CURSO BÁSICO DE ANÁLISIS ESTADÍSTICO EN SPSS.

FRANCISCO PARRA RODRÍGUEZ
JUAN ANTONIO VICENTE VÍRSEDA
MAURICIO BELTRÁN PASCUAL

EL PROGRAMA ESTADÍSTICO SPSS

1. EL PROGRAMA ESTADÍSTICO SPSS

1.1 INTRODUCCIÓN

El programa informático SPSS es en la actualidad el más extendido entre los investigadores y profesionales, no sólo en el campo de las ciencias sociales, sino también en las humanas, biomédicas, en la economía, y, en general, en múltiples campos de actividad en el que se precise realizar un tratamiento estadístico de la información.

La gran acogida dispensada al programa SPSS es debido a su flexibilidad y facilidad de manejo. El SPSS incluye una amplia y variada gama de análisis estadísticos y de gestión de datos en un entorno gráfico.

El programa se maneja a través de menús descriptivos y cuadros de diálogo, pero también se pueden realizar todas las tareas a través de un lenguaje de comandos (programación). Señalar que algunas de las opciones disponibles sólo son accesibles a través de dicho lenguaje.

El paquete estadístico se puede adquirir e instalar de forma modular. Los módulos disponibles son: **Base, Técnicas estadísticas Profesionales, Técnicas Estadísticas Avanzadas, Tablas, Tendencias, Categorías, CHAID, Pruebas exactas, Redes Neuronales, Mapinfo y AllCLEAR III.**

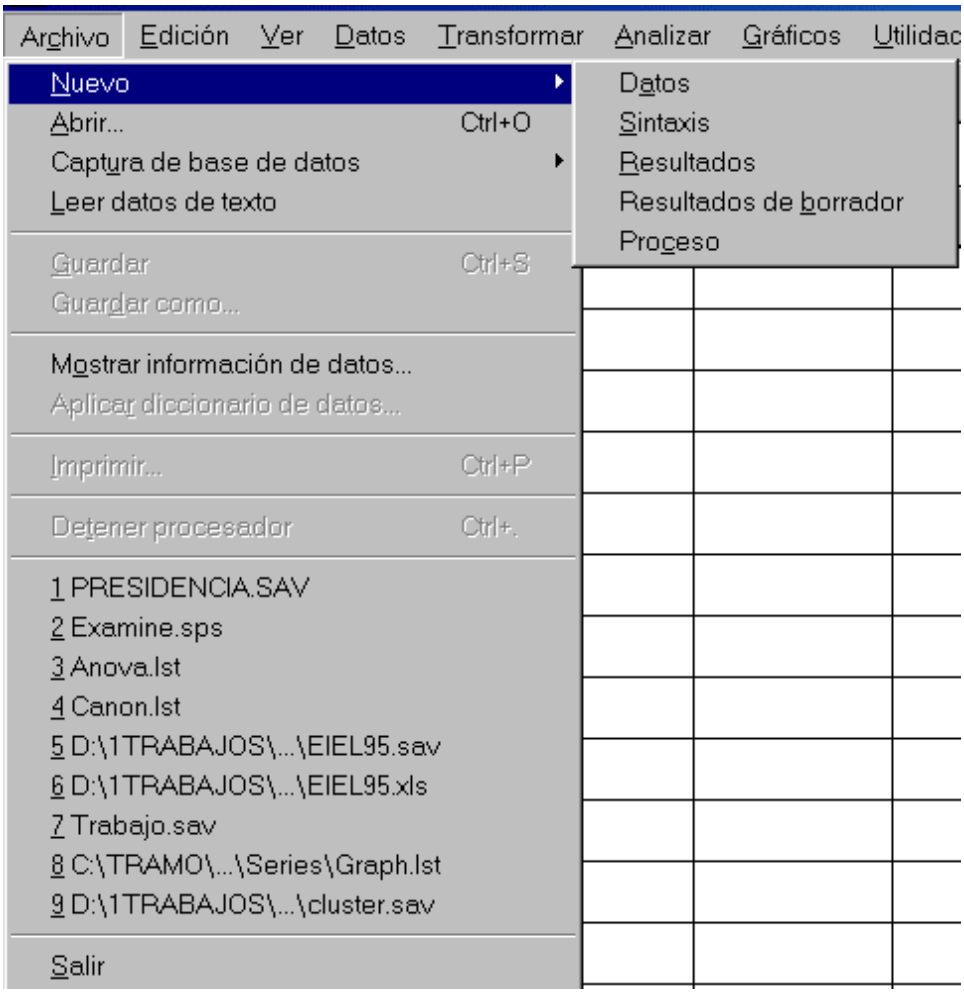
A continuación se presentan de forma gráfica las principales opciones del programa a través del sistema de menús descriptivos y cuadros de dialogo. A lo largo del desarrollo del curso ofreceremos una visión global de las posibilidades del programa y comentaremos los contenidos de los diferentes menús descriptivos.

1.2 LOS MENÚS EN SPSS

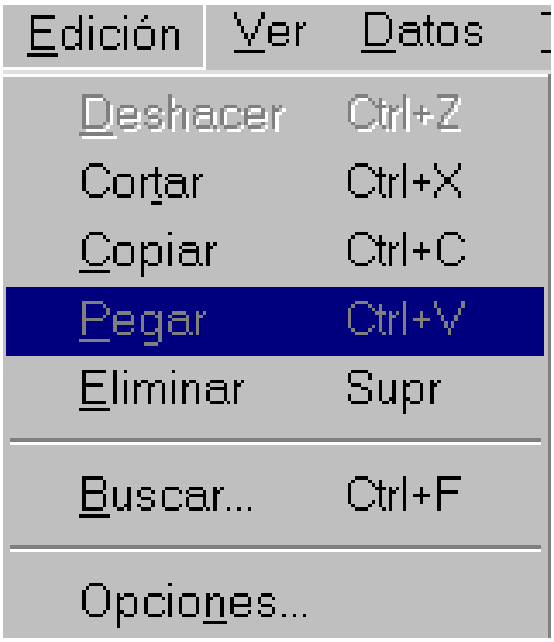
MENÚ GENERAL DEL PROGRAMA

En la barra de menús se encuentran 10 opciones con el siguiente contenido:

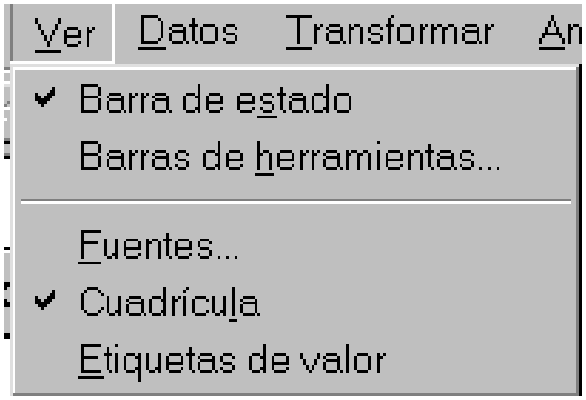
ARCHIVO A través de este menú se realizan las operaciones de abrir, crear o grabar ficheros, que pueden ser de datos, instrucciones, resultados o procesos. También se controlan las tares de impresión



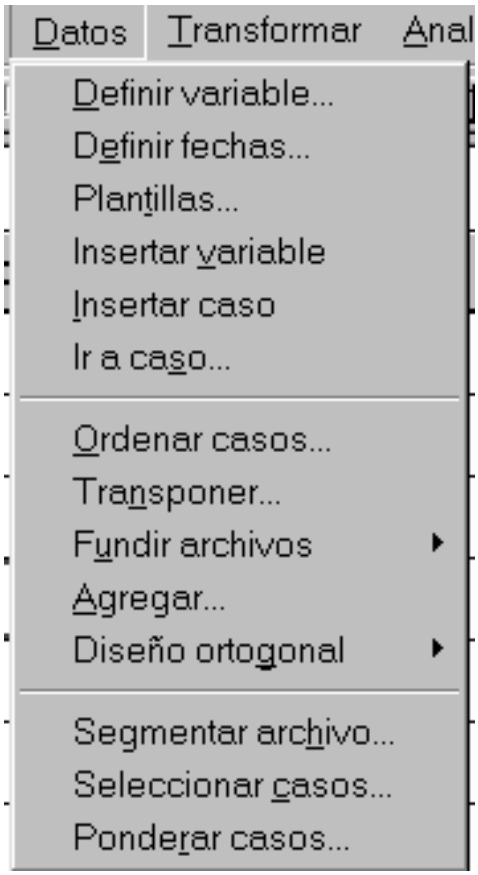
EDICIÓN Se realizan las tareas habituales de edición de texto, disponibles en la mayor parte de los programas del entorno Windows: modificar, borrar, copiar, pegar, seleccionar, buscar, etcétera.



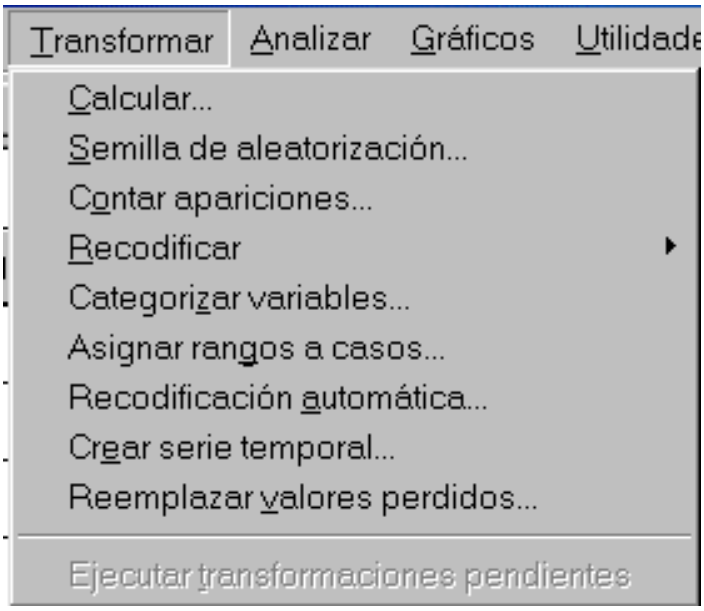
VER Desde esta opción se controlan diversos parámetros de visualización.



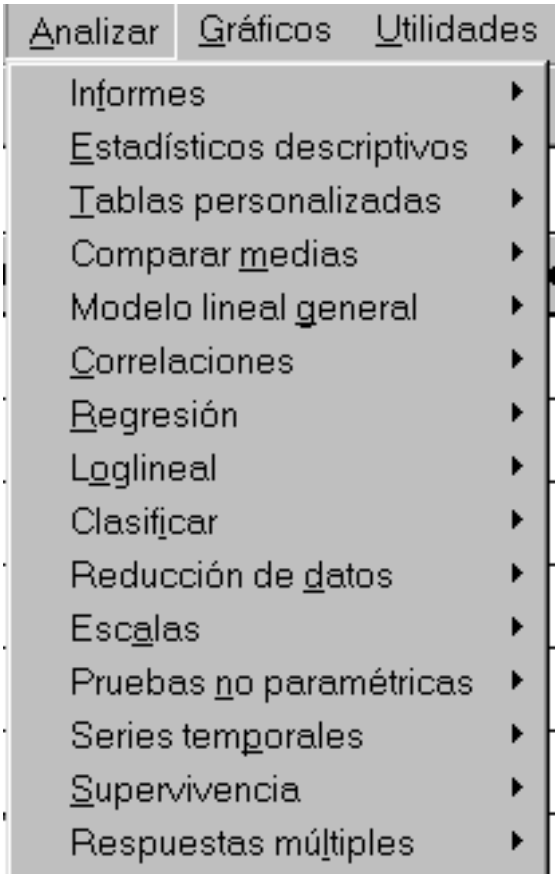
DATOS Permite definir variables y fechas, modificar los ficheros de datos activos, segmentar archivos, seleccionar y ponderar casos, etc... Las funciones de este menú son temporales y sólo permanecen activas mientras dure la sesión.



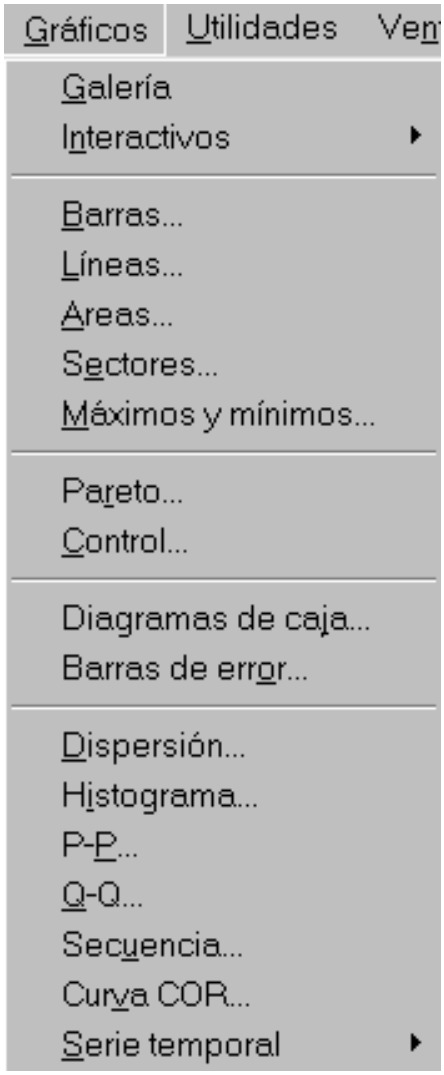
TRANSFORMAR Permite, en el fichero de datos activo, calcular nuevas variables a partir de las existentes, recodificar, asignar etiquetas a casos, y diversas operaciones relativas a la modificación y creación de nuevas variables. También las modificaciones son temporales y si éstas se quieren conservar hay que grabar los cambios.



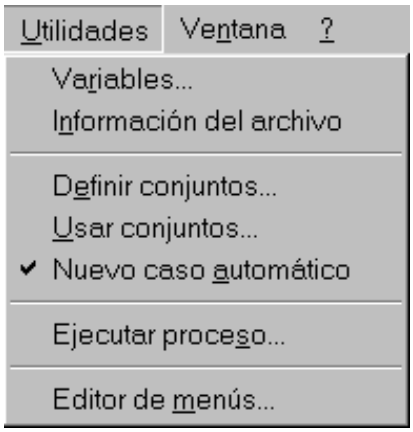
ANALIZAR Mediante este menú se accede a los diferentes análisis estadísticos disponibles en SPSS.



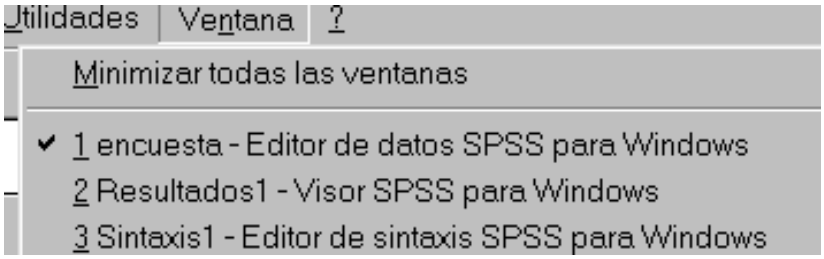
GRÁFICOS Desde aquí se accede a las posibilidades gráficas



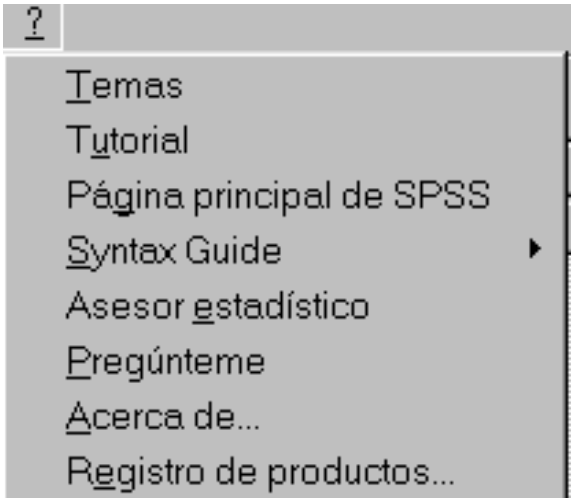
UTILIDADES Incluye diferentes opciones para visualizar el contenido de los ficheros de datos y crear subconjuntos de variables.



VENTANA Desde esta opción podemos controlar la ventana que queremos tener activa (ver apartado 1.3).



AYUDA El programa permite acceder al manual de ayuda a través de un completo menú de opciones:



1.3 EL SISTEMA DE VENTANAS EN EL SPSS

El SPSS dispone de ocho tipos de ventanas desde las cuales se pueden efectuar diversas operaciones.

- **Ventana del editor de datos.** En esta ventana están los datos del fichero con el que se está trabajando. Sólo puede haber un conjunto de datos activo (un solo fichero). Los ficheros de datos en SPSS se nombran: *.sav
- **Ventana del visor de resultados.** En esta ventana se guardan los diferentes resultados que generamos: salidas de los diferentes procedimientos, listados, subprogramas, mensajes de error, gráficos, etcétera. Una ventana de este tipo se abre automáticamente cuando se genera el primer resultado de la sesión. Se pueden tener tantas ventanas abiertas como se quiera. Los ficheros de resultados en SPSS se nombran: *.spo

- **Ventana del visor de resultados de borrador.** Es posible enviar los resultados a un visor preestablecido al que se accede a través de: "archivo"/"nuevo"/"resultados de borrador". También se pueden mantener abiertas tantas como se deseen.
- **Ventana del editor de tablas pivote.** Permiten editar y modificar las tablas pivote. Estas tablas disponen de la posibilidad de editar texto, intercambiar los datos transponiendo filas y columnas, modificar colores, etcétera.
- **Ventana del editor de gráficos.** En todos los gráficos que se generan en SPSS se pueden realizar modificaciones cambiando colores, fuentes y tamaños, ejes, rotaciones, etc.
- **Ventana del editor de resultados de texto**. Para modificar aquellos resultados de texto generados por el programa.
- **Ventana del editor de sintaxis**. Visualiza los ficheros de sintaxis o de lenguaje de comandos, que se pueden modificar desde este editor. Los ficheros de sintaxis se nombran: *.sps.

  Este editor es de gran utilidad especialmente en tres casos:

  - Algunas posibilidades del SPSS sólo son accesibles a través del lenguaje de comandos.
  - En operaciones que habitualmente se repiten es más adecuado grabar el programa completo y ejecutarlo desde esta ventana.
  - Si el ordenador tiene que ser compartido por muchos usuarios.

- **Ventana del editor de procesos.** También es posible automatizar y personalizar procesos aplicando la tecnología OLE y el lenguaje BASIC.

ESTADÍSTICA DESCRIPTIVA

2. ESTADÍSTICA DESCRIPTIVA

2.1. INTRODUCCIÓN

2.2. MEDIDAS DE POSICIÓN

- **Medidas de posición central**
  - **Media aritmética**
  - **Media geométrica**
  - **Media armónica**
  - **La mediana**
  - **La moda**
- **Medidas de posición no central**
  - **Cuartiles**
  - **Deciles**
  - **Percentiles**

2.3. MEDIDAS DE DISPERSIÓN

- **Medidas de dispersión absoluta**
  - **Recorrido**
  - **Recorrido intercuartilico**
  - **Desviación absoluta media respecto a la media aritmética**
  - **Desviación absoluta media respecto a la mediana**
  - **La varianza**
  - **La desviación típica o estándar**
- **Medidas de dispersión relativa**
  - **Coeficiente de apertura**
  - **Recorrido relativo**
  - **Recorrido semi-intercuartílico**
  - **Coeficiente de variación de Pearson**
  - **Índice de dispersión de la mediana**

2.4. TIPIFICACIÓN DE VARIABLES

2.5. MEDIDAS DE FORMA: ASIMETRÍA Y CURTOSIS

- **Medidas de asimetría**
  - **Coeficiente de asimetría de Fisher**
  - **Coeficiente de asimetria de Bowley**
  - **Medida de asimetría de Pearson**
- **Medidas de apuntamiento o curtosis**
  - **Coeficiente de apuntamiento o curtosis**

2.6. MEDIDAS DE CONCENTRACIÓN

- **Índice de Gini**
- **Curva de Lorenz**

2.7. LA REPRESENTACIÓN GRÁFICA DE LOS DATOS

- **Diagrama de Pareto**
- **Gráficos de barras**
- **Histograma**
- **Gráficos de series temporales**
- **Gráficos de sectores**
- **Gráficos de dispersión**
- **Diagramas de caja**
- **Diagramas de tallos y hojas (Stem an Leaf)**
- **Otras representaciones gráficas**
- **Creación de gráficos con Excel**

2.1. INTRODUCCIÓN

La Estadística Descriptiva es el primer paso en la investigación de poblaciones o conjunto de datos procedentes del recuento o de experimentos. Nos proporciona herramientas que nos permiten resumir la información obtenida y pasar así de un gran volumen de datos a otro más reducido.

La Estadística Descriptiva cubre un amplio conjunto de técnicas y métodos. En este capítulo contemplamos sólo algunos conceptos, los más elementales.

Las principales medidas que se estudian en la Estadística Descriptiva son:

- Medidas de Posición
- Medidas de Dispersión
- Medidas de Asimetría y Curtosis
- Medidas de Concentración

Los datos estadísticos sobre los que vamos a realizar los análisis que nos proporciona la Estadística Descriptiva se suelen presentar en tres situaciones diferentes:

a) Los valores no se repiten en ningún caso.
b) Cada valor de la característica medida se repite un determinado número de veces.
c) En numerosas ocasiones, clasificamos las observaciones en intervalos. Así, al preguntar a una persona por su edad su respuesta puede ser clasificada en uno de los siguientes intervalos:

0 a 18 años
19 a 30 años
31 a 45 años
45 a 60 años
Más de 60 años

lo habitual es que dichos intervalos se construyan basándose en un estudio previo o mediante algún criterio científico o técnico específico. Por ejemplo, si mediante un estudio anterior realizado por la Organización Mundial de la Salud sabemos que el consumo de seis cigarrillos diarios no ejerce ningún tipo de influencia

en la salud, de siete a diez cigarrillos se considera un consumo moderado, de once a veinte es un consumo de riesgo y con más de veinte tenemos un consumo excesivo, estableceremos los intervalos de acuerdo a dicho criterio. Es decir, definimos los siguientes grupos de consumo:

0 a 6 cigarrillos
7 a 10 cigarrillos
11 a 20 cigarrillos
Más de 20 cigarrillos

Establecer intervalos de una forma arbitraria puede conducirnos a falsas conclusiones, de ahí la importancia de utilizar un criterio reconocido para definir los estratos.

Según estemos en una u otra situación se utilizará una forma distinta de presentación de los datos, denominándose de tipo I a la primera situación, de tipo II a la segunda y de tipo III a la tercera.

A) PRESENTACIÓN DE TIPO I

La notación más utilizada en Estadística y que se asume en este trabajo es la siguiente:

- N -> Número de unidades en las cuales efectuamos la medición.
- X_i -> Valor que toma la característica en el individuo i.

Por lo tanto, los datos se representan como la sucesión: $X_1, X_2, X_3, ..., X_N$.

B) PRESENTACIÓN DE TIPO II

La notación utilizada es la siguiente:

- N -> Número de unidades en las cuales efectuamos la medición.
- k -> Número de valores distintos.
- Xi -> Valor que toma la característica en el individuo i, $i = 1, ..., k$.

- n_i -> Número de veces que aparece el valor X_i, es decir, frecuencia del valor

  X_i, $i = 1, ..., k$.

Por lo tanto, cada dato X_i, tendrá asociada su frecuencia de aparición, n_i.

Bajo esta notación, la suma de todos los n_i será igual al número de datos, es decir, N.

$$N = \sum_{i=1}^{k} n_i$$

C) PRESENTACIÓN DE TIPO III

Si tenemos los datos clasificados en intervalos, utilizaremos la siguiente notación:

- N -> Número de unidades en las cuales efectuamos la medición.
- K -> Número de intervalos considerados.
- L_{i-1}-L_i -> Intervalo i, siendo L_{i-1} el límite inferior y L_i el límite superior.
- n_i -> Número de unidades comprendidas en el intervalo i.
- N_i -> Número acumulado de unidades hasta el intervalo i.

Así pues, a cada intervalo se le asocia el número de valores que contiene, verificándose por tanto:

$$N = \sum_{i=1}^{k} n_i$$

Se muestra a continuación una tabla resumen en la que aparecen los tres tipos de presentaciones:

Tabla 2.1.
Formas de presentación de los datos

Tipo I	**Tipo II**			**Tipo III**		
X_i	X_i	n_i	N_i	Intervalo i	n_i	N_i
----	----	----	----	----------	----	----
X_1	X_1	n_1	N_1	L_0, L_1	n_1	N_1
X_2	X_2	n_2	N_2	L_1, L_2	n_2	N_2
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
X_i	X_i	n_i	N_i	L_{i-1}, L_i	n_i	N_i
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
X_N	X_k	n_k	N_k	L_{k-1}, L_k	n_k	N_k

En los apartados que siguen utilizaremos esta nomenclatura para las sucesivas definiciones.

2.2. MEDIDAS DE POSICION

Medidas de posición central

Las medidas de posición central más comunes son: la media, la mediana, y la moda. La media, a su vez, puede ser definida como media aritmética, geométrica y armónica. Cada una de ellas presenta sus ventajas e inconvenientes y su elección depende tanto de la naturaleza de la estadística como del propósito para el que se utiliza.

a) La media aritmética. Es la suma de todos los valores de la variable dividida por el número total de los datos.

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \cdots + x_n n_n}{N} = \sum_{i=1}^{n} \frac{x_i n_i}{N}$$

Las propiedades de la media aritmética son:

1. La suma de las desviaciones de los valores de la variable respecto al valor de la media es cero.

$$\sum_{i=1}^{n} (x_i - \bar{x}) n_i = 0$$

2. La media de las desviaciones elevadas al cuadrado de los valores de la variable respecto a una constante cualquiera es mínima si esta constante es la media.

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 \frac{n_i}{N} \text{ es mínima}$$

3. Si a todos los valores de la variable les sumamos una cantidad constante k, la media aritmética de la variable queda aumentada en esa constante. Lo mismo puede decirse respecto de la multiplicación. Si a todos los valores de la variable les multiplicamos por una constante, la media de esa variable se multiplica por esa constante.

Bajo esta propiedad, la variable *X'*, definida de la siguiente forma:

$$x_i' = \frac{x_i - o}{c}$$

siendo **o** y **c** dos constantes cualesquiera, se cumple que:

$$\bar{x}' = \frac{\bar{x} - o}{c} \Rightarrow c\bar{x}' = \bar{x} - o \Rightarrow \bar{x} = c\bar{x}' + o$$

Las ventajas de utilizar la media aritmética son:

- En el calculo intervienen todos los valores de la variable
- Es única
- Es calculable
- Es el centro de gravedad de la distribución.

Sin embargo está muy afectada por los valores extremos que presenten los datos, lo que puede originar que a veces las conclusiones no sean muy atinadas.

b) La media geométrica. Es la raíz N-ésima del producto de los valores de la variable elevados por sus respectivas frecuencias.

$$G = \sqrt[N]{x_1^{n_1} x_2^{n_2} \ldots x_n^{n_n}}$$

La propiedad fundamental de esta media es que el logaritmo de la media geométrica es igual a la media aritmética de los logaritmos de los valores de la variable.

La principal ventaja que ofrece esta media respecto a la media aritmética es su menor sensibilidad respecto a los valores extremos de la variable. La desventaja es que no está determinada si alguno de los valores de la variable es negativo. También tiene un significado menos intuitivo que la media aritmética.

Su utilización más frecuente es promediar porcentajes, y también se aconseja su uso cuando se presupone que la variable analizada se ha formado a partir de variaciones acumulativas.

c) La media armónica. La media armónica es la media aritmética de los inversos de los valores de la variable.

$$H = \frac{N}{\frac{1}{X_1}n_1 + \frac{1}{X_2}n_2 + \ldots\ldots + \frac{1}{X_n}n_n} = \frac{N}{\sum_{i=1}^{n}\frac{1}{X_i}n_i}$$

En ciertos casos la media armónica es más representativa que la media aritmética.

Tiene como inconvenientes que está muy influenciada por los valores pequeños y no está determinada cuando algún valor de la variable es igual a cero.

d) La mediana. La mediana es el valor de la distribución que divide la distribución de la variable en dos partes iguales, es decir deja a la izquierda y a la derecha igual número de valores si el numero de datos es impar. Cuando el número de valores es par se toma la media aritmética de los dos valores centrales. En términos de frecuencia se define la mediana como aquel valor de la distribución cuya frecuencia acumulada es $\frac{N}{2}$. Para distribuciones agrupadas en intervalos aplicamos la siguiente fórmula:

$$M_e = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i}c_i$$

siendo c_i es la amplitud del intervalo donde se encuentra la mitad de la distribución y N_{i-1} es la frecuencia acumulada inmediatamente anterior al intervalo donde se encuentra la mitad de la distribución ($N/2$) y n_i la frecuencia del intervalo.

La propiedad fundamental de la mediana es que la suma de todas las desviaciones en valor absoluto de la variable respecto de la mediana es mínima.

La mediana adquiere mayor importancia cuando las variables son ordinales, o susceptibles de ser ordenadas, en cuyo caso la mediana es la medida de tendencia central más representativa.

d) La moda. Es el valor de la variable que más veces se repite. Para distribuciones agrupadas en intervalos se utiliza la siguiente fórmula. $\frac{N}{4}$

$$M_o = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} c_i$$

donde n_{i-1} n_{i+1} son las frecuencias asociadas a los intervalos anterior y posterior del intervalo que más se repite.

Si los intervalos no tienen la misma amplitud debemos calcular las densidades de frecuencia, que se obtienen dividiendo las frecuencias absolutas de cada valor de la variable por las amplitudes de cada intervalo.

$$M_o = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} c_i$$

Siendo $d_i = \frac{n_i}{c_i}$

Medidas de posición no central

Son medidas de posición no central los cuartiles, deciles y percentiles. Las medidas de posición no centrales dividen la distribución en partes iguales. Los cuartiles son tres valores y dividen la distribución en cuatro partes iguales . Los deciles son nueve y dividen la distribución en diez partes. Los percentiles son 99 y dividen la distribución en cien partes.

Así, el primer cuartil C_1 es el valor que ocupa el lugar

el primer decil D_1 es el valor que ocupa el $\frac{N}{10}$ lugar

y el primer percentil P_1 es el valor que ocupa el lugar $\frac{N}{100}$

Para distribuciones agrupadas en intervalos utilizamos la siguiente fórmula

$$Q_{r/k} = L_{i-1} + \frac{\frac{r}{k}.N - N_{i-1}}{n_i} \cdot c_i$$

Si k=4 y r = 1, 2, 3 obtenemos los cuartiles

Si k=10 y r = 1, 2,.....,9 obtenemos los deciles

Si k=100 y r = 1, 2,....,99 obtenemos los percentiles

A continuación vamos a calcular la media, la mediana y la moda de la distribución de salarios de la empresa XXX,SA (tabla 2.2), constituida por 1.000 trabajadores:

Tabla 2.2.
Distribución de los salarios que paga la empresa XXX S.A.

Salario Mensual ¡ X_i	*Marca de clase*	*Nº de Trabajadores* n_i	*Nº acumulado de trabajadores* N_i	*Total de Salarios* $X_i n_i$
60.000-80.000	70.000	160	160	11.200.000
80000-100000	90.000	200	360	18.000.000
100000-120000	110.000	100	460	11.000.000
120000-140000	130.000	110	570	14.300.000
140000-160000	150.000	100	670	15.000.000
160000-180000	170.000	85	755	14.450.000
180000-200000	190.000	10	765	1.900.000
200000-220000	210.000	14	779	2.940.000
220000-240000	230.000	25	804	5.750.000
240000-260000	250.000	47	851	11.750.000
260000-280000	270.000	24	875	6.480.000
280000-30000	290.000	40	915	11.600.000
320000-340000	310.000	85	1.000	26.350.000
				150.720.000

La media aritmética se calcularía:

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_n n_n}{N} = \sum_{i=1}^{n} \frac{x_i n_i}{N} = \frac{150.720.000}{1000} = 150.720$$

Para calcular la mediana partimos del intervalo central, el intervalo 120.000-140.000, en donde sabemos que ha de estar la mitad de nuestra distribución (*N/2*). Esto

implica que 460 sea el valor que toma N_{i-1} (frecuencia acumulada del intervalo inmediatamente anterior), y 110 el valor de n_i (frecuencia relativa del intervalo).

$$M_e = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} c_i = 120.000 + \frac{500 - 460}{110} 20.000 = 127.273$$

En el cálculo de la moda dado que el intervalo que más se repite es el de 80.000-100.000, el valor de n_{i-1} es 160 y el de n_{i+1} es 100.

$$M_o = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} c_i = 80.000 + \frac{100}{160 + 100} 20.000 = 87.692$$

2.3. MEDIDAS DE DISPERSION

Medidas de dispersión absoluta

Las medidas de dispersión o de variabilidad miden la representatividad de las medidas de tendencia central, obteniéndose como desviación de los valores de la distribución respecto a estas medidas.

