

The logo for UNED (Universidad Nacional de Educación a Distancia) is a dark green square with the letters 'UNED' in white, bold, sans-serif font.The background features a light gray grid with a white candlestick chart at the top and a dark gray bar chart at the bottom. A semi-transparent dark gray rectangle is centered over the grid, containing the title and author's name.

# Estadística e informática (SPSS) en la investigación descriptiva e inferencial

Juan Antonio Gil Pascual

A close-up, low-angle shot of a dark brown computer keyboard. The keys are slightly worn and have white text. Visible keys include 'Retroceso', 'Entrar', 'Inicio', 'Fin', 'Supr', 'Re Pág', 'Av Pág', 'Mayús', and 'Bloq Num'. The keyboard is positioned at the bottom of the page, partially overlapping the dark gray rectangle.

# *Estadística*



*e informática (SPSS) en la  
investigación descriptiva e  
inferencial*

*Juan Antonio Gil Pascual*

**Juan Antonio Gil Pascual**

# ESTADÍSTICA E INFORMÁTICA (SPSS) EN LA INVESTIGACIÓN DESCRIPTIVA E INFERENCIAL

UNIVERSIDAD NACIONAL DE EDUCACIÓN A  
DISTANCIA

*ESTADÍSTICA E INFORMÁTICA (SPSS)  
EN LA INVESTIGACIÓN DESCRIPTIVA E INFERENCIAL*

*Quedan rigurosamente prohibidas, sin la autorización escrita de los titulares del Copyright, bajo las sanciones establecidas en las leyes, la reproducción total o parcial de esta obra por cualquier medio o procedimiento, comprendidos la reprografía y el tratamiento informático, y la distribución de ejemplares de ella mediante alquiler o préstamo públicos.*

© Universidad Nacional de Educación a Distancia

Madrid, 2015

/

/ [www.uned.es/publicaciones](http://www.uned.es/publicaciones)

/

© Juan Antonio Gil Pascual

/

/

ISBN electrónico electrónico7042-6

/

/

Edición digital: mayo de 2015

## Índice

UNIDAD DIDÁCTICA 1

[PROCEDIMIENTOS Y PREPARACIÓN DE LOS DATOS.....](#)

[.....15](#)

1. GENERALIDADES .....	17
1.1. Introducción .....	17
1.2. Ventanas .....	17
1.2.1. Menús .....	18
1.2.2. Barra de herramientas y barra de estado .....	20
1.2.3. Botones .....	21
1.3. Entrada y ejecución del SPSS .....	23
1.4. Pasos básicos para realizar un análisis de datos .....	24
2. ICONOS .....	24
2.1. Iconos de la ventana de datos .....	24
2.2. Iconos de la ventana de sintaxis .....	29
2.3. Iconos de la ventana de resultados .....	30
3. DATOS .....	31
3.1. Matriz de datos .....	31
3.2. Introducción o modificación de datos con el SPSS .....	32
3.2.1. Definición de las variables .....	33
3.2.2. Introducción de datos .....	36
3.2.2.1. Introducir los datos correspondientes a una variable ..	36
3.2.2.2. Introducir los datos correspondientes a una unidad de análisis .....	36
3.2.2.3. Restricciones de los valores de datos .....	37
3.3. Almacenar datos .....	37
3.4. Lectura del fichero de datos .....	39
3.4.1. Abrir base de datos .....	40
3.4.2. Lectura de un fichero de datos de texto .....	

41	3.4.3. Lectura de otros tipos de fichero de datos .....	
42		
	3.5. Un ejemplo de construcción de un fichero de datos .....	
... 43		
	3.5.1. Denominación de variables .....	43
	3.5.2. Codificación de los valores .....	44
	3.5.3. Codificación de los valores ausentes .....	44
	3.5.4. Etiquetado de variables y valores .....	45
	3.5.5. Elaboración del plan de codificación y de la matriz de datos ..	45
	3.5.6. Fichero de procedimiento .....	
48		
	3.5.6.1. Fichero de comandos. EJEMPLO.SPS .....	48
	4. TRANSFORMACIONES DE LOS DATOS .....	
..... 49		
	4.1. Crear variables .....	
... 50		
	4.1.1. Calcular variables: Si los casos deben cumplir una condición lógica .....	51
	4.1.1.1. Indicar tipo y etiqueta a la variable creada .....	51
	4.2. Recodificar valores .....	
... 52		
	4.2.1. Recodificación en las mismas variables .....	52
	4.2.1.1. Recodificación en las mismas variables: valores anti guos y nuevos .....	53
	4.2.2. Recodificación en distintas variables .....	53
	4.2.3. Recodificación automática .....	
55		
	4.3. Contar apariciones de valores dentro de los casos .....	
55		
	4.4. Asignar rangos a los casos .....	
..... 57		
	4.5. Categorizador visual .....	
... 58		
	4.6. Semilla de aleatorización .....	
... 60		
	5. OPERACIONES EN Y CON LOS ARCHIVOS .....	
..... 60		

5.1. Ordenar casos .....	60
5.2. Fusión de archivos .....	61
5.2.1. Añadir casos .....	61
5.2.2. Añadir variables .....	63
5.3. Agregar datos .....	63
5.4. Segmentar archivos .....	65
5.5. Seleccionar casos .....	66
5.5.1. Seleccionar casos: Sí .....	67
5.5.2. Seleccionar casos: Muestra aleatoria .....	68
5.5.3. Seleccionar casos: Rango .....	68
5.6. Ponderar casos .....	69
5.7. Transponer .....	70
5.8. Reestructurar .....	71
5.9. Identificar casos duplicados .....	72
EJERCICIOS DE AUTOCOMPROBACIÓN .....	75
SOLUCIÓN A LOS EJERCICIOS DE AUTOCOMPROBACIÓN .....	76
BIBLIOGRAFÍA .....	83
<b>UNIDAD DIDÁCTICA 2</b>	
<b><u>ANÁLISIS DESCRIPTIVO DE DATOS</u></b> .....	85
<b>1. INTRODUCCIÓN A LA ESTADÍSTICA</b> .....	87

1.1. Introducción .....	87
1.2. Conceptos elementales: población, observación, muestra .....	87
1.3. Caracteres, rango y modalidades .....	88
1.4. Variable estadística y carácter .....	89
1.4.1. Tipos de variables y datos .....	89
2. DISTRIBUCIONES ESTADÍSTICAS UNIDIMENSIONALES .....	91
2.1. Distribuciones de frecuencias unidimensionales .....	91
2.1.1. Distribución de frecuencias en el caso de caracteres cualitativos ..	93
2.1.2. Distribución de frecuencias en el caso de variable discreta ....	93
2.1.3. Distribución de frecuencias en el caso de variable continua ....	94
2.2. Tablas estadísticas .....	96
3. REPRESENTACIONES GRÁFICAS .....	99
3.1. Introducción .....	99
3.2. Caracteres cualitativos .....	99
3.3. Caracteres cuantitativos .....	102
3.3.1. Gráficas de variables discretas .....	102
3.3.2. Gráficas para variables continuas .....	105
3.3.3. Gráfico de tallos y hojas .....	108
4. MEDIDAS DE POSICIÓN .....	109
4.1. Introducción .....	109
4.2. Medidas de tendencia central .....	110
4.2.1. Media aritmética .....	110

4.2.1.1. Medias generalizadas .....	113
4.2.1.2. Relación entre las medias .....	114
4.3. Mediana .....	
.... 114	
4.4. Moda .....	
..... 116	
4.5. Relación empírica entre media, mediana y moda .....	
.... 117	
4.6. Cuartiles, deciles y percentiles .....	
... 118	
4.7. Ejemplo de cálculo de la media, mediana, moda y cuartil .....	
119	
5. MEDIDAS DE DISPERSIÓN .....	
..... 120	
5.1. Introducción .....	
.... 120	
5.2. Recorridos .....	
..... 121	
5.3. Medidas de dispersión en torno a la media .....	
.... 121	
5.4. Medidas de dispersión en torno a la mediana .....	
..... 125	
5.5. Tipificación de una variable .....	
.. 126	
5.5.1. Definición de variable tipificada .....	
126	
6. MEDIDAS DE FORMA .....	
..... 127	
6.1. Medidas de simetría .....	
..... 127	
6.2. Medidas de apuntamiento .....	
..... 130	
6.3. Momentos .....	
..... 131	
7. LA EXPLORACIÓN DESCRIPTIVA DE DATOS .....	
..... 133	
7.1. Datos atípicos .....	
.... 133	
7.2. Gráfico caja (Box Plot) .....	
..... 133	
7.3. Comparación de distribuciones .....	
.. 134	
8. REGRESIÓN Y CORRELACIÓN .....	
..... 135	

8.1. Relación entre variables: aspectos generales .....	135
8.2. Regresión .....	135
8.2.1. Estudio particular de la regresión lineal: ajuste de una recta por el método de mínimos cuadrados .....	136
8.2.1.1. Recta de mínimos cuadrados .....	136
8.3. Correlación .....	138
8.3.1. Coeficiente de correlación .....	138
9. ANÁLISIS DESCRIPTIVO DE DATOS: COMANDOS .....	141
9.1. Introducción .....	141
9.2. La exploración de datos. Comando EXPLORAR .....	141
9.3. La distribución de frecuencias. FRECUENCIAS .....	150
9.4. Descriptivos. La opción DESCRIPTIVOS .....	156
9.5. Los comandos CORRELATIONS y NOPAR CORR .....	160
9.6. Análisis descriptivo de datos: GRÁFICOS .....	163
9.6.1. Tipos de gráficos .....	163
9.6.2. Los gráficos de barras, líneas, áreas y sectores .....	163
9.6.3. Diagramas de caja y barras de error .....	166
9.6.4. Histograma .....	167
9.6.5. Diagrama de dispersión .....	167
9.6.6. Gráficos P-P y Q-Q .....	168
9.6.7. Desarrollo de algunos ejemplos .....	169
9.6.7.1. Gráficos de barras, líneas, áreas y sectores .....	170

9.6.7.2. Diagramas de cajas y barras de error .....	171
9.6.7.3. Diagrama de dispersión .....	173
9.6.7.4. Gráficos P-P .....	174
EJERCICIOS DE AUTOCOMPROBACIÓN .....	
..... 176 SOLUCIÓN A LOS EJERCICIOS DE	
AUTOCOMPROBACIÓN .....	179
BIBLIOGRAFÍA .....	
..... 184	
<b>UNIDAD DIDÁCTICA 3</b>	
<b>MUESTREO Y ESTIMACIÓN .....</b>	
<b>..... 185</b>	
<b>1. TEORÍA ELEMENTAL DEL MUESTREO. TIPOS DE MUESTREOS</b>	
<b>..... 187</b>	
1.1. Introducción .....	
.... 187 1.2. Diseños muestrales .....	
..... 189	
1.3. Selección de la muestra en diseños probabilísticos .....	
189	
1.4. Muestreo aleatorio simple .....	
... 190	
1.4.1. Estimación de media, total y proporción .....	190
1.4.2. Selección del tamaño de la muestra para la estimación de la	
media, total y proporción .....	195
1.5. Muestreo estratificado .....	
198	
1.5.1. Estimadores de la media, el total y la proporción poblacional	
199	
1.5.2. Cálculo del tamaño muestral para la estimación de la media, total y	
proporción poblacional .....	202
1.5.3. Afijación .....	
... 202 1.6. Muestreo por conglomerados .....	
..... 204	
1.6.1. Estimación de la media, total y proporción poblacional .....	205
1.6.2. Tamaño de la muestra para la estimación de la media, total	
y proporción poblacional .....	209
1.7. Muestreo por conglomerados en dos etapas .....	
210	
1.8. Muestreo sistemático .....	

....	211	1.9. Muestreo por cuotas .....	212
<u>2. ESTIMACIÓN DE HIPÓTESIS. FUNDAMENTOS ESTADÍSTICOS</u> .....			
.....	213		
<u>2.1. Inferencia estadística .....</u>			
....	213		
<u>2.2. Distribuciones asociadas al proceso de muestreo .....</u>			
214	2.2.1. Distribución en el muestreo de algunos estadísticos .....	225	
<u>2.3. Estimación estadística .....</u>			
....	228		
<u>2.4. Estimadores por intervalos de confianza para parámetros de la población .....</u>			
	232		
<u>EJERCICIOS DE AUTOCOMPROBACIÓN .....</u>			
.....	242	SOLUCIÓN A LOS EJERCICIOS DE AUTOCOMPROBACIÓN .....	243
<u>BIBLIOGRAFÍA .....</u>			
.....	250		
<b>UNIDAD DIDÁCTICA 4</b>			
<b><u>ANÁLISIS INFERENCIAL DE DATOS</u> .....</b>			
.....	251		
<u>1. DECISIÓN ESTADÍSTICA. PRUEBAS PARAMÉTRICAS .....</u>			
.....	253		
<u>1.1. Introducción .....</u>			
.....	253		
<u>1.2. Hipótesis estadística .....</u>			
....	253		
<u>1.2.1. Hipótesis nula .....</u>			
	254		
<u>1.2.2. Hipótesis alternativa .....</u>			
	254		
<u>1.3. Formulación de hipótesis. Hipótesis simple frente alternativa simple .....</u>			
	255		
<u>1.3.1. Regiones críticas y de aceptación .....</u>			
	255		
<u>1.3.2. Error tipo I, error tipo II y potencia de una prueba .....</u>			
	256		

1.4. Hipótesis simple frente alternativa compuesta .....	
.257	1.5. Potencia de una prueba de hipótesis .....
..... 258	1.5.1. Función de potencia y curvas características operativas .....
.... 260	1.5.2. Cálculo de n para un a y b dados .....
.... 264	1.6. Principales pruebas paramétricas .....
..... 265	1.6.1. Muestras grandes: media y proporción muestrales .....
..... 267	1.6.2. Muestras pequeñas: media, varianza y correlaciones muestrales .. 271
	1.7. Pruebas para comparación de medias. Los comandos T-TEST Y

MEANS .....	
..... 277	

1.7.1. Prueba T para una muestra .....	
.277	
1.7.2. Prueba T para muestras independientes .....	279
1.7.3. Prueba T para muestras relacionadas .....	
282	
1.7.4. El comando MEANS (Medias) .....	
284	

2. DECISIÓN ESTADÍSTICA. PRUEBAS NO PARAMÉTRICAS .....	
..... 287	

2.1. Introducción .....	
.... 287	

2.2. Pruebas no paramétricas a partir de una sola muestra .....	
288	

2.2.1. De bondad de ajuste .....	
288	

2.2.1.1. Kolmogorov-Smirnov .....	288
-----------------------------------	-----

2.2.1.2. Contraste de $c^2$ de bondad de ajuste .....	291
---	-----

2.2.2. La prueba binomial .....	
.. 293	

2.2.3. De aleatoriedad .....	
295	

2.2.3.1. Prueba de rachas .....	295
---------------------------------	-----

2.3. Pruebas bimuestrales (muestras relacionadas) .....	
..... 297	2.3.1. La prueba de Mc.Nemar para la significación de los

cambios .. 298	
----------------	--

2.3.2. La prueba de los signos .....	
--------------------------------------	--

300	
	<a href="#">2.3.3. Prueba de pares igualados de Wilcoxon .....</a>
302	
	<a href="#">2.4. Pruebas bimuestrales (muestras independientes) .....</a>
305	
	<a href="#">2.4.1. La prueba de la probabilidad exacta de Fisher ..... 306</a>
	<a href="#">2.4.2. Contraste de Kolmogorov .....</a>
308	
	<a href="#">2.4.3. Contraste de <math>c^2</math> .....</a>
310	
	<a href="#">2.4.3.1. El procedimiento Tablas de Contingencia ..... 313</a>
	<a href="#">2.4.4. Test de la mediana .....</a>
318	
	<a href="#">2.4.5. Prueba U de Mann-Whitney .....</a>
320	
	<a href="#">2.4.6. Test de Wald-Wolfowitz o de rachas .....</a>
323	
	<a href="#">2.5. Pruebas para k-muestras relacionadas .....</a>
325	
	<a href="#">2.5.1. Prueba Q de Cochran .....</a>
326	
	<a href="#">2.5.2. Análisis de la varianza de Friedman .....</a>
328	
	<a href="#">2.6. Pruebas para k-muestras independientes .....</a>
331	
	<a href="#">2.6.1. Extensión de la prueba de la mediana ..... 331</a>
	<a href="#">2.6.2. Análisis de la varianza de Kruskal- Wallis .....</a>
334	
	<a href="#">2.7. Correlación no paramétrica .....</a>
336	
	<a href="#">2.7.1. El coeficiente de contingencia: C .....</a>
336	
	<a href="#">2.7.2. El coeficiente de correlación de rangos de Spearman: <math>r_s</math> ..... 338</a>
	<a href="#">2.7.3. El coeficiente de correlación de rango de Kendall: <math>t</math> ..... 341</a>
	<a href="#">2.7.4. El coeficiente de correlación parcial de rango de Kendall: <math>t_{xy.z}</math> ... 343</a>
	<a href="#">2.7.5. El coeficiente de concordancia de Kendall: W ..... 345</a>

2.8. Pruebas no paramétricas. El comando NPAR TEST .....	
.. 347 2.8.1. Pruebas para una muestra .....	
.. 348 2.8.2. Pruebas para dos muestras independientes .....	
.. 353 2.8.3. Pruebas para dos muestras relacionadas .....	
.. 355 2.8.4. Pruebas para varias muestras independientes .....	
356 2.8.5. Pruebas para varias muestras relacionadas .....	
358	

EJERCICIOS DE AUTOCOMPROBACIÓN .....	
..... 362 SOLUCIÓN A LOS EJERCICIOS DE	
AUTOCOMPROBACIÓN .....	371

BIBLIOGRAFÍA .....	
..... 394	

## **UNIDAD DIDÁCTICA 5**

<b>HACIA UN ESTUDIO DEL MODELO</b> .....	
..... 395	

<b>1. EL MODELO DE REGRESIÓN. EL COMANDO REGRESION</b> .....	
..... 397	

<b>1.1. El modelo de regresión</b> .....	
..... 397	

<b>1.1.1. Introducción</b> .....	
397	

<b>1.1.2. Procedimiento</b> .....	
398	

<b>1.1.3. Variables de intervención</b> .....	
409	

<b>1.2. El comando REGRESSION</b> .....	
..... 409	

<b>1.2.1. Regresión lineal simple</b> .....	
409	

<b>1.2.2. Regresión lineal múltiple</b> .....	
413	

<b>1.2.2.1. Estadísticos adicionales</b> .....	417
<b>1.2.2.2. Estudio de los supuestos del modelo de regresión</b> .....	422

<b>1.2.3. Estudio de casos influyentes y extraños</b> .....	
433	

<b>1.3. La elaboración de pronósticos</b> .....	
.. 436	

<b>2. ANÁLISIS DE LA VARIANZA</b> .....	
---	--

.....	437
<u>2.1. El significado de la experimentación .....</u>	<u>.....</u>
<u>.437 2.2. Elementos básicos del diseño de experimentos .....</u>	<u>.....</u>
<u>..... 438</u>	
<u>2.3. Los diseños factoriales .....</u>	<u>.....</u>
<u>... 440</u>	
<u>2.4. Análisis de la varianza factorial .....</u>	<u>.....</u>
<u>..... 442</u>	
<u>2.4.1. Análisis de la varianza con un factor. Modelo de efectos fijos .</u>	<u>443</u>
<u>2.4.2. Análisis de la varianza de dos factores .....</u>	<u>449</u>
<u>2.5. El comando ONEWAY .....</u>	<u>.....</u>
<u>..... 453 2.6. El comando UNIANOVA .....</u>	<u>.....</u>
<u>..... 458</u>	
<u>EJERCICIOS DE AUTOCOMPROBACIÓN .....</u>	<u>.....</u>
<u>..... 464 SOLUCIÓN A LOS EJERCICIOS DE</u>	<u>.....</u>
<u>AUTOCOMPROBACIÓN .....</u>	<u>465 BIBLIOGRAFÍA .....</u>
<u>.....</u>	<u>.....</u>
<u>..... 471</u>	

# UNIDAD DIDÁCTICA 1

## PROCEDIMIENTOS Y PREPARACIÓN DE LOS DATOS

### Objetivos

Conocer los rudimentos operativos del paquete estadístico SPSS. Saber manejar y transformar los datos con SPSS.

Operar adecuadamente con archivos de datos.

### 1. GENERALIDADES

#### 1.1. Introducción

El SPSS-PC es un paquete estadístico de Análisis de datos. En sus comienzos funcionaba sobre MS-DOS y en la actualidad trabaja en el entorno Windows. Se utiliza como apoyo en la investigación en ciencias sociales, económicas y de la salud. Contiene programas capaces de realizar desde un simple análisis descriptivo hasta diferentes tipos de análisis multivariante como: análisis discriminante, análisis de regresión, análisis de cluster, análisis de varianza, series temporales, etc. Está formado por un conjunto de módulos:

— **Básico** (herramientas de análisis exploratorio de datos, gráficos interactivos, tablas de contingencia, estadística descriptiva, comparación de medias, análisis de la varianza, pruebas no-paramétricas, correlación, regresión lineal múltiple, escalado multidimensional, análisis de fiabilidad, análisis factorial, análisis discriminante, análisis de conglomerados o cluster y análisis de proximidades).

— **Modelos de Regresión** (regresión logística binomial y multinomial, regresión no lineal, cuadrados mínimos ponderados, cuadrados mínimos en dos fases, modelos logit y probit).

— **Modelos Avanzados** (modelo lineal general (GLM), análisis loglineal, hiloglineal y genlog, estimación de componentes de la varianza, análisis de supervivencia: estimación Kaplan-Meier, regresión de Cox, modelos MANOVA).

— **Tablas** (para tratamiento y presentación de datos en forma tabular).

— **Tendencias** (análisis de series temporales: métodos de medias móviles, suavizado, Box-Jenkin, etc.).

— **Categorías** (procedimiento de escalamiento óptimo, análisis de correspondencias y mapas perceptuales).

Aparte de estos módulos, el paquete estadístico SPSS contiene una serie de programas que trabajan de forma independiente al mismo:

— **Análisis de segmentación-Answer Tree** (segmenta la información en grupos homogéneos respecto a una variable criterio).

— **Análisis relacional-AMOS** (análisis factorial confirmatorio, análisis de ecuaciones estructurales lineales).

— **QI-Análisis** (herramienta para el control de calidad que incluye estadísticos y gráficos de control interactivos para determinar las capacidades del proceso y detectar patrones de variación o tendencia de los datos).

— **Neural Connection** (herramienta para mejorar los modelos de segmentación, factorial, conglomerados, regresión, series temporales, etc. con técnicas de redes neuronales).

En este manual nos centraremos en la versión 12 del programa SPSS y haremos referencia a procedimientos estadísticos incluidos en el módulo base.

## 1.2. Ventanas

El SPSS utiliza diferentes ventanas:

Editor de datos (Datos nuevos) Contiene el fichero de datos. Con este

editor se puede

crear un fichero de datos o modificar uno ya existente. Esta ventana se abre automáticamente cuando se ejecuta el SPSS.

Salida (Resultado 1) Contiene los resultados de la ejecución de los comandos

del SPSS, tanto de lectura de datos, transformación de datos, etc. como de análisis de datos. Esta ventana se abre automáticamente cuando se ejecuta el SPSS. Se pueden abrir varias ventanas de salida.

Sintaxis (Sintaxis 1) Esta ventana se utiliza para ejecutar comandos del

SPSS. Estos comandos pueden pertenecer a un fichero que ha sido creado con el SPSS, o pueden escribirse directamente, o pueden proceder de ejecutar la opción **Pegar** de alguna de las ventanas de diálogo. Los mencionados comandos se pueden grabar en un fichero para utilizarlos en otro momento de la sesión de SPSS. Se pueden abrir varias ventanas de sintaxis. Para ejecutar los comandos SPSS de una ventana de sintaxis, se marca con el *ratón* los comandos que se desean ejecutar y se *pincha* en el icono



(ejecutar).

Proceso (Proceso 1) Ejecuta programas en Sax Basic sobre un objeto de

SPSS, por ejemplo una tabla pivote. Hay programas que contienen procesos standard almacenados en el subdirectorio SPSS y con extensión sbs.

En todas las ventanas del SPSS están disponibles unos menús, una barra de herramientas, una barra de estado, y unos *botones*, algunos de los cuales son comunes a todas las ventanas y otros específicos de cada una de ellas. Las acciones que se pueden realizar con los iconos se irán explicando a lo largo del manual. En cuanto a los menús, se comentarán brevemente a continuación, y con más detalle a lo largo del manual.

### **1.2.1. Menús**

Los menús más importantes del SPSSWIN son los siguientes:

Archivo Se utiliza para nuevo (datos, sintaxis, resultados, resultados borrador,

proceso), abrir, abrir base de datos, leer datos de texto, cerrar, guardar,

guardar como, mostrar información de datos, hacer caché de datos, imprimir, presentación preliminar, cambiar servidor, detener procesador, datos usados recientemente, archivos usados recientemente, salir de SPSS.

Edición Se utiliza para *deshacer, rehacer, cortar, copiar, pegar, pegar varia*

*bles, eliminar* en los ficheros de datos y procedimientos y *buscar* en el fichero de datos y *opciones* para definir las condiciones de entorno para tablas y datos.

Ver Se utiliza para visualizar la *barra de estado*, visualizar los iconos de la

*barra de herramientas* , cambiar las *fuentes*, poner o quitar la *cuadrícula* en los datos y para visualizar las *etiquetas de valor*, y finalmente *variables* (si se está en la vista de datos).

Datos Se utiliza para realizar cambios en el fichero de datos, tales como: defi

nir propiedades de variables, copiar propiedades de datos, definir fechas, insertar variable, insertar caso, ir a caso, ordenar casos, transponer, reestructurar, fundir archivos (añadir casos, añadir variables), agregar —es decir, crear grupos de casos para un análisis—, diseño ortogonal, segmentar archivos, seleccionar casos y ponderar casos. Estos cambios, excepto el caso de agregar que crea un archivo independiente, son temporales a no ser que se graben antes de salir del programa.

Transformar Se utiliza para realizar cambios en las variables seleccionadas del

fichero de datos y crear nuevas variables basadas en valores de otras variables que ya existen. Estos cambios son temporales y se guardan al salvar el fichero de datos. Las opciones del menú transformar son las siguientes: *calcular, semilla de aleatorización, contar apariciones, recodificar (en las mismas variables, en distintas variables), categorizar variables, asignar rango a casos, recodificación automática, crear serie temporal, reemplazar valores perdidos*. Además ejecutar transformaciones pendientes.

Analizar Se utiliza para realizar el análisis estadístico deseado. Las opciones

son: informes, estadísticos descriptivos, tablas, comparar medias, modelo lineal general, modelos mixtos, correlaciones, regresión, loglineal,

clasificar, reducción de datos, escalas, pruebas no paramétricas, series temporales, supervivencia, respuestas múltiples. Cada opción tiene otras sub-opciones que más adelante se comentarán.

**Gráficos** Se utiliza para crear gráficos de barras, histogramas, series temporales, etc.

Vienen en grupos de gráficos según homogeneidad de objetivos: Galería y Interactivos. Barras, Líneas, Áreas, Sectores, Máximos y mínimos. Pareto y Control. Diagramas de caja y Barras de error. Dispersión, Histograma, P-P, Q-Q, Secuencia, Curva COR y Serie temporal.

**Utilidades** Se utiliza para visualizar información sobre el contenido del fichero

de datos, del fichero de parámetros, o definir grupos de variables. Las distintas opciones disponibles son: *variables, información del archivo, definir conjuntos, usar conjuntos, ejecutar proceso y editor de menús.*

**Ventana** Para minimizar las ventanas y donde van apareciendo los distintos ficheros que se ejecutan.

? Este menú abre una ventana de ayuda que contiene información sobre el uso de cualquiera de las características del SPSS.

### **1.2.2. Barra de herramientas y barra de estado**

El SPSSWIN tiene una barra de herramientas donde se reflejan mediante iconos las distintas opciones de la ventana activa. Consiste, por decirlo de otra manera, un método abreviado de acceder a los menús. La barra de estado, situada en la parte inferior de la pantalla, tiene varias secciones:

— Área del procesador: nos indica el estado del procesador — Área de recuento: iteraciones realizadas en el fichero de datos — Área de filtrado: si existen datos filtrados — Área de ponderación: si los datos están ponderados — Área de segmentación de archivos

Solo se reflejará información cuando el procesador esté realizando su actividad y cuando se ejecute la misma.

ejmedias.sav - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

12:

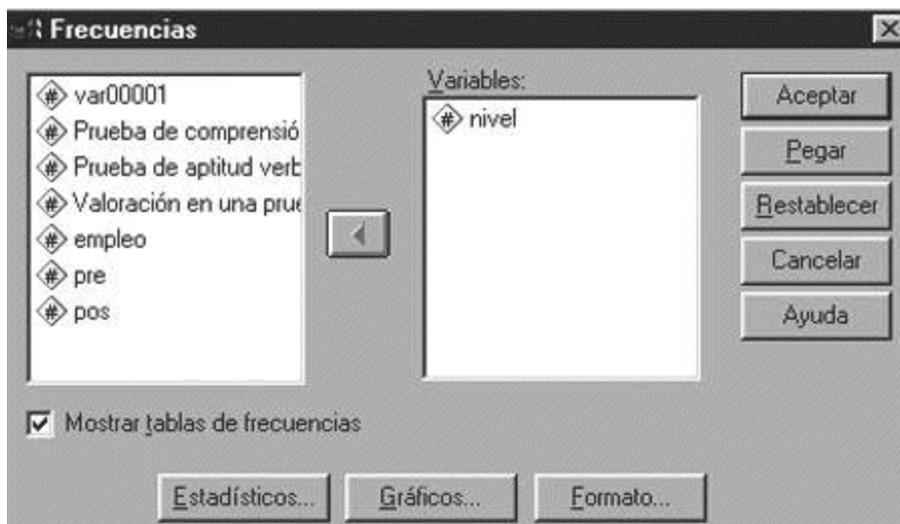
	var00001	item1	item2	item3	nivel	empleo	pre	pos	var	var
1	1.00	1.40	6.79	6.75	5.00	1.00	2.00	8.00		
2	3.00	4.31	4.41	7.51	5.00	3.00	5.00	5.00		
3	4.00	6.12	4.11	1.68	5.00	2.00	7.00	5.00		
4	5.00	2.91	4.99	7.62	5.00	1.00	4.00	6.00		
5	6.00	1.56	2.68	8.24	1.00	2.00	3.00	4.00		
6	7.00	7.00	.63	3.09	7.00	2.00	8.00	2.00		
7	8.00	3.46	5.60	5.88	3.00	2.00	4.00	7.00		
8	9.00	4.46	7.59	1.69	5.00	2.00	5.00	9.00		
9	10.00	.52	.60	5.97	1.00	3.00	2.00	2.00		
10	11.00	1.03	7.17	2.04	6.00	3.00	2.00	8.00		
11	12.00	1.41	1.59	.93	7.00	2.00	2.00	3.00		
12	13.00	.43	4.12	2.62	7.00	2.00	1.00	5.00		
13	14.00	6.22	9.69	8.55	7.00	2.00	7.00	11.00		
14	15.00	1.54	7.20	9.90	2.00	1.00	3.00	8.00		
15	16.00	7.15	3.61	8.08	4.00	2.00	8.00	5.00		
16	1.00	9.28	6.01	.80	4.00	2.00	10.00	7.00		
17	1.00	5.78	6.65	1.56	5.00	3.00	7.00	8.00		
18	2.00	2.62	3.28	1.07	7.00	3.00	4.00	4.00		
19	2.00	7.25	4.42	1.90	4.00	2.00	8.00	5.00		
20	2.00	.37	5.28	9.69	1.00	1.00	1.00	6.00		
21	3.00	1.01	.28	8.48	5.00	2.00	2.00	1.00		
22	5.00	7.33	5.05	2.60	1.00	1.00	8.00	6.00		

Vista de datos Vista de variables / SPSS El procesador está preparado

Inicio Dirección 09:38

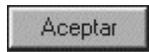
### 1.2.3. Botones

Además de las ventanas explicadas anteriormente, existen otras que se abren cuando se ejecuta alguna acción. Estas ventanas contienen distintos campos que hay que especificar, por ejemplo:



A los diferentes menús se puede acceder pinchando con el ratón en el campo deseado, o pulsando simultáneamente la tecla ALT y la tecla de la letra que figura subrayada.

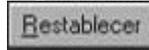
Obsérvese que estas ventanas disponen también de unos botones, unos que están activados (se puede pinchar en ellos), y otros que se activarán cuando se completen los campos de las ventanas. Estos botones son:



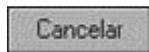
Almacena los campos definidos y cierra la ventana.



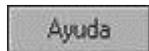
Crea una ventana de sintaxis que incluye el comando de SPSS equivalente a la acción que se está realizando.



Borra los campos que se acaban de definir y no cierra la ventana.



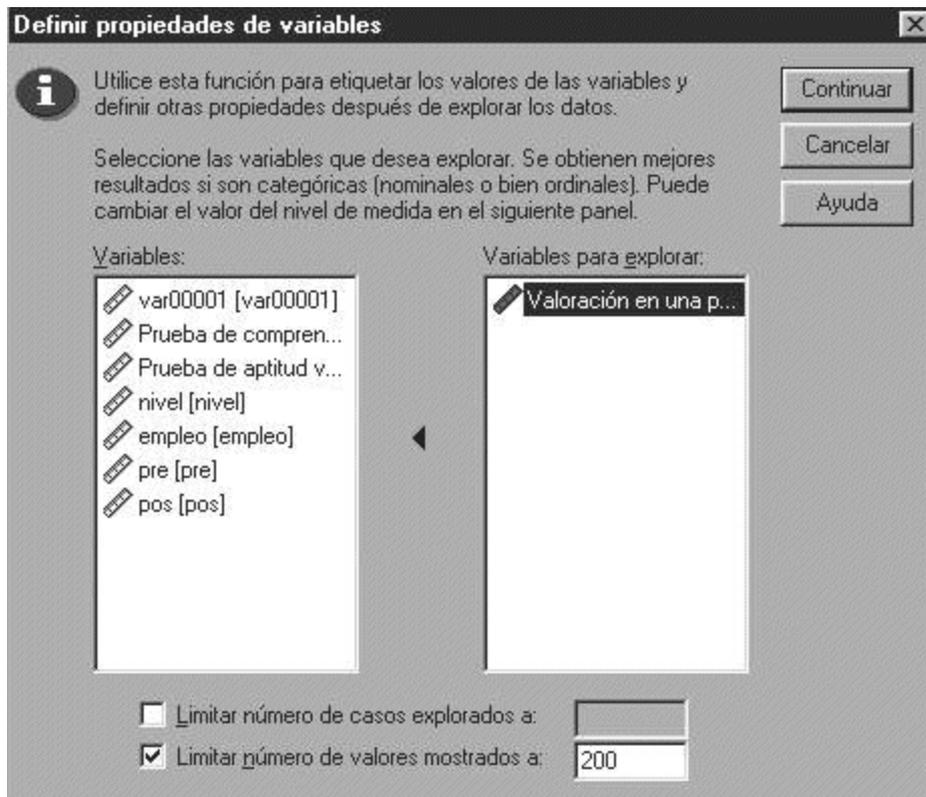
Cierra la ventana sin grabar.



Abre una ventana de ayuda que da información sobre la ventana que está abierta.

Los botones de Añadir, Cambiar, Borrar y se explicarán posteriormente

Hay ventanas que contiene otros botones diferentes, como por ejemplo:



donde el botón de **Continuar** almacena los campos definidos y vuelve a la ventana anterior. Los otros dos botones funcionan igual que en los casos anteriores.

### 1.3. Entrada y ejecución del SPSS

Para entrar en SPSSWIN se «pinchará» con el «ratón» 2 veces seguidas en el icono correspondiente al programa; entonces aparece la siguiente ventana:



aquí se puede seleccionar: un archivo de datos, un fichero de resultados, de sintaxis o de procedimientos. Otra forma de entrar en SPSSWIN es «pulsar» 2 veces sobre algún fichero de sintaxis (.SPS), de datos (.SAV) o de resultados (.SPO).

Además de estas formas de entrar en SPSS, existe dos formas de ejecutar SPSS: una pulsando el botón **ACEPTAR** en el menú correspondiente del procedimiento estadístico que estemos realizando; y otra marcando con el ratón en el fichero de procedimientos (.SPS) los comandos que se quieran ejecutar y pulsar

el icono



## 1.4. Pasos básicos para realizar un análisis de datos

— **Introducir los datos en SPSS.** Se puede abrir un archivo de datos previamente grabado en SPSS, importar de una hoja de cálculo o una base de datos, leer de un archivo de texto, de datos de programas estadísticos (SYSTAT o SAS) o introducir datos directamente en el editor de datos del SPSS.

— **Seleccionar un procedimiento** para calcular algún estadístico o crear un gráfico.

— **Seleccionar las variables para el análisis.** Cuando se elija un procedimiento estadístico hay que seleccionar las variables a las que se aplicará el mismo.

— **Ejecutar el procedimiento y seleccionar los resultados** que se necesiten como salida. Esta selección se realizará en el navegador de resultados.

Todos los comandos ejecutados en una sesión de SPSS se guardan en el fichero **spss.jnl** que está en el subdirectorio TEMP del Windows. Este fichero, en ASCII, puede ser editado. Se aconseja el cambio de nombre si se quieren conservar los comandos utilizados en la sesión para una posterior utilización.

