Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

Lección

Estadística

ANÁLISIS EXPLORATORIO DE DATOS (A.E.D.)

© Citar como: Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

Presentación:

La finalidad del Análisis Exploratorio de Datos (AED) es examinar los datos previamente a la aplicación de cualquier técnica estadística. De esta forma el analista consigue un entendimiento básico de sus datos y de las relaciones existentes entre las variables analizadas.

El AED proporciona métodos sencillos para organizar y preparar los datos, detectar fallos en el diseño y recogida de datos, tratamiento y evaluación de datos ausentes, identificación de casos atípicos y comprobación de los supuestos subyacentes en la mayor parte de las técnicas multivariantes.

En esta lección se va a dar una breve visión general de dicho conjunto de técnicas exponiendo, brevemente, cuál es su finalidad, ilustrada con ejemplos.

Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

Introducción

¿Existe algún tipo de estructura (normalidad, multimodalidad, asimetría, curtosis, linealidad, homogeneidad entre grupos, homocedasticidad, etc.) en los datos que voy a analizar?

¿Existe algún sesgo en los datos recogidos?

¿Hay errores en la codificación de los datos?

¿Cómo se sintetiza y presenta la información contenida en un conjunto de datos?

¿Existen datos atípicos (outliers)? ¿Cuáles son? ¿Cómo tratarlos?

¿Hay datos ausentes (missing)? ¿Tienen algún patrón sistemático? ¿Cómo tratarlos?

EN ESTA LECCIÓN SE HACE UNA BREVE REVISIÓN DE TÉCNICAS ESTADÍSTICAS PARA ABORDAR ESTE TIPO DE PROBLEMAS.

Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

Objetivos

1) Definir qué es el Análisis Exploratorio de Datos (A.E.D.) y cuáles son sus objetivos.
2) Indicar cuáles son las etapas a seguir en la realización de un A.E.D.
3) Seleccionar los métodos gráfico y numérico apropiados para examinar las características de los datos y/o relaciones de interés.
4) Comprobar si se verifican algunas hipótesis de interés en los datos (normalidad, linealidad, homocedasticidad).
5) Identificar casos atípicos univariantes, bivariantes y multivariantes.
6) Comprender los diferentes tipos de datos ausentes y evaluar su impacto potencial.

Salvador Figueras, M y **Gargallo, P.** (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

Apartados

1) ¿Qué es el Análisis Exploratorio de Datos (A.E.D.)?

2) Etapas del A.E.D.

3) Preparación de los Datos

4) Análisis Estadístico Unidimensional.

5) Estudio de la Normalidad

6) Análisis Estadístico Bidimensional

7) Datos Atípicos (outliers)

8) Datos Ausentes (missing)

Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

Contenidos

1.- ¿QUÉ ES EL ANÁLISIS EXPLORATORIO DE DATOS?

El Análisis Exploratorio de Datos (A.E.D.) es un conjunto de técnicas estadísticas cuya finalidad es conseguir un entendimiento básico de los datos y de las relaciones existentes entre las variables analizadas. Para conseguir este objetivo el A.E.D. proporciona métodos sistemáticos sencillos para organizar y preparar los datos, detectar fallos en el diseño y recogida de los mismos, tratamiento y evaluación de datos ausentes (missing), identificación de casos atípicos (outliers) y comprobación de los supuestos subyacentes en la mayor parte de las técnicas multivariantes (normalidad, linealidad, homocedasticidad).

El examen previo de los datos es un paso necesario, que lleva tiempo, y que habitualmente se descuida por parte de los analistas de datos. Las tareas implícitas en dicho examen pueden parecer insignificantes y sin consecuencias a primera vista, pero son una parte esencial de cualquier análisis estadístico.

Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

2.- ETAPAS DEL A.E.D.

Para realizar un A.E.D. conviene seguir las siguientes etapas:

1) Preparar los datos para hacerlos accesibles a cualquier técnica estadística.
2) Realizar un examen gráfico de la naturaleza de las variables individuales a analizar y un análisis descriptivo numérico que permita cuantificar algunos aspectos gráficos de los datos.
3) Realizar un examen gráfico de las relaciones entre las variables analizadas y un análisis descriptivo numérico que cuantifique el grado de interrelación existente entre ellas.
4) Evaluar, si fuera necesario, algunos supuestos básicos subyacentes a muchas técnicas estadísticas como, por ejemplo, la normalidad, linealidad y homocedasticidad.
5) Identificar los posibles casos atípicos (outliers) y evaluar el impacto potencial que puedan ejercer en análisis estadísticos posteriores.
6) Evaluar, si fuera necesario, el impacto potencial que pueden tener los datos ausentes (missing) sobre la representatividad de los datos analizados.

Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

3.- PREPARACIÓN DE LOS DATOS

El primer paso en un A.E.D. es hacer accesible los datos a cualquier técnica estadística. Ello conlleva la selección del método de entrada (por teclado o importados de un archivo) y codificación de los datos así como la de un paquete estadístico adecuado para procesarlos.

Los paquetes estadísticos son conjuntos de programas que implementan diversas técnicas estadísticas en un entorno común. Algunos de los más utilizados son SAS, BMDP, SPSS, SYSTAT, STATISTICA, STATA y últimamente MINITAB, S-PLUS, EVIEWS, STATGRAPHICS y MATLAB.

La codificación de los datos depende del tipo de variable. Los paquetes estadísticos existentes en el mercado proporcionan diversas posibilidades (datos tipo cadena, numéricos, nominales, ordinales, etc).

La inmensa mayoría de los paquetes estadísticos permite realizar manipulaciones de los datos previas a un análisis de los mismos. Algunas operaciones útiles son las siguientes:

- Combinar conjuntos de datos de dos archivos distintos
- Seleccionar subconjuntos de los datos
- Dividir el archivo de los datos en varias partes
- Transformar variables
- Ordenar casos
- Agregar nuevos datos y/o variables
- Eliminar datos y/o variables
- Guardar datos y/o resultados

Finalmente, y con el fin de aumentar la inteligibilidad de los datos almacenados, conviene asociar a la base de datos utilizada, un libro de códigos en el que se detallen los nombres de las variables utilizadas, su tipo y su rango de valores, su significado así como las fuentes de donde se han sacado los datos. Todos los paquetes anteriormente citados permiten esta posibilidad.

Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

4.- ANÁLISIS ESTADÍSTICO UNIDIMENSIONAL

Una vez organizados los datos, el segundo paso de un A.E.D. consiste en realizar una análisis estadístico gráfico y numérico de las variables del problema con el fin de tener una idea inicial de la información contenida en el conjunto de datos así como detectar la existencia de posibles errores en la codificación de los mismos.

El tipo de análisis a realizar depende de la escala de medida de la variable analizada. En la Tabla 1 se sugieren las representaciones gráficas y resúmenes descriptivos numéricos más aconsejables para realizar dicho análisis. En dicha Tabla se sobreentiende que las escalas más informativas pueden utilizar las medidas numéricas y representaciones gráficas de las escalas menos informativas además de las suyas propias (razón > intervalo > ordinal > nominal).

Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

Tabla 1

Medidas Descriptivas Numéricas y Representaciones Graficas aconsejadas en función de la escala de medida de la variable

Escala de medida	Representaciones gráficas	Medidas de tendencia central	Medidas de dispersión
Nominal	Diagrama de barras Diagrama de líneas Diagrama de sectores	Moda	
Ordinal	Boxplot	Mediana	Rango Intercuartílico
Intervalo	Histogramas Polígono de frecuencias	Media	Desviación Típica
Razón		Media Geométrica	Coeficiente de Variación

4.1.- Variables cualitativas

Las variables cualitativas son aquellas que no aparecen en forma numérica, sino como categorías o atributos como, por ejemplo, el sexo o la profesión de una persona. En dichas categorías puede haber un orden subyacente (variable ordinal) o no (variable nominal).

Los datos correspondientes a variables cualitativas se agrupan de manera natural en diferentes categorías o clases y se cuenta el número de datos que aparecen en cada una de ellas.

Se suelen representar mediante *diagrama de barras*, *sectores* o *líneas*.

Ejemplo 1 (Encuesta en un supermercado)

En la Tabla 2 se muestra la tabla de frecuencias del Estado Civil de una muestra extraída de los clientes de un Supermercado. Así mismo, en la Figura 1 se muestra el diagrama de sectores correspondiente a esta variable. (Todos las Tablas y Gráficos aquí presentados han sido hechos utilizando el paquete estadístico SPSS 10.0).

Se observa que la mayor parte de los clientes (75.9%) son casados que constituye el valor modal de la distribución de frecuencias, y que apenas acuden personas separadas (4%)

Tabla 2
Tabla de frecuencias del Estado Civil

Estado Civil

	Frecuencia	Porcentaje
Soltero	77	19.2
Casado	305	75.9
Viudo	16	4.0
Separado	4	1.0
Total	402	100.0

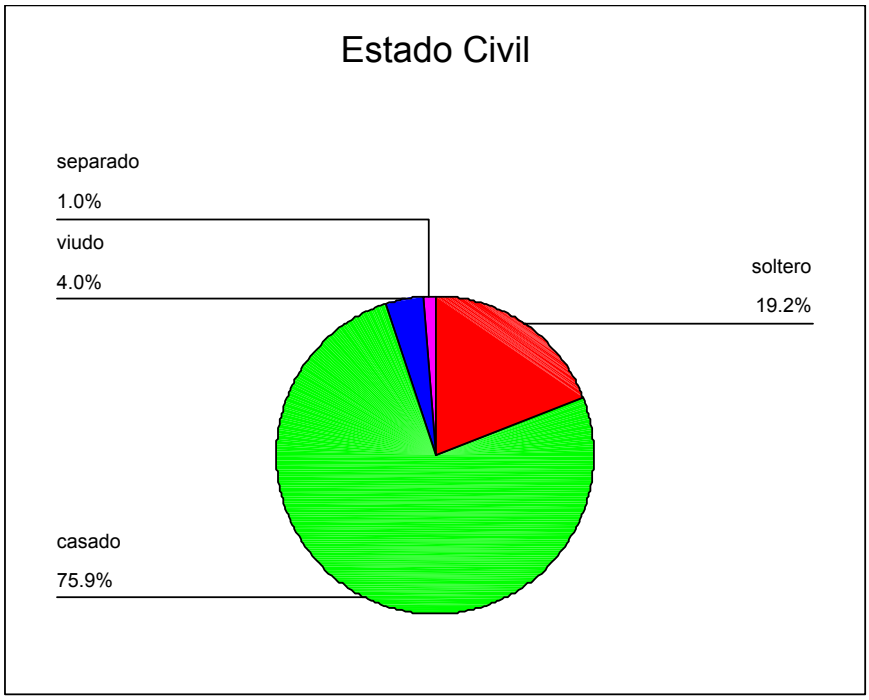

Figura 1: Diagrama de Sectores del Estado Civil

Ejemplo 2 (Padrón de Zaragoza)

En la Tabla 3 se muestra la distribución de frecuencias del Nivel de Estudios de la Población de Zaragoza en 1996 según los datos del padrón municipal realizado ese año. Así mismo, en la Figura 2, se muestra el diagrama de barras correspondiente a dicha variable.

Tabla 3
Tabla de frecuencias del Nivel de Estudios

Nivel de estudios

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Sin estudios	217199	36.1	36.3	36.3
	Bachiller elemental	199625	33.2	33.4	69.7
	Bachiller superior	104726	17.4	17.5	87.2
	Diplomado	36573	6.1	6.1	93.3
	Licenciado	40261	6.7	6.7	100.0
	Total	598384	99.5	100.0	
Perdidos	Sistema	3288	.5		
Total		601672	100.0		

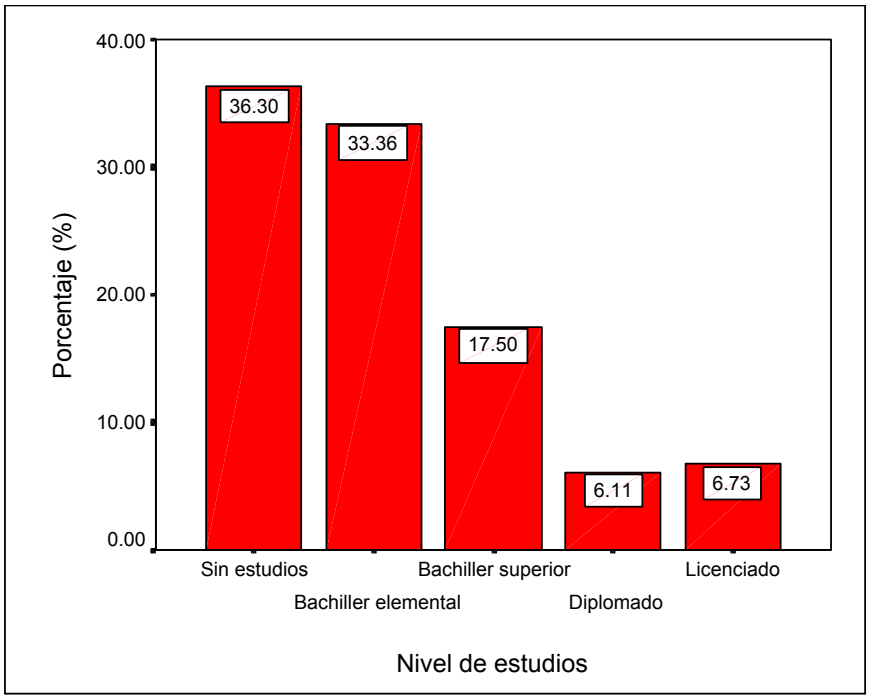

Figura 2: Diagrama de Barras del Nivel de Estudios

El nivel de estudios más frecuente es la categoría Sin Estudios, mientras que la categoría mediana es Bachiller Elemental. Se observa así mismo que, tan sólo un 12.84% tienen estudios superiores poniéndose de manifiesto, además, que existen más licenciados que diplomados en la ciudad debido, muy probablemente, al escaso número de Escuelas Universitarias existentes en la misma.

4.2- Variables cuantitativas

Las variables cuantitativas son las que pueden expresarse numéricamente. Una primera clasificación, basada en el tipo de valores que puede tomar, permite distinguir entre **variables cuantitativas discretas** – que son, frecuentemente el resultado de contar y, por tanto, toman sólo valores enteros – y **continuas**, que resultan de medir y pueden contener cifras decimales. Variables discretas son el número de lavadoras producidas por una empresa en un año. Variables continuas son aquellas cuyos valores pueden ser cualquier cantidad en un intervalo, como la temperatura, el peso o la altura de una persona o la superficie de las viviendas.

Las variables cuantitativas discretas con un número pequeño de valores se tratarían de manera similar a las variables cualitativas antes descritas.

Ejemplo 3 (Encuesta en un supermercado)

En la Tabla 4 se muestra la distribución de frecuencias del Número de Miembros que viven en la casa de una muestra de clientes de un supermercado. Así mismo, la Tabla 5 presenta algunas medidas descriptivas numéricas de dicha distribución y la Figura 3 su diagrama de barras.

Tabla 4
Tabla de frecuencias del Número de Miembros que viven en casa

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	1	.2	.3	.3
	1	30	7.5	7.5	7.8
	2	91	22.6	22.8	30.5
	3	87	21.6	21.8	52.3
	4	129	32.1	32.3	84.5
	5	43	10.7	10.8	95.3
	6	12	3.0	3.0	98.3
	7	7	1.7	1.8	100.0
	Total	400	99.5	100.0	
Perdidos	Sistema	2	.5		
Total		402	100.0		

Tabla 5
Estadísticos descriptivos de la variable
Número de Miembros que viven en casa

Estadísticos

miembros que viven en casa

N	Válidos	400
	Perdidos	2
Media		3.31
Mediana		3.00
Moda		4
Desv. típ.		1.33
Asimetría		.234
Error típ. de asimetría		.122
Curtosis		-.107
Error típ. de curtosis		.243
Mínimo		0
Máximo		7
Percentiles	25	2.00
	50	3.00
	75	4.00

El número medio de miembros que viven en casa es 3.31, siendo 4 el valor más frecuente y 3 el valor mediano. En el 50% de los casos el número de miembros oscila entre

2 y 4 (ver Tabla 5). Además, se observa que uno de los encuestados entendió incorrectamente la pregunta al contestar que nadie vivía en su casa (ver Tabla 4).

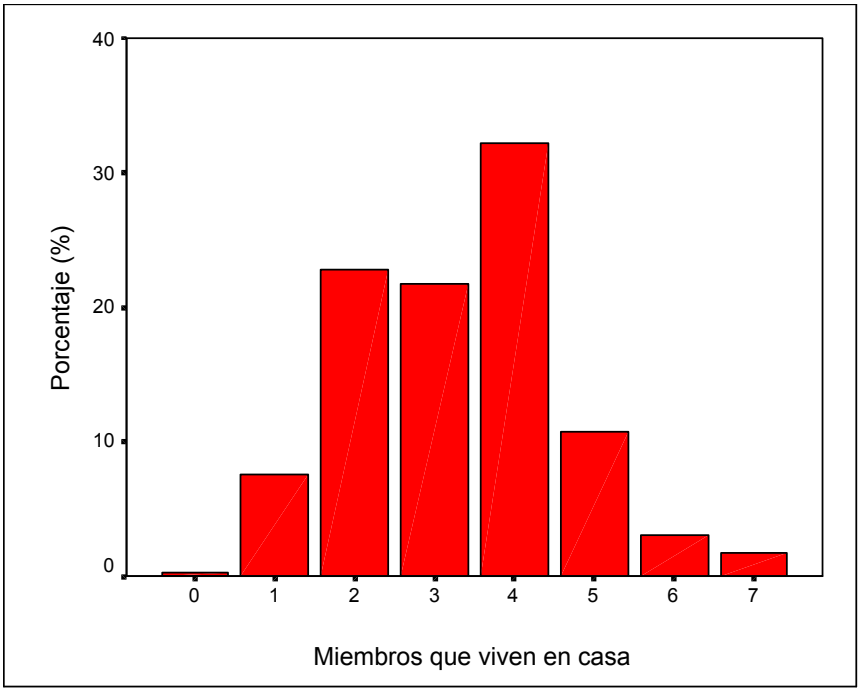

Figura 3: Diagrama de Barras del Número de Miembros que viven en casa

Si la variable analizada es continua o discreta con un elevado número de valores distintos se tabula como una distribución de frecuencias agrupadas y se representa gráficamente mediante histogramas, diagramas de tallos y hojas y boxplots con el fin de estudiar la forma de la distribución y analizar, en particular, la posible existencia de varias modas en la misma que pongan de manifiesto la presencia de diversos grupos homogéneos en la muestra.