Las medidas de dispersión o de variabilidad son: el recorrido, el recorrido intercuartílico, la desviación absoluta media respecto a la media aritmética, la desviación absoluta media respecto a la mediana, la varianza y la desviación típica o estándar.

a) Recorrido. Es la diferencia entre el valor mayor y el valor menor de la distribución

$$R = X_n - X_1$$

b) Recorrido intercuartílico. Es la diferencia que exite entre el tercer cuartil y el primer cuartil

$$R_1 = C_3 - C_1$$

c) Desviación absoluta media respecto a la media aritmética

$$D_{\bar{x}} = \sum_{i=1}^{n} \left| X_i - \overline{X} \right| \frac{n_i}{N}$$

d) Desviación absoluta media respecto a la mediana

$$D_{M_e} = \sum_{i=1}^{n} \left| X_i - M_e \right| \frac{n_i}{N}$$

e) La Varianza

$$S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 \frac{n_i}{N}$$

e) La Desviación típica o estándar

$$S = \sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2 \frac{n_i}{N}}$$

Las propiedades de la desviación típica son:

- Es siempre mayor o igual que cero
- Es una medida de dispersión óptima
- Está acotada superior e inferiormente
- No está afectada por cambios de origen
- Si que está afectada por cambios de escala (queda multiplicada por el factor de escala)

Medidas de dispersión relativa

Las medidas de dispersión relativa tratan de hacer comparables distribuciones diferentes, es decir, distribuciones que no vienen expresadas en las mismas medidas. A diferencia de la medidas de variabilidad, las medidas de dispersión relativa son medidas adimensionales y las más utilizadas son: el coeficiente de apertura, el recorrido relativo, el recorrido semi-intercuartílico y el coeficiente de variación de Pearson.

f) Coeficiente de apertura. Es la relación entre el mayor y el menor valor de la distribución

$$A = \frac{X_n}{X_1}$$

g) Recorrido relativo. Es el cociente entre el recorrido y la media. Esta expresión mide el número de veces que el recorrido contiene a la media aritmética

$$RR = \frac{R_e}{\overline{X}}$$

h) Recorrido semi-intercuartílico. Es el cociente entre el recorrido intercuartílico y la suma del primer y tercer cuartil

$$R_s = \frac{C_3 - C_1}{C_3 + C_1}$$

g) Coeficiente de Variación de Pearson. Resuelve el problema de comparar medias aritméticas provenientes de distribuciones medidas con unidades diferentes. Es el cociente entre la desviación típica y la media aritmética

$V = \frac{S}{\overline{X}}$ también se puede expresar en porcentaje: $V = \frac{S}{\overline{X}}.100$

Al venir expresados tanto la desviación típica como la media en las mismas unidades, el coeficiente de variación de Pearson es adimensional. También es invariable respecto a los cambios de origen.

Dado que el coeficiente representa el número de veces que la desviación típica contiene a la media, entonces si V=0 la representatividad de la media sería máxima y si V>0,5 indicaría una baja representatividad de la media

j) Índice de dispersión de la mediana

$$V_{Me} = \frac{D_{Me}}{M_e} = \frac{\sum_{i=1}^{n} |X_i - M_e| n_i}{N \cdot M_e}$$

2.4. LA TIPIFICACIÓN DE VARIABLES

La tipificación de variables consiste en expresar la diferencia entre la media y los valores de la variable en términos de desviación típica.

$$Z = \frac{X - \overline{X}}{S}$$

Cuando tipificamos una variable, la media de la variable tipificada *Z* es igual a 0 y su desviación típica 1.

Veamos con un ejemplo, el uso de esta técnica. Supóngase que los alumnos de primer curso de matemáticas están distribuidos en un centro en dos aulas distintas (Clase A y Clase B) y que para una misma asignatura, análisis matemático por ejemplo, tienen dos profesores distintos. Supóngase además que dentro de cada aula no ha habido ninguna selección de alumnos previa y puede esperarse un mismo nivel de aprendizaje en las dos aulas.

Después de realizar el mismo examen de análisis matemático, las notas de los alumnos para cada aula son las siguientes:

Tabla 2.3.
Notas de las clases A y B

NOTAS	Clase A	Clase B
Alumno 1	5,00	5,50
Alumno 2	2,00	7,00
Alumno 3	6,75	7,25
Alumno 4	9,00	5,00
Alumno 5	7,50	8,25
Alumno 6	6,75	2,80
Alumno 7	3,50	7,75
Alumno 8	5,30	8,25
Alumno 9	8,50	6,75
Alumno 10	2,75	7,25
Alumno 11	4,00	8,75
Alumno 12	2,75	6,75
Alumno 13	4,75	9,50
Alumno 14	3,00	8,25
Alumno 15	4,00	7,50
Alumno 16	3,00	5,25
Alumno 17	4,50	6,25
Alumno 18	4,75	6,50
Alumno 19	6,50	8,50
Alumno 20	5,00	5,75
Alumno 21	5,00	5,25
Alumno 22	4,50	4,75
Alumno 23	7,25	6,75
Alumno 24	6,00	8,50
Alumno 25	5,50	8,00

Las notas medias y las desviaciones típicas para cada aula son las siguientes:

	Clase A	Clase B
Media	5,10	6,88
Desviación típica	1,80	1,52

Puede observarse, que en la Clase A la nota media ha sido más baja que en la Clase B, dándose en la Clase A una mayor variabilidad. Esto puede deberse a que el profesor de la clase A ha sido algo más exigente a la hora de corregir.

Si queremos comparar dos alumnos, uno de una clase y otro de otra, con el objetivo de comprobar cuál de ellos ha alcanzado un mayor nivel de aprendizaje, utilizar la nota obtenida por cada uno en el examen puede llevarnos a falsas conclusiones, conscientes de que el profesor de la Clase A ha sido, probablemente, más duro a la hora de corregir..

Con el objetivo de eliminar esta influencia, tipificamos las notas de ambas clases. Los resultados obtenidos son los siguientes:

Tabla 2.4.
Notas tipificadas de las clases A y B

NOTAS	Clase A	Clase B
Alumno 1	-0,06	-0,91
Alumno 2	-1,72	0,08
Alumno 3	0,92	0,24
Alumno 4	2,17	-1,24
Alumno 5	1,33	0,90
Alumno 6	0,92	-2,69
Alumno 7	-0,89	0,57
Alumno 8	0,11	0,90
Alumno 9	1,89	-0,09
Alumno 10	-1,31	0,24
Alumno 11	-0,61	1,23
Alumno 12	-1,31	-0,09
Alumno 13	-0,20	1,72
Alumno 14	-1,17	0,90
Alumno 15	-0,61	0,41
Alumno 16	-1,17	-1,07
Alumno 17	-0,33	-0,42
Alumno 18	-0,20	-0,25
Alumno 19	0,78	1,06
Alumno 20	-0,06	-0,75
Alumno 21	-0,06	-1,07
Alumno 22	-0,33	-1,40
Alumno 23	1,19	-0,09
Alumno 24	0,50	1,06
Alumno 25	0,22	0,74

	Clase A	Clase B
Media de variable tipificada	0,00	0,00
Desviación típica de variable tipificada	1,00	1,00

Si comparamos el primer alumno de cada clase, podemos observar lo siguiente:

Alumno 1	**Clase A**	**Clase B**	**Diferencia**
Nota	5,00	5,50	-0,50
Nota tipificada	-0,06	-0,91	0,85

A pesar de que el alumno de la clase A tiene una nota ligeramente inferior al de la clase B (5 frente a 5,5), en relación al nivel medio de su clase no cabe duda del mayor nivel de aprendizaje del primer alumno de la clase A, ya que su nota tipificada es mayor (-0,06 frente a -0,91). Además, gracias a la tipificación comprobamos que ambos alumnos están por debajo de la media de aprendizaje.

2.5. MEDIDAS DE FORMA: ASIMETRIA Y CURTOSIS

Medidas de asimetría

Las medidas de asimetría son indicadores que permiten establecer el grado de simetría de una distribución de valores estadísticos sin necesidad de realizar el gráfico de la distribución.

a) Coeficiente de asimetría de Fisher.

$$g_1 = \frac{\frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^3 n_i}{(\sum_{i=1}^{n}(x_i - \bar{x})^2 \frac{n_i}{N})^{\frac{3}{2}}}$$

Según el valor de g_1, se deduce:

Si $g_1 = 0$ la distribución es simétrica

Si $g_1 < 0$ la distribución es asimétrica a la izquierda

Si $g_1 > 0$ la distribución es asimétrica a la derecha

b) Coeficiente de asimetría de Bowley. Está basado en los valores de los cuartiles y la mediana.

$$A_B = \frac{C_3 + C_1 - 2M_e}{C_3 - C_1}$$

Dependiendo del valor de A_B concluimos que:

Si $A_B=0$ la distribución es simétrica

Si $A_B>0$ la distribución es asimétrica a la derecha

Si $A_B<0$ la distribución es asimétrica a la izquierda

c) Medida de Asimetría de Pearson.

$$A_p = \frac{\bar{x} - M_o}{S}$$

Dado que para distribuciones campaniformes, unimodales y moderadamente asimétricas, se verifica que $\bar{X} - M_o \approx 3(\bar{X} - M_e)$, algunos autores prefieren utilizar esta otra medida de asimetría:

$$A_p = \frac{3(\bar{x} - M_e)}{S}$$

Dependiendo del valor que tome A_p, señalamos que:

Si $A_p=0$ la distribución es simétrica

Si $A_p>0$ la distribución es asimétrica a la derecha

Si $A_p<0$ la distribución es asimétrica a la izquierda

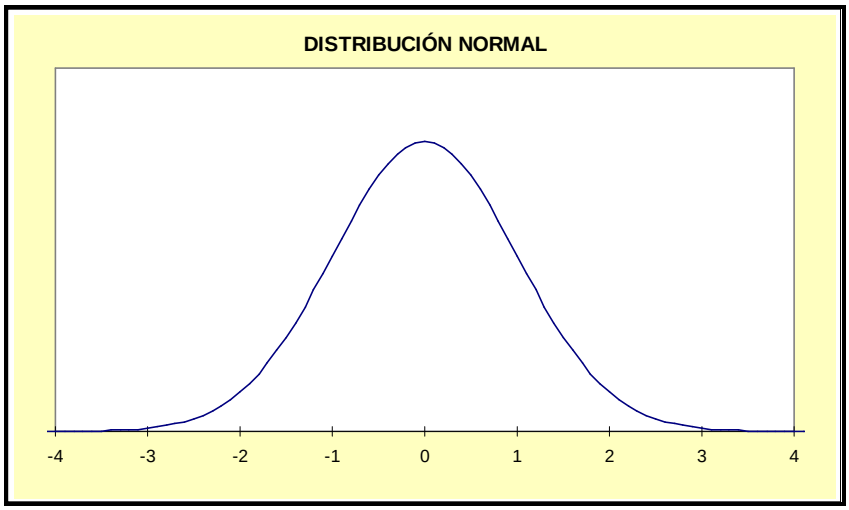
Medidas de apuntamiento o curtosis

Por su parte, las medidas de apuntamiento o curtosis tratan de estudiar la distribución de frecuencias en la zona media. El mayor o menor número de valores de la variable alrededor de la media dará lugar a una distribución más o menos apuntada.

Para estudiar el apuntamiento hay que definir una distribución tipo que nos sirva de referencia. Esta distribución es conocida como distribución Normal o curva de Gauss y se corresponde con numerosos fenómenos de la naturaleza. Su forma es la de una campana en donde la gran mayoría de los valores se encuentran concentrados alrededor de la media, siendo escasos los valores que están muy distanciados de ésta.

La representación gráfica de la distribución normal es:

Gráfico 2.1.
Representación gráfica de la distribución normal



Al tomar como referencia la distribución normal se dice que otra distribución es más apuntada que la distribución normal (leptocúrtica) o menos apuntada (platicúrtica). A las distribuciones que se asemejan a la distribución normal se les denomina mesocúrticas.

Dado que en una distribución Normal se verifica siempre que :

$$m_4 = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^4 \frac{n_i}{N} = 3\left(\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \frac{n_i}{N} \right)^2 = 3\left(S^2\right)^2 = 3S^4$$

El coeficiente de apuntamiento o curtosis utilizado es el siguiente:

$$g_2 = \frac{m_4}{S^4} - 3$$

Dependiendo entonces del valor del coeficiente g_2 llamamos

Mesocúrtica(Normal)	**si**	$g_2 = 0$
Leptocúrtica	**si**	$g_2 > 0$
Platicúrtica	**si**	$g_2 < 0$

2.6. MEDIDAS DE CONCENTRACION

Denominamos concentración de una variable a la mayor o menor equidad en el reparto de la suma total de esa variable. Las medidas de concentración intentan, pues, mostrarnos el mayor o menor grado de igualdad en el reparto del total de los valores de una variable Estas medidas tienen mucho interés en algunas distribuciones donde ni la media ni la varianza son significativas.

Las medidas de concentración más utilizadas son el **índice de concentración de Gini y la curva de Lorenz.**

Para proceder a su cálculo se utiliza la siguiente tabla

Tabla 2.5.
Cálculo del Índice de Gini y la curva de Lorenz

x_i	n_i	$x_i n_i$	N_i	U_i	$p_i = \frac{N_i}{N}100$	$q_i = \frac{U_i}{U_n}100$
$x_{(1)}$	n_1	$x_1 n_1$	N_1	U_1	p_1	q_1
$x_{(2)}$	n_2	$x_2 n_2$	N_2	U_2	p_2	q_2
.	.		.	.	.	.
.	.		.	.	.	.
$x_{(i)}$	n_i	$x_i n_i$	N_i	U_i	p_i	q_I
.	.		.	.	.	.
$x_{(n)}$	n_n	$x_n n_n$	N_n	U_n	p_n	q_n
	N	u_n				

Primero, se ordenan los valores de la variable *X;* $x_{(1)} \le x_{(2)} \le x_{(3)} \le ... \le x_{(n)}$ se calculan los productos $x_i\, n_i$; y las frecuencias acumuladas N_i.

Los valores U_i se calculan de la siguiente forma:

$$U_1 = X_1 n_1$$
$$U_2 = X_1 n_1 + X_2 n_2$$
.
$$U_n = X_1 n_1 + X_2 n_2 + ... + X_n n_n = \sum_{i=1}^{n} X_i n_i$$

Índice de Gini

El índice de Gini se calcula a través de la siguiente expresión:

$$I_G = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i}$$

El índice de Gini toma valores entre 0 y 1. Si la variable está distribuida homogeneamente la concentración es mínima $p_i = q_i$ lo que implica que el Índice de Gini tome un valor próximo a cero. Por el contrario, si el total de la distribución está muy concentrado en el último valor de la variable el índice se aproximaría a 1.

Curva de Lorenz

La curva de Lorenz es la representación gráfica de los coeficientes p_i y q_i. En el eje de abcisas se representa p_i y en el de ordenadas q_i. En la representación gráfica de la curva de Lorenz se incluye la diagonal principal, que indica la mayor igualdad en el reparto de la variable (p_i=q_i). En consecuencia, cuando la curva de Lorenz se aproxime a la diagonal principal mayor homogeneidad habrá en la distribución de la variable. El índice de Gini se calcula como el cociente del área comprendida entre la diagonal principal y la curva de Lorenz, y el área que está por debajo de la diagonal principal.

Vamos a calcular el índice de Gini y la curva de Lorenz utilizando como ejemplo la distribución de salarios de la tabla 2.2.

Tabla 2.6.
Índice de Gini y curva de Lorenz de la distribución de salarios

Salario Mensual	Marca de clase	Nº de Trabajadores	Nº acumulado de trabajadores	Total de Salarios	$U_i = \sum_{i=1}^{n} X_i n_i$	$p_i = \frac{N_i}{N} 100$	$q_i = \frac{U_i}{U_n} 100$
	X_i	n_i	N_i	$X_i n_i$	U_i	p_i	q_i
60.000-80.000	70.000	160	160	11.200.000	11.200.000	16	7
80.000-100.000	90.000	200	360	18.000.000	29.200.000	36	19
100.000-120.000	110.000	100	460	11.000.000	40.200.000	46	27
120.000-140.000	130.000	110	570	14.300.000	54.500.000	57	36
140.000-160.000	150.000	100	670	15.000.000	69.500.000	67	46
160.000-180.000	170.000	85	755	14.450.000	83.950.000	76	56
180.000-200.000	190.000	10	765	1.900.000	85.850.000	77	57
200.000-220.000	210.000	14	779	2.940.000	88.790.000	78	59
220.000-240.000	230.000	25	804	5.750.000	94.540.000	80	63
240.000-260.000	250.000	47	851	11.750.000	106.290.000	85	71
260.000-280.000	270.000	24	875	6.480.000	112.770.000	88	75
280.000-300.000	290.000	40	915	11.600.000	124.370.000	92	83
320.000-340.000	310.000	85	1.000	26.350.000	150.720.000	100	100

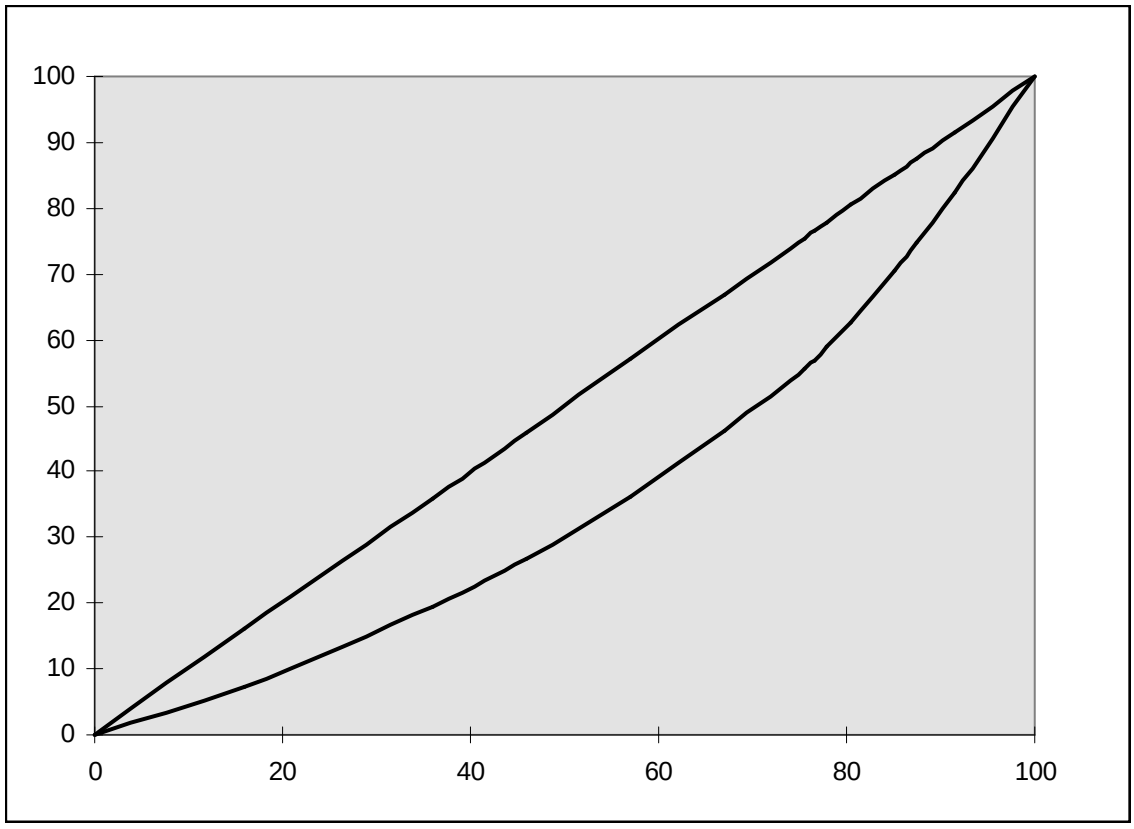
Al analizar las dos últimas columnas se observa que el 16% de los trabajadores se reparte el 7% de los salarios de la empresa y que el 46% de los trabajadores perciben solamente el 27% del total de los salarios .Si los salarios estuvieran equidistribuidos entonces el 16% de los trabajadores recibiría el 16% de los salarios, el 46% recibiría el 46% del total de los salarios, etc. Comprobamos a través del índice de Gini que los salarios no están equidistribuidos.

$$I_G = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i} = 0{,}25$$

A la misma conclusión llegamos al realizar la curva de Lorenz

Gráfico 2.2.

Curva de Lorenz de la distribución de salarios



2.7. LA REPRESENTACIÓN GRÁFICA DE LOS DATOS

El estudio de las distribuciones estadísticas resulta más atractivo cuando va acompañado, no sólo de las medidas descriptivas señaladas anteriormente, sino también de gráficos y diagramas que realcen las características que tratamos de describir.

Existen diversos tipos de análisis gráficos. Aquí se describen aquellos más utilizados:

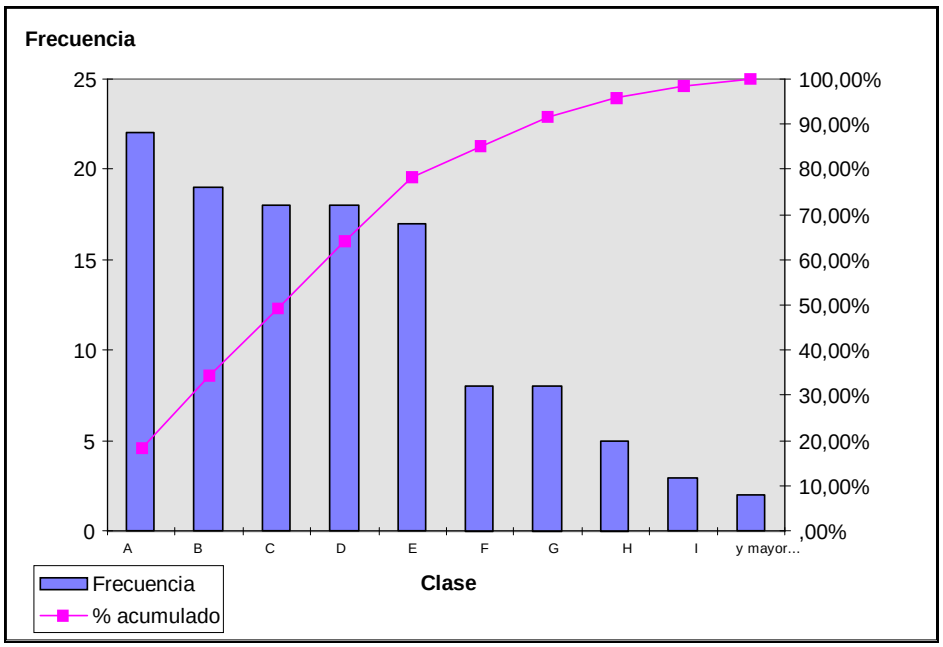
- **Diagramas de Pareto**

Se emplea para representar datos cualitativos y su construcción se realiza en dos pasos:

a) Ordenamos las clases o categorías según la frecuencia relativa de su aparición
b) Cada clase se representa por un rectángulo con una altura igual a la frecuencia relativa

El diagrama de Pareto representa los valores de las variables en el eje de abscisas y las frecuencias absolutas y relativas acumuladas en el eje de ordenadas.

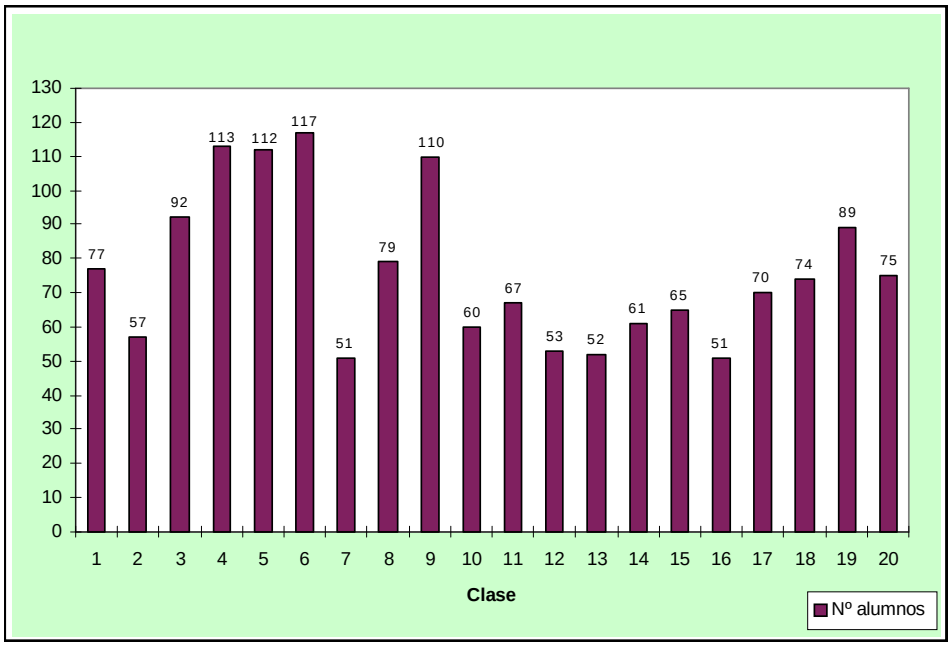
Gráfico 2.3.
Ejemplo de diagrama de Pareto



- **Gráficos de barras**

En general, se emplean para variables discretas en distribuciones de frecuencias de datos sin agrupar. Su mayor utilidad es comparar valores discretos a partir de dos o más series. Estos diagramas representan los valores de las variables en el eje de abscisas y en el de ordenadas se levanta, para cada punto, una barra con un valor igual a la frecuencia absoluta o relativa.

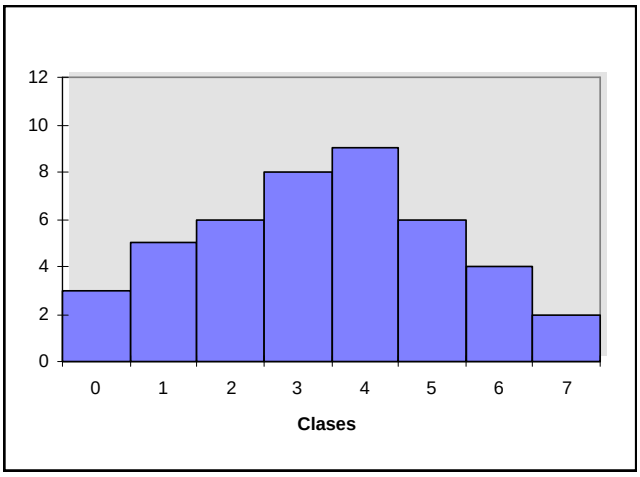
Gráfico 2.4.
Ejemplo de diagrama de barras



- **Histograma**

Los histogramas son las representaciones más frecuentes para ver los datos agrupados. Esta representación es un conjunto de rectángulos donde cada uno representa una clase. La base de los rectángulos es igual a la amplitud del intervalo y la altura se determina de tal forma que el área del rectángulo sea proporcional a la frecuencia de cada clase.

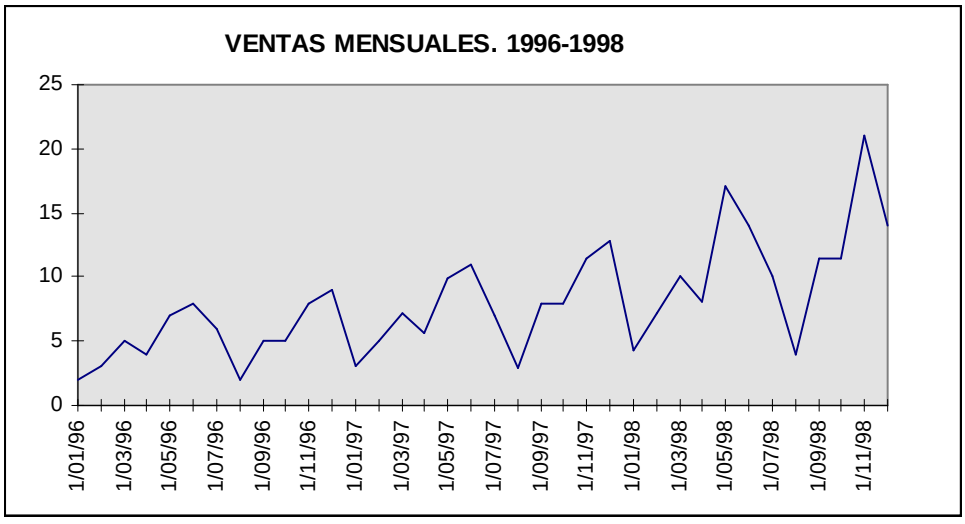
Gráfico 2.5.
Ejemplo de histograma



- **Gráficos de series temporales**

Una secuencia de valores a intervalos regulares de tiempo constituye una serie temporal. En los gráficos de series temporales se representan los valores ordenados según la secuencia temporal, la cual figura en ordenadas, en tanto que los valores obtenidos se representan en el eje de abscisas.

Gráfico 2.6.
Ejemplo de serie temporal



- **Gráficos de sectores**

Estos gráficos se utilizan para mostrar las contribuciones relativas de cada punto de los datos al total de la serie. En un gráfico de sectores sólo se representa una serie.

Gráfico 2.7.
Ejemplo de gráfico de sectores

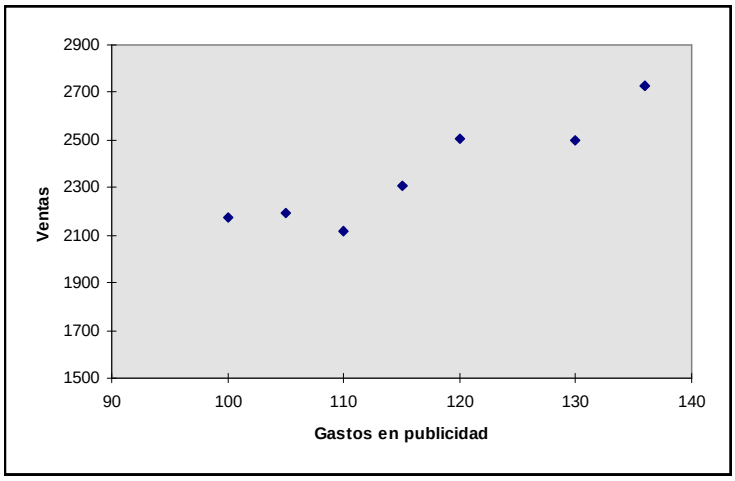
- **Gráficos de dispersión**

En este tipo de gráficos se visualizan dos series. Habitualmente el eje de ordenadas o eje *y*, es el eje de valores, y el de abcisas o eje *x*, es el de categorías. En los gráficos de dispersión ambos ejes tienen valores medibles, y normalmente se utilizan para ver la relación que existe, entre las series de datos que se representan.

En el ejemplo que sigue se muestra la relación entre gastos en publicidad y ventas en una empresa durante siete años.

Gráfico 2.8.

Ejemplo de gráfico de dispersión



- **Diagramas de caja**

Los diagramas de caja son representaciones semigráficas de un conjunto de datos que muestran las características principales de la distribución y señalan los datos atípicos (outliers).

Para la construcción de un diagrama de caja hay que seguir ciertos pasos.

a) Se ordenan los datos y se calcula el valor mínimo, el máximo y los tres cuartiles Q_1, Q_2,Q_3

b) Dibujamos un rectángulo cuyos extremos sean Q_1 y Q_3 y se indica dentro de la caja mediante una línea la posición de la mediana Q_2.

c) Calculamos los valores atípicos, que serán aquellos valores que caen fuera de los siguientes límites:

- Límite inferior

$$L_I = Q_1 - 1{,}5\left(\frac{Q_3 - Q_1}{2}\right)$$

- Límite superior

$$L_5 = Q_3 + 1{,}5\left(\frac{Q_3 - Q_1}{2}\right)$$

d) Se dibuja una línea que vaya desde cada extremo del rectángulo central hasta el valor más alejado y que no sea atípico.

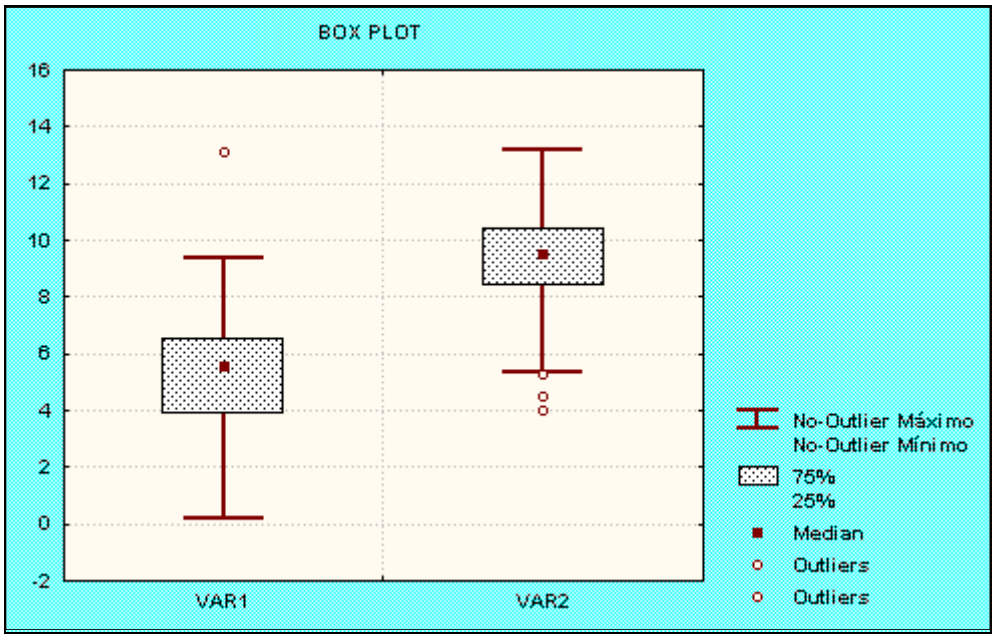
e) Se marcan todos los valores que sean atípicos

La utilización de la mediana en este tipo de gráficos, en vez de la media, como medida central de los datos, viene justificada porque la mediana es poco influenciable por los valores atípicos.

En el siguiente gráfico se han representado dos variables, la primera con una media de 5 y con una desviación típica de 3, y la segunda con la misma desviación típica y con una media de 9:

Gráfico 2.9.

Ejemplo de diagrama de caja (box plot)



- **Diagramas de tallos y hojas. (Stem and Leaf)**

Estos diagramas son procedimientos semi-gráficos cuyo objetivo es presentar los datos cuantitativos de una forma sintética, siempre y cuando, éstos no sean muy numerosos.

Para su construcción seguiremos los siguientes pasos.

a) Se redondean los datos expresándolos en unidades convenientes
b) Se disponen en una tabla. A la izquierda se escribe, para datos con dos cifras, el primer número, que será el tallo, y a la derecha, las unidades que formarán las hojas. Si el número es el 54 se escribe 5/4
c) Cada tallo definirá una clase y sólo se escribe una vez. El número de hojas representa la frecuencia de dicha clase.