## 2. ICONOS

El programa SPSS, como cualquier otro programa de Windows, dispone de una serie de iconos que permiten realizar ciertas operaciones directamente, sin necesidad de utilizar los menús.

Las tres barras de iconos más utilizadas son: la correspondiente a la ventana de datos, la de sintaxis y la de resultados.

### 2.1. Iconos de la ventana de datos

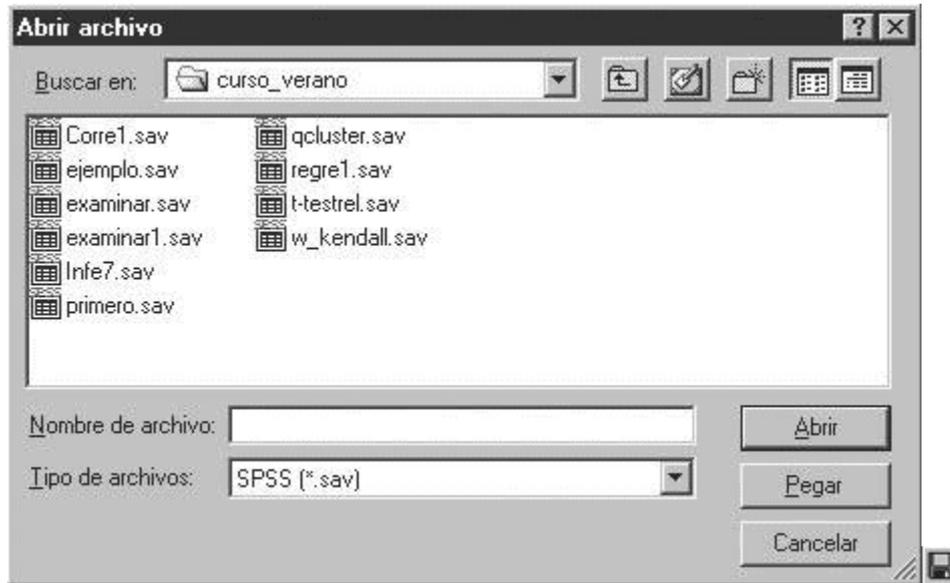
La barra de iconos de la ventana de datos es:



A continuación se explica la operación que se realiza cuando se *pincha* con el botón izquierdo del ratón alguno de estos iconos:



**Abrir archivo** . Muestra la ventana **Abrir archivo** para el tipo de documento que esté en la ventana activa. Según el tipo de ventana que esté activa, se puede utilizar este icono para abrir un archivo de datos, sintaxis, resultados, proceso y otro. Por ejemplo, si la ventana que está activa es la del Editor de datos, cuando se pincha este icono se abre la siguiente ventana:

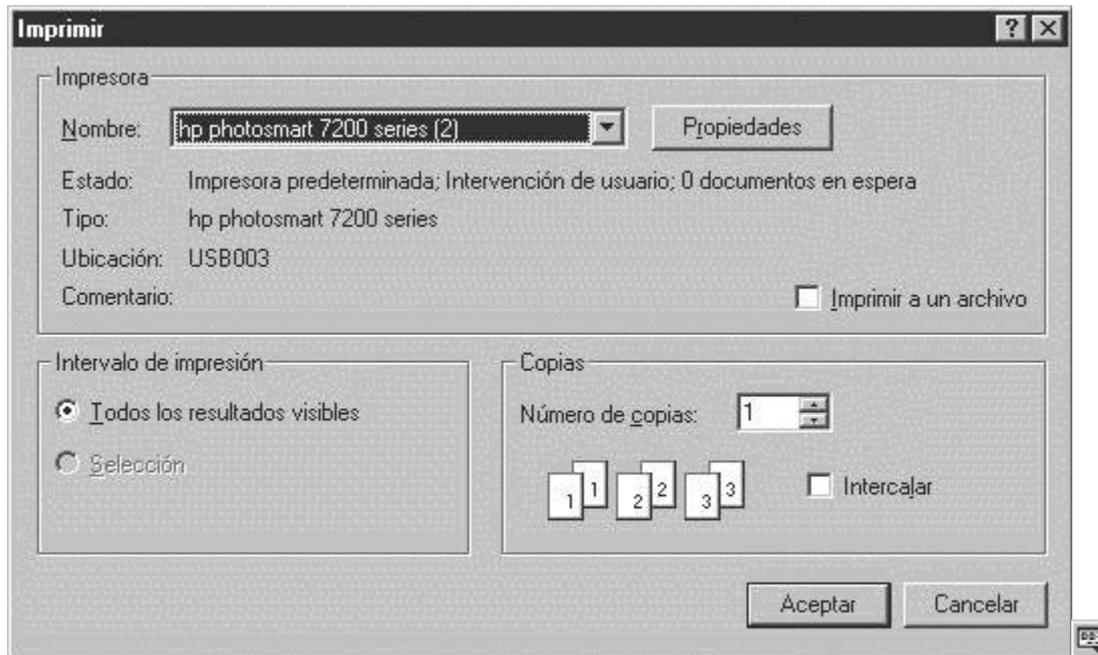


**Guardar archivo.** Guarda el archivo de la ventana que está activa. Puede guardar el documento completo o las líneas de texto seleccionadas. Si el archivo que se va a guardar no tiene nombre, muestra la ventana de **Guardar como**. Por ejemplo, si la ventana que está activa es la de resultados, aparecerá la siguiente ventana:



**Imprimir archivo.** Muestra la ventana de **Imprimir** para el tipo de documento que esté en la ventana activa. En el caso de archivos de

resultados, sintaxis y datos, puede imprimir el documento completo o un área seleccionada. Por ejemplo, si la ventana que está activa es la de resultados, cuando se pincha en este icono aparecerá la siguiente ventana:



**Recuperar cuadro de diálogo.** Muestra una lista de las últimas ventanas abiertas (aunque no se hayan abierto en la misma sesión). Cuando se *pincha* en este icono aparece el siguiente menú:



donde se selecciona la ventana operación que se desea abrir.

**Deshacer/rehace.** Se utiliza para hacer/deshacer las operaciones realizadas con los datos.

**Recorrer gráficos.** Este icono se utiliza para activar la ventana de gráficos.



**Ir a caso.** Este icono se utiliza para desplazarse a un caso en el Editor de datos. Cuando se *pincha* en este icono aparece la siguiente ventana:



**Información de variables.** Muestra una ventana que contiene la lista de variables y la información de la variable seleccionada. Este icono también se utiliza para posicionar el cursor en dicha variable, basta con

*pinchar* en el botón **Ir a**



**Buscar texto.** Se utiliza para buscar una cadena de caracteres en la variable del fichero de datos seleccionada. Cuando se *pincha* en este icono aparece la siguiente ventana:



En el campo **Buscar qué:** se especifica la cadena de caracteres que se quiere localizar.



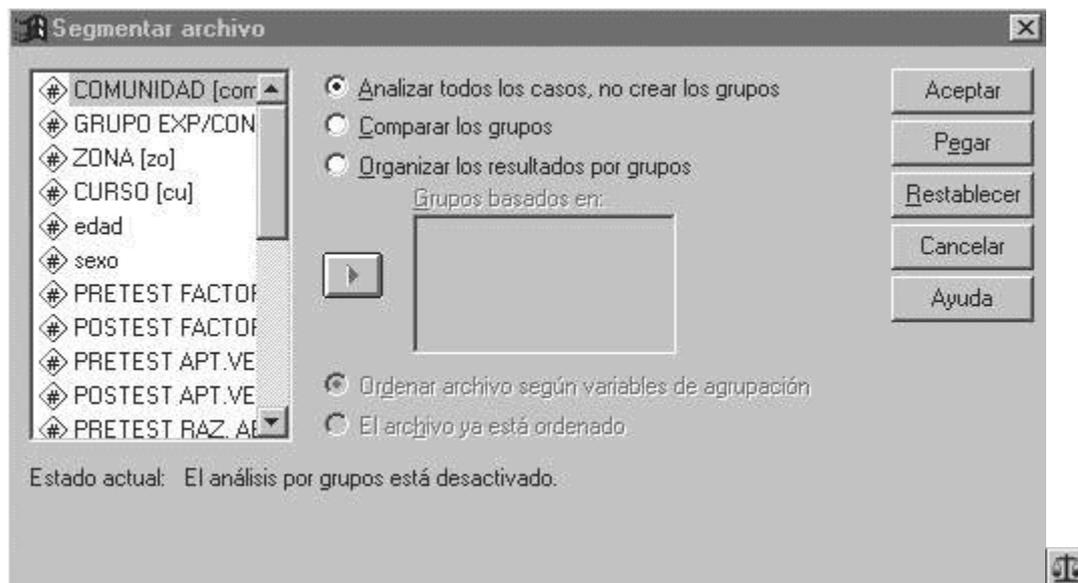
**Insertar caso.** En el Editor de datos, al *pinchar* en este icono se inserta un nuevo caso. Tiene el mismo efecto que la selección de **Insertar caso** del menú **Datos**.



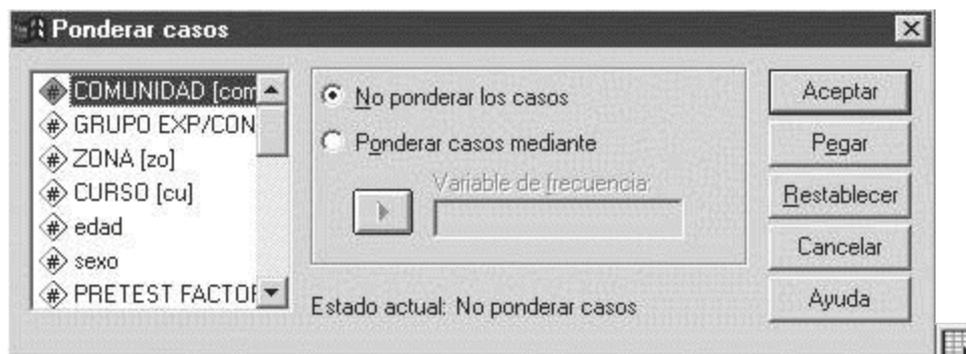
**Insertar variable.** En el Editor de datos, al *pinchar* en este icono se inserta una variable a la izquierda de la variable que contenga la celda activa. Tiene el mismo efecto que la selección de **Insertar variable** del menú **Datos**.



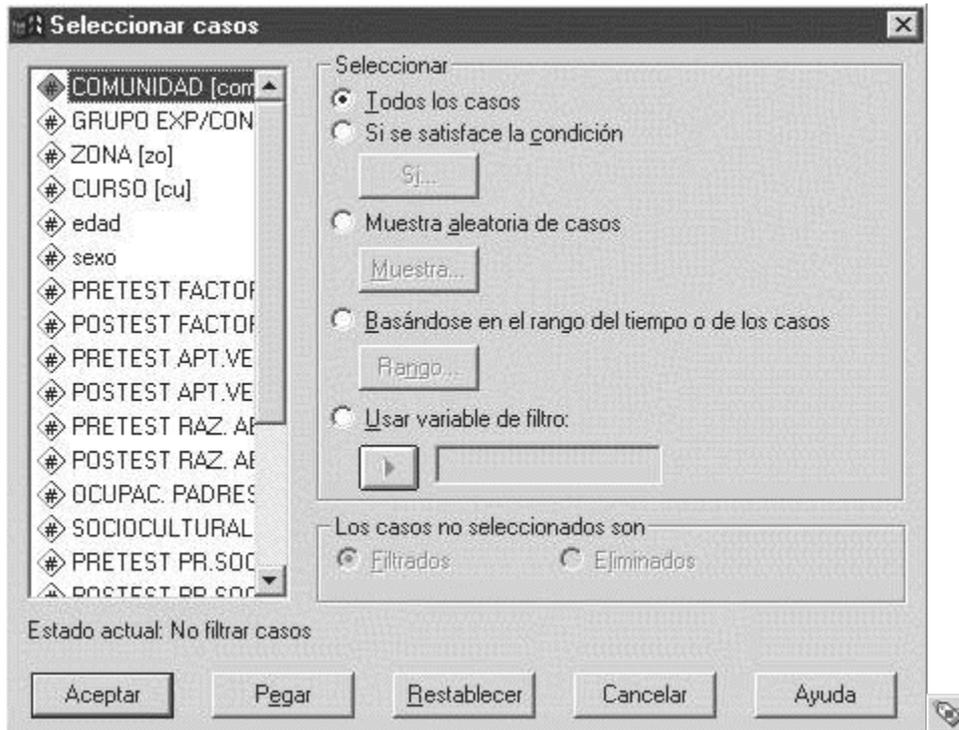
**Segmentar archivo.** Sirve para segmentar un archivo de datos en función de una o varias variables.



**Ponderar.** Cuando se *pincha* en este icono, muestra una ventana donde se puede seleccionar la variable utilizada para ponderar los casos en función de las frecuencias de la misma.

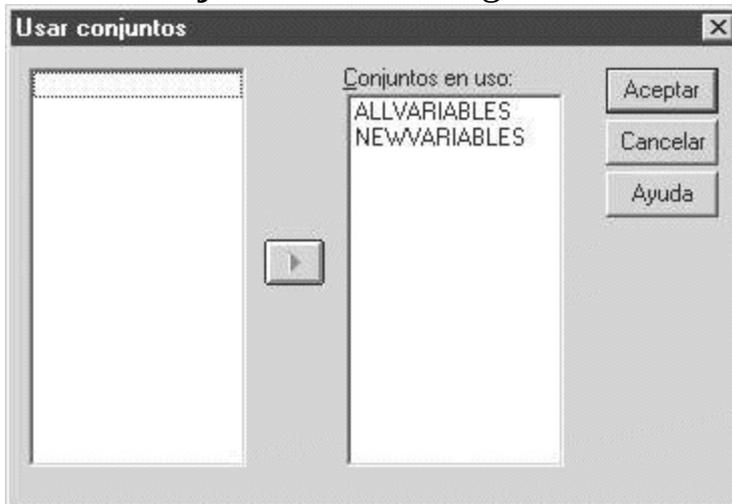


**Seleccionar casos.** Permite la selección de casos según distintas circunstancias.



**Etiquetas de valor.** Conmuta entre los valores actuales y las etiquetas de valor en el Editor de datos.

**Usar conjuntos.** Abre la siguiente ventana:



Aquí se puede seleccionar los conjuntos de variables que van a aparecer en las ventanas que se utilizan para realizar los análisis estadísticos o transformaciones.

Si se va a utilizar algún conjunto de variables diferente a los que tiene por defecto (ALLVARIABLES y NEWVARIABLES) éstos no pueden figurar en el campo **Conjuntos en uso:**

## 2.2. Iconos de la ventana de sintaxis

A continuación se presenta la barra de iconos que se corresponden con

operaciones de la ventana de sintaxis.



Veamos alguno de los iconos que no se han comentado en el caso de la ventana de datos, por no figurar entre los últimos:



**Ejecutar sintaxis.** En una ventana de sintaxis, este icono se utiliza para ejecutar los comandos de SPSS previamente seleccionados. Si no se ha seleccionado ningún comando se ejecuta el comando en el que aparezca el cursor.



**Ayuda de sintaxis.** Pulsando este botón accedemos a una ayuda de sintaxis. Si pasamos el ratón sobre un comando determinado y pulsamos se abrirá la siguiente ventana:

```
EXAMINE VARIABLES=varlist [[BY varlist] [varname BY varname]]
[/COMPARE={GROUP** }]
{VARIABLE}
[/ID={%CASENUM**}]
{varname }
[/PERCENTILES[ (value list)=[{HAVERAGE } ] [NONE]]
{WAVERAGE }
{ROUND }
{AEMPIRICAL}
{EMPIRICAL}
]
[/PLOT={STEMLEAF**} [BOXPLOT**] [NPLOT]]
[SPREADLEVEL(value)] [HISTOGRAM]
[{ALL } ]
{NONE}
]
[/STATISTICS={DESCRIPTIVES**} [EXTREME({5})] ]
[{ALL } ]
{NONE}
]
[{/TOTAL } ]
{/NOTOTAL}
]
[/MESTIMATOR=[ {NONE**}]]
{ALL }
]
[HUBER({1.339})] [ANDREW({1.34 Pi})]
{c } {c }
]
[HAMPEL({1.7,3.4,8.5})]
{a ,b ,c }
]
[TUKEY({4.685})]
{c }
]
[/MISSING={LISTWISE**} [INCLUDE]]
{REPORT }
{PAIRWISE }
**Default if the subcommand is omitted.
```

Como vemos, podemos mostrar la sintaxis del comando, imprimir o cancelar.



**Designar ventana.** Si tiene abierta más de una ventana del Editor de sintaxis, la sintaxis de comandos se pega en la ventana designada del Editor de sintaxis. Las ventanas designadas se indican por un signo de exclamación (!) en la barra de estado. Puede cambiar las ventanas designadas en cualquier momento.

La ventana designada no debería confundirse con la ventana activa, que es la ventana actualmente seleccionada. Si tiene ventanas superpuestas, la ventana activa aparecerá en primer plano. Si abre una nueva ventana del Editor de sintaxis o del Navegador de resultados, esa ventana se convertirá automáticamente en la ventana activa y la ventana designada.

### 2.3. Iconos de la ventana de resultados

Se muestra a continuación la barra de iconos que se corresponden con operaciones de la ventana de resultados.



Veamos alguno de los iconos que no se han comentado al hablar de los mismos en las ventanas de datos y sintaxis:

 **Presentación preliminar.** Permite ver los resultados en pantalla completa, de esta forma podremos predecir el formato que tendrán al salir por impresora.  **Exportar.** Guarda las tablas pivote y los resultados de texto en formato html y de texto y los gráficos en diversas opciones.

 **Seleccionar últimos resultados.** Marca los últimos resultados realizados por el procesador de SPSS.

 **Insertar encabezado.** Permite colocar un encabezado en el diagrama del navegador de resultados para que resulten más clarificadores. 

**Insertar títulos.** Como su nombre indica, podemos insertar en los resultados algún texto esclarecedor no relacionado con alguna tabla pivote.

 **Insertar texto.** Coloca un texto en cualquier parte de los resultados no relacionados con una tabla pivote.

 **Ocultar.** Deja de presentar un resultado, texto o gráfico en la pantalla de resultados.

 **Mostrar.** Presenta el resultado oculto mediante el icono anterior.



**Contraer/Expandir.** Permite agrupar un conjunto de resultados de la ejecución de uno o varios comandos, o por el contrario, expandir los resultados previamente contraídos.



**Ascender/Degradar.** Permite, en el navegador de resultados, realizar un proceso similar al tabulador en un procesador de textos. Es decir, llevar a izquierda o derecha un conjunto de resultados.

### 3. DATOS

#### 3.1. Matriz de datos

Para realizar un análisis estadístico es necesario disponer de una matriz de datos. Dicha matriz se puede estructurar de la siguiente forma:

<b>Variable 1</b>	<b>Variable 2</b>	.....	<b>Variable m</b>	Caso 1
Caso 2				
Caso 3				
.				
.				
.				
Caso n				

Es decir, es una matriz rectangular de dimensión  $n \times m$ , donde  $n$  es el número de filas y corresponde a las unidades o elementos de análisis, y  $m$  que es el número de columnas y corresponde a las variables.

Para introducir los datos y después poder analizarlos con SPSS, se puede utilizar el editor de datos de dicho programa, importarlos de una base de datos o una hoja de cálculo, importarlos de otros programas estadísticos, o bien con un procesador de textos crear un fichero ASCII que contenga dicha matriz de datos.

Con introducir o importar los datos no se ha terminado la labor. Para crear un archivo de datos de SPSS (.sav), se necesita además «*definir variables*». Esta labor incluye: el nombre de la variable, el tipo, el formato de lectura, la etiqueta de la variable, la etiqueta de las categorías (cuando la variable es discreta) y la definición de los valores «*missing*».

#### 3.2. Introducción o modificación de datos con el SPSS

Como se ha comentado, los datos se pueden introducir directamente en el editor del SPSS. En este apartado se explica cómo formar la matriz de datos y la definición de las variables mediante dicho editor del programa. Es aconsejable definir primero las variables y después introducir los datos. Ambas operaciones se realizan desde la ventana «**Nuevo Datos**».

La ventana de datos tiene dos visiones: a) **vista de datos**, donde se muestran los valores reales o las etiquetas de valor definidas y b) **vista de**

**variables**, con la información relativa a las mismas.

The screenshot shows the 'Vista de variables' window in SPSS. The window title is 'examinar1.sav - Editor de datos SPSS'. The menu bar includes 'Archivo', 'Edición', 'Ver', 'Datos', 'Transformar', 'Analizar', 'Gráficos', 'Utilidades', and 'Ventana'. The toolbar contains various icons for file operations and data manipulation. The main area is a table with the following columns: 'Nombre', 'Tipo', 'Anchura', 'Decimales', 'Etiqueta', 'Valores', 'Perdidos', 'Columnas', 'Alineación', and 'Medida'. The table contains 11 rows of variable definitions:

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida
1	item1	N Numérico	8	2	comprensión l	{Ninguno	{Ninguno	8	Derecha	Escala
2	item2	N Numérico	8	2	aptitud verbal	{Ninguno	{Ninguno	8	Derecha	Escala
3	item3	N Numérico	8	2	prueba de ingl	{Ninguno	{Ninguno	8	Derecha	Escala
4	nivel	N Numérico	8	2	nivel	{1,00, primario	{Ninguno	8	Derecha	Escala
5	selecc	N Numérico	8	2	selección	{1,00, seleccio	{Ninguno	8	Derecha	Escala
6	zitem1	N Numérico	11	5	Puntua: comp	{Ninguno	{Ninguno	8	Derecha	Escala
7	zitem2	N Numérico	11	5	Puntua: aptitu	{Ninguno	{Ninguno	8	Derecha	Escala
8	zitem3	N Numérico	11	5	Puntua: prueb	{Ninguno	{Ninguno	8	Derecha	Escala
9	sexo	N Numérico	2	0	sexo	{1, hombre}...	{Ninguno	4	Derecha	Nominal
10	zona	N Numérico	8	2	zona	{1,00, rural}...	{Ninguno	8	Derecha	Ordinal
11	norma	N Numérico	8	2		{Ninguno	{Ninguno	8	Derecha	Escala
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										
25										
26										
27										
28										
29										
30										
31										
32										

The status bar at the bottom shows 'Vista de datos' and 'Vista de variables'. The taskbar at the bottom includes 'Inicio', 'Explorando - C:\juan\curs...', 'spss\_1.doc - Microsoft Word', and 'examinar1.sav - Edito...'. The system clock shows '18:23'.

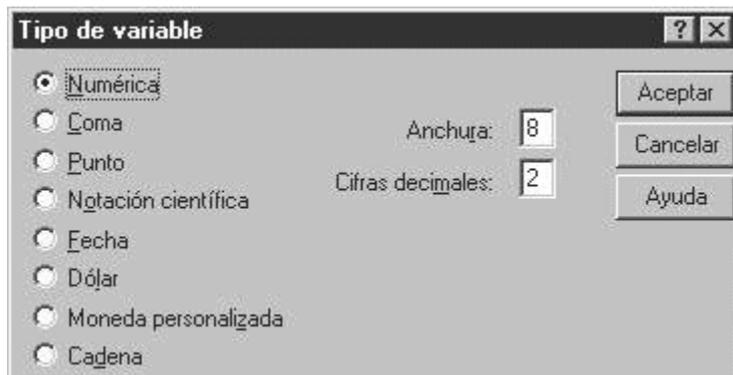
### 3.2.1. Definición de las variables

Una vez activada la ventana **Nuevo Datos**, si se trata de crear un fichero de datos nuevo, o estando en un fichero de datos ya creado cuando se pretende su modificación. Si se está en la **vista de variables**, se puede definir las características de cualquier variable: nombre, tipo, anchura, decimales, etiqueta, valores, perdidos, columnas, alineación y medida.

Estas propiedades de las variables son:

**1. Nombre:** en nuestro idioma puede tener como máximo 64 caracteres, que pueden ser alfabéticos o numéricos o el símbolo de subrayado (    ); la única restricción es que el primer carácter debe ser alfabético, \$ o #, no puede contener el blanco, ni caracteres especiales, no debe acabar en punto. No se puede duplicar el nombre de una variable en el fichero de datos.

**2. Tipo de variable:** Se utiliza para modificar el tipo de la variable y el formato. Por defecto se toma la variable como numérica de Ancho: 8, Decimales:2 (definido en cuadro de diálogo de opciones).



donde se *pinchará* en el tipo que se desee.

Los tipos de variables son:

— *Numéricas*: Sólo toman valores numéricos en formato estándar o en notación científica.

— *Coma*: Variable numérica con valores expresados con comas que delimitan cada tres posiciones y con el punto como delimitador decimal.

— *Punto*: Variable numérica con valores expresados con puntos que delimitan cada tres posiciones y con la coma como delimitador decimal.

— *Notación científica*: Variable numérica cuyos valores se muestran con una E intercalada y un exponente con signo que representa una potencia de base diez.

— *Fecha*: Variable numérica con valores expresados en uno de los diferentes formatos de fecha-calendario u hora-reloj. Al indicar fecha aparece una lista con diferentes formatos. Habrá que seleccionar uno.

— *Moneda personalizada*: Variable numérica con valores expresados en uno de los formatos de moneda personalizados que se habrán definido previamente en la pestaña Moneda del cuadro de diálogo Opciones.

— *Cadena*: Variable cuyos valores no son numéricos y, por tanto, no se utilizan en los cálculos. Pueden contener cualquier carácter siempre que no se exceda la longitud definida en el cuadro de diálogo que aparece al seleccionar esta opción. Las mayúsculas y las minúsculas se consideran diferentes. También se conoce como variable alfanumérica.

**3. Etiqueta:** Se utiliza para definir la etiqueta de la variable, no puede exceder a 256 caracteres pueden contener espacios y caracteres reservados que no se permiten en los nombres de las variables.

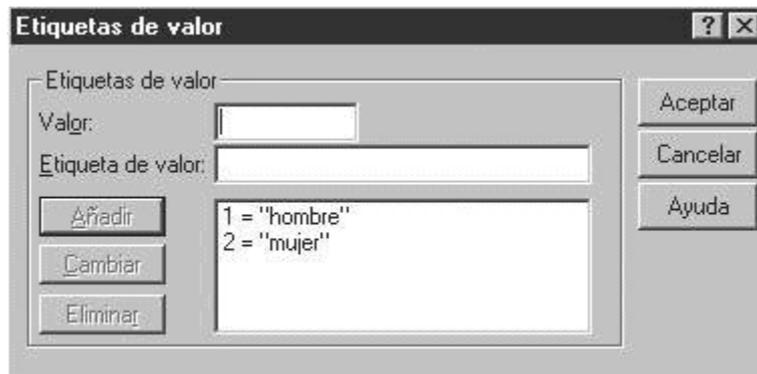
**4. Valores:** Son las etiquetas de valor donde a cada valor de una variable se le pueden asignar etiquetas descriptivas de dicho valor. Es particularmente útil si el archivo de datos utiliza códigos numéricos para

representar categorías no numéricas (por ejemplo, los códigos 1 y 2 para *hombre* y *mujer*).

Las etiquetas de valor pueden tener hasta 60 caracteres.

Estas etiquetas no están disponibles para las variables de cadena larga (variables de cadena con más de 8 caracteres).

Al pulsar Valor se entra en el cuadro de diálogo siguiente:



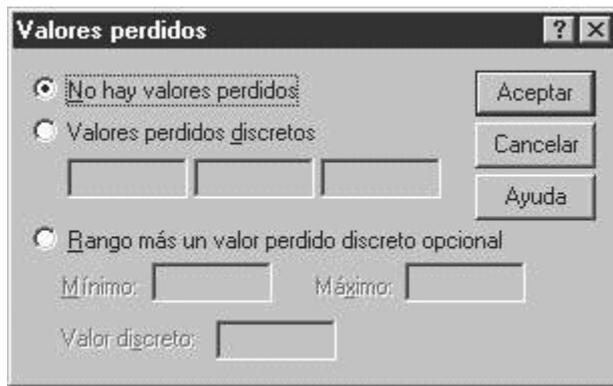
Para introducir una etiqueta de valor se procede de la siguiente manera: una vez rellenos los campos **Valor** y **Etiqueta de valor** se *pincha* en el botón de **Añadir**. Si se desea eliminar alguna de estas etiquetas se selecciona dicha etiqueta (*pinchando* con el *ratón* en ella) y se pulsa en el botón de **Borrar**. Si se desea hacer alguna modificación en la definición de los valores y sus etiquetas, se selecciona la etiqueta a modificar, se *pincha* en el campo que se va a modificar (valor o etiqueta), y una vez hecha la modificación se *pincha* en el botón de **Cambiar** (Estos botones funcionan igual en todas las ventanas del SPSS, por lo que no se volverán a comentar. Naturalmente dichos botones deben estar activos para poder *pinchar* en los mismos).

**5. Perdidos:** Se utiliza para definir los valores *missing* (ausencia de dato) de la variable que se está describiendo, estos valores no se utilizan en la mayoría de cálculos. El SPSS distingue entre 2 tipos de valores *missing*:

— *System missing*: valores *missing* declarados por el SPSS, para los espacios en blanco y caracteres ilegales que pueda haber en el fichero de datos. Estos aparecerán en los listados representados por puntos.

— *User missing*: valores *missing* declarados por el usuario, que se puede modificar conforme la ventana siguiente:

Cuando se *pincha* en **Perdidos**, aparece la siguiente ventana de los valores perdidos por el usuario:



No se pueden definir valores perdidos para variables de cadena larga (más de 8 caracteres).

**6. Columnas:** Se utiliza para definir el número de caracteres para el ancho de la columna donde se introducirán los valores de la variable.

Los anchos de columna también se pueden cambiar en la Vista de datos pulsando y arrastrando los bordes de la columna.

Los formatos de columna afectan sólo a la presentación de valores en el Editor de datos. Al cambiar el ancho de columna no se cambia el ancho definido de una variable. Si el ancho real y definido de un valor es más ancho que la columna, aparecerán asteriscos (\*) en la ventana Vista de datos.

**7. Alineación:** La alineación se utiliza para controlar la presentación de los valores de los datos y/o de las etiquetas de valor en la Vista de datos. La alineación por defecto es derecha para las variables numéricas e izquierda para las variables de cadena. Esta disposición afecta sólo a la presentación en la Vista de datos.

**8. Medida:** Sirve para especificar el nivel de medida de la variable: Escala (datos numéricos de una escala de intervalo o de razón), Ordinal o Nominal. Los datos nominales y ordinales pueden ser de cadena (alfanuméricos) o numéricos. La especificación de medida toma importancia en:

— Procedimientos de gráficos o de tablas personalizadas que identifican las variables como de escala o categóricas. Las variables nominales y ordinales se tratan como categóricas.

— Los archivos de datos con formato SPSS utilizados con AnswerTree y otros módulos.

### **3.2.2. Introducción de datos**

Una vez que se han definido las variables, se pasará a introducir la matriz de datos. Estos se introducen en el panel de la ventana **Datos nuevos**.

Los datos se pueden introducir de dos formas diferentes: metiendo todos los valores de una determinada variable, es decir, introducir los datos por columna, o metiendo todos los valores de una determinada unidad de análisis, es decir, introducir los datos por fila, no obstante el sistema no es rígido y se puede saltar de una variable a otra y de una columna a otra. También se puede con los datos, como en cualquier programa de Windows, utilizar las opciones de cortar y pegar.

#### *3.2.2.1. Introducir los datos correspondientes a una variable*

Por defecto, cuando se teclea el primer valor y se pulsa la tecla de Intro éste se almacena en la celda correspondiente a la primera unidad de análisis de la primera variable, el segundo valor se almacena en la celda correspondiente a la segunda unidad de análisis de la primera variable, y así sucesivamente hasta completar todos los datos de la primera variable. Después se activa la segunda variable, y se procede de la misma forma. Y así sucesivamente, hasta completar todos los datos de todas las variables.

#### *3.2.2.2. Introducir los datos correspondientes a una unidad de análisis*

A veces es más cómodo introducir los datos por unidad de análisis en vez de por variable. En este caso, se activará la fila correspondiente a la unidad de análisis y después se pasa a introducir los datos. Es decir, una vez activada una determinada fila, el valor tecleado se almacena en la primera columna de la fila activada, el segundo valor se almacena en la segunda columna, y así sucesivamente hasta completar todos los datos del primer caso.

Si no se ha activado ninguna fila ni ninguna columna, y después de introducir un valor se pulsa la tecla de tabulación, en vez de la tecla de Intro, los datos se almacenarán por fila en lugar de por columna.

*3.2.2.3. Restricciones de los valores de datos* El ancho y el tipo de variable definidos determinan el tipo de valor que se puede introducir en una casilla.

Si se escribe un carácter no permitido por el tipo de variable definido, no se introducirá y se producirá una señal acústica.

Para variables de cadena (string), no se permite que el número de caracteres exceda del ancho definido.

Para variables numéricas se pueden introducir valores enteros que excedan del ancho definido, no obstante se presentará en pantalla asteriscos indicándonos esta circunstancia.

### **3.3. Almacenar datos**

Una vez que se ha introducido la matriz de datos utilizando el SPSS, ésta se puede almacenar en un fichero. El SPSSWIN permite almacenar los datos y la definición de las variables en distintos tipos de ficheros; el tipo que se elija dependerá de lo que se vaya a hacer posteriormente.

Para almacenar los datos, se pincha en el menú **Archivo** y se selecciona la opción **Guardar datos o Guardar como...** En ambos casos se abrirá la siguiente ventana:

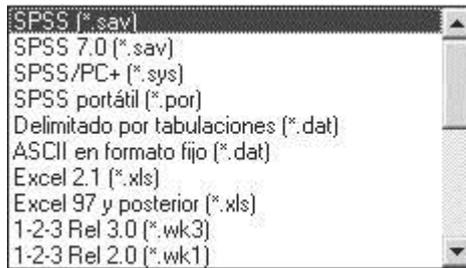


Esta ventana tiene 3 partes:

**1. Nombre de archivo :** es el campo donde se especifica el nombre del fichero donde se almacenarán los datos. En cuanto a la extensión del nombre del fichero, diremos que en el caso de un fichero SPSS, no es obligatorio mantener las extensiones que figuran por defecto (éstas dependen del tipo de fichero), pero si es aconsejable utilizarlas, pues dan idea del tipo de fichero a tratar.

**2. Guardar como archivos de tipo:** se utiliza para especificar el tipo del fichero en el que se desea almacenar los datos. Al pinchar en el símbolo

▾ de dicho campo aparece un menú con los tipos de ficheros disponibles:



Los tipos de ficheros más utilizados son:

— *SPSS (\*.sav)*: Formato SPSS. Los archivos de datos guardados con formato SPSS no se pueden leer en versiones anteriores a la 7.5. Al utilizar archivos de datos con nombres de variable con longitud superior a ocho bytes en SPSS 10.X u 11.X, se utilizan versiones únicas de ocho bytes de los nombres de variable, pero se mantienen los nombres originales para su utilización en la versión 12.0 o posteriores. En versiones anteriores a SPSS 10, los nombres de variable largos originales se pierden si se guarda el archivo de datos.

— *SPSS 7.0 (\*.sav)*: Formato SPSS 7.0 para Windows. Para la versión 7.0 del SPSS y versiones anteriores del SPSS para Windows.

— *SPSS/PC+ (\*.sys)*: Es un fichero binario que sólo puede crearse y leerse con el SPSS para Ms-Dos. Si el fichero de datos contiene más de 500 variables, sólo se guardarán las primeras 500. Sólo admite un valor perdido por variable; cuando exista más de uno se cogerá sólo el primero.

— *SPSS portátil (\*.por)*: Archivo de SPSS portátil que pueden leer otras versiones de SPSS en otros sistemas operativos (Macintosh, UNIX). Los nombres de variable se limitan a ocho bytes, y convertirán a nombres únicos de ocho bytes si es preciso.

— *Delimitado por tabuladores (\*.dat)*: Es un fichero ascii que sólo contiene la matriz de datos, por lo tanto se podrá leer con cualquier programa de análisis de datos, y en general, con cualquier programa que lea ficheros ascii. Tiene formato fijo y no existe tabuladores ni espacios en blanco entre los campos de variables.

— *Excel (\*.xls)*: Archivo de hoja de cálculo de Microsoft Excel 4.0. El número máximo de variables es 256.

— *1-2-3 versión 3.0 (\*.wk3)*: Archivo de hoja de cálculo de Lotus 1-2-3, versión 3.0. El número máximo de variables que puede guardar es 256.

— *SYLK (\*.slk)*: Formato de vínculo simbólico para archivos de hojas de cálculo de Microsoft Excel y de Multiplan. El número máximo de

variables que puede guardar es 256.

- *dbase IV* (\*.dbf): Formato dbase IV.
- *dbase III* (\*.dbf): Formato dbase III.
- *dBASE II* (\*.dbf): Formato dBASE II.
- *SAS v6 para Windows* (\*.sd2): Formato de archivo de SAS v6 para Windows/OS2.
- *SAS v6 para UNIX* (\*.ssd01): Formato de archivo de SAS v6 para UNIX (Sun, HP, IBM).
- *SAS v6 para Alpha/OSF* (\*.ssd04): Formato de archivo de SAS v6 para Alpha/OSF (DEC UNIX).
- *SAS v7+ Windows extensión corta* (\*.sd7): Versión 7-8 de SAS para Windows con formato de nombre de archivo corto.
- *SAS v7+ Windows extensión larga* (\*.sas7bdat): Versión 7-8 de SAS para Windows con formato de nombre de archivo largo.
- *SAS v7+ para UNIX* (\*.ssd01): SAS v8 para UNIX.
- *Transporte de SAS* (\*.xpt): Archivo de transporte de SAS.