La Figura 4 presenta de forma estilizada, algunos de los patrones que más frecuentemente presentan los histogramas. Las distribuciones (a) y (b) son simétricas alrededor de un valor central. El caso (a) presenta un único máximo – se dice que es una distribución unimodal – que necesariamente ha de coincidir con el centro de simetría y, en este caso, las medidas de tendencia central son una síntesis adecuada de la información contenida en la variable. La distribución (b) tiene dos máximos o modas – uno a cada lado del centro de simetría – Este patrón aparece cuando los datos responden a una mezcla de dos grupos heterogéneos y, siempre que sea posible, conviene estudiar ambos grupos por separado. Las formas que aparecen en (c) y (d) se denominan asimétrica a la derecha y a la

izquierda, respectivamente, e indican la presencia de un número significativo de valores muy altos (c) y bajos (d) susceptibles de distorsionar los resultados de análisis estadísticos posteriores. En estos dos últimos casos, los coeficientes de asimetría son significativamente distintos de cero - positivo en el caso (c) y negativo en el caso (d).

Conviene hacer notar finalmente, que aunque una distribución sea unimodal, no está libre de la presencia de valores anormalmente altos y bajos en ambas colas de la distribución que puedan distorsionar los resultados de un análisis estadístico. Para detectar este hecho se utiliza el coeficiente de curtosis, de forma que si la distribución es leptocúrtica (curtosis muy elevada), indica que sus colas son "muy pesadas" y, por lo tanto, se corre el riesgo antes nombrado.

Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

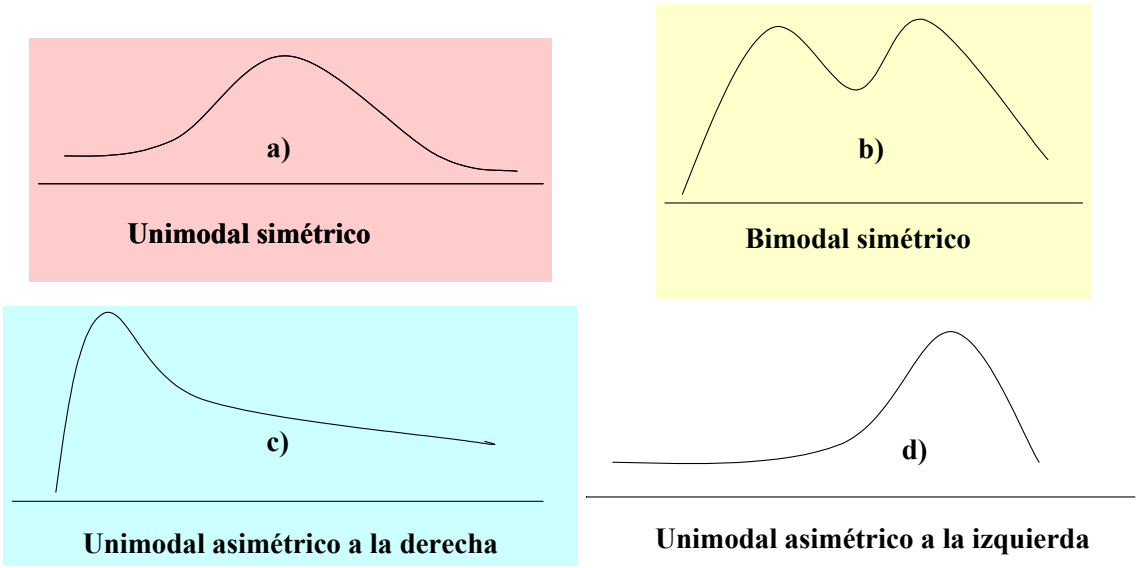

Figura 4:
Tipología de las distribuciones de frecuencias agrupadas

Ejemplo 4 (Datos macroeconómicos)

En este ejemplo analizamos diversas variables de una base de datos que contiene información macroeconómica de una muestra de países del mundo.

Exportaciones

En la Tabla 6 y las Figuras 5 y 6 se muestran los resultados del análisis estadístico de las Exportaciones de los países de la muestra. Así, en la Tabla 6 se muestran las medidas descriptivas numéricas de dicha variable y en las Figuras 5 y 6 su histograma y su diagrama de cajas, respectivamente. La media de las exportaciones ha sido 66.718 miles de millones de $ y su mediana 23.4. Esta diferencia refleja el elevado grado de asimetría hacia la derecha que se pone claramente de manifiesto con el histograma (Figura 4, tipo c) y sus coeficientes de asimetría (2.434) y curtosis (5.588).

Tabla 6
Medidas Descriptivas de las Exportaciones

Descriptivos

			Estadístico	Error típ.
Exportaciones (Miles de millones de $)	Media		66.7180	14.7693
	Intervalo de confianza para la media al 95%	Límite inferior	37.0223	
		Límite superior	96.4136	
	Media recortada al 5%		51.3139	
	Mediana		23.4000	
	Varianza		10688.493	
	Desv. típ.		103.3852	
	Mínimo		.73	
	Máximo		449.00	
	Rango		448.27	
	Amplitud intercuartil		57.5000	
	Asimetría		2.434	.340
	Curtosis		5.588	.668

Dicha asimetría se debe a las diferencias existentes entre los países en cuanto a tamaño económico tal y como se aprecia en la Figura 6 en la que los países más desarrollados del planeta (esencialmente los países del G7) tienen un número de exportaciones mucho mayores que el resto.

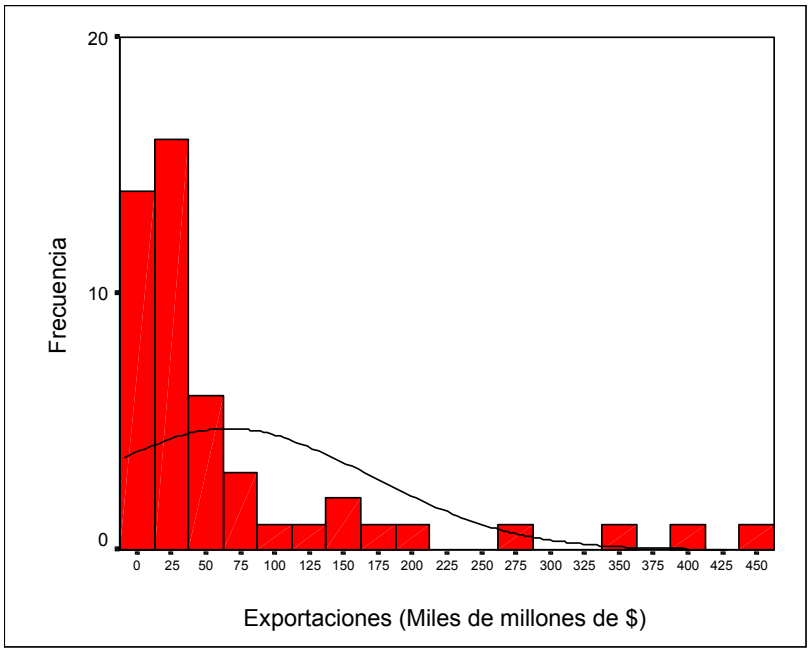

Figura 5: Histograma de las Exportaciones

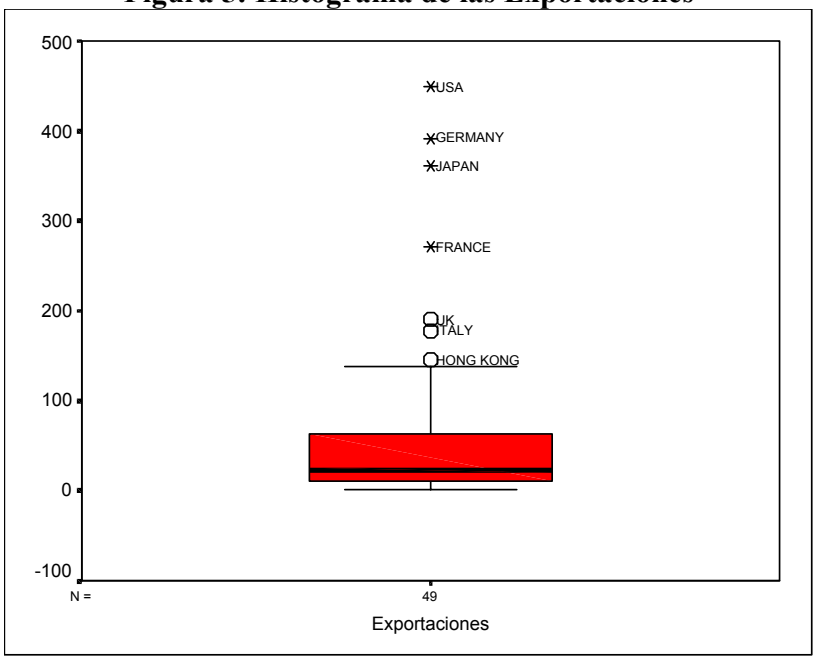

Figura 6: Diagrama de cajas de las Exportaciones

Esperanza de Vida

En la Tabla 7 y las Figuras 7 y 8 se muestran los resultados del análisis estadístico de la Esperanza de Vida de los países de la muestra. Así, en la Tabla 7 se muestran las medidas descriptivas numéricas de dicha variable y en las Figuras 7 y 8 su histograma y su diagrama de cajas, respectivamente. La esperanza de vida media es 72.98 años y su mediana es 74.99. Esta diferencia refleja la existencia de asimetría hacia la izquierda la

cual se pone claramente de manifiesto con el histograma (Figura 4 tipo d) y su coeficientes de asimetría (-0.883).

Tabla 7
Medidas Descriptivas de la Esperanza de Vida

Descriptivos

			Estadístico	Error típ.
Esperanza de vida (en años)	Media		72.9784	.7436
	Intervalo de confianza para la media al 95%	Límite inferior	71.4833	
		Límite superior	74.4734	
	Media recortada al 5%		73.2882	
	Mediana		74.9900	
	Varianza		27.091	
	Desv. típ.		5.2049	
	Mínimo		58.58	
	Máximo		80.09	
	Rango		21.51	
	Amplitud intercuartil		8.4200	
	Asimetría		-.883	.340
	Curtosis		-.063	.668

Dicha asimetría se debe a la existencia de países con una esperanza de vida mucho menor que el resto tal y como se observa en el histograma y en el mínimo valor de la variable (58.58 años) que corresponde a la India.

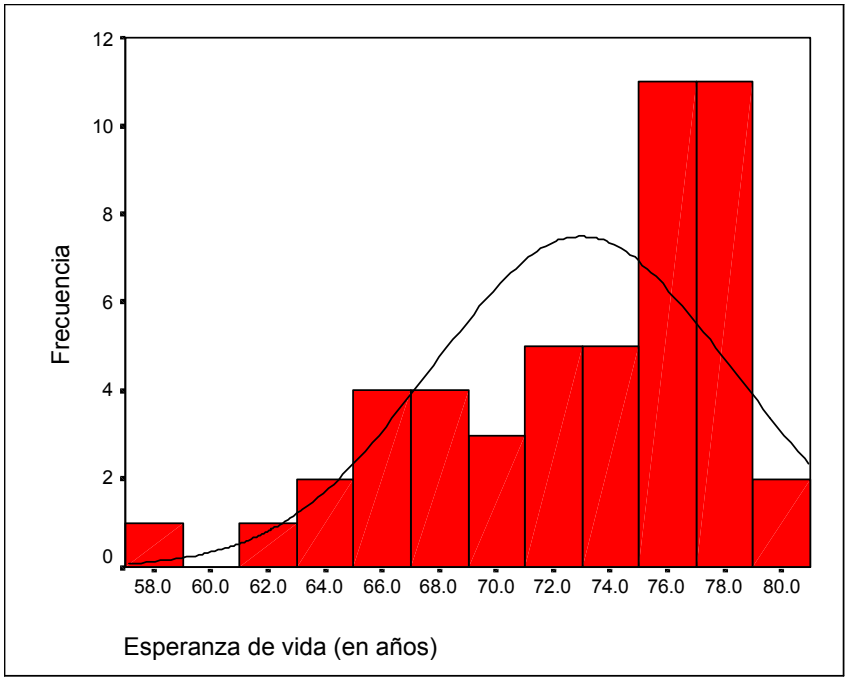

Figura 7: Histograma de la Esperanza de Vida

Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

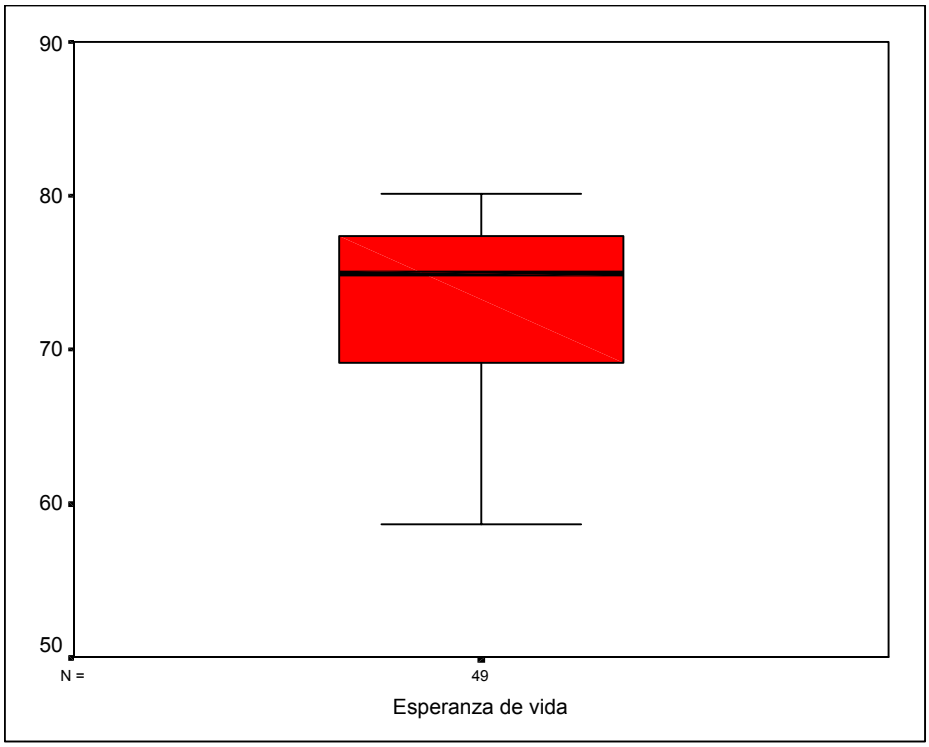

Figura 8: Diagrama de cajas de la Esperanza de Vida

Renta per Cápita

En la Tabla 8 y las Figuras 9 y 10 se muestran los resultados del análisis estadístico de la Renta per Cápita de los países de la muestra. Así, en la Tabla 8 se muestran las medidas descriptivas numéricas de dicha variable y en las Figuras 9 y 10 su histograma y su diagrama de cajas, respectivamente. La Renta per Per media es 9348.19$ y su mediana es 7692.58 $. Esta distribución es, sin embargo, claramente multimodal (Figura 4 tipo b) tal y como refleja su histograma (ver Figura 9) y la curtosis negativa (-1.421).

Tabla 8
Medidas Descriptivas de la Renta per Cápita

Descriptivos

			Estadístico	Error típ.
Renta per Cápita (en $)	Media		9348.1982	965.6046
	Intervalo de confianza para la media al 95%	Límite inferior	7406.7199	
		Límite superior	11289.6764	
	Media recortada al 5%		9178.6583	
	Mediana		7692.5837	
	Varianza		45687225	
	Desv. típ.		6759.2325	
	Mínimo		373.81	
	Máximo		22198.12	
	Rango		21824.31	
	Amplitud intercuartil		13101.1444	
	Asimetría		.259	.340
	Curtosis		-1.421	.668

Esta multimodalidad es debida, esencialmente, a las diferencias existentes entre los países del primer mundo respecto al resto, tal y como se refleja en la Figura 28 en la que se compara la Renta per Cápita de los países de la OCDE de la muestra con el resto.

Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

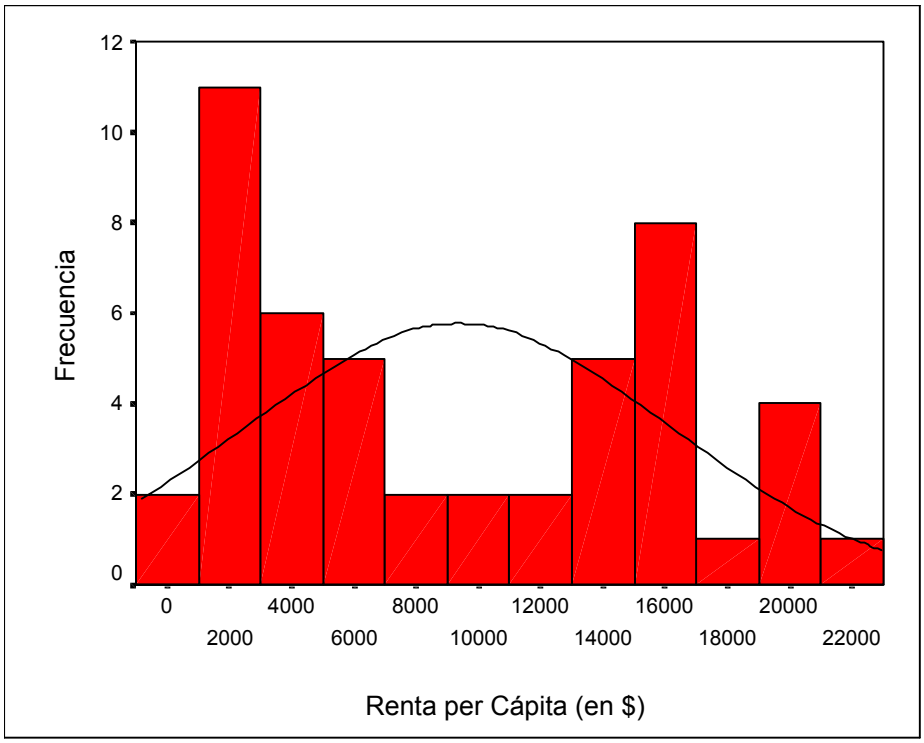

Figura 9: Histograma de la Renta per Cápita

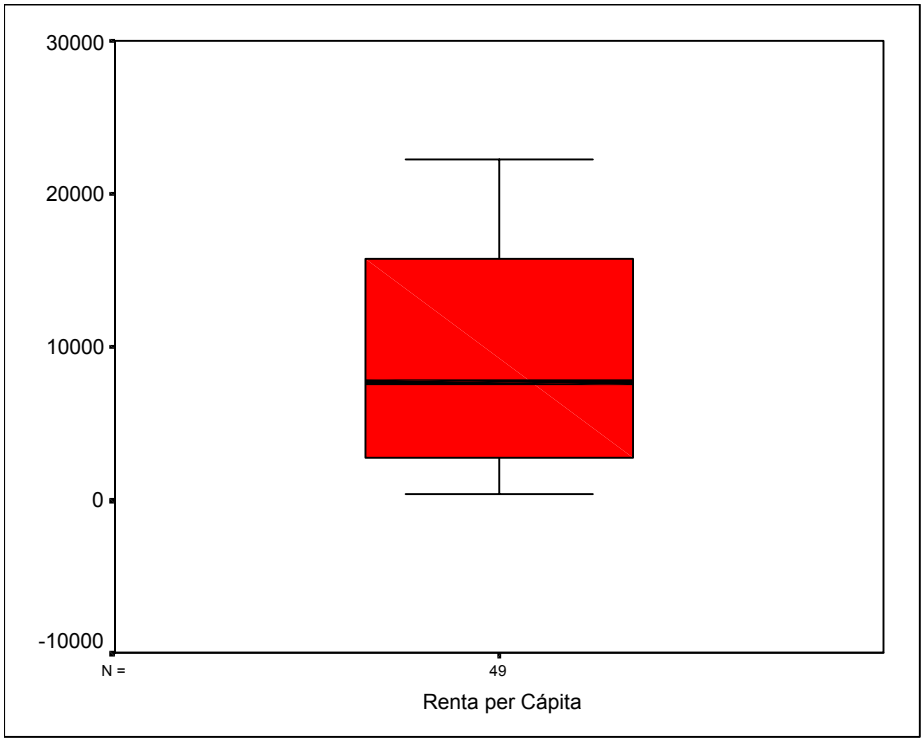

Figura 10: Diagrama de cajas de la Renta per Cápita

Crecimiento del PIB

En la Tabla 9 y las Figuras 11 y 12 se muestran los resultados del análisis estadístico de la Tasa Real del Crecimiento del PIB de los países de la muestra. Así, en la Tabla 9 se muestran las medidas descriptivas numéricas de dicha variable y en las Figuras 11 y 12 su histograma y su diagrama de cajas, respectivamente. La tasa real de crecimiento media es 2.33% y su mediana es 2%. Esta distribución es simétrica y unimodal (Figura 4 tipo a) tal y como refleja su histograma (ver Figura 11). Sin embargo su diagrama de cajas (ver Figura 12) muestra la existencia de 3 atípicos: dos con una elevada tasa de crecimiento (Kuwait y China) y uno con una baja tasa (Rusia). Este hecho es responsable del alto nivel de curtosis de la variable (3.701)

Tabla 9
Medidas Descriptivas de la Tasa de Crecimiento del PIB

Descriptivos

			Estadístico	Error típ.
Tasa real de crecimiento del PIB	Media		2.331E-02	5.962E-03
	Intervalo de confianza para la media al 95%	Límite inferior	1.132E-02	
		Límite superior	3.529E-02	
	Media recortada al 5%		2.224E-02	
	Mediana		2.000E-02	
	Varianza		1.741E-03	
	Desv. típ.		4.173E-02	
	Mínimo		-.12	
	Máximo		.15	
	Rango		.27	
	Amplitud intercuartil		4.850E-02	
	Asimetría		.167	.340
	Curtosis		3.701	.668

Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

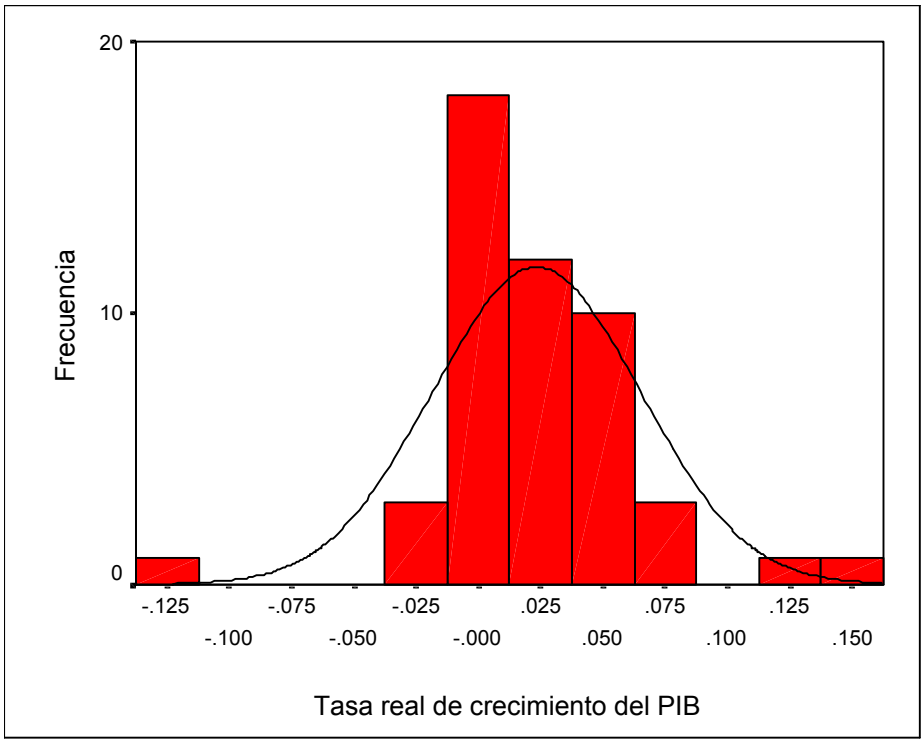

Figura 11: Histograma de la Tasa de Crecimiento del PIB

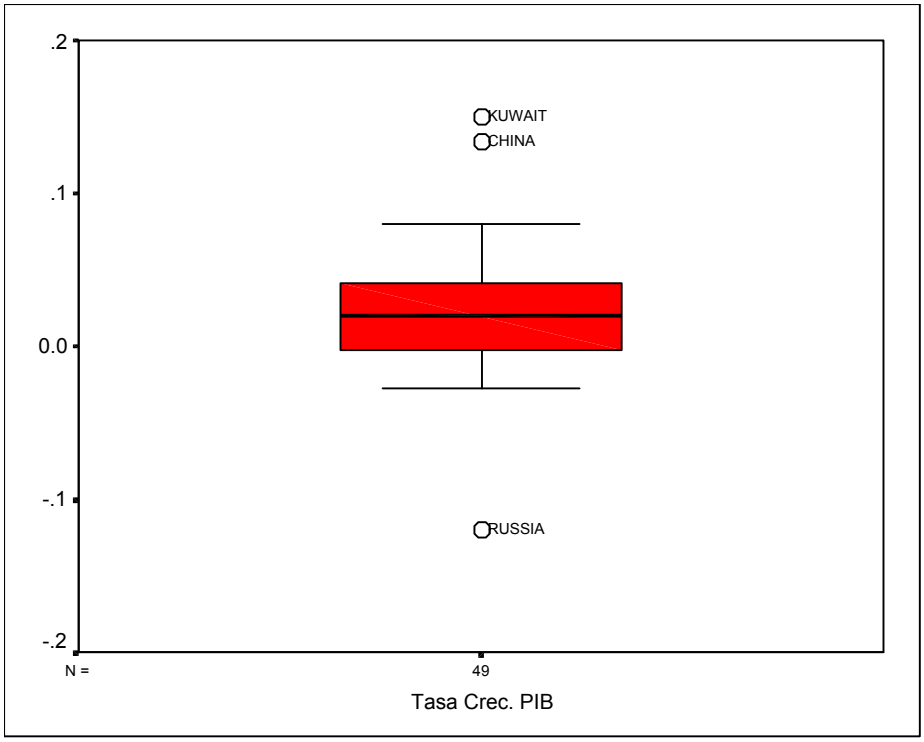

Figura 12: Diagrama de cajas de la Tasa de Crecimiento del PIB

5.- ESTUDIO DE LA NORMALIDAD

Muchos métodos estadísticos se basan en la hipótesis de normalidad de la variable objeto de estudio. De hecho, si la falta de normalidad de la variable es suficientemente fuerte, muchos de los contrastes utilizados en los análisis estadístico-inferenciales no son válidos. Incluso aunque las muestras grandes tiendan a disminuir los efectos perniciosos de la no normalidad, el investigador debería evaluar la normalidad de todas las variables incluidas en el análisis.

Existen varios métodos para evaluar la normalidad de un conjunto de datos que pueden dividirse en dos grupos: los ***métodos gráficos*** y los ***contrastes de hipótesis***.

5.1 Métodos gráficos

El método gráfico univariante más simple para diagnosticar la normalidad es una comprobación visual del ***histograma*** que compare los valores de los datos observados con una distribución normal. Aunque atractivo por su simplicidad, este método es problemático para muestras pequeñas, donde la construcción del histograma puede distorsionar la representación visual de tal forma que el análisis sea poco fiable.

Otras posibilidades, también basadas en información gráfica, consisten en realizar ***diagramas de cuantiles*** (Q-Q plots).

Los diagramas de cuantiles comparan en un sistema de coordenadas cartesianas, los cuantiles muestrales (eje X) con los cuantiles esperados bajo la hipótesis normalidad. Si la distribución de partida es normal dichos diagramas tenderán a ser rectas que pasan por el origen. Cuanto más se desvíen de una recta menos normales serán los datos. En la Figura 13 se muestran posibles diagramas de cuantiles según la forma de la distribución de frecuencias.

Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

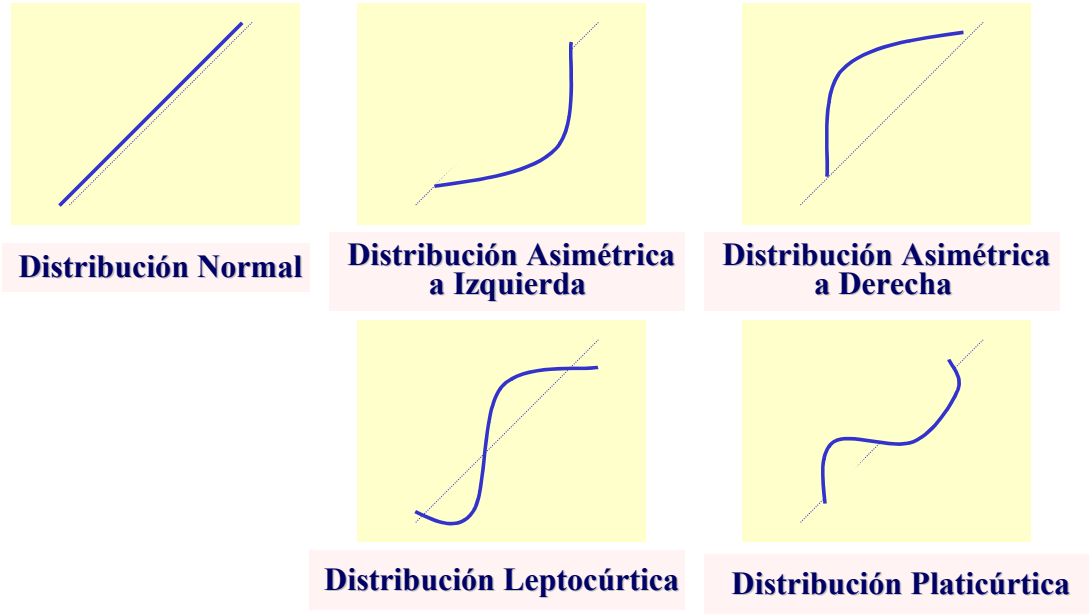

Figura 13: Diagramas de cuantiles correspondientes a distintos tipos de distribuciones

5.2 Contrastes de Hipótesis

La segunda de las formas para comprobar la normalidad de una distribución se efectúa a través de un ***contraste de hipótesis***. No existe un contraste óptimo para probar la hipótesis de normalidad. La razón es que la potencia relativa depende del tamaño muestral y de la verdadera distribución que genera los datos. Desde un punto de vista poco riguroso, el contraste de Shapiro y Wilks es, en términos generales, el más conveniente en muestras pequeñas ($n < 30$), mientras que el contraste de Kolmogorov-Smirnov, en la versión modificada de Lilliefors es adecuado para muestras grandes.

En el ***test de Kolmogorov-Smirnov*** la hipótesis nula que se pone a prueba es que los datos proceden de una población con distribución normal frente a una alternativa de que no es así. Este contraste calcula la distancia máxima entre la función de distribución empírica de la muestra y la teórica. Si la distancia calculada es mayor que la encontrada en las tablas, fijado un nivel de significación, se rechaza el modelo normal.

El ***contraste de Shapiro y Wilks*** se utiliza para muestras pequeñas ($n<30$) y utiliza el hecho de que si $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ es una muestra ordenada de la $N(\mu,\sigma)$ entonces:

$$E\left[\frac{x_{(i)}-\mu}{\sigma}\right] = C_{i,n} \quad \text{donde } C_{i,n} = \phi^{-1}\left(\frac{i-3/8}{n+1/4}\right)$$

por lo que $E[x_{(i)}] = \mu + \sigma C_{i,n}$ y el gráfico de $x_{(i)}$ frente a $C_{i,n}$ será una recta.

Dado que $\phi(C_{i,n}) + \phi(C_{n+1-i,n}) = 1$ $i=1,...,n/2$ se tiene que $C_{i,n} = -C_{n+1-i,n}$ por lo que $C_{1,n} + ... + C_{n,n} = 0$. El test de Shapiro-Wilks se basa en calcular el coeficiente de correlación entre $x_{(i)}$ y $C_{i,n}$ y cuanto más cerca de 1 esté, mayor será el grado de normalidad de la distribución y viene dado por la expresión:

$$r^2 = \frac{\left(\sum_{i=1}^{n} x_{(i)} C_{i,n}\right)^2}{ns^2\left(\sum_{i=1}^{n} C_{i,n}^2\right)}$$

Shapiro y Wilks evalúan la distribución del estadístico r^2 bajo hipótesis de normalidad y proporcionan un test que rechaza dicha normalidad cuando el ajuste es bajo, es decir, cuando el estadístico toma valores pequeños.

Otros contrastes muy utilizados son los **tests de asimetría y curtosis** cuyos estadísticos muestrales vienen dados por:

$$z_{asimetria} = \frac{asimetria}{\sqrt{\frac{6}{N}}} \text{ y } z_{curtosis} = \frac{curtosis}{\sqrt{\frac{24}{N}}}$$

donde N es el tamaño muestral. Si es cierta la hipótesis de normalidad ambos se distribuyen asintóticamente según una N(0,1).

Ejemplo 5 (Test realizado a estudiantes universitarios)

En la Tabla 10 y en las Figuras 14 y 15 se muestran los resultados obtenidos al realizar un análisis estadístico univariante a las puntuaciones conseguidas por un grupo de 40 estudiantes universitarios en un Test de Inteligencia. Así mismo, en la Figura 16 se muestra el diagrama de cuantiles y en la Tabla 11 los resultados del test de Kolmogorov-Smirnov con corrección de Lilliefors y del test Shapiro-Wilks. Tanto gráfica como numéricamente no se aprecian desviaciones significativas de la hipótesis de normalidad. Así, el histograma (en el que aparece superpuesta la densidad de la normal) es unimodal simétrico (ver Figura 14) y sin atípicos (ver Figura 15) y el diagrama de cuantiles (ver Figura 16) es lineal.