A continuación vamos a representar un diagrama de tallos y hojas, utilizando como variable las medidas en centímetros de una pieza de metal que se han obtenido a partir de una muestra de todas las piezas fabricadas por una unidad de fabricación:

160,2	170,4	158,9	160,7	161,2	158,2
160,4	170,6	166,2	158,1	160,9	155,1
160,4	157,2	170,1	170,4		
158,3	161,4	170,7	166,5		

Redondeamos los datos a milímetros

160	170	159	161	161	158
160	171	166	158	161	155
160	157	170	170		
158	161	171	166		

Representamos el diagrama de tallos y hojas

Gráfico 2.10.

Ejemplo de diagrama de tallos y hojas

15	5 7 8 8 8 9
16	0 0 0 1 1 1 1
16	6 6
17	0 0 0 1 1

- **Otras representaciones gráficas**

Dado que el verdadero interés de los gráficos es describir la información de los datos, la naturaleza de las variables nos pueden sugerir otras representaciones distintas de las anteriores. Dos ejemplos de ello se muestran a continuación:

Gráfico 2.11.
Ejemplo de mapa o cartograma

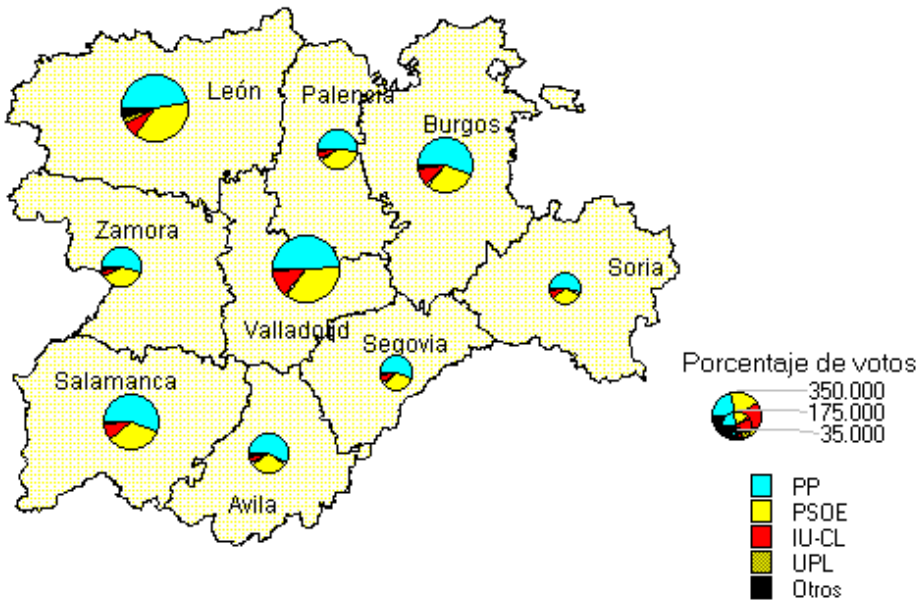
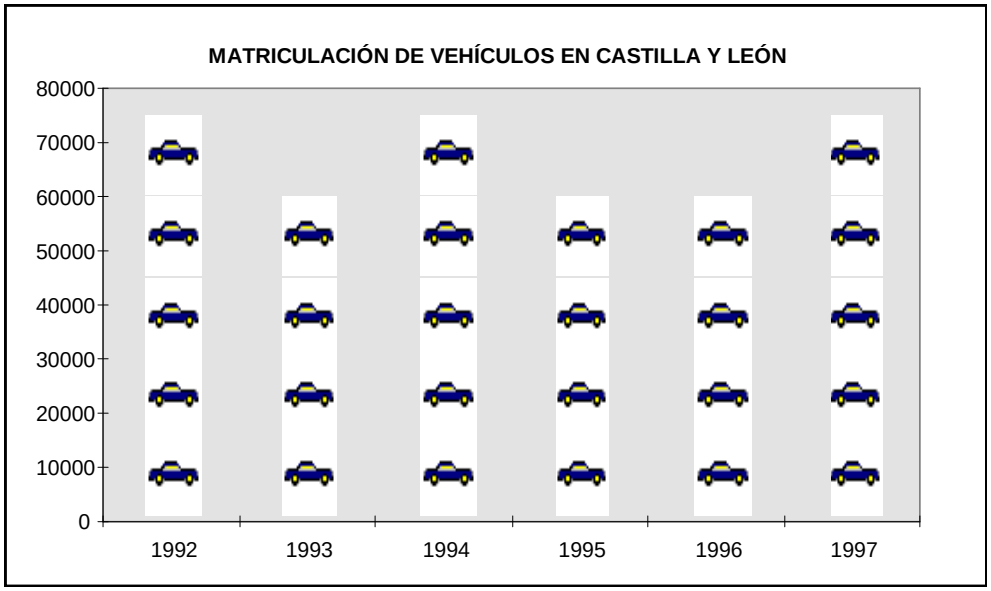


Gráfico 2.12.
Ejemplo de pictograma

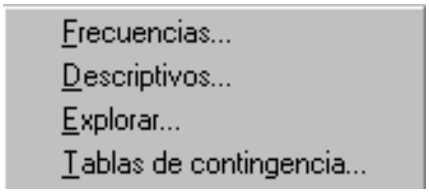


2.8 Ejemplo en SPSS

En el SPSS el análisis descriptivo de los datos en SPSS es sencillo. El menú "análisis"/"estadísticos descriptivos" dispone de variar opciones para analizar las variables del fichero de datos.

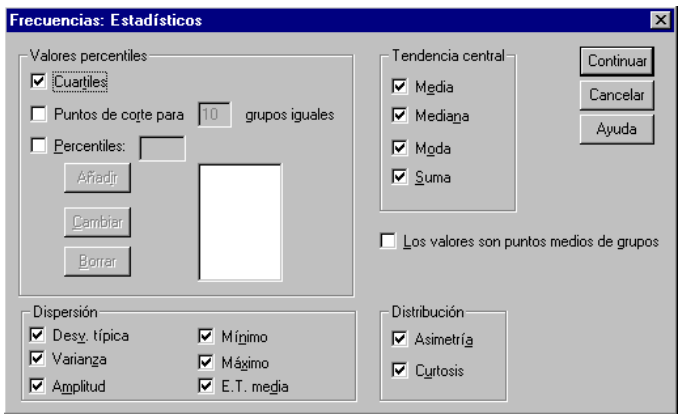
Vamos a utilizar la variable ***EDAD*** de la encuesta descrita en el Anexo nº 1.

Desde el menú de SPSS ***Analizar / Estadísticos Descriptivos*** accedemos a la siguiente pantalla:



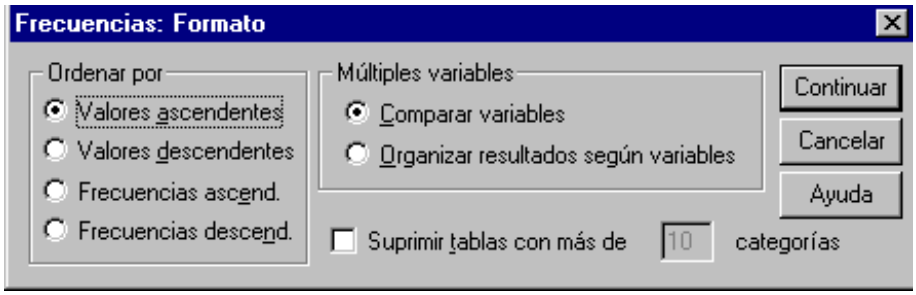
Si optamos por ***Frecuencias*** obtenemos los siguientes cuadros de diálogo

En ***Estadísticos*** podemos elegir diferentes medidas de posición, dispersión y forma:



En la opción de ***Gráficos*** podemos optar por realizar un histograma, un gráfico de barras o de sectores

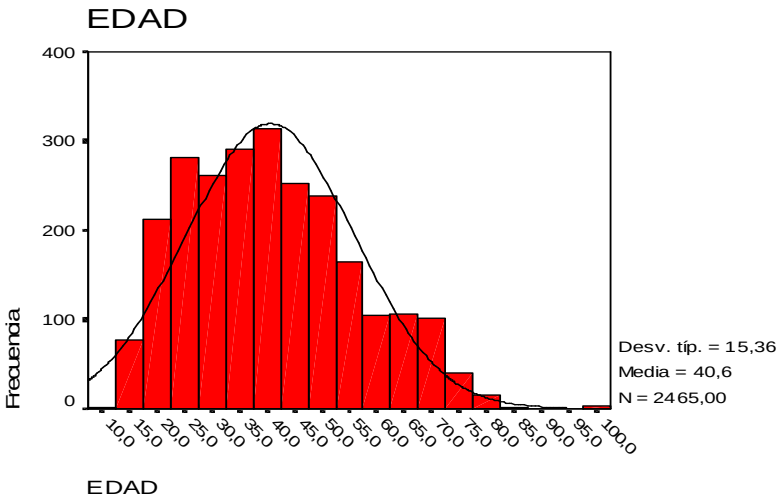
También podemos dar formato a la salida de resultados:



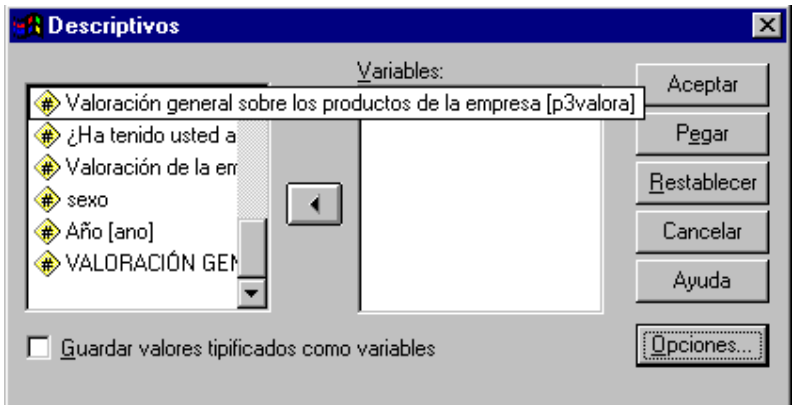
La salida de resultados se presenta a continuación:

EDAD

N	**Válidos**	**2465**
	Perdidos	169
Media		40,62
Error típ. de la media		,31
Mediana		39,00
Moda		40
Desv. típ.		15,36
Varianza		235,85
Asimetría		,466
Error típ. de asimetría		,049
Curtosis		-,377
Error típ. de curtosis		,099
Rango		90
Mínimo		9
Máximo		99
Suma		100134
Percentiles	25	28,00
	50	39,00
	75	50,00



En ***Descriptivos*** tenemos dos pantallas una para introducir las variables y otra para elegir las medidas estadísticas



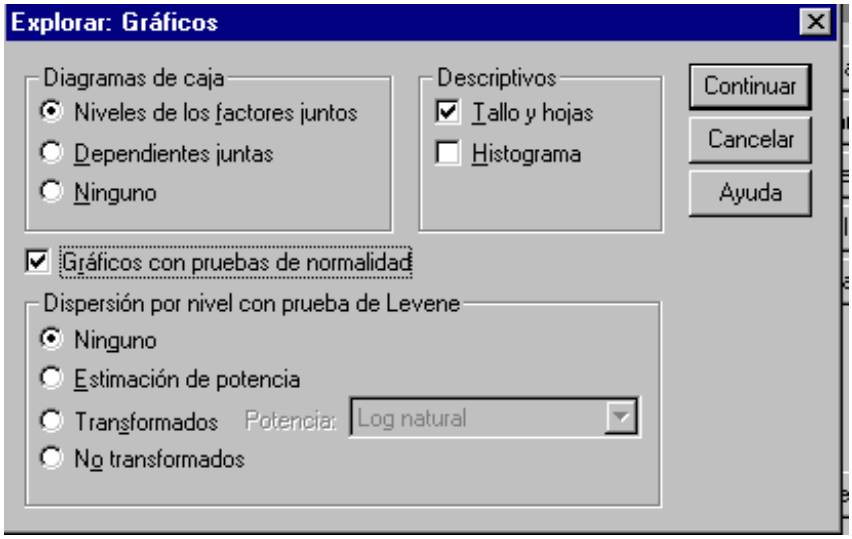
En ***Opciones*** nos muestra esta pantalla

El análisis exploratorio de datos es un conjunto de técnicas que se utilizan desde hace poco tiempo y que persiguen los mismos objetivos que la estadística descriptiva, pero incidiendo de forma especial en la detección de anomalías y errores en la distribución de las variables. Estos análisis se basan en análisis gráficos y en estadísticos robustos relacionados con el orden y la mediana. Esta opción permite realizar

A continuación se muestran las pantallas de SPSS relacionadas con el análisis exploratorio de los datos.

Descriptivos

	SEXO			Estadístico	Error típ.
EDAD	Hombre	Media		43,12	,40
		Intervalo de confianza para la media al 95%	Límite inferior	42,33	
			Límite superior	43,91	
		Media recortada al 5%		42,85	
		Mediana		43,00	
		Varianza		239,960	
		Desv. típ.		15,49	
		Mínimo		11	
		Máximo		84	
		Rango		73	
		Amplitud intercuartil		22,00	
		Asimetría		,200	,063
		Curtosis		-,669	,127
	Mujer	Media		36,53	,45
		Intervalo de confianza para la media al 95%	Límite inferior	35,64	
			Límite superior	37,41	
		Media recortada al 5%		35,74	
		Mediana		34,00	
		Varianza		195,239	
		Desv. típ.		13,97	
		Mínimo		9	
		Máximo		99	
		Rango		90	
		Amplitud intercuartil		17,00	
		Asimetría		,850	,079
		Curtosis		,439	,157
	9	Media		56,50	6,99
		Intervalo de confianza para la media al 95%	Límite inferior	38,52	
			Límite superior	74,48	
		Media recortada al 5%		56,28	
		Mediana		55,00	
		Varianza		293,500	
		Desv. típ.		17,13	
		Mínimo		41	
		Máximo		76	
		Rango		35	
		Amplitud intercuartil		31,25	
		Asimetría		,074	,845
		Curtosis		-3,110	1,741



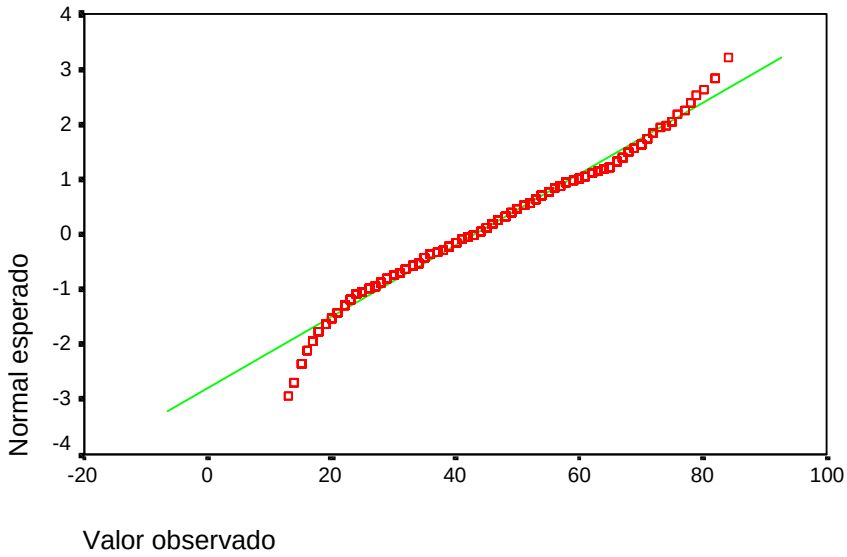
EDAD Stem-and-Leaf Plot for
SEXO= Hombre

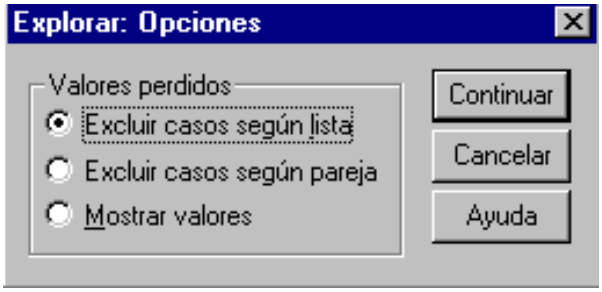
GRÁFICO DE TALLOS Y HOJAS

```
Frequency    Stem &  Leaf

     7,00        1 .  4&
    79,00        1 .  556667778889999
   130,00        2 .  000111112222223333344444
   116,00        2 .  5555666777788888999999
   136,00        3 .  00111112222222333344444444
   173,00        3 .  55555555556666777778888889999999999
   150,00        4 .  000000000111222222233344444444
   198,00        4 .  555555556666666666777777888888899999999
   157,00        5 .  000000000111112222223333344444444
   108,00        5 .  555555666677778888899
    65,00        6 .  000012223344
    89,00        6 .  555556667777888889
    50,00        7 .  00001122234
    28,00        7 .  55678&
     7,00        8 .  2&

 Stem width:  10
 Each leaf:       5 case(s)


 & denotes fractional leaves.
```

Gráfico Q-Q normal de EDAD

Para SEXO= Hombre





Algunos de los resultados que nos presenta SPSS, con las opciones elegidas, se :muestran a continuación

Resumen del procesamiento de los casos

		Casos					
		Válidos		Perdidos		Total	
	SEXO	N	Porcentaje	N	Porcentaje	N	Porcentaje
EDAD	Hombre	1493	96,0%	62	4,0%	1555	100,0%
	Mujer	963	95,3%	48	4,7%	1011	100,0%
	9	6	12,8%	41	87,2%	47	100,0%

DISTRIBUCIONES DE PROBABILIDAD Y CONTRASTES DE HIPÓTESIS

3. DISTRIBUCIONES DE PROBABILIDAD Y CONTRASTE DE HIPÓTESIS

3.1. INTRODUCCIÓN

3.2. DEFINICIÓN DE PROBABILIDAD, VARIABLE ALEATORIA Y VALOR ESPERADO

- **Definición de probabilidad**
- **Definición de variable aleatoria**
- **Definición de valor esperado: esperanza y varianza**

3.3. DISTRIBUCIONES DE PROBABILIDAD

- **Distribución binomial**
- **Distribución hipergeométrica**
- **Distribución normal o de Gauss**

3.4. DISTRIBUCIONES DERIVADAS DE LA NORMAL

- **Distribución χ^2 de Pearson**
- **Distribución t de Student**
- **Distribución F de Fisher-Snedecor**

3.5. TEOREMA CENTRAL DEL LÍMITE

3.6. DISTRIBUCIONES MUESTRALES

- **Distribución de la media muestral**
- **Distribución de la diferencia entre dos medias muestrales**
- **Distribución de la proporción muestral**
- **Distribución de la diferencia entre dos proporciones muestrales**
- **Distribución de la varianza muestral**
- **Distribución de la razón de varianzas muestrales**

3.7. INTERVALOS DE CONFIANZA

3.8. CONTRASTE DE HIPÓTESIS

3.9. DISTRIBUCIONES BIDIMENSIONALES

- **Distribuciones marginales**
- **Distribuciones condicionadas**
- **Dependencia lineal**

3.10. TABLAS DE CONTINGENCIA

- **Estadístico χ^2 de Pearson**

- **Medidas de asociación**
  - **Odds Ratio**
  - **Coeficiente de contingencia**
  - **Coeficiente V de Cramer**
  - **Q de Yule**

3.1. INTRODUCCIÓN

En este capítulo desarrollaremos los conceptos fundamentales de la teoría de la probabilidad, los contrastes de hipótesis, el análisis de la varianza y la teoría de la regresión. Dichos conceptos se juzgan claves a la hora de enfrentarnos con la teoría del muestreo, la cual se expone en el capítulo siguiente.

Uno de los objetivos de la ciencia consiste en describir y predecir sucesos que ocurren a nuestro alrededor de forma cotidiana. Una manera de hacerlo es mediante la construcción de modelos matemáticos. Así, las distribuciones de probabilidad se definen como modelos matemáticos que describen una infinidad de sucesos que acontecen a nuestro alrededor.

Si nos planteamos predecir el sexo de cada uno de los nacidos en un determinado lugar, vemos que la construcción de una ecuación que lo determine con total exactitud sería excesivamente compleja, hasta tal punto que no se ha descubierto aún ninguna, pero sin embargo, aproximarse al número total de nacidos de cada sexo si puede realizarse satisfactoriamente a través de un modelo matemático. Se exponen en este capítulo los modelos matemáticos más importantes que nos describen grupos de sucesos .

Asimismo, estos modelos matemáticos nos permitirán, como veremos, contrastar hipótesis realizadas sobre un grupo de sucesos.

3.2. DEFINICIÓN DE PROBABILIDAD, VARIABLE ALEATORIA Y VALOR ESPERADO

Definición de probabilidad

Se exponen a continuación las tres definiciones existentes de *probabilidad*, en orden al desarrollo histórico de la teoría:

a) *Probabilidad clásica o a priori*: es una probabilidad inicial. Por ejemplo, antes de lanzar una moneda al aire, se supone que la probabilidad de que salga cara va a ser igual a ½, de igual forma, podría decirse que si un dado se arroja muchas veces la probabilidad de obtener un *uno* o un *dos* será de 2/6, ya que este suceso puede aparecer 2 veces de 6. Esta probabilidad la definen Mood y Graybill (1978) de la siguiente manera:

> *si un suceso puede ocurrir de n maneras mutuamente excluyentes e igualmente verosímiles y si* n_A *de éstas poseen un atributo A, la probabilidad de A es la fracción* n_A / n.

b) *Probabilidad a posteriori o frecuencial*: es una probabilidad experimental. Por ejemplo, si sospechamos que un dado no está equilibrado, la única manera de probar esta sospecha es arrojándolo *muchas veces* y observar si la frecuencia relativa de obtener un uno se aproxima a un sexto.

La probabilidad a posteriori se define como *el límite de la frecuencia relativa* cuando el número de experimentos realizados tiende a infinito, y se enuncia formalmente de la siguiente manera:

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n}$$

donde A sería el suceso obtener un uno y:

n → número de veces que se repite el experimento (lanzamiento del dado)

n_A → número de veces que aparece el resultado A.

$\frac{n_A}{n}$ → denota, por tanto, la frecuencia relativa

lim $\rightarrow$ denota el límite de la frecuencia relativa a medida que el número de lanzamientos se aproxima a infinito.

La probabilidad frecuencial parte del supuesto de que los distintos posibles resultados o sucesos que pueden derivarse de un experimento no tienen por que ser igualmente verosímiles.

c) De las definiciones anteriores se deducen tres axiomas que son los que constituyen la definición axiomática de la probabilidad.

Sea *S* un espacio muestral (conjunto de todos los posibles sucesos de un determinado experimento) y *A* un determinado suceso de *S* (cualquier elemento o subconjunto de *S*), diremos que *P* es una *función de probabilidad* en el espacio muestral *S* si se satisfacen los tres axiomas siguientes:

Axioma 1. *P(A)* es un número real tal que $P(A) \geq 0$ para todo suceso *A* de S, es decir, la probabilidad de cualquier suceso en un experimento es siempre mayor o igual que 0.

Axioma 2. $P(S) = 1$, es decir, la probabilidad de todos los sucesos posibles de un experimento es igual a 1.

Axioma 3. Si *A*, *B*, *C*,… es una sucesión de sucesos mutuamente excluyentes de S, la probabilidad asociada a la unión de todos ellos (que en un experimento ocurra cualquiera de ellos) es igual a la suma de sus probabilidades.

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

De estos tres axiomas se deducen los siguientes teoremas:

Teorema 1: Si definimos suceso complementario de *A, A'*, como aquel que está formado por todos los puntos o sucesos del espacio muestral S que no están en *A*, entonces la probabilidad de *A'* será igual a:

$$P(A') = 1 - P(A)$$

ya que $P(A \cup A') = P(S) = P(A) + P(A') = 1 \Rightarrow P(A') = 1 - P(A)$

Teorema 2. Sea A un suceso de S. Entonces se verifica:

$$0 \leq P(A) \leq 1$$

ya que por el *axioma 1* $p(A) \geq 0$ y por el teorema anterior sabemos que:
$P(A) + P(A') = 1$
siendo por el *axioma 1* $P(A)$ y $P(A') \geq 0 \Rightarrow P(A) = 1 - P(A') \leq 1$

Teorema 3. Si ϕ es el suceso nulo, entonces se verifica que:

$$P(\phi) = 0$$

ya que ϕ es el suceso complementario de S.

Señalar por último que el conjunto de todos los sucesos posibles, *S*, puede ser:

- *Discreto*: si toma solamente un número finito o numerable de valores.
- *Continuo*: puede tomar cualesquiera de los infinitos valores de un intervalo.

Definición de variable aleatoria

Sea *S* un espacio muestral en el que se define una función de probabilidad. Sea *X* una función de valores reales definida en *S*. Si la función *X* transforma puntos de *S* en

puntos del eje *X* y es medible, entonces se dice que *X* es una variable aleatoria (variable aleatoria unidimensional). Una variable aleatoria es, por tanto, una regla o mecanismo que asigna un valor numérico a todos y cada uno de los sucesos asociados a un experimento.

Se muestra a continuación un ejemplo sencillo de variable aleatoria.

Supongamos que se lanzan al aire tres monedas, entonces los sucesos posibles son los siguientes:

CCC XCC CXC CCX CXX XCX XXC XXX

La probabilidad de cada uno de estos sucesos es igual a 1/8. Si asignamos un valor numérico a sacar una cara, por ejemplo, un 1, y un valor numérico a sacar una cruz, por ejemplo un 0, estamos construyendo la siguiente variable aleatoria:

S	*X(S)*
CCC	3
XCC	2
CXC	2
CCX	2
CXX	1
XCX	1
XXC	1
XXX	0

La *función de probabilidad de la variable aleatoria* sería entonces la siguiente:

$$P(X = 0) = P(X(S) = 0) = P(XXX) = 1/8$$
$$P(X = 1) = P(X(S) = 1) = P(CXX) + P(XCX) + P(XXC) = 1/8 + 1/8 + 1/8 = 3/8$$

P(X = 2) = P(X(S) = 2) = P(XCC) + P(CXC) + P(CCX) = 1/8 + 1/8 + 1/8 = 3/8
P(X = 3) = P(X(S) = 3) = P(CCC) = 1/8

Señalar que una variable aleatoria se dice *discreta* si toma solamente un número finito o numerable de valores y *continua* si puede tomar cualesquiera de los infinitos valores de un intervalo.

Definición de valor esperado

El *valor esperado, esperanza o media de una variable aleatoria* se obtiene calculando el valor medio de la distribución de valores de la variable aleatoria.

En el ejemplo anterior, el valor esperado de la variable aleatoria construida sería:

$$E(X) = 0 * \frac{1}{8} + 1 * \frac{3}{8} + 2 * \frac{3}{8} + 3 * \frac{1}{8} = 1{,}5$$

Este valor esperado o esperanza de la variable aleatoria se expresa formalmente de la siguiente manera:

$$E(X) = \sum_{-\infty}^{+\infty} x * f(x) \qquad \text{si la variable aleatoria es discreta}$$

$$E(X) = \int_{-\infty}^{\infty} x * f(x)dx \qquad \text{si la variable aleatoria es continua}$$

siendo *f(x)* la función de probabilidad y denotando el intervalo (-∞,+∞) el conjunto de todos los posibles valores que toma la variable aleatoria.

El *valor esperado, esperanza o media de una función de una variable aleatoria* se obtiene calculando el valor medio de la distribución de valores de la función de la variable aleatoria. Formalmente toma la siguiente expresión:

$$E(h(X)) = \sum_{-\infty}^{+\infty} h(x) * f(x) \qquad \text{si la variable aleatoria es discreta}$$

$$E(h(X)) = \int_{-\infty}^{\infty} h(x) * f(x)dx$$ si la variable aleatoria es continua

Nótese que si h(x) = x estaríamos calculando el *valor esperado, esperanza o media de la variable aleatoria*

Si calculamos el valor para h*(x)* = x^2 estaríamos calculando el valor esperado, esperanza o media de la variable aleatoria al cuadrado, y si a este valor le restamos $(E(X))^2$, obtendremos la varianza de la variable aleatoria, es decir:

$$Var(X) = \sum_{-\infty}^{+\infty} x^2 f(x) - E(X)^2$$ si la variable aleatoria es discreta

$$Var(X) = \int_{-\infty}^{\infty} x^2 f(x)dx - E(X)^2$$ si la variable aleatoria es continua

Continuando con el ejemplo anterior, la varianza de la variable aleatoria sería:

$$Var(X) = \sum_{-\infty}^{+\infty} x^2 P(x) - E(X)^2 = \left(0^2 * \frac{1}{8} + 1^2 * \frac{3}{8} + 2^2 * \frac{3}{8} + 3^2 * \frac{1}{8}\right) - \left(\frac{1}{5}\right)^2 = 0{,}75$$

3.3. DISTRIBUCIONES DE PROBABILIDAD

Se denomina distribución de probabilidad a cualquier regla o mecanismo que nos permita determinar la probabilidad de que la variable aleatoria ξ tome cada uno de los posibles valores *x*. Esta regla o mecanismo puede ser una tabla, una fórmula o un gráfico. La función de probabilidad, será la fórmula que se emplee para calcular *P[ξ=x]*.

Cualquier distribución de probabilidad, como se deduce del apartado anterior, ha de tener dos características necesarias:

1) $P(\xi \leq x) \geq 0 \quad \forall x$ si es discreta o bien $\int_{-\infty}^{x} f(x)dx \geq 0 \quad \forall x$ si es continua.

2) $\sum P(\xi = x) = 1$ si es discreta o bien $\int_{-\infty}^{\infty} f(x)dx = 1$ si es continua

La función de distribución es la probabilidad de que la variable aleatoria tome todos los valores menores o iguales a *x*.

Las distribuciones de probabilidad pueden ser discretas o continuas. No obstante, podríamos definir funciones de probabilidad mixtas, es decir, en unos tramos discretas y en otros continuas.

Las distribuciones de probabilidad discretas que vamos a analizar en este capítulo son: la distribución binomial y la distribución hipergeométrica. Entre las distribuciones continuas se estudian la distribución normal o de Gauss, la distribución χ^2 de Pearson, la distribución t de Student y la distribución F de Fisher-Snedecor.

Distribución binomial

Para comenzar el estudio de la distribución binomial vamos a considerar una variable aleatoria ξ_i que puede tomar únicamente los valores **1 y 0** con probabilidades **p y q** respectivamente.

$$P[\xi_i = 1] = p \qquad P[\xi_i = 0] = q \qquad p + q = 1$$

Ejemplos concretos de estos fenómenos aleatorios son: el lanzamiento de una moneda a cara o cruz, los resultados de un examen de aprobado o suspenso o el lanzamiento de un dado con posibles resultados de sólo par o impar.

Estos experimentos donde se producen resultados mutuamente excluyentes se denominan ensayos de Bernouilli, en honor al matemático suizo Jakob Bernouille (1654-1705).

Las condiciones que se deben de satisfacer son:

1. La probabilidad de éxito *p* permanece constante de un experimento a otro.
2. Los ensayos son independientes.

La esperanza matemática o la media de esta distribución es :

$$E(\xi_i) = 1p + 0q = p$$

La varianza se calcula mediante la siguiente expresión:

$$\sigma^2 = E(\xi_i - p)^2 = pq$$

siendo $q = 1-p$

La distribución binomial de parámetros n , p se basa en una prueba conocida como experimento de Bernouilli o problema de las pruebas repetidas, que consiste en averiguar la probabilidad de que en las n extracciones o pruebas se hayan conseguido x valores *1* o/y $n-x$ valores *0*.

La distribución binomial de parámetros n , p se construye, pues, como una suma de variables independientes distribuidas como las anteriores. La variable ξ puede tomar todos los valores comprendidos entre *0* y *n*

$$\xi = \xi_1 + \xi_2 + \ldots\ldots + \xi_n$$

La función de cuantía o de probabilidad viene expresada por la siguiente función:

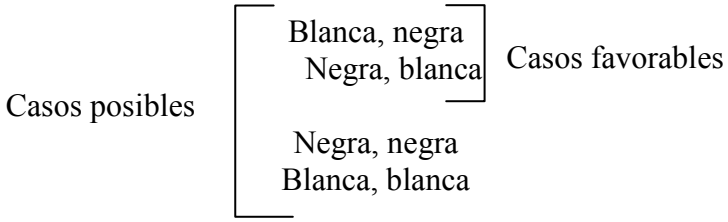
$$P[\xi = x] = \binom{n}{x} p^x q^{n-x}$$

Para facilitar la comprensión de esta función nos vamos a apoyar en el siguiente ejemplo. Supóngase la extracción de *r* bolas en una urna, de las cuales n_1 son blancas y n_2 negras, siendo el suceso a medir el número de bolas blancas extraídas. Cada vez que efectuemos una extracción se volverá a introducir la bola dentro de la urna.

La función de cuantía o probabilidad tiene una sencilla deducción en el ejemplo expuesto. En primer lugar, dado que partimos de sucesos independientes, la probabilidad se obtiene multiplicando las probabilidades de los sucesos, es decir, si la urna contiene únicamente cuatro bolas, dos blancas y dos negras, y efectuamos dos extracciones, la probabilidad de que una sea blanca y la otra no, siendo la probabilidad de obtener bola blanca *p = 0,5* (ya que tenemos dos bolas blancas sobre cuatro), sería:

$$p(blanca, negra) = p(blanca) * p(negra) = p(1-p) = pq=0,25$$

Ahora bien, hemos de tener en cuenta que el orden no influye y, por tanto, obtenemos dos casos favorables sobre los cuatro posibles, teniendo cada uno probabilidad *pq*:



Al ser dos los casos favorables se deberá multiplicar *pq* por 2, es decir la probabilidad del suceso a evaluar es la siguiente: *2pq*.

El número de estos casos favorables se calcula a través del número combinatorio $\binom{n}{x}$, siendo en el caso que nos ocupa n = 2 y x = 1.

