luego merece la pena

estudiar con más atención. Posiblemente haya otras variables que deban ser tenidas en cuenta para explicar la aparente relación que se observa.

Si estamos en la fase exploratoria de un estudio, podemos tener múltiples variables numéricas, observadas sobre diferentes muestras. Recomendamos vivamente hacer una rápida inspección visual con diagramas de dispersión matriciales. En SPSS se encuentran en la opción de menú "Gráficos – Dispersión… – Matricial".



Ejemplo: Se tiene una base de datos de vehículos, con variables numéricas como el *consumo*, la *cilindrada* y la *potencia*. También disponemos de la variable categórica "*país de origen*". Un diagrama de dispersión matricial nos permite reconocer visualmente qué variables parecen estar relacionadas. Lo apreciamos glbalmente, y también por cada uno de los estratos que define la variable cualitativa.