Tabla 10
Medidas Descriptivas de las Puntuaciones de un Test

Descriptivos

			Estadístico	Error típ.
TEST	Media		101.28	1.77
	Intervalo de confianza para la media al 95%	Límite inferior	97.69	
		Límite superior	104.86	
	Media recortada al 5%		101.28	
	Mediana		101.50	
	Varianza		125.640	
	Desv. típ.		11.21	
	Mínimo		79	
	Máximo		124	
	Rango		45	
	Amplitud intercuartil		15.75	
	Asimetría		-.015	.374
	Curtosis		-.626	.733

En cuanto a los test de hipótesis ni el test de Kolmogorov-Smirnov ni el de Shapiro-Wilks son significativos al 5% (sus p-valores son > 0.2 y 0.789, respectivamente. Lo mismo ocurre con los contrastes de asimetría y curtosis. En estos casos los estadísticos toman los valores:

$$z_{\text{asimetria}} = \frac{-0.015}{0.374} = -0.04 \text{ y } z_{\text{curtosis}} = \frac{-0.626}{0.733} = -0.854$$

cuyos p-valores son 0.968 y 0.393, respectivamente.

A la luz de estos resultados, cabe pensar que la variable *test* se distribuye normalmente.

Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

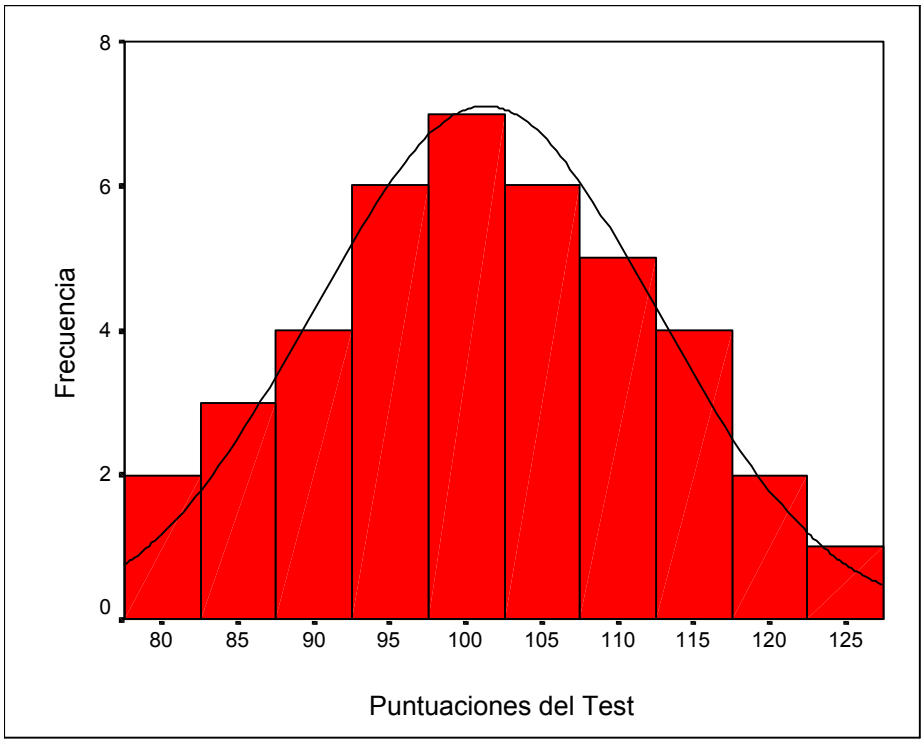

Figura 14: Histograma de las Puntuaciones del Test

Tabla 11
Contrastes de normalidad de las Puntuaciones de un Test

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Punt. test	.055	40	.200*	.981	40	.789

*. Este es un límite inferior de la significación verdadera.

a. Corrección de la significación de Lilliefors

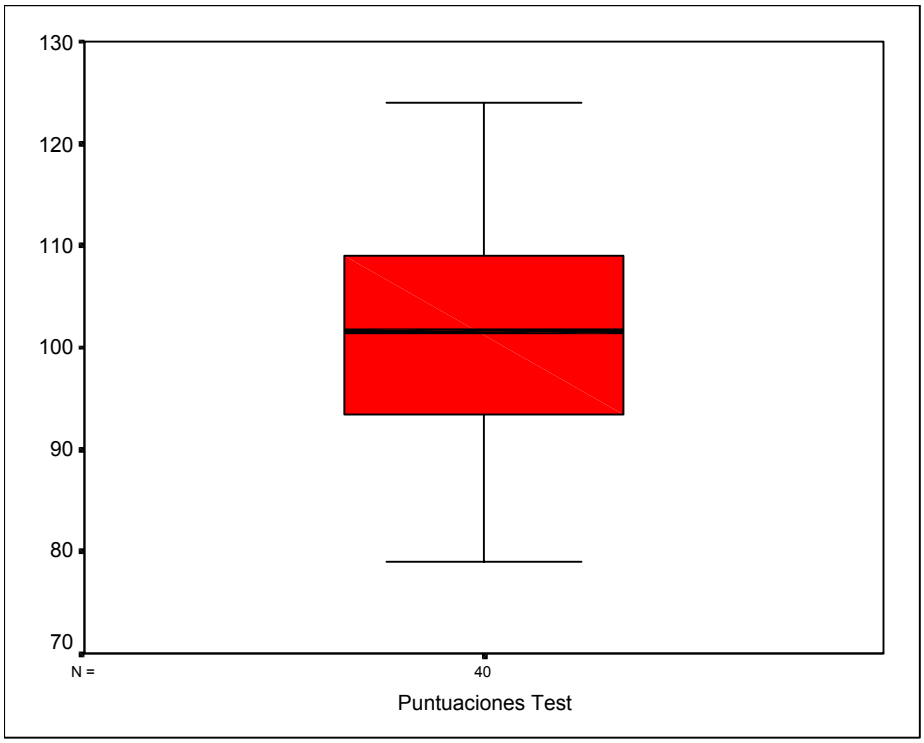

Figura 15: Diagrama de cajas de las Puntuaciones del Test

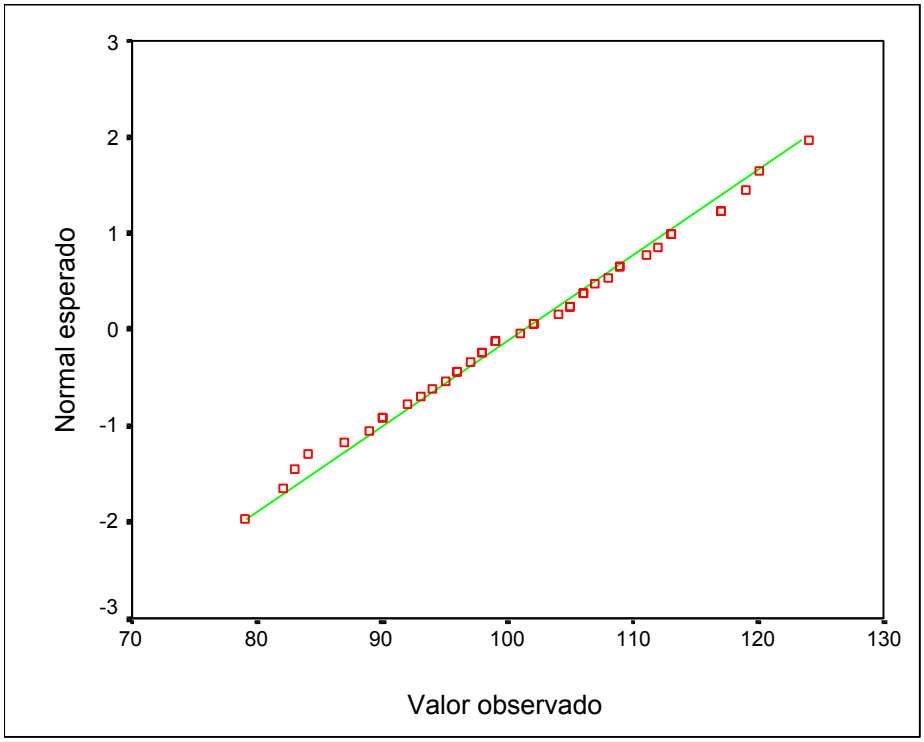

Figura 16: Diagrama de cuantiles de las Puntuaciones del Test

5.3 Transformaciones para alcanzar normalidad

En ocasiones la falta de normalidad de una variable puede arreglarse mediante una transformación de la misma. En la Tabla 12 se muestran algunas de las transformaciones más utilizadas.

Tabla 12
Transformaciones para conseguir normalidad

Forma de la Distribución	Transformación aconsejada
Asimetría Positiva	Log(X+C)
Asimetría Negativa	Log(C-X)
Leptocurtosis	1/X
Platicurtosis	X^2

Ejemplo 4 (Datos macroeconómicos, continuación)

Normalidad de las Exportaciones

En la Tabla 13 y la Figura 17 se muestran los resultados de analizar la normalidad de las Exportaciones.

Tabla 13
Análisis numérico de la normalidad de las Exportaciones

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Exportaciones (Miles de millones de $)	.297	49	.000	.635	49	.010**

**. Este es un límite superior de la significación verdadera.

a. Corrección de la significación de Lilliefors

Los resultados obtenidos concuerdan con el análisis estadístico realizado anteriormente volviendo a poner de manifiesto la falta de normalidad de la variable, siendo la razón de la misma, su elevado grado de asimetría positiva. Este hecho se pone de manifiesto tanto gráficamente, ya que el diagrama de cuantiles tiene forma de C abierta hacia abajo, la cual es típica de las distribuciones asimétricas a derecha (ver Figura 13), como numéricamente, puesto que tanto el test de Kolmogorov-Smirnov como el de Shapiro-Wilks rechazan la hipótesis de normalidad. Lo mismo ocurre con los contrastes de asimetría y curtosis cuyos estadísticos toman los valores:

$$z_{asimetria} = \frac{2.434}{0.340} = 7.159 \text{ y } z_{curtosis} = \frac{5.588}{0.688} = 8.365$$

con p-valores 0.000 en ambos casos.

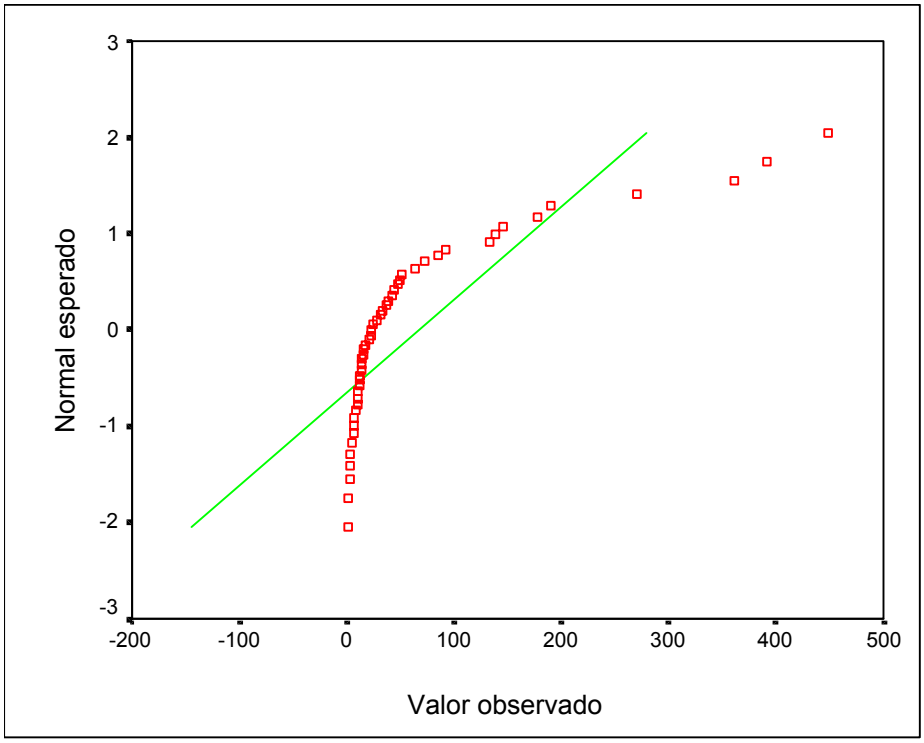

Figura 17: Diagrama de cuantiles de las Exportaciones

Un posible remedio a la falta de normalidad de esta variable es transformarla logarítmicamente. En la Tabla 14 y en la Figura 18 se muestran los resultados obtenidos al analizar la normalidad de la variable transformada, no observándose desviaciones significativas de la hipótesis de normalidad para esta variable.

Tabla 14
Análisis numérico de la normalidad
del logaritmo de las Exportaciones

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
LEXPORT	.061	49	.200*	.975	49	.523

*. Este es un límite inferior de la significación verdadera.

a. Corrección de la significación de Lilliefors

$$z_{asimetria} = \frac{-0.122}{0.340} = -0.359 \text{ y } z_{curtosis} = \frac{0.049}{0.688} = 0.073$$

(p-valores 0.720 y 0.942, respectivamente)

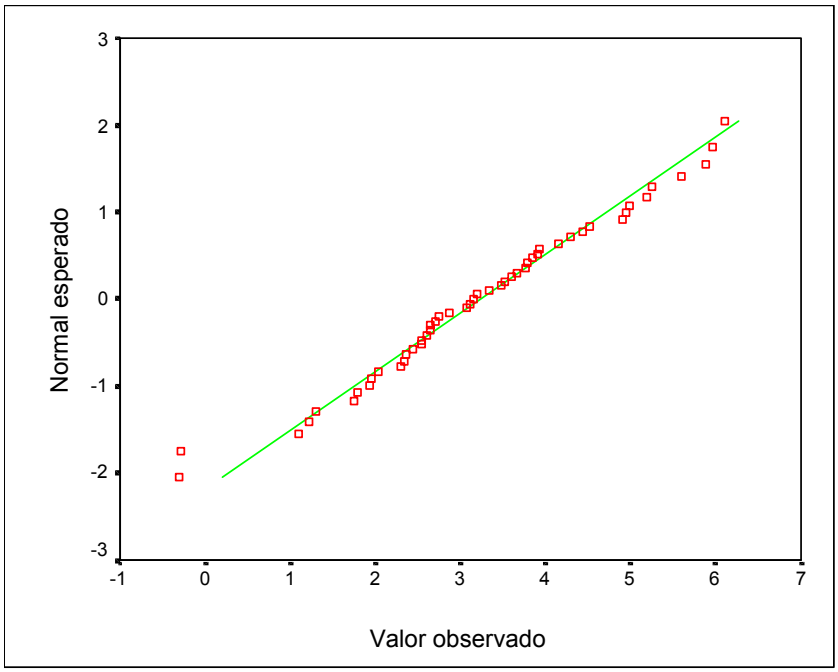

Figura 18: Diagrama de cuantiles del logaritmo de las Exportaciones

Normalidad de la Esperanza de Vida

En la Tabla 15 y en la Figura 19 se muestran los resultados de analizar la normalidad de la Esperanza de Vida.

Tabla 15
Análisis numérico de la normalidad de la Esperanza de Vida

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Esperanza de vida (en años)	.161	49	.003	.903	49	.010**

**. Este es un límite superior de la significación verdadera.

a. Corrección de la significación de Lilliefors

Los resultados obtenidos concuerdan con el análisis estadístico realizado anteriormente volviendo a poner de manifiesto la falta de normalidad de la variable siendo la razón de la misma su asimetría negativa. Este hecho se pone de manifiesto tanto gráficamente, ya que el diagrama de cuantiles tiene forma de C abierta hacia arriba, la cual es típica de las distribuciones asimétricas a derecha (ver Figura 13), como numéricamente, puesto que tanto el test de Kolmogorov-Smirnov como el de Shapiro-Wilks y el de asimetría rechazan la hipótesis de normalidad. Los contrastes de asimetría y curtosis toman los valores:

$$z_{asimetria} = \frac{-0.883}{0.340} = -2.597 \text{ y } z_{curtosis} = \frac{-0.063}{0.688} = -0.0943$$

cuyos p-valores son 0.009 y 0.925, respectivamente.

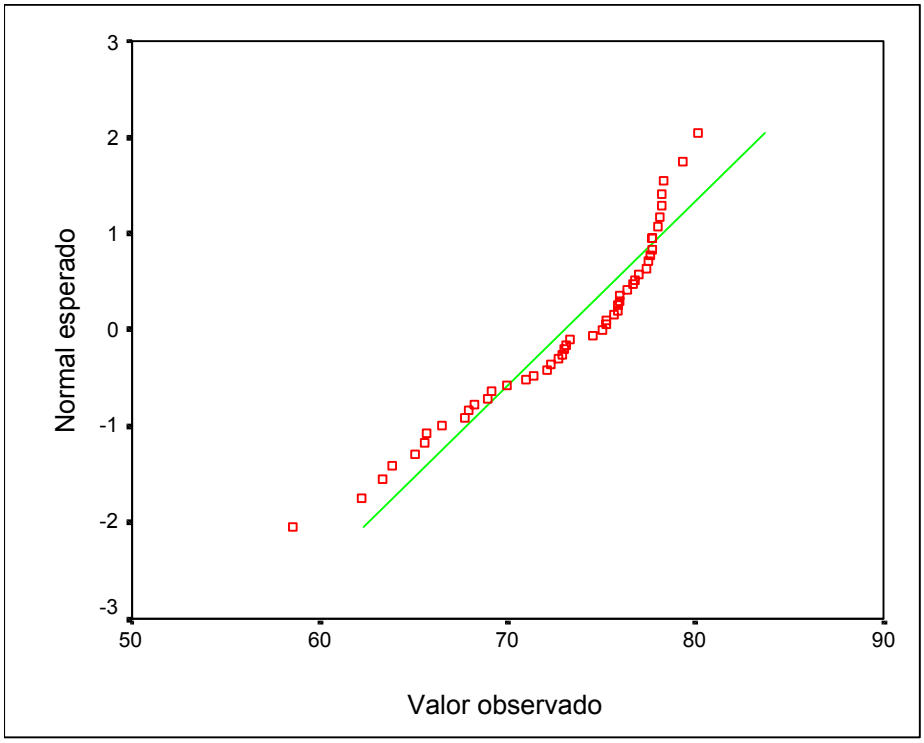

Figura 19: Diagrama de cuantiles de la Esperanza de Vida

Un posible remedio a la falta de normalidad de esta variable es transformarla logarítmicamente. En este caso la transformación aplicada es log(81-Esperanza). En la Tabla 16 y en la Figura 20 se muestran los resultados obtenidos al analizar la normalidad de dicha variable no observándose desviaciones significativas de la hipótesis de normalidad.