**3. Guardar en :** En este campo figura la estructura de subdirectorios de la unidad especificada y permite seleccionar la ubicación donde se almacenará el fichero de datos.

### **3.4. Lectura del fichero de datos**

Hay veces que ya se dispone del fichero de datos para realizar el análisis estadístico y no es necesario utilizar el SPSS para crearlo. Este fichero puede ser un fichero ascii, en cuyo caso sólo contiene datos, o puede haberse creado con el SPSS (versión de Windows o de Ms-Dos), con una base de datos, o con una hoja de cálculo, en cuyo caso contiene los datos y la definición de variables. Por lo tanto, la forma de leer el fichero de datos dependerá de las características del mismo.

#### **3.4.1. Abrir base de datos**

Se puede leer cualquier formato de base de datos del cual se disponga de controlador, tanto de forma local como en modo red. En el primer caso el controlador deberá estar en el ordenador local y en el segundo en el servidor de la red.

Se puede:

- a) Leer archivos de la base de datos:
  - Se seleccionará: Archivo->Abrir base de datos->Nueva consulta.
  - El origen de la base de datos.

- Seleccionar la tabla y campos de la base.
- Especificar cualquier relación existente entre las tablas.
- b) Editar una consulta de base de datos guardada.
  - Se seleccionará: Archivo->Abrir base de datos->Editar consulta.
  - Seleccionar el archivo de consulta (\*.spq) a editar.
  - Seguir las instrucciones para crear la consulta.
- b) Leer archivos de bases de datos con una consulta guardada.
  - Se seleccionará: Archivo->Abrir base de datos->Ejecutar consulta.
  - Seleccionar el archivo de consulta (\*.spq) que se desea ejecutar.
- Algunos archivos de base de datos pueden tener una palabra de paso para el acceso. Será necesario introducir dicha palabra.
- En algunos caso además tendrá una solicitud de entrada de datos adicionales, como por ejemplo el mes.

Cuando se realiza la entrada a una base de datos el sistema activa el asistente para bases de datos que mediante seis pasos guía al usuario para introducir el archivo en el sistema. En el primer paso se selecciona la fuente de datos; se habilita la entrada a la base, si tiene contraseña; y se selecciona los campos de datos. El segundo permite seleccionar los datos (campos, nombres de los mismos, ordenar éstos, etc.). El tercer paso se dedica a especificar las relaciones entre tablas. El cuarto a limitar la recuperación de casos, es decir, si se quiere recuperar la base en su totalidad o parte de la misma. En el quinto se definen las variables, etiquetas, etc. El sexto y último permite guardar la consulta para futuras utilizaciones, ver los resultados del proceso realizado y pegar la sintaxis.

**3.4.2. Lectura de un fichero de datos de texto** Cuando el fichero de datos es un fichero de texto el sistema permite diferentes formatos:

- Datos delimitados por tabulaciones.
- Datos delimitados por espacios.
- Datos delimitados por comas.
- Datos con formato de campos fijos.

En los archivos delimitados, también se puede especificar otros caracteres como delimitadores entre valores.

Para leer un archivo de datos de texto se necesita:

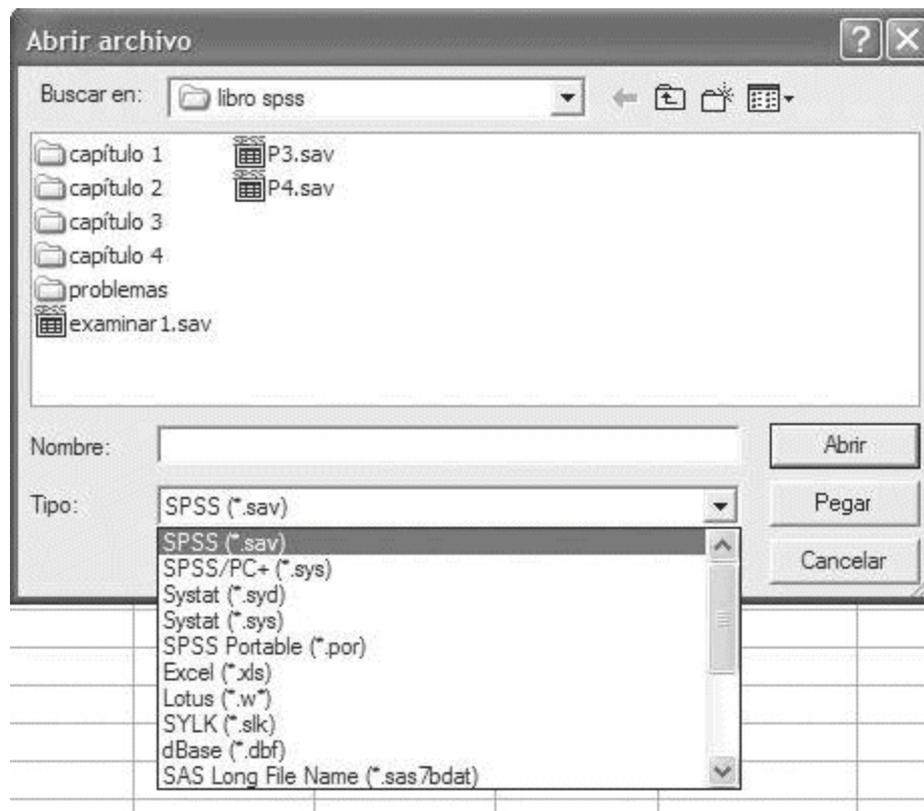
- Se seleccionará: Archivo->Leer datos de texto.
- Seleccionar el fichero de texto.
- Seguir las instrucciones del asistente para la importación de texto.

El asistente para la importación de texto utiliza seis pasos para realizar

el proceso. En el primero pregunta si el archivo de texto es de un formato predefinido. En el segundo, se selecciona la forma de organizar las variables (delimitas o de ancho fijo) y si están incluidas los nombres de las variables en la parte superior del archivo. El paso tercero está dedicado a seleccionar las características de los archivos delimitados o de ancho fijo, según la opción elegida en el segundo paso. El paso cuarto completa la visión del archivo de datos según la opción elegida de archivos delimitados o de ancho fijo. El quinto paso permite modificar las características de nombre y formato de datos de las variables del fichero a importar. Finalmente el paso sexto tiene la opción de pegar la sintaxis y de guardar el formato de archivo para futuros usos.

### 3.4.3. Lectura de otros tipos de fichero de datos

En general se puede leer fichero de datos en distintos formatos, comentados anteriormente al hablar de **Guardar como...** Se procede, Archivo->Abrir y aparecerá la pantalla que se muestra a continuación donde se selecciona el formato elegido, la ubicación y el nombre del fichero de datos que se desea abrir.



### 3.5. Un ejemplo de construcción de un fichero de datos

Para comprender los pasos relativos a los ficheros de datos, nos parece interesante comentar el proceso de construcción de un fichero de datos.

El presente ejemplo nace a caballo entre datos reales de una investigación sobre la aplicación de un programa metacognitivo de comprensión lectora a alumnos de EGB y Secundaria Obligatoria, y datos simulados para atender a otras necesidades.

Como se ha visto la primera tarea a abordar con los datos es precisamente, la **codificación** de las variables.

A su vez, y para SPSSWIN, ello conlleva varias acciones:

- Nombrar las variables.
- Codificar sus valores.
- Asignar un código específico a los valores ausentes.
- Asignar etiquetas significativas a variables y valores de algunas de ellas. — Construir la nueva matriz de datos.
- Grabar la misma sobre un soporte magnético.

Veamos ahora cada una de estas acciones.

### **3.5.1. Denominación de variables**

En primer lugar, se relacionan las variables estadísticas utilizadas en el ejemplo, agrupadas en cuatro bloques:

1. Variables de clasificación:

- a) Comunidad Autónoma.
- b) Grupo.
- c) Zona.
- d) Curso.
- e) Edad.
- f) Sexo.

2. Variables cognitivas:

- a) Inteligencia general.
- b) Aptitud verbal.
- c) Razonamiento abstracto.

3. Variables sociales y ocupacionales:

- a) Status sociocultural.
- b) Ocupación de los padres.

4. Variables pedagógicas:

- a) Aprendizaje en CC. Sociales.
- b) Vocabulario.

A las variables antes indicadas es preciso asignarles un nombre para que el sistema pueda identificarlas correctamente a lo largo de las diferentes etapas del proceso de los datos.

En general es recomendable asignar nombres significativos a las variables con objeto de facilitar las labores de proceso y, sobre todo, de interpretación de resultados.

Además, para completar su identificación, el sistema permite asignar adicionalmente etiquetas explicativas a los nombres de las variables. Pero, como luego veremos, las etiquetas no son más que una facilidad para la lectura de listados. El sistema identifica a las variables **solamente** y **exclusivamente** mediante el nombre de las mismas.

### **3.5.2. Codificación de los valores**

Tal codificación está en función de los tipos de variables que antes hemos comentado.

Al margen de esta recomendación, no existe ninguna regla específica para realizar la codificación de los valores.

Decir simplemente que conviene hacer un planteamiento lo más económico posible, de forma que se minimice el número de caracteres a introducir sin que ello implique, claro está, pérdida de información.

Al igual que con las variables, el sistema permite también la asignación de etiquetas a valores de las variables con idénticos fines.

En cualquier caso, siempre que sea posible, se considera como más recomendable realizar una codificación numérica de todas las variables.

### **3.5.3. Codificación de los valores ausentes**

Es habitual en cualquier trabajo que implique recogida de datos que en ocasiones, y para algunas variables y sujetos, carezcamos de la información a ellos referida.

A estos **valores ausentes** (*MISSING*) hay que asignarles un código específico con el fin de que el sistema los procese adecuadamente. La manera más habitual de proceder en estos casos es asignar un valor que estemos seguros nunca podrá aparecer como valor real de la variable en cuestión.

Por ejemplo, si para algunos sujetos carecemos de la información relativa a su comunidad, en nuestro caso podemos asignar a estos casos el valor 9, pues estamos seguros de que al tratarse sólo de tres comunidades este código nunca va a aparecer como valor real.

En cualquier caso, la asignación de valores ausentes debe ser coherente con los tipos de variables. Es decir, a una variable numérica le podemos asignar como valor ausente cualquier número, pero sólo números. De la misma manera, a una variable alfanumérica sólo podemos asignarle como

valor ausente un conjunto de caracteres (*string*).

Si, como antes se afirmó, se realiza una codificación numérica para todas las variables, es útil emplear un mismo esquema de codificación para los valores ausentes.

De esta manera se suele emplear series de nueves como códigos de los valores ausentes. Y así, a las variables de un solo dígito se les asigna el 9, a las de dos el 99, etc. Es una convención útil que facilita las tareas de preparación de los datos. Tiene como problema que, en ocasiones, el 9 — por ejemplo— representa un valor real de la variable. Imaginemos que trabajamos con sujetos de nueve comunidades. En tal caso, basta con recodificar los valores reales a dos dígitos (01, 02, etc.) y asignar el 99 como valor ausente.

Sea de esta manera, o de otra cualquiera, lo importante es prever esta posibilidad, y allá donde se requiera, asignar a estos valores códigos específicos que estemos seguros que nunca podrán aparecer como reales.

#### **3.5.4. Etiquetado de variables y valores**

Como se ha dicho, SPSSWIN permite asignar a las variables **etiquetas** con el fin de facilitar la lectura de los listados de salida.

Lógicamente, debe realizarse sólo para las variables cualitativas que deseemos, pues para las cuantitativas no tendría sentido.

Si hemos realizado una codificación numérica de todas las variables, la posibilidad de etiquetar los valores de las cualitativas se nos revela de mayor utilidad incluso que el etiquetado de las variables, pues contribuye en alto grado a facilitar la lectura e interpretación de los resultados.

#### **3.5.5. Elaboración del plan de codificación y de la matriz de datos**

Una vez tomada esta serie de decisiones, conviene registrarlas en un plan de codificación. En el mismo deben aparecer para cada variable, su **nombre**, **códigos** de los valores, **valores ausentes** y **etiquetas** si las hubiese.

A continuación, se procede a cumplimentar la matriz definitiva de datos.

Como se ha afirmado en varias ocasiones, la forma más usual de trabajo en SPSSWIN es con matrices rectangulares. Esto es, todos los sujetos tienen las mismas variables y ocupan el mismo número de líneas.

Además el formato de organización suele ser de los denominados «*fijos*». A saber: para todos los sujetos, las variables ocupan siempre el

mismo número de columnas (o si se prefiere, todos los valores de una variable tienen el mismo número de dígitos), y se encuentran siempre en las mismas columnas.

Por otra parte, hay que tener cuidado en el número de columnas (o dígitos) que se reservan para cada variable. Además del caso de los valores ausentes que ya hemos comentado, hay que tener en cuenta que todos los valores deben tener el mismo número de dígitos. Por tanto, visto el máximo de cada variable, por ejemplo 124, los valores inferiores a 100 deben completarse con tantos ceros o espacios en blanco a la izquierda como sea preciso (074, 005).

Además, a la hora de asignar columnas a las variables, conviene tener en cuenta si con posterioridad se van a incorporar al fichero nuevos datos relativos a las mismas variables pero a otros sujetos.

En tal caso, podría ocurrir que el número de dígitos reservado resultase insuficiente para algún sujeto, lo que implicaría costosas reformas del fichero.

Pensemos, por ejemplo, en la variable «curso». En nuestros datos actuales vemos que el máximo es 8 y, en consecuencia, asignamos una columna a esta variable. Si en una muestra posterior incorporamos otros cursos de Bachiller y queremos codificarlos con dos dígitos para distinguirlos nos causaría problemas y habría que modificar todo el fichero.

Anticipamos ya que en nuestro caso, y por aquello de predicar con el ejemplo, a esta variable le hemos asignado una sola columna. Es igual, porque si hay datos futuros, serán igualmente inventados y ya nos cuidaremos muy mucho de la creación de tales cursos.

Si los datos que disponemos se refieren a muchas variables, o éstas requieren muchas columnas, es posible que a la hora de grabar la matriz, ésta no quepa en las 80 columnas habituales. En tal caso, se puede optar por prolongar la línea hasta donde sea preciso aunque no podamos verla entera en la pantalla o, mejor, asignar a cada sujeto más de una línea.

Por último, con respecto a hipotéticos valores decimales, decir, en primer lugar, que SPSSWIN emplea la notación sajona, aunque se puede modificar por nuestra coma. Segundo, a la hora de registrar este tipo de valores no es preciso grabar el punto decimal pues los comandos relativos a la definición de los datos permiten especificar tal cuestión. Si se desea puede escribirse, pero en tal caso recordar que de cara a la grabación de los

datos, cualquier carácter incluido el punto ocupa una columna.

Dicho esto, sólo queda el arduo trabajo de escribir los datos codificados. La matriz de datos puede escribirse primero en papel y luego pasarla a soporte magnético mediante el empleo de cualquier tratamiento de textos. O puede optarse, por obviar la primera fase y grabar los datos directamente.

Sea como fuera el fichero de datos es la materia prima fundamental, por lo que conviene extremar las precauciones con el fin de evitar errores.

Con respecto a la entrada de los datos al sistema, lo habitual es que éstos estén grabados en un fichero de datos independiente, tal y como estamos explicando, en este caso se trata de un fichero de texto con extensión DAT. SPSSWIN permite también la entrada de datos «*en línea*», es decir, dentro de una sesión y mediante la introducción de los datos por el teclado. Decir que, en nuestra opinión, no es práctica recomendable excepto cuando se trate de una matriz pequeña o de un ejemplo.

Acabada pues la exposición, veamos ahora como ejemplo lo relativo a nuestros datos.

### **Ejemplo**

El siguiente plan de codificación recoge para cada variable, su nombre, la columna inicial y final que ocupan en la matriz de datos, su etiqueta, si la tiene, y los valores de la misma (mínimo y máximo, si es cuantitativa; códigos y etiquetas si es cualitativa).

#### **Variable Posición Etiqueta Valores**

COM 2 «COMUNIDAD»

GR 7 «GRUPO EXP/CONTROL»

ZO 9 «ZONA»

CU 15 «CURSO»

EDAD 19-20

SEXO 22

1 «ANDALUCIA»

2 «CANTABRIA»

3 «GALICIA»

1 «EXPERIMENTAL» 2 «CONTROL»

1 «URBANA»

2 «SEMIURBANA»

3 «RURAL»

6 «SEXTO»  
7 «SÉPTIMO»  
8 «OCTAVO»

MIN.: 10  
MÁX.: 15  
1 «HOMBRE»  
2 «MUJER»

FGPRE 23-24  
FGPOS 25-26  
AVPRE 27-28  
AVPOS 29-30 «PRETEST FACT. G» «POSTEST FACT. G» «PRETEST  
APT.VERBAL» «POSTEST APT.VERBAL»

**Variable Posición Etiqueta Valores** RAPRE 32-33 «PRETEST RAZ.  
ABST.» RAPOS 34-35 «POSTEST RAZ. ABST.» OP 37-38 «OCUPAC.  
PADRES» SC 39-40 «SOCIOCULTURAL» SOCPRE 41-42 «PRETEST  
PR.SOCIALE» SOCPOS 43-44 «POSTEST PR.SOCIALE» VOCPRE 45-  
46 «PRETEST PR.VOCAB.» VOCPOS 47-48 «POSTEST PR.VOCAB.»  
LECPRE 49-50 «PRETEST COMP.LECT.» LECPOS 51-52 «POSTEST  
COMP.LECT.»

### **3.5.6. Fichero de procedimiento**

Los comandos que corresponden con el proceso realizado con los datos del ejemplo nos da el primer fichero de procedimiento del programa SPSSWIN.

#### **3.5.6.1. Fichero de comandos. EJEMPLO.SPS**

Contiene todos los comandos de definición de datos y el de grabación del fichero de sistema (SAVE). Además, se ejecutará el comando LIST con el fin de listar cada variable y sus valores.

TITLE «PROGRAMA DE COMANDOS INICIALES».

DATA LIST FILE=«EJEMPLO.DAT»  
/COM 2 GR 7 ZO 9 CU 15 EDAD 19-20 SEXO 22  
FGPRE 23-24 FGPOS 25-26 AVPRE 27-28 AVPOS 29-30 RAPRE 32-33  
RAPOS 34-35 OP 37-38 SC 39-40 SOCPRE 41-42 SOCPOS 43-44  
VOCPRE 45-46 VOCPOS 47-48 LECPRE 49-50 LECPOS 51-52.

VARIABLE LABELS  
COM «COMUNIDAD»  
/GR «GRUPO EXP/CONTROL»  
/ZO «ZONA»

/CU «CURSO»  
/FGPRE «PRETEST FACTOR G»  
/FGPOS «POSTEST FACTOR G»  
/AVPRE «PRETEST APT.VERBAL»  
/AVPOS «POSTEST APT.VERBAL»  
/RAPRE «PRETEST RAZ. ABST.» /RAPOS «POSTEST RAZ. ABST.»  
/OP «OCUPAC. PADRES»  
/SC «SOCIOCULTURAL»  
/SOCPRE «PRETEST PR.SOCIALES» /SOCPOS «POSTEST  
PR.SOCIALES» /VOCPRE «PRETEST PR.VOCAB.» /VOCPOS  
«POSTEST PR.VOCAB.» /LECPRE «PRETEST COMP.LECT.»  
/LECPOS «POSTEST COMP.LECT.».

#### VALUE LABELS

COM

- 1 «ANDALUCIA»
- 2 «CANTABRIA»
- 3 «GALICIA»

/GR

- 1 «EXPERIMENTAL»
- 2 «CONTROL»

/ZO

- 1 «URBANA»
- 2 «SEMIURBANA»
- 3 «RURAL»

/CU

- 6 «SEXTO»
- 7 «SEPTIMO»
- 8 «OCTAVO»

/EDAD

- 10 «DIEZ AÑOS»
- 11 «ONCE AÑOS»
- 12 «DOCE AÑOS»
- 13 «TRECE AÑOS»
- 14 «CATORCE AÑOS»
- 15 «QUINCE AÑOS»

/SEXO

- 1 «HOMBRE»
- 2 «MUJER».

MISSING VALUE CU(99) SEXO(3). LIST.

SAVE OUTFILE='EJEMPLO.SAV'. EXECUTE.

#### 4. TRANSFORMACIONES DE LOS DATOS

Una de las principales ventajas que aporta SPSSWIN al proceso de datos es que —previamente al análisis estadístico de los mismos, o como resultado de alguno en concreto— es posible *transformar los datos originales* presentes en el fichero de datos.

Tales transformaciones pueden estar orientadas hacia la recodificación de los valores de una(s) variable(s) o hacia la generación de nuevas mediante operaciones a realizar sobre las preexistentes.

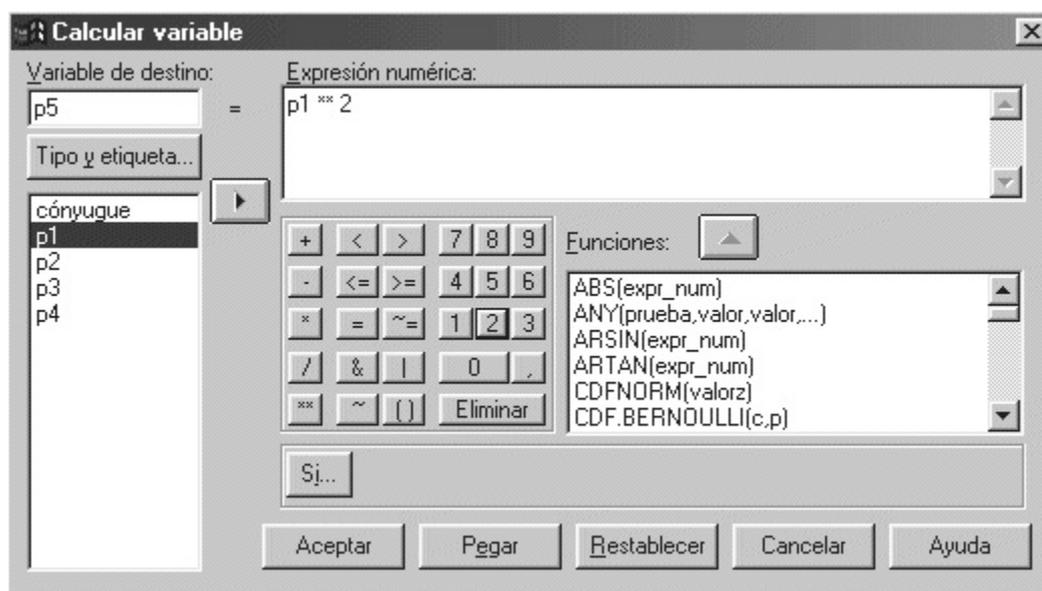
Los cinco comandos relativos a esta cuestión son: RECODE, AUTORECODE, COMPUTE, RANK y COUNT.

Los dos primeros recodifican los valores de una variable. Mediante COMPUTE se pueden generar nuevas variables empleando para ello transformaciones de las antiguas. El comando RANK genera nuevas variables asignando rango a la primera.

Por último, COUNT permite crear nuevas variables en función de la ocurrencia o no de ciertos valores de una lista dada de variables originales.

##### 4.1. Crear variables

Se pueden crear variables mediante transformación numérica o de cadena de otras variables. En las nuevas variables podemos especificar el tipo y la etiqueta de la variable. Se puede calcular valores de forma selectiva según cumplan unas condiciones lógicas. El SPSS incorpora más de 70 funciones aritméticas, estadísticas, de distribución y de cadena.



Para

calcular variables debemos:

— Pulsar en:

Transformar.

Calcular variable.

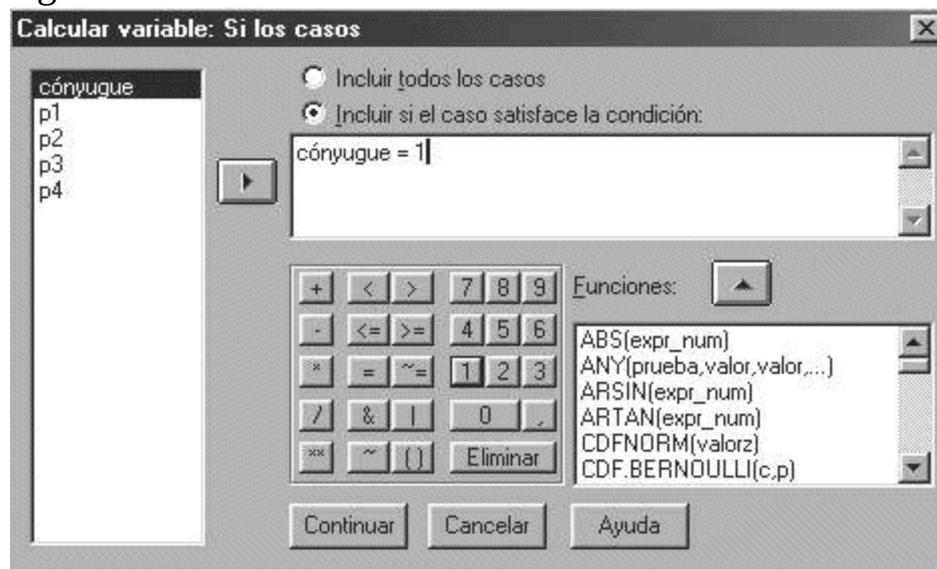
— Elegir una nueva variable o la existente.

— Pegar las funciones elegidas y tener en cuenta las siguientes observaciones:

- Las constantes de cadena deben estar entre comillas o apóstrofes.
- Para nuevas variables de cadena es conveniente incluir **tipo y etiqueta**.

#### 4.1.1. Calcular variables: Si los casos deben cumplir una condición lógica

Si observamos la figura anterior, pulsando el botón **Si**, entramos en el siguiente menú:



— Si el resultado es **cierto**, se aplicará la transformación.

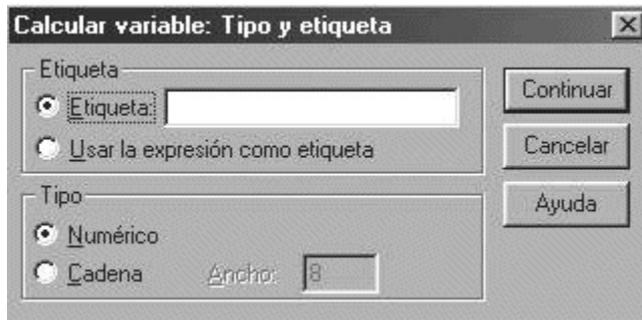
— Si el resultado es **falso o missing**, la transformación no se aplicará.

Las expresiones condicionales pueden incluir nombres de variables, constantes, operaciones aritméticas, funciones numéricas, variables lógicas y operadores relacionales.

##### 4.1.1.1. Indicar tipo y etiqueta a la variable creada

Por defecto las nuevas variables creadas son numéricas. Cuando se necesite una variable cadena hay que indicar el tipo de dato y el ancho. Las variables de cadena no se pueden utilizar en los cálculos.

Como siempre, se puede poner **etiqueta** a la nueva variable creada. La longitud máxima admisible son 120 caracteres. También se puede usar los 110 primeros caracteres de la expresión como etiqueta.



## 4.2. Recodificar valores

Los valores de datos se pueden modificar mediante recodificación. Esta operación es sobre todo interesante para agrupar categorías. Se puede recodificar los valores de variables existentes o crear unas nuevas variables con los valores recodificados de las variables existentes.

### 4.2.1. Recodificación en las mismas variables

Se pueden recodificar variables numéricas y de cadena. Si se seleccionan varias variables para su recodificación, todas deben ser del mismo tipo.



Para recodificar los valores de una variable:

— Se pulsará:

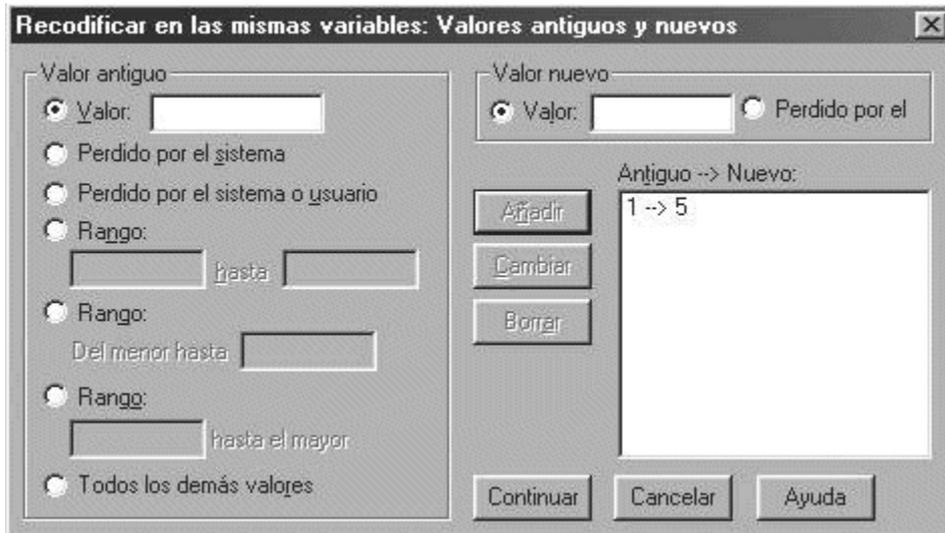
Transformar.

Recodificar.

En las mismas variables.

— Después se seleccionará las variables, **especificando** los **valores antiguos** y los **nuevos**. Si se desea, se puede definir un conjunto de casos que se van a modificar.

4.2.1.1. Recodificación en las mismas variables: valores antiguos y nuevos

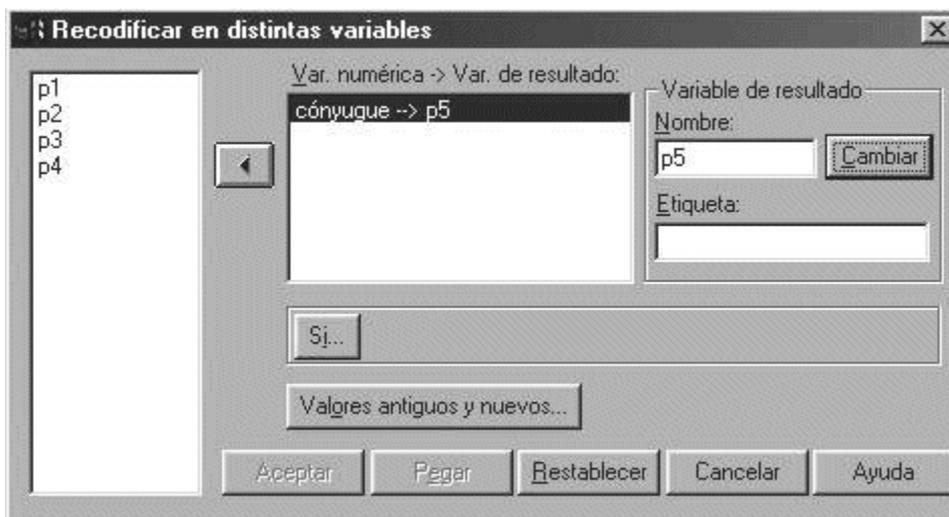


Aparecerá un cuadro de dialogo donde se pondrá:  
**Valor antiguo.** Se indicará un solo valor, un rango de valores o los valores perdidos por el sistema o indicados como perdidos por el usuario.  
**Valor nuevo.** Se puede indicar un valor o indicar perdido por el sistema.

**Antiguo-Nuevo** . Se va introduciendo los valores recodificados, pulsando añadir, cambiar o borrar según interese. El SPSS va ordenando los valores conforme se añaden.

#### 4.2.2. Recodificación en distintas variables

La recodificación sirve para reasignar los valores o rango de valores de las variables existentes a una nueva variable. Permite, por ejemplo, pasar de una variable de cadena a una variable numérica o al contrario. Si seleccionamos un conjunto de variables, todas tienen que tener el mismo rango. No se pueden recodificar juntas variables numéricas y de cadena.



Para recodificar realizaremos los siguientes pasos:

— Elegir en el menú:  
Transformar.

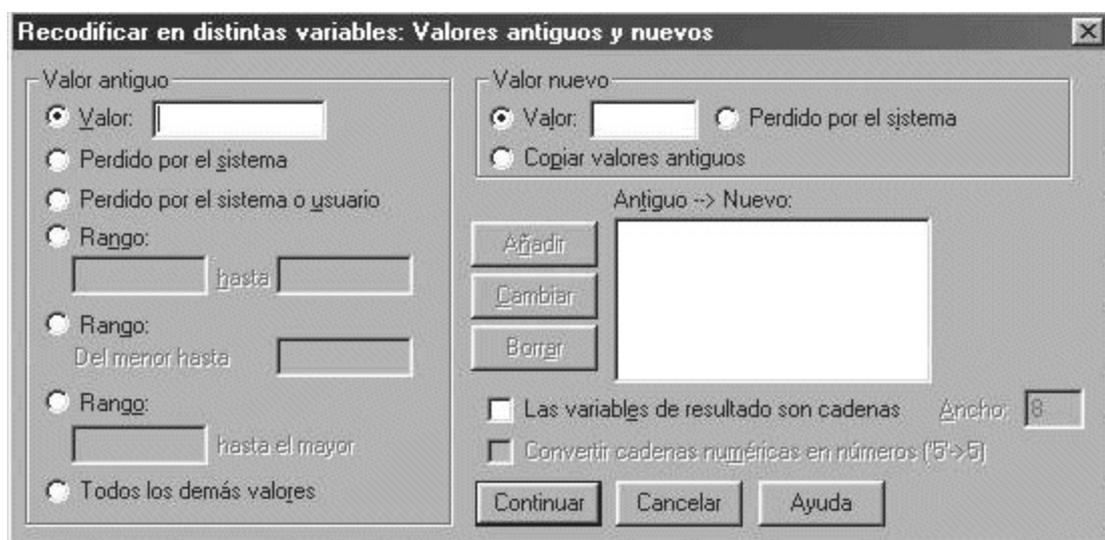
Recodificar.

En distintas variables.

— Seleccionar las variables que se quieren recodificar y pulsar **Cambiar**.

— Finalmente pulsar **valores antiguos** y **nuevos** para recodificar los valores.

Esta última opción nos llevará a un cuadro de dialogo similar al que se ha comentado anteriormente:



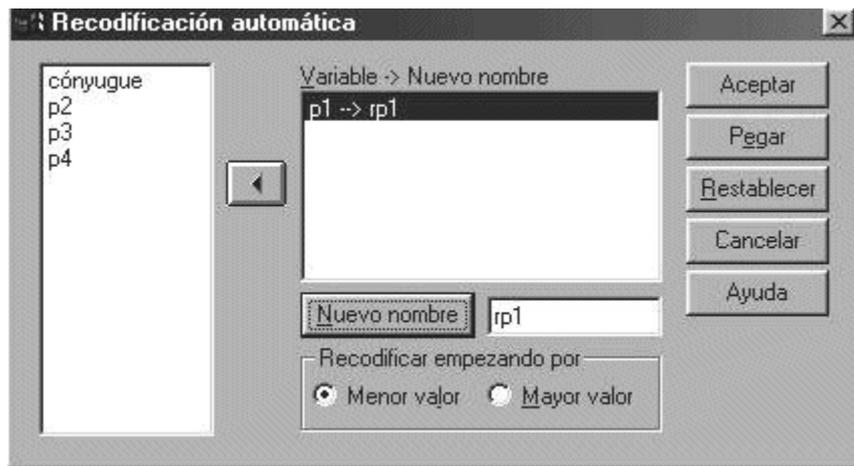
Sólo existen dos novedades:

**Copiar valores antiguos y las variables de resultado son cadenas.**

Esta última para facilitar la conversión de variables numérica a variables de cadena y la primera para cambiar **todos los demás valores** de la variable antigua y **copiar valores antiguos** por el **valor nuevo** seleccionado.

#### 4.2.3. Recodificación automática

También existe la posibilidad de recodificar de forma automática en otras variables que conservarán las etiquetas de valor de las variables antiguas. Aquellas variables que no tengan etiquetas en la recodificación, cogerán como etiqueta el valor original de la variable origen. Como resultado de la operación de recodificación se presentará una tabla de valores antiguos y nuevos y las etiquetas de valor.



Para recodificar de forma automática valores enteros o de cadena en valores enteros consecutivos. Será necesario:

— Elegir en los menús:  
Transformar.