La esperanza matemática de la distribución es:

$$E(\xi) = np$$

La varianza de la distribución es:

$$\sigma^2 = npq$$

Por ejemplo, la probabilidad de obtener *x* caras en 10 lanzamientos de una moneda será igual a:

$$P[\xi = x] = \binom{n}{x} p^x q^{10-x} = \binom{10}{x} 0{,}5^x 0{,}5^{10-x} \text{ siendo } 0 \leq x \leq 10$$

Si *x = 0* estaríamos calculando la probabilidad de obtener 10 cruces o ninguna cara, que sería igual a 0,00097. A su vez, la probabilidad de obtener 5 caras y 5 cruces (*x = 5*) sería de 0,24609.

Esta distribución se aplica a sondeos exhaustivos con reemplazamiento, constituyendo la base teórica de las formulaciones desarrolladas en el muestreo aleatorio con reemplazamiento.

Distribución hipergeométrica

Una variable aleatoria, ξ, que toma todos los valores comprendidos entre *0* y *n*, se dice que sigue una distribución hipergeométrica cuando:

$$P\left[\xi_n = r\right] = \frac{\binom{Np}{r}\binom{Nq}{n-r}}{\binom{N}{r}}$$

donde *Np* y *Nq* son números enteros, siendo *Np* + *Nq* = *N*.

Un ejemplo de esta distribución lo encontramos cuando queremos saber cual es la probabilidad de extraer de una urna que contiene *N* bolas, de las cuales n_1 (*Np* en la fórmula) son blancas y n_2 (*Nq*) son negras, *r* bolas blancas y *n-r* bolas negras al hacer *n* extracciones. Cada vez que se efectúe una extracción, la bola no se repone de nuevo en la urna, es decir, no entrará a formar parte de la siguiente extracción.

La esperanza matemática de la distribución es: $E[\xi_n] = np$

La varianza de la distribución viene dada por la siguiente expresión:

$$\sigma^2 = \frac{N-n}{N-1} npq$$

Esta distribución es la base teórica del muestreo aleatorio sin reposición.

Distribución normal o de Gauss

En un buen número de sucesos aleatorios, la distribución de probabilidad, sigue una distribución específica en forma de campana, llamada curva normal o curva de

Gauss[1]. Esta distribución es la más común y útil de la estadística, dado que muchos fenómenos se suelen ajustar a esta distribución: errores de observación, procesos de medición sin errores sistemáticos, medidas físicas del cuerpo humano, etc.

Decimos que una variable aleatoria ξ que toma los valores *x* (desde -∞ hasta +∞) se distribuye normalmente con parámetros (0, 1), es decir, con media 0 y varianza 1, cuando su función de distribución viene dada por la siguiente expresión:

$$P[\xi \leq x] = F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}x^2} dx \qquad -\infty < x < \infty$$

La función de densidad la obtenemos derivando la función de distribución:

$$f(X) = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}x^2}$$

A su vez, decimos que una variable aleatoria η se distribuye normalmente con parámetros (α, σ) cuando está ligada a la distribución normal de parámetros (0, 1) por la siguiente expresión:

$$\eta = \sigma\xi + \alpha \quad \text{siendo } \sigma > 0$$

La función de densidad de la distribución normal de parámetros (α, σ) toma, entonces, la siguiente expresión:

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\alpha}{\sigma}\right)^2}$$

La representación gráfica de la función de densidad de la distribución normal de parámetros (α, σ), tiene las siguientes características:

[1] Karl Friedrich Gauss, 1977-1855, investigó el comportamiento de los errores de medida y llegó a la expresión matemática que se conoce como Ley de los errores o Ley de Gauss.

a) $y = 0$ asíntota para $x \rightarrow \infty$ y $x \rightarrow -\infty$

b) Simetría respecto a $x = \alpha$

c) Creciente cuando $x < \alpha$

d) Decreciente cuando $x > \alpha$

e) Hay un Máximo en $x = \alpha$

Gráfico 3.1.

Función de densidad y de distribución de la normal (0,1)

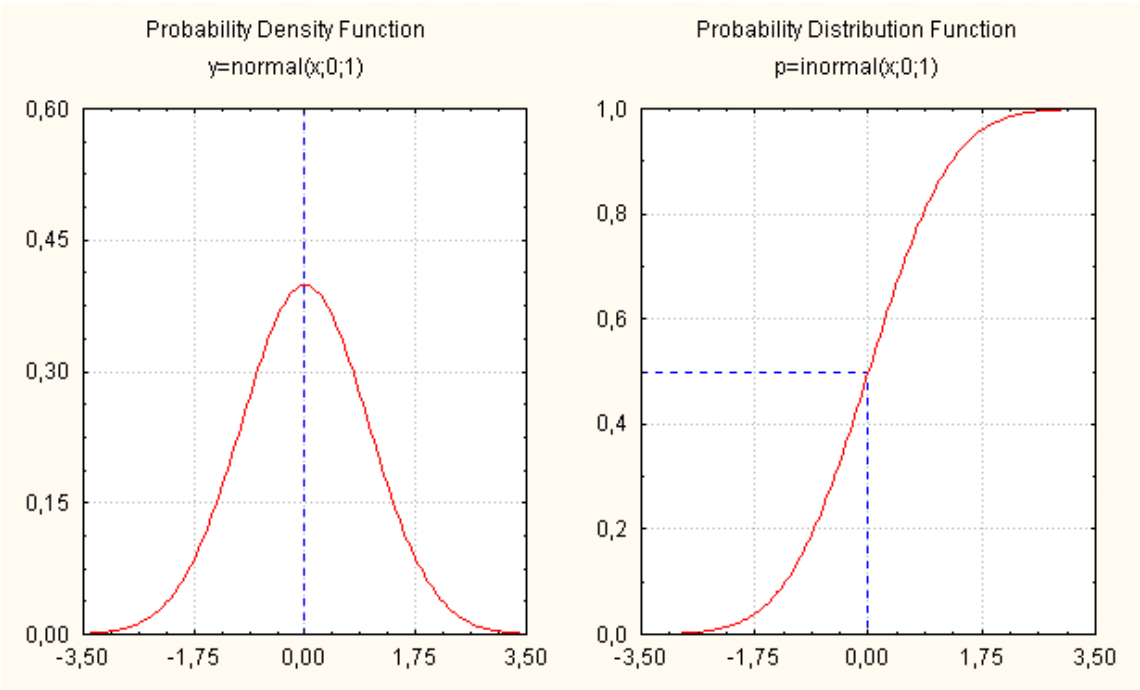


Gráfico 3.2.
Probabilidad comprendida entre los valores -1 y 1 en una distribución normal de parámetros (0,1)

En el gráfico 3.1. hemos representado las funciones de densidad y de probabilidad de una distribución normal con parámetros (0, 1). En el gráfico 3.2. aparece la probabilidad de que la variable aleatoria ξ distribuida como una normal con parámetros (0, 1) esté comprendida entre los valores 1 y -1, probabilidad que es igual a 0,6823, que representa el 68,23%.

Estas probabilidades pueden calcularse utilizando tablas estadísticas construidas al efecto. Las tablas de la distribución normal figuran en el Anexo II. Así, si queremos responder a la pregunta ¿cuál es la probabilidad de que un valor sacado al azar, de una población que sigue una distribución normal de media 0 y varianza 1, esté comprendida entre -2,05 y 0?, debemos, en virtud del carácter simétrico de la función de densidad, buscar en la tabla la probabilidad correspondiente al intervalo de valores $-\infty$ y 2,05. Esta probabilidad es igual a 0,979820. Dado que hay que descontar la probabilidad correspondiente al intervalo de valores que va de $-\infty$ a 0, es decir, la mitad de la distribución, el resultado final será:

$$p(-2,05 \leq \xi \leq 0) = p(0 \leq \xi \leq 2,5) = p(\xi \leq 2,5) - p(\xi \leq 0) = 0,979820 - 0,5 = 0,479820$$

A continuación se ofrece una tabla de la distribución normal con la probabilidad que corresponde a diversos valores de la variable a contrastar:

Tabla 3.1.

Distribución normal estándar acumulativa

Valor	Probabilidad	Valor	Probabilidad	Valor	Probabilidad
0,0	0,50000000	**1,5**	0,93319277	**3,0**	0,99865003
0,1	0,53982790	**1,6**	0,94520071	**3,1**	0,99903233
0,2	0,57925969	**1,7**	0,95543457	**3,2**	0,99931280
0,3	0,61791136	**1,8**	0,96406973	**3,3**	0,99951652
0,4	0,65542170	**1,9**	0,97128351	**3,4**	0,99966302
0,5	0,69146247	**2,0**	0,97724994	**3,5**	0,99976733
0,6	0,72574694	**2,1**	0,98213564	**3,6**	0,99984085
0,7	0,75803642	**2,2**	0,98609660	**3,7**	0,99989217
0,8	0,78814467	**2,3**	0,98927592	**3,8**	0,99992763
0,9	0,81593991	**2,4**	0,99180247	**3,9**	0,99995188
1,0	0,84134474	**2,5**	0,99379032	**4,0**	0,99996831
1,1	0,86433390	**2,6**	0,99533878	**4,1**	0,99997933
1,2	0,88493027	**2,7**	0,99653298	**4,2**	0,99998665
1,3	0,90319945	**2,8**	0,99744481	**4,3**	0,99999145
1,4	0,91924329	**2,9**	0,99813412	**4,4**	0,99999458

3.4. DISTRIBUCIONES DERIVADAS DE LA NORMAL

Las distribuciones que a continuación se comentan se obtienen como combinaciones de funciones de variables aleatorias independientes que siguen una distribución normal. Las distribuciones derivadas de la normal que se explican son la distribución χ^2 de Pearson, la distribución t de Student y la distribución F de Fisher-Snedecor.

Distribución χ^2 de Pearson.

Consideramos la siguiente variable $\chi_n^2 = \eta_1^2 + ... + \eta_n^2$ donde las variables η_i son distribuciones normales e independientes. El número de distribuciones normales utilizadas para construir la variable χ^2 recibe el nombre de grados de libertad.

La función de densidad de la nueva variable así definida viene dada por la siguiente expresión:

$$\chi_n^2(x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{1}{2}x}$$

La esperanza matemática es igual a: $E(\chi_n^2) = n$

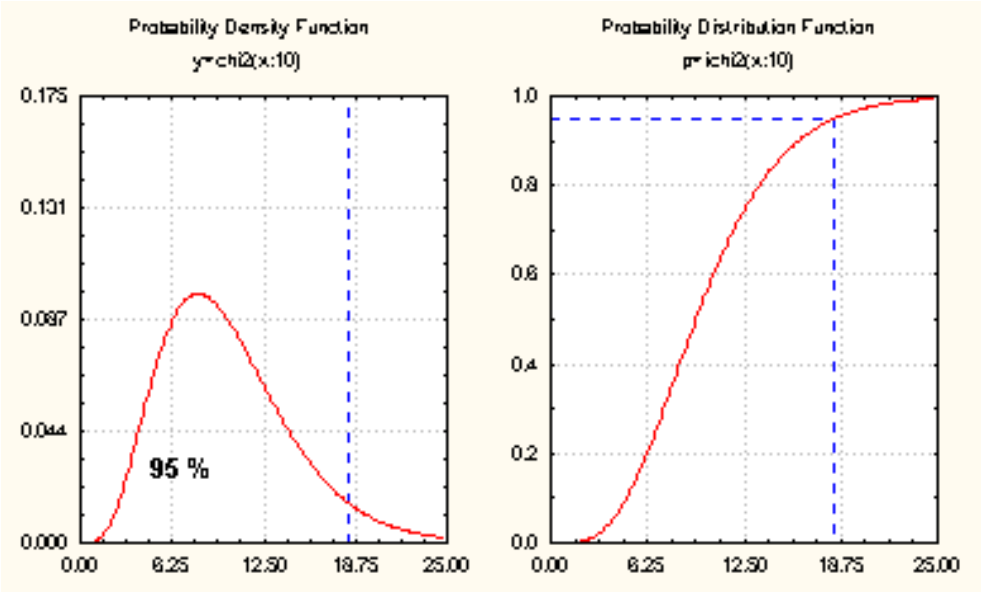
La varianza de la variable aleatoria es: $Var(\chi_n^2) = 2n$

La distribución χ^2 de Pearson es asimétrica (ver gráfico 3.3.). Su propiedad fundamental es que si sumamos dos χ^2 independientes de grados de libertad n_1 y n_2, se obtiene una nueva variable χ^2 con grados de libertad igual a la suma de n_1 y n_2.

Esta propiedad aditiva posibilita los contrastes de hipótesis que se explican más adelante, así como también la combinación de varios estadígrafos o de otros valores en el mismo contraste.

Gráfico 3.3.

Función de densidad y de distribución de la χ^2



En la tabla siguiente se especifican los valores *x* tales que la probabilidad de que la variable aleatoria *sea mayor* es igual a *p*.

Tabla 3.2.

Distribución Chi-cuadrado de Pearson

$$\Pr(X_n^2 > x) = p$$

VALORES DE	1
REFERENCIA	0,05

=PRUEBA.CHI.INV(B2;B1) **P**

Grados de libertad	**0,005**	**0,010**	**0,025**	**0,05**	**0,10**	**0,25**	**0,50**
1	7,88	6,63	5,02	3,84	2,71	1,32	0,45
2	10,60	9,21	7,38	5,99	4,61	2,77	1,39
3	12,84	11,34	9,35	7,81	6,25	4,11	2,37
4	14,86	13,28	11,14	9,49	7,78	5,39	3,36
5	16,75	15,09	12,83	11,07	9,24	6,63	4,35
6	18,55	16,81	14,45	12,59	10,64	7,84	5,35
7	20,28	18,48	16,01	14,07	12,02	9,04	6,35
8	21,95	20,09	17,53	15,51	13,36	10,22	7,34
9	23,59	21,67	19,02	16,92	14,68	11,39	8,34
10	25,19	23,21	20,48	18,31	15,99	12,55	9,34
11	26,76	24,73	21,92	19,68	17,28	13,70	10,34
12	28,30	26,22	23,34	21,03	18,55	14,85	11,34
13	29,82	27,69	24,74	22,36	19,81	15,98	12,34
14	31,32	29,14	26,12	23,68	21,06	17,12	13,34
15	32,80	30,58	27,49	25,00	22,31	18,25	14,34
16	34,27	32,00	28,85	26,30	23,54	19,37	15,34
17	35,72	33,41	30,19	27,59	24,77	20,49	16,34
18	37,16	34,81	31,53	28,87	25,99	21,60	17,34
19	38,58	36,19	32,85	30,14	27,20	22,72	18,34
20	40,00	37,57	34,17	31,41	28,41	23,83	19,34
21	41,40	38,93	35,48	32,67	29,62	24,93	20,34
22	42,80	40,29	36,78	33,92	30,81	26,04	21,34
23	44,18	41,64	38,08	35,17	32,01	27,14	22,34
24	45,56	42,98	39,36	36,42	33,20	28,24	23,34
25	46,93	44,31	40,65	37,65	34,38	29,34	24,34
26	48,29	45,64	41,92	38,89	35,56	30,43	25,34
27	49,65	46,96	43,19	40,11	36,74	31,53	26,34
28	50,99	48,28	44,46	41,34	37,92	32,62	27,34
29	52,34	49,59	45,72	42,56	39,09	33,71	28,34
30	53,67	50,89	46,98	43,77	40,26	34,80	29,34
40	66,77	63,69	59,34	55,76	51,81	45,62	39,34
50	79,49	76,15	71,42	67,50	63,17	56,33	49,33
60	91,95	88,38	83,30	79,08	74,40	66,98	59,33
70	104,21	100,43	95,02	90,53	85,53	77,58	69,33
80	116,32	112,33	106,63	101,88	96,58	88,13	79,33
90	128,30	124,12	118,14	113,15	107,57	98,65	89,33
100	140,17	135,81	129,56	124,34	118,50	109,14	99,33

Distribución t de Student.

Esta distribución fue obtenida por el método de Montecarlo en 1908 por el químico W.S. Gosset.

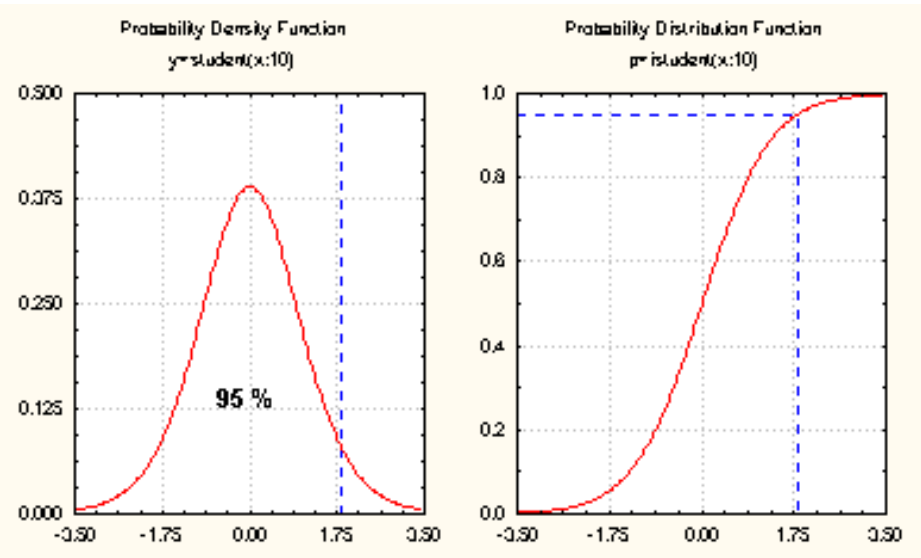
Consideramos la variable $t = \dfrac{\eta}{\sqrt{\dfrac{1}{n}\left(\eta_1^2 + \eta_2^2 + \ldots + \eta_n^2\right)}}$ donde las variables que intervienen, η_i y η son *independientes y normales (0, 1)*. La función de densidad de la variable t viene dada por la siguiente expresión**:**

$$t_n(x) = \frac{\Gamma\left(\dfrac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\dfrac{n}{2}\right)}\left[1 + \frac{x^2}{n}\right]^{-\frac{n+1}{2}}$$

Esta distribución es simétrica presentando mayor dispersión que la curva normal estándar para valores pequeños de *n*. A medida que *n* aumenta ($n > 100$) es prácticamente igual que la normal.

Gráfico 3.4.

Función de densidad y de distribución de la t de Student



En la tabla siguiente se presentan los valores donde $\Pr(|T| < x) = p$

Tabla 3.3.

Distribución t de Student (dos colas)

VALORES DE	1
REFERENCIA	0,05

$$\Pr(|T| < x) = p$$

=DISTR.T.INV(B3;B2)

	0,005	**0,010**	**0,025**	**0,05**	**0,10**	**0,25**	**0,50**
1	127,32	63,66	25,45	12,71	6,31	2,41	1,00
2	14,09	9,92	6,21	4,30	2,92	1,60	0,82
3	7,45	5,84	4,18	3,18	2,35	1,42	0,76
4	5,60	4,60	3,50	2,78	2,13	1,34	0,74
5	4,77	4,03	3,16	2,57	2,02	1,30	0,73
6	4,32	3,71	2,97	2,45	1,94	1,27	0,72
7	4,03	3,50	2,84	2,36	1,89	1,25	0,71
8	3,83	3,36	2,75	2,31	1,86	1,24	0,71
9	3,69	3,25	2,69	2,26	1,83	1,23	0,70
10	3,58	3,17	2,63	2,23	1,81	1,22	0,70
11	3,50	3,11	2,59	2,20	1,80	1,21	0,70
12	3,43	3,05	2,56	2,18	1,78	1,21	0,70
13	3,37	3,01	2,53	2,16	1,77	1,20	0,69
14	3,33	2,98	2,51	2,14	1,76	1,20	0,69
15	3,29	2,95	2,49	2,13	1,75	1,20	0,69
16	3,25	2,92	2,47	2,12	1,75	1,19	0,69
17	3,22	2,90	2,46	2,11	1,74	1,19	0,69
18	3,20	2,88	2,45	2,10	1,73	1,19	0,69
19	3,17	2,86	2,43	2,09	1,73	1,19	0,69
20	3,15	2,85	2,42	2,09	1,72	1,18	0,69
21	3,14	2,83	2,41	2,08	1,72	1,18	0,69
22	3,12	2,82	2,41	2,07	1,72	1,18	0,69
23	3,10	2,81	2,40	2,07	1,71	1,18	0,69
24	3,09	2,80	2,39	2,06	1,71	1,18	0,68
25	3,08	2,79	2,38	2,06	1,71	1,18	0,68
26	3,07	2,78	2,38	2,06	1,71	1,18	0,68
27	3,06	2,77	2,37	2,05	1,70	1,18	0,68
28	3,05	2,76	2,37	2,05	1,70	1,17	0,68
29	3,04	2,76	2,36	2,05	1,70	1,17	0,68
30	3,03	2,75	2,36	2,04	1,70	1,17	0,68
40	2,97	2,70	2,33	2,02	1,68	1,17	0,68
50	2,94	2,68	2,31	2,01	1,68	1,16	0,68
60	2,91	2,66	2,30	2,00	1,67	1,16	0,68
70	2,90	2,65	2,29	1,99	1,67	1,16	0,68
80	2,89	2,64	2,28	1,99	1,66	1,16	0,68
90	2,88	2,63	2,28	1,99	1,66	1,16	0,68
100	2,87	2,63	2,28	1,98	1,66	1,16	0,68

Distribución F de Fisher- Snedecor

Consideramos ahora $n + m$ variables aleatorias independientes y normalmente distribuidas con parámetros (0, σ). Definimos la variable F como:

$$F = \frac{\frac{1}{m}\left(\eta_1^2 + \ldots + \eta_m^2\right)}{\frac{1}{n}\left(\eta_1'^2 + \ldots + \eta_n'^2\right)} = \frac{Y_m}{Y_n}$$

La distribución de probabilidad de la variable F tiene la siguiente función de densidad:

$$F_{n,m} = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \frac{x^{\frac{m}{2}-1}}{[1+x]^{\frac{(m+n)}{2}}}$$

Las curvas de densidad dependen de n y de m, grados de libertad del numerador y del denominador. Por definición se verifica que $F_{n,m} = F^{-1}{}_{m,n}$. En el Anexo II aparecen los valores que toma la distribución F para los diferentes grados de libertad. Se muestra a continuación un ejemplo de una tabla con los valores que toma la distribución para los percentiles 50, 75, 90, 95, 97,5, 99 y 99,5 con grados de libertad que van de 1 a 6 tanto en numerador como denominador.

Tabla 3.4.

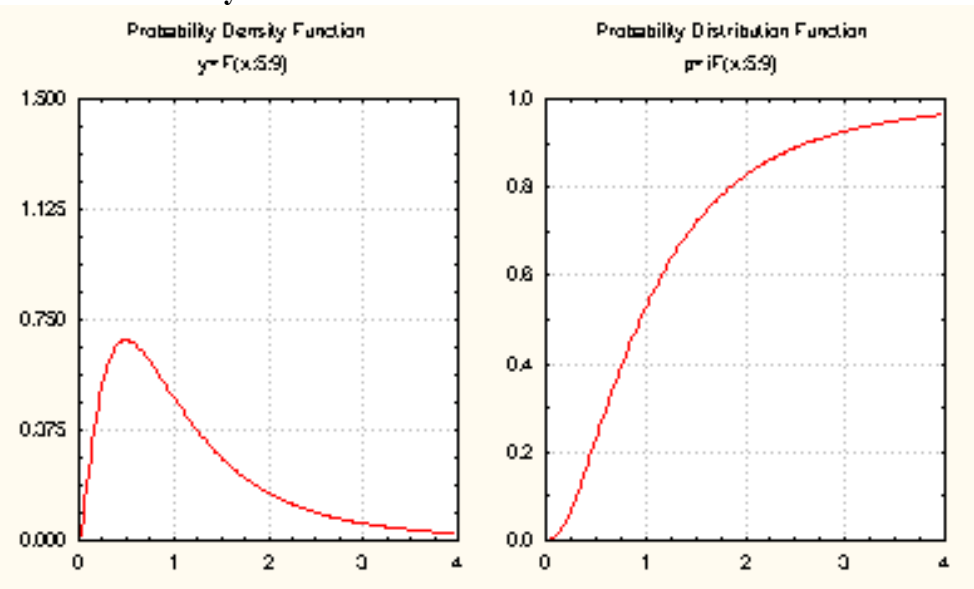
Distribución F *de Fisher-Snedecor*

$$\Pr(F < x) = p$$

		n					
	p	1	2	3	4	5	6
1	0,500	1,0000	1,5000	1,7092	1,8227	1,8937	1,9422
	0,750	5,8284	7,5000	8,1999	8,5809	8,8198	8,9832
	0,900	39,8636	49,5002	53,5933	55,8330	57,2400	58,2045
	0,950	161,4462	199,4995	215,7067	224,5833	230,1604	233,9875
	0,975	647,7931	799,4822	864,1509	899,5994	921,8347	937,1142
	0,990	4052,1845	4999,3396	5403,5336	5624,2570	5763,9554	5858,9503
	0,995	16212,4634	19997,3583	21614,1343	22500,7534	23055,8217	23439,5266
2	0,500	0,6667	1,0000	1,1349	1,2071	1,2519	1,2824
	0,750	2,5714	3,0000	3,1534	3,2321	3,2799	3,3121
	0,900	8,5263	9,0000	9,1618	9,2434	9,2926	9,3255
	0,950	18,5128	19,0000	19,1642	19,2467	19,2963	19,3295
	0,975	38,5062	39,0000	39,1656	39,2483	39,2984	39,3311
	0,990	98,5019	99,0003	99,1640	99,2513	99,3023	99,3314
	0,995	198,5027	199,0120	199,1575	199,2448	199,3030	199,3321
3	0,500	0,5851	0,8811	1,0000	1,0632	1,1024	1,1289
	0,750	2,0239	2,2798	2,3555	2,3901	2,4095	2,4218
	0,900	5,5383	5,4624	5,3908	5,3427	5,3091	5,2847
m	0,950	10,1280	9,5521	9,2766	9,1172	9,0134	8,9407
	0,975	17,4434	16,0442	15,4391	15,1010	14,8848	14,7347
	0,990	34,1161	30,8164	29,4567	28,7100	28,2371	27,9106
	0,995	55,5519	49,8003	47,4683	46,1951	45,3911	44,8381
4	0,500	0,5486	0,8284	0,9405	1,0000	1,0367	1,0617
	0,750	1,8074	2,0000	2,0467	2,0642	2,0723	2,0766
	0,900	4,5448	4,3246	4,1909	4,1072	4,0506	4,0097
	0,950	7,7086	6,9443	6,5914	6,3882	6,2561	6,1631
	0,975	12,2179	10,6490	9,9792	9,6045	9,3645	9,1973
	0,990	21,1976	17,9998	16,6942	15,9771	15,5219	15,2068
	0,995	31,3321	26,2844	24,2599	23,1539	22,4563	21,9752
5	0,500	0,5281	0,7988	0,9071	0,9646	1,0000	1,0240
	0,750	1,6925	1,8528	1,8843	1,8927	1,8947	1,8945
	0,900	4,0604	3,7797	3,6195	3,5202	3,4530	3,4045
	0,950	6,6079	5,7861	5,4094	5,1922	5,0503	4,9503
	0,975	10,0069	8,4336	7,7636	7,3879	7,1464	6,9777
	0,990	16,2581	13,2741	12,0599	11,3919	10,9671	10,6722
	0,995	22,7847	18,3136	16,5301	15,5560	14,9394	14,5133
6	0,500	0,5149	0,7798	0,8858	0,9419	0,9765	1,0000
	0,750	1,6214	1,7622	1,7844	1,7872	1,7852	1,7821
	0,900	3,7760	3,4633	3,2888	3,1808	3,1075	3,0546
	0,950	5,9874	5,1432	4,7571	4,5337	4,3874	4,2839
	0,975	8,8131	7,2599	6,5988	6,2271	5,9875	5,8197
	0,990	13,7452	10,9249	9,7796	9,1484	8,7459	8,4660
	0,995	18,6346	14,5442	12,9166	12,0276	11,4637	11,0731

Gráfico 3.5.

Función de densidad y de distribución de la F de Fisher-Snedecor



3.5. TEOREMA CENTRAL DEL LÍMITE

El Teorema Central del Límite demuestra que dado un conjunto de variables aleatorias independientes, $X_1, X_2, \ldots, X_n$, distribuidas con media μ_i y varianza σ_i^2, la variable suma:

$$Y = X_1 + X_2 + \ldots + X_n$$

cuando el número de variables (*n)* crece, tiende a una distribución normal con parámetros ($\sum \mu_i$, $\sum \sigma_i^2$), y por tanto la variable tipificada:

$$Z = \frac{Y - \sum \mu_i}{\sqrt{\sum \sigma_i^2}}$$

tiende a una distribución normal con parámetros (0, 1).

Este teorema es de vital importancia porque justifica que en la práctica variables aleatorias de las que no conocemos su distribución de frecuencias puedan ser aproximadas a una distribución normal siempre y cuando *n* sea suficientemente grande.

3.6. DISTRIBUCIONES MUESTRALES

Se llama distribución muestral a la distribución de probabilidad de un estadístico muestral que ha sido calculado a partir de todas las muestras posibles de tamaño *n* que han sido elegidas al azar.

Si la población es finita podemos calcular una distribución muestral experimental. Para ello, procederíamos del siguiente modo:

a) Sacamos todas las muestras de un tamaño dado.
b) Calculamos para cada muestra el valor del estadístico que nos interesa.
c) Enumeramos los diferentes valores junto con sus probabilidades de ocurrencia.

Para entender este procedimiento vamos a utilizar una distribución muestral experimental de medias calculadas a partir de todas las muestras posibles de tamaño 2 que se pueden sacar de una población pequeña. Esta distribución muestral experimental se ha tomado de Daniel W. (1981).

Tabla 3.5.

Distribución de valores de la variable X para una determinada población

Población	**1 2 3 4 5 6 7 8 9 10**
Variable X	**2 5 7 3 4 1 10 9 8 6**

Calculamos el valor medio y la varianza de la distribución de probabilidad asociada a la variable X:

$$\overline{X} = \frac{\sum X_i}{N} = \frac{55}{10} = 5{,}5 \qquad \sigma^2 = \sum \frac{(X_i - \mu)^2}{N} = 8{,}25$$

Todas las posibles muestras de 2 elementos de la variable X, junto con sus diferentes medias muestrales son:

Tabla 3.6.

Muestras posibles de tamaño 2 para la variable X

Primera	**Segunda muestra**									
muestra	1	2	3	4	5	6	7	8	9	10
1	**1, 1**	1, 2	1, 3	1, 4	1, 5	1, 6	1, 7	1, 8	1, 9	1, 10
	(1)	(1,5)	(2)	(2,5)	(3)	(3,5)	(4)	(4,5)	(5)	(5,5)
2	2, 1	**2, 2**	2, 3	2, 4	2, 5	2, 6	2, 7	2, 8	2, 9	2, 10
	(1,5)	**(2)**	(2,5)	(3)	(3,5)	(4)	(4,5)	(5)	(5,5)	(6)
3	3, 1	3, 2	**3, 3**	3, 4	3, 5	3, 6	3, 7	3, 8	3, 9	3, 10
	(2)	(2,5)	**(3)**	(3,5)	(4)	(4,5)	(5)	(5,5)	(6)	(6,5)
4	4, 1	4, 2	4, 3	**4, 4**	4, 5	4, 6	4, 7	4, 8	4, 9	4, 10
	(2,5)	(3)	(3,5)	**(4)**	(4,5)	(5)	(5,5)	(6)	(6,5)	(7)
5	5, 1	5, 2	5, 3	5, 4	**5, 5**	5, 6	5, 7	5, 8	5, 9	5, 10
	(3)	(3,5)	(4)	(4,5)	**(5)**	(5,5)	(6)	(6,5)	(7)	(7,5)
6	6, 1	6, 2	6, 3	6, 4	6, 5	**6, 6**	6, 7	6, 8	6, 9	6, 10
	(3,5)	(4)	(4,5)	(5)	(5,5)	**(6)**	(6,5)	(7)	(7,5)	(8)
7	7, 1	7, 2	7, 3	7, 4	7, 5	7, 6	**7, 7**	7, 8	7, 9	7, 10
	(4)	(4,5)	(5)	(5,5)	(6)	(6,5)	**(7)**	(7,5)	(8)	(8,5)
8	8, 1	8, 2	8, 3	8, 4	8, 5	8, 6	8, 7	**8, 8**	8, 9	8, 10
	(4,5)	(5)	(5,5)	(6)	(6,5)	(7)	(7,5)	**(8)**	(8,5)	(9)
9	9, 1	9, 2	9, 3	9, 4	9, 5	9, 6	9, 7	9, 8	**9, 9**	9, 10
	(5)	(5,5)	(6)	(6,5)	(7)	(7,5)	(8)	(8,5)	**(9)**	(9,5)
10	10, 1	10, 2	10, 3	10, 4	10, 5	10, 6	10, 7	10, 8	10, 9	**10, 10**
	(5,5)	(6)	(6,5)	(7)	(7,5)	(8)	(8,5)	(9)	(9,5)	**(10)**

Como vemos en la tabla, el número de muestras posibles de tamaño 2 es de N^2 = 100. Entre paréntesis figura el cálculo de la media de cada muestra.

Vamos a construir ahora las distribuciones de la media muestral $\bar{x}$ del ejemplo que figura en la tabla 3.6.

a) Muestreo con reposición

Bajo el supuesto de que el orden influye, es decir, no somos indiferentes al orden en que se extraen las muestras, lo que significa que la muestra que contiene el elemento 1, 2 no es la misma que la que contiene el elemento 2, 1, las muestras posibles son todas las que figuran en la tabla 3.6.. Los pasos que han de seguirse son los siguientes:

a) Determinación del número de muestras posibles: N^2=100.

b) Calculo de todas las medias posibles (entre paréntesis en la tabla 3.6.).

c) Cálculo de la probabilidad de ocurrencia, que se muestra en la siguiente tabla:

Tabla 3.7.