Tabla 16
Análisis numérico de la normalidad
del logaritmo de la Esperanza de Vida

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
LESPER	.074	49	.200*	.967	49	.353

*. Este es un límite inferior de la significación verdadera.

a. Corrección de la significación de Lilliefors

$$z_{asimetria} = \frac{-0.311}{0.340} = -0.915 \text{ y } z_{curtosis} = \frac{-0.159}{0.688} = -0.238$$

(p-valores 0.360 y 0.812, respectivamente)

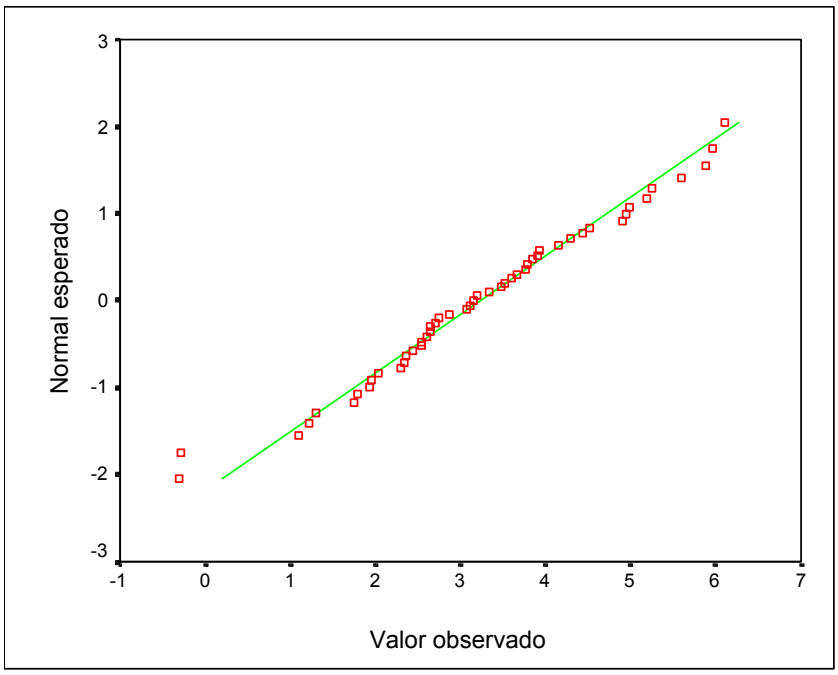

Figura 20: Diagrama de cuantiles del logaritmo de la Esperanza de Vida

En ocasiones la falta de normalidad va ligada a un problema de heterocedasticidad. En estos casos una transformación útil es la **transformación de Box-Cox** que estudiamos en el siguiente apartado.

6.- ANÁLISIS ESTADÍSTICO BIDIMENSIONAL

Una vez realizado el estudio unidimensional de cada variable por separado, el siguiente paso consiste en analizar la existencia de posibles relaciones entre ellas. Dicho estudio puede realizarse desde una óptica bidimensional o multidimensional. Éste último podría llevarse a cabo utilizando técnicas multivariantes (ver la página http://ciberconta.unizar.es/LECCION/anamul en la que se proporciona una visión general de dichas técnicas). En este apartado centraremos nuestra atención en el análisis bidimensional.

Las tres situaciones generales que pueden presentarse en este caso son las siguientes:

1) Ambas variables son cualitativas.
2) Ambas variables son cuantitativas.
3) Una variable es cuantitativa y la otra cualitativa.

6.1. Análisis de dos variables cualitativas

Se utiliza una tabla de contingencia que contiene en cada casilla la correspondiente frecuencia conjunta que representa el número de datos que pertenecen a la modalidad i-ésima de la primera variable y a la modalidad j-ésima de la segunda. A partir de dicha tabla podemos estudiar si las dos variables son o no independientes. Si son independientes no existe relación alguna entre ellas; en caso contrario analizaríamos el tipo y el grado de su dependencia tanto gráfica como numéricamente.

Ejemplo 6 (Encuesta en un Supermercado)

En la Tabla 17 y en la Figura 21 se muestra el análisis numérico y gráfico de la relación existente entre Poseer o no la tarjeta de compra de un Supermercado y la Frecuencia de Compra en el mismo.

La Tabla 17 contiene las frecuencias cruzadas de dichas variables, sus porcentajes fila y columna y los residuos tipificados corregidos con respecto a la hipótesis de independencia. Así mismo, se proporcionan los resultados de aplicar el test de la χ^2 de Pearson de independencia.

Tabla 17
Tabla de contingencia de Tarjeta vs Frecuencia de Compra

Tabla de contingencia Tarjeta * frecuencia

			frecuencia							
			esporadicamente	cada mes	cada 15 días	1 vez a la semana	2 veces por semana	de 3 a 5 días a la semana	todos los días	Total
Tarjeta	no	Recuento	42	16	10	19	13	12	5	117
		% de Tarjeta	35.9%	13.7%	8.5%	16.2%	11.1%	10.3%	4.3%	100.0%
		% de frecuencia	70.0%	50.0%	21.7%	21.6%	18.1%	21.4%	10.4%	29.1%
		Residuos corregidos	7.6	2.7	-1.2	-1.8	-2.3	-1.4	-3.0	
	si	Recuento	18	16	36	69	59	44	43	285
		% de Tarjeta	6.3%	5.6%	12.6%	24.2%	20.7%	15.4%	15.1%	100.0%
		% de frecuencia	30.0%	50.0%	78.3%	78.4%	81.9%	78.6%	89.6%	70.9%
		Residuos corregidos	-7.6	-2.7	1.2	1.8	2.3	1.4	3.0	
Total		Recuento	60	32	46	88	72	56	48	402
		% de Tarjeta	14.9%	8.0%	11.4%	21.9%	17.9%	13.9%	11.9%	100.0%
		% de frecuencia	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	73.004[a]	6	.000
Razón de verosimilitud	68.956	6	.000
Asociación lineal por lineal	54.259	1	.000
N de casos válidos	402		

a. 0 casillas (.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 9.31.

La Figura 21, por su parte, muestra los perfiles fila de dicha Tabla que comparan la frecuencia de compra entre los que poseen la tarjeta de compra y los que no la poseen.

La hipótesis de independencia es rechazada claramente (ver Tabla 17). Analizando, además, los residuos tipificados corregidos (Tabla 17) y el gráfico de los perfiles fila (Figura 21) se observa que las personas que poseen tarjeta tienden a comprar más frecuentemente en dicho Supermercado que aquéllas que no la poseen.

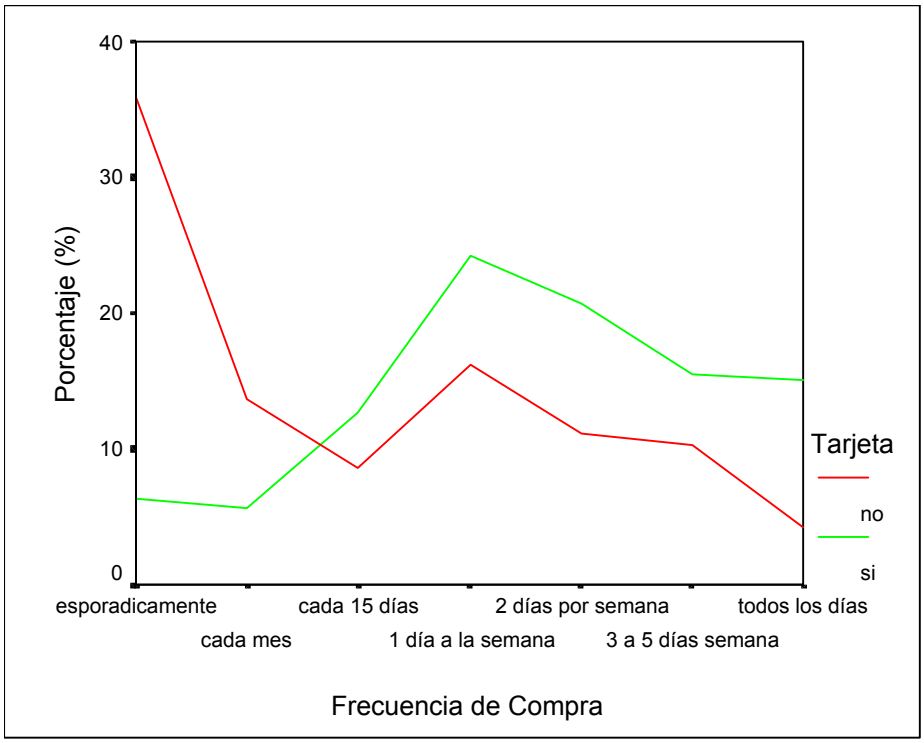

Figura 21: Perfiles fila de la Tabla Tarjeta vs Frecuencia de Compra

El análisis de la dependencia proseguiría cuantificando el grado de asociación entre las variables. Existen muchas medidas de asociación. Remitimos al lector a las páginas (PONER REFERENCIA A PÁGINAS DE ANÁLISIS DE CORRESPONDENCIAS Y DE ANÁLISIS DE TABLAS DE CONTINGENCIA DE CIBERCONTA) para profundizar sobre estos temas.

6.2. Análisis de dos variables cuantitativas

La distribución conjunta de dos variables puede expresarse gráficamente mediante un ***diagrama de dispersión*** que proporciona una buena descripción de la relación entre las dos variables.

La relación entre las variables también puede expresarse de forma numérica. Una medida de la relación entre dos variables que resuma la información del gráfico de dispersión y que no dependa de las unidades de medida es el ***coeficiente de correlación lineal***. Cuando las variables están relacionadas linealmente de forma exacta, el coeficiente de correlación lineal será igual a uno en valor absoluto. Cuando las variables no están relacionadas linealmente entre sí, el coeficiente de correlación lineal es cero. Para interpretar este coeficiente conviene mirar siempre el diagrama de dispersión de los datos para comprobar que son homogéneos y que no existen datos atípicos. La existencia de

correlación no implica una relación de causalidad entre las variables ni, en general, la no existencia de correlación permite deducir falta de causalidad.

Cuando se estudia la relación entre dos variables es importante asegurarse de que los individuos estudiados son homogéneos respecto a dichas variables. La Figura 22 muestra dos casos frecuentes de heterogeneidad.

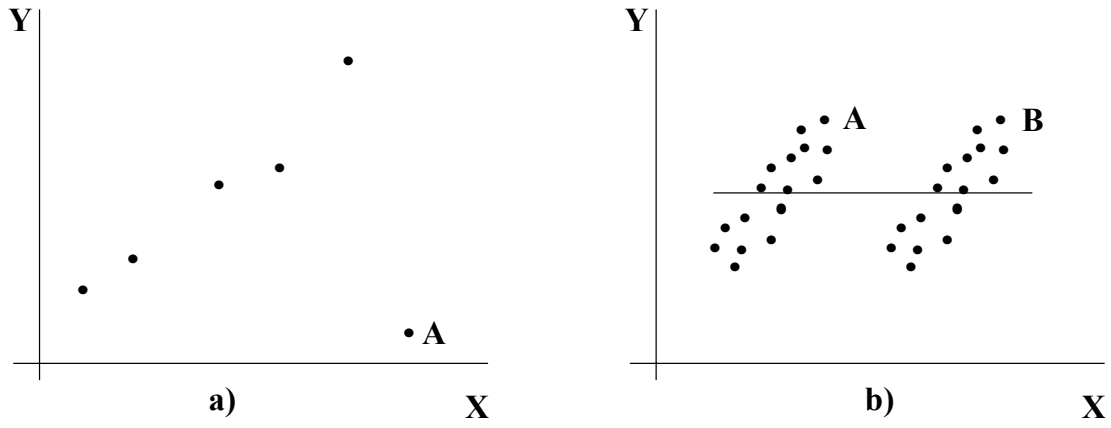

Figura 22: Dos casos frecuentes de heterogeneidad

En el caso a) hay un dato atípico o discordante con el resto, que modifica el signo de la correlación. Puede comprobarse que si el punto A no existiese, el coeficiente de correlación sería positivo, mientras que su presencia hace la correlación negativa. Ante una situación como ésta conviene asegurase de que no se ha cometido un error de medida o de trascripción del dato y que el individuo de la población al que le corresponde el dato atípico es homogéneo con respecto a los demás.

La Figura b) presenta otro caso de heterogeneidad. En este caso el gráfico indica que la relación entre las variables es distinta para los individuos del grupo A que para los del B y si calculamos un coeficiente de correlación para todos los datos obtendremos un valor muy pequeño. Sin embargo, si obtenemos los coeficientes para los grupos A y B separadamente, encontraremos que dentro de cada grupo hay una relación fuerte.

La conclusión fundamental de este análisis es que conviene asegurarse mirando el gráfico de dispersión que el coeficiente es un buen resumen del mismo. Tratar de interpretar un coeficiente de correlación sin haber visto previamente el gráfico de las variables puede ser muy peligroso.

6.2.1 Linealidad

La ***linealidad*** es un supuesto implícito de todas las técnicas multivariantes basadas en medidas de correlación, tales como la regresión múltiple, regresión logística, análisis factorial y los modelos de ecuaciones estructurales. Es, además, una forma indirecta de contrastar la normalidad conjunta de dos variables dado que si dicha hipótesis es cierta la relación existente entre ellas deberá ser lineal.

Dado que las correlaciones representan sólo la asociación lineal entre variables, los efectos no lineales no estarán representados en el valor de la correlación. Como resultado, es siempre prudente examinar todas las relaciones para identificar cualquier desplazamiento de la linealidad que pueda impactar la correlación.

La forma más común de evaluar la linealidad es examinar los gráficos de dispersión de las variables e identificar cualquier pauta no lineal en los datos. Una aproximación alternativa es ir a un análisis de regresión múltiple y examinar los residuos que reflejan la parte no explicada de la variable dependiente; por tanto, cualquier parte no lineal de la relación quedará reflejada en los residuos.

Ejemplo 7 (Salarios en un banco)

En la Figura 23 se muestra el diagrama de dispersión de los logaritmos de los Salarios Inicial y Actual correspondientes a una muestra de 474 empleados de un banco y superpuesta, la línea de regresión lineal.

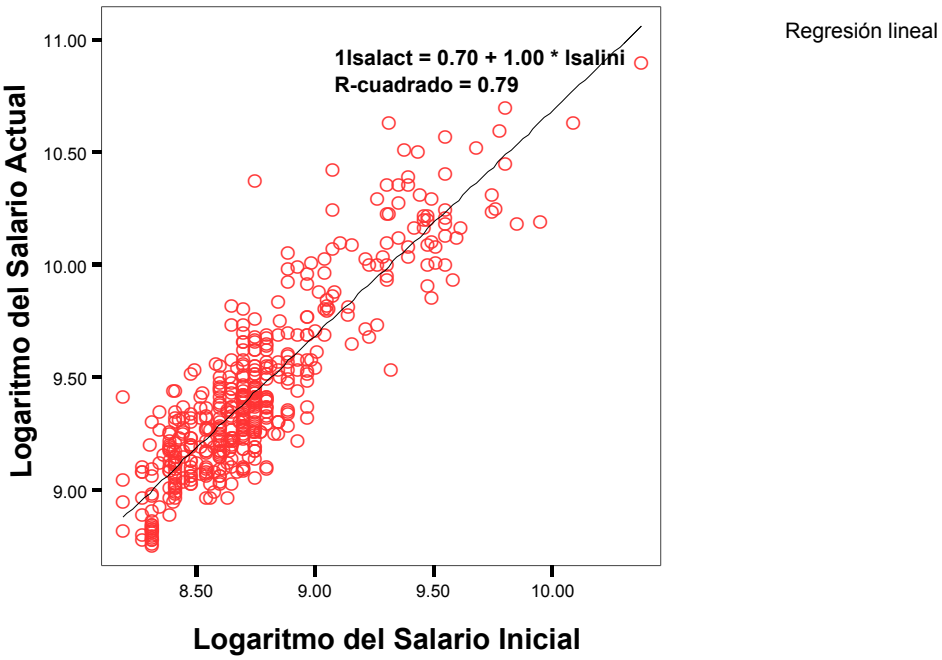

Figura 23: Diagrama de dispersión de Salarios

El gráfico muestra que existe una relación lineal directa entre dichas variables y que, por lo tanto, trabajadores con salario inicial elevado tienden a tener elevado salario

actual y viceversa. Dicha relación lineal es fuerte con un coeficiente de determinación del 79% y viene dada por la ecuación:

$$\text{Log(Salario Actual)} = 0.7 + \text{Log(Salario Inicial)}$$

Por lo tanto, los salarios han crecido, en media, un 100(exp(0.7)-1) = 101.37% respecto al salario inicial.