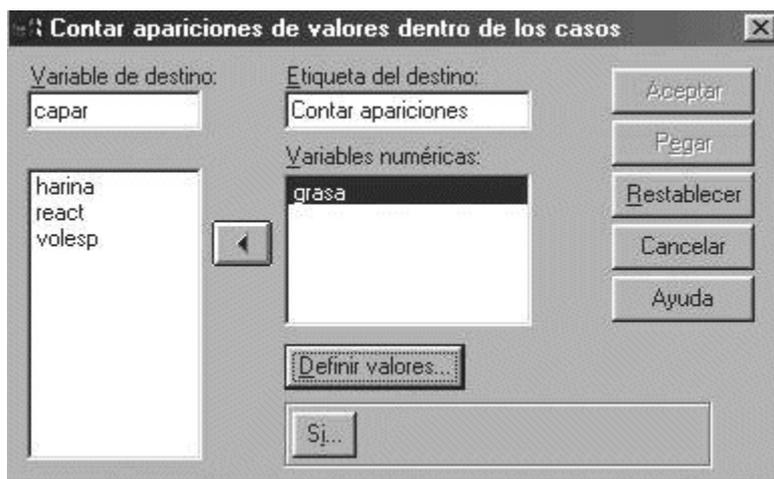
Recodificación automática.

— Seleccionar la variable o variables que desee recodificar.

— Para cada variable seleccionada, introducir un nombre para la nueva variable y pulsar en *Nuevo nombre*.

### 4.3. Contar apariciones de valores dentro de los casos

Nos sirve para contar las apariciones del mismo valor o valores en una lista de variables. Por ejemplo si queremos contar los diferentes tipos de pan, que tenemos en el fichero de datos, según tengan una cantidad definida de grasa (1) y harina (1). Procederemos de la siguiente manera:



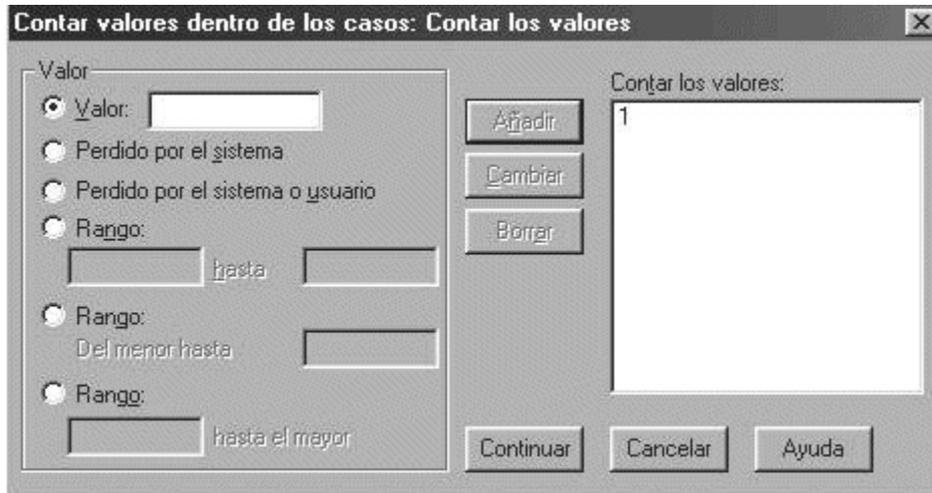
Para contar las apariciones:

— Elegir en el menú:  
Transformar.

Contar apariciones.

— Introducir el nombre de la variable de destino.

— Seleccionar dos o más variables del mismo tipo (numéricas o de cadena). — Pulsar definir valores, con lo cual entraremos en el siguiente cuadro de dialogo:

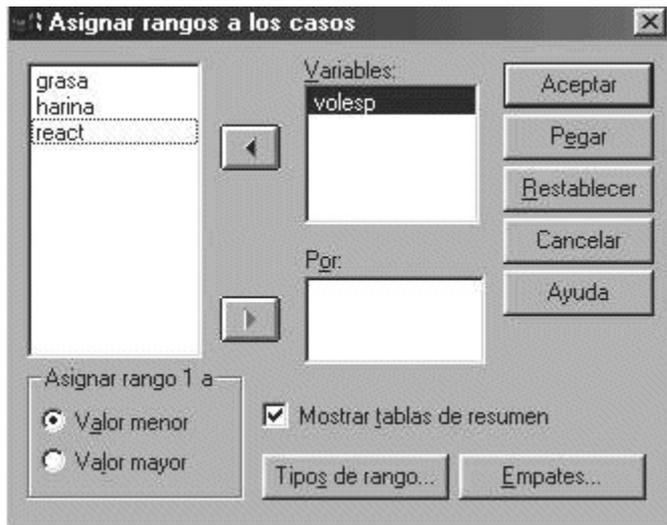


Se irá añadiendo valores según interese.

La forma de incrementar la frecuencia de la variable de destino será la siguiente: se incrementará en 1 cada vez que una de las variables seleccionada coincida con una especificación de la lista **Contar valores**. Los diferentes valores que puede tener la variable de destino coinciden con las diferentes combinaciones de valores que tengan las variables que queramos contar. Naturalmente la variable de destino siempre será numérica.

#### 4.4. Asignar rangos a los casos

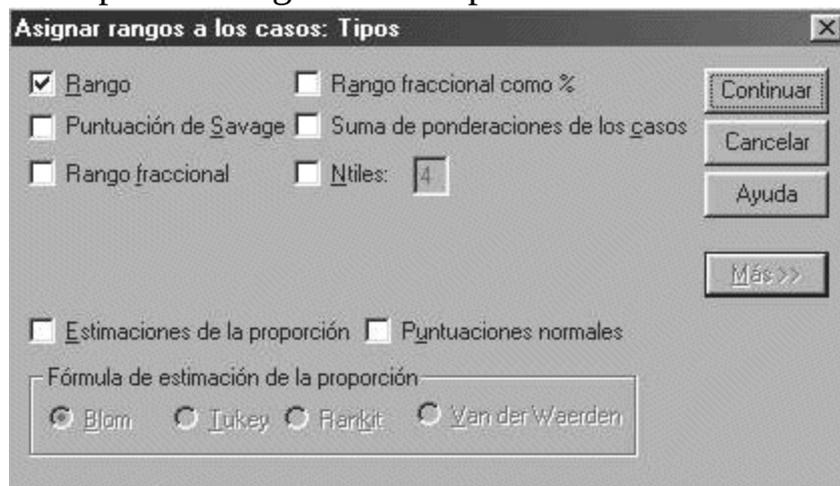
Sirve para asignar rangos, puntuaciones de Savage y normales, ntiles y percentiles, para las variables numéricas creando automáticamente nuevas variables que tendrán un nombre similar a la variable a la que se asigne el rango, anteponiendo algún carácter esclarecedor; por ejemplo, en el caso de los ntiles se antepone la n, así los ntiles de la variable peso se denominará npeso.



Los pasos que realizaremos para asignar rangos a los casos serán los siguientes: — Elegir en el menú: Transformar.

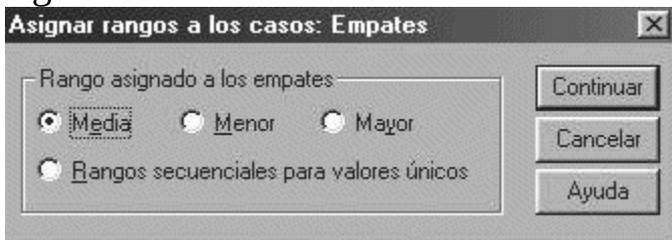
Asignar rangos a casos.

— Seleccionar las variables. Sólo se podrán seleccionar variables numéricas. Los tipos de asignaciones aparecen en el cuadro de dialogo



siguiente:

Para los empates entre valores de la variable los rangos pueden tomar los valores siguientes:



Las diferentes maneras de asignación de rangos en los empates se explican en el cuadro siguiente:

**Valor Media Menor Mayor Secuencial**

20 1 1 1 1  
26 3 2 4 2

26 3 2 4 2

26 3 2 4 2

28 5 5 5 3

30 6 6 6 4

Para comprender la tabla téngase en cuenta que la columna Valor muestra los diferentes valores de la variable, y las columnas: Media, Menor, Mayor y Secuencial, los distintos rangos asignados a la variable según el método de resolver los empates.

#### **4.5. Categorizador visual**

El categorizador visual se utiliza para crear nuevas variables a partir de otras. Se puede utilizar para categorizar una variable continua o para agrupar las categorías, es decir, para reducir el número de las mismas en una variable ordinal.

El programa SPSS no muestra como variables seleccionables aquellas variables nominales o de cadena.

Para realizar la categorización visual habrá que:

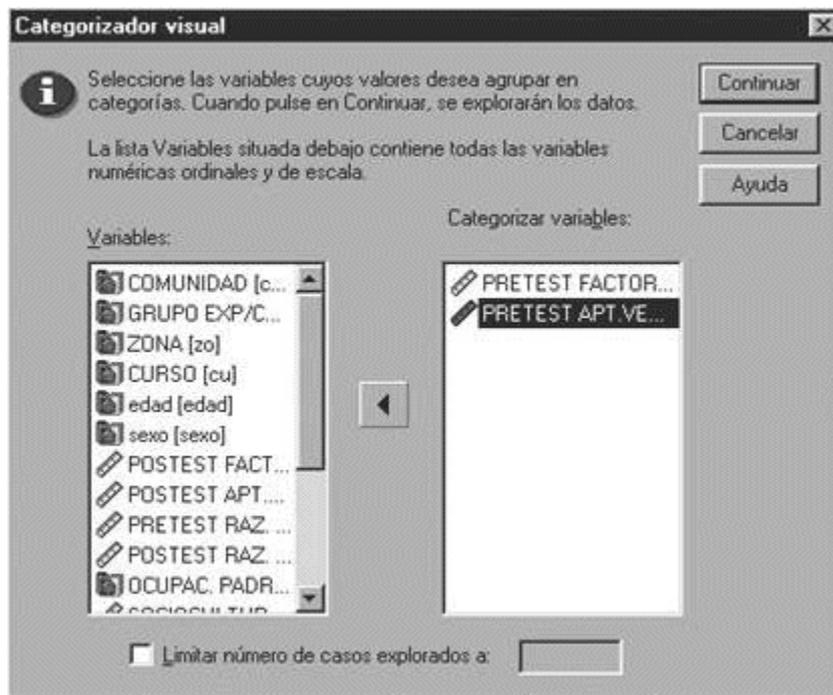
— Seleccionar en los menús:

Transformar.

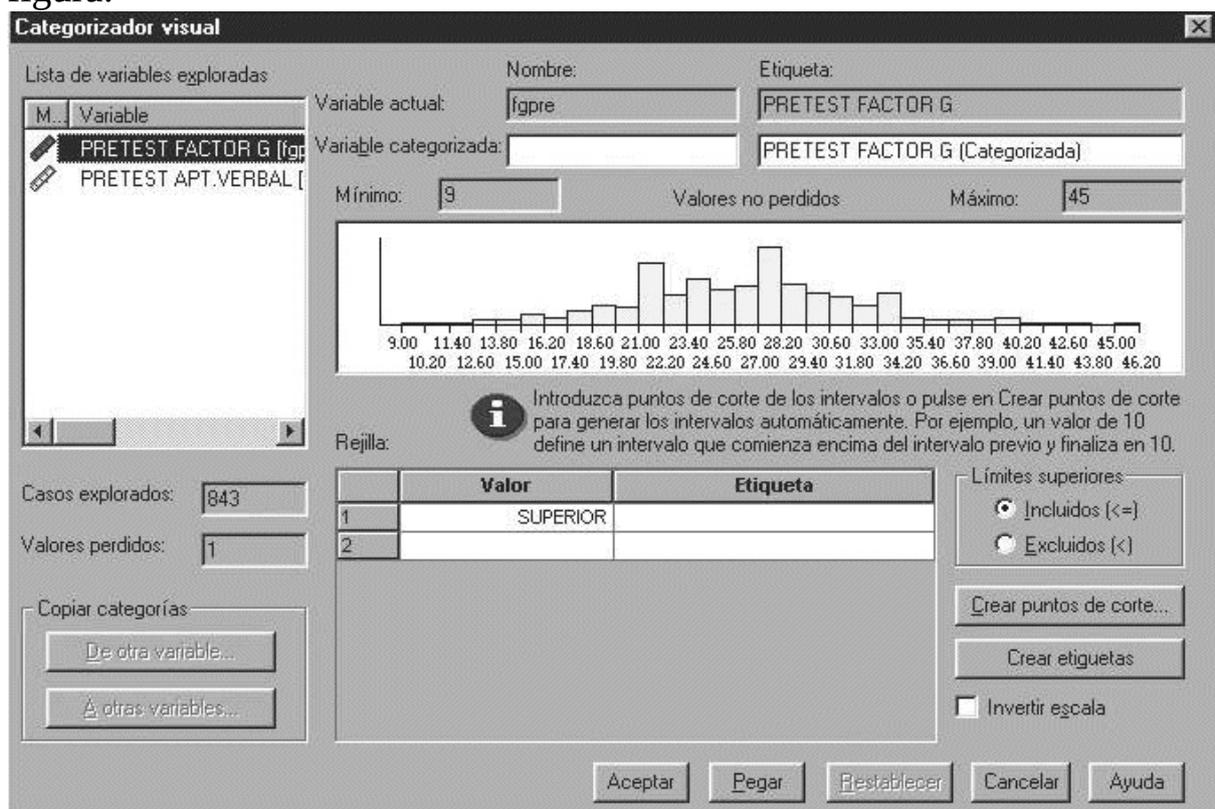
Categorizador visual.

— Seleccionar las variables numéricas de escala u ordinales para las que desee crear nuevas variables categóricas (en intervalos).

Seleccionar los casos que se desean utilizar en el proceso o en su defecto, si no hay selección utilizar todos.



A continuación al pulsar aceptar se entrará en el cuadro de diálogo que muestra la siguiente figura:



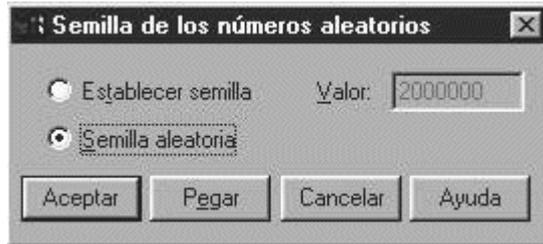
Aquí habrá que:

- Seleccionar una variable de la Lista de variables exploradas.
- Introducir un nombre para la nueva variable categorizada (recuérdese que no se pueden repetir el nombre de las variables en el

fichero de datos). — Definir los criterios de categorización para la nueva variable.

#### 4.6. Semilla de aleatorización

La semilla de aleatorización sirve para establecer un valor específico como origen utilizado por el generador de números pseudo-aleatorios, de modo que se pueda reproducir una secuencia concreta de números pseudo-aleatorios.



La semilla de aleatorización cambia cada vez que se genera un número aleatorio, por ejemplo en transformaciones de las variables. Si se quiere replicar una secuencia de números aleatorios, se utilizará el cuadro de diálogo mostrado en la figura anterior para restituir la semilla a un valor específico, antes de realizar alguna transformación, o análisis que utilice los números aleatorios.

Por defecto, la semilla de aleatorización se restablece automáticamente con 2.000.000 cada vez que se inicia una nueva sesión de SPSS.

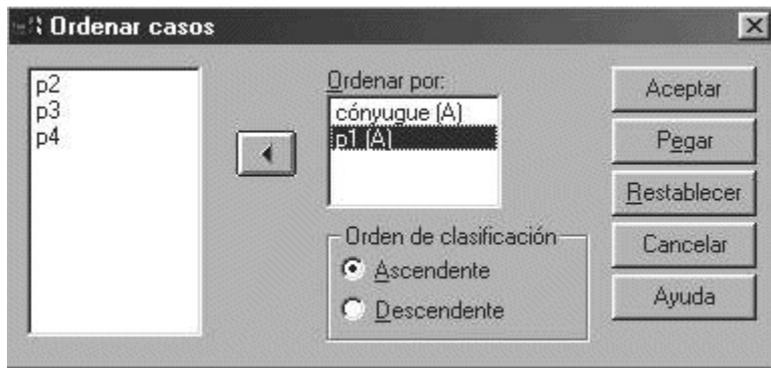
### 5. OPERACIONES EN Y CON LOS ARCHIVOS

#### 5.1. Ordenar casos

Nos facilita la ordenación de casos tanto de modo ascendente como descendente en función de los valores de una o varias variables.

Si hay más de una variable criterio para ordenar, entonces se procede de forma anidada, es decir, primero se ordena la primera variable y después se ordenará la segunda para cada caso de la primera variable.

Se debe tener en cuenta que en las variables de cadena las letras mayúsculas preceden a las minúsculas.



Para ordenar los casos tenemos que realizar los siguientes pasos: —  
Elegir en el menú:

Datos.

Ordenar casos.

— Seleccionar la variable o variables de ordenación.

## 5.2. Fusión de archivos

Se puede unir dos ficheros de datos de dos maneras distintas:

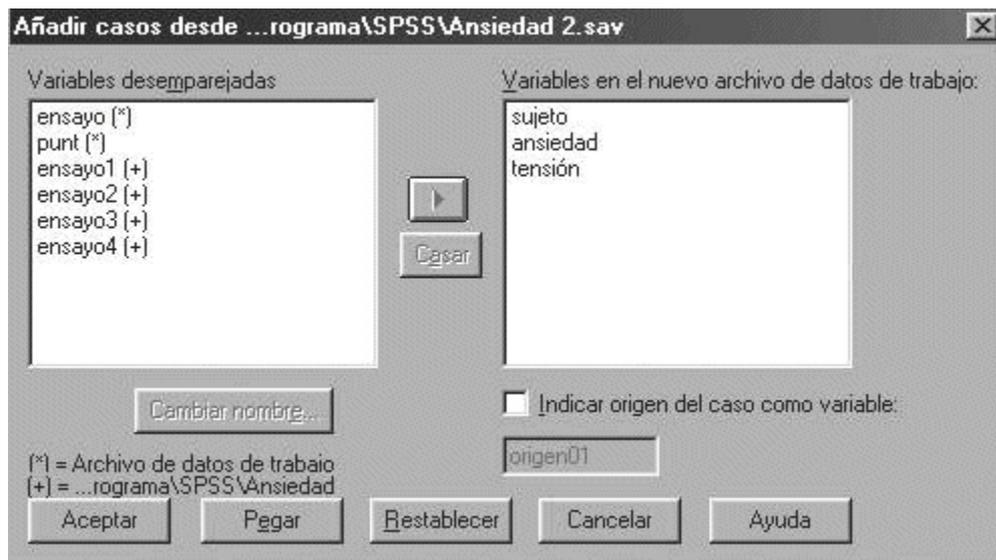
**Añadiendo casos.** Es decir, si coinciden en las variables se añaden los casos de unos ficheros con los datos de otro.

**Añadiendo variables.** Si los casos coinciden se pueden anexar las variables de un fichero de datos con las de otro.

### 5.2.1. Añadir casos

Como hemos dicho se unen los casos de dos ficheros que coinciden en las variables. En los dos ficheros de datos no tiene que coincidir el orden de colocación de las variables, sin embargo los casos si deben estar ordenados en ambos ficheros.

Cualquier información sobre etiquetas de variable y de valor, valores perdidos por el usuario, formatos de visualización en el archivo de datos de trabajo se aplicará al archivo de datos fusionado.



En el cuadro de diálogo aparecen:

**Variables desemparejadas** . Son las variables que quedan excluidas del fichero de datos fusionado. Las variables del archivo de datos de trabajo se identifican con asterisco (\*). Las variables del archivo de datos externo se identifican mediante un signo más (+). En esta lista pueden estar:

- Variables definidas como datos numéricos en un archivo y como datos de cadena en el otro.
- Variables de cadena de longitud diferente.

**Variables en el nuevo archivo de datos de trabajo** . Las variables que se incluirán en el nuevo archivo de datos fusionado. Por defecto en esta lista estarán las variables que coinciden en el nombre y el tipo de datos. Se pueden eliminar variables de la lista que no queramos que estén en el fichero de datos fusionado.

Para fusionar los archivos con las mismas variables y distintos casos tendremos: — Abrir uno de los archivos de datos.

— Elegir en el menú:

Datos.

Fundir archivos.

Añadir casos.

— Seleccionar el archivo de datos que se desee fundir con el fichero anteriormente abierto.

— Eliminar todas las variables que se desee del nuevo archivo de datos. —

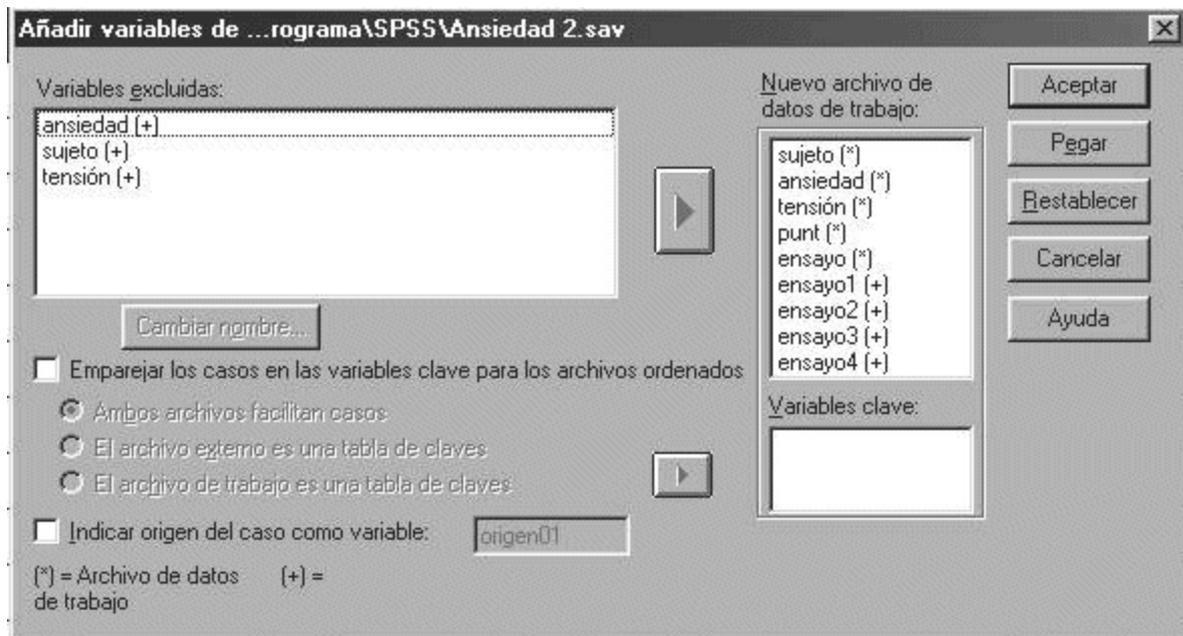
Añadir cualquier pareja de variables de la lista de **variables desemparejadas** que aún teniendo distinto nombre represente lo mismo.

Se puede cambiar el nombre de alguna variable del fichero de datos de trabajo y del fichero de datos externo pulsando la tecla correspondiente.

### 5.2.2. Añadir variables

En este caso los dos ficheros a fusionar coinciden en el número de casos, pero difieren en las variables de cada fichero de datos. Para realizar esta labor se debe cumplir: — Los casos deben tener el mismo orden en los dos ficheros y respecto a las mismas variables clave.

— Si dos o más variables coinciden en el nombre se suprimirán del fichero de datos externo al incorporarse al fichero fusionado, ya que se supone contienen la misma información.



Para fusionar los archivos con los mismos casos y distintas variables tendremos: — Abrir uno de los archivos de datos.

— Elegir en el menú:

Datos.

Fundir archivos.

Añadir variables.

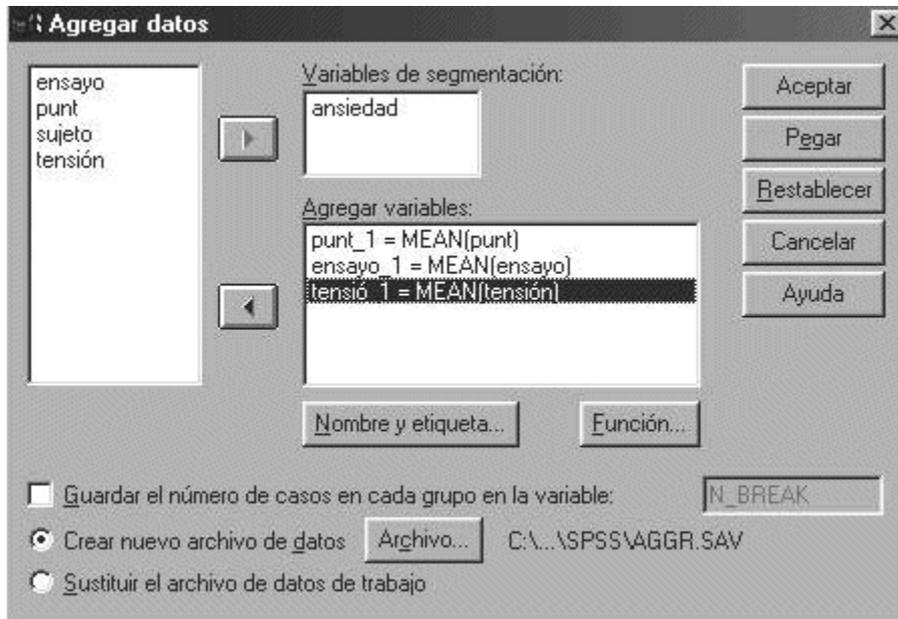
— Seleccionar el archivo de datos que se desee fundir con el fichero anteriormente abierto.

— Se puede incluir alguna variable (+) del archivo de datos externo de la lista de variables excluidas, para esto podemos cambiar el nombre, pulsando la tecla correspondiente.

### 5.3. Agregar datos

Agregar datos permite crear un fichero de datos con el resumen o

agrupamiento de casos basándose en el valor de una o más variables. El nuevo archivo contiene un caso por cada grupo. Por ejemplo se puede crear un fichero agrupado con la media de notas de los alumnos en distintas clases.

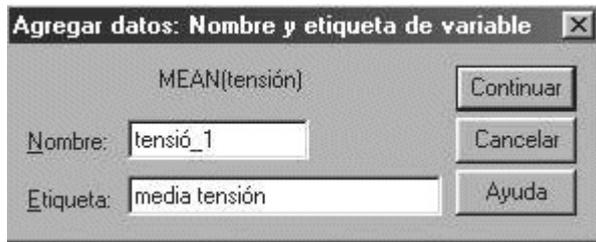


En este cuadro de diálogo hay que tener en cuenta:

**Variables de segmentación** . Utilizadas para agrupar los casos en función de las mismas. Cada valor de la variable de segmentación genera un grupo al que se aplicará la función de agregación dando como resultado un sujeto del fichero agregado. Todas las variables de segmentación se guardan en el fichero segmentado con su definición. Pueden ser numéricas o de cadena.

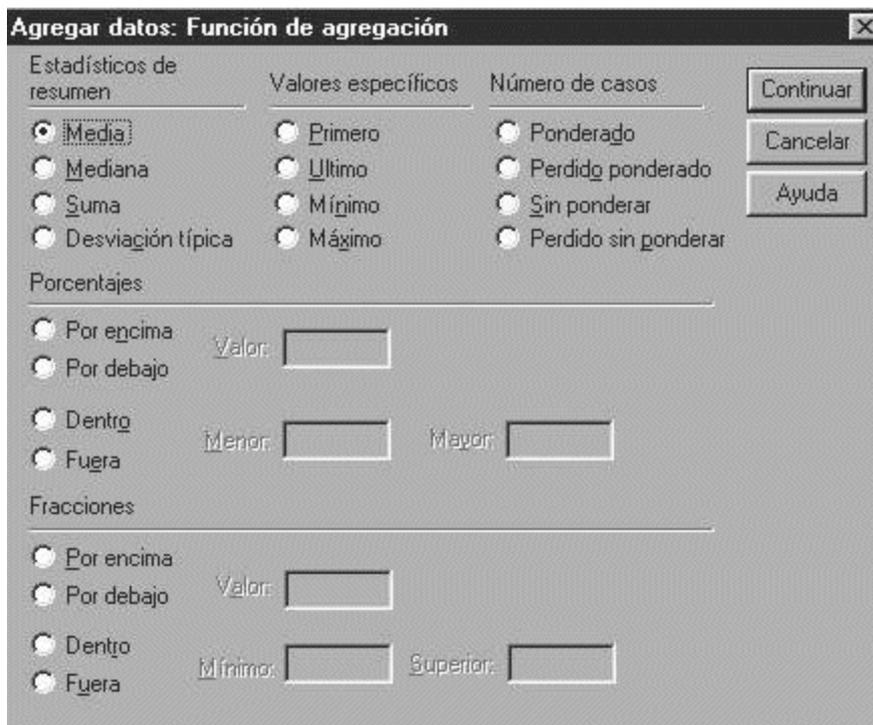
**Variables agregadas** . Variables que resultan de realizar en los grupos la función de agregación. SPSS las nombra con los primeros caracteres de las variables del fichero de datos origen seguidos de un subrayado y un número secuencial de uno o dos dígitos. Las variables origen deben ser numéricas.

**Nombre y etiqueta.** Podemos cambiar el nombre y la etiqueta de la variable agregada, pero siempre con las limitaciones comentadas anteriormente en relación al número de caracteres para el nombre de variable y para las etiquetas de las mismas.



**Función de agregación.** Las funciones de agregación que podemos utilizar son las siguientes:

- **Estadísticos de resumen:** media, mediana, suma y desviación típica.
- **Valores específicos:** primero, último, mínimo y máximo.
- **Número de casos:** ponderado, perdido ponderado, sin ponderar y perdido sin ponderar.
- **Porcentaje o fracción de valores:** por encima o debajo de un valor determinado. Porcentaje o fracción de valores dentro o fuera de un rango.



Para agregar un archivo de datos debemos:

— Elegir en el menú:

Datos.

Agregar.

— Seleccionar una o más variables que definan los grupos a formar.

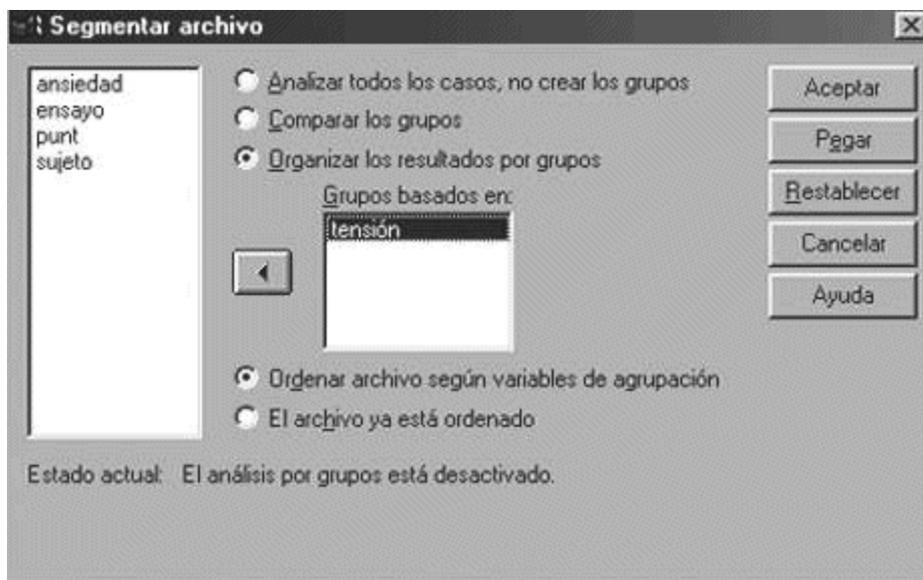
— Seleccionar una o más variables agregadas que se van a incluir en el fichero de datos agregado.

— Seleccionar una función de agregación para cada variable agregada. — Opcionalmente podemos cambiar el nombre y la etiqueta de las variables agregadas, así como el nombre del fichero de destino de la agregación.

#### 5.4. Segmentar archivos

Segmentar archivos divide el archivo de datos en grupos diferentes para el análisis basándose en los valores de una o más variables de agrupación. Se debe tener en cuenta:

- Si agrupamos por una variable de cadena, sólo los 8 primeros caracteres sirven como criterio de agrupación.
- Se puede especificar hasta 8 variables de agrupación.
- Es conveniente ordenar los casos según las variables de agrupación.



Conviene destacar del cuadro de diálogo:

**Comparar los grupos** . Los grupos de segmentación se presentan juntos para compararlos. Las tablas pivote que se generen en los análisis tendrán los resultados juntos, es decir, habrá una sola tabla pivote. Para los gráficos, se crea un gráfico para cada grupo pero se presentan juntos en el navegador de resultados.

**Organizar los resultados por grupos.** Los resultados de cada análisis se presentan por separado para cada grupo de segmentación.

Para segmentar un archivo de datos debemos:

— Elegir en el menú:

Datos.

Segmentar archivo.

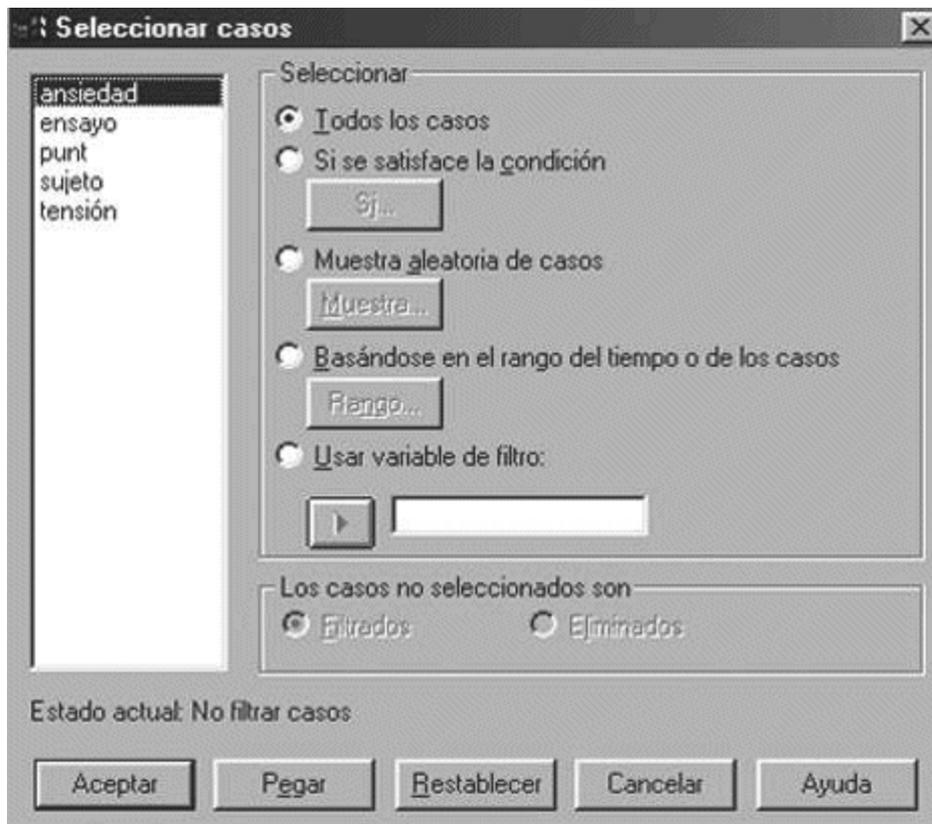
— Seleccionar comparar los grupos u organizar los resultados por grupos.

- Seleccionar una o más variables de agrupación.
- Ordenar archivo según variable de agrupación.

## 5.5. Seleccionar casos

Seleccionar casos nos permite seleccionar un subconjunto de datos según diversos métodos: si satisfacen una condición, una muestra aleatoria de casos, basándose en el rango del tiempo o de los casos, etc.

Los casos no seleccionados se pueden o bien filtrar o bien eliminar. SPSS crea una variable de filtro (filter\_\$) para indicar el estado de filtrado de un determinado caso. Los casos seleccionados tiene el valor 1 en la variable de filtro y los no seleccionados el valor 0. Los casos filtrados también están indicados con una barra transversal sobre el número de fila del editor de datos. Si damos a la opción **seleccionar todos los casos**, desactivamos el filtro de datos.



Para seleccionar un conjunto de casos debemos:

- Elegir en el menú:

Datos.

Seleccionar casos.

- Seleccionar uno de los métodos de selección de casos.
- Especificar si los casos van a ser filtrados o eliminados. Por defecto

está en la opción de filtrar.

### 5.5.1. *Seleccionar casos: Sí*

El cuadro de diálogo que figura a continuación nos muestra la opción que tenemos de seleccionar casos conforme al cumplimiento de una expresión condicional.

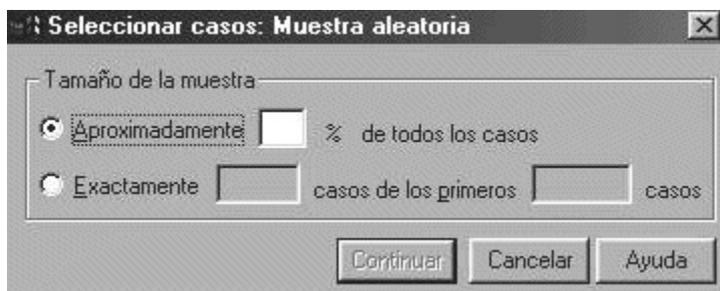


### 5.5.2. *Seleccionar casos: Muestra aleatoria*

Selecciona aleatoriamente un número de casos mediante dos opciones excluyentes:

- **Aproximadamente** . Un porcentaje aproximado de casos del fichero de datos. Conforme mayor número de casos tenga el fichero de datos el número de casos generados se aproximará más exactamente al porcentaje que hayamos seleccionado.

- **Exactamente** . Un número seleccionado por el usuario. Si seleccionamos una cantidad de casos mayor del número total de casos del fichero de datos, sólo se cogerá una fracción del fichero de datos.