Distribución muestral de la media de las muestras de tamaño n=2

$\bar{x}$	**Probabilidad**
1,0	1/100
1,5	2/100
2,0	3/100
2,5	4/100
3,0	5/100
3,5	6/100
4,0	7/100
4,5	8/100
5,0	9/100
5,5	10/100
6,0	9/100
6,5	8/100
7,0	7/100
7,5	6/100
8,0	5/100
8,5	4/100
9,0	3/100
9,5	2/100
10,0	1/100
	100/100

En la tabla 3.7. observamos que se cumplen las condiciones que se exigen a una distribución de probabilidad: que cada una de las probabilidades sea mayor o igual que 0, y que la suma de todas las probabilidades sea igual a 1.

Calculamos ahora la media y la varianza de las 100 medias muestrales.

$$\mu_{\bar{x}} = \frac{\sum \bar{X}_i}{100} = \frac{550}{100} = 5{,}5 \qquad \sigma^2_{\bar{x}} = \frac{\sum \left(\bar{X}_i - \mu_{\bar{x}}\right)^2}{100} = 4{,}125$$

Se observa que la media de todas las medias muestrales, es exactamente igual a la media poblacional y que la varianza de las medias muestrales es igual a la varianza de la población dividida por el tamaño de la muestra. Podemos calcular el error típico de las medias muestrales como:

Error típico de la media: $\sigma_{\bar{x}} = \sqrt{\sigma^2_{\bar{x}}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

b) Muestreo sin reposición

Si el muestreo se realiza sin reposición, es decir, sin que aparezca el elemento seleccionado en la extracción anterior (en la tabla 3.6., todas las muestras que están por debajo o por encima de la diagonal principal), y bajo el supuesto de que el orden en que se sacan las muestras no tiene importancia, el número de muestras viene dado por las combinaciones de *N* elementos tomados de *n* en *n* $\binom{N}{n}$, que para nuestro ejemplo serían:

$$\binom{10}{2} = \frac{10!}{8!2!} = 45$$ (las 45 muestras que figuran por encima o por debajo de la diagonal principal, excluyendo ésta)

La media de las 45 muestras es:

$$\mu_{\bar{x}} = \frac{247{,}5}{45} = 5{,}5$$

De nuevo observamos que la media de las medias muestrales es igual a la media de la población.

La varianza de la media de las 45 muestras se obtiene:

$$\sigma^2_{\bar{x}} = \frac{(1{,}6 - 5{,}5)^2 + \ldots + (9{,}5 - 5{,}5)^2}{45} = 3{,}67$$

Como se aprecia, la varianza de las 45 medias muestrales no coincide con la varianza obtenida en el muestreo con reposición. La varianza del muestreo sin reposición se obtiene multiplicando $\frac{\sigma^2}{n}$ por el factor $\frac{N-n}{N-1}$ llamado *factor de corrección de poblaciones finitas*. Cuando el tamaño de la población es muy grande con relación al tamaño de la muestra, es decir, *n/N* es menor o igual a 0,05, se suele prescindir del factor de corrección.

A continuación exponemos de forma breve las distribuciones asociadas a los estadísticos: media muestral, diferencia de medias muestrales, proporción muestral, diferencia de proporciones muestrales, varianza muestral y razón de varianzas muestrales.

Distribución de la media muestral

Si $\bar{x}$ es la media de una muestra de tamaño *n*, obtenida de forma aleatoria de una población distribuida normalmente con media μ y varianza σ^2, entonces la media muestral, $\bar{x}$, se distribuye normalmente con media μ y varianza σ^2/n, es decir:

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

y, por tanto, la variable tipificada:

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

se distribuye como una normal de parámetros (0, 1).

Cuando efectuamos un muestreo en una población que no está distribuida normalmente, podemos utilizar el Teorema Central del Límite si la muestra es suficientemente grande.

En este caso, se puede afirmar que sin tener en cuenta la forma funcional de la población de donde se extrae la muestra, la distribución de la media muestral, calculada con muestras de tamaño *n* extraídas de una población con media μ y varianza σ^2, se distribuye como una distribución normal con media μ y varianza σ^2/n. Si *n* es grande, la distribución de las medias muestrales pueden aproximarse mucho a una distribución normal. Muchos expertos sugieren que tamaños muestrales superiores a 30 justifican el uso del Teorema Central del Límite.

Si no se conoce la desviación típica de la distribución, se utiliza la variable $\frac{\bar{x}-\mu}{S/\sqrt{n}}$,

siendo *S* la desviación típica muestral, la cual sigue una distribución t de Student con *n-1* grados de libertad.

Distribución de la diferencia entre dos medias muestrales

La distribución muestral de la diferencia de dos medias muestrales, calculadas a partir de muestras alternativas independientes de tamaño n_1 y n_2 extraídas de dos poblaciones distribuidas normalmente, también estará distribuida normalmente.

Bajo el supuesto de que las *varianzas poblacionales son conocidas*:

$\bar{x}_1 - \bar{x}_2$ se distribuye como una normal con parámetros $\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$, es decir:

$$\bar{x}_1 - \bar{x}_2 \rightarrow N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

y por tanto la variable tipificada seguirá una distribución normal con parámetros (0, 1).

$$Z = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightarrow N(0,1)$$

Cuando *no se conocen las varianzas poblacionales, pero pueden suponerse iguales*:

$$T = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - \left(\mu_1 - \mu_2\right)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightarrow t_{n_1+n_2-2}$$

donde Sp es la varianza muestral combinada

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

Si *no se conocen las varianzas y son desiguales*:

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \rightarrow t_{n_1+n_2-2-\Delta}$$

siendo Δ un número positivo corrector que se calcula tomando el entero más próximo a:

$$\frac{\left((n_2-1)S_1 - (n_1-1)S_2\right)^2}{(n_2-1)S_1^2 + (n_1-1)S_2^2}$$

donde $S_i = \frac{s_i^2}{n_i}$ (i = 1, 2)

Distribución de la proporción muestral

La proporción muestral, $\hat{p}$, calculada con muestras aleatorias de tamaño *n*, extraídas de una población en la que *p* es la proporción poblacional, también se distribuye normalmente si *n* es grande.

Si la población es finita y de tamaño *N*, la media $\mu_{\hat{p}}$ de la distribución de $\hat{p}$ será $\mu_{\hat{p}} = p$ y la desviación típica:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}} \ .$$

Si la población de la que se extrae la muestra es infinita, la media y la desviación típica de la distribución de $\hat{p}$ serán iguales a p y $\sqrt{\frac{p(1-p)}{n}}$, respectivamente.

La distribución muestral de $\hat{p}$ será aproximadamente normal si tanto *np* como *n(1-p)* son mayores que 5.

Distribución de la diferencia entre dos proporciones muestrales.

La distribución muestral de $\hat{p}_1 - \hat{p}_2$, o diferencia entre dos proporciones muestrales, donde $\hat{p}_1 - \hat{p}_2$ se calcula a partir de dos muestras aleatorias de tamaño n_1 y n_2 tienen como media y varianza:

$$\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2 \qquad \sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Si n_1 y n_2 son grandes, la distribución muestral de $\hat{p}_1 - \hat{p}_2$ es aproximadamente normal.

Distribución de la varianza muestral

Si $s^2 = \sum \frac{(x_i - \bar{x})^2}{n-1}$ es la varianza de una muestra aleatoria de tamaño *n* de una población distribuida normalmente con media μ y varianza σ^2, entonces:

$$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma^2}$$

sigue una distribución chi-cuadrado con *n-1* grados de libertad.

Distribución de la razón de varianzas muestrales

Dadas S_1^2 y S_2^2, o varianzas muestrales calculadas a partir de muestras aleatorias independientes de tamaño n_1 y n_2, extraídas de poblaciones distribuidas normalmente con varianzas σ_1^2 y σ_2^2 respectivamente, entonces:

$$F = \frac{S_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$$

sigue una distribución F con $n_1 - 1$ y $n_2 - 1$, grados de libertad.

Este resultado viene derivado de la relación que existe entre la distribución F y la χ^2, ya que:.

$$F_{n_1,n_2} \equiv \frac{\frac{\chi_1^2}{(n_1 - 1)}}{\frac{\chi_2^2}{(n_2 - 1)}}$$

3.7. INTERVALOS DE CONFIANZA

En la estimación estadística generalmente sólo tenemos una muestra, a través de la cual obtenemos una estimación de los parámetros poblacionales. El número de estimaciones que podemos realizar de una población, a través de un procedimiento de muestro, con una muestra prefijada, es generalmente muy grande, porque cada una de las muestras posibles que se pueden sacar de la población arrojaría una estimación.

Por esta razón, a la estimación que obtenemos en una investigación por muestreo la acompañamos con un intervalo de valores posibles. La amplitud de dicho intervalo dependerá del grado de confianza que establezcamos.

El grado o nivel de confianza nos expresa el número de veces que la media verdadera de la población está incluida en cien intervalos de cien muestras extraidas de la población. El nivel de confianza más utilizado es el 95%, lo que quiere decir que 95 de cada 100 intervalos construidos contendrán el verdadero valor de la media.

El intervalo de confianza para la media de una población normalmente distribuida se construye en base a la probabilidad de que dicha media esté comprendida entre dos valores, $\overline{X}_a$ y $\overline{X}_b$ equidistantes a ella:

$$P[\bar{X}_a \leq \mu_{\bar{x}} \leq \bar{X}_b] = 1 - \alpha \quad (1)$$

siendo *1-* α el nivel o grado de confinza asociado a dicho intervalo.

Tomando como estimador de la media poblacional la media muestral, sabemos por el apartado 3.4. que:

$$\frac{\bar{x} - \mu_{\bar{x}}}{\sigma}\sqrt{n}$$

se distribuye como una normal de parámetros (0, 1) y, por tanto, puede determinarse, a través de la tabla de la normal, un valor *K* que verifique lo siguiente:

$$P(-K \leq \frac{\bar{x} - \mu_{\bar{x}}}{\sigma}\sqrt{n} \leq K) = 1 - \alpha \quad (2)$$

Dada la propiedad de simetría de la distribución normal, este valor *K* será aquel que deje a su izquierda una probabilidad de *(1- α) /2*. Por ejemplo, si el nivel de confianza que fijamos es del 95% el valor de *K* será aquel que deje a su izquierda una probabilidad de 0,975, tomando en este caso el valor 1,96 (ver tabla de la normal en el Anexo II).

Multiplicando por σ a los términos de la ecuación (2), obtenemos lo siguiente:

$$P(-K\sigma \le (\bar{x} - \mu_{\bar{x}})\sqrt{n} \le K\sigma) = 1 - \alpha$$

Dividiendo por $\sqrt{n}$:

$$P(\frac{-K\sigma}{\sqrt{n}} \le (\bar{x} - \mu_{\bar{x}}) \le \frac{K\sigma}{\sqrt{n}}) = 1 - \alpha$$

Restando a todos los términos $\bar{x}$:

$$P(-\bar{x} - \frac{K\sigma}{\sqrt{n}} \le -\mu_{\bar{x}} \le -\bar{x} + \frac{K\sigma}{\sqrt{n}}) = 1 - \alpha$$

Por último, si multiplicamos por *-1*, y ordenando el intervalo de confianza:

$$P(\bar{x} - \frac{K\sigma}{\sqrt{n}} \le \mu_{\bar{x}} \le \bar{x} + \frac{K\sigma}{\sqrt{n}}) = 1 - \alpha \quad (3)$$

Si sustituimos en (1), $\bar{X}_a$ y $\bar{X}_b$ serían:

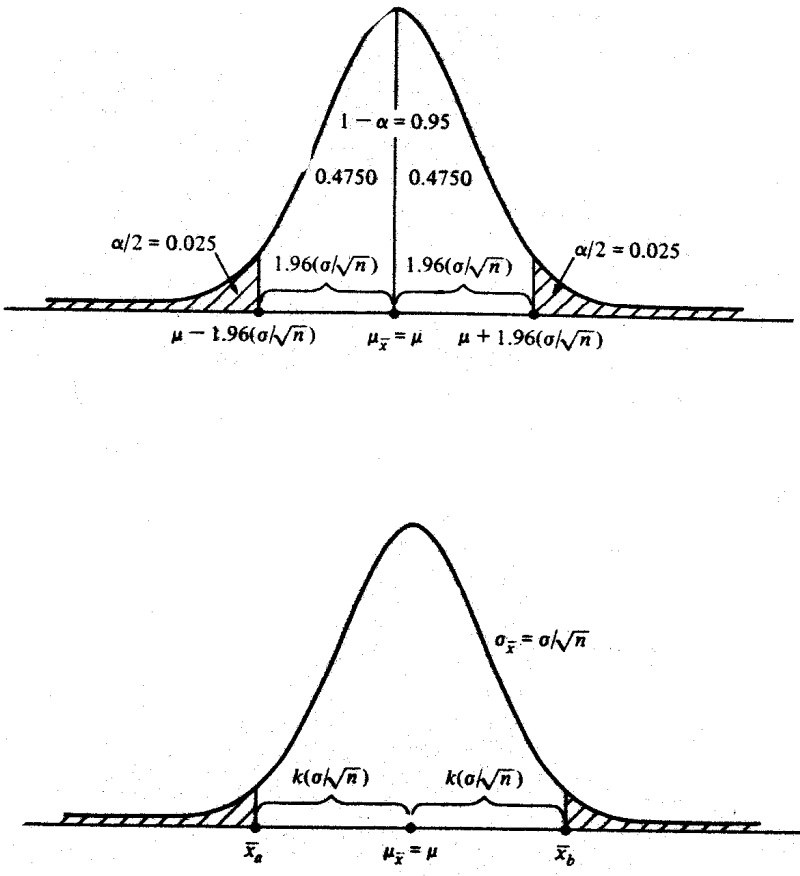
$$\bar{X}_a = \mu - k\frac{\sigma}{\sqrt{n}} \qquad \bar{X}_b = \mu + k\frac{\sigma}{\sqrt{n}}$$

Como vemos los extremos del intervalo se acaban expresando en función del error típico de la distribución del estadístico y de *K*. A éste último se le denomina *factor de confiabilidad*.

Se muestra en el gráfico 3.6. el esquema de construcción de intervalos de confianza.

Gráfico 3.6.

Esquema de construcción de intervalos de confianza



Supóngase como ejemplo la construcción de un intervalo con un nivel de confianza del 95% para la media de una distribución normal con desviación típica σ. = 3 En este caso, *K* toma el valor 1,965 (valor que deja una probabilidad de 0,975 a la izquierda en una distribución normal estándar).Una vez extraída una muestra de tamaño igual a 100, la media toma un valor de 5,5. El intervalo de confianza resultante es el siguiente:

$$P[5{,}5 - 1{,}965\frac{3}{\sqrt{100}} \le \mu_{\bar{x}} \le 5{,}5 + 1{,}965\frac{3}{\sqrt{100}}] = 0{,}95$$

y decimos que la probabilidad de que el parámetro desconocido esté entre los puntos

$$\overline{X}_a = 5{,}5 - 1{,}965\frac{3}{\sqrt{100}} = 4{,}91 \quad \text{y} \quad \overline{X}_b = 5{,}5 + 1{,}965\frac{3}{\sqrt{100}} = 6{,}09$$ es igual a 0,95.

Como hemos visto, a diferentes valores de 1- α le corresponden diferentes valores k. Si 1-α = 0,99, entonces k = 2,58.

Si para cada muestra posible obtenemos los intervalos de confianza para la media

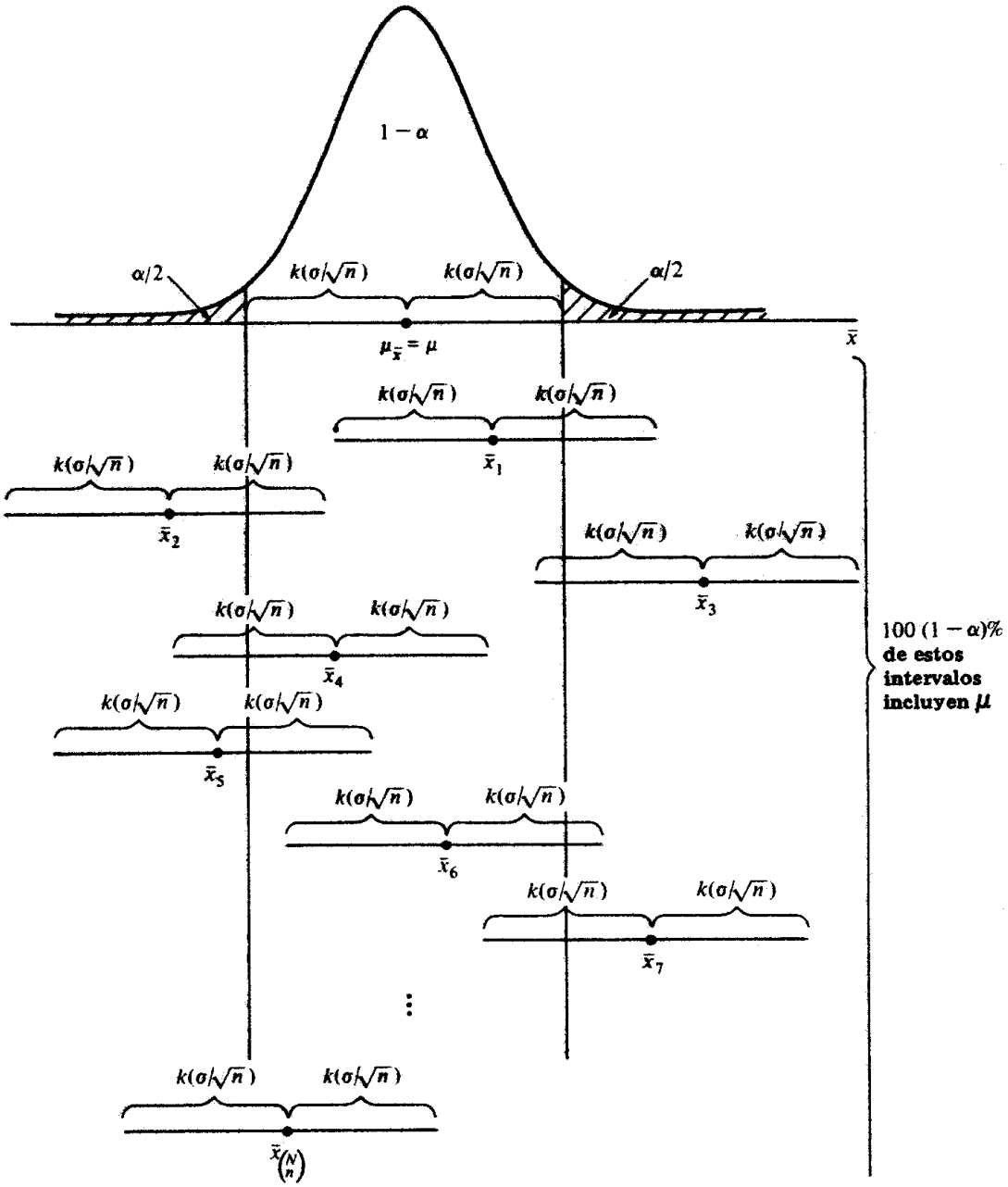
$$\bar{X}_1 \pm k\frac{\sigma}{\sqrt{n}} \ , \ \bar{X}_2 \pm k\frac{\sigma}{\sqrt{n}} \ , \ \bar{X}_3 \pm k\frac{\sigma}{\sqrt{n}} \ldots\ldots$$

La ecuación anterior nos indica que a la larga 100(1 - α)% de los intervalos así construidos contendrán la media poblacional desconocida μ, siendo éste, como se señaló anteriormente, el significado del nivel de confianza.

En la práctica como sólo disponemos de una muestra, construimos el intervalo de la media $\bar{X}_0$, sumando y restando k veces la desviación típica de la media. A este intervalo le llamamos intervalo de confianza, y dentro de este intervalo la verdadera media poblacional, puede estar o no. Se muestra a continuación un gráfico explicativo.

Gráfico 3.7.

Significado del nivel de confianza



En términos generales, los intervalos de confianza para los estadísticos muestrales se expresan como:

Estimador ± (factor de confiabilidad)(error típico del estimador)*

Así, el intervalo de confianza para la proporción, sería:

$$\hat{p} \pm k\left(\sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}\right)$$

3.8. CONTRASTE DE HIPÓTESIS

Una buena parte de las investigaciones estadísticas están orientadas al desarrollo de procesos encaminados a la contrastación de hipótesis que previamente se han establecido.

Una hipótesis es una afirmación que está sujeta a verificación o comprobación. Hay que tener presente que una hipótesis no es un hecho establecido o firme, las hipótesis están basadas en la experiencia, en la observación, en la experimentación o en la intuición del sujeto que las formula.

Cuando las hipótesis se plantean de tal modo que se pueden comprobar por medio de los métodos estadísticos reciben el nombre de hipótesis estadísticas. Estas hipótesis son afirmaciones que se efectúan sobre uno o más parámetros de una o más poblaciones. Las hipótesis estadísticas son de dos tipos: hipótesis nula e hipótesis altenativa. La hipótesis nula, simbolizada por H_0, es la hipótesis que se debe de comprobar. Esta hipótesis recibe también el nombre de hipótesis de ninguna diferencia, dado que generalmente se afirma que no hay ninguna diferencia entre la hipótesis nula y la alternativa.

Para contrastar una hipótesis nula examinamos los datos de la muestra tomados de la población y determinamos si son o no compatibles con dicha hipótesis. Si no son compatibles entonces H_0 se rechaza, en caso contrario no se rechaza. Si no se rechaza la hipótesis nula afirmamos que los datos de esa muestra en concreto no dan suficiente evidencia para que concluyamos que la hipótesis nula sea falsa. Si se rechaza decimos que los datos particulares de la muestra si evidencian que la hipótesis nula es falsa y la hipótesis alternativa, H_1, es verdadera.

El criterio que permite decidir si rechazamos o no la hipótesis nula es siempre el mismo. Definimos un estadístico de prueba y unos límites que dividen el espacio muestral en una región en donde se rechaza la hipótesis establecida y otra región en la que no se rechaza, llamada región de aceptación. A la región donde se rechaza la hipótesis nula se le llama región crítica. Esta región es un subconjunto del espacio muestral, y si el valor del estadístico de prueba pertenece a él se rechaza la hipótesis nula.

El límite entre la región crítica y la región de aceptación viene determinado por la información previa relativa a la distribución del estadístico, mediante la especificación de la hipótesis alternativa y por las consideraciones en los costes de obtener conclusiones incorrectas.

Señalar que un estadístico de prueba es una fórmula que nos dice como confrontar la hipótesis nula con la información de la muestra, y es por tanto una variable aleatoria cuyo valor cambia de muestra a muestra.

En la contrastación de hipótesis estadísticas aparecen dos tipos de errores:

- *Error de tipo I*. Rechazar la hipótesis nula siendo cierta.
- *Error de tipo II.* Aceptar la hipótesis nula siendo falsa.

La situación ideal sería poder minimizar los dos tipos de errores al mismo tiempo, pero dado que esto es imposible, normalmente lo que se hace es fijar la probabilidad del error de tipo I o *nivel de significación* y se realiza el contraste. Así, si por ejemplo si se utiliza para el nivel de significación un valor de 0,05, esto equivale a decir que si para realizar un contraste tomáramos infinitas muestras de la población, rechazaríamos la hipótesis nula de forma incorrecta un 5 % de las veces.

En resumen, en la contrastación de hipótesis hay que considerar tres factores:

1. La opinión a priori acerca de la validez del contraste
2. Las consecuencias que se pueden derivar de equivocarnos
3. La evidencia aportada por la muestra

Los contrastes de hipótesis estadísticas se clasifican en *paramétricos* y *no paramétricos*. Las pruebas estadísticas paramétricas requieren que los valores de las características de la población analizada sean producto de una medición en una escala de intervalo, de tal forma que sea posible utilizar operaciones aritméticas (sumas, productos, medias, etc.). Las no paramétricas se utilizan cuando el modelo no especifica las condiciones de los parámetros de la población de donde se sacó la muestra.

Se escoge una prueba paramétrica cuando se satisfacen al menos, además del requisito de la medición, las condiciones asociadas al modelo estadístico: las observaciones deben de ser independientes y hacerse en poblaciones distribuidas normalmente. Las poblaciones deben de tener la misma varianza (o en algunos casos se exige tener una proporción de varianza conocida).

Las suposiciones que se asocian a las pruebas no paramátricas son pocas y más débiles que las de las pruebas paramétricas.

Los principales test de hipótesis que se pueden aplicar son los siguientes:

A) CONTRASTES DE HIPÓTESIS PARAMÉTRICAS.

1.- Contrastes sobre los estadísticos de una población:

- Contrastes de significación para la media.
- Contrastes sobre la varianza.
- Contrastes sobre la proporción poblacional.

2.- Contrastes de comparación de dos poblaciones:

- Contraste de la T de Student para la igualdad de medias.
- Contraste de diferencias entre las medias problacionales.
- Test de Barlett para contraste de igualdad de varianzas.
- Contraste Kolmogorov-Smirnov para dos muestras.

3.- Contrastes de comparación de más de dos poblaciones.

- Método Scheffé de comparaciones múltiples.

4.- Contrastes de bondad en el ajuste:

- Contraste χ^2 de Pearson de bondad de ajuste.
- Contraste Kolmogorov-Smirnov de bondad de ajuste.
- Contraste de normalidad de Lilliefors.
- Contraste de normalidad de Shapiro-Wilk.

B) CONTRASTES DE HIPÓTESIS NO PARAMÉTRICAS.

1.- Contrastes de comparación de dos poblaciones:

- Contraste de la mediana.
- Contraste de Wilcoxon-Mann-Whitney.

- Contraste de Siegel-Tukey.

2.- Contrastes de comparación de más de dos poblaciones:

- Contraste de Kruskal-Wallis.
- Comparaciones Múltiples.

3.- Contrastes de independencia:

- Contraste X^2 de independencia.
- Contraste G^2 de independencia.
- Test de Tocher.
- Test binomial.
- Test de McNemar.
- Test de Gart.

4.- Contrastes de aletoriedad:

- Contraste de Rachas de Wald-Wolfowitz.
- Contraste del cuadrado medio de diferencias sucesivas.

5.- Contrastes de localización:

- Contraste de signos de la mediana.
- Contraste de signos para una muestra apareada.
- Contraste de rangos-signos de Wilcoxon para una muestra.

6.- Contrastes de homogeneidad:

- Contraste X^2 de homogeneidad.

En la formalización del procedimiento de contrastación podemos distinguir siete pasos principales:

1.- Planteamiento de las hipótesis.

2.- Selección del nivel de significación.

3.- Descripción de la población y tamaño de la muestra.

4.- Selección del estadístico de prueba y su distribución.

5.- Especificación de las regiones de aceptación y de rechazo.

6.- Recolección de datos y cálculo del estadístico.

7.- Decisión estadística.

A continuación se desarrolla un ejemplo que nos sirve para ilustrar algunos de los conceptos anteriormente descritos.

En el ejemplo consideramos una región donde se piensa que al menos el 25 % de la población toma cierta bebida. Con el fin de verificar si esta suposición es razonable un investigador selecciona una muestra aleatoria de 120 personas. De las 120 personas seleccionadas 20 afirmaron tomar la bebida, es decir, el 16,7%.

1º.- Planteamiento de la hipótesis

Se contrasta la hipótesis nula de que el 25% o más de la población toma dicha bebida, frente a la hipótesis alternativa de que menos de un 25% la toma.

$$H_0: p \geq 0{,}25 \qquad H_1: p<0{,}25$$

2º.- Nivel de significación o error de tipo I.

Sea α=0,05.

3º.- Descripción de la población

La población es binomial, ya que está compuesta por bebedores y no bebedores de dicha bebida. La población es suficientemente grande en relación con la muestra para que podamos pasar por alto el factor de corrección y la muestra es suficientemente grande para que podamos aplicar la aproximación a la distribución normal en la verificación de la hipótesis.

4º.- El Estadístico pertinente.

Bajo la hipótesis nula, la distribución muestral de $\hat{p}$ es de forma aproximadamente normal con una media $p = p_0 = 0{,}25$ y un error típico de 0,0395.

$$\sigma_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0{,}25x0{,}75}{120}} = 0{,}0395$$

Nótese que empleamos p_0 en lugar de p ya que se supone que la hipótesis nula es verdadera hasta que haya suficiente evidencia para rechazarla.

5º.- Regiones de aceptación y de rechazo.

El valor crítico es -1,645, que es el valor correspondiente de la distribución normal estándar que deja el 5 % de la distribución a la izquierda, de modo que la región de rechazo consta de todos los valores *Z* iguales o menores que -1,645. La región de aceptación corresponde a todos los valores de *Z* mayores que -1,645, siendo *Z*:

$$Z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}}$$

6º.- Recolección de datos y cálculo del estadístico

$$Z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}} = \frac{0{,}167 - 0{,}25}{0{,}0395} = -2{,}108$$

7º.- Decisión estadística.

Dado que -2,108 es menor que -1,645 rechazamos la H_0: $p \geq 0{,}25$ y, por tanto, concluimos que menos del 25 % de la población ha probado alguna vez la bebida.

3.9. DISTRIBUCIONES BIDIMENSIONALES

En el estudio de una población, colectivo o simplemente conjunto de elementos, podemos estar interesados en medir no sólo una, sino varias variables, lo que en la práctica es muy frecuente. En principio y para simplificar, suponemos que se han observado sólo dos variables a las que denotamos por *x* e *y*.

Se llama distribución conjunta de frecuencias de las dos variables (*x*, *y*) a la tabla que representa los valores observados de ambas variables y las frecuencias relativas de aparición de cada una de las variables.

Si la muestra consta de *r* elementos para la variable *X* y *k* elementos para la variable *Y*, se tendrán *r* * *k* pares de elementos que expresaremos por (x_i, y_j) para *i = 1,2,...,r* y *j = 1,2, ...,k* donde O_{ij} es la frecuencia correspondiente a las dos variables x_i e y_j, $O_{i\cdot}$ es la frecuencia del elemento x_i, $O_{\cdot j}$ la del elemento y_j y $O_{\cdot\cdot}$ la frecuencia total. Esto se transcribe a una tabla tal como se muestra en la siguiente figura.

Y / X	y_1	y_2	y_j	y_k	$\sum_{j=1}^{k} O_{ij}$
x_1	O_{11}	O_{12}	O_{1j}	O_{1k}	$O_{1\cdot}$
x_2	O_{21}	O_{22}	O_{2j}	O_{2k}	$O_{2\cdot}$
x_i	O_{i1}	O_{i2}	O_{ij}	O_{ik}	$O_{i\cdot}$
x_r	O_{r1}	O_{r2}	O_{rj}	O_{rk}	$O_{r\cdot}$
$\sum_{i=1}^{r} O_{ij}$	$O_{\cdot 1}$	$O_{\cdot 2}$	$O_{\cdot j}$	$O_{\cdot k}$	$\sum_{i=1}^{r}\sum_{j=1}^{k} O_{ij} = O_{\cdot\cdot}$

Si llamamos $fr(x_i, y_j) = \frac{O_{ij}}{O_{\cdot\cdot}}$ a la *frecuencia relativa* del elemento (x_i, y_j), entonces se verifica que $\sum_i \sum_j fr(x_i, y_j) = 1$, es decir, la suma de todas las frecuencias relativas es igual a 1.

En el análisis conjunto de las dos variables nos interesa de forma especial la relación existente entre ambas variables.

Distribuciones marginales

Las distribuciones marginales aparecen cuando se estudian aisladamente las variables con independencia del resto. Éstas se obtienen con las siguientes fórmulas:

$$fr(x_i) = \sum_j fr(x_i, y_j) = \sum_j \frac{O_{ij}}{O_{..}} = \frac{1}{O_{..}} \sum_j O_{ij} = \frac{O_{i.}}{O_{..}}$$

$$fr(y_j) = \sum_i fr(x_i, y_j) = \sum_i \frac{O_{ij}}{O_{..}} = \frac{1}{O_{..}} \sum_i O_{ij} = \frac{O_{.j}}{O_{..}}$$

Distribuciones condicionadas

Son las distribuciones de una determinada variable, condicionada ésta por los valores de la otra, es decir:

$$fr(y_j / x_i) = \frac{fr(x_i, y_j)}{fr(x_i)} = \frac{O_{ij} / O_{..}}{O_{i.} / O_{..}} = \frac{O_{ij}}{O_{i.}}$$

Ha de verificarse que $\sum_j fr(y_j / x_i) = 1$ ya que:

$$\sum_j fr(y_j / x_i) = \sum_j \frac{fr(x_i, y_j)}{fr(x_i)} = \frac{1}{fr(x_i)} \sum_j fr(x_i, y_j) = \frac{1}{fr(x_i)} * fr(x_i) = 1$$

Dependencia lineal

Las dos medidas de las que dispone la estadística descriptiva para medir la relación lineal que hay entre cada par de variables son: la *covarianza* y el *coeficiente de correlación*.

La *covarianza* entre dos variables viene reflejada por la siguiente expresión:

$Cov\ (x,y) = \sum_i \sum_j (x_i - \bar{x})\ (y_j - \bar{y})\ fr\ (x_i, y_j)$

La covarianza es, como vemos, el promedio del producto de las desviaciones de los puntos respecto a su media.

Si los valores altos de una variable están asociados con los valores altos de la otra variable, la covarianza será positiva; y su valor será negativo cuando los valores altos de una variable se asocian con los valores bajos de la otra variable.

Si entre los valores de ambas variables no hay relación, la covarianza tenderá a cero.

El inconveniente que presenta la covarianza es su dependencia de las unidades de medida de las variables.