Ejemplo 8 (Relación entre Tasa de Mortalidad y Esperanza de Vida)

En la Figura 24 se muestra el diagrama de dispersión de la Tasa de Mortalidad Infantil (medida en número de muertos por cada mil nacimientos) y la Esperanza de Vida (en años) para una muestra de 49 países del mundo

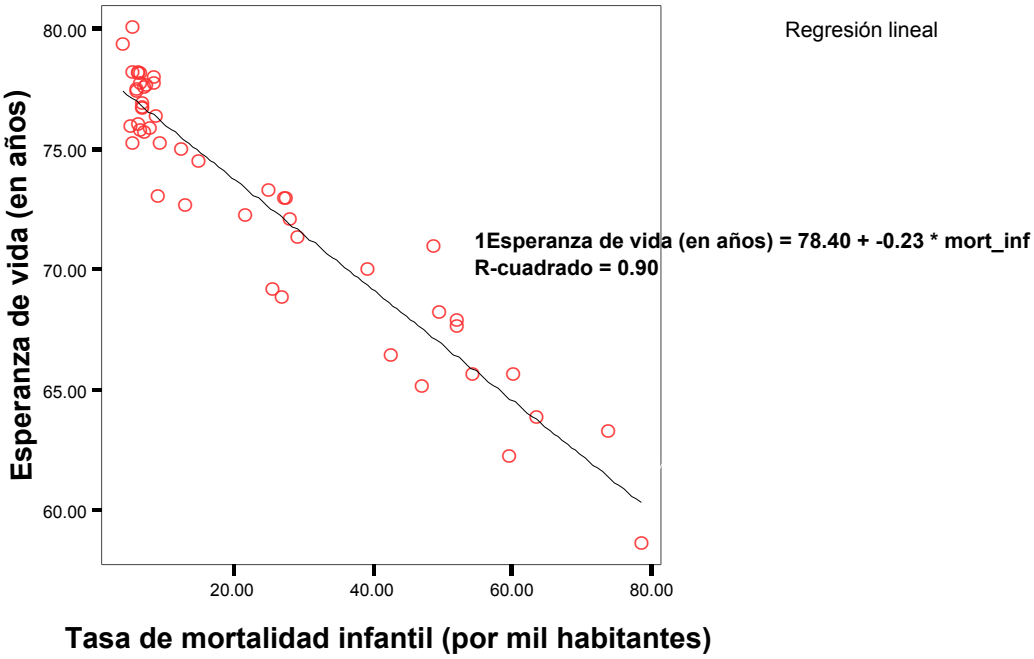

Figura 24: Diagrama de dispersión de la Tasa de Natalidad vs Esperanza de Vida

El gráfico muestra que existe una relación lineal inversa entre dichas variables y que, por lo tanto, los países con mayor mortalidad infantil tienen menor esperanza de vida, y al revés. Dicha relación es muy fuerte con un coeficiente de determinación del 90% y refleja, de forma implícita, la influencia del nivel de desarrollo de un país.

Ejemplo 9 (Relación entre Edad y Veteranía en el Trabajo)

En la Figura 25 se muestra el diagrama de dispersión de la Edad del empleado y la Veteranía en el puesto para un grupo de trabajadores. Se observa que no existe relación de ningún tipo y, en particular, lineal entre ambas variables.

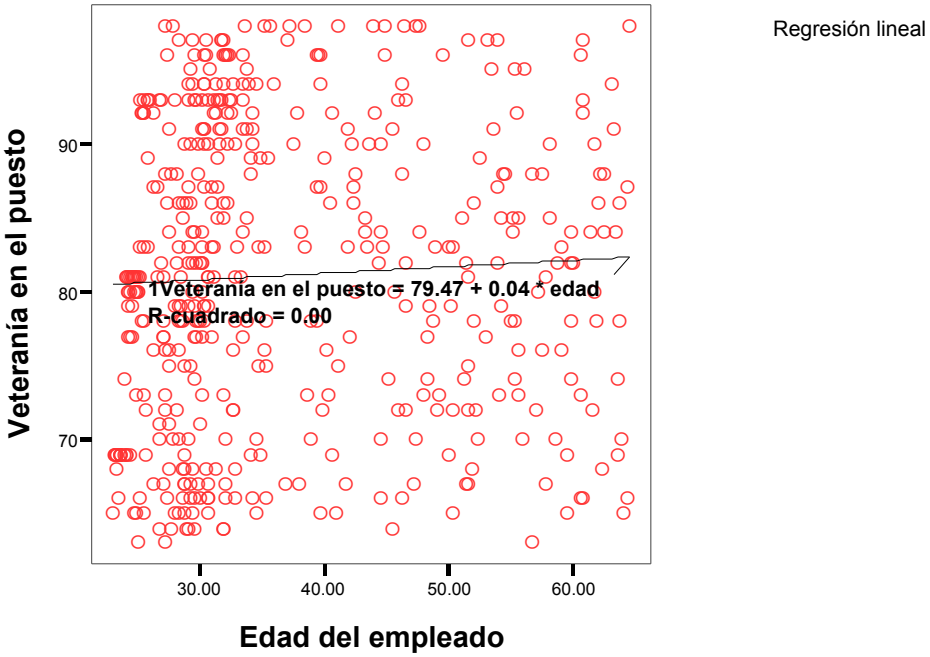

Figura 25: Diagrama de dispersión de la edad vs veteranía

Ejemplo 10 (Renta per Cápita y Esperanza de Vida)

La Figura 26 muestra el diagrama de dispersión de la Renta per Cápita vs la Esperanza de Vida para un grupo de países así como la estimación obtenida mediante un ajuste logarítmico.

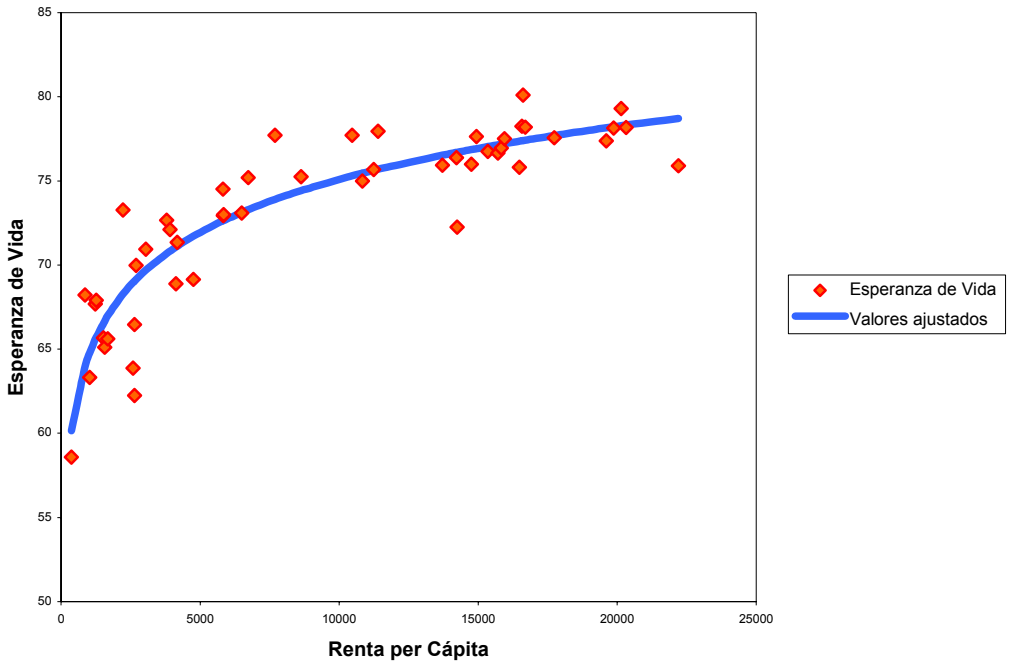

Figura 26: Diagrama de dispersión de la Renta per Cápita vs Esperanza de Vida

La relación estimada es creciente pero no lineal y viene dada por la expresión:

$$\text{Esperanza de Vida} = 33.24 + 4.54*\log(\text{Renta per Cápita})$$

siendo su coeficiente de determinación igual a 0.82. Se estima, por lo tanto, que un crecimiento de la renta per cápita de un 1% implicaría un aumento de 4.54 años en la esperanza de vida.

6.2.2. Diagramas de dispersión matriciales

Existen muchos tipos de gráficos de dispersión, pero un formato que se ajusta particularmente cuando se aplican técnicas multivariantes son los llamados ***diagramas de dispersión matriciales*** que permiten analizar, de forma simultánea, las relaciones existentes entre un grupo de variables cuantitativas. Consisten en representar los diagramas de dispersión para todas las combinaciones de las variables analizadas. Con p variables existen, por lo tanto, p(p-1)/2 gráficos posibles, que pueden disponerse en forma de matriz para entender el tipo de relación existente entre los distintos pares de variables. En

particular, estos gráficos son importantes para apreciar si existen relaciones no lineales, en cuyo caso la matriz de covarianzas puede no ser un buen resumen de la dependencia entre variables.

Ejemplo 11 (Análisis de variables demográfico-económicas)

En la Figura 27 se muestra la matriz de diagramas de dispersión correspondiente a un grupo de variables demográficas y económicas de una muestra de países. Superpuestas se muestran, además, las rectas de regresión estimadas. Se observa que, con la única excepción de la renta per cápita, las relaciones existentes entre las variables son lineales. El tipo de relación de la renta per cápita con el resto de las variables es, sin embargo, logarítmico, indicando, por lo tanto, la necesidad de considerar su logaritmo como variable objeto de estudio si se requiere la hipótesis de linealidad para todas las variables.

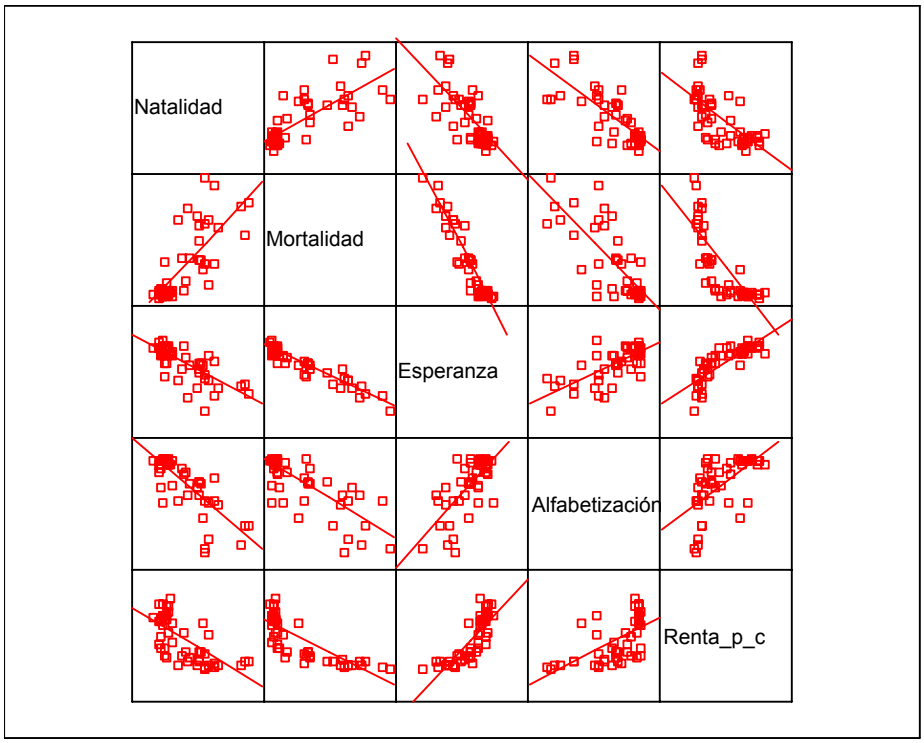

Figura 27: Matriz de diagramas de dispersión

6.3. Análisis de una variable cuantitativa y otra cualitativa.

Cuando se dispone de una variable cuantitativa y otra cualitativa, el estudio se enfoca como un problema de comparación del comportamiento de la variable numérica en las diferentes subpoblaciones que define la variable cualitativa. Ignorar la heterogeneidad debida a la presencia de subpoblaciones puede conducir a conclusiones equivocadas en el análisis.

Una forma de realizar dicho análisis es mediante los diagramas de cajas y los test de diferencias de medias, tal y como se muestra en el siguiente ejemplo.

Ejemplo 11 (Datos macroeconómicos)

En el Ejemplo 4 se analizó la renta per cápita de una muestra de países del mundo encontrándose que la distribución de dicha variable era multimodal. En la Figura 28 y la Tabla 18 se muestran los resultados de un estudio comparativo de dicha renta entre países pertenecientes y no pertenecientes a la OCDE. Se observa que existen diferencias significativas en la renta media de dichos grupos que tiende a situarse en torno a las modas encontradas en el Ejemplo 4 justificando el por qué de dicha multimodalidad.

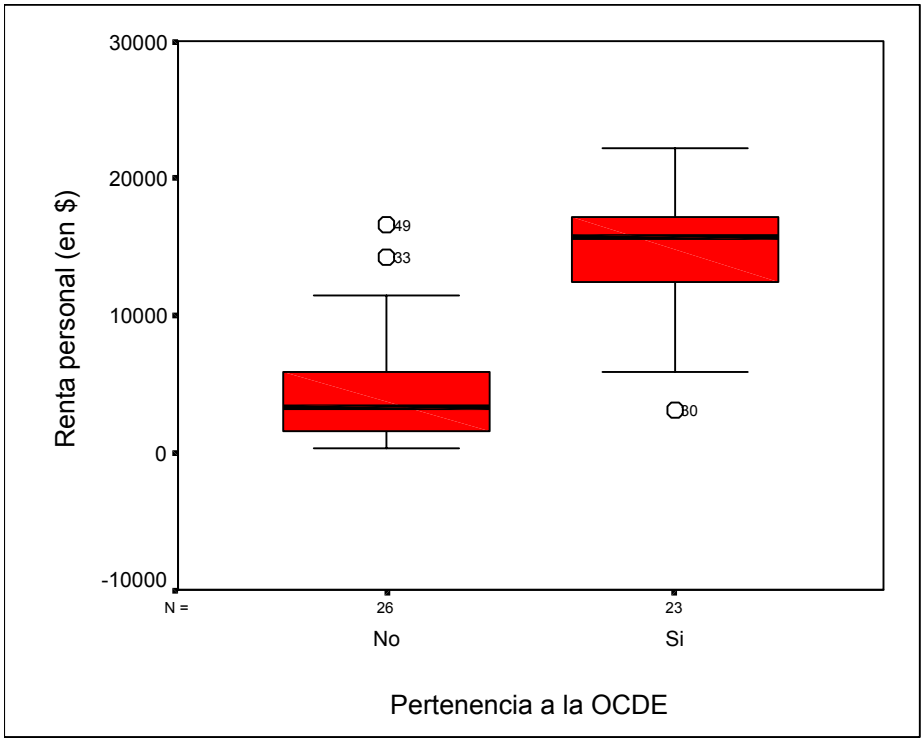

Figura 28: Comparación de Renta per Cápita entre Países Pertenecientes y No Pertenecientes a la OCDE

En este caso la renta per cápita media (9348.2$) no es un valor representativo de la distribución debido a las diferencias existentes entre estos dos grupos. Valores más representativos serían 4730.37 $ y 14568.36 $ que son las rentas per cápita medias de los dos grupos anteriormente citados.

Tabla 18
Comparación de Renta per Cápita
entre Países Pertenecientes y No Pertenecientes a la OCDE

Estadísticos de grupo

	Pertenencia a organizaciones:OCDE	N	Media	Desviación típ.	Error típ. de la media
Renta personal (en $)	No	26	4730.3656	4296.7840	842.6687
	Si	23	14568.36	5001.1037	1042.8022

Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
									95% Intervalo de confianza para la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	Inferior	Superior
Renta personal (en $	Se han asumido varianzas iguales	.436	.512	-7.407	47	.000	-9837.9912	1328.1472	-12509.9	-7166.10
	No se han asumido varianzas iguales			-7.338	43.711	.000	-9837.9912	1340.7188	-12540.5	-7135.44

6.3.1. Estudio de la homocedasticidad

La ***homocedasticidad*** es una hipótesis muy habitual en algunas técnicas estadísticas como el Análisis de la Varianza, el Análisis Discriminante y el Análisis de Regresión. Dicha hipótesis se refiere a suponer la igualdad de varianzas de las variables dependientes en diversos grupos formados por los distintos valores de las variables independientes. Si dicha hipótesis no se verifica puede alterar la potencia y el nivel de significación de los contrastes utilizados por dichas técnicas y de ahí el interés de analizar si se verifica o no y, en éste último caso, poner los remedios oportunos.

Para ello se utilizan contrastes de hipótesis cuya finalidad es analizar la existencia de esta igualdad que, en muchas ocasiones, va ligada a una falta de normalidad de las variables analizadas. En la literatura se han propuesto diversos tests (ver, por ejemplo, Jobson, 1991, Volumen 1). Uno de los más utilizados es el test de Levene basado en aplicar un ANOVA a las diferencias absolutas respecto a una medida de tendencia central de los diversos grupos. Dicho test toma como hipótesis nula la de homocedasticidad y como alternativa la de heterocedasticidad.

Un posible remedio contra la heterocedasticidad es transformar los datos originales. Un grupo de transformaciones muy utilizadas son las de Box-Cox que vienen dadas por la expresión

$$\begin{cases} \dfrac{(X+C)^{\lambda}-1}{\lambda} & \text{si } \lambda \neq 0 \\ \ln(X+C) & \text{si } \lambda = 0 \end{cases}$$

El valor de C se elige de forma que X + C sea positiva. El valor de λ se suele determinar de forma empírica.