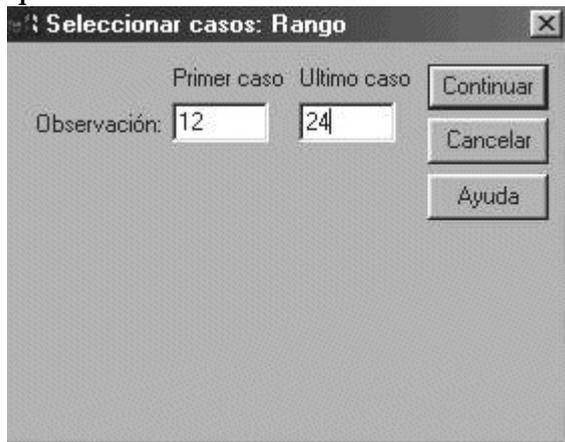
### 5.5.3. *Seleccionar casos:*

#### **Rango**

Se selecciona los casos conforme a un rango de números de casos o de fechas u horas.

El número de caso se toma del número de fila que ocupa el caso en el fichero de datos.

Los rangos de fechas y horas sólo se pueden realizar con variables fecha que fundamentalmente se utilizan en series temporales.



## 5.6. Ponderar casos

Permite ponderar los casos en función de los valores de una variable. Los casos con valores perdidos, negativos o cero para la variable de ponderación se excluyen del análisis.

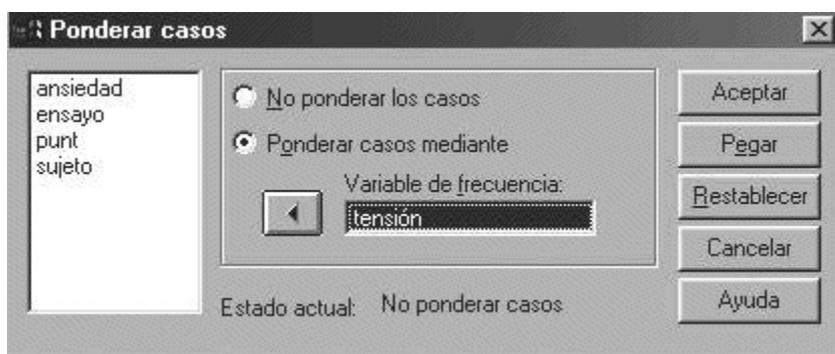
La sintaxis del comando que realiza la ponderación es:

**Weight by** variable.

Supongamos, por ejemplo, que todos los cines de las zonas rurales ofrecen siempre un programa doble, mientras que los de las ciudades sólo proyectan una película (y además son más caros).

En tal caso, debemos ponderar a los sujetos de las zonas rurales doblando sus puntuaciones en horas de cine.

Una vez aplicada una ponderación, permanece activada hasta que se selecciona otra variable de ponderación o se desactive la misma.



Para ponderar caso debemos:

— Elegir en el menú:

Datos.

Ponderar casos.

— Seleccionar **ponderar casos mediante**.

— Especificar una variable de frecuencia. Así por ejemplo si la variable de frecuencia toma el valor 4 nos indicará que habrá 4 casos en el archivo de datos ponderado, es decir, que el valor de la variable se repite 4 veces.

## 5.7. Transponer

La opción Transponer de SPSS crea un archivo de datos nuevo en el que se transponen las filas y las columnas del archivo de datos original de manera que los casos (las filas) se convierten en variables, y las variables (las columnas) se convierten en casos. De igual forma crea automáticamente nombres de variable y presenta una lista de dichos nombres.



El proceso de transponer genera:

- Una nueva variable de cadena, *case\_lbl*, que contiene el nombre de variable original.

- Si el archivo de datos de trabajo contiene una variable de identificación o de nombre con valores únicos, se puede utilizar como variable de nombre: sus valores se emplearán como nombres de variable en el archivo de datos transpuesto. Si se trata de una variable numérica, los nombres de las variables comenzarán por la letra *V*, seguida de un valor numérico.

- Los valores perdidos definidos por el usuario se convierten en el valor perdido del sistema en el archivo de datos transpuesto.

Para transponer variables se debe:

— Elegir en el menú:

Datos.

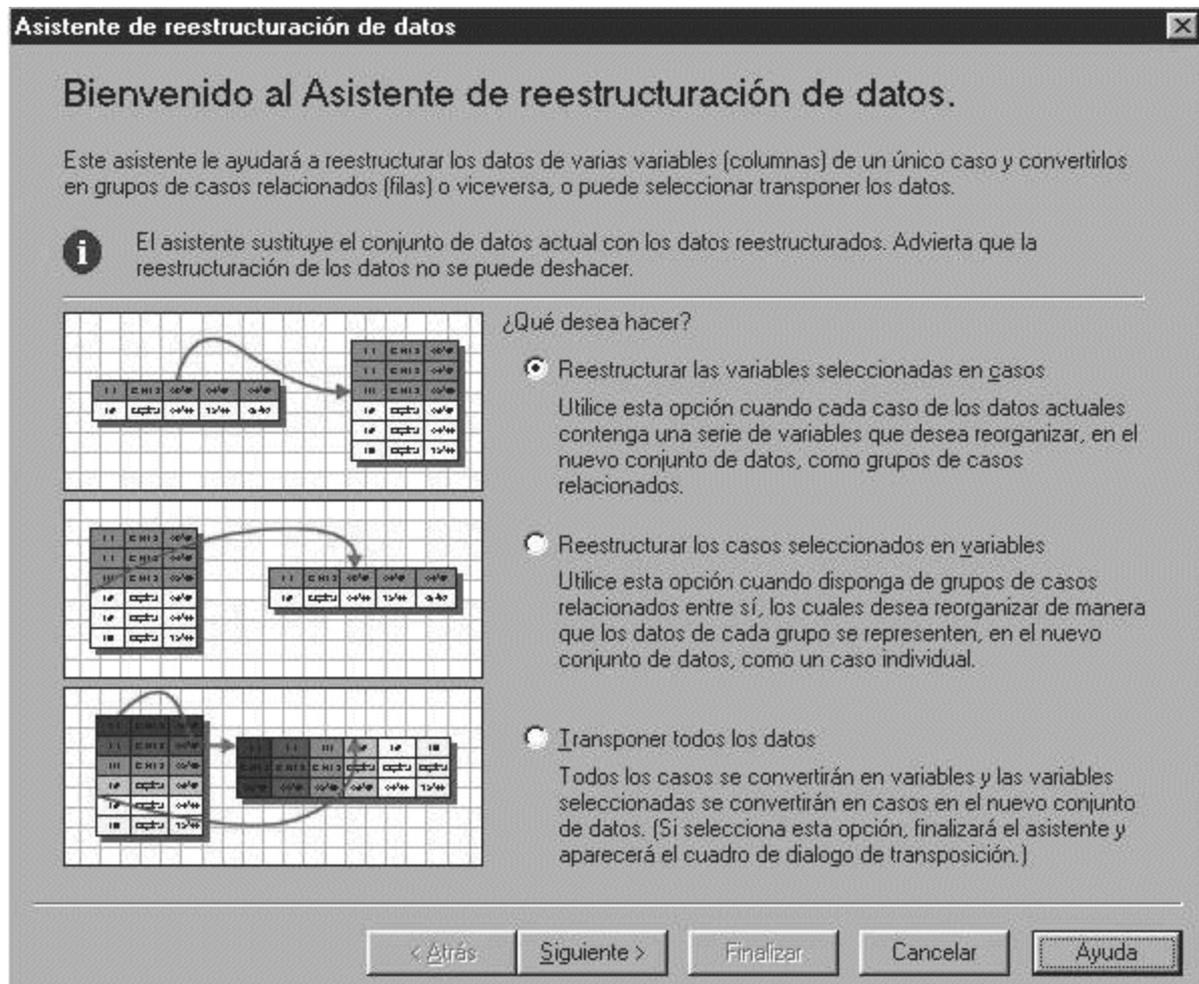
Transponer.

— Seleccionar la variable o variables a transponer y, en su caso, la

variable de nombres.

## 5.8. Reestructurar

SPSS tiene un Asistente de reestructuración de datos para realizar dicha operación. En el cuadro de diálogo que aparece a continuación, se debe seleccionar el tipo de reestructuración a realizar:



**Reestructurar variables seleccionadas en casos** . Se utiliza esta opción cuando se dispone, en los datos, de grupos de columnas relacionadas y se desea que aparezcan en el nuevo archivo de datos como grupos de filas.

Por ejemplo:  
var\_1 var\_2 12 5  
14 2

En este ejemplo, las dos primeras columnas son un grupo de variables que están relacionadas. Contienen datos para la misma variable, var\_1 para el nivel 1 del factor y var\_2 para el nivel 2 del factor. En el análisis de datos de SPSS, si los datos se estructuran de esta manera, se está hablando

de medidas repetidas.

**Reestructurar casos seleccionados en variables** . Esta opción sirve cuando, en los datos, se dispone de grupos de filas relacionadas y se desea que aparezcan en el nuevo archivo de datos como grupos de columnas.

Por ejemplo:

var factor

10 1

12 1

4 2

5 2

En este ejemplo, las dos primeras filas son un grupo de casos que están relacionadas. Contienen datos para el mismo nivel del factor. En el análisis de datos de SPSS, cuando los datos están estructurados de esta manera, se está hablando de un factor como variable de agrupación.

**Transponer todos los datos** . Se selecciona esta opción cuando se desea transponer los datos. Todas las filas se convertirán en columnas y todas las columnas en filas, en el nuevo archivo de datos, según se ha visto anteriormente. Esta opción cierra el Asistente de reestructuración de datos y abre el cuadro de diálogo Transponer datos.

Para reestructurar los datos se debe:

— Elegir en el menú:

Datos.

Reestructurar.

— Seguir las indicaciones del asistente de reestructuración.

### **5.9. Identificar casos duplicados**

Los casos «duplicados» pueden existir por distintos motivos, entre ellos:

— Errores en la entrada de datos: cuando por accidente se introduce el mismo caso más de una vez.

— Casos múltiples que comparten un valor de identificador primario común pero tienen valores diferentes de un identificador secundario, como los alumnos de un colegio que pertenecen a la misma clase.

— Casos múltiples que representan el mismo caso pero con valores diferentes para variables que no sean las que identifican el caso, como en el caso de varias notas del mismo alumno o grupo en diferentes materias o en diferentes momentos.

La identificación de los casos duplicados permite definir prácticamente como se quiera lo que se considera duplicado, y proporciona cierto control

sobre la determinación automática de los casos primarios frente a los duplicados.

Para identificar casos duplicados se debe:

— Elegir en el menú:

Datos.

Identificar casos duplicados.

— A continuación seleccionar:

- **Definir casos coincidentes por.** Los casos se consideran duplicados si sus valores coinciden para todas las variables seleccionadas. Si se desea identificar únicamente aquellos casos que coincidan al 100% en todos los aspectos, habrá que seleccionar todas las variables.

- **Ordenar dentro de los bloques coincidentes por.** Los casos se ordenan automáticamente por las variables que definen los casos coincidentes. Se puede seleccionar otras variables de ordenación que determinarán la secuencia de los casos en cada bloque de coincidencia.

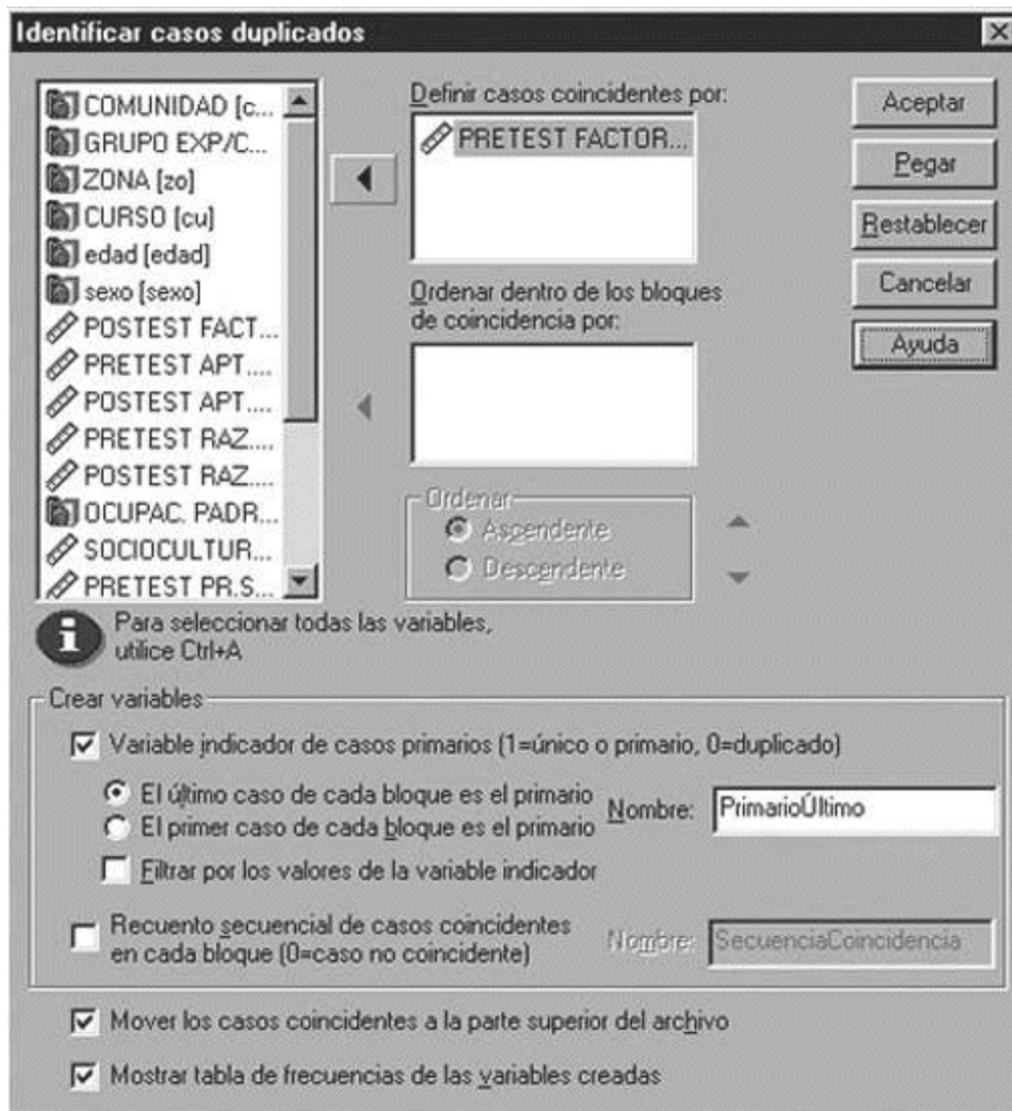
— También se puede crear una variable:

**Variable indicador de casos primarios.** Crea una variable con un valor de 1 para todos los casos únicos y para el caso identificado como caso primario en cada bloque de casos coincidentes y un valor de 0 para los duplicados no primarios de cada bloque. Puede:

- El caso primario ser el primer o el último caso de cada bloque de coincidencia, según determine el orden del bloque de coincidencia. Si no especifica ninguna variable de ordenación, el orden del archivo original determina el orden de los casos dentro de cada bloque.

- Puede utilizar la variable indicador como una variable filtro para excluir los duplicados que no sean primarios de los informes y los análisis sin eliminar dichos casos del archivo de datos.

**Recuento secuencial de casos coincidentes en cada bloque.** Crea una variable con un valor secuencial de 1 a n para los casos de cada bloque de coincidencia. La secuencia se basa en el orden actual de los casos de cada bloque, que puede ser el orden del archivo original o el orden determinado por las variables de ordenación especificadas (ver help de SPSS).



Además se puede:

**Mover los casos coincidentes a la parte superior del archivo** . Para ordenar el archivo de datos de manera que todos los bloques de casos coincidentes estén en la parte superior del archivo de datos.

**Mostrar tabla de frecuencias de las variables creadas.** Las tablas de frecuencias contienen los recuentos de cada valor de las variables creadas.

**Valores perdidos.** En el caso de variables numéricas, los valores perdidos del sistema se tratan como cualquier otro valor. En el caso de variables de cadena, los casos que no tengan ningún valor para una variable de identificación se tratarán como si tuviesen valores coincidentes para dicha variable.

## EJERCICIOS DE AUTOCOMPROBACIÓN

1. Entrar en el programa SPSS abrir un fichero de sintaxis nuevo y

escribir como si se trata del bloc de notas de windows la sintaxis del fichero EJEMPLO.SPS (apartado 3.5.6.1), salvar dicho fichero. Debe observar el lector que la sintaxis llama a un fichero de datos (EJEMPLO.DAT), presente en la página del autor ([www.uned.es/personal/jgil](http://www.uned.es/personal/jgil)), y que deberá copiar en algún subdirectorio de trabajo. En consecuencia, en la sintaxis se debe referenciar la ubicación de dicho fichero de datos. Ir al fichero de sintaxis y ejecutar la misma, para lo cual deberá seleccionar todo el texto y pulsar el icono



(ejecutar comando actual). Salvar el fichero de datos creado como EJEMPLO.SAV

2. Abrir SPSS. Con el fichero de datos EJEMPLO.SAV crear una nueva variable llamada FGPRE1 que tome el valor 1 si FGPRE es menor o igual a 30 y 2 para el resto de valores. Pegar la sintaxis en el programa P2.SPS.

3. Se quiere construir un fichero de datos con 80 valores simulados. Las nuevas variables serán item1, que toma valores entre 20 y 60; item2 con valores entre 40 y 99; y finalmente item3 con valores entre 0 y 99. Suponiendo una distribución uniforme para cada variable. Se quiere construir el fichero de sintaxis, P3.SPS y el fichero de datos, P3.SAV.

4. Las notas finales en Lengua (LG) de 40 estudiantes de una Escuela Superior son las siguientes:

```
11 46 58 25 48 18 41 35 59 28
35 2 37 68 70 31 44 84 64 82
26 42 51 29 59 92 56 5 52 8
1 12 21 6 32 15 67 47 61 47
```

Construir la variable LGQ que tome los valores 1,2,3 y 4 según los cuartiles que se forman con la variable ordenada. Pegar el fichero de sintaxis como P4.SPS

### **SOLUCIÓN A LOS EJERCICIOS DE AUTOCOMPROBACIÓN**

1. Para recoger un fichero de datos en formato texto como EJEMPLO.DAT se puede hacer de dos procesos:

a) importar el fichero de datos EJEMPLO.DAT siguiendo un proceso cuyo resultado, una vez pegada la sintaxis, se salva en un fichero

```
GET DATA /TYPE = TXT
/FILE = 'c:\Ejemplo.dat'
```

```
/DELCASE = LINE
/DELIMITERS = " "
/ARRANGEMENT = DELIMITED
/FIRSTCASE = 1
/IMPORTCASE = ALL
/VARIABLES =
V1 F1.0
V2 F1.0
V3 F1.0
V4 F5.2
V5 F1.0
V6 F4.2
V7 F9.2
V8 F36.2
V9 F32.2
V10 F32.2
V11 F30.2
V12 F11.2
V13 F3.2
V14 F1.0
V15 F1.0
V16 F1.0
V17 F1.0
V18 F1.0
V19 F1.0
V20 F1.0
V21 F1.0
```

.

CACHE.

EXECUTE.

Después será necesario renombrar las variables según los nombres que toman en la investigación

b) escribir directamente la sintaxis según los comandos de lectura de datos (DATA LIST FILE), etiquetado de variables (VARIABLE LABELS) y etiquetados de valores de variables (VALUE LABELS).

TITLE «PROGRAMA DE COMANDOS INICIALES».

DATA LIST FILE=«EJEMPLO.DAT» TABLE

/COM 2 GR 7 ZO 9 CU 15 EDAD 19-20 SEXO 22

FGPRE 23-24 FGPOS 25-26 AVPRE 27-28 AVPOS 29-30 RAPRE 32-33

RAPOS 34-35

OP 37-38 SC 39-40 SOCPRE 41-42 SOCPOS 43-44 VOCPRE 45-46

VOCPOS 47-48

LECPRE 49-50 LECPOS 51-52.

VARIABLE LABELS

COM «COMUNIDAD»

/GR «GRUPO EXP/CONTROL»

/ZO «ZONA»

/CU «CURSO»

/FGPRE «PRETEST FACTOR G»

/FGPOS «POSTEST FACTOR G»

/AVPRE «PRETEST APT.VERBAL»

/AVPOS «POSTEST APT.VERBAL»

/RAPRE «PRETEST RAZ. ABST.»

/RAPOS «POSTEST RAZ. ABST.»

/OP «OCUPAC. PADRES»

/SC «SOCIOCULTURAL»

/SOCPRE «PRETEST PR.SOCIALES»

/SOCPOS «POSTEST PR.SOCIALES»

/VOCPRE «PRETEST PR.VOCAB.»

/VOCPOS «POSTEST PR.VOCAB.»

/LECPRE «PRETEST COMP.LECT.»

/LECPOS «POSTEST COMP.LECT.».

VALUE LABELS

COM

1 «ANDALUCIA»

2 «CANTABRIA»

3 «GALICIA»

/GR

1 «EXPERIMENTAL»

2 «CONTROL»

/ZO

1 «URBANA»

2 «SEMIURBANA»

3 «RURAL»

/CU

6 «SEXTO»

7 «SEPTIMO»

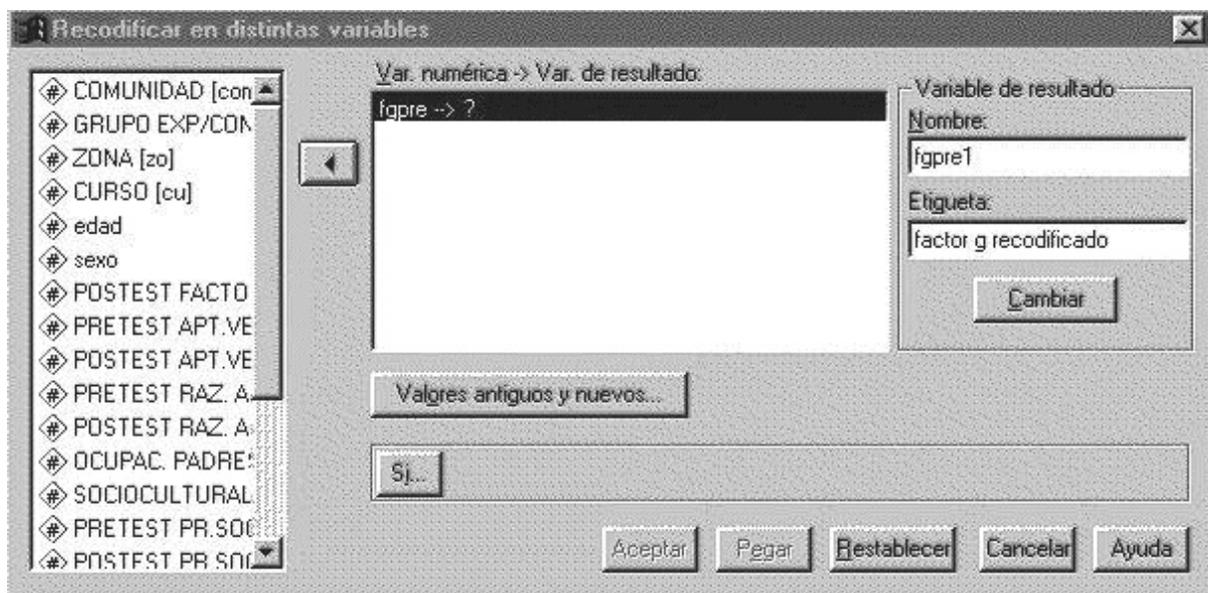
8 «OCTAVO»

/EDAD

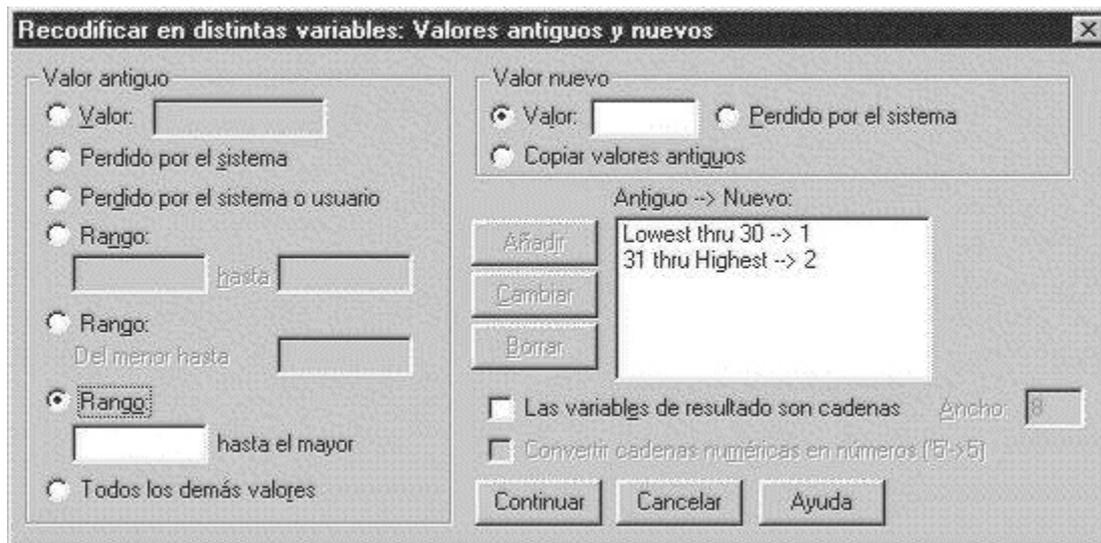
10 «DIEZ AÑOS»  
11 «ONCE AÑOS»  
12 «DOCE AÑOS»  
13 «TRECE AÑOS»  
14 «CATORCE AÑOS»  
15 «QUINCE AÑOS»  
/SEXO  
1 «HOMBRE»  
2 «MUJER»

.  
MISSING VALUE CU(99) SEXO(3).  
SUBTITLE 'LISTA TODAS LAS VARIABLES Y TODOS LOS SUJETOS'. LIST.  
SAVE OUTFILE=«EJEMPLO.SAV»  
/COMPRESO.

2. La secuencia de operaciones necesarias para crear el fichero de sintaxis con la recodificación de los valores de la variable FGPRE y la creación de una nueva variable FGPRE1 con esta recodificación se esquematiza en la siguiente secuencia de pantallas:



Se escribirá el nombre de la nueva variable y su etiqueta. A continuación dar «Cambiar».



Se escribirá los valores antiguos y los nuevos. Después se pulsará «Continuar». La sintaxis de P2.SPS será:

RECODE

fgpre

(Lowest thru 30=1) (31 thru Highest=2) INTO fgpre1 .

VARIABLE LABELS fgpre1 'factor g recodificado'.

EXECUTE.

3. Para crear el fichero de datos, una de las formas más utilizadas, será crear una variable auxiliar, que en nuestro caso hemos denominado ID e introducir cualquier valor, por ejemplo 1, y el número de casos necesarios (en el ejercicio 80 casos). Después se ejecuta la sintaxis:

COMPUTE item1 = RV.UNIFORM(20,60) .

EXECUTE .

COMPUTE item2 = RV.UNIFORM(40,99) .

EXECUTE .

COMPUTE item3 = RV.UNIFORM(0,99) .

EXECUTE .

Y se tendrá como resultado el siguiente fichero de datos:

**id item1 item2 item3**

1 25.59 53.49 15.78

1 37.25 82.85 40.77

1 44.49 92.07 95.91

1 31.63 50.75 71.31

1 26.23 56.50 35.71

1 47.98 81.04 59.49

1 33.85 93.25 65.80  
1 37.83 47.25 32.52  
1 22.10 69.95 43.71  
1 24.13 46.92 52.25  
1 25.65 94.56 2.81  
1 21.72 70.52 49.95  
1 44.87 56.01 3.84  
1 26.14 78.03 3.03  
1 48.61 77.98 4.22  
1 57.13 90.32 88.47  
1 43.13 43.84 64.86  
1 30.47 64.00 .60  
1 48.99 73.44 6.57  
1 21.48 89.75 3.91  
1 24.02 50.92 .35  
1 49.31 54.31 21.68  
1 29.02 51.05 19.05  
1 28.86 98.50 94.17  
1 44.40 56.08 66.78

**id item1 item2 item3**

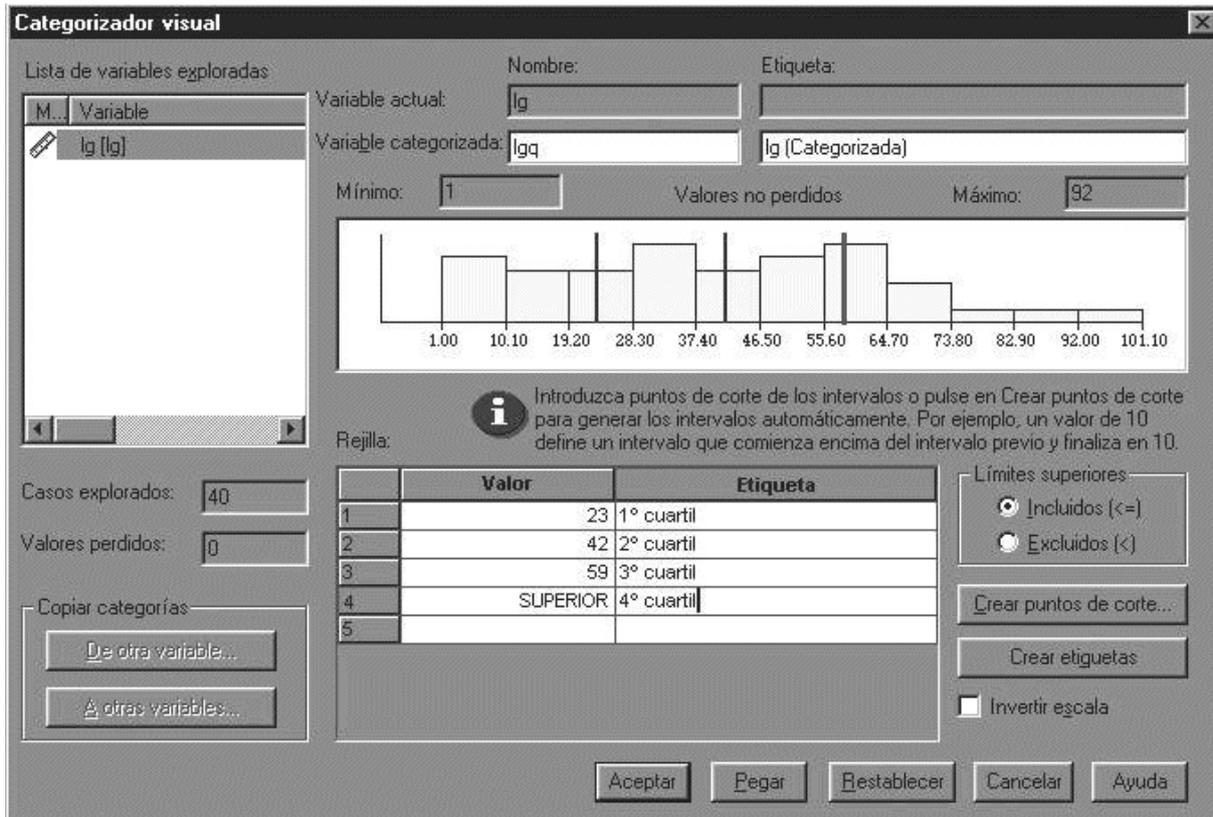
1 33.54 75.47 9.73  
1 22.11 94.88 21.25  
1 49.53 89.34 92.03  
1 24.79 95.88 81.65  
1 22.36 71.78 81.88  
1 27.52 91.47 60.84  
1 23.98 81.73 33.61  
1 47.03 62.69 92.58  
1 47.71 68.56 23.28  
1 54.67 40.86 81.92  
1 34.05 51.21 26.45  
1 39.86 74.33 9.50  
1 48.08 97.60 13.73  
1 28.24 78.97 37.17  
1 50.94 83.81 24.51  
1 22.68 56.99 25.13  
1 41.08 61.86 89.79  
1 46.47 78.70 52.56  
1 57.43 42.31 42.16

1 32.63 87.94 16.55  
1 40.91 73.00 91.91  
1 50.63 75.54 .99  
1 25.08 84.61 41.40  
1 42.37 46.36 95.62  
1 59.30 62.14 43.90  
1 33.01 84.15 86.45  
1 41.42 95.95 74.41  
1 40.31 66.24 76.88  
1 31.28 82.11 43.34  
1 33.89 74.05 18.64  
1 56.84 42.79 24.18  
1 36.33 71.04 14.41  
1 23.83 60.95 69.87  
1 33.36 68.03 39.29  
1 44.83 97.59 80.71  
1 45.65 80.04 18.37  
1 50.09 81.31 93.70  
1 28.20 41.48 85.68  
1 25.73 40.73 47.26  
1 37.84 86.12 62.20  
1 34.76 92.35 76.02  
1 39.41 71.41 64.54  
1 23.97 51.67 17.91  
1 49.03 80.07 1.48  
1 21.49 65.99 72.32

**id item1 item2 item3**

1 46.75 55.84 75.31  
1 50.45 43.69 44.37  
1 52.13 73.03 47.51  
1 48.11 84.81 46.22  
1 21.92 43.55 73.65  
1 35.77 82.32 53.37

Nótese que los valores de cada ítem son aleatorios, en consecuencia, cada vez que se ejecute la sintaxis resultará distinto fichero de datos.  
4. Si se utiliza el categorizador visual de SPSS se puede realizar la división en intervalos que muestra la figura:



La sintaxis será:

\*Visual Bander.

\*lg.

RECODE lg

( MISSING = COPY )

( LO THRU 23 =1)

( LO THRU 41.5 =2 )

( LO THRU 58.5 =3 )

( LO THRU HI = 4 )

( ELSE = SYSMIS ) INTO lgq.

VARIABLE LABELS lgq 'lg (Categorizada)'. FORMAT lgq (F5.0).

VALUE LABELS lgq

1 '1° cuartil'

2 '2° cuartil'

3 '3° cuartil'

4 '4° cuartil'.

MISSING VALUES lgq ( ).

VARIABLE LEVEL lgq ( ORDINAL ). EXECUTE.

## BIBLIOGRAFÍA

ARCE, CONSTANTINO; REAL, EULOGIO (2002): *Introducción al análisis*

*estadístico con SPSS*. PPU. Barcelona.

CAMACHO ROSALES, JUAN (2002): *Estadística con SPSS versión 11 para Windows*. Ra-Ma, Librería y Editorial Microinformática. Madrid.

FERRÁN ARANAZ, MAGDALENA (2002): *Curso de SPSS para Windows*. McGraw-Hill/ Interamericana de España. Madrid.

LIZASOAIN HERNÁNDEZ, LUIS; JOARISTI OLARRIAGA, LUIS (2003): *Gestión y análisis de datos con SPSS*. Thomson Paraninfo. Madrid.

PARDO MERINO, ANTONIO; RUIZ DÍAZ, MIGUEL ÁNGEL (2002): *SPSS 11. Guía para el análisis de datos*. McGraw-Hill/ Interamericana de España. Madrid.

P ÉREZ, CÉSAR (2001): *Técnicas estadísticas con SPSS*. Pearson Educación. Madrid. RÍOS, S. (1974): *Métodos estadísticos*. Ed. del Castillo. Madrid.

SPIEGEL, M. R.(1992):*Estadística*. McGraw-Hill.Madrid.

VISAUTA VINACUA, B. (2002): *Análisis estadístico con SPSS 11.0 para Windows*. McGraw-Hill/ Interamericana de España. Madrid.

## **UNIDAD DIDÁCTICA 2**

### **ANÁLISIS DESCRIPTIVO DE DATOS**

#### **Objetivos**

Describir el comportamiento de datos estadísticos con la ayuda de tablas de frecuencias y gráficos.

Determinar los principales estadísticos de posición y dispersión.  
Determinar los principales estadísticos de simetría y apuntamiento.

Realizar con soltura el proceso de exploración y filtrado de datos estadísticos.

Describir el comportamiento de datos bivariados a partir de tabla de doble entrada.

Determinar los estadísticos de correlación y regresión en datos bivariados.

Utilizar comandos SPSS para realizar los estudios exploratorio, descriptivo y relacional de datos.

#### **1. INTRODUCCIÓN A LA ESTADÍSTICA**

##### **1.1. Introducción**

Cualquier análisis de datos suele iniciarse por una descripción de las variables observadas. Esto suele incluir, principalmente, medidas de tendencia central, variabilidad, asimetría y curtosis. Con estos análisis

cumplimos diversas funciones. En primer lugar nos permite situarnos en la realidad de los datos que poseemos, sus características, descubrir individuos con puntuaciones extremas o situados fuera del rango de la variable («outliers»), descubrir errores en la codificación de las variables, etc. Este proceso descriptivo se conoce con el nombre de «**depuración de los datos**» y es el paso previo a cualquier análisis, pues si existen errores en la matriz de datos es necesaria su corrección para no arrastrar e incrementar los mismos durante los análisis posteriores.

A continuación repasaremos ciertos conceptos de estadística descriptiva y presentaremos los comandos que realizan estos cálculos mediante SPSSWIN.