El *coeficiente de correlación* es la covarianza dividida por el producto de las desviaciones típicas de ambas variables. De esta forma, dicho coeficiente es independiente de las unidades de medida.

$$r = \frac{\text{Cov}\,(x, y)}{S_x\, S_y}$$

Se puede comprobar que:

a) Al ser adimensional, el coeficiente de correlación no varía al multiplicar x_i por una constante k_1 e y_j por otra constante k_2.
b) Cuando la relación lineal entre las dos variables es exacta, lo que implica que todos los puntos están en la recta $y = a + bx$, el coeficiente de correlación es igual a 1 ó -1, si b es positivo o negativo.
c) El coeficiente de correlación está entre -1 y 1 cuando no hay relación lineal exacta.

3.10. TABLAS DE CONTINGENCIA

Las tablas de contingencia se utilizan para realizar contrastes no paramétricos de independencia de poblaciones, es decir, saber si existe o no relación entre variables de tipo cualitativo. Este tipo de variables pueden ser nominales (por ejemplo el sexo de los encuestados), de atributos (marcas de un producto) u ordinales (por ejemplo la medición del grado de satisfacción de los encuestados en una determinada escala). El empleo de las tablas de contingencia está especialmente indicado si las variables son de tipo nominal.

Una tabla de contingencia se utiliza para mostrar la existencia de relaciones entre dos variables en una encuesta estadística. También, mediante una tabla de contingencia podemos establecer una medición del grado de relación que se da entre ambas variables.

Supongamos que mediante una encuesta estadística estamos estudiando determinado atributo de la población (opina a favor o en contra), y deseamos saber si existen diferencias en las respuestas de los encuestados en función de su sexo.

Para ello, realizamos una tabla cruzada de doble entrada en donde resumimos los resultados obtenidos en la encuesta:

Tabla 3.9.

Opiniones a favor y en contra en función del sexo

	Varón	Mujer	Total
A favor	32	10	42
En contra	11	27	38
Total	43	37	80

Las tablas de la forma de la ejemplo anterior reciben el nombre de Tablas de Contingencia, y sobre ellas contrastamos la hipótesis de independencia entre las respuestas dadas a las preguntas realizadas en relación con el sexo, utilizando el estadístico χ^2, y podemos evaluar el grado de relación que se da entre las opiniones y el sexo a partir de diferentes coeficientes de asociación como la Odds Ratio, el coeficiente de contingencia, el coeficiente V de Crammer o la Q de Yule .

Estadístico χ^2

La hipótesis nula, H_0, que implicaría que existe independencia entre los factores (en el ejemplo anterior el sexo y las opiniones), se prueba a través de :

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Siendo:

- *r* el número de filas
- *k* el número de columnas
- *Oij* (frecuencia observada) el número de casos observados clasificados en la fila *i* de la columna *j*
- *Eij* (frecuencia esperada) el número de casos esperados, en el supuesto de independencia, correspondientes a la fila *i* de la columna *j*.

Se define la *frecuencia esperada* como aquella frecuencia que se daría si los sucesos fueran independientes. Para encontrar la frecuencia esperada o teórica de cada casilla (E_{ij}), se multiplican los dos totales marginales (fila y columna) y se divide este producto por el número total de casos. En el ejemplo anterior la tabla de frecuencias esperadas sería:

Tabla 3.10.

Frecuencias esperadas para las opiniones a favor y en contra en función del sexo

	Varón	Mujer	Total
A favor	23	19	42
En contra	20	18	38
Total	43	37	80

calculándose del siguiente modo:

$$E_{11} = \frac{O_{1\cdot}O_{\cdot 1}}{O_{\cdot\cdot}} = (42*43)/80 = 23$$

$$E_{12} = \frac{O_{1\cdot}O_{\cdot 2}}{O_{\cdot\cdot}} = (42*37)/80 = 19$$

$$E_{21} = \frac{O_{2\cdot}O_{\cdot 1}}{O_{\cdot\cdot}} = (38*43)/80 = 20$$

$$E_{22} = \frac{O_{2\cdot}O_{\cdot 2}}{O_{\cdot\cdot}} = (38*37)/80 = 18$$

Si la hipótesis nula se verifica los valores del estadístico χ^2 están distribuidos como una χ^2 con grados de libertad = (r-1)(k-1), en donde r es el número de filas y k el número de columnas de la tabla de contingencia.

Recordar que para que el test de la χ^2 ofrezca resultados concluyentes el 80% de las frecuencias esperadas ha de presentar un valor superior a 5 y ninguna ha de ser menor o igual que 1. Si esto ocurre se debe proceder a recodificar las respuesta dadas a la encuesta.

El valor de χ^2 para el ejemplo es:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = \frac{(32-23)^2}{23} + \frac{(10-19)^2}{19} + \frac{(11-20)^2}{20} + \frac{(27-18)^2}{18} = 17{,}91$$

el cual bajo la hipótesis nula sigue una distribución χ^2 con 1 grado de libertad.

El percentil 95 de la distribución χ^2 con 1 grado de libertad toma un valor de 3,84. Como el valor del estadístico χ^2 = 17,91, es mayor que este percentil se rechaza la hipótesis de independencia lo que significa que el sexo tiene influencia a la hora de estar a favor o en contra.

Medidas de asociación

La *Odds Ratio* se define como el cociente de las siguientes probabilidades:

$$OR = \frac{\frac{O_{11}}{O_{1\bullet}} / \frac{O_{12}}{O_{1\bullet}}}{\frac{O_{21}}{O_{2\bullet}} / \frac{O_{22}}{O_{2\bullet}}} = \frac{O_{11}O_{22}}{O_{12}O_{21}}$$

Si OR > 1 entonces la probabilidad de "a favor" es mayor en los hombres que en las mujeres, si OR = 1 ambas probabilidades son iguales (independencia en las opiniones de hombres y mujeres) y si OR < 1 la probabilidad de "a favor" es menor en los hombres que en las mujeres.

El valor de esta medida está comprendido en el intervalo (0; ∞).

Las propiedades más relevantes son las siguientes:

1. Es invariante ante los cambios de escala en filas y columnas, o tan sólo en filas o en columnas.
2. Alcanza sus valores extremos, 0 e ∞, bajo asociación perfecta.
3. OR y 1/OR indican igual intensidad de la asociación, pero en direcciones opuestas.

Con objeto de lograr una interpretación más fácil, se define la siguiente medida:

$$OR' = ln(OR)$$

la cual es una medida *simétrica* cuyo rango de variación es (-∞, +∞), tomando el valor 0 en el caso de independencia y -∞ o +∞ en el caso de asociación perfecta.

En el ejemplo, *OR* y *OR'* toman los siguientes valores:

$$OR = \frac{32*27}{10*11} = 7{,}85 \qquad OR' = \ln\left(\frac{32*27}{10*11}\right) = 2{,}06$$

lo cual quiere decir que los hombres muestran una opinión más favorable que las mujeres.

El *coeficiente de contingencia* C es una medida del grado de asociación entre dos conjuntos de atributos, están ordenados o no, e independiente de la naturaleza de la variable (continua o discreta). Es un estadístico que se obtiene de la tabla de contingencia mediante la siguiente formula:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

El valor de este coeficiente está entre 0 y 1, los valores más próximos a 1 indicarían un mayor grado de interdependencia entre variables. Lógicamente, este coeficiente nunca puede alcanzar el valor 1, aunque haya completa asociación.

En el ejemplo anterior el valor del coeficiente de contingencia es igual a:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{17{,}91}{17{,}91 + 80}} = 0{,}4277$$

lo cual indica un grado de asociación medianamente alto.

El *coeficiente V de Cramer*, es otro estadístico que se obtiene a partir de la χ^2. Su valor oscila entre 0 y 1, siendo 0 cuando la independencia es completa y 1 cuando se da una completa asociación. Se obtiene a partir de:

$$V = \sqrt{\frac{\chi^2}{n * min(k-1, r-1)}}$$

En nuestro ejemplo el valor *V* será de:

$$V = \sqrt{\frac{\chi^2}{n * min(k-1, r-1)}} = \sqrt{\frac{17{,}91}{80}} = 0{,}4732$$

Otra medida de asociación es la *Q de Yule* que se calcula sobre las diferencias entre las frecuencias observadas (O_{ij}) y esperadas (E_{ij}). En una tabla 2×2 la medida Q de Yule se calcula a través de la siguiente expresión:

$$Q = \frac{nD_{11}}{O_{11}O_{22} - O_{12}O_{21}} \qquad \text{donde} \quad D_{11}=O_{11}-E_{11}$$

La *Q* de Yule está comprendida entre -1 y 1, siendo los criterios interpretativos:

- Q=0 independencia
- Q>0 asociación positiva
- Q<0 asociación negativa

Tabla 3.11.

Diferencias entre las frecuencias observadas y esperadas para las opiniones a favor y en contra en función del sexo

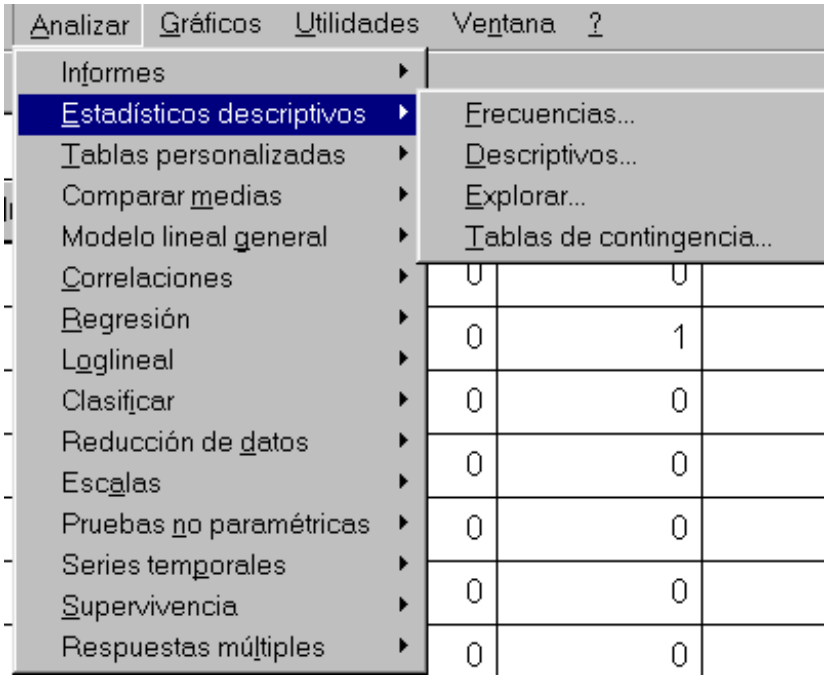
	Nivel 1	Nivel 2
Nivel 1	$D_{11}=O_{11}-E_{11}= 9$	$D_{12}=O_{12}-E_{12}= -9$
Nivel 2	$D_{21}=O_{21}-E_{21}= -9$	$D_{22}=O_{22}-E_{22}= 9$

En la tabla 3.11. se muestran las diferencias entre las frecuencias esperadas y observadas en el ejemplo que seguimos. La *Q* toma en este caso el valor:

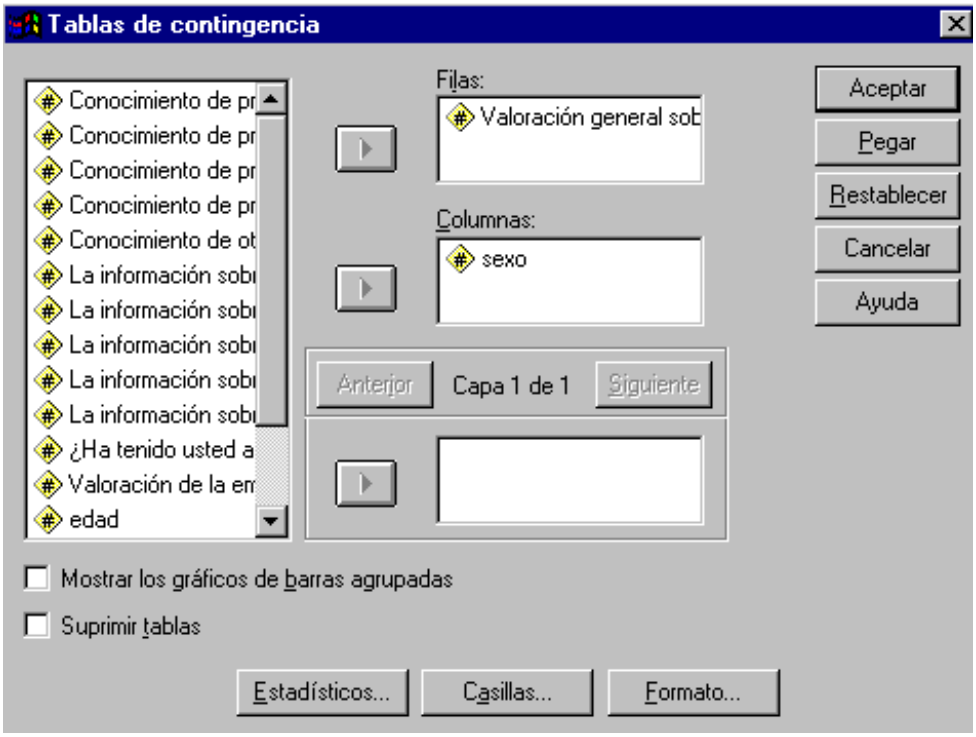
$$Q = \frac{nD_{11}}{O_{11}O_{22} - O_{12}O_{21}} = \frac{80 * 9}{32 * 27 + 10 * 11} = 0{,}74$$

4.11 Ejemplo en el SPSS

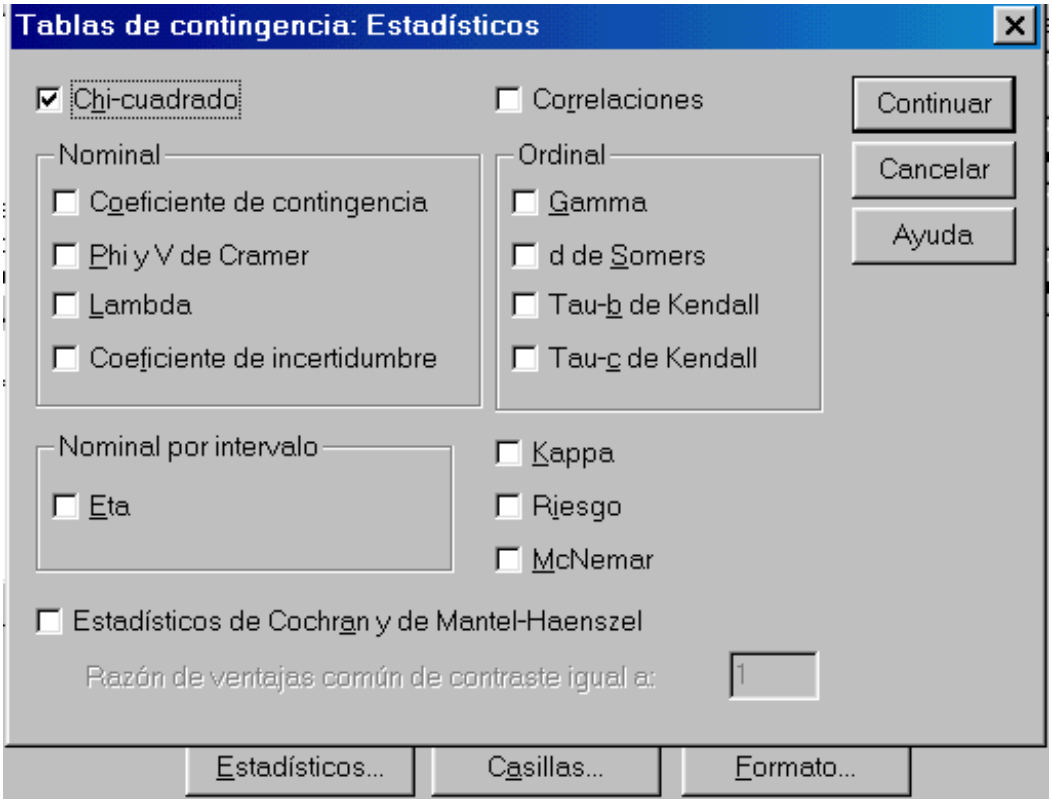
Para realizar una tabla de contingencia, en el SPSS seleccionamos el menú **Analizar /Estadísticos Descriptivos /Tablas de Contingencia**



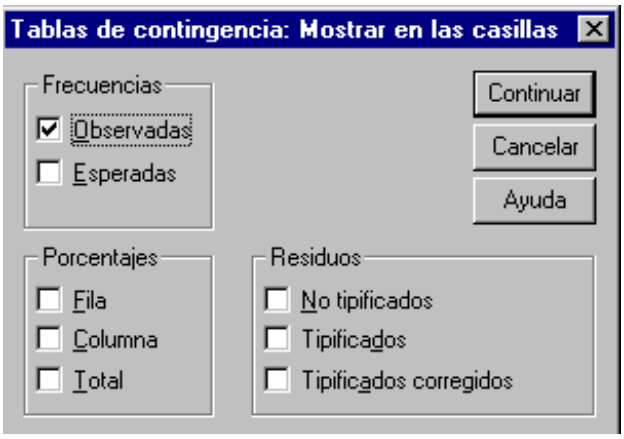
Con los datos del ejemplo del anexo nº 1, se eligen las variables ***Valoración general sobre los productos de la empresa y el Sexo.*** Nuestro objetivo es probar si existe dependencia entre las respuestas a las preguntas de valoración de los productos de la empresa y el sexo de los encuestados.



Se escogen los estadísticos para proceder a contrastar la hipótesis nula (En este caso que la valoración sobre los productos de la empresa es idéntica para hombres y mujeres)



En ***Casillas*** podemos elegir varias opciones relacionadas con la presentación de frecuencias, porcentajes y los residuos:



Los resultados que nos ofrece el SPSS son los siguientes:

Resumen del procesamiento de los casos

	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Valoración general sobre los productos de la empresa * SEXO	2613	99,2%	21	,8%	2634	100,0%

Tabla de contingencia Valoración general sobre los productos de la empresa * SEXO

Recuento

		SEXO			
		Hombre	Mujer	9	Total
Valoración general sobre los productos de la empresa	NS/NC	47	16	8	71
	Muy buena	207	135	1	343
	Buena	548	381	9	938
	Normal	615	393	26	1034
	Mala	111	61	2	174
	Muy deficiente	27	25	1	53
Total		1555	1011	47	2613

Pruebas de chi-cuadrado

	Valor	gl	Sig. asint. (bilateral)
Chi-cuadrado de Pearson	56,135[a]	10	,000
Razón de verosimilitud	40,315	10	,000
Asociación lineal por lineal	,042	1	,837
N de casos válidos	2613		

a. 3 casillas (16,7%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es ,95.

ANALISIS DE LA VARIANZA, REGRESIÓN Y SERIES TEMPORALES

4.1.ANÁLISIS DE LA VARIANZA

- **Modelo matemático**
- **Fases del análisis**
- **Estimación de los parámetros del modelo**
- **Tabla de análisis de la varianza**
- **Análisis de los residuos**

4.2 EJEMPLO EN SPSS

4.3ANÁLISIS DE REGRESIÓN

4.4 SERIES TEMPORALES

- **Tendencia**
- **Variaciones cíclicas y estacionales**

4.5 EJEMPLO EN SPSS

4.1. ÁNALISIS DE LA VARIANZA

PROBLEMA QUE SE PLANTEA

Se mide la producción de trigo por hectárea en cuatro parcelas distintas. Se realizan en cada parcela 50 mediciones respectivamente. Las cuatro parcelas están situadas en la misma zona y por tanto han estado sometidas a las mismas condiciones climáticas. El abono utilizado en cada una de ellas es distinto y lo que se desea es contrastar si la utilización de los diferentes abonos da lugar a distintas producciones.

Las producciones en una y otra parcela serán también dependientes de una serie de factores no controlables y muchas veces desconocidos, como por ejemplo diferencias en la maquinaria utilizada, cualificación de los agricultores, variaciones en la calidad de la parcela de cultivo, etc. Estos factores están englobados en un término al que denominamos *error experimental* o *perturbación*.

Por tanto, se parte de la hipótesis de que cada tipo de abono tendrá asociada una producción, la cual es desconocida, y los valores observados se determinan como la suma de esta producción y el error experimental o perturbación.

Los objetivos, pues, que pretendemos son los siguientes:

1. Comprobar si todos los abonos dan lugar a una misma producción.
2. Si las producciones asociadas son distintas, determinar qué tipo de abono da lugar a una mayor producción.

MODELO MATEMÁTICO

El planteamiento anteriormente expuesto da lugar a la formulación de un modelo matemático. Dicho modelo es el siguiente:

$$y_{ij} = \mu + \alpha_i + u_{ij}$$

siendo:

y_{ij} -> producción en la j ésima observación de la parcela *i* (*i=1,2,3,4; j=1,...,50*)

$\mu + \alpha_i$ -> producción de la parcela i, siendo por tanto μ la producción media de las cuatro parcelas. De esto se deduce que: $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0$.

u_{ij} -> error experimental o perturbación

Observando el modelo, se comprende que si α_i es igual a 0 para todo *i*, las cuatro parcelas tendrán igual producción, tomando ésta el valor μ.

El modelo estimado sería el siguiente:

$$y_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + e_{ij}$$

siendo $\bar{y}_{..}$ la estimación de la media general $\bar{y}_{i.} - \bar{y}_{..}$ la estimación del efecto de cada grupo y e_{ij} la de las perturbaciones.

Los residuos se estiman del siguiente modo:

$$e_{ij} = y_{ij} - (\bar{y}_{i.} - \bar{y}_{..}) - \bar{y}_{..} = y_{ij} - \bar{y}_{i.} + \bar{y}_{..} - \bar{y}_{..} = y_{ij} - \bar{y}_{i.}$$

TABLA DE ANÁLISIS DE LA VARIANZA

El contraste de medias se efectúa mediante la descomposición de la variabilidad total (varianza de todos los datos) en varianza explicada y no explicada.

Así, se define la variabilidad explicada (VE) y la variabilidad no explicada (VNE) como:

$$VE = \sum_{i=1}^{L} n_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$VNE = \sum_{i=1}^{L} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

verificándose que la variabilidad total (VT) es igual a:

$$VT = VE + VNE = \sum_{i=1}^{L} \sum_{j=1}^{n_i} (\bar{y}_{ij} - \bar{y}_{..})^2$$

Todo esto se resume en la denominada *tabla de análisis de la varianza*, la cual presenta el siguiente aspecto:

Tabla 4.1.

Esquema de Tabla de análisis de la varianza con un factor

Origen de las variaciones	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*F*
Entre grupos	VE	L -1	VE / (L-1) = S_e^2	S_R^2 / S_e^2
Dentro de los grupos	VNE	n - L	VNE / (n-L) = S_R^2	
Total	VT	n-1	VT / (n-1) = S_y^2	

En la última columna aparece el estadístico F del análisis de la varianza, el cual bajo la hipótesis nula (H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$, o lo que es lo mismo: todas las medias son iguales) sigue una distribución *F* con (*L - 1*) (3 en nuestro caso) y (*n - L*) (en el ejemplo 200 - 4 = 196) grados de libertad.

4.2. EJEMPLO EN SPSS

PARA OBTENER UN ANÁLISIS DE VARIANZA DE UN FACTOR

Elija en los menús:

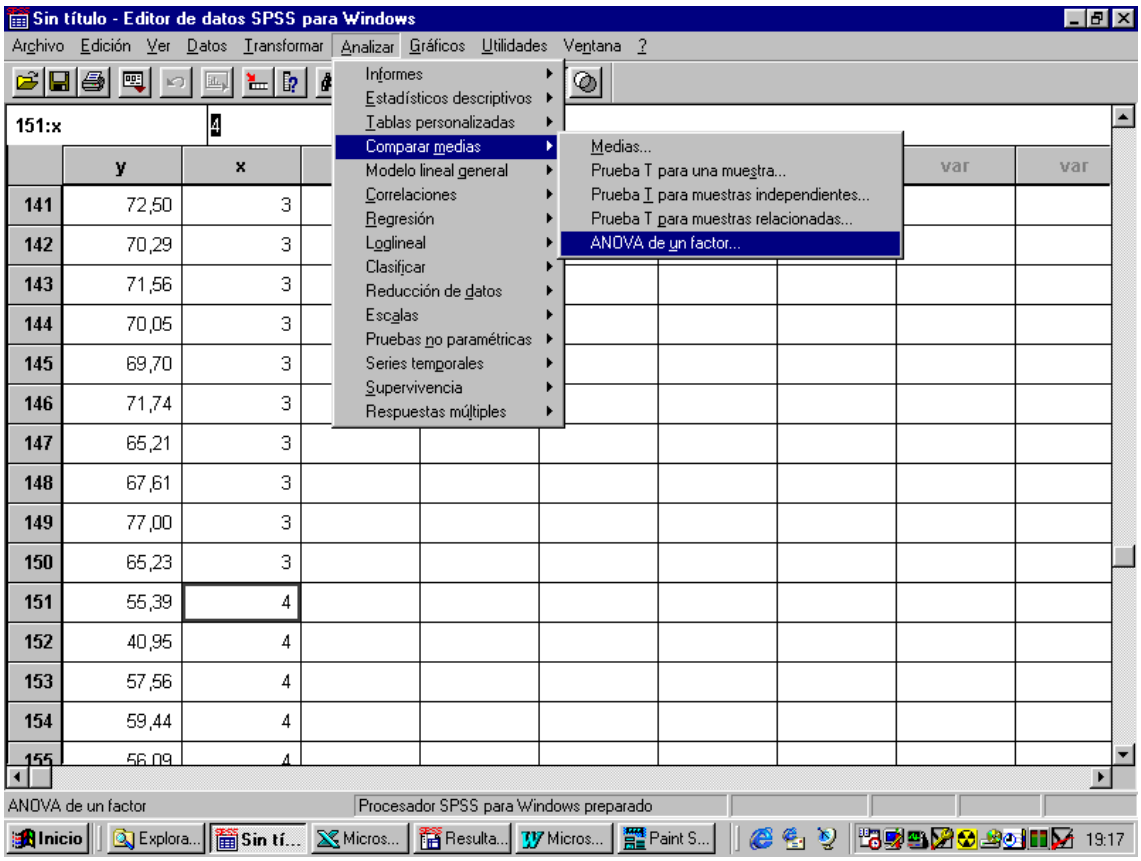
Analizar

Comparar medias

ANOVA de un factor...

Seleccione una o más variables dependientes.

Seleccione una sola variable de factor independiente.



ANÁLISIS DE LA VARIANZA DE UN FACTOR

El procedimiento ANOVA de un factor genera un análisis de varianza de un factor para una variable dependiente cuantitativa respecto a una única variable de factor (la variable independiente). El análisis de varianza se utiliza para contrastar la hipótesis de que varias medias son iguales. Esta técnica es una extensión de la prueba t para dos muestras.

Además de determinar que existen diferencias entre las medias, es posible que desee saber qué medias difieren. Existen dos tipos de contrastes para comparar medias: los contrastes a priori y las pruebas post hoc. Los contrastes a priori se plantean antes de ejecutar el experimento y las pruebas post hoc se realizan después de haber llevado a cabo el experimento. También puede contrastar las tendencias existentes a través de las categorías.

Ejemplo. Las rosquillas absorben diferentes cantidades de grasa cuando se fríen. Se plantea un experimento utilizando tres tipos de grasas: aceite de cacahuete, aceite de maíz y manteca de cerdo. El aceite de cacahuete y el aceite de maíz son grasas no saturadas y la manteca es una grasa saturada.

Además de determinar si la cantidad de grasa absorbida depende del tipo de grasa utilizada, también se podría preparar un contraste a priori para determinar si la cantidad de absorción de la grasa difiere para las grasas saturadas y las no saturadas.

Estadísticos. Para cada grupo: número de casos, media, desviación típica, error típico de la media, mínimo, máximo, intervalo de confianza al 95% para la media. Prueba de Levene sobre la homogeneidad de varianzas, tabla de análisis de varianza para cada variable dependiente, contrastes a priori especificados por el usuario y las pruebas de rango y de comparaciones múltiples post hoc: Bonferroni, Sidak, diferencia honestamente significativa de Tukey, GT2 de Hochberg, Gabriel, Dunnett, prueba F de Ryan-Einot-Gabriel-Welsch (R-E-G-W F), prueba de rango de Ryan-Einot-Gabriel-Welsch (R-E-G-W Q), T2 de Tamhane, T3 de Dunnett, Games-Howell, C de Dunett, prueba de rango múltiple de Duncan, Student-Newman-Keuls (S-N-K), Tukey b, Waller-Duncan, Scheffé y diferencia menos significativa.

CONSIDERACIONES SOBRE LOS DATOS

Datos. Los valores de la variable de factor deben ser enteros y la variable dependiente debe ser cuantitativa (nivel de medida de intervalo).

Supuestos. Cada grupo es una muestra aleatoria independiente procedente de una población normal. El análisis de varianza es robusto a las desviaciones de la normalidad, aunque los datos deberán ser simétricos. Los grupos deben proceder de poblaciones con varianzas iguales. Para contrastar este supuesto, utilice la prueba de Levene de homogeneidad de varianzas.

CONTRASTES A PRIORI

Puede dividir las sumas de cuadrados inter-grupos en componentes de tendencia o especificar contrastes a priori.

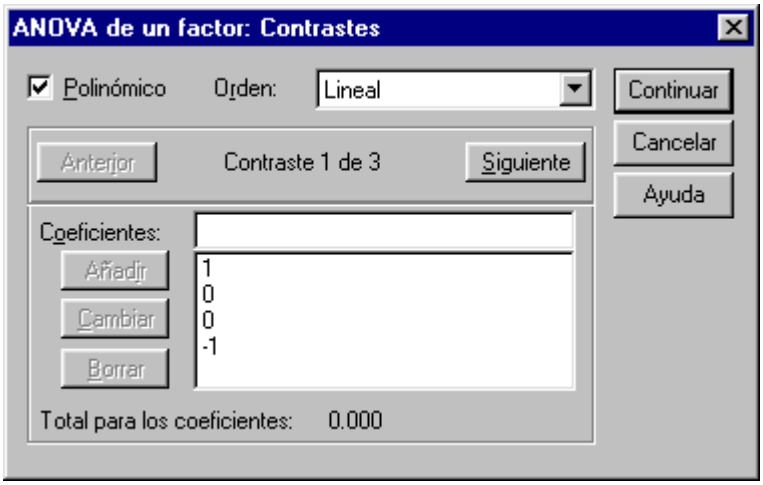
Polinómico. Divide las sumas de cuadrados inter-grupos en componentes de tendencia. Puede contrastar la existencia de tendencia en la variable dependiente a través de los niveles ordenados de la variable de factor. Por ejemplo, podría contrastar si existe una tendencia lineal (creciente o decreciente) en el salario, a través de los niveles ordenados de la titulación mayor obtenida.

Orden. Se puede elegir un orden polinómico 1º, 2º, 3º, 4º o 5º.

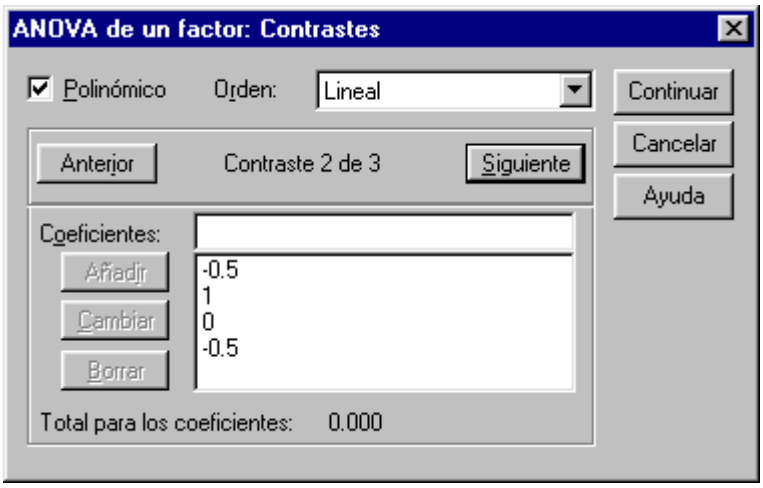
Coeficientes. Contrastes a priori especificados por el usuario que serán contrastados mediante el estadístico t. Introduzca un coeficiente para cada grupo (categoría) de la variable factor y pulse en Añadir después de cada entrada. Cada nuevo valor se añade al final de la lista de coeficientes. Para especificar conjuntos de contrastes adicionales, pulse en Siguiente. Utilice Siguiente y Previo para desplazarse entre los conjuntos de contrastes.

El orden de los coeficientes es importante porque se corresponde con el orden ascendente de los valores de las categorías de la variable de factor. El primer coeficiente en la lista se corresponde con el menor de los valores de grupo en la variable de factor y el último coeficiente se corresponde con el valor más alto. Por ejemplo, si existen seis categorías en la variable factor, los coeficientes -1, 0, 0, 0, 0,5 y 0,5 contrastan el primer grupo con los grupos quinto y sexto. Para la mayoría de las aplicaciones, la suma de los coeficientes debería ser 0. Los conjuntos que no sumen 0 también se pueden utilizar, pero aparecerá un mensaje de advertencia.
Sospechamos en nuestro caso que los abonos 1 y 4 tienen un nivel de producción similar ya que ambos incluyen un mismo compuesto. Si esto se cumple, queremos observar la diferencia entre la media de ambas producciones con la de los otros dos abonos.