En general este tipo de transformaciones suelen ser efectivas si no hay un número excesivo de outliers y si el cociente de la desviación típica dividida por la media es mayor que ¼ o si el cociente de la observación mas grande dividida por la más pequeña es mayor que 2. Es importante que la transformación elegida sea fácilmente interpretable y, en caso de duda, se aconseja repetir el análisis con los datos transformados y sin transformar y observar si los resultados obtenidos difieren demasiado. En éste último caso y si el procedimiento utilizado es poco robusto a la hipótesis de normalidad, utilizar los resultados con los datos transformados.

Ejemplo 12 (Datos macroeconómicos)

En la Figura 29 se muestra el diagrama de cajas correspondiente a las exportaciones de un grupo de países clasificados de acuerdo a un índice de estabilidad política. Así mismo en la Tabla 19 se presentan los resultados del test de Levene. En ambos casos es claramente rechazada la hipótesis de homocedasticidad.

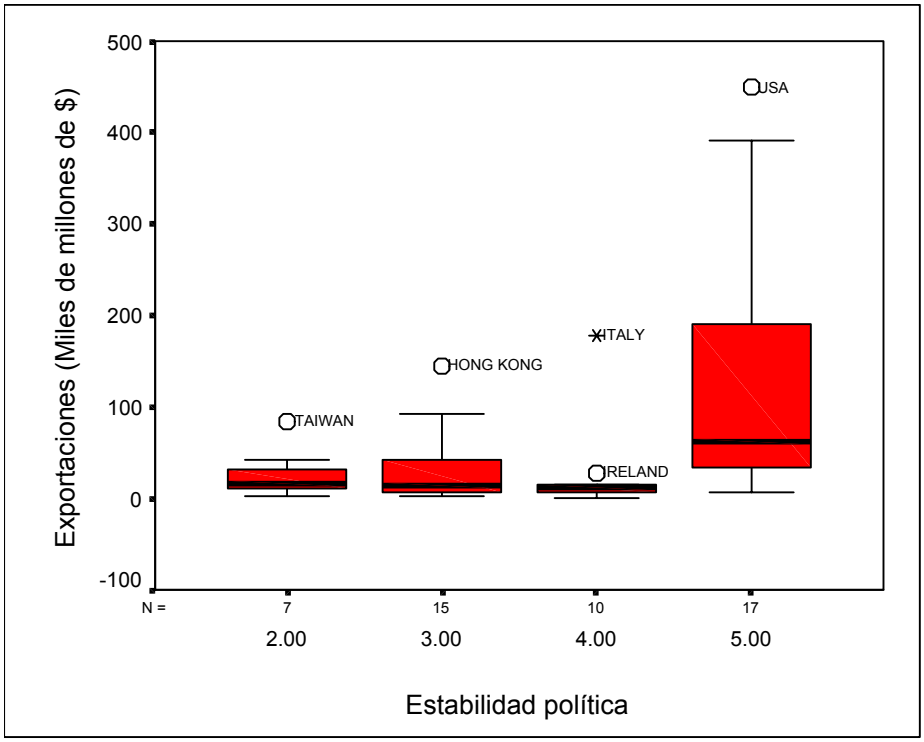

Figura 29: Diagrama de cajas de las Exportaciones de países clasificados por Estabilidad Política

Tabla 19

Resultados del test de Levene para las exportaciones

Prueba de homogeneidad de la varianza

		Estadístico de Levene	gl1	gl2	Sig.
Exportaciones (Miles de millones de $)	Basándose en la media	9.706	3	45	.000
	Basándose en la mediana.	3.869	3	45	.015
	Basándose en la mediana y con gl corregido	3.869	3	21.853	.023
	Basándose en la media recortada	8.773	3	45	.000

En la Tabla 20 y la Figura 30 se muestran los resultados obtenidos al analizar el logaritmo de las exportaciones. Se observa que dicha transformación (que ya corregía la falta de normalidad de la distribución como se ve en el Ejemplo 4) también resuelve la heterocedasticidad de la misma.

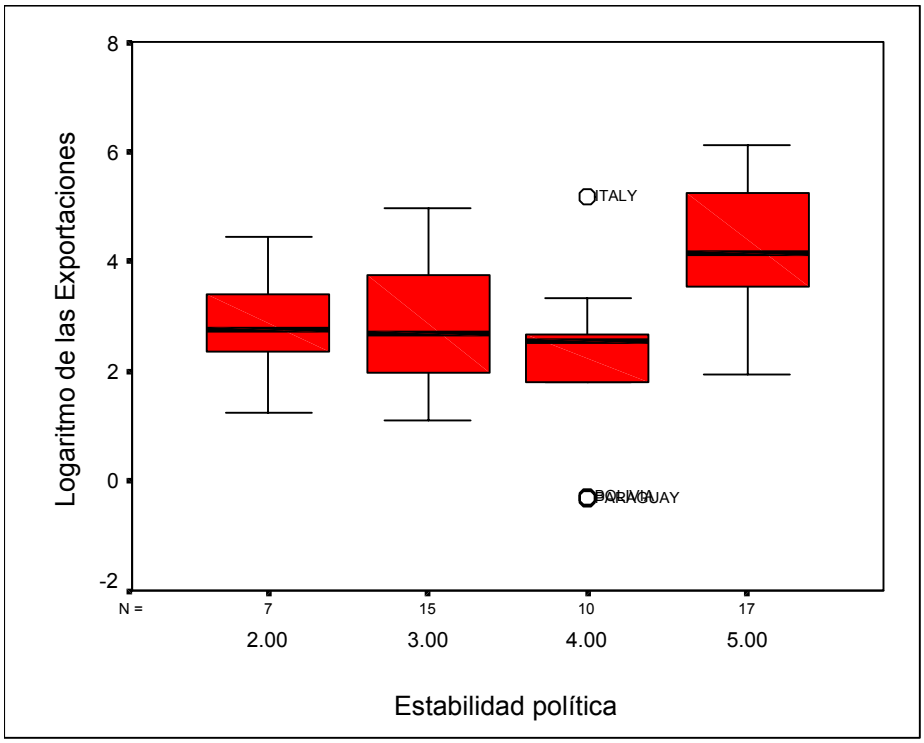

Figura 30: Diagrama de cajas del logaritmo de las Exportaciones de países clasificados por Estabilidad Política

Tabla 20
Resultados del test de Levene para el logaritmo de las Exportaciones

Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

Prueba de homogeneidad de la varianza

		Estadístico de Levene	gl1	gl2	Sig.
Logaritmo de las Exportaciones	Basándose en la media	.282	3	45	.838
	Basándose en la mediana.	.174	3	45	.914
	Basándose en la mediana y con gl corregido	.174	3	31.060	.913
	Basándose en la media recortada	.303	3	45	.823

Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

7.- DATOS ATÍPICOS (OUTLIERS)

Los casos atípicos son observaciones con características diferentes de las demás. Este tipo de casos no pueden ser caracterizados categóricamente como benéficos o problemáticos sino que deben ser contemplados en el contexto del análisis y debe evaluarse el tipo de información que pueden proporcionar. Su principal problema radica en que son elementos que pueden no ser representativos de la población pudiendo distorsionar seriamente el comportamiento de los contrastes estadísticos. Por otra parte, aunque diferentes a la mayor parte de la muestra, pueden ser indicativos de las características de un segmento válido de la población y, por consiguiente, una señal de la falta de representatividad de la muestra.

7.1 Tipos de outliers

Los casos atípicos pueden clasificarse en 4 categorías.

La primera categoría contiene aquellos casos atípicos que surgen de un error de procedimiento, tales como la entrada de datos o un error de codificación. Estos casos atípicos deberían subsanarse en el filtrado de los datos, y si no se puede, deberían eliminarse del análisis o recodificarse como datos ausentes.

La segunda clase es la observación que ocurre como consecuencia de un acontecimiento extraordinario. En este caso, el outlier no representa ningún segmento válido de la población y puede ser eliminado del análisis.

La tercera clase contiene las observaciones cuyos valores caen dentro del rango de las variables observadas pero que son únicas en la combinación de los valores de dichas variables. Estas observaciones deberían ser retenidas en el análisis pero estudiando qué influencia ejercen en los procesos de estimación de los modelos considerados.

La cuarta y última clase comprende las observaciones extraordinarias para las que el investigador no tiene explicación. En estos casos lo mejor que se puede hacer es replicar el análisis con y sin dichas observaciones con el fin de analizar su influencia sobre los resultados. Si dichas observaciones son influyentes el analista debería reportarlo en sus conclusiones y debería averiguar el por qué de dichas observaciones.

7.2 Identificación de outliers

Los casos atípicos pueden identificarse desde una perspectiva univariante o multivariante.

La *perspectiva univariante* examina la distribución de observaciones para cada variable, seleccionando como casos atípicos aquellos casos cuyos valores caigan fuera de los rangos de la distribución. La cuestión principal consiste en el establecimiento de un umbral para la designación de caso atípico. Esto se puede hacer gráficamente mediante histogramas o diagramas de caja o bien numéricamente, mediante el cálculo de puntuaciones tipificadas. Para muestras pequeñas (de 80 o incluso menos observaciones), las pautas sugeridas identifican como atípicos aquellos casos con valores estándar de 2.5 o superiores. Cuando los tamaños muestrales son mayores, las pautas sugieren que el valor umbral sea 3.

Ejemplo 13 (Datos macroeconómicos)

En la Figura 31 se muestra de nuevo el diagrama de cajas de la Tasa Real de Crecimiento del PIB para una muestra de países del mundo considerada en el Ejemplo 4 obtenido mediante el paquete estadístico SPSS 10.0. Dicho paquete distingue dos tipos de atípicos: débiles y extremos. Un atípico débil (resp. extremo) es aquél que dista del cuartil más cercano más de 1.5 (resp. 3) veces el recorrido intercuartílico. Los atípicos débiles se marcan con o y los extremos con *. En la Figura 31 se observa la existencia de 3 atípicos débiles: dos con una elevada tasa de crecimiento (Kuwait y China) y uno con una baja tasa (Rusia). Dichos atípicos son los responsables de la elevada curtosis de la variable y, por lo tanto, de su falta de normalidad (ver Tabla 9).

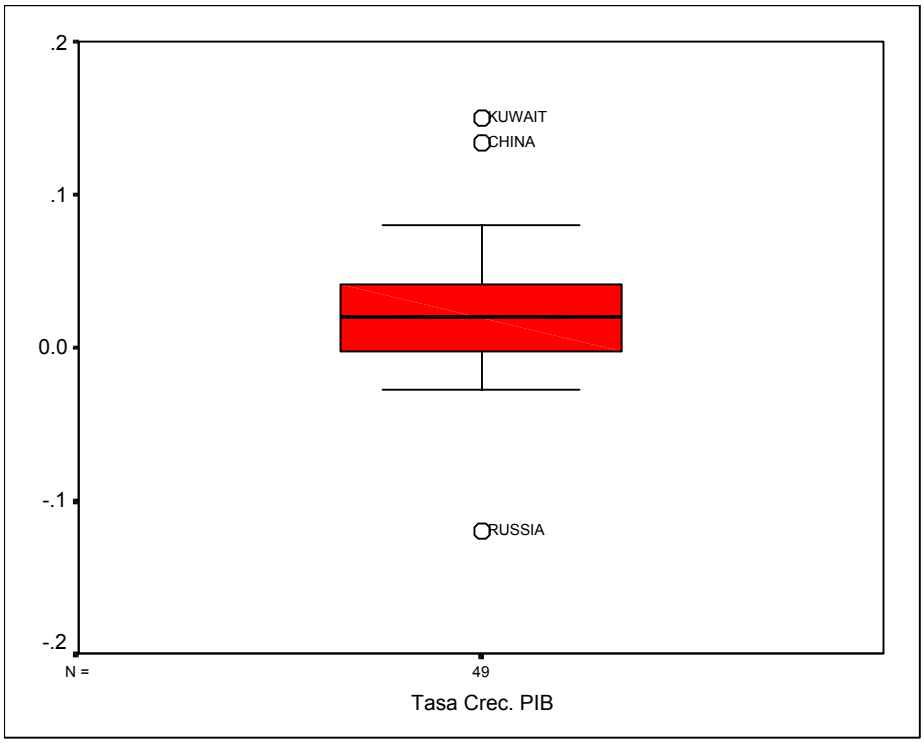

Figura 31: Diagrama de cajas del logaritmo de las Exportaciones de países clasificados por Estabilidad Política

En la Tabla 21 y la Figura 32 se muestran los resultados obtenidos al analizar la normalidad de esta variable una vez eliminados los 3 atípicos anteriores. Los contrastes de asimetría y curtosis toman los valores:

$$z_{asimetria} = \frac{0.353}{0.350} = 1.01,\ z_{curtosis} = -\frac{0.698}{0.688} = -1.014$$

cuyos p-valores son 0.157 y 0.155, respectivamente. Se observa que la falta de normalidad de esta variable se debía a la presencia de los 3 atípicos. Una vez eliminados se resuelve el problema y la variable se puede considerar normal.

Tabla 21
Análisis de la normalidad de la Tasa de Crecimiento del PIB

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Tasa real de crecimiento del PIB	.101	46	.200*	.958	46	.196

*. Este es un límite inferior de la significación verdadera.

a. Corrección de la significación de Lilliefors

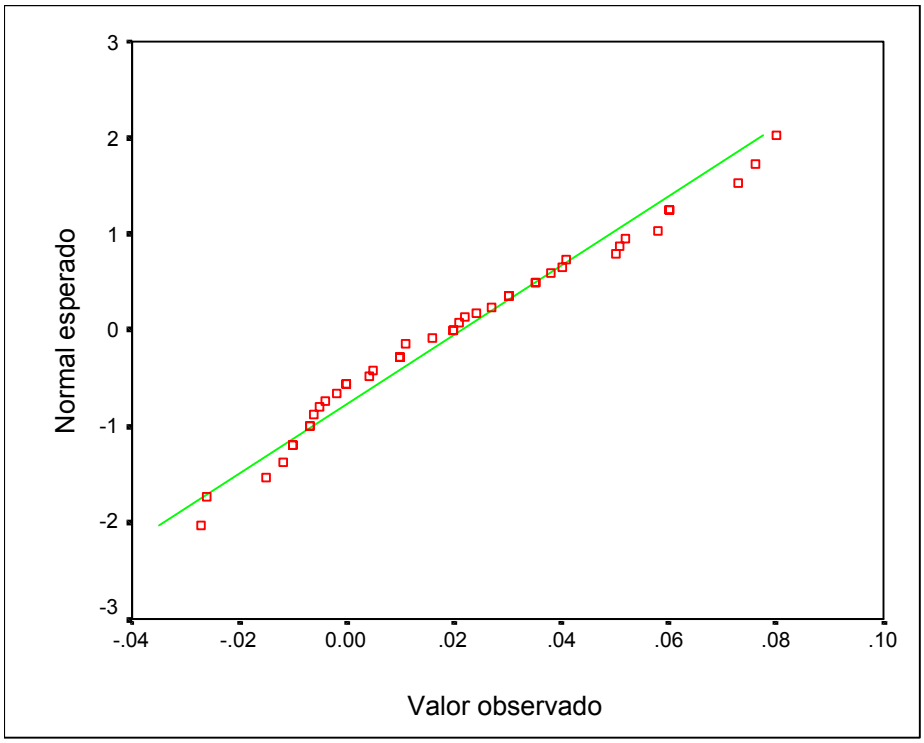

Figura 32: Diagrama de cuantiles de la Tasa de Crecimiento del PIB

Además de la evaluación univariante, pueden analizarse conjuntamente pares de variables mediante un gráfico de dispersión. Casos que caigan manifiestamente fuera del rango del resto de las observaciones pueden identificarse como puntos aislados en el gráfico de dispersión. Para ayudar a determinar el rango esperado de las observaciones, se puede superponer sobre el gráfico de dispersión una elipse que represente un intervalo de confianza especificado para una distribución normal bivariante (ver Figura 33). Esto proporciona una representación gráfica de los límites de confianza y facilita la identificación de casos atípicos.

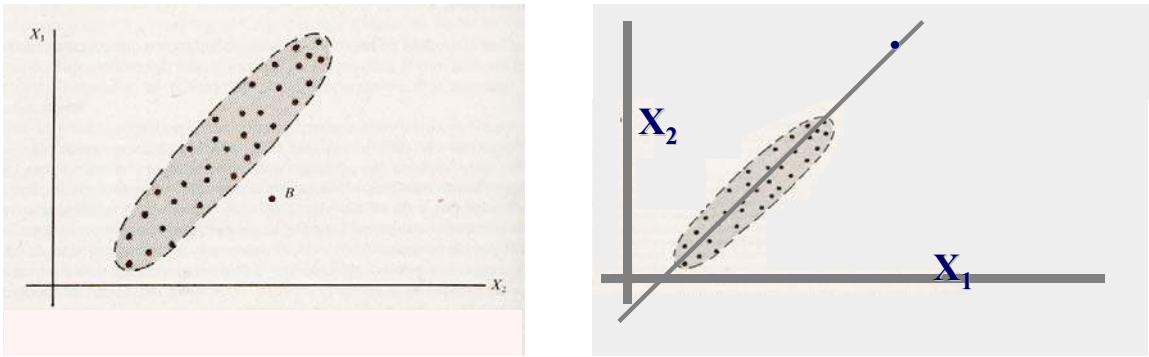

Figura 33: Detección bivariante de atípicos

Finalmente existen procedimientos para detectar atípicos multivariantes. Dicha detección se puede hacer mediante un Análisis de Componentes Principales (PONER

ENLACE A LA FUTURA PAGINA WEB DE COMPONENTES PRINCIPALES) o con el cálculo de las distancias de Mahalanobis al centroide de la distribución.

Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

8.- DATOS AUSENTES (MISSING)

Los datos ausentes son algo habitual en el Análisis Multivariante; de hecho, rara es la investigación en la que no aparece este tipo de datos.

En estos casos la ocupación primaria del investigador debe ser determinar las razones que subyacen en el dato ausente buscando entender el proceso principal de esta ausencia para seleccionar el curso de acción más apropiado.

Para ello se debe determinar cuál es el proceso de datos ausentes, entendido como cualquier evento sistemático externo al encuestado (errores en la introducción de datos) o acción por parte del encuestado (tales como rehusar a contestar) que da lugar a la ausencia de datos. En particular, el investigador debe analizar si existe algún patrón no aleatorio en dicho proceso que pueda sesgar los resultados obtenidos debido a la pérdida de representatividad de la muestra analizada.

8.1 Tipos de valores ausentes

Se distinguen las dos situaciones siguientes:

1) **Datos ausentes prescindibles:** son resultado de procesos que se encuentran bajo el control del investigador y pueden ser identificados explícitamente. En estos casos no se necesitan soluciones específicas para la ausencia de datos dado que dicha ausencia es inherente a la técnica usada.

   Ejemplos de estas situaciones son aquellas observaciones de una población que no están incluidas en la muestra o los llamados **datos censurados** que son observaciones incompletas como consecuencia del proceso de obtención de datos seguido en el análisis.

2) **Datos ausentes no prescindibles:** son resultado de procesos que no se encuentran bajo el control del investigador y/o no pueden ser identificados explícitamente.

   Ejemplos de estas situaciones son los errores en la entrada de datos, la renuncia del encuestado a responder a ciertas cuestiones o respuestas inaplicables.

   En estos casos se debe analizar si existen o no patrones sistemáticos en el proceso que puedan sesgar los resultados obtenidos.

Si los datos ausentes son no prescindibles conviene, por lo tanto, analizar el grado de aleatoriedad presente en los mismos. Según este grado el proceso de datos ausentes se puede clasificar del siguiente modo:

1) **Datos ausentes completamente aleatorios (MCAR):** este es el mayor grado de aleatoriedad y se da cuando los datos ausentes son una muestra aleatoria simple de la muestra sin un proceso subyacente que tiende a sesgar los datos observados. En este caso se podría solucionar el problema sin tener cuenta el impacto de otras variables

3) **Datos ausentes aleatorios (MAR)**: en este caso el patrón de los datos ausentes en una variable Y no es aleatorio sino que depende de otras variables de la muestra X. Ahora bien, para cada valor de X, los valores observados de Y sí representan una muestra aleatoria de Y.

Así, por ejemplo, si X es el sexo del encuestado e Y es su renta, un proceso MAR se tendría si existen más valores ausentes de Y en hombres que en mujeres y, sin embargo, los datos son aleatorios para ambos sexos en el sentido de que, tanto en los hombres como en las mujeres, el patrón de ausentes es completamente aleatorio. Si, además, tampoco existen diferencias por sexos los datos ausentes serían MCAR.

Si los datos ausentes son MAR cualquier solución al problema deberá tener en cuenta los valores de X dado que afectan al proceso generador de datos ausentes.

3) Datos ausentes no aleatorios: en este caso existen patrones sistemáticos en el proceso de datos ausentes y habría que evaluar la magnitud del problema calibrando, en particular, el tamaño de los sesgos introducidos por dichos patrones. Si éstos son grandes habría que atacar el problema directamente intentando averiguar cuáles son dichos valores.

8.2 Localización de datos ausentes

El primer paso en el tratamiento de datos ausentes consiste en evaluar la magnitud del problema. Para ello se comienza analizando el porcentaje de datos ausentes por variables y por casos.

Si existen casos con un alto porcentaje de datos ausentes se deberían excluir del problema. Así mismo si existe una variable con un alto porcentaje de este tipo de casos su exclusión dependerá de la importancia teórica de la misma y la posibilidad de ser reemplazada por variables con un contenido informativo similar.

Como regla general, sin embargo, si dicha variable es dependiente debería ser eliminada ya que cualquier proceso de imputación de valores puede distorsionar la significación estadística y práctica de los modelos estimados para ella.

8.3 Diagnóstico de la aleatoriedad en el proceso de datos ausentes

Existen 3 métodos:

a) Para cada variable Y formar dos grupos (observaciones ausentes y presentes en Y) y aplicar contrastes de comparación de dos muestras para determinar si existen diferencias significativas entre los dos grupos sobre otras variables de interés. Si se encuentran diferencias significativas el proceso de datos ausentes no es aleatorio.

b) Utilizar correlaciones dicotomizadas para evaluar la correlación de los datos ausentes en cualquier par de valores. Estas correlaciones indicarían el grado de asociación entre los valores perdidos sobre cada par de variables. Bajas correlaciones implican aleatoriedad en el par de variables y que los datos ausentes pueden clasificarse como MCAR. En caso contrario son MAR.

c) Realizar contrastes conjuntos de aleatoriedad que determinen si los datos ausentes pueden ser clasificados como MCAR. Estos contrastes analizan el patrón de datos ausentes sobre todas las variables y las compara con el patrón esperado para un proceso de datos ausentes aleatorio. Si no se encuentran diferencias significativas el proceso puede clasificarse como MCAR; en caso contrario deben utilizarse los procedimientos a) y b) anteriores para identificar los procesos específicos de datos ausentes que no son aleatorios.

8.4 Aproximaciones al tratamiento de datos ausentes

Si se encuentran procesos de datos ausentes MAR o no aleatorios, el investigador debería aplicar sólo el método diseñado específicamente para este proceso. Sólo si el investigador determina que el proceso de ausencia de datos puede clasificarse como MCAR pueden utilizarse las siguientes aproximaciones:

a) Utilizar sólo los casos completos: conveniente si el tamaño muestral no se reduce demasiado

b) Supresión de casos y/o variables con una alta proporción de datos ausentes. Esta supresión deberá basarse en consideraciones teóricas y empíricas. En particular, si algún caso tiene un dato ausente en una variable dependiente, habitualmente excluirlo puesto que cualquier proceso de imputación puede distorsionar los modelos estimados. Así mismo una variable independiente con muchos datos ausentes podrá eliminarse si existen otras variables muy similares con datos observados.

c) Imputar valores a los datos ausentes utilizando valores válidos de otras variables y/o casos de la muestra

8.3.1 Métodos de imputación

Los métodos de imputación pueden ser de tres tipos:

1) **Métodos de disponibilidad completa** que utilizan toda la información disponible a partir de un subconjunto de casos para generalizar sobre la muestra entera. Se utilizan habitualmente para estimar medias, varianzas y correlaciones
2) **Métodos de sustitución** que estiman valores de reemplazo para los datos ausentes, sobre la base de otra información existente en la muestra. Así se podría sustituir observaciones con datos ausentes por observaciones no maestrales o sustituir dichos datos por la media de los valores observados o mediante regresión sobre otras variables muy relacionadas con aquella a la que le faltan observaciones
3) **Métodos basados en modelos** que construyen explícitamente el mecanismo por el que se producen los datos ausentes y lo estiman por máxima verosimilitud. Entran en esta categoría el algoritmo EM o los procesos de aumento de datos.

Ejemplo 14 (Análisis de costes marginales financieros)

Para ilustrar el tratamiento de datos ausentes consideraremos datos pertenecientes a una muestra de 1628 empresas españolas sobre la que se ha obtenido información acerca de sus costes marginales en su deuda bancaria a largo (CMDBL) y a corto plazo (CMDBC) así como los correspondientes a otras deudas (CMREST) y algunas características adicionales como su edad (EDAD), sector (SECTOR), forma jurídica (FORJUR), tamaño (NTRAB) y si produce productos estandarizados (PROEST). En la Tabla 22 se muestran las estadísticas correspondientes a cada variable en cuanto al número de datos ausentes. Se observa que los mayores problemas corresponden a las variables PROEST (5.1%) y CMDBC (7.8%) no teniendo el resto de las variables graves problemas por este aspecto.

Tabla 22
Estadísticas de datos ausente por variables

	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
SECTOR	1628	100.0%	0	.0%	1628	100.0%
EDAD	1608	98.8%	20	1.2%	1628	100.0%
FORJUR	1628	100.0%	0	.0%	1628	100.0%
PROEST	1545	94.9%	83	5.1%	1628	100.0%
NTRAB	1628	100.0%	0	.0%	1628	100.0%
CMDBL	1614	99.1%	14	.9%	1628	100.0%
CMREST	1593	97.9%	35	2.1%	1628	100.0%
CMDBC	1501	92.2%	127	7.8%	1628	100.0%

Así mismo, en la Tabla 23 se muestra el número de datos ausentes por caso. El número máximo de datos ausentes es 3 y tan sólo 4 empresas alcanzan este número. Conviene notar, sin embargo, que a un 14.3% de los casos les falta al menos un dato.

Tabla 23
Estadísticas de datos ausente por caso

		Frecuencia	Porcentaje
Válidos	0	1396	85.7
	1	189	11.6
	2	39	2.4
	3	4	.2
	Total	1628	100.0

En la Tabla 24 se muestra las diversas pautas de datos ausentes tabuladas. Se observa que las más frecuentes son las correspondientes a la ausencia de los costes marginales a corto plazo (6.33%) y las de productos estandarizados (3.5%) sin que, en ningún caso el problema sea excesivamente grave.

Tabla 24
Pautas de datos ausentes tabuladas

EDAD	PROEST	CMDBL	CMDBC	CMREST	FRECUENCIA	PORCENTAJE
					1396	85.75
				X	22	1.35
			X		103	6.33
				X	11	0.68
		X			7	0.43
				X	1	0.06
			X		5	0.31
	X				57	3.50
				X	1	0.06
			X		4	0.25
		X			1	0.06
X	X				16	0.98

			X		4	0.25

En la Tabla 25 se analiza el grado de aleatoriedad de los datos ausentes correspondientes a las variables PROEST y CMDBC utilizando contrastes de comparaciones de dos muestras. Más concretamente se ha utilizado el test de independencia de la χ^2 para tablas de contingencia en el caso de las variables cualitativas (SECTOR, FORJUR, PROEST) indicando si dicha hipótesis es aceptada (celda en blanco), rechazada al 5% (*) y al 1% (**). Para el resto de las variables se ha realizado un test de la t de Student para comparación de medias en dos muestras independientes indicando si la diferencia de medias de los datos ausentes es superior (+) o inferior (-) a la de los datos no ausentes y si el resultado es significativo al 5% (+ ó -) o al 1% (++ ó --).

Se observa que, tanto en el caso de la variable PROEST como en el de la variable CMDBC los patrones de datos ausentes no son aleatorios.

En el caso de la variable PROEST ambos grupos difieren significativamente en cuanto al Sector, Edad y Costes Marginales Bancarios a Largo Plazo y casi significativamente en la Forma Jurídica y en los Costes a Corto Plazo. Se observa, en particular, una tendencia en las empresas con datos ausentes en PROEST a ser más jóvenes y soportar menores costes marginales bancarios tanto a corto como a largo plazo. Además tienden a pertenecer a sectores con gran soporte tecnológico y a tener un mayor número de sociedades limitadas.

Tabla 25
Evaluación de la aleatoriedad de los datos ausentes a través de test de comparaciones de dos muestras

Grupos formados por datos ausentes sobre:	PROEST	CMDBC
SECTOR	**	
EDAD	--	++
FORJUR	*	
PROEST		**
NTRAB		++
CMDBL	--	--
CMDBC	-	
CMREST		

En el caso de la variable CMDBC ambos grupos difieren significativamente en cuanto a la Edad, Tamaño, Costes Marginales Bancarios a Largo Plazo y la fabricación de productos estandarizados. Se observa, en particular, una tendencia en las empresas con

datos ausentes en CMDBC a ser más viejas, grandes y soportar menores costes marginales bancarios a largo plazo. Además las empresas que producen productos estandarizados tienden a tener un mayor número datos missing que las que no.

Por lo tanto los procesos de datos ausentes de estas dos variables son no aleatorios aunque, afortunadamente, son un porcentaje muy bajo del total (ver Tabla 22) por lo que el problema no es tan grave aunque debería intentar solucionarse utilizando alguno de los procedimientos descritos anteriormente y, en todo caso, hacerse constar en el informe final del análisis.

Finalmente, en la Tabla 26 se muestran las correlaciones entre las variables indicadoras de datos ausentes para cada una de las variables de la Tabla 22 en las que existe este problema. No se observa ninguna correlación especialmente fuerte (superior, en valor absoluto a 0.5). La más significativa es la correspondiente a las variables indicadoras de PROEST y la EDAD observándose una cierta tendencia a no contestar a ambas variables. Este patrón no es muy importante, sin embargo, puesto que, tal y como se enseña en la Tabla 24, solamente un 0.98% de las empresas del análisis muestra este patrón.

Tabla 26
Evaluación de la aleatoriedad de los datos ausentes a través de las correlaciones de una variable dicotomizada

Correlaciones

		MEDAD	MPROEST	MCDBL	MCDBC	MCREST
MEDAD	Correlación de Pearson	1	.481**	-.010	.051*	-.017
	Sig. (bilateral)	.	.000	.675	.041	.505
	N	1628	1628	1628	1628	1628
MPROEST	Correlación de Pearson	.481**	1	.009	.016	-.015
	Sig. (bilateral)	.000	.	.727	.522	.543
	N	1628	1628	1628	1628	1628
MCDBL	Correlación de Pearson	-.010	.009	1	.097**	.032
	Sig. (bilateral)	.675	.727	.	.000	.196
	N	1628	1628	1628	1628	1628
MCDBC	Correlación de Pearson	.051*	.016	.097**	1	.131**
	Sig. (bilateral)	.041	.522	.000	.	.000
	N	1628	1628	1628	1628	1628
MCREST	Correlación de Pearson	-.017	-.015	.032	.131**	1
	Sig. (bilateral)	.505	.543	.196	.000	.
	N	1628	1628	1628	1628	1628

**. La correlación es significativa al nivel 0,01 (bilateral).

*. La correlación es significante al nivel 0,05 (bilateral).

Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

Resumen

El Análisis Exploratorio de Datos (AED) es un conjunto de técnicas estadísticas uni y multivariantes cuya finalidad es examinar los datos previamente a la aplicación de cualquier técnica estadística. De esta forma el analista consigue un entendimiento básico de sus datos y de las relaciones existentes entre las variables analizadas.

El AED proporciona métodos sencillos para organizar y preparar los datos, detectar fallos en el diseño y recogida de datos, el tratamiento y evaluación de datos ausentes, la identificación de casos atípicos y comprobación de los supuestos subyacentes en la mayor parte de las técnicas multivariantes (linealidad, normalidad, homocedasticidad).

En esta lección se han mostrado los pasos a seguir para llevarlo a cabo ilustrando su aplicación mediante ejemplos sacados de problemas reales analizados por los autores.

Conviene hacer notar, finalmente, la importancia de estas técnicas y la necesidad de "perder el tiempo" en aplicarlas. Nuestra experiencia es que un A.E.D. hecho en profundidad muestra mucha información acerca de los datos objeto de análisis y que, en muchas ocasiones, la aplicación de técnicas estadísticas más sofisticadas del Análisis Multivariante no hace más que confirmar impresiones iniciales obtenidas a partir de un A.E.D.

Salvador Figueras, M y Gargallo, P. (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

Bibliografía

No existe un número excesivo de libros dedicados exclusivamente al tópico de A.E.D. En español tenemos noticias de los siguientes:

ESCOBAR, M. (2000) *Análisis Gráfico/Exploratorio.* Cuadernos de Estadística. Editorial La Muralla.
RIAL, A.; VARELA, J. y ROJAS, A. (2001). *Depuración y Análisis Preliminares de Datos en SPSS.* Sistemas Informatizados para la Investigación del Comportamiento. RA-MA.

ambos muy orientado al paquete estadístico SPSS 10.0

En inglés un libro clave es el siguiente:

TUKEY, J.W. (1977). *Exploratory Data Analysis.* Addison-Wesley

que dio, históricamente, un impulso muy importante a esta parte tradicionalmente despreciada del análisis estadístico aplicado.

El siguiente libro contiene un buen capítulo dedicado al A.E.D. y es el que hemos tomado como patrón a la hora de diseñar la página.

HAIR, J., ANDERSON, R., TATHAM, R. y BLACK, W. (1999). *Análisis Multivariante.* 5ª Edición. Prentice Hall.

Si estáis interesados en el tratamiento de datos ausentes y queréis profundizar en el tema os recomendamos la lectura de los dos libros siguientes:

LITTLE, R.J.A. and RUBIN, D. (1987) *Statistical Analysis with Missing Data.* New York. Wiley.
SCHAFER, J.L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman & Hall.

Un tópico que ha adquirido fuerza últimamente es el diseño de procedimientos exploratorios en grandes bases de datos. Son algoritmos y técnicas estadístico-informáticas que buscan la extracción de patrones de comportamiento y de conocimiento en conjuntos de datos muy grandes. Dichas técnicas se conocen bajo el nombre de *Data Mining.* Si queréis haceros una idea de en qué consisten y cómo funcionan, un buen libro introductorio es

Berry, M. and Linoff, G. (1997). *Data Mining Techniques for Marketing, Sales and Customer Support.* John Wiley & Sons, Inc, New York.

Finalmente, otro libro al que se ha hecho referencia en el apartado dedicado a homocedasticidad es

Salvador Figueras, M y **Gargallo, P.** (2003): **"Análisis Exploratorio de Datos"**, [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/aed> [y añadir fecha consulta].

JOBSON, J.D. (1992) *Applied Multivariate Data Analysis. Volume I: Regression and Experimental Design*. Springer-Verlag.