## **1.2. Conceptos elementales: población, observación, muestra**

Cuando se realizan mediciones un *individuo* (objeto o caso) se caracteriza por las siguientes propiedades:

- Es irreplicable.
- Es observable y, por tanto, objeto de estudio.
- Es semejante (similar, equivalente) a otros respecto a criterios fijados a priori.
- Se comporta de acuerdo con leyes objetivas (es decir, independientes del observador).

**Población** es el conjunto de todos los valores posibles de una característica particular de un grupo especificado de objetos. Tal grupo especificado de objetos se llama *universo*, evidentemente podemos asociar diversas poblaciones a un determinado universo.

Por ejemplo sea el universo formado por el conjunto de individuos cuya altura está comprendida entre 160 y 170 cm; una población asociada a dicho universo será el conjunto de individuos de altura comprendida entre 160 y 170 cm cuya edad sea inferior a 24 años.

Un **experimento** es cualquier proceso (estrategia, mecanismo) que permite asociar a cada individuo de una población un símbolo, o grupo de ellos, numérico o no, de entre todos los símbolos de un conjunto dado a priori (+/-, colores, letras, números, etc.). Son experimentos, p. ej., la determinación de la edad y el sexo de los habitantes de una ciudad; la obtención de la velocidad o el rendimiento de una reacción química; la realización de protocolos de enfermos; la medición de factores ambientales, etc.

La realización concreta de un experimento recibe el nombre de *ensayo*

o prueba, y el símbolo (o los símbolos) obtenido del ensayo o prueba se llama *resultado*, *observación* o *dato*.

El conjunto (colección, colectivo) de símbolos asociables a cada individuo de una población por un experimento se llama *espacio muestral* o *población estadística*, o *suceso seguro*, y lo representaremos por E. Por ejemplo, en la determinación del sexo en los humanos, el espacio muestral está reducido al conjunto macho/hembra o en el lanzamiento de un dado, el espacio muestral limitado por el conjunto de símbolos 1, 2, 3, 4, 5 y 6.

Cada elemento de E recibe cualquiera de estos nombres: *observación*, o *resultado*, o *dato*.

Una **muestra** es una parte del espacio muestral; es, por tanto, una colección de observaciones o datos. Cualquier proceso de obtención de muestras recibe el nombre de *muestreo*. El número de elementos que forman la muestra se llama *tamaño muestral* (n). Sus características se definen por *estadísticos*.

A la muestra se le exige *dimensión* y *representatividad*.

### 1.3. Caracteres, rango y modalidades

Siempre que se estudia una población estamos prestando atención a alguna característica de cada elemento o individuo de la población. A dicha característica la llamaremos *carácter*.

Ej: Si en el lanzamiento de un dado 100 veces observamos cuál es el número obtenido, el carácter en estudio es «resultado obtenido en el lanzamiento de un dado». Si en el grupo de 20 alumnos, preguntamos a cada uno por su talla, el carácter es «la talla».

Si en el mismo grupo se preguntara por el color de los ojos, el carácter en estudio es el «color de ojos».

En los ejemplos anteriores se ha preguntado a cada individuo por características bien diferentes. En los primeros experimentos, ambas son características medibles, la respuesta es un número. La tercera característica, color de ojos, es una cualidad del individuo imposible de medir con un patrón, su respuesta no es un número.

Este ejemplo nos sirve para intuir la clasificación de los caracteres en dos grandes grupos:

a) **Caracteres cuantitativos o variables estadísticas**: son caracteres medibles de la población en estudio. A cada uno de los posibles resultados del experimento se le llamará valor de la variable estadística. Al conjunto

de todos ellos se le llamará *rango* o *recorrido de la variable*.

b) **Caracteres cualitativos o atributos:** son características no medibles (no son numéricas) para los individuos de la población. A cada uno de los posibles resultados del experimento se le llamará *modalidad del atributo*.

#### 1.4. Variable estadística y carácter 1.4.1. Tipos de variables y datos

En la fig. 2.1.1 se recogen las relaciones entre los distintos tipos de variables, que a continuación se desarrollará.

Cuantitativas Acotadas

Discretas

Bernouilli

Binomial Otras

No acotadas Poisson Binomial negativa Otras

UNIVARIANTES

VARIABLES Y

DATOS

Cualitativas Continuas

Ordinales

Nominales Normal

Log-normal

Exponencial Chi-cuadrado t de Student F. de Snedecor Otras

MULTIVARIANTES<sup>Bidimensional</sup> Tridimensional n-dimensional

#### FIGURA 2.1.1. Relación entre distintos tipos de variables.

**Variables cuantitativas** son aquellas que toman valores numéricos: pesos, longitudes, densidades, proporciones, números de nacimientos o de habitantes, etc.

Cuando la variable toma valores numéricos fijos sin valores intermedios se llama *discreta*. Cuando el espacio muestral de una variable discreta contiene una cantidad finita de elementos (por ejemplo, 1/2, 1, 3/2, 2, 7) la variable se llama *acotada*. En caso contrario es *no acotada*.

Las variables discretas se caracterizan por los saltos o interrupciones en su campo de variación. Así, el número de hijos de una pareja es una variable discreta que puede tomar los valores 0,1,2,3,4,... pero ningún

valor entre 0 y 1, entre 1 y 2, etc. Son también discretas las variables: número de hojas de una flor, número de habitantes de una ciudad, etc.

El número de lanzamientos de un dado que hay que realizar antes de obtener por primera vez un 6, es un ejemplo de variable *discreta no acotada*.

Las variables cuyo espacio muestral es todo un intervalo (a,b) se denominan *continuas*. Se caracterizan por tomar «todos» los valores comprendidos entre a y b; es decir, si x es cualquier valor mayor que a y menor que b, hay por lo menos un individuo de la población para el que la variable toma el valor x. Así pues, una variable continua no presenta «huecos» en su campo de variación.

Por extensión se llaman también continuas aquellas variables cuyo espacio muestral es la unión de varios intervalos que no se superponen. Los pesos y las longitudes son variables continuas (si se supone que no existe límite de precisión en el aparato de medida).

**Caracteres cualitativos** son aquellos cuyo espacio muestral es un conjunto de cualidades o atributos (colores, actividades profesionales, presencia o ausencia de un atributo, estado de ánimo, etc).

Cuando es posible ordenar los valores de un carácter cualitativo con algún criterio (raro, frecuente, abundante, por ejemplo), el carácter se llama *ordinal*. En caso contrario, el carácter, también a veces denominada variable diremos que es *nominal* (nombres de especies, clases de enfermedad, etc).

La *escala* elegida y el *grado de precisión* en las observaciones influyen en la calificación de las variables. Así, si el aparato de medida utilizado para medir longitudes de personas sólo aprecia centímetros, la variable se considerará, a efectos prácticos, discreta con valores 0,1,2,...(aunque «potencialmente» sea continua).

De la misma forma una variable cualitativa «color» puede considerarse cuantitativa continua si cada clase de color se caracteriza por su longitud de onda.

Cuando de cada individuo de una población se observan dos o más características, el experimento correspondiente es multivariante. Cada realización del experimento conduce a un *dato múltiple* o conjunto de datos simples. Las variables asociadas a este tipo de experimentos se llaman *multivariantes*.



## NS E CI HE LR N OA

Individuo<sub>Sexo</sub> Edad Cociente<sup>Horas</sup> Lugar de Nota Opinión sobre n.º  
Intelectual<sup>estudio</sup> residencia final la asignatura<sub>semanales</sub>

11 h 19 128 2.64 III 6.49 4  
12 v 20 120 3.08 I 8.30 4  
13 v 19 137 3.79 V 7.38 0  
14 v 20 128 3.42 IV 6.59 3  
15 h 19 130 3.78 II 6.64 3  
16 h 19 118 2.32 IV 5.72 4  
17 h 22 109 2.31 III 7.01 1  
18 v 30 133 2.46 IV 4.43 2  
19 h 20 130 2.30 I 4.46 4  
20 v 29 118 2.68 III 6.15 3  
21 v 23 104 3.37 IV 6.61 4  
22 h 19 128 3.44 II 7.22 3  
23 v 19 123 3.30 III 7.66 1  
24 h 20 117 2.61 IV 6.31 4  
25 v 19 110 2.14 VI 3.90 2  
26 h 19 121 2.45 III 4.50 2  
27 v 19 142 2.33 II 4.76 3  
28 h 20 139 2.50 IV 6.03 3  
29 h 19 187 2.43 IV 8.05 4  
30 v 21 125 2.16 I 5.23 1  
31 v 19 115 2.12 V 5.88 3  
32 h 22 132 2.08 III 5.84 2  
33 h 19 158 2.15 IV 4.45 4  
34 v 19 120 2.21 VI 5.00 3  
35 v 19 155 2.44 III 6.61 4  
36 v 21 106 2.34 V 6.80 2  
37 v 30 135 2.69 IV 7.97 4  
38 h 19 135 3.02 I 5.34 1  
39 h 23 144 2.68 III 5.41 4  
40 v 19 131 3.23 II 6.05 4  
41 h 20 124 3.04 IV 7.85 3  
42 v 20 166 3.17 III 6.98 3  
43 v 19 140 3.32 I 7.96 4  
44 v 24 178 3.15 I 8.04 4  
45 h 19 130 2.91 III 8.09 3  
46 h 25 113 3.12 III 7.26 4

47 h 20 105 3.16 II 6.71 0  
 48 h 20 107 3.32 IV 9.09 4  
 49 h 26 186 5.46 II 4.84 3  
 50 v 19 133 3.11 IV 8.31 2

I:  $\geq 100.000$  habitantes

II: 50.000 a 100.000 habitantes III: 25.000 a 50.000 habitantes IV: 10.000 a 25.000 habitantes V: 5.000 a 10.000 habitantes VI:  $\leq 5.000$  habitantes

Variable OA: 4=excelente; 3=buena; 2=normal; 1=regular; 0=mala

### 2.1.1. Distribución de frecuencias en el caso de caracteres cualitativos

Cuando se ha obtenido una muestra de una población los datos se encuentran, en general:

- Desordenados (no van de menor a mayor).
- Repetidos (dos o más individuos pueden proporcionar los mismos valores para una o más variables).

Es usual utilizar una serie de reglas para *agrupar y resumir* los datos de la muestra.

Supongamos que se tiene una población P con n individuos en los que se estudia el carácter cualitativo C. Para dicho atributo supongamos que existen k modalidades diferentes:  $c_1, c_2, \dots, c_k$ .

Llamamos **frecuencia absoluta** de una modalidad  $c_i$  al número de veces que aparece dicha modalidad en la población de los n individuos. Se tratará evidentemente de un número natural al que denotaremos por  $n_i$ . Al conjunto de las frecuencias absolutas de cada una de estas modalidades que se presenta en la población de los n individuos, es decir, a  $\{n_1, n_2, \dots, n_k\}$  se le denomina **distribución de frecuencias absolutas** del atributo c en la población P.

Esta distribución de frecuencias absolutas verificará, como es obvio, que

$$\sum_{i=1}^k \hat{A}_i = n$$

donde n y k son los números descritos anteriormente.

Esta igualdad tiene una interpretación clara, cada individuo de una población debe presentar una y sólo una modalidad de las k iniciales.

Llamaremos **frecuencia relativa** de una modalidad  $c_i$  al número  $f_i = n_i/n$  donde  $n_i$  y n son los anteriormente descritos.

La definición de  $f_i$  nos indicará la proporción de veces que aparece la modalidad  $c_i$  en la población inicial. Al conjunto de números  $\{f_1, f_2, \dots, f_k\}$  se le denomina distribución de frecuencias relativas del carácter  $c$  en la población  $P$ . Por otra parte si

$k$

$$\sum_{i=1}^k \hat{A}_i = 1$$

$i=1$

### 2.1.2. Distribución de frecuencias en el caso de variable discreta

Sea una variable discreta  $X$  que se estudia en una población  $P$  con  $n$  individuos y que tiene como resultados posibles

$x_1, x_2, \dots, x_k$ .

Las definiciones de distribución de **frecuencia absoluta** y **relativa** para  $X$  son análogas a las dadas para atributos con sólo cambiar modalidad por valor de la variable. Es evidente que estas también verifican las igualdades:

$k$

$$\sum_{i=1}^k \hat{A}_i = 1$$

$i=1$

Además en este caso se van a poder definir dos nuevos tipos de distribuciones de frecuencias:

— **Frecuencia absoluta acumulada** es un valor de la variable,  $x_i$ : es el número de individuos de la población que presenta un valor de la variable menor o igual que  $x_i$ ; lo denotaremos por  $N_i$ . Si suponemos a los  $x_1, x_2, \dots, x_k$  ordenados de menor a mayor, entonces es claro que,

$i$

$$N_i = \sum_{r=1}^i n_r \hat{A}_r$$

$r=1$

Al conjunto de números  $\{N_1, N_2, \dots, N_k\}$  se le denomina, distribución de frecuencias absolutas acumuladas de  $X$  en la población inicial. Por definición se tendrá que  $N_1 = n_1$  y  $N_k = n$ .

— **Frecuencia relativa acumulada** es un valor de la variable,  $x_i$ : es la

proporción de población que presenta un valor de la variable menor o igual que  $x_i$ ; lo notaremos por  $F_i$ . Si nuevamente suponemos que  $x_1, x_2, \dots, x_k$  están ordenados en el orden natural,

$$\hat{A}_i = \sum_{r=1}^i f_r$$

Al conjunto de números  $\{F_1, F_2, \dots, F_k\}$  se le llama, distribución de frecuencias relativas acumuladas de la variable  $X$  en la población  $P$ . Por definición será cierto siempre que  $F_1 = f_1$  y  $F_k = 1$ .

Nota 1. En algunos libros aparece el concepto de frecuencia acumulada en un  $x_i$  como la proporción de individuos que presentan un valor de la variable menor estrictamente que  $x_i$ . Si no se advierte lo contrario, nosotros siempre utilizaremos la definición inicial (la del  $\leq$ ).

Nota 2. Está claro que hablar de frecuencias acumuladas en el caso de atributos carece de todo sentido por la propia definición de frecuencia acumulada.

### 2.1.3. Distribución de frecuencias en el caso de variable continua

Recordemos que en el caso de variable continua, el rango o recorrido de la variable es todo un intervalo de  $\mathbf{R}$  (conjunto de los números reales), y por tanto, los posibles resultados del experimento es un conjunto infinito no numerable. Para poder presentar, de forma compacta, la información recogida para una variable de estas características, nos hace falta el concepto de «clase».

**Definición 1:** Llamaremos partición de las infinitas modalidades, de una variable continua, a un número finito de intervalos, de tal forma que cubran todas las posibilidades de la variable, y que dichos intervalos no se solapen (tengan intersección vacía). A estos intervalos les denominaremos *clases*.

Ej: En el experimento de la talla, el rango de la variable era el intervalo  $[1; 2]$ , damos a continuación dos posibles particiones:

- a)  $[1; 1,2] (1,2; 1,4] (1,4; 1,6] (1,6; 1,8] (1,8; 2]$
- b)  $[1; 1,5] (1,5; 1,6] (1,6; 2]$

Generalmente estas clases no vendrán dadas con los datos del problema. Cuando no sea así, será la lógica y la intuición las que nos digan

cómo tomarlas. En el texto, casi todas las particiones serán de intervalos abiertos inferiormente y cerrados superiormente; la notación general para una partición será  $[l_0, l_1] (l_1, l_2], \dots, (l_{k-1}, l_k]$ , y se supondrán ordenadas, es decir  $l_0 \leq l_1 \leq \dots \leq l_k$ .

**Definición 2:** Llamaremos *extremos inferior y superior de la clase*  $(l_{i-1}, l_i]$  a los números  $l_{i-1}$  y  $l_i$  respectivamente.

**Definición 3:** Llamaremos *marca de clase i-ésima*  $(l_{i-1}, l_i]$  al número  $c_i$  donde

$c_i = \frac{l_i + l_{i-1}}{2}$   
 iii  
 2

es decir, el punto medio de dicho intervalo. Aún cuando  $x_i$  usualmente representa la modalidad i-ésima, para datos agrupados en intervalos, denota la marca de clase o punto medio, es decir son equivalentes  $c_i$  y  $x_i$ .

**Definición 4:** Llamaremos *amplitud de la clase i-ésima*  $(l_{i-1}, l_i]$  al número  $a_i = l_i - l_{i-1}$ , es decir, la longitud del segmento que ocupa dicha clase en . En una partición no todas las clases tienen que tener la misma amplitud como se observa en el ejemplo de la talla.

Una vez enfocado el estudio de las variables continuas, estamos en condiciones de definir las distribuciones de frecuencias para una de estas variables. Sea  $X$  una variable continua, con sus correspondientes clases, que se estudia en una población  $P$ .

**Frecuencia absoluta** de la clase i-ésima  $(l_{i-1}, l_i]$ : es el número de individuos que presentan un valor de la variable mayor que  $l_{i-1}$  pero menor o igual que  $l_i$ . A dicho número se le nota por  $n_i$ . Al conjunto  $\{n_1, n_2, \dots, n_k\}$  se le denomina distribución de frecuencias absolutas de  $X$  en la población  $P$ . Sigue siendo cierto que la suma de todas las frecuencias absolutas es  $n$ .

**Frecuencia relativa** de la clase i-ésima  $(l_{i-1}, l_i]$ : se denomina así, a la proporción de población que presenta un valor de la variable mayor que  $l_{i-1}$  y menor o igual que  $l_i$ ; la notaremos por  $f_i$ . Por lo anteriormente definido  $f_i = n_i/n$ . Al conjunto de números  $\{f_1, f_2, \dots, f_k\}$  se la denomina distribución de frecuencias relativas de la variable  $X$  en la población  $P$ . Dicha distribución de frecuencias relativas deberá verificar, para ser tal, que la suma de todas las frecuencias relativas sea 1.

**Frecuencia absoluta acumulada** hasta el extremo de clase  $l_i$ : llamaremos así al número de individuos que presenta un valor de la variable menor o igual que  $l_i$ ; los denotaremos por  $N_i$ . Por definición de  $n_i$  y por cómo se han tomado las clases,

$$N_i = \sum_{r=1}^i n_r$$

Por tanto, dada la distribución de frecuencias absolutas de una variable continua, podemos dar la distribución de frecuencias absolutas acumuladas en los extremos de clase. Podríamos preguntarnos por la frecuencia absoluta acumulada de un valor de la variable que no sea extremo de clase, pero ese problema lo resolveremos más adelante.

A los números  $\{N_1, N_2, \dots, N_k\}$  se le denomina, distribución de frecuencias absolutas acumuladas de  $X$  en la población inicial.

**Frecuencia relativa acumulada** en  $l_i$ : es la proporción de la población que presenta un valor de la variable menor o igual que  $l_i$  lo notaremos por  $F_i$ . Por la definición anterior

$$F_i = \frac{N_i}{n} = \frac{\sum_{r=1}^i n_r}{n}$$

A  $\{F_1, F_2, \dots, F_k\}$  se le denomina distribución de frecuencias relativas acumuladas de  $x$  en  $P$ .

## 2.2. Tablas estadísticas

Toda la información referente a un carácter, y las distribuciones de frecuencias de éste, en una población en estudio, suele venir expresada en forma de tabla del siguiente modo:

a) **Para carácter cualitativo:** la tabla se construye escribiendo, en primer lugar, una columna correspondiente a todas las modalidades posibles. A continuación, otra columna con la distribución de frecuencias

absolutas, y una tercera con la distribución de frecuencias relativas, de manera que por fila vayamos leyendo cada modalidad, su frecuencia absoluta, y su frecuencia relativa. La notación general para esta tabla será:

$$\begin{array}{ccc}
 c & n_i & f_i \\
 c_1 & n_1 & f_1 \\
 c_2 & n_2 & f_2 \\
 \dots & \dots & \dots \\
 c_k & n_k & f_k
 \end{array}$$

Ej: Supongamos que se pregunta por el color de ojos de 20 alumnos y que las respuestas se resumen en:

- 4 alumnos con ojos azules.
- 7 alumnos con ojos negros.
- 8 alumnos con ojos marrones.
- 1 alumno con ojos verdes.

La tabla para este carácter en la población mencionada sería entonces:

TABLA 2.2.2. Para carácter cualitativo.

Color de ojos	$n_i$	$f_i$
Azules	4	4/20
Negros	7	7/20
Marrones	8	8/20
Verdes	1	1/20

b) **Para variables estadísticas discretas:** la tabla es análoga a la de un atributo, pero en lugar de aparecer las modalidades, tendríamos que especificar los valores de la variable. Además, en el caso de variables estadísticas, se dispone de un par de distribuciones de frecuencias acumuladas que se añadirán a la tabla. Esta queda entonces como sigue:

$$\begin{array}{ccccccc}
 \mathbf{X} & \mathbf{n_i} & \mathbf{f_i} & \mathbf{N_i} & \mathbf{F_i} & \mathbf{x_1} & \mathbf{n_1} & \mathbf{f_1} & \mathbf{N_1} & \mathbf{F_1} \\
 \mathbf{x_2} & \mathbf{n_2} & \mathbf{f_2} & \mathbf{N_2} & \mathbf{F_2} & & & & & \\
 \dots & \dots & \dots & \dots & \dots & \mathbf{x_k} & \mathbf{n_k} & \mathbf{f_k} & \mathbf{N_k} & \mathbf{F_k}
 \end{array}$$

Ej: Supongamos que en el lanzamiento 20 veces de un dado, los resultados han sido los siguientes:

- el 1 aparece 3 veces
- el 2 aparece 5 veces
- el 3 aparece 5 veces
- el 4 no aparece
- el 5 aparece 4 veces

— el 6 aparece 3 veces

Esta información daría lugar a la siguiente tabla:

TABLA 2.2.3. Variables estadísticas discretas.

<b>X</b>	<b>n<sub>i</sub></b>	<b>f<sub>i</sub></b>	<b>N<sub>i</sub></b>	<b>F<sub>i</sub></b>
1	3	3/20	3	3/20
2	5	5/20	8	8/20
3	5	5/20	13	13/20
4	0	0	13	13/20
5	4	4/20	17	17/20
6	3	3/20	20	1

c) **Para variables estadísticas continuas:** la única novedad que presentan estas tablas con respecto a las variables discretas es que, en la columna de los valores de la variable lo que aparecen son sus clases, pudiendo aparecer además, una columna que da las marcas de clase y otra que da las amplitudes de clase:

**X<sup>Marca</sup> Amplitud n<sub>i</sub> f<sub>i</sub> N<sub>i</sub> F<sub>i</sub> de clase**

$[l_0 \ l_1]$	$x_1$	$a_1$	$n_1$	$f_1$	$N_1$	$F_1$
$(l_1 \ l_2]$	$x_2$	$a_2$	$n_2$	$f_2$	$N_2$	$F_2$
.....						
$(l_{k-1} \ l_k]$	$x_k$	$a_k$	$n_k$	$f_k$	$N_k$	1

Ej: Supongamos se les ha preguntado por su talla a un grupo de 20 alumnos. Las clases para la variable son:

$[1 \ 1,25]$   $(1,25 \ 1,50]$   $(1,50 \ 1,60]$   $(1,60 \ 1,70]$   $(1,70 \ 1,75]$  Los 20 alumnos contestan con sus tallas y se cuenta el número de ellos que se quedan en cada clase, resultando que:

- hay 1 individuo que mide entre 1 y 1,25
- hay 2 individuos que miden entre 1,25 y 1,50
- hay 3 individuos que miden entre 1,50 y 1,60
- hay 7 individuos que miden entre 1,60 y 1,70
- hay 7 individuos que miden entre 1,70 y 1,75

La tabla que construiríamos para esta variable sería la siguiente:

TABLA 2.2.4. Variables estadísticas continuas.

<b>X</b>	<b>x<sub>i</sub></b>	<b>a<sub>i</sub></b>	<b>n<sub>i</sub></b>	<b>f<sub>i</sub></b>	<b>N<sub>i</sub></b>	<b>F<sub>i</sub></b>
$[1 \ 1,25]$	1,125	0,25	1	1/20	1	1/20
$(1,25 \ 1,50]$	1,375	0,25	2	2/20	3	3/20
$(1,50 \ 1,60]$	1,55	0,1	3	3/20	6	6/20

(1,60 1,70] 1,65 0,1 7 7/20 13 13/20 (1,70 1,75] 1,725 0,05 7 7/20 20 1

### 3. REPRESENTACIONES GRÁFICAS

#### 3.1. Introducción

Trataremos ahora de representar los datos de una tabla de frecuencias a través de una gráfica, sin que por ello se pierda información alguna. Es decir, dada la tabla de frecuencias de un carácter, podemos dibujar la gráfica y viceversa, dada cualquier gráfica vamos a ser capaces de escribir la tabla de frecuencias que representa.

#### 3.2. Caracteres cualitativos

Sea  $C$  un carácter cualitativo que se estudia en una población  $P$ . Supongamos que este estudio nos da la tabla de frecuencias 2.3.1. Para representar gráficamente esta información caben varias posibilidades:

— **Diagrama de sectores:** En un círculo, cada modalidad ocupará un sector circular con ángulo proporcional a la frecuencia absoluta de dicha modalidad. Para dibujarlo, nos planteamos con cada modalidad una simple regla de tres, esta es, si  $n$  individuos van a repartirse  $360^\circ$ ,  $n_i$  individuos de la modalidad  $c_i$  ocuparán  $x_i$  grados.

A la vista de un diagrama de sectores y conociendo  $n$ , es fácil escribir la tabla que presenta.

— **Diagrama de barras:** Se dibujan unos ejes de coordenadas. En el eje de abscisas se representarán las modalidades mediante intervalos separados, y todos de la misma amplitud. Sobre ellos construiremos rectángulos de altura proporcional a la frecuencia de cada modalidad; para la modalidad  $i$ -ésima  $h_i = c \cdot f_i$ . La constante de proporcionalidad  $c$ , se puede elegir de manera que las alturas  $h_i$  sean fáciles de dibujar, y siempre que  $c$  sea la misma para todas las modalidades.

Para una gráfica de este tipo, conocer la distribución de frecuencias de la que proviene es sencillo sólo con notar que

$$f_i = \frac{h_i}{c} \text{ y que } c = \frac{\sum_{i=1}^k h_i}{n}$$

— **Pictogramas:** La idea es la misma que la del diagrama de barras pero dibujándole a cada rectángulo una figura representativa de la modalidad. Ej: Dada la variable cualitativa «color de ojos», que presenta la tabla 2.3.1,

vamos a construir las gráficas anteriores para esta distribución.

TABLA 2.3.1

**Color de ojos  $n_i$   $f_i$**  Azules 4 4/20

Negros 7 7/20

Marrones 8 8/20

Verdes 1 1/20

• Diagrama de sectores:

— Para el color de ojos azules, corresponderá un sector de ángulo  $a_1$ , para calcularlo

$$360^\circ \frac{4}{20}$$

$$a_1 = \frac{4}{20} \cdot 360^\circ = 72^\circ$$

— Para el color de ojos negros, el ángulo lo notaremos por  $a_2$ , donde  $360^\circ$

$$\frac{7}{20} \cdot 360^\circ$$

$$a_2 = \frac{7}{20} \cdot 360^\circ = 126^\circ$$

$$20$$

— Para el color de ojos marrones, el ángulo lo notaremos por  $a_3$ ,

donde  $360^\circ \frac{8}{20}$

$$a_3 = \frac{8}{20} \cdot 360^\circ = 144^\circ$$

— Para el color de ojos verdes, el ángulo lo notaremos por  $a_4$ , donde

$$360^\circ \frac{1}{20}$$

$$\frac{1}{20} \cdot 360^\circ$$

$$a_4 = \frac{1}{20} \cdot 360^\circ = 18^\circ$$

$$20$$

El diagrama de sectores quedaría entonces como sigue.

**Color de ojos**

Verdes 18% Azules 72%

Marrones 144%

Negros 126%

• Diagrama de barras:

— Para el color de ojos azules, corresponderá un rectángulo de altura  $h_1$  tal que

— Para el color de ojos negros,  $h_2$  con  $h_2 = c \cdot f_2 = c \cdot 7/20$

— Para el color de ojos marrones,  $h_3$   $h_3 = c \cdot f_3 = c \cdot 8/20$

— Para el color de ojos verdes,  $h_4$

$$h_4 = c \cdot f_4 = c \cdot 1/20$$

Como constante de proporcionalidad podemos utilizar  $c = 20$ . Con ella el diagrama de barras queda como:

### Color de ojos

8  
6  
4  
2  
0

Azul Negro Marrón Verde

- **Pictograma:** Para construir este diagrama podríamos aprovechar el diagrama anterior colocando encima de cada rectángulo un dibujo. En este ejemplo parece lo más apropiado una figura de un ojo del color correspondiente a cada modalidad.

### 3.3. Caracteres cuantitativos

#### 3.3.1. Gráficas de variables discretas

Para estos caracteres existen dos tipos de gráficas:

— **Diagrama diferencial o diagrama de barras (utiliza las frecuencias relativas):** Se dibujará sobre unos ejes de coordenadas. En el eje de abscisas colocaremos los valores de la variable, y sobre cada uno de ellos, levantaremos una línea recta de altura proporcional a la frecuencia relativa de cada valor. Así, para  $x_i$ , la línea recta tendrá altura  $h_i = c \cdot f_i$ . Como antes, la constante  $c$  será la misma para todos los valores. Dado el diagrama de barras de una variable, es fácil escribir la tabla de frecuencias que está representando.

— **Diagrama integral o función de distribución (utiliza las frecuencias acumuladas).** Como su nombre indica se trata de la representación de la función de distribución  $F(x)$ . Por las características de  $F$  sabemos que se tratará de una función escalonada, determinada por los valores de la variable y las frecuencias con que los tome.

Ej: Se lanza 20 veces un dado y se obtiene la tabla 2.3.2, vamos a construir para ella los dos diagramas anteriores:

TABLA 2.3.2

$X$	$n_i$	$f_i$	$N_i$	$F_i$
1	3	3/20	3	3/20
2	5	5/20	8	8/20
3	5	5/20	13	13/20

4 0 0 13 13/20  
 5 4 4/20 17 17/20  
 6 3 3/20 20 1

• Diagrama diferencial o diagrama de barras:

— Para el primer valor de la variable, el 1, tendremos que levantar una línea recta de altura  $h_1$ , con

$$h_1 = c \cdot f_1 = c \cdot 3/20$$

— Para el segundo valor de la variable, el 2, esta altura sería  $h_2 = c \cdot f_2 = c \cdot 5/20$

— Para el 3,

$$h_3 = c \cdot f_3 = c \cdot 5/20$$

— Para el 4,

$$h_4 = c \cdot f_4 = c \cdot 0/20$$

— Para el 5,

— Para el 6,  $h_6 = c \cdot f_6 = c \cdot 3/20$

A la vista de las alturas que queremos levantar, podríamos utilizar como constante de proporcionalidad  $c = 40$ , y el diagrama quedaría

**20 lanzamientos de un dado 12**

10  
 8  
 6  
 4  
 2  
 0 123456

• Diagrama integral o función de distribución:

Podemos construir la función de distribución para este ejemplo que será:

0  $F_3$   
 20 12  
 8  
 20 23  
 13  
 $F(x) =$  20 34  
 13  
 20 45  
 17  
 20 56

16 Su gráfica será entonces:

## 20 lanzamiento de un dado 1,2

1

0,8

0,6

0,4

0,2

0  $x < 1$   $1 < x \leq 2$   $2 < x \leq 3$   $3 < x \leq 4$   $4 < x \leq 5$   $5 < x \leq 6$   $x > 6$

### 3.3.2. Gráficas para variables continuas

También en este caso se podrán dibujar dos tipos de gráficas.

— **Diagrama diferencial o histograma (utilizando frecuencias relativas).** Sobre unos ejes de coordenadas, dibujaremos en el eje de abscisas las clases de la variable. Sobre cada clase levantaremos un rectángulo de superficie proporcional a la frecuencia relativa de la clase. Dada la clase  $(l_{i-1}, l_i]$ , un rectángulo de altura  $h_i$ , tendría superficie  $s_i = h_i \cdot a_i$ , donde  $a_i$  es la amplitud de esta clase. Si queremos que esta superficie sea proporcional a la frecuencia relativa  $f_i$ , tendría que ser cierta la igualdad  $h_i \cdot a_i = c \cdot f_i$ . Y por tanto la altura que debemos ponerle al rectángulo vendrá dada por

$$h_i = c \cdot f_i / a_i$$

Como siempre la constante de proporcionalidad  $c$ , puede ser cualquier número, pero el mismo para todas las clases.

Dado un histograma, si queremos construir la tabla de la variable que representa, bastará notar que la suma de todas las superficies es:

$\sum_{i=1}^k$

$\hat{A}$

$\sum_{i=1}^k h_i a_i$

$= \sum_{i=1}^k c f_i$

$$\hat{A} = \sum_{i=1}^k h_i a_i = c \sum_{i=1}^k f_i = c$$

$\sum_{i=1}^k$

y conocida  $c$ , podemos calcular las frecuencias relativas con la expresión:

$f_i$

$h_i a_i$

$=$

$\sum_{i=1}^k h_i a_i = c$

— **Diagrama integral o función de distribución (utiliza las frecuencias acumuladas).** Como hemos dicho anteriormente, encontrar el

valor exacto de la función de distribución para cada  $x_j$ . Para dibujar la gráfica uniremos, mediante un trazo regular, los puntos  $(x_j, F(x_j))$ . Por definición se le llama función de distribución a la curva monótona creciente, continua a la derecha y regular, que une a estos puntos. En la práctica, la curva que se utiliza es la poligonal que une los puntos  $(x_j, F(x_j))$ .

De otro lado, dada la poligonal, función de distribución de alguna variable, es fácil determinar las clases de la variable y sus frecuencias relativas correspondientes. Ej: Dada la tabla 2.3.3, perteneciente a la talla de 20 alumnos, vamos a construir los diagramas anteriores para la misma

TABLA 2.3.3. Variables estadísticas continuas.

$X$	$x_j$	$a_j$	$n_j$	$f_j$	$N_j$	$F_j$
[1, 1,25]	1,125	0,25	1	1/20	1	1/20
(1,25, 1,50]	1,375	0,25	2	2/20	3	3/20
(1,50, 1,60]	1,55	0,1	3	3/20	6	6/20
(1,60, 1,70]	1,65	0,1	7	7/20	13	13/20
(1,70, 1,75]	1,725	0,05	7	7/20	20	1

• Diagrama diferencial o histograma:

— Sobre la clase [1, 1,25], tendremos que levantar un rectángulo de altura  $h_1$ , tal que su área  $s_1 = c \cdot f_1$ , es decir,  $0,25 \cdot h_1 = c \cdot f_1$  fi  $h_1 = c \cdot f_1 / 0,25 = c \cdot 1 / (20 \cdot 0,25) = c \cdot 1/5$

— Sobre la clase (1,25, 1,5], la altura  $h_2$  con  $0,25 \cdot h_2 = c \cdot f_2$  fi  $h_2 = c \cdot f_2 / 0,25 = c \cdot 2 / (20 \cdot 0,25) = c \cdot 2/5$

— Sobre la clase (1,5, 1,6], la altura  $h_3$  con  $0,1 \cdot h_3 = c \cdot f_3$  fi  $h_3 = c \cdot 3/2$

— Sobre la clase (1,6, 1,7], la altura  $h_4$  con  $h_4 = c \cdot 7/2$

— Sobre la clase (1,7, 1,75], la altura  $h_5$  con  $h_5 = c \cdot 7$

Si usamos la constante de proporcionalidad  $c = 1$ , el histograma quedaría:

**Variable altura**

7
6
5
4
3
2
1
0
[1, 1,25] [1,25, 1,5] [1,5, 1,6] [1,6, 1,7] [1,7, 1,75]

• Diagrama integral o función de distribución:

Para dibujar esta gráfica disponemos de cinco puntos pertenecientes a la misma:

(1,25 1/20) (1,5 3/20) (1,6 6/20) (1,7 13/20) (1,75 1) Si dibujamos estos puntos en unos ejes de coordenadas, y los unimos a través de una curva regular, la gráfica resultante sería:

Variable altura 1,2

1

0,8

0,6

0,4

0,2

0

[1 1,25] [1,25 1,5] [1,5 1,6] [1,6 1,7] [1,7 1,75]

### 3.3.3. Gráfico de tallos y hojas

Un procedimiento semi-gráfico de presentar la información para variables cuantitativas, que es especialmente útil cuando el número total de datos es pequeño (menor que 50), es el diagrama de tallo y hojas de Tukey.

Supongamos que tratamos con variables cuantitativas discretas o continuas y que el número de datos no es muy grande (menor que 50). Si la variable es cuantitativa continua en primer lugar la agrupamos en clase.

Decidir el número de clases depende del investigador, una regla utilizada es elegir el número entero más próximo a  $2\sqrt{n}$ . Sin embargo puede ser cualquier número comprendido entre 5 y 20.

Una vez que tenemos agrupados los valores los pasos a seguir son los siguientes: a) Redondear los datos con dos o tres cifras significativas de forma que la clasificación sea lo más sencilla posible.

b) Dibujar el diagrama de tallos y hojas. Para ello primero dibujamos los tallos y después separadas por una línea vertical dibujamos las hojas. Los tallos son todos los dígitos de la marca de clase, excepto las unidades. Las hojas son las unidades de todos los datos y se representan en la línea correspondiente al tallo donde está la marca de clase a la que corresponde cada dato.