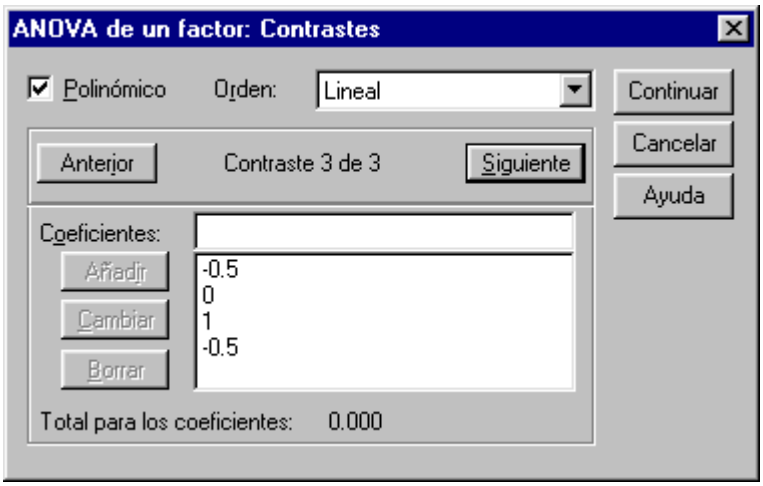
Contraste del 1º y 4º abono



Contraste del 2º con la media del 1º y 4º abono



Contraste del 3º con la media del 1º y 4º abono



CONTRASTES POST HOC

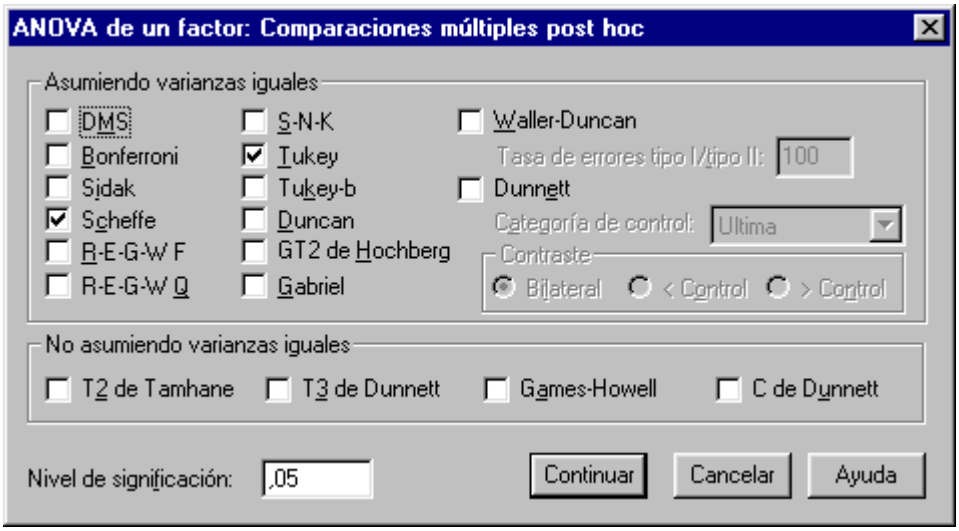
Pruebas. Una vez que se ha determinado que existen diferencias entre las medias, las pruebas de rango post hoc y las comparaciones múltiples por parejas permiten determinar qué medias difieren. Las pruebas de rango identifican subconjuntos homogéneos de medias que no se diferencian entre sí. Las comparaciones múltiples por parejas contrastan la diferencia entre cada pareja de medias y dan lugar a una matriz

donde los asteriscos indican las medias de grupo significativamente diferentes a un nivel alfa de 0,05.

La prueba de la diferencia honestamente significativa de Tukey, la GT2 de Hochberg, la prueba de Gabriel y la prueba de Scheffé son pruebas de comparaciones múltiples y pruebas de rango. Otras pruebas de rango disponibles son Tukey b, S-N-K (Student-Newman-Keuls), Duncan, R-E-G-W F (prueba F de Ryan-Einot-Gabriel-Welsch), R-E-G-W Q (prueba de rango de Ryan-Einot-Gabriel-Welsch) y Waller-Duncan. Las pruebas de comparaciones múltiples disponibles son Bonferroni, Diferencia honestamente significativa de Tukey, Sidak, Gabriel, Hochberg, Dunnett, Scheffé, y DMS (diferencia menos significativa). Las pruebas de comparaciones múltiples que no suponen varianzas iguales son T2 de Tamhane, T3 de Dunnett, Games-Howell y C de Dunnett.

Nota: Posiblemente le resulte más fácil interpretar el resultado de los contrastes post hoc si desactiva Ocultar filas y columnas vacías en el cuadro de diálogo Propiedades de tabla (en una tabla pivote activada, seleccione Propiedades de tabla en el menú Formato).

En nuestro caso, realizaremos las pruebas de Scheffé y la diferencia honestamente significativa de Tukey.

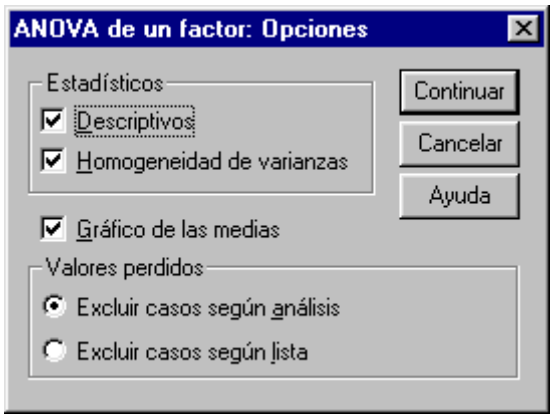


OPCIONES

Estadísticos. Elija uno o más entre los siguientes:

- Descriptivos. Calcula los siguientes estadísticos: Número de casos, Media, Desviación típica, Error típico de la media, Mínimo, Máximo y los Intervalos de confianza al 95% de cada variable dependiente para cada grupo.
- Homogeneidad de varianzas. Calcula el estadístico de Levene para contrastar la igualdad de las varianzas de grupo. Esta prueba no depende del supuesto de normalidad.
- Gráfico de medias. Muestra un gráfico que representa la medias de los subgrupos (las medias para cada grupo definido por los valores de la variable factor).
- Valores perdidos. Controla el tratamiento de los valores perdidos.
- Excluir casos según análisis. Un caso que tenga un valor perdido para la variable dependiente o la variable de factor en un análisis determinado, no se utiliza en ese análisis. Además, los casos fuera del rango especificado para la variable de factor no se utilizan.
- Excluir casos según lista. Se excluyen de todos los análisis los casos con valores perdidos para la variable de factor o para cualquier variable dependiente incluida en la lista de variables dependientes en el cuadro de diálogo principal. Si no se han especificado varias variables dependientes, esta opción no surte efecto.

En nuestro análisis señalaremos las siguientes opciones:



RESULTADOS

Descriptivos

Producción por hectárea de trigo

	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
1	50	49,4650	5,8170	,8227	47,8118	51,1182	39,08	61,88
2	50	61,3217	6,0250	,8521	59,6094	63,0340	48,51	74,90
3	50	70,2058	5,8848	,8322	68,5334	71,8783	54,39	81,60
4	50	49,5379	6,4286	,9091	47,7109	51,3649	37,03	60,78
Total	200	57,6326	10,5989	,7495	56,1547	59,1105	37,03	81,60

Prueba de homogeneidad de varianzas

Producción por hectárea de trigo

Estadístico de Levene	gl1	gl2	Sig.
,353	3	196	,787

Se asume la igualdad de varianzas entre los grupos.

ANOVA

Producción por hectárea de trigo

			Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	(Combinadas)		15196,488	3	5065,496	138,691	,000
	Término lineal	Contraste	207,154	1	207,154	5,672	,018
		Desviación	14989,334	2	7494,667	205,200	,000
Intra-grupos			7158,652	196	36,524		
Total			22355,140	199			

Según el test de la F, existen diferencias significativas entre las producciones de los distintos abonos.

Coeficientes de contraste

	Abono utilizado			
Contraste	1	2	3	4
1	1	0	0	-1
2	-,5	1	0	-,5
3	-,5	0	1	-,5

Pruebas de contraste

		Contraste	Valor de contraste	Error típico	t	gl	Sig. (bilateral)
Producción por hectárea de trigo	Suponer igualdad de varianzas	1	-7,29E-02	1,2087	-,060	196	,952
		2	11,8203	1,0468	11,292	196	,000
		3	20,7044	1,0468	19,779	196	,000
	No asume igualdad de varianzas	1	-7,29E-02	1,2261	-,059	97,037	,953
		2	11,8203	1,0497	11,261	99,409	,000
		3	20,7044	1,0337	20,030	101,511	,000

Se observa que, efectivamente, las medias para el primer y cuarto abono no son significativamente diferentes.

Las medias del segundo y tercer abono son significativamente mayores que las del primero y el cuarto.

Comparaciones múltiples

Variable dependiente: Producción por hectárea de trigo

	(I) Abono utilizado	(J) Abono utilizado	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
						Límite inferior	Límite superior
HSD de Tukey	1	1					
		2	-11,8567*	1,2087	,000	-14,9619	-8,7515
		3	-20,7408*	1,2087	,000	-23,8460	-17,6357
		4	-7,2904E-02	1,2087	1,000	-3,1781	3,0323
	2	1	11,8567*	1,2087	,000	8,7515	14,9619
		2					
		3	-8,8841*	1,2087	,000	-11,9893	-5,7789
		4	11,7838*	1,2087	,000	8,6786	14,8890
	3	1	20,7408*	1,2087	,000	17,6357	23,8460
		2	8,8841*	1,2087	,000	5,7789	11,9893
		3					
		4	20,6679*	1,2087	,000	17,5628	23,7731
	4	1	7,290E-02	1,2087	1,000	-3,0323	3,1781
		2	-11,7838*	1,2087	,000	-14,8890	-8,6786
		3	-20,6679*	1,2087	,000	-23,7731	-17,5628
		4					
Scheffé	1	1					
		2	-11,8567*	1,2087	,000	-15,2652	-8,4483
		3	-20,7408*	1,2087	,000	-24,1493	-17,3324
		4	-7,2904E-02	1,2087	1,000	-3,4814	3,3355
	2	1	11,8567*	1,2087	,000	8,4483	15,2652
		2					
		3	-8,8841*	1,2087	,000	-12,2926	-5,4757
		4	11,7838*	1,2087	,000	8,3754	15,1923
	3	1	20,7408*	1,2087	,000	17,3324	24,1493
		2	8,8841*	1,2087	,000	5,4757	12,2926
		3					
		4	20,6679*	1,2087	,000	17,2595	24,0764
	4	1	7,290E-02	1,2087	1,000	-3,3355	3,4814
		2	-11,7838*	1,2087	,000	-15,1923	-8,3754
		3	-20,6679*	1,2087	,000	-24,0764	-17,2595
		4					

*. La diferencia entre las medias es significativa al nivel .05.

De la tabla anterior, se deduce que también el 2º y 3º abono presentan medias distintas, correspondiendo la mayor producción al 3º.

Producción por hectárea de trigo

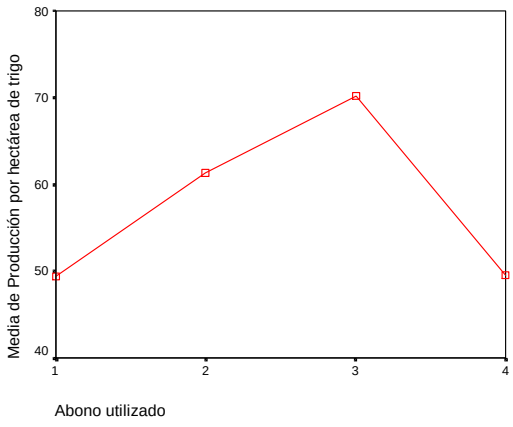
	Abono utilizado	N	Subconjunto para alfa = .05		
			1	2	3
HSD de Tukey[a]	1	50	49,4650		
	4	50	49,5379		
	2	50		61,3217	
	3	50			70,2058
	Sig.		1,000	1,000	1,000
Scheffé[a]	1	50	49,4650		
	4	50	49,5379		
	2	50		61,3217	
	3	50			70,2058
	Sig.		1,000	1,000	1,000

Se muestran las medias para los grupos en los subconjuntos homogéneos.

a. Usa tamaño de la muestra de la media armónica = 50,000.

La tabla inferior nos muestra los subconjuntos homogéneos detectados en el análisis, utilizando por una parte el test de la diferencia honestamente significativa de Tukey y por otra el de Scheffé. Se observa un subconjunto homogéneo formado por dos abonos, el 1º y el 4º.

Por último, se muestra a continuación el gráfico de medias para los distintos abonos.



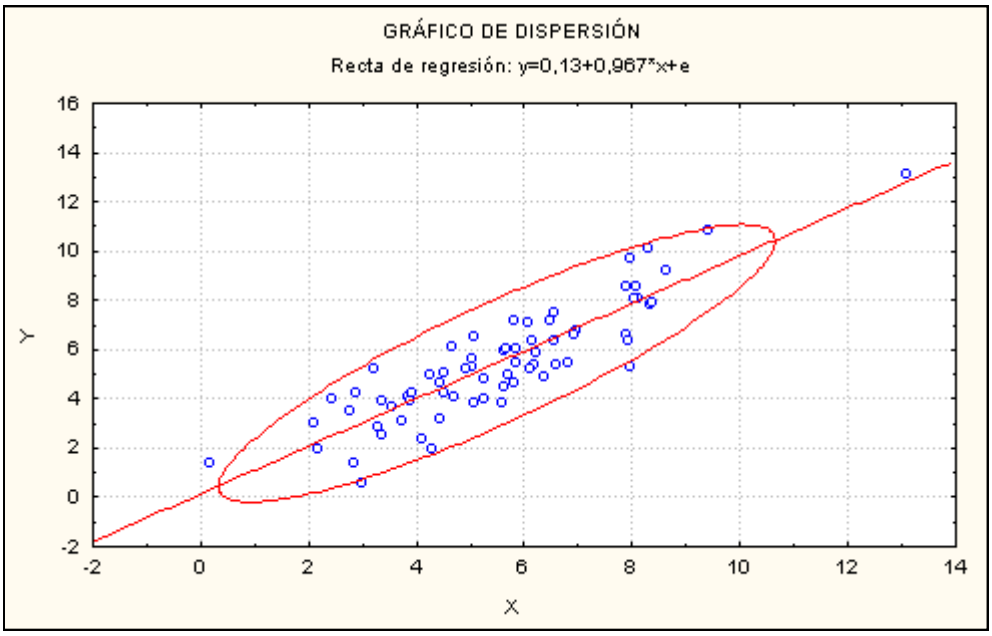
4.4. ANÁLISIS DE REGRESIÓN

El *análisis de regresión* de dos variables nos permite ajustar una línea a la serie de observaciones (x_i, y_i) que obtenemos con dos variables aleatorias *X*, Y. Cuando realizamos un diagrama de dispersión con ellas, obtenemos una representación que se denomina nube de puntos (gráfico 3.12.), que nos ayuda a conocer si las dos variables están relacionadas. Si suponemos la existencia de una relación lineal, la función a ajustar será:

$$y_i = a + bx_i$$

Gráfico 4.1.

Nube de puntos o gráfico de dispersión con variables relacionadas linealmente



Para calcular los coeficientes *a* y *b* de esta recta de regresión tenemos que minimizar las distancias al cuadrado de los puntos a la recta (estimación por mínimos cuadrados), es decir:

$$\text{Minimizar } \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Derivando esta expresión respecto a los coeficientes *a* y *b* e igualando a cero obtenemos el siguientes sistema de ecuaciones:

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} y_i x_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

A los términos $e_i = y_i - a - bx_i$ se les denomina *residuos*, y expresan la diferencia entre los valores observados para la variable y los predichos a través de la recta de regresión.

Despejando *a* en la primera ecuación: $a = \bar{y} - b\bar{x}$

Sustituyendo *a* por su valor en la segunda ecuación:

$\sum_{i=1}^{n} y_i x_i = n\overline{yx} - b\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} + b\sum_{i=1}^{n} x_i^2$ de lo cual se obtiene que:

$\sum_{i=1}^{n} y_i x_i - n\overline{yx} = b\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$ y dividiendo ambos términos por "n":

$$Cov(x,y) = bS_x^2 \quad => \quad b = \frac{Cov(x,y)}{S_x^2}$$

Siendo S_x^2 la varianza de la variable *x* o regresor.

Sustituyendo *b* por su valor en la ecuación $a = \bar{y} - b\bar{x}$ obtenemos el valor de *a*.

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

El modelo de regresión requiere que se cumplan las siguientes hipótesis sobre los residuos:

$E(e_i) = 0 \quad i=1,\ldots,n$

$Var(e_i) = \sigma^2$ constante $\quad i=1,\ldots,n \quad$ *(Supuesto de homocedasticidad)*

$Cov(e_i,e_j) = 0 \quad \forall\, i,j \quad i \neq j \quad i = 1,\ldots,n \quad j = 1,\ldots,n \quad$ *(Ausencia de autocorrelación)*

Estas hipótesis, como vemos, inciden en el carácter aleatorio de los residuos.

La variabilidad del modelo viene expresada por la desviación estándar de los residuos (diferencia entre el valor de y_i menos la recta estimada de regresión). Esta medida se calcula por la siguiente fórmula:

$$S_R = \sqrt{\frac{\sum_i^n (y_i - a - bx_i)^2}{n - 2}}$$

El divisor de la fórmula anterior viene determinado por el número de observaciones menos el número de parámetros a estimar en el modelo (*a* y *b*).

Sin embargo, hay que tener presente que esta medida no es útil para comparar rectas de regresión de variables distintas ya que depende de las unidades de medida de la variable *y*.

La medida utilizada para medir el ajuste del modelo a los datos es el coeficiente de determinación R^2, que se define como el cociente entre la variabilidad explicada por el modelo ajustado y la variabilidad total, y cuya expresión es la siguiente.

$$R^2 = \frac{VE}{VT} = 1 - \frac{VNE}{VT} = 1 - \frac{(n-2)S_R^2}{(n-1)S_y^2}$$

En el caso del análisis de regresión con una única variable dependiente este coeficiente coincide con el coeficiente de correlación al cuadrado.

Destacar que mediante el uso de transformaciones en los datos se pueden estimar relaciones no lineales.

Por ejemplo, dada la siguiente ecuación no lineal $y_i = ax_i^b$, aplicando logaritmos en ambos términos se obtiene la siguiente relación lineal:

$$log\, y_i = log\, a + b\, log\, x_i$$

Si denominamos:

$$Y_i = \log y_i$$
$$A = log\, a$$
$$X_i = log\, x_i$$

entonces

$$Y_i = A + bX_i$$

se puede estimar por el procedimiento de mínimos cuadrados ordinarios.

Si observamos una relación no lineal como la definida por la ecuación $y = x^2$ (ver ejemplo gráfico 3.13.), aplicando la transformación logarítmica obtenemos:

$$Y = log\, y$$
$$X = log\, x$$

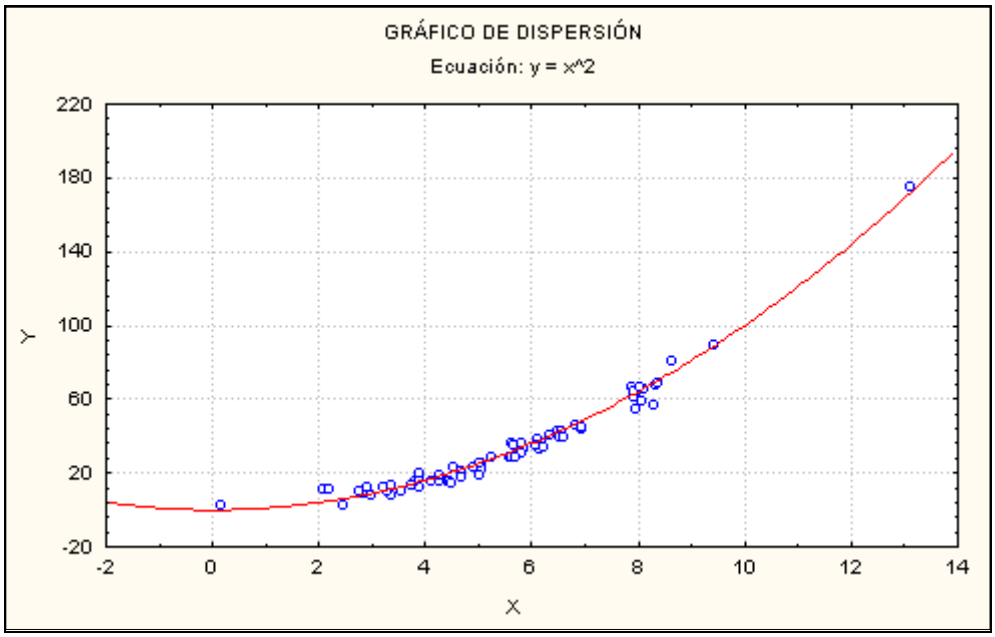
La recta resultante mediante esta transformación es la siguiente:

$$Y = 2X$$

Deshaciendo la transformación, llegaríamos a la relación realmente existente.

Gráfico 4.2.

Nube de puntos o gráfico de dispersión con variables relacionadas de forma no lineal



Si tenemos más de una variable explicativa, se supone que cada una de ellas está incorrelacionada con el resto. Si esto no es así, existen tres modos de proceder:

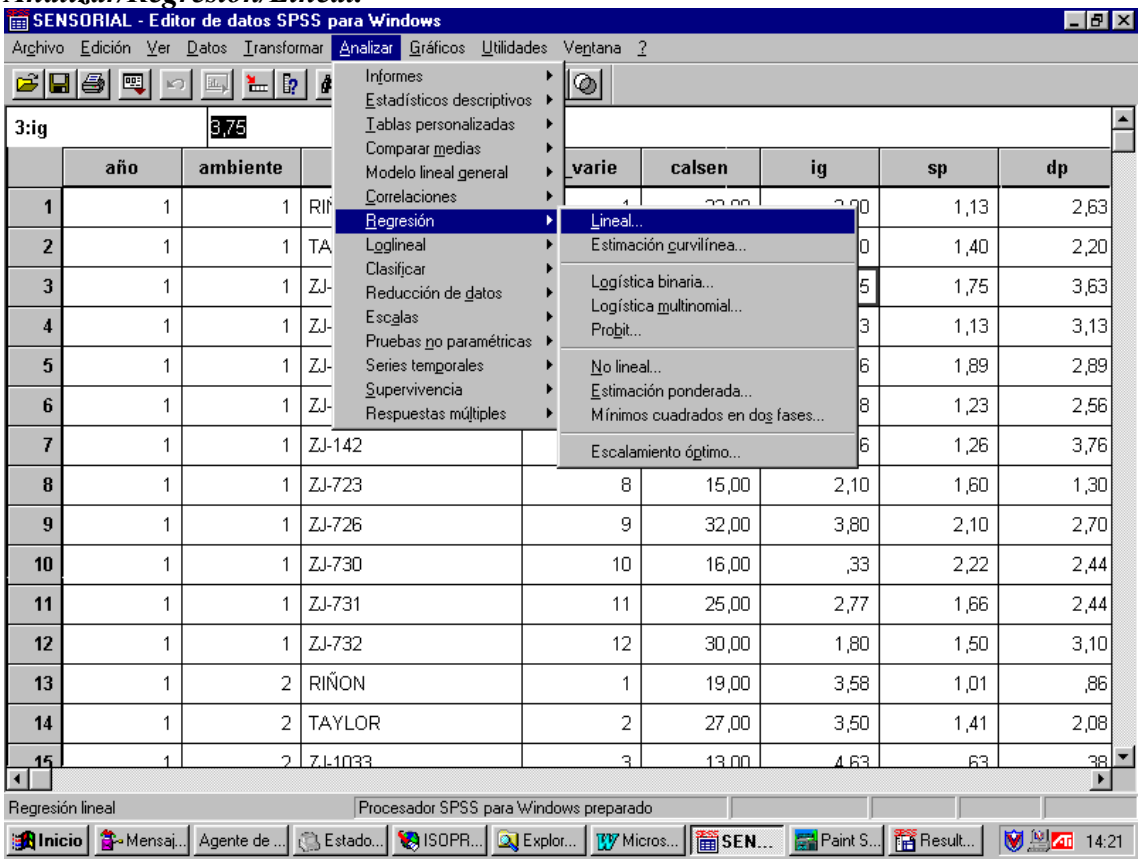
- Eliminar regresores (métodos de selección de variables).
- Incluir información externa a los datos.
- Regresión Ridge o contraída.
- Regresión en componentes principales.

En el ejemplo que se desarrolla a continuación se tienen tres regresores altamente correlacionados, y se utiliza el denominado método de selección paso a paso o stepwise, el cual, en cada iteración añade la variable más relevante para el modelo o suprime la menos relevante.

4.14. EJEMPLO EN SPSS

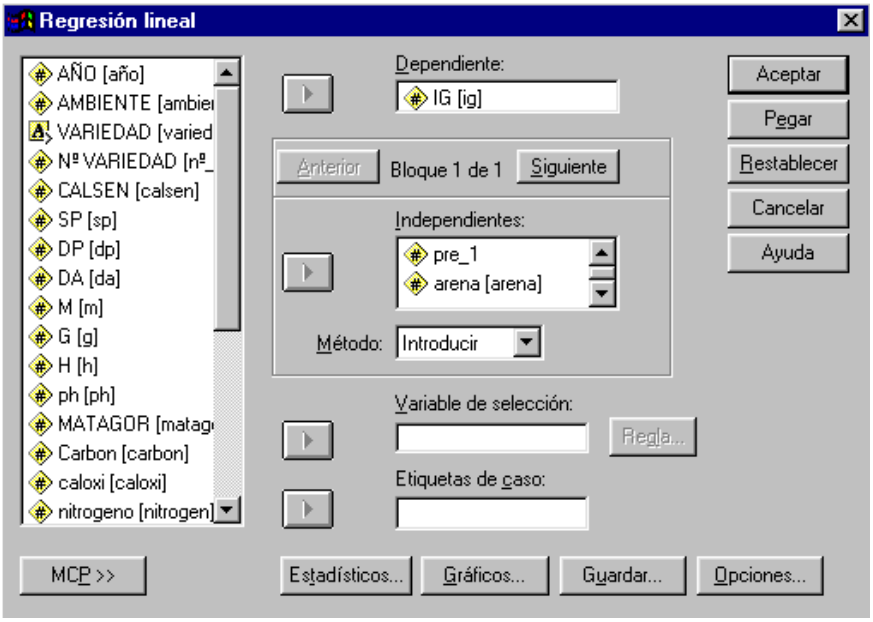
A continuación realizamos en SPSS un ejemplo de regresión multiple de una variable dependiente (integridad del grano de diferentes variedades de alubias) y dos variables independientes: valores medios por variedades y el % de arena en el suelo. Mediante este ejercicio pretendemos averiguar como ambos aditivos influyen en la viscosidad que hemos observado en dicho fluido.

Para realizar un análisis de regresión en SPSS habrá que seleccionar el menú ***Analizar/Regresión/Lineal.***

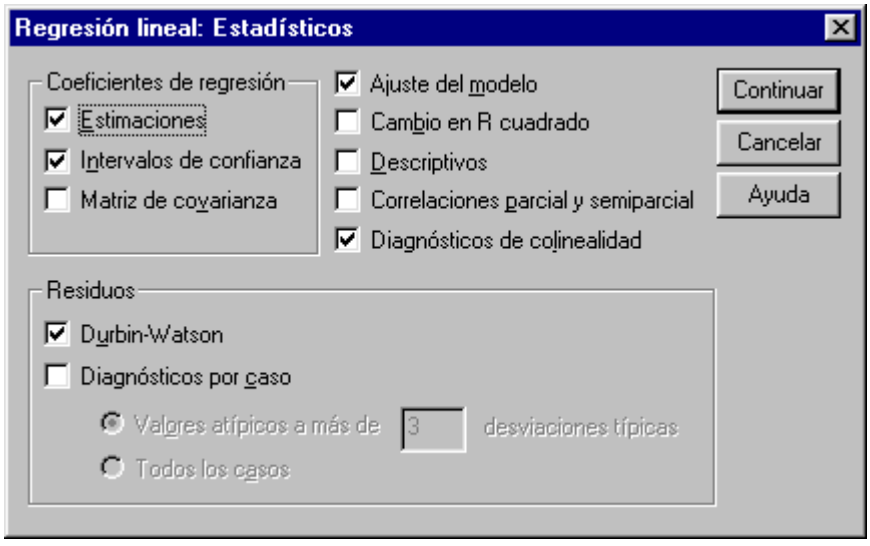


	año	ambiente		varie	calsen	ig	sp	dp
1	1	1	RIÑ	1			1,13	2,63
2	1	1	TA				1,40	2,20
3	1	1	ZJ-				1,75	3,63
4	1	1	ZJ-				1,13	3,13
5	1	1	ZJ-				1,89	2,89
6	1	1	ZJ-				1,23	2,56
7	1	1	ZJ-142				1,26	3,76
8	1	1	ZJ-723	8	15,00	2,10	1,60	1,30
9	1	1	ZJ-726	9	32,00	3,80	2,10	2,70
10	1	1	ZJ-730	10	16,00	,33	2,22	2,44
11	1	1	ZJ-731	11	25,00	2,77	1,66	2,44
12	1	1	ZJ-732	12	30,00	1,80	1,50	3,10
13	1	2	RIÑON	1	19,00	3,58	1,01	,86
14	1	2	TAYLOR	2	27,00	3,50	1,41	2,08
15	1	2	ZJ-1033	3	13,00	4,63	,63	38

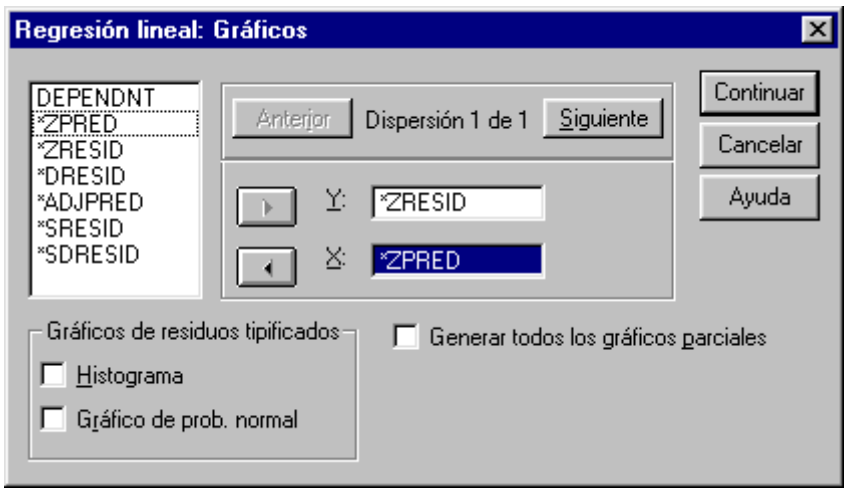
Especificamos en la siguiente pantalla la variable dependiente, las independientes y en método elegimos el método de selección Pasos sucesivos (Pasos suc.).



A continuación pulsamos en el botón Estadísticos y seleccionamos las estimaciones de los coeficientes de regresión, los intervalos de confianza, el ajuste del modelo, el diagnóstico de multicolinealidad (detección de la existencia de correlación en las variables predictoras)y el estadístico de Durbin-Watson para detectar la correlación entre los residuos.



A través del botón gráfico, realizamos el gráfico de residuos por valores pronosticados.



Por último pulsamos aceptar y obtenemos los siguientes resultados:

Regresión

Variables introducidas/eliminadas[b]

Modelo	Variables introducidas	Variables eliminadas	Método
1	arena, Valor medio por variedades[a]	,	Introducir

a. Todas las variables solicitadas introducidas

b. Variable dependiente: IG

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,926[a]	,857	,854	,5196

a. Variables predictoras: (Constante), arena, Valor medio por variedades

b. Variable dependiente: IG

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	145,374	2	72,687	269,230	,000[a]
	Residual	24,298	90	,270		
	Total	169,672	92			

a. Variables predictoras: (Constante), arena, Valor medio por variedades

b. Variable dependiente: IG

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	1,355	,483		2,807	,006		
	Valor medio por variedades	,990	,044	,909	22,706	,000	,993	1,007
	arena	-1,90E-02	,007	-,117	-2,917	,004	,993	1,007

a. Variable dependiente: IG

Diagnósticos de colinealidad[a]

Modelo	Dimensión	Autovalor	Indice de condición	Proporciones de la varianza		
				(Constante)	Valor medio por variedades	arena
1	1	2,871	1,000	,00	,02	,00
	2	,122	4,852	,01	,93	,02
	3	6,672E-03	20,745	,99	,05	,98

a. Variable dependiente: IG

Estadísticos sobre los residuos[a]

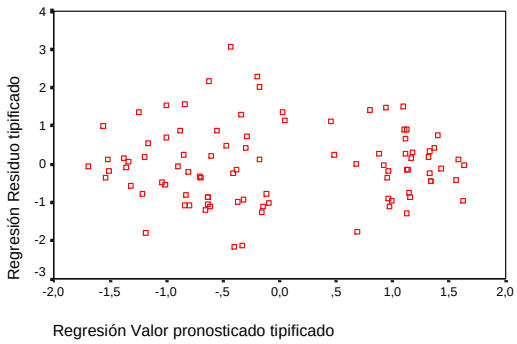
	Mínimo	Máximo	Media	Desviación típ.	N
Valor pronosticado	,6164	4,8105	2,7529	1,2570	93
Residual	-1,1221	1,5898	1,504E-16	,5139	93
Valor pronosticado tip.	-1,700	1,637	,000	1,000	93
Residuo tip.	-2,160	3,060	,000	,989	93

a. Variable dependiente: IG

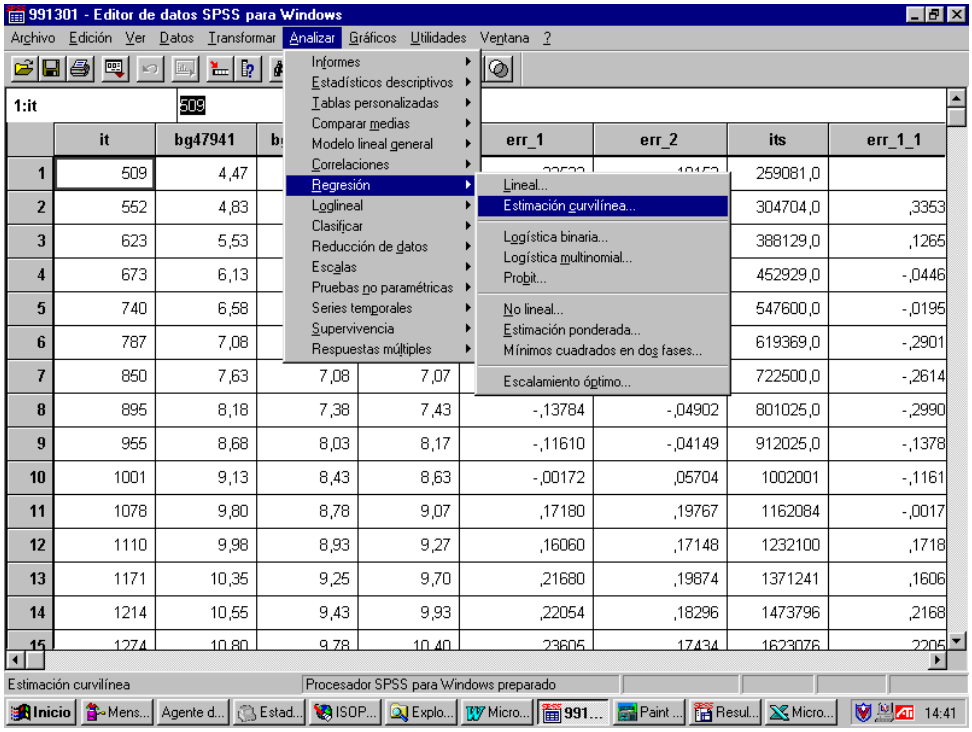
Gráficos

Gráfico de dispersión

Variable dependiente: IG

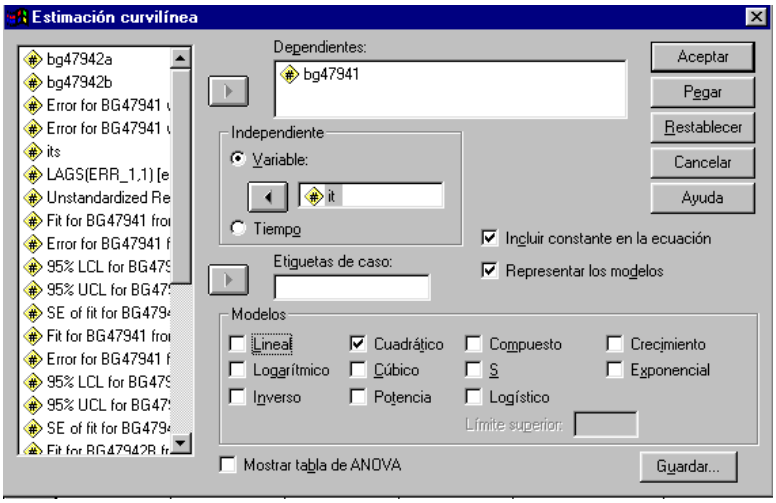


A continuación vamos a estimar una relación no lineal entre el numero de hojas de una planta y su integral térmica (IT).