Veamos el siguiente ejemplo, supongamos los siguientes datos recogidos:

45 57 59 58 61 56 63 57 69  
59 59 57 63 72 59 62 60 41  
59 63 59 65 57 61 58 59 63

El primer problema que debemos abordar es calcular cuantas clase vamos a construir, según la información precedente en este caso serán:  $27=10,39^{*}10$ , no obstante sólo vamos a crear 9 clases.

A continuación formaremos las clases y sus correspondientes marcar de clase:

**Intervalo Centro del intervalo: marca de clase**

- 40-43 41,5
- 44-47 45,5
- 48-51 49,5
- 52-55 53,5
- 56-59 57,5
- 60-63 61,5
- 64-67 65,5
- 68-71 69,5
- 72-75 73,5

Ahora se redondearía los datos para que el número de cifras significativas fuese dos o tres, pero en nuestro ejemplo esto no es necesario ya que son números de dos cifras. Por último dibujaremos el gráfico de tallos y hojas.

```

    41
45
4
5
5 7 9 8 6 7 9 9 7 9 9 9 7 8 9
6 1 3 3 2 0 3 1 3
65
69
72
```

decenas unidades

Observación: cuanto menor sea el número de clases menor será la apreciación de los valores extremos.

Cuando el primer dígito de la clasificación varía poco, la mayoría de los datos tienden a agruparse alrededor de un tallo y el diagrama resultante tiene poco detalle. En este caso es conveniente subdividir cada tallo en dos o más partes introduciendo algún signo arbitrario.

Una de las ventajas del gráfico de tallos y hojas es que hace los papeles de tabla de frecuencias agrupadas e histograma al mismo tiempo.

Otra ventaja de estos gráficos es que los valores de la variable son conservados, cosa que no ocurre con la distribución de frecuencias agrupadas. Esta ventaja permite realizar cálculos con los valores de la variable sobre el mismo gráfico.

## 4. MEDIDAS DE POSICIÓN

### 4.1. Introducción

Hasta ahora el tratamiento dado a un carácter ha sido el de resumir la información (en forma de tablas) y presentarla gráficamente. A la vista de dichas gráficas, podíamos hacernos una idea cualitativa de lo que estaba pasando con el carácter.

Centrémonos, para este apartado, en variables estadísticas, es decir, caracteres cuantitativos. Para estas variables nos interesa un resumen numérico de lo que ocurre con ellas en la población. Con este objetivo, vamos a sustituir el conjunto total de números de la tabla por otros llamados estadísticos o características estadísticas. Estos estadísticos, según su función, se clasifican en tres grupos:

— En primer lugar están los de **tendencia central**: van a indicarnos un valor

entorno al cual se reparten los valores que toma la variable, para la población en estudio.

— En segundo lugar los de **dispersión**: medirán el grado de dispersión que hay entre los valores que toma la variable realmente y un valor central.

— En tercer lugar están los que hablarán de la **forma** en que se presentan la dispersión anterior. Son los llamados estadísticos de forma.

Conviene también recordar dos aspectos terminológicos:

1. Un **estadístico** es una medida descriptiva calculada de los datos de la muestra.

2. Un **parámetro** es una medida descriptiva calculada de todos los datos de la población.

### 4.2. Medidas de tendencia central

La posición o tendencia central de una distribución hace referencia al lugar donde se centra un conjunto de valores.

Existe diferente criterio para definir esta tendencia central. Nos detendremos en las siguientes:

- la media aritmética
- la media geométrica, armónica y cuadrática
- la mediana

— la moda

No se puede afirmar que una de ellas sea mejor que las demás. Simplemente, habrá casos en los que una medida de tendencia central sea más adecuada que otra. Lo que si podemos decir es que la media aritmética es la más utilizada, y por ello será la más estudiada.

#### 4.2.1. Media aritmética

**Definición de media aritmética:** Dada la variable X que toma valores  $x_i$  con frecuencias relativas  $f_i$ , se define su media como

$$\bar{X} = \sum_{i=1}^k f_i x_i$$

Si expresamos la frecuencia relativa en términos de frecuencia absoluta, la expresión anterior resulta

$$\bar{X} = \frac{\sum_{i=1}^k n_i x_i}{n}$$

$$\bar{X} = \frac{\sum_{i=1}^k x_i n_i}{n}$$

$$\bar{X} = \frac{\sum_{i=1}^k x_i n_i}{n}$$

$$\bar{X} = \frac{\sum_{i=1}^k x_i n_i}{n}$$

y esta es la fórmula que intuitivamente encontrábamos adecuada.

En la definición anterior hemos hablado de valores de la variable y sus frecuencias relativas. No olvidemos que también son variables estadísticas las variables continuas, ¿cómo entender para ellas la expresión de la media? La respuesta es simple, en lugar de utilizar valores de la variable, utilizaremos las marcas de clase. Como frecuencia relativa para cada marca de clase, utilizaremos la frecuencia relativa de la clase a la que pertenece.

Ej: Con los datos de la tabla 2.2.3 del lanzamiento de un dado 20 veces, la media sería

3 5 5 4 3

$x = + + + + = 1. \dots$ ,  
 20 20 20 40 5 20 20 33

Con la tabla 2.2.4 de la talla de 20 alumnos, la media será:

1 2 3 7 7

$x = + + + + 20 155 20 165, \dots, \dots, \dots^a, 161 20 20 20$

El cálculo de la media a través de tablas de datos agrupados lleva implícito una *perdida de precisión* respecto del cálculo directo a partir de la muestra en bruto, ya que valores distintos en la muestra se identifican con un único representante (o marca de clase)  $x_i$  de la clase  $A_i$ .

**Propiedades de la media aritmética:** Con la definición de media aritmética para una variable estadística  $X$ , que acabamos de ver, se verifican las siguientes propiedades:

i) La media ponderada de las diferencias de la variable a su media es cero

$$\sum_{i=1}^k \hat{A}_i$$

$$f_x X() = 0$$

ii) Si definimos la desviación de la variable  $X$  a un número real «a» como el resultado de la operación

$\sum_{i=1}^k \hat{A}_i f_x a$

$$ii ()$$

$i=1$

—

esta desviación alcanza su mínimo cuando hacemos  $a = X$ .

iii) Si dada la variable  $X$ , le aplicamos una transformación lineal para pasar a una variable  $Y$ , es decir,  $Y = (a+X)/b$ , la media de la variable  $Y$  resulta ser

—

$$- Y aX = b$$

Esta tercera propiedad nos permite calcular la media de una variable directamente, o bien a través de la media aritmética de cualquier transformación lineal de la variable inicial. Este hecho resulta

especialmente interesante, en el caso de medias para variables continuas con amplitud de clase constantes, ya que, marcas de clase con varias cifras decimales las podemos transformar en números enteros (restando la marca de clase central y dividiendo por la amplitud de clase

=

$u$   $x$   $a$

$c$

Son aplicaciones de esta propiedad las llamadas *fórmulas abreviadas de la media* para variables continuas:

1) Si  $d_i = x_i - A$  fi  $x_i = d_i + A$  por lo sustituyendo en la fórmula de la media tendremos

$$\bar{X} = \frac{\sum f d_i}{\sum f} + A$$

iii

2) En el caso que  $d_i = c \cdot u_i$  entonces  $cu_i = x_i - A$ ;  $x_i = cu_i + A$

$\bar{X} =$

$\frac{\sum f cu_i}{\sum f} + A$

$$\bar{X} = \frac{\sum f cu_i}{\sum f} + A$$

iiii

Ej: Dada la tabla siguiente, perteneciente a la talla de 20 alumnos, vamos a efectuar las transformaciones expresadas anteriormente:

$X$	$x_i$	$d_i$	$f_i$	$d_i f_i$	$u_i$	$u_i f_i$
[1, 1,25]	1,125	-0,25	1/20	-0,0125	-1	-1/20
(1,25, 1,50]	1,375	0	2/20	0	0	0
(1,50, 1,75]	1,625	0,25	17/20	0,2125	1	17/20

De forma directa la media será:  $\bar{X} = \frac{1}{20} (1,125 + 2 \cdot 1,375 + 17 \cdot 1,625) = 1,575$  Siendo en este caso  $A = 1,375$  y  $c = 0,25$

—

Con lo cual  $\bar{X} = 1,375 + (0,2125 - 0,0125) = 1,575$

—

$$\bar{X} = 1,375 + 0,25 \cdot 0,8 = 1,575$$

iv) — Unicidad. Para un conjunto de datos, existe una sola media.

— Puesto que cada dato de la muestra entra en su cálculo, su valor está *afectado* por todos los datos. Así, los valores extremos (mínimo y máximo) de la muestra pueden alterar gravemente el valor de la media si

difieren mucho de los restantes, aún cuando tengan una baja frecuencia.

Esta última observación sugiere la necesidad de introducir otras medidas de centralización.

#### 4.2.1.1. Medias generalizadas

— A veces asociamos con los números  $X_1, X_2, \dots, X_k$  ciertos factores peso (o pesos)  $w_1, w_2, \dots, w_k$  dependientes de la relevancia asignada a cada número. En tal caso,

$$\bar{X} = \frac{w_1 X_1 + w_2 X_2 + \dots + w_k X_k}{w_1 + w_2 + \dots + w_k}$$

$$w_1 X_1 + w_2 X_2 + \dots + w_k X_k$$

se llama la media aritmética ponderada.

Ej.: si en el examen final de un curso cuenta tres veces más que una evaluación parcial, y un estudiante tiene calificación 85 en el examen final y 70 y 90 en los dos parciales, la calificación media es:

$$\bar{X} = \frac{3 \cdot 85 + 70 + 90}{3 + 1 + 1} = \frac{355 + 160}{5} = \frac{515}{5} = 103$$

—

— La **media geométrica**, que notaremos por  $X_G$ , se definirá como el resultado de hacer la operación:

$$\frac{n}{n} = 1$$

$$n \cdot \left( \prod_{i=1}^n x_i \right)^{1/n}$$

$$\sqrt[n]{x_1 x_2 \dots x_n}$$

$i=1$   
 $=$

Ej.: la media geométrica de 2, 4 y 8 es:

$$\sqrt[3]{2 \cdot 4 \cdot 8} = \sqrt[3]{64} = 4$$

— La **media armónica**, que notaremos por  $X_H$  se definirá como:

$$X_H = \frac{n}{\sum_{k=1}^n \frac{1}{x_k}}$$

$$\hat{A} \hat{A} \hat{A}$$

$$f(x) = \sum_{i=1}^n \frac{1}{x_i}$$

Ej.: la media armónica de los números 2, 4 y 8 es:  $\frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{8}} = \frac{3}{\frac{7}{8}} = \frac{24}{7} \approx 3.43$

$$X_H = \frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{8}} = \frac{24}{7} \approx 3.43$$

$$\frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{8}} = \frac{24}{7} \approx 3.43$$

— La **media cuadrática**, que notaremos por  $X_C$  se define por:

$$\sqrt{\frac{1}{n} \sum_{k=1}^n x_k^2}$$

$$X_C = \sqrt{\frac{1}{n} \sum_{k=1}^n x_k^2}$$

$$\hat{A} \hat{A} \hat{A}$$

$$\sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2}$$

Ej.: la media cuadrática de los números 2, 4 y 8 será:

$\bar{X}$

$C$

Á

3 2

$\sqrt{\frac{2^2 + 4^2 + 8^2}{3}} = \sqrt{\frac{4 + 16 + 64}{3}} = \sqrt{\frac{84}{3}} = \sqrt{28} \approx 5.29$

$\bar{X} = \frac{2 + 4 + 8}{3} = \frac{14}{3} \approx 4.67$

$\bar{X} < C < \bar{X}$

#### 4.2.1.2. Relación entre las medias

Algunos resultados, que no demostraremos, nos permiten establecer el siguiente orden entre estas tres últimas medias y la media aritmética:

$\bar{X} < C < \bar{X}$

### 4.3. Mediana

La *mediana* de un conjunto finito de datos es aquel valor que deja por debajo (o por encima) a la mitad de las puntuaciones de la distribución.

Veamos como se calcula, en cada caso, el valor de la mediana que notaremos por  $M_e$ :

a) **Variable discreta con un número impar de observaciones  $n = 2r + 1$ :** si esto es así, y ordenamos las observaciones (no los valores de la variable) de menor a mayor, habrá una de ellas que ocupe el lugar  $r + 1$ . El valor de la variable a que corresponda dicha observación será la mediana.

b) **Variable discreta con un número par de observaciones  $n = 2r$ :** ordenadas las observaciones de menor a mayor, la mediana se obtendrá como la media aritmética de las observaciones que ocupan los lugares  $r$  y  $r + 1$ .

c) **Variable continua:** buscaremos la mediana en un punto tal que  $F(M_e) = 1/2$ . Se pueden dar dos situaciones posibles:

— Que algún extremo de clase  $l_i$  verifique  $F(l_i) = 1/2$ , en tal caso  $l_i$  sería la mediana.

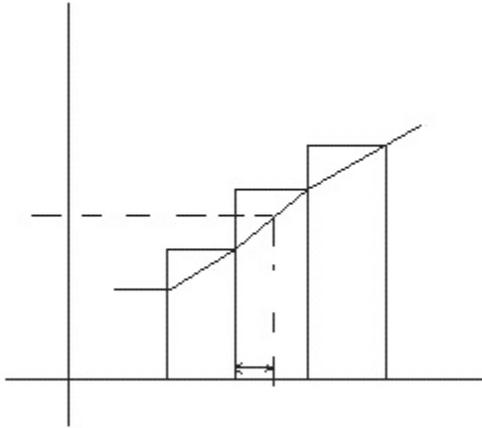
— Que no exista un extremo de clase que verifique la condición de mediana. Puesto que la función de distribución es creciente, existirá una clase  $(l_{i-1}, l_i]$  tal que  $F(l_{i-1}) < 1/2$  y  $F(l_i) > 1/2$ . Esto nos indicará que la mediana que vamos buscando es un punto de ese intervalo

$M_e$

$e_{i-1}$

Supongamos la poligonal de frecuencias acumuladas correspondiente al intervalo de la mediana y a sus dos contiguos:

$N_i$   
 $n/2$



$N_{i+1}$

$N_{i-1}$

$N_i$

1

k

$l_{i-2} \quad l_{i-1} \quad M_e \quad l_i \quad l_{i+1} \quad x_i$

Tenemos  $l_i - l_{i-1} = a_i$

$M_e - l_{i-1} = k$  luego  $M_e = l_{i-1} + k$

Planteamos la siguiente proporción:

$$\frac{n - N_{i-1}k}{a_i} = \frac{n}{a_i} - \frac{N_{i-1}k}{a_i} = 2$$

$a_i n_i$

^

Á

ii

luego  $k = \frac{2a_i}{n - N_{i-1}k}$  entonces finalmente:

$$k = \frac{2a_i n_i}{n - N_{i-1}k}$$

-  $n_i$

donde:

$l_{i-1}$ : es el extremo inferior de la clase mediana.

$a_i$ : es la amplitud de la clase mediana.

$N_{i-1}$ : es la frecuencia acumulada de la clase anterior a la clase mediana.  $n_i$ : es el número de observaciones en la clase mediana.

La mediana tiene varias cualidades que no posee la media aritmética:  
 — Su cálculo no se ve afectado en el caso de variable continua con algún extremo de clase infinito.

— El hecho de que exista alguna observación extraña no afecta al valor de la mediana.

— Unicidad. Cada conjunto de datos posee sólo una mediana.

#### 4.4. Moda

La *moda* de un conjunto de datos es el valor con mayor frecuencia.

— En el caso de variable discreta, la moda será aquel valor de la variable que presenta mayor frecuencia. La notaremos por  $M_0$  y puede no existir o existir más de una.

— En el caso de variable continua, llamaremos clase modal a aquella que presenta mayor frecuencia por unidad de longitud, es decir, aquella en que  $h_i = n_i/a_i$  sea mayor. De otra forma, la moda estará más cerca de aquel intervalo contiguo cuya frecuencia sea mayor.

$n_i$



$n_{i+1}$

$n_i$

$n_{i-1}$

$k$

$l_{i-}$

2  
 li-  
 1  
 Mo  
 li  
 li+ x 1<sup>1</sup>  
 Sea:

$$a_{i-1} = l_{i-1} - l_{i-2} \quad a_{i+1} = l_{i+1} - l_i \quad a_i = l_i - l_{i-1}$$

Pero  $k = ni+1$  y por la propiedad de las proporciones  $iak n_{i-1}$   
 $k iak - i akk = k$  entonces  $k = a_i \cdot ni+1 = =>$

n

n nn n nn i+ + +11 1 1 1 ii11

Por tanto  $Ma^{ni+1} o_{i-1} i nn ii11$

Si la amplitud del intervalo no es constante entonces  $h_i = ni$  por tanto la fórmula de la moda queda así:  $a_i$

$Ma^{hi+1} o_{i-1} i hh ii11$

A diferencia de la media y la mediana, un conjunto de datos puede poseer más de una moda o no tenerla.

#### 4.5. Relación empírica entre media, mediana y moda

Para curvas de frecuencia unimodales que sean poco asimétricas tenemos la siguiente relación empírica:

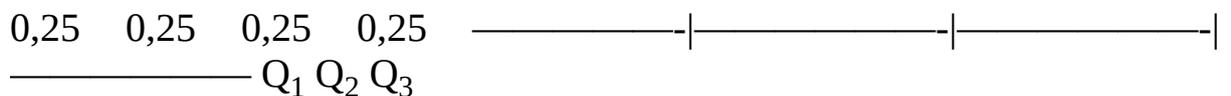
$$\text{Media} - \text{moda} = 3(\text{media} - \text{mediana})$$

Para curvas asimétricas a la derecha se cumple: Para curvas asimétricas a la izquierda se cumple: Para curvas simétricas se cumple la igualdad:  
 Moda < Mediana < Media    Media < Mediana < Moda    Media = Mediana = Moda

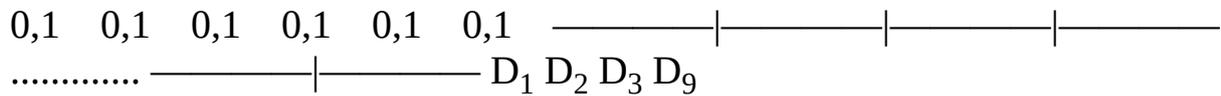
#### Asimétrica a la derecha Asimétrica a la izquierda

#### 4.6. Cuartiles, deciles y percentiles

Si un conjunto de datos está ordenado por magnitud, el valor central (o la media de los centrales) que divide al conjunto en dos mitades iguales, es la mediana. Extendiendo esa idea podemos pensar en aquellos valores que dividen al conjunto en cuatro partes iguales. Esos valores denotados por  $Q_1$ ,  $Q_2$  y  $Q_3$ , se llaman primer, segundo y tercer **cuartiles**, respectivamente. El  $Q_2$  coincide con la mediana.



Análogamente, los valores que dividen a los datos en 10 partes iguales se llaman **deciles**  $D_1, D_2, \dots, D_9$ , mientras que los valores que los dividen en 100 partes iguales se llaman **percentiles**, denotados por  $P_1, P_2, \dots, P_{99}$ . El 5.º decil y el 50.º percentil coinciden con la mediana. Los 25.º y 75.º percentiles coinciden con el primer y tercer cuartiles.



Colectivamente, cuartiles, deciles y percentiles se denominan **cuantiles**.

El cálculo de cuantiles es distinto según de trate de variable discreta o continua. En el caso continuo, se realiza de forma similar a como hemos hecho anteriormente para la mediana.

Para el caso discreto existen distintos procedimientos.

Si  $w$  es la suma de las ponderaciones para todos los casos válidos,  $p$  es el percentil dividido entre 100 y  $x_i$  es el  $i$ -ésimo menor valor. Veamos un ejemplo desarrollado con cada método para el cálculo de los cuartiles:

Sea la distribución  $x$ : 1, 2, 3, 7, 8, 9

a) **Haverage**: es el que utilizaremos a lo largo del resto de los temas. Si  $i$  es la parte entera de  $(w+1)p$  y  $d$  es la parte decimal, el valor del  $p$ -ésimo percentil es el promedio ponderado  $(1-d)x_i + d x_{i+1}$ .

$$w + 1 = 6 + 1 = 7$$

Para  $Q_1$   $p = 1/4 = 0,25$ ;  $(w + 1) p = 7 \cdot 0,25 = 1,75$  luego  $i = 1$   $d = 0,75$ ;  $x_i = 1$   $x_{i+1} = 2$   
 $Q_1 = (1-0,75) \cdot 1 + 0,75 \cdot 2 = 1,75$

Para  $Q_3$   $p = 3/4 = 0,75$ ;  $(w + 1) p = 7 \cdot 0,75 = 5,25$  luego  $i = 5$ ;  $d = 0,25$ ;  $x_i = 8$   $x_{i+1} = 9$   
 $Q_3 = (1 - 0,25) \cdot 8 + 0,25 \cdot 9 = 8,25$

b) **Waverage**: si  $i$  es la parte entera de  $wp$  y  $d$  es la parte decimal, el valor del  $p$ -ésimo percentil es el promedio ponderado  $(1-d)x_i + d x_{i+1}$ .  $w = 6$

Para  $Q_1$   $p = 1/4 = 0,25$ ;  $wp = 6 \cdot 0,25 = 1,5$  luego  $i = 1$ ;  $d = 0,5$ ;  $x_i = 1$ ;  $x_{i+1} = 2$   $Q_1 = (1-0,5) \cdot 1 + 0,5 \cdot 2 = 1,5$

Para  $Q_3$   $p = 3/4 = 0,75$ ;  $wp = 6 \cdot 0,75 = 4,5$  luego  $i = 4$ ;  $d = 0,5$ ;  $x_i = 7$ ;  $x_{i+1} = 8$   $Q_3 = (1-0,5) \cdot 7 + 0,5 \cdot 8 = 7,5$

c) **Round**: es el que suele aparece en los textos. Si  $i$  es la parte entera

de  $(w_p + 0,5)$ , el valor del  $p$ -ésimo percentil es  $x_i$ .

$$w = 6$$

Para  $Q_1$   $p = 1/4 = 0,25$ ;  $w_p + 0,5 = 6 \cdot 0,25 + 0,5 = 2$  luego  $i = 2$ ;  $x_i = 2$   $Q_1 = 2$

Para  $Q_3$   $p = 3/4 = 0,75$ ;  $w_p + 0,5 = 6 \cdot 0,75 + 0,5 = 5$  luego  $i = 5$   $x_i = 8$   $Q_3 = 8$

d) **Empirical**: si  $i$  es la parte entera de  $w_p$ , el valor del  $p$ -ésimo percentil será  $x_i$  o  $x_{i+1}$  dependiendo de si la parte decimal es cero o mayor que cero.  $w = 6$

Para  $Q_1$   $p = 1/4 = 0,25$ ;  $w_p = 6 \cdot 0,25 = 1,5$  luego  $i = 1$ ;  $x_i = 1$ ;  $x_{i+1} = 2$   $Q_1 = 2$  (por tener la parte decimal mayor que cero)

Para  $Q_3$   $p = 3/4 = 0,75$ ;  $w_p = 6 \cdot 0,75 = 4,5$  luego  $i = 4$   $x_i = 7$   $x_{i+1} = 8$   $Q_3 = 8$

e) **Aempirical**: si  $i$  es la parte entera de  $w_p$ , el valor del  $p$ -ésimo percentil será  $(x_i + x_{i+1})/2$  o  $x_{i+1}$  dependiendo de si la parte decimal es cero o mayor que cero.

$$w = 6$$

Para  $Q_1$   $p = 1/4 = 0,25$ ;  $w_p = 6 \cdot 0,25 = 1,5$  luego  $i = 1$ ;  $x_i = 1$ ;  $x_{i+1} = 2$   $Q_1 = 2$  (por tener la parte decimal mayor que cero)

Para  $Q_3$   $p = 3/4 = 0,75$ ;  $w_p = 6 \cdot 0,75 = 4,5$  luego  $i = 4$   $x_i = 7$   $x_{i+1} = 8$   $Q_3 = 8$

#### 4.7. Ejemplo de cálculo de la media, mediana, moda y cuartil

En la tabla siguiente se recogen los pesos en libras de 40 estudiantes varones de una universidad, con precisión de 1 libra. Construir una distribución de frecuencias, agrupándolos en intervalos de amplitud 10 libras y calcular la media, mediana, moda y primer cuartil ( $Q_1$ ) de los datos.

138 164 150 132 144 125 149 157 146 158 140 147  
136 148 152 144 168 126 138 176 163 189 154 165  
146 173 142 147 135 153 140 135 161 145 135 142  
150 156 145 128

Peso $X_i$	$n_i$	$f_i$	$x_i f_i$	$F_i$
(124-134]	129	4	0,1	12,9
(134-144]	139	12	0,3	41,7
(144-154]	149	13	0,325	48,425
(154-164]	159	6	0,15	23,85
(164-174]	169	3	0,075	12,675
(174-184]	179	1	0,025	4,475
(184-194]	189	1	0,025	4,725
	140			148,55

$$X = 148,55$$

$$M_e = 144 + (0,1 \cdot 10) / 0,325 = 144 + 3,077 = 147,077$$

$$M_o = 144 + 10 \cdot (6) / [12 + 6] = 147,33 \text{ (estará más cerca de 144 que de 154 porque la frecuencia del intervalo (134 – 144] es 12 y la del (154 – 164] es 6)}$$
$$Q_1 = 134 + (6 \cdot 10) / 12 = 139$$

## 5. MEDIDAS DE DISPERSIÓN

### 5.1. Introducción

La dispersión de un conjunto de datos hace referencia a la variabilidad entre estos. Si todos los datos son iguales, no existe dispersión; si no son todos iguales, se presenta dispersión. La «cantidad de dispersión» puede ser pequeña cuando los valores, aunque diferentes, están próximos entre sí. Si los valores están ampliamente separados, la dispersión será grande.

Hay varias medidas de dispersión:

— Medidas de dispersión basadas en el recorrido de la variable:

- Rango.
- Rango intercuartílico.

— Medidas de dispersión entorno a la media:

- Desviación media absoluta a la media.
- Varianza y desviación típica.
- Coeficiente de variación de Pearson.

— Medida de dispersión entorno a la mediana:

- Desviación media absoluta a la mediana.

### 5.2. Recorridos

Una medida de dispersión global es el *rango*, que es la diferencia entre la mayor y la menor de las observaciones. Si representamos el rango por  $R$ , la observación mayor por  $x_{\max}$ , y la observación menor por  $x_{\min}$ , resulta:

$$R = x_{\max} - x_{\min}$$

Cuando los valores de un conjunto de observaciones están próximos a la media muestral la dispersión es menor que cuando están repartidos en un amplio intervalo.

Basadas en la misma idea que la del recorrido de la variable aparecen, como medidas de dispersión, la distancia entre los dos puntos que dejan el  $r\%$  de la población central entre ellos. Para localizar estos dos puntos se utiliza la función de distribución del modo siguiente:

a) Localizamos en primer lugar, el punto que deja al  $(100 - r)/2\%$  de la

población por debajo de él. A ese punto se le suele notar por  $P_r$ .

b) Localizamos el punto que deja por encima de él a esa misma cantidad de población.

Entre las medidas definidas de este modo destaca el llamado **rango intercuartílico**:  $R_i = Q_3 - Q_1$ , donde  $Q_1$  es el punto que deja por debajo de él el 25% de la población, es decir,  $F(Q_1) = 0,25$ ;  $Q_1$  recibe el nombre de *primer cuartil*.  $Q_3$  sería el punto que deja por encima de él el 25% de la población, dicho de otro modo  $F(Q_3) = 0,75$ ; éste recibe el nombre de *tercer cuartil*.

**5.3. Medidas de dispersión en torno a la media** Supongamos la variable  $X$  con valores  $x_i$  (marcas de clase en el caso continuo) y frecuencias relativas  $f_i$ . Si calculamos su media:

$k$

$$\bar{X} = \sum_{i=1}^k f_i x_i$$

$i=1$

se trata de medir como de lejos están los valores  $x_i$  de  $X$ .

— **Desviación media absoluta a la media:** Para conseguir el objetivo anterior, parece que lo más fácil es sumar las diferencias entre  $x_i$  y la media. Esto nos podría conducir a un absurdo pues, aunque estas diferencias sean muy grandes, se podrían compensar al tener signo opuesto. Así llegaríamos a la conclusión de que no existe dispersión aún cuando ningún  $x_i$  coincida con la media. La definición de desviación media absoluta viene a evitarlo tomando valores absolutos. La expresión que la define es:

$k$

$$D.M. = \sum_{i=1}^k f_i |x_i - \bar{X}|$$

$i=1$

Por ejemplo:

Sea  $x$ : 1, 2, 3, 7, 8, 9

$x_i$  1 2 3 7 8 9

—

$|x - \bar{X}|$  4 3 2 2 3 4 18

—

$$D.M. = \sum_{i=1}^k f_i |x_i - \bar{X}| = 18/6 = 3$$

El poner a cada sumando  $f_i$  es como siempre, para pesar cada diferencia en la medida en que aparece en los datos.

— **Varianza y desviación típica:** Otra manera de soslayar el problema de los signos es tomando cuadrados. Esto nos conduce a otra medida de dispersión llamada **varianza**, y que notaremos por  $V_x$ :

$k_2 k$

$$V_x = \frac{\sum_{i=1}^k f_i (x_i - \bar{x})^2}{n}$$

$i=1$

Por ejemplo: sea  $x$ : 1, 2, 3, 7, 8, 9

$x_i$  1 2 3 7 8 9

—

$$(x_i - \bar{x})^2 \quad 16 \quad 9 \quad 4 \quad 4 \quad 9 \quad 16 \quad 58$$

$$V_x = 58/6 = 9,6$$

—

Si llamamos *desviación del dato*  $x_i$  respecto de la media a la diferencia  $(x_i - \bar{x})$ , la varianza se puede interpretar como «la media de los cuadrados de las desviaciones de los datos a la media».

Cuando tenemos que calcular la varianza de datos agrupados es aconsejable utilizar una corrección debida a Sheppard:  $\text{varianza corregida} = \text{varianza datos agrupados} - c^2/12$  donde  $c$  = anchura del intervalo.

La *varianza modificada*, o corregida, también llamada por algunos *cuasivarianza* de una muestra se define por la fórmula:

$$s^2 = \frac{\sum_{i=1}^k f_i (x_i - \bar{x})^2}{n} - \frac{c^2}{12}$$

ii - (0)

$$S^2 = \sum_{i=1}^k f_i (x_i - \bar{x})^2$$

—

La razón principal para introducir la varianza modificada es la siguiente: se ha observado que cuando se utilizan los estadísticos para sacar conclusiones sobre la población de la que proviene la muestra, la media de ésta es una buena estimación de la media de la población, sin embargo, la varianza muestral calculada como en [4] no da una buena

estimación de la población. Se ha demostrado que este error se reduce dividiendo la suma de los cuadrados por  $k-1$  en vez de  $k$ . Esta modificación es especialmente relevante en el caso de tamaños muestrales reducidos. El término  $k-1$  se denomina *grados de libertad*.

Como es usual en la mayor parte de la bibliografía, en lo sucesivo, el término *varianza muestral* se aplicará al estadístico calculado con  $k-1$  grados de libertad en el denominador.

La varianza se expresa como el cuadrado de las unidades en que se midió la variable original; no es, por tanto, una medida de dispersión totalmente apropiada. Se puede obtener otra medida de dispersión con las mismas unidades que los datos, simplemente tomando la raíz cuadrada de la varianza. El resultado se llama *desviación típica*:  $s_x$

La *desviación típica* se utiliza como una medida de dispersión intrínseca dentro de un conjunto de datos. Pero no se pueden comparar dos desviaciones típicas distintas cuando los datos están dados en distinta escala de medida.

La varianza y la desviación típica son las medidas de dispersión que más se utilizarán, de ahí que se estudie algunas de sus propiedades:

a) La varianza permanece invariante frente a traslaciones.

Demostración:

Si definimos la variable  $Y = X + A$  fi  $y_i = x_i + A$  con frec.  $f_i$

$k_2$

$$V_{f_y Y} = V_{f_x X + A} = V_{f_x X} = \sum_{i=1}^k f_i (x_i - \bar{x})^2$$

$i=1$

b) La varianza no permanece constante frente a cambios de escala.

Demostración:

Definimos la variable  $Y = A.X$  fi  $y_i = A.x_i$  con frec.  $f_i$

$k$

$k$

$$V_{f_y Y} = V_{f_x A X} = A^2 V_{f_x X} = A^2 \sum_{i=1}^k f_i (x_i - \bar{x})^2$$

$$\hat{A}_{y_{ii}} = \hat{A}_{x_{ii}} = \hat{A}$$

$i=1$

y por tanto  $s_y = A \cdot s_x$ . En el caso de variables que están representadas en intervalos, basándonos en esta propiedad vamos a ver alguna fórmula para calcular la varianza:

$$V_{\hat{E}}^2 = \sum_{i=1}^n f_i x_i^2 - (\sum_{i=1}^n f_i x_i)^2$$

~ si hacemos el cambio  $x_i = d_i + A$  entonces:  $\hat{E} = \sum_{i=1}^n f_i d_i + A = \bar{d} + A$

$$V_{\hat{E}}^2 = \sum_{i=1}^n f_i (d_i + A)^2 - (\sum_{i=1}^n f_i (d_i + A))^2$$

Á  $\hat{E} = \bar{d} + A$   
 si el cambio es de la forma  $x_i = A + c \cdot u_i$   
 $\hat{E} = \bar{d} + A = \bar{c} \cdot \bar{u} + A$

$$V_{\hat{E}}^2 = \sum_{i=1}^n f_i (c \cdot u_i + A)^2 - (\sum_{i=1}^n f_i (c \cdot u_i + A))^2$$

Veamos un ejemplo de cálculo de la varianza para datos agrupados. Sean los pesos de 100 estudiantes dados por la siguiente distribución:

**Peso  $n_i$   $x_i$   $d_i = x_i - A$   $d_i^2 n_i$**  60-62 5 61 -6 -30 180

63-65 18 64 -3 -54 162 66-68 42 67 0 0 0

69-71 27 70 3 81 243

72-74 8 73 6 48 288 100 45 873

$$V_x = 873/100 - (45/100)^2 = 8,52; s_x = 2,92$$

Si utilizamos la fórmula para amplitud constante tenemos:

**Peso  $n_i$   $x_i$   $u_i$   $u_i^2 n_i$**  60-62 5 61 -2 -10 20

63-65 18 64 -1 -18 18 66-68 42 67 0 0 0

69-71 27 70 1 27 27

72-74 8 73 2 16 32 100 15 97

$V_x = 3^2[97/100 - (15/100)^2] = 8,52; s_x = 2,92$  c) Entre la desviación típica de una variable y la desviación media absoluta, existe siempre la relación

--

D.M.  $\leq s_x$  ya que  $|x_i - X| \leq (x_i - X)^2$

— **Coefficiente de variación de Pearson:** Consideremos una misma variable estudiada en dos poblaciones distintas, y supongamos que las medias de estas variables, en las dos poblaciones, son muy diferentes.

Ej: El peso en una población de caballos y otra población de gatos. Es evidente que una misma desviación, digamos  $s = 1$  kg, representa una dispersión insignificante en la población de caballos. En cambio una desviación de 1 kg para el peso medio de los gatos indica una dispersión importante. Es decir, si sabemos que la desviación ha sido 1 kg en ambas poblaciones, se debe aportar alguna información más sobre la dispersión.

Las consideraciones anteriores sugieren la conveniencia de introducir una medida de *variación relativa* mejor que una de *variación absoluta*. Dicha medida se llama coeficiente de variación de Pearson.

El *coeficiente de variación* es la desviación típica medida en unidades de la

media.  $\text{Coeficiente de variación} = \frac{s}{X}$

Cuando se desea expresar la desviación típica como un porcentaje de la media

se emplea la fórmula:  $\text{coeficiente de variación} \cdot 100$ . Es importante tener en cuenta que este coeficiente tiene problemas de cálculo cuando la media se aproxima a 0.

—

Este coeficiente, por el comportamiento de  $s$  y  $X$ , resulta invariante frente a cambios de escala, pero no frente a traslaciones.

#### 5.4. Medidas de dispersión en torno a la mediana

Se define la desviación media absoluta respecto a la mediana como una medida de la dispersión, de una variable  $X$ , entorno a su mediana como:

$$\sum_{k=1}^n |x_k - M|$$

$$= \sum_{i=1}^n |x_i - \hat{A}|$$

$i=1$

Sobre ella sólo enunciaremos la relación que guarda con la desviación media absoluta a la media aritmética.

D  $\neq$  D · M Otra de las medidas de dispersión en torno a la mediana más utilizadas es la mediana de las desviaciones absolutas a la mediana, denominada como **MEDA** y definida por:

$$\text{MEDA} = \text{mediana } |x_i - M_e|$$

que tiene la ventaja, como la mediana, de no verse afectada por datos extremos. A las medidas que tienen esta propiedad las llamaremos *medidas robustas* o *resistentes*. Si conocemos la mediana y la MEDA de datos no agrupados sabemos que, al menos, el 50% de los datos están en el intervalo  $(M_e \pm \text{MEDA})$ .