Elegimos el siguiente modelo explicativo cuadrático para evaluar dicha relación:

$$IH = a + b.IT + c.IT^2 + \mu$$



Obteniendo los siguientes resultados en SPSS:

Estimación curvilínea

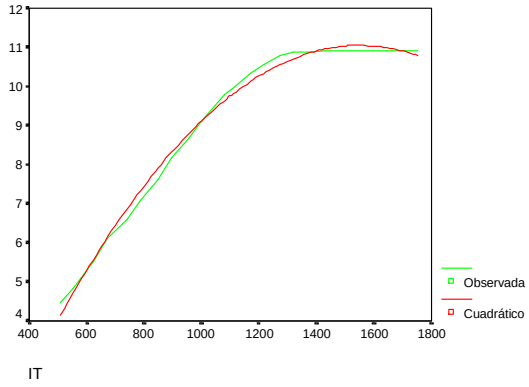
```
MODEL:  MOD_1.
```

—

```
Independent:  IT
```

Dependent	Mth	Rsq	d.f.	F	Sigf	b0	b1	b2
BG47941	QUA	,994	21	1620,59	,000	-4,2644	,0198	-6,E-06

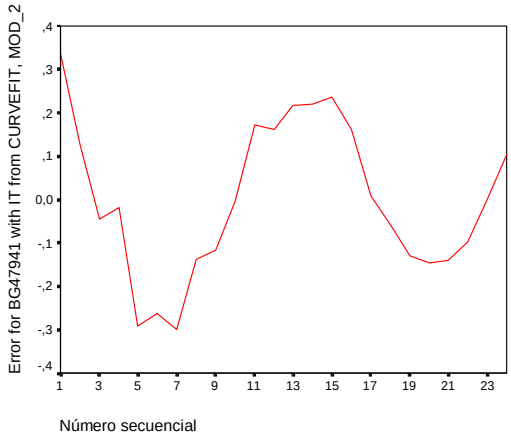
BG47941



A pesar de que el ajuste es bueno los residuos son poco aceptables porque están correlacionados:

TSPLOT

```
MODEL:  MOD_3.
```



4.4. SERIES TEMPORALES

Las series temporales son conjuntos de observaciones sobre una variable realizadas a intervalos regulares de tiempo. La serie temporal puede estar generada con datos continuos o discretos[2], flujos o stocks[3], valorados en pesetas o en magnitudes físicas, con periodicidad de datos diarios, semanales, mensuales, trimestrales o anuales, bianuales, etc. pero lo que caracteriza a la serie temporal es la presencia de una referencia cronológica concreta y determinada.

El análisis tradicional de una serie temporal se basa en considerar que la serie temporal se puede dividir en cuatro componentes diferenciados, llamados tendencia (*T*), fluctuación cíclica (*C*), variación estacional (*S*) y movimientos irregulares (*I*). La tendencia de una serie temporal representa la evolución de la serie en el largo plazo. El ciclo refleja las fluctuaciones a corto y medio plazo en torno a la tendencia. La variación estacional se produce cuando la serie temporal muestra un comportamiento regular y repetitivo a lo largo de un período de tiempo (un año) y por último, denominamos movimientos irregulares a las fluctuaciones que son ocasionales o tienen carácter impredecible o aleatorio. La asociación de los cuatro componentes en la serie temporal (*Y*) puede ser aditiva:

$$Y=T+C+S+I$$

multiplicativa

$$Y=TCSI$$

---

[2] Un ejemplo de una serie temporal de datos discretos es la que se genera a partir de las opiniones de los empresarios en la Encuesta Trimestral de Opiniones de Castilla y León. En dicha encuesta se pregunta, entre otras cosas, a una muestra regional de empresarios si sus ventas: aumentan, disminuyen o se mantienen. La serie como se ve se genera a partir de respuestas categóricas o datos discretos.

[3] Son datos flujo datos generados en un período determinado de tiempo : un día, un mes, un año, etc... y datos stock los referidos a una fecha determinada: 31 de diciembre de cada año. Un ejemplo de datos flujos son las ventas de una empresa ya que éstas tendrán un valor si se toma al cabo de un día, una semana, un mes ó un año; sin embargo, el valor de las acciones de esa misma empresa solo puede ser registrado a una fecha determinada por ejemplo a 31 de diciembre. Nótese que con datos stock también se puede tomar una serie diaria, semanal, mensual o anual, lo que dependerá de la frecuencia con la que registremos el dato, si lo hacemos cuando cierra la jornada de la bolsa generaremos una serie diaria, si lo hacemos únicamente un día determinado de la semana estaremos generando una serie semanal, si fuera a determinada fecha de cada mes una mensual o si lo hacemos al finalizar el año una serie anual.

o una combinación de ambas, por ejemplo:

$$Y=TCS+I$$

- **LA TENDENCIA**

La tendencia es el componente de la serie temporal que representa la evolución a largo plazo de la serie. La tendencia se asocia al movimiento uniforme o regular observado en la serie durante un período de tiempo extenso. La tendencia es la información más relevante de la serie temporal ya que informa si dentro de cinco, diez o quince años tendrá un nivel mayor, menor o similar al que la serie tiene hoy día.

Analizamos la tendencia con dos objetivos diferentes: para conocer cuales son las pautas de comportamiento a lo largo del tiempo de la variable objeto de estudio, y para predecir sus valores futuros. En este apartado se examinarán los métodos clásicos de análisis de la tendencia: los semipromedios, ajustes de una función por mínimos cuadrados y el método de los promedios móviles.

Las tendencias suelen representarse mediante funciones de tiempo continuas y diferenciables. Las funciones de tendencia más utilizadas son:

1. Lineal.
2. Polinómica.
3. Exponencial.
4. Modelo autoregresivo
5. Función
6. Curva de Gompertz
7. Modelo logarítmico recíproco

Si una serie temporal X_t se ajusta a una tendencia lineal, la función de tiempo que se plantea es la siguiente:

$$X_t = \alpha + \beta t_{nt} \quad t= 1, 2, ..,t$$

Una tendencia polinómica de grado p se ajustará a una función del siguiente tipo:

$$f(t) = \alpha+\beta_1 + \beta_2 t^2 + \ldots+\beta_p t^p$$

Si la tendencia sigue una ley exponencial, entonces la función de ajuste será:

$$f(t) = ae^{rt}$$

donde *a* y *r* son constantes.

Un modelo autoregresivo ajusta la tendencia de la forma siguiente

$$Xt = \gamma_0+\gamma_1 x_{t-1} + u_t \qquad \text{siendo } \gamma>0$$

La curva logística se representa mediante la función:

$$T(t) = \frac{T}{1-be^{-rt}}$$

donde *t*, *b* y *r* son constantes positivas.

La curva de Gompertz responde a la siguiente ecuación:

$$T(t) = T.b^{e-rt}$$

donde *T, r, b* son parámetros positivos.

Finalmente, señalar que el modelo logaritmo recíproco, viene definido por la relación:

$$T(t) = a + b\ 1/t \qquad B<0$$

Para calcular las funciones de tendencia, lo habitual es linearizar las formas de las funciones no lineales y proceder a su estimación como si fuera una función de tendencia lineal.

- **VARIACIONES CÍCLICAS Y ESTACIONALES.**

Entendemos por variación cíclica las variaciones regulares que se producen en las series temporales con periodo superior a un año. De hecho una serie temporal puede

estar originada por diversos ciclos: un ciclo de medio plazo, otro ciclo de largo plazo, etc.

Un ciclo tiene dos componentes básicos: la amplitud o la distancia que media entre el cero y el máximo valor que alcanza el ciclo, y el periodo o el tiempo que tarda en ocurrir un ciclo completo.

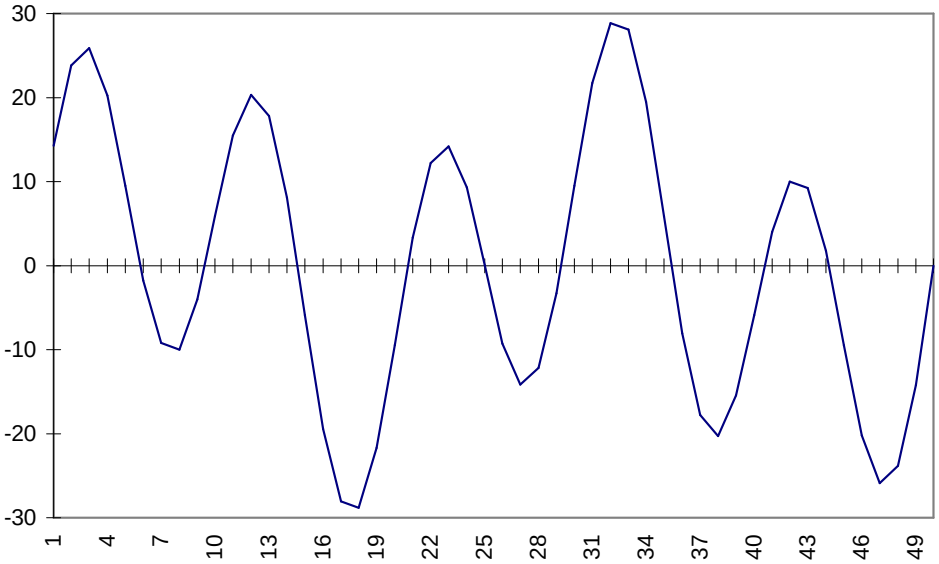
En teoría cabe entender una serie temporal como una suma de un número indeterminado de ciclos de amplitud y período diferentes, y puede demostrarse que la varianza que muestra en el tiempo una serie temporal se obtiene a partir de la suma de las amplitudes de los diferentes ciclos en que se descompone la serie temporal (relación de Parseval).

En el gráfico 6.5. representamos una serie temporal construida a partir dos ciclos de seno, uno de período 4 y amplitud 10, y otro de período 10 y amplitud 20. La representación gráfica de la serie reproduce los dos ciclos, el que tiene lugar cada 25 periodos, es decir, 4 veces cada 100 periodos y el que tiene lugar cada 10 periodos, 10 veces cada 100 periodos. La serie temporal descrita obedece a la siguiente ecuación:

$$x(t) = 10\,sen\left(2\pi t/25\right) + 20\,sen\left(2\pi t/10\right)$$

Gráfico 4.3.

Serie temporal formada por dos ciclos de frecuencia (4/100) y (10/100).



En la serie temporal representada, la varianza sería: 10+20.

Para conocer los ciclos que dominan la evolución temporal de la serie temporal se utiliza las Transformadas de Fourier, que Excel incluye en la macro de Herramientas para el Análisis. Dicha Transformada de Fourier es una función de números complejos que puede operarse en Excel a través de las funciones de ingeniería del menú Función.

La Transformada de Fourier, *F(u),* se define para una función continua de variable real, *f(x),* mediante la siguiente formula:

$$F(u) = \int_{-\infty}^{\infty} f(x) e^{[-2\pi i x u]} dx$$

siendo $i = \sqrt{-1}$, $e^{[2\pi iux]} = \cos(2\pi ux) + \text{isen}(2\pi ux)$ y u una variable que representa las distintas frecuencias.

Esta función tiene transformada inversa, lo que significa que ha partir de la función *F(u)* podemos calcular la función *f(x)*:

$$f(x) = \int_{-\infty}^{\infty} F(u)e^{[-2\pi ixu]}du$$

Para que una función tenga Transformada de Fourier han de verificarse algunas condiciones (Condiciones de Dieterlich). No obstante, hay que destacar que, por regla general, las funciones con las que tratamos los problemas reales verifican todas las condiciones que es necesario imponer para que las expresiones anteriores puedan calcularse.

Como ya se ha señalado, la Transformada de Fourier es una función compleja con una parte real y otra parte imaginaria, es decir:

$$F(u) = R(u) + I(u)$$

donde R(u) es la parte real y I(u) es la parte imaginaria.

La representación gráfica de la función de magnitud $|F(u)|$ se le denomina Espectro de Fourier y se expresa en términos del modulo del número complejo:

$$|F(u)| = \sqrt{R^2(u) + I^2(u)}$$

y al cuadrado de dicha función $|F(u)|^2$ se le denomina Espectro de potencias.

Por su parte, la representación gráfica de su ángulo de fase recibe el nombre de Función de fase:

$$\phi(u) = arctg\left[\frac{I(u)}{R(u)}\right]$$

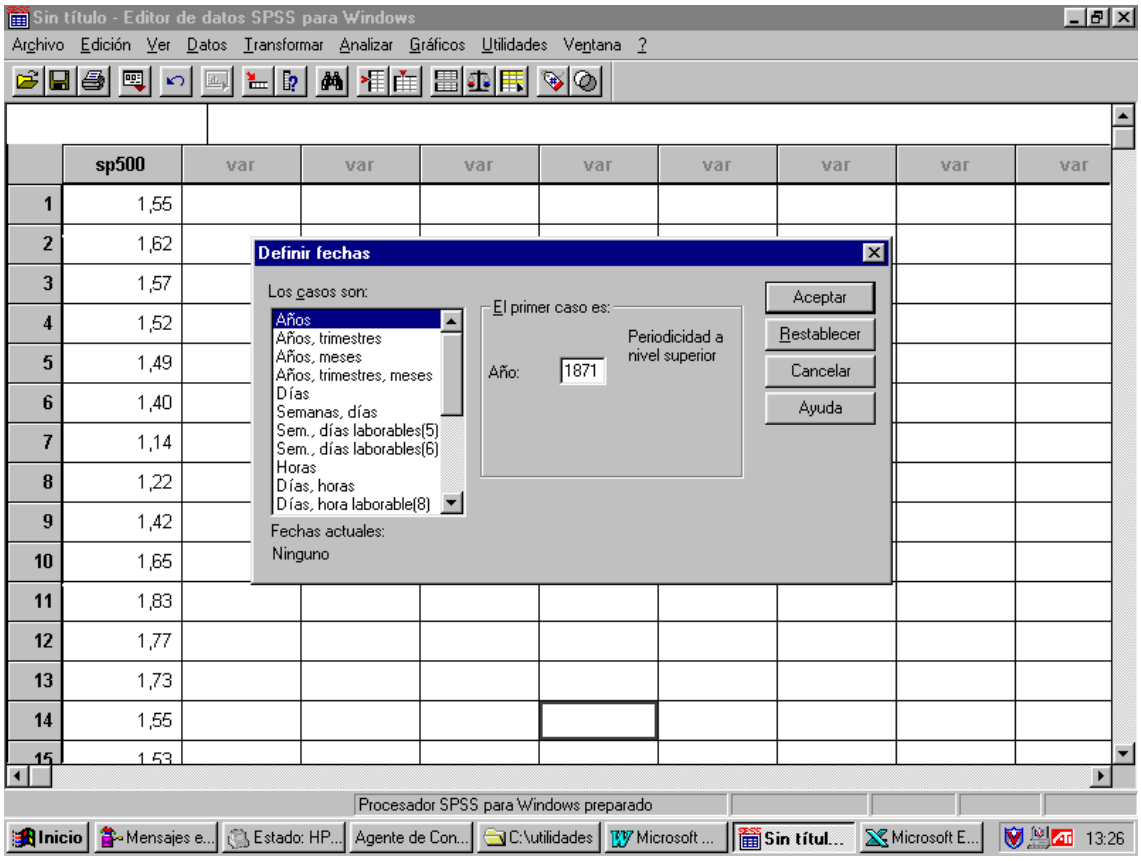
Las series temporales no son consideradas funciones continuas como tal, sino muestras de señales continuas tomadas a una misma distancia temporal a partir de un valor inicial x_0. El par de Transformadas de Fourier Discretas asociadas a una sucesión finita de valores se obtiene entonces a través de las siguientes expresiones:

$$F(u) = \frac{1}{N} \sum_{x=0}^{N-1} f(x)\, e^{[-2i\pi ux/N]} \quad \text{para } u=0,1,...,N-1$$

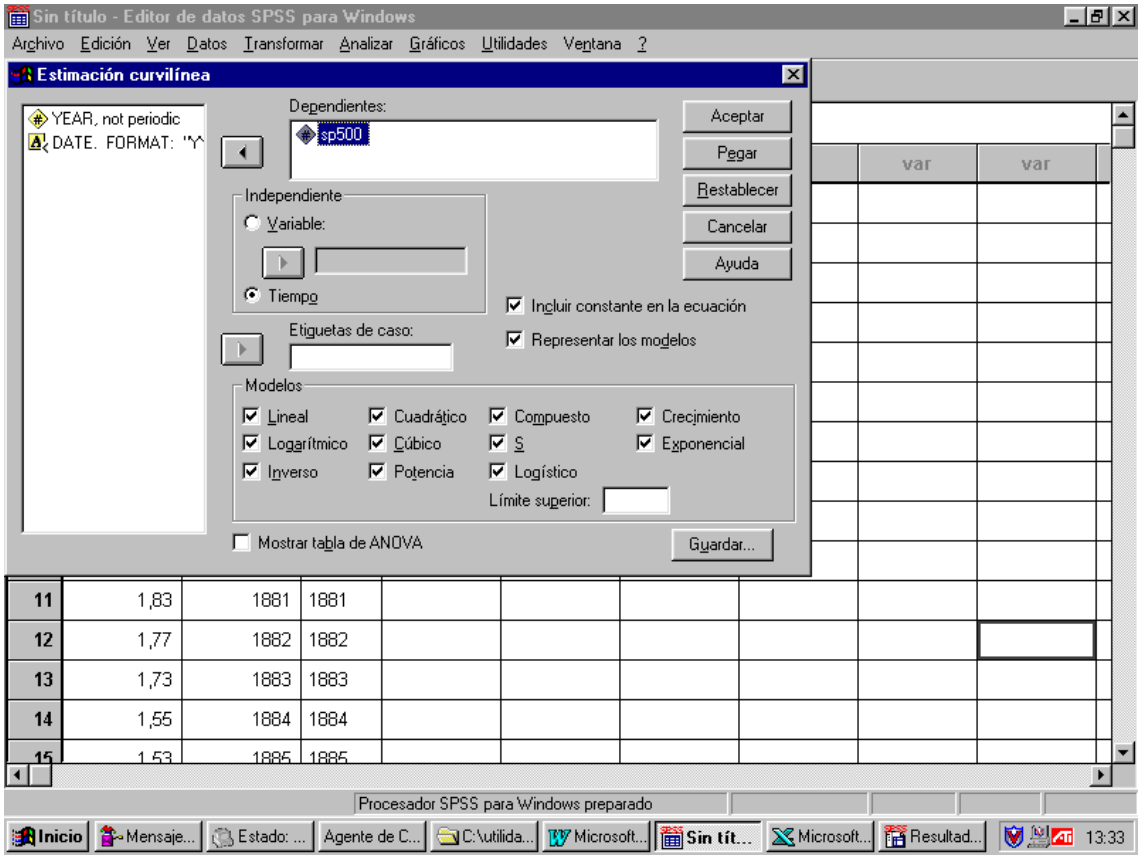
$$f(x) = \sum_{x=0}^{N-1} f(u)\, e^{[2i\pi ux/N]} \quad \text{para } x=0,1,...,N-1$$

4.5 Ejemplo de series temporales en SPSS

A continuación vamos a calcular las tendencias en SPSS del índice de bolsa sp500, para el que definimos unas fechas que empiezan en el año 1871.



Para ajustar tendencias en SPSS se utiliza el menú regresión, estimación curvilinea. Dicho menú nos permite seleccionar diferentes modelos de tendencia (lineal, exponencial, cuadrático, cúbico, etc...), y realizar una comparativa para ver cual de ellos es el que mejor se ajusta.



Como se puede apreciar, hemos realizado una selección de todos los modelos, incluyendo una constante en la función y utilizando como regresor la variable temporal. A continuación se presentan los resultados obtenidos.

Independent: Time

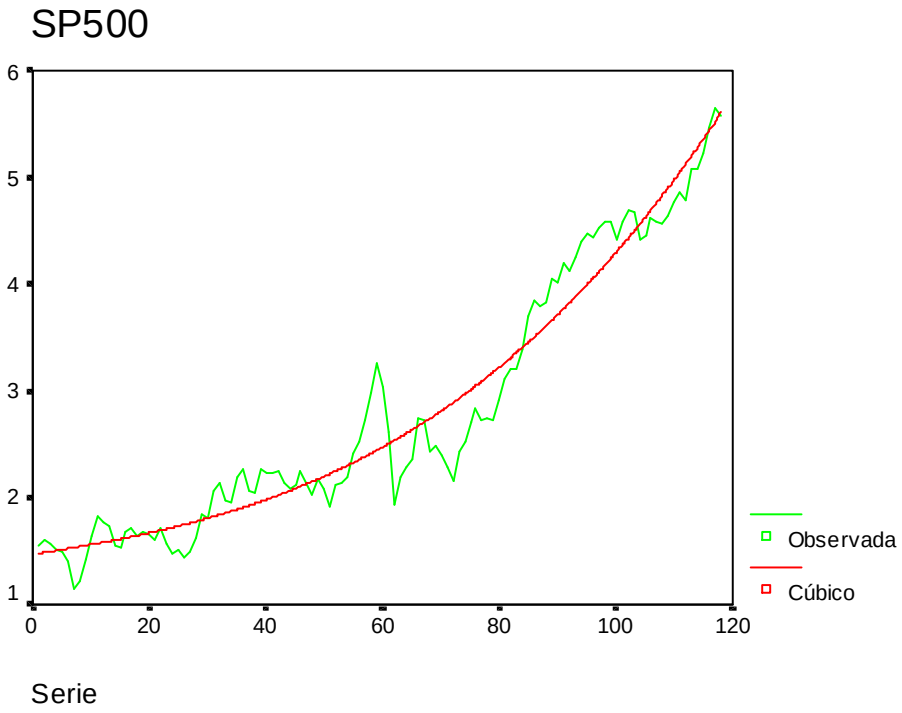
Dependent	Mth	Rsq	d.f.	F	Sigf	Upper bound	b0	b1	b2	b3
SP500	LIN	,874	116	801,26	,000		,8489	,0333		
SP500	LOG	,554	116	144,20	,000		-,8525	,9688		
SP500	INV	,104	116	13,48	,000		2,9911	-3,5992		
SP500	QUA	,947	115	1024,67	,000		1,6018	-,0044	,0003	
SP500	CUB	,948	114	697,67	,000		1,4719	,0084	4,8E-05	1,5E-06
SP500	COM	,919	116	1316,38	,000		1,2922	1,0118		
SP500	POW	,657	116	221,96	,000		,6565	,3613		
SP500	S	,140	116	18,82	,000		1,0166	-1,4277		
SP500	GRO	,919	116	1316,38	,000		,2564	,0117		
SP500	EXP	,919	116	1316,38	,000		1,2922	,0117		
SP500	LGS	,919	116	1316,38	,000	,	,7739	,9884		

El análisis de estos resultados nos lleva a concluir que la mejor representación de la tendencia de la serie es la cubica. Su especificación formal sería :

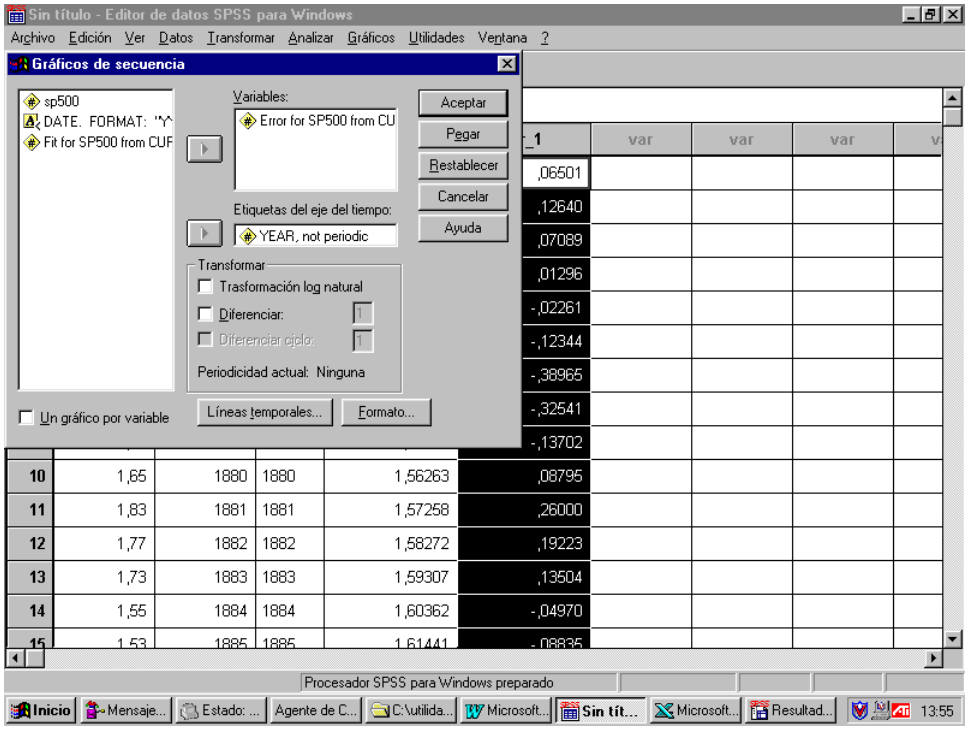
$$SP500 = 1.4719+0.0084t+0.00005t^2+0.0000015t^3$$

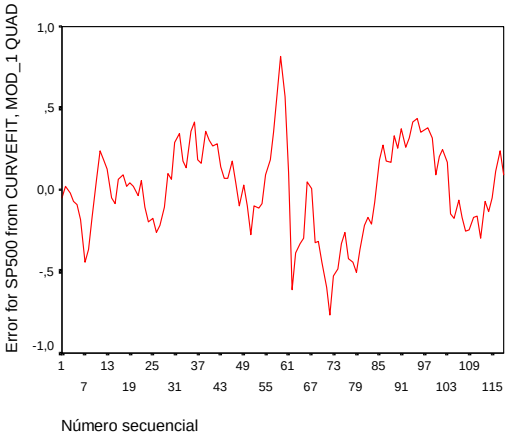
El análisis gráfico de dichos resultados lo presentamos a continuación:

YEAR, not periodic

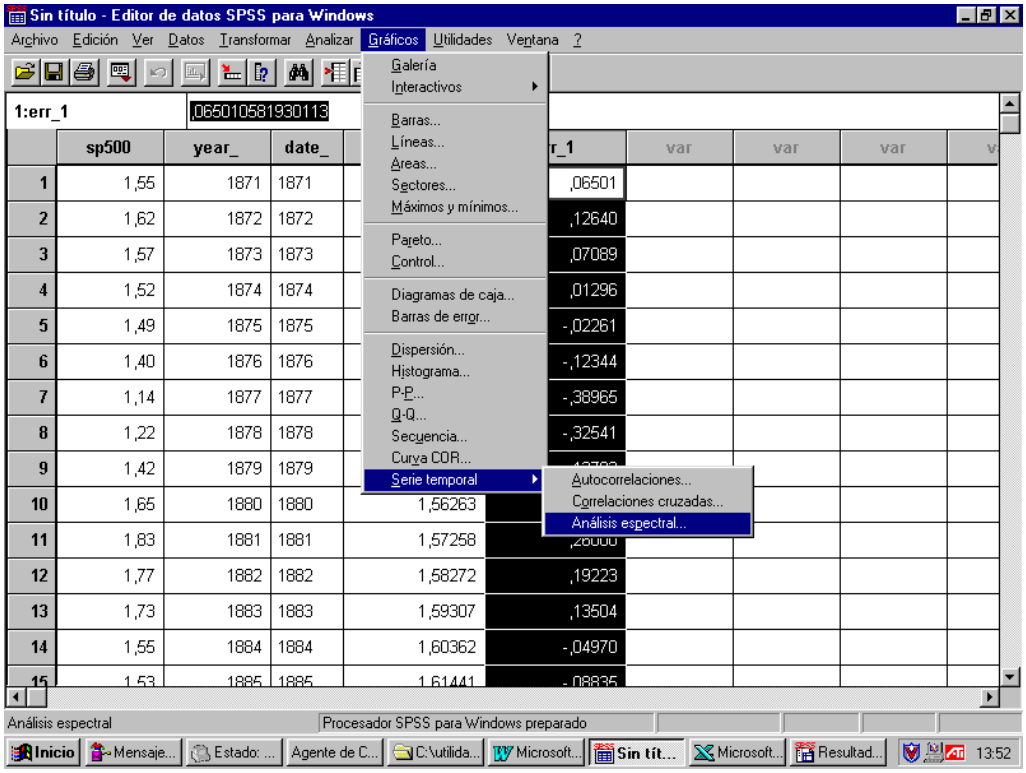


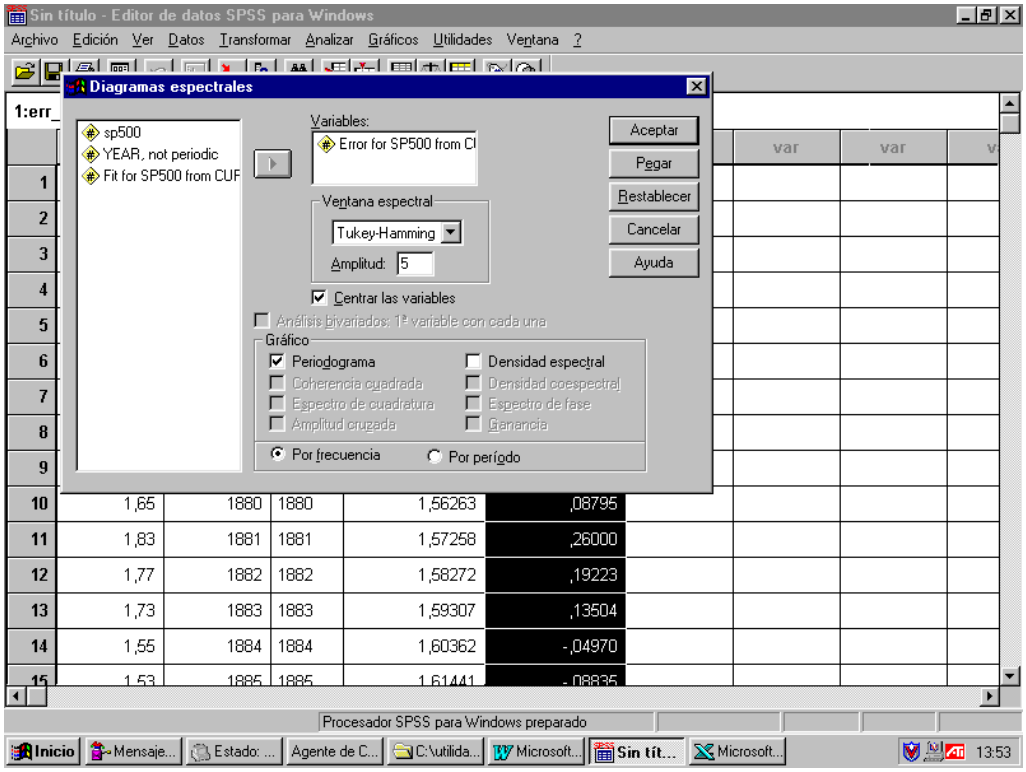
El componente cíclico de la serie se obtiene como residuo de la serie original de datos y la serie de tendencia. Este se puede observar realizando un gráfico de secuencia:





Finalmente, podemos ver los ciclos dominantes realizando un gráfico de análisis espectral:





Análisis espectral

```
MODEL:  MOD_2.
```

—