Veamos, como ejemplo, el cálculo de los valores D y MEDA para la variable x cuya distribución es: 1, 2, 3, 7, 8, 9

$$x_i \quad 1 \quad 2 \quad 3 \quad 7 \quad 8 \quad 9$$

$$|x_i - M_e| \quad 4 \quad 3 \quad 2 \quad 2 \quad 3 \quad 4 \quad 18$$

$$D = 18/6 = 3$$

Ordenando las diferencias  $|x_i - M_e|$  tenemos 2 2 3 3 4 4 luego MEDA = 3

Como podemos comprobar se cumple  $(M_e \pm \text{MEDA})$  está al menos el 50% de los casos. En nuestro ejemplo el 67%.

### 5.5. Tipificación de una variable

Supongamos que dos estudiantes discuten sobre el significado de sus respectivas notas obtenidas en clases diferentes. Una comparación directa de estas notas, podría carecer de significado, pues bastaría con el baremo utilizado para corregir fuese distinto en una y otra clase. Este problema se puede resolver estudiando la posición relativa, que ha obtenido cada alumno en su clase. Para obtener estas posiciones relativas, restaríamos a cada nota su correspondiente media, y dividiríamos por su desviación típica. A la nueva variable construida se la denomina variable tipificada de la inicial.

#### 5.5.1. Definición de variable tipificada

Dada la variable estadística X que toma valores  $x_i$  con frecuencias relativas  $f_i$ , se define la variable tipificada de X como la variable transformación lineal de la inicial

=

$$Z = \frac{x - \bar{x}}{s}$$

$s_x$

esta variable tiene las siguientes propiedades: tiene media 0 y desviación típica 1. La media y la desviación típica nos aporta una información sobre la distribución de los datos:

— En una distribución cualquiera:  $(\bar{X} - s, \bar{X} + s)$  está entre el 50% y el 75% de los casos. Si la distribución es normal el 68,27%.

— En una distribución cualquiera:  $(\bar{X} - 2s, \bar{X} + 2s)$  está entre el 90% y el 100% de los casos. Si la distribución es normal el 95,45%.

Como ejemplo vamos a calcular los valores  $z_i$  de la variable  $x$ : 1, 2, 3, 7, 8,

9

$x$

$i$

$z$

$i$

$= (x$

$-$

$- \bar{x})/s_x$

1 -1,29

2 -0,97

3 -0,64

7 0,64

8 0,97

9 1,29

Podemos comprobar que

—

$\bar{X} \pm s_x = 5 \pm 3,1$  están el 67% de los casos.

## 6. MEDIDAS DE FORMA

Una vez estudiadas la tendencia central de la población y en qué medida es ésta fiable, nos disponemos a estudiar de qué forma se reparte esta dispersión entorno al promedio.

### 6.1. Medidas de simetría

Una inspección somera de los histogramas dibujados más abajo (ver fig. 2.6.1) permite apreciar diferentes situaciones de simetría(o asimetría) respecto de la recta vertical que pasa por la media. Los tres histogramas poseen la misma media

—

$\bar{x}^a 0$  y el mismo tamaño  $n = 95$ .

Simetría

Asimétrica a la derecha

Asimétrica hacia la izquierda

FIGURA 2.6.1. Simetría.

— **Definición de distribución de frecuencias simétricas respecto de un punto a:** La distribución de frecuencias de  $X$  resulta simétrica respecto de  $a$ , si la gráfica del diagrama de barras o histograma resulta simétrica respecto de la vertical que pasa por  $a$ .

— **Definición de distribución de frecuencias simétrica:** Se dice que una distribución de frecuencias es simétrica si lo es respecto a su mediana.

— **Proposición:** Si  $X$  posee una distribución de frecuencias simétrica, la media, la mediana y la moda (si la distribución es unimodal) coinciden.

Diremos que una distribución de frecuencias es asimétrica a la izquierda, o *sesgada a la izquierda*, si su diagrama diferencial está más estirado a la izquierda. Análogamente, diremos de una distribución que es asimétrica a la derecha, o *sesgada a la derecha*, si el diagrama diferencial presenta la cola más larga a la derecha.

Puesto que estudiar la simetría mediante la gráfica de cada distribución resulta un método poco operativo, definiremos algunos estadísticos para estudiar esta característica:

a) **Primer coeficiente de sesgo de Pearson:** Basándose en la idea de que en una distribución sesgada, la media tiende a situarse, con respecto a la moda, al mismo lado que la cola más larga:

$$s_1 = \frac{3M - M_o}{s_x}$$

Su interpretación es:

—  $s_1 = 0$  si y sólo si la distribución es simétrica.

—  $s_1 > 0$  si y sólo si la distribución es sesgada a la derecha. —  $s_1 < 0$  si y sólo si la distribución es sesgada a la izquierda.

b) **Segundo coeficiente de sesgo de Pearson:** Utilizando la fórmula empírica aproximada, para distribuciones moderadamente asimétricas,  $X - M_o = 3 \cdot (X - M_e)$

se define  $s_2$  como:

$$s_2 = \frac{3(X - M_o) - (X - M_e)}{s_x^2}$$

$s_{XM()}$

=◇

$s_x$

y naturalmente su interpretación es la misma que la de  $s_1$ .

c) **Coefficiente de simetría de Fisher:** Se trata de uno de los coeficientes más utilizados para medir la simetría. Su expresión es la siguiente:

$g$

$1$

=

$m_3$

$()^3$

$k^3$  donde  $m_3$  es el momento centrado de orden 3, es decir,  $m_3 = \int \hat{A}^{f_{x-X}}$

$i()$

$i=1$

Su interpretación es la siguiente:

— Es nulo para todo el histograma simétrico respecto de la media.

— Es positivo para los histogramas cuya cola derecha (rectángulos situados a la derecha de la media) es más larga que la cola izquierda: asimetría positiva. — Es negativo para los histogramas cuya cola izquierda es más larga que su cola derecha: asimetría negativa.

1. El coeficiente de asimetría es adimensional e invariante por cambios de escala en los datos.

2. Igualmente es invariante por desplazamiento del histograma a derecha o izquierda.

3. Las condiciones 1 y 2 son aplicables al coeficiente de apuntamiento.

Los coeficientes de asimetría para los tres gráficos de la fig. 2.6.1 son, respectivamente:  $g_1=0$   $g_1=0,81$  y  $g_1=- 1,45$ .

## 6.2. Medidas de apuntamiento

Los histogramas más simétricos (o aproximadamente simétricos) pueden presentar una de las formas que aparecen en fig. 2.6.2.

Caso 1

Caso 2

FIGURA 2.6.2. Apuntamiento.

El histograma del caso 1 es más «bajo» y «ancho» que el histograma del caso 2.

Las características bajo/alto y ancho/estrecho de los histogramas simétricos se estudian comparándolos con la forma de una curva standard: la distribución normal. Esta curva tiene una forma característica (llamada campana de Gauss) que se presenta en la fig. 2.6.3.

FIGURA 2.6.3. Curva normal.

El coeficiente que se utiliza para la comparación se llama coeficiente de apuntamiento o curtosis.

El *coeficiente de apuntamiento o curtosis de Fisher* es la media de las cuartas potencias de las desviaciones de los datos a la media muestral ( $m_4$ ) dividida por  $s^4$  y disminuido el cociente en 3 unidades:

$m$

$$g_2 = \frac{m_4}{3s^4}$$

Dicho coeficiente posee las siguientes características:

— Si el histograma es muy parecido a la curva normal,  $g_2$  es casi nulo; y la llamaremos *mesocúrtica*.

— Si el histograma es más alto y estrecho que la curva normal,  $g_2$  es positivo. A estas distribuciones las llamaremos *leptocúrticas*.

— Si el histograma es más bajo y ancho que la curva normal,  $g_2$  es negativo. Las llamaremos *platicúrticas*.

### 6.3. Momentos

La media, la varianza y las componentes de los coeficientes de asimetría y apuntamiento de un conjunto de datos forman parte de una clase más amplia de estadísticos llamados momentos muestrales.

Si  $r$  es un número entero positivo, el momento de orden  $r$  respecto del origen es la media de las potencias de orden  $r$  de los datos de la muestra:

$$\hat{A}X_{ir} = \frac{1}{n} \sum_{i=1}^n X_i^r$$

$$m_{r, \cdot}^2 = \frac{1}{n} \sum_{i=1}^n X_i^r. \text{ Observemos que } m_1 = \bar{X} \text{ y } m_2 = s^2$$

Si  $r$  es un número entero positivo, el momento central de orden  $r$ , o momento de orden  $r$  respecto de la media, es «la media de las potencias de orden  $r$  de las desviaciones respecto de la media muestral»:

$n \hat{A}_{XX}$

**$i 0$**

$m$   
 $r$   
 $=$   
 $i=1$   
 $n$

Observemos que:  $m_1 = 0$

$$m_2 = s^2$$

$$g_1 = m_3/s^3$$

$$g_2 = (m_4/s^4) - 3$$

Las siguientes relaciones entre los momentos  $m_1, m_2, m_3, m_4$  y  $m_1, m_2, m_3, m_4$  son de utilidad en el cálculo de los coeficientes de  $g_1$  y  $g_2$ ,

a)  $m_2 = m_2 - m_{12}$

b)  $m_3 = m_3 - 3m_2m_1 + 2m_{13}$

c)  $m_4 = m_4 - 4m_3m_1 + 6m_2m_{12} - 3m_{13}$

El momento de orden  $(r,s)$  respecto al origen de un variable bidimensional  $(x,y)$  toma la expresión:

$m$   
 $rs$   
 $=$   
 $1$   
 $h k$

**$\hat{A}$**

$\hat{A}$   
 $r$   
 $s$   
 $xyn$   
 $n$   
 $i j i=1 j=1$

Cuando se trata de estudiar dos variables  $(x,y)$  los primeros momentos con relación al origen toman la expresión y son:

$m_{10}$  la media con relación a  $x$   
 $m_{01}$  la media con relación a  $y$   
 $m_{20}$  el momento de segundo orden con relación a  $x$   
 $m_{02}$  el momento de segundo orden con relación a  $y$   
 $m_{11}$  el momento de segundo orden con relación a  $x$  e  $y$

Un ejemplo aclaratorio de esta nomenclatura se puede encontrar en Martín (1994, pág. 231).

## 7. LA EXPLORACIÓN DESCRIPTIVA DE DATOS

### 7.1. Datos atípicos

Los datos que obtenemos al realizar una investigación pueden ser clasificados en dos categorías excluyentes:

- a) Datos correctos
- b) Datos erróneos

Los datos erróneos son debidos a múltiples factores (errores de transcripción, distintos instrumentos utilizados en la recogida de datos, etc). Hay otros datos *atípicos* denominados *outliers* y *extremos*, que tenemos que tener en cuenta a la hora de hacer cualquier estudio pues existen estimadores (funciones de los estadísticos) que son sensibles frente a estos datos y la información que nos proporcionan, si existen datos de este tipo, es errónea.

Solamente los expertos podrán discernir entre datos atípicos y datos erróneos. En el texto se considerará ambos como atípicos.

### 7.2. Gráfico caja (Box Plot)

El diagrama de caja es una representación semigráfica de una distribución, construida para mostrar sus características principales y señalar los posibles datos atípicos, es decir, aquellas observaciones que parecen ser distintas de las demás.

Un diagrama de caja se construye como sigue:

- Ordenar los datos de la muestra y obtener el valor mínimo, el máximo, y los tres cuartiles  $Q_1$ ,  $Q_2$ ,  $Q_3$ .
- Dibujar un rectángulo cuyos extremos son  $Q_1$  y  $Q_3$  e indicando la posición de la mediana ( $Q_2$ ) mediante una línea o un asterisco.

Si denominamos  $IQR = Q_3 - Q_1$  podemos calcular los límites admisibles, superior e inferior, que van a servir para identificar los valores atípicos. Estos límites se calculan con:

$$LI = Q_2 - 1,5 (IQR) \quad FI = Q_2 - 3 (IQR)$$

$$LS = Q_2 + 1,5 (IQR) \quad FS = Q_2 + 3 (IQR)$$

Considerar como valores atípicos *outlier* los situados entre LI y FI o entre LS y FS y como atípicos *extremos* aquellos fuera del intervalo (FI,FS). Vendrán representados en SPSS en el gráfico boxplot mediante: (°) en el caso de outlier y (\*) para el extremo.

— Dibujar una línea que vaya desde cada extremo del rectángulo central hasta

el *valor más alejado no atípico*, es decir, que está dentro del intervalo (LI, LS). — Identificar todos los datos que están fuera del intervalo (LI,LS), marcándolos como *atípicos* (outlier o extremos).

La razón de utilizar la mediana como medida de centralización y el rango intercuartílico como medida de dispersión es que estas medidas, al depender del orden de los datos y no de su magnitud, son poco influenciadas por unos pocos datos atípicos.

Por ejemplo, si cometemos un error al transcribir un dato, la media y la desviación típica pueden cambiar mucho, mientras que la mediana y los cuartiles cambiarán poco o nada. En consecuencia, si existe una pequeña fracción de datos en la muestra heterogénea con el resto (por errores de medida, cambio de las condiciones de medición, distinta población, etc.) los cuartiles variarán poco, y pueden servirnos para identificar los valores extremos.

Los diagramas de caja son especialmente útiles para comparar la distribución de una variable en distintas poblaciones.

Si la distribución se aproxima a la normal los valores contenidos en el rango intercuartílico (IQR) son el 50% del total. También podemos utilizar para detectar observaciones erróneas la siguiente regla relacionada con la  $z$  (variable tipificada): si la  $z$  para una medición  $x$  es igual o menor que  $-2$  o bien igual o mayor que  $2$ , debe ser considerada como un valor inusual y tendrá que verificarse por si se tratara de una observación errónea (outlier o extremo). Así, por ejemplo, consideremos los siguientes datos: 3,2,0,15,2,3,4,0,1,3. Si observamos, a priori parece inusual 15. Si tipificamos este valor tendremos el correspondiente  $z = (15 - 3,3)/4,32 = 2,71$  luego podemos decir que este valor se puede considerar como atípico.

### 7.3. Comparación de distribuciones

Supongamos el siguiente ejemplo donde comparamos la distribución de los países del Mercado Común con el resto. Tenemos:

**M.C.** 2,2 2,3 3,2 3,9 4,6 5,3 5,5 5,9 7,4 9,1 15,9 20,1 **Resto** 1,7 2,9 3,2 3,4 4,1 5,1 5,8 6,7 7,6 16,3 32,7 40,5

Para los países del mercado común la mediana es  $5,4 = [(5,3 + 5,5)/2]$  y los cuartiles  $Q_1 = 3,375$   $Q_3 = 8,675$ . Para el resto estos valores son  $Q_1 = 3,25$ ;  $Q_2 = 5,45$ ;  $Q_3 = 14,125$ . Los intervalos de admisibilidad serán  $(5,4 \pm 7,95)$  para el Mercado Común y  $(5,45 \pm 16,3125)$  para el resto. La figura inferior muestra los diagramas de caja para ambos colectivos. Se observa que en ambos hay dos valores atípicos y que la variabilidad es mucho más baja entre los países del Mercado Común.

2,2 3,375 5,4 8,675 9,1 Mer. Común  
Resto 1,7 3,25 5,4 14,125 16,3

## 8. REGRESIÓN Y CORRELACIÓN

### 8.1. Relación entre variables: aspectos generales

El estudio de una variable bidimensional (con dos dimensiones o componentes) se hace, en la mayoría de los casos, para intentar buscar una relación entre sus componentes. Existen dos enfoques para esta búsqueda:

— **Regresión:** es una técnica dirigida a encontrar una expresión que explique una de las componentes en función de la otra  $Y = f(X)$  o  $X = g(Y)$

Las técnicas de regresión intentarán a partir de unas variables independientes o explicativas encontrar la función que mejor explique el comportamiento de una variable dependiente o explicada.

Según que tipo de función se utilice para ello, estaremos ante la regresión de tipo lineal, parabólica, exponencial, etc.

El término de regresión se debe a Galton, que lo introdujo cuando estudiaba la relación entre la talla de padres e hijos.

— **Correlación:** con ella vamos a dar una medida numérica del grado de dependencia que hay entre las componentes de la variable bidimensional. Dependiendo del tipo de dependencia que estemos estudiando, nos encontraremos con los distintos coeficientes de correlación.

### 8.2. Regresión

El análisis de regresión trata el problema de describir o pronosticar el valor de una variable, llamada variable *dependiente*, sobre la base de otra variable u otras variables llamadas *independientes*.

El problema puede concretarse, por ejemplo, cuando se quiere predecir la calificación de un alumno en matemáticas sabiendo su comportamiento en otras materias. O también cuando se sabe la existencia de cierta influencia en los resultados de la mencionada materia de un conjunto de habilidades intelectuales, y se quiere saber la configuración de dicha relación.

### **8.2.1. Estudio particular de la regresión lineal: ajuste de una recta por el método de mínimos cuadrados**

Cuando se tiene una nube de puntos formado por pares  $(x_i, y_i)$  y nos planteamos el problema de ajuste de  $y$  mediante una función de  $x$ , es muy útil hacer una representación gráfica previa con la idea de darnos una visión intuitiva del tipo de función que mejor se ajusta a nuestros datos.

Las funciones más utilizadas en los ajustes son las siguientes:

- a) Línea recta  $Y = a_0 + a_1X$
- b) Parábola  $Y = a_0 + a_1X + a_2X^2$
- c) Curva de grado  $n$   $Y = a_0 + a_1X + a_2X^2 + \dots + a_nX^n$
- d) Curva exponencial  $Y = ab^X$
- e) Curva geométrica  $Y = aX^b$

El método utilizado para tales ajustes es minimizar la suma de las distancia al cuadrados entre los puntos y los valores de la curva para los respectivos datos de la variable independiente.

Veamos a continuación la forma de proceder en el caso de la línea recta y de forma similar se realizará para otras curvas.

#### **8.2.1.1. Recta de mínimos cuadrados**

Supongamos  $N$  puntos  $(x_1, y_1)$   $(x_2, y_2)$  .....  $(x_N, y_N)$  (ver fig. 2.8.1) queremos calcular la ecuación de una recta que se aproxima a esos puntos  $y = ax + b$ , y tal que la distancia  $\sum e_i^2 = \sum [y_i - (ax_i + b)]^2$  sea mínima. Donde  $e_i = y_i - (ax_i + b)$ , expresa el error en la predicción. Supongamos  $f(a, b) = \sum e_i^2$ , es decir, la suma de errores al cuadrado es función de los parámetros  $a$  y  $b$ , como tenemos que minimizar esta función hallaremos las derivadas parciales respecto a  $a$  y  $b$  respectivamente e igualando a cero cada una de ellas tenemos las siguientes ecuaciones normales:

$$\sum y_i = bN + a\sum x_i$$

$$\sum x_i y_i = b\sum x_i + a\sum x_i^2$$

Este sistema de ecuaciones se puede resolver por cualquier método algebraico, como por ejemplo Cramer, con lo cual después de diversas

operaciones tenemos finalmente:

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}$$

$$b = \frac{S_{xy}}{S_{xx}}$$

En esta ecuación  $b$ , la *pendiente* de la recta, será:

$$b = \frac{S_{xy}}{S_{xx}}$$

$S_{xy}$  Covarianza / Varianza de  $x$

$$S_{xx} = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

$a$ , la *ordenada en el origen*, valdrá:

$$a = \bar{y} - b \bar{x}$$

Si  $x$  es la variable dependiente entonces la ecuación anterior y las normales resultan de forma parecida, solamente debemos cambiar la  $x$  por la  $y$ .

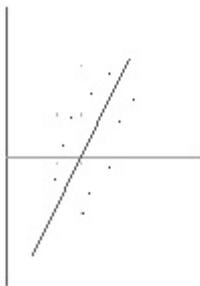


FIGURA 2.8.1. Recta mínimo cuadrática.

Hay que resaltar que la estimación de  $y$  sólo es cierta para valores dentro del margen utilizado en la predicción y nunca fuera de este intervalo, es decir nunca para valores de extrapolación.

### El método de mínimos cuadrados:

- a) Es un método de ajuste de curvas y por tanto no exige suposiciones respecto a las distribuciones de las variables, ni de los errores.
- b) Puede ser aplicado en regresiones no lineales.

### 8.3. Correlación

Cuando tratamos de establecer el grado de asociación entre dos

variables estamos refiriéndonos a su correlación, que podemos entender como el conjunto de causas comunes a ambas, es decir, el grado de dependencia mutua entre las variables. Así, si una variable (por ejemplo,  $y$ ) se supone <<motivada>> o <<explicada>> por otra (por ejemplo,  $x$ ), la correlación entre ambas nos hablará de su asociación, del grado en que los fenómenos representados por ambas poseen causas comunes (por ejemplo, como sucede cuando se quiere examinar la variación del rendimiento académico ante variaciones producidas en el cociente intelectual). Evidentemente, existe proximidad entre este problema y el problema de regresión pero hay una diferencia clara entre ambos, pues mientras ésta trata de representar el comportamiento de una variable según el atribuido a la otra, aquella se limita a explicar la asociación de las dos variables.

En la representación gráfica de la nube de puntos, si todos los puntos estuvieran sobre la función que los representa, la correspondencia sería uno a uno y por tanto el grado de dependencia sería máximo. Cuanto más se alejan los puntos de la función, es decir, conforme mayor sea la diferencia entre el valor de  $y$ , y su estimación mediante la función  $y^*$ , menor será la intensidad de la asociación, o de otra forma, mayor será el error ( $e_i = y_i - y_i^*$ ).

### 8.3.1. Coeficiente de correlación

Si indicamos por  $y_i^*$  el valor de  $y$  para los valores estimados con la recta de mínimos cuadrados, una medida de dispersión respecto de la recta de regresión de  $y$  sobre  $x$  será:

$$s_e = \sqrt{\frac{\sum (y_i - y_i^*)^2}{N}}$$

que se denomina *error típico* de  $y$  sobre  $x$ .

El error típico tiene propiedades similares a la desviación típica y nos permite construir regiones de confianza trazando rectas paralelas a la recta de regresión, donde para distancias verticales  $S_e$ ,  $2S_e$  y  $3S_e$  estarán el 68%, 95% y 99,7% respectivamente para  $N$  lo suficientemente grande.

Por otra parte, la *variación total* de  $y$  se define como:

$$s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{N}}$$

iii)  $\sum (y_i - \hat{y}_i)^2 = \sum (y_i - \bar{y})^2 - \sum (\hat{y}_i - \bar{y})^2$

es decir, variación total = variación no explicada + variación explicada. Si se eleva al cuadrado y se realiza la suma la descomposición queda así:

$$\sum (y_i - \hat{y}_i)^2 = \sum (y_i - \bar{y})^2 - \sum (\hat{y}_i - \bar{y})^2$$

, es decir,  $SCT (S_{y^2}) = SCE (S_{e^2}) + SCR (S_{R^2})$

Donde SCT será la suma de cuadrados total, SCE la suma de cuadrados de los errores y SCR la suma de cuadrados debido a la regresión.

Denominaremos *coeficiente de determinación* al cociente entre la SCR (variación explicada por la regresión) dividido entre la variación total (SCT) y varía entre 0 y 1. En el caso de la regresión lineal, su raíz cuadrada se llamará *coeficiente de correlación (r)* y como es natural variará entre -1 y 1 con  $r = 0$  cuando las variables están incorreladas, y  $r = 1$  o  $r = -1$  si la correlación es máxima.

Podemos poner

$$r = \pm \frac{S_{R^2}}{S_{y^2}}$$

$R^2$  y como  $1 = e R e$

$$\frac{S_{R^2}}{S_{y^2}}$$

+ tenemos  $r = \pm \frac{S_{R^2}}{S_{y^2}}$

o de otra forma

SS 1r2

$r_{xy} = \frac{S_{xy}}{S_x S_y}$

Las ecuaciones [4] y [5] son aplicables tanto en relaciones lineales como no lineales, pero en el caso lineal la fórmula utilizada es la denominada de productomomento o correlación de Pearson cuya expresión es la siguiente:

$S_{xy}$  donde  $S_{xy}$  representa la covarianza de las variables  $x$  e  $y$  y  $S_x$   $S_y$  son las desr =<sub>SS</sub>

$S_x$   $S_y$  son las desr =<sub>SS</sub> variaciones típicas de las variables.

### Ejemplo

Dada la siguiente distribución:

$x_i$	$y_j$	$n_{ij}$
2	3	5
2	4	12
3	5	15
4	5	20
6	6	20
7	8	16
9	9	8
9	6	4
		100

Calcular:

- La recta de regresión de  $y$  sobre  $x$ .
- Estudiar el grado de dependencia lineal entre ambas variables.

Solución:  $S_{xy}$  a) Como sabemos  $y = a + bx$  donde  $a = \frac{S_y}{S_x} - b \frac{S_x}{S_x}$ ;  $b = \frac{S_{xy}}{S_x^2}$

$x$

Para calcular la varianza y covarianza planteamos la siguiente tabla de correlación:

$x \backslash y$	3	4	5	6	8	9	$n_i$	$x_i n_i$	$x_i^2 n_i$
2	5	12	17	34	68				
3	15	15	45	135					
4	20	20	80	320					
6	20	20	120	720					
7	16	16	112	784					

$n_j$ : 5 12 35 24 16 8 100 499 2999  
 $y_j n_j$ : 15 48 175 144 128  
 72 582  
 $y_j^2 n_j$ : 45 192 875 864 1024 648 3648

En primer lugar:

$$\hat{A} = \frac{\sum y_j n_j}{\sum n_j} = \frac{582}{100} = 5,82$$

$$\hat{A} = \frac{\sum x_i n_i}{\sum n_i} = \frac{499}{100} = 4,99$$

$$S = \frac{\sum x_i^2 n_i - \frac{(\sum x_i n_i)^2}{N}}{N} = \frac{2999 - \frac{499^2}{100}}{100} = 509$$

$$\hat{A} = \frac{\sum y_j^2 n_j - \frac{(\sum y_j n_j)^2}{N}}{N} = \frac{261 - \frac{582^2}{100}}{100} = 2,55$$

$\hat{A}$

$$\hat{A} = \frac{\sum x_j y_j n_j}{\sum n_j} = \frac{3096}{100} = 30,96$$

luego  $b = \frac{30,96 - 5,82(4,99)}{2,55} = 5,09$

$a = 5,82 - 0,5(4,99) = 3,33$   
 en consecuencia la recta de regresión será:

$$y = 3,33 + 0,5x$$

b) La dependencia lineal vendrá dada por el coeficiente de correlación lineal:

s

r

=  $\frac{r_{xy}}{s_x s_y}$

$\frac{2,55}{2,55}$

$\frac{261,070}{261,070}$

Como se observa un coeficiente de correlación positivo e importante

## 9. ANÁLISIS DESCRIPTIVO DE DATOS: COMANDOS

### 9.1. Introducción

Como complemento a la estadística descriptiva y para facilitar el cálculo se va a detallar las opciones de SPSS que realizan dichos cálculos, y que en resumen, nos permiten una depuración, descripción y representación de los datos. Los comandos EXAMINE, FREQUENCIES, DESCRIPTIVES y CORRELATION, además del conjunto de comandos relacionados con la representación gráfica. Estos comandos, cuando SPSS no funcionaba aún bajo Windows, era imprescindible conocerlos en profundidad para realizar los análisis. No obstante sin el apremio anterior, hay dos motivos para no olvidarse de los mismos: a) algunas opciones de los comandos no están contempladas en los menús y b) el manejo de los comandos permite componer programas que facilitan al investigador la realización de estudios con características similares con ahorro de tiempo y esfuerzo. Recordemos que cuando en el menú de SPSS se da la opción **PEGAR**, aparecerán en el fichero de sintaxis los comandos que realizan las opciones seleccionadas en el menú.

### 9.2. La exploración de datos. Comando EXPLORAR

La opción de **EXPLORAR** de SPSS proporciona diversos tipos de gráficos y estadísticos para el análisis exploratorio de datos. Permite el análisis bien de todos los casos o bien de un grupo de casos por separado. Esta exploración de los datos permite detectar valores atípicos, comprobar los supuestos paramétricos de los datos, comprobar supuestos y características entre subpoblaciones (grupos de casos formados en relación a una variable o factor), indagar posibles transformaciones para conseguir la normalidad de la distribución, etc.

#### *Problema-ejemplo*

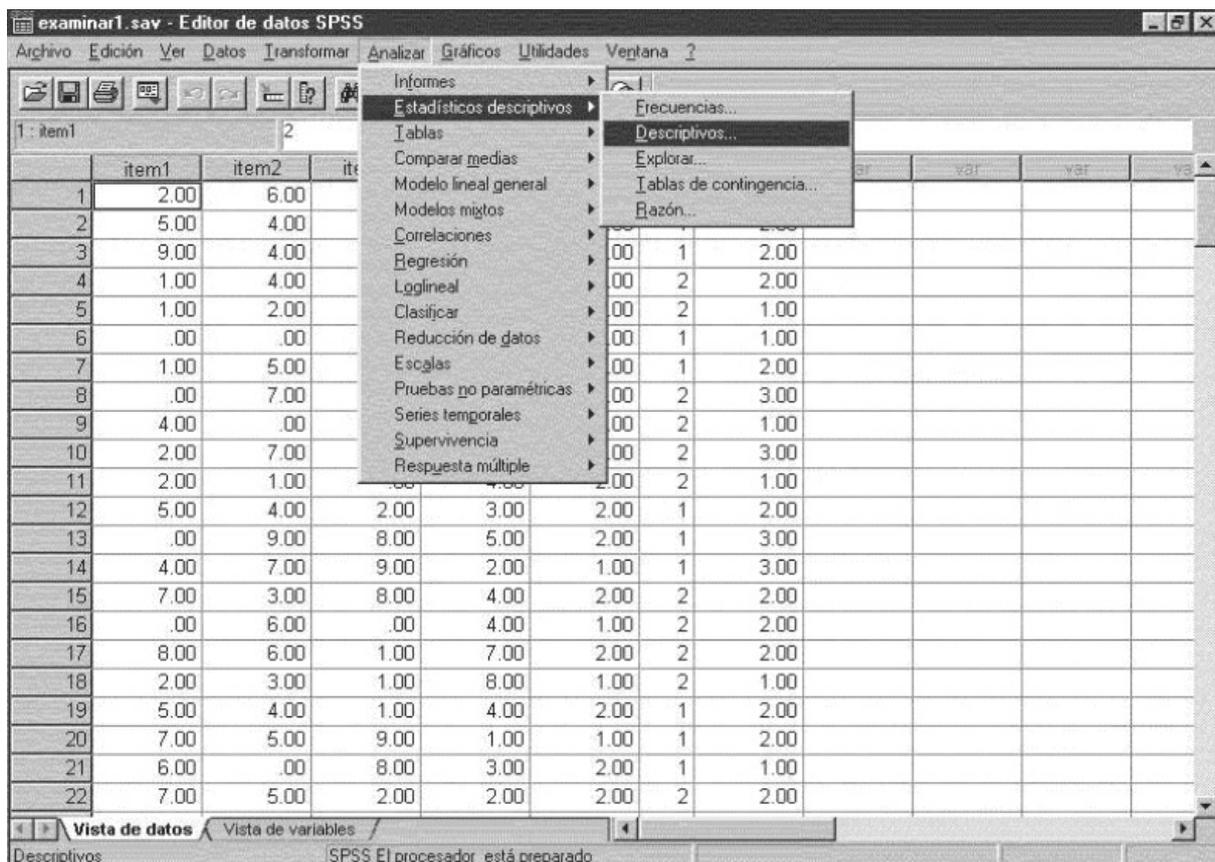
Supongamos el siguiente *problema-ejemplo*.

Se pretende examinar el comportamiento de tres ítems: ITEM1 «Prueba de comprensión lectora», ITEM2 «Prueba de aptitud verbal» y ITEM3 «Valoración en una prueba de ingles», resultados de las pruebas realizadas a 150 alumnos de un instituto de enseñanza secundaria, medidos en una escala de 0 a 10 puntos. También tenemos los datos de la variable NIVEL:

- 1 «Primarios sin c. escolaridad»
- 2 «Certificado escolaridad»
- 3 «Graduado escolar»
- 4 «FP-I»
- 5 «BUP/COU»
- 6 «FP-II»
- 7 «Titulado grado medio»
- 8 «Titulado grado superior»

#### *Desarrollo del ejemplo*

Si utilizamos el SPSS para realizar el problema-ejemplo propuesto la secuencia de ventanas del menú será las siguientes: **Analizar > Estadísticos descriptivos > Explorar.**



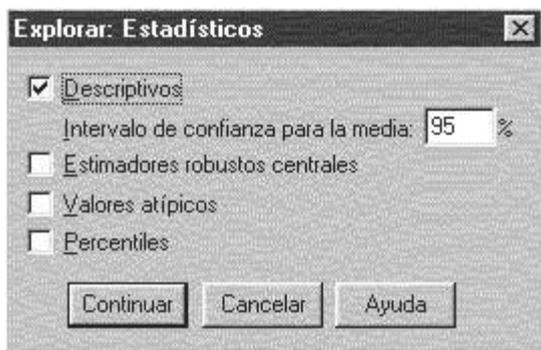
Que abre la ventana siguiente:



## VARIABLES:

Permite seleccionar las variables dependientes e independientes o factores. Además, se puede utilizar una variable de cadena para etiquetar los casos atípicos en los diagramas de caja.

## ESTADÍSTICOS:



**Descriptivos:** Media, mediana, moda, media recortada el 5%, error típico, varianza, desviación típica, mínimo, máximo, rango, rango intercuartílico, y coeficiente de asimetría y curtosis con sus errores típicos. Intervalo de confianza para la media al 95% aunque se puede especificar otro valor.

### Estimadores robustos:

Proporciona estimadores robustos de la media y la mediana muestral.

**HUBER** M-estimador de Huber.

**ANDREW** Estimador de onda de Andrew.

**HAMPEL** M-descendente de Hampel.

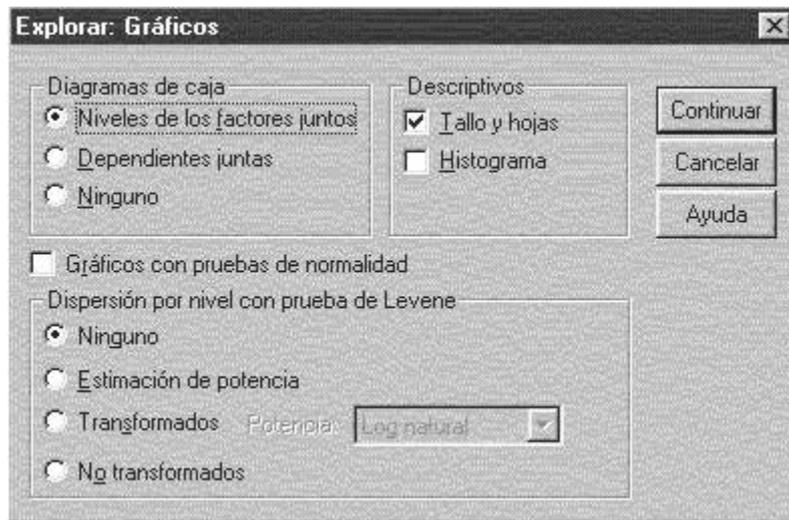
**TUKEY** Estimador bponderado de Tukey.

## Valores

**atípicos:** Muestra los cinco valores mayores y los cinco menores, con las etiquetas de caso.

**Percentiles:** Permite el cálculo de percentiles. Por defecto muestra los percentiles: 5, 10, 25, 50, 75, 90 y 95.

## GRÁFICOS:



Proporciona distintos tipos de gráficos para el análisis exploratorio de datos. **Diagrama de caja:**

### Descriptivos:

#### Gráficos con pruebas de normalidad:

Presenta los gráficos de caja (BOXPLOT). Existen diferentes alternativas para controlar los mismos cuando existen más de una variable dependiente. *Niveles de los factores juntos* genera una presentación para cada variable dependiente. *Dependientes juntas* genera una presentación para grupo definido por una variable o factor.

Presenta los gráficos de tallo y hojas (STEAMLEAF) y los histogramas.

Muestra gráficos de probabilidad normal y de probabilidad normal sin tendencia y estadísticos de Shapiro-Wilks y de Kolmogorov-Smirnov para contrastar la normalidad (NPLOT).

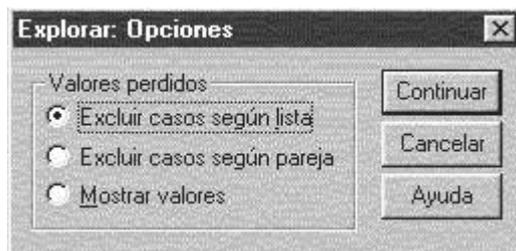
**Dispersión por nivel:** Gráfico de nivel y dispersión,  $p$  es el poder de transformación; si

no se especifica, será estimado a partir de los datos. Además, proporciona el estadístico de Levene para contrastar la homogeneidad de

varianzas (SPREADLEVEL(p)), todos los gráficos disponibles (ALL) y ninguno de los gráficos disponibles(NONE). Si no selecciona ninguna variable de factor, no se generará ningún diagrama de dispersión por nivel.

### **OPCIONES:**

Como en otros comandos de SPSS, el subcomando MISSING controla el tratamiento de los valores omitidos. Tiene cuatro opciones de las cuales las tres primeras se pueden seleccionar en el cuadro de diálogo siguiente, la cuarta se debe hacer por comando.



**Excluir casos según lista (LISTWISE)**

**Excluir casos según pareja (PAIRWISE)**

**Mostrar valores (REPORT)**

**INCLUDE**

Excluye los casos con valor omitido en cualquiera de las variables de la lista.

En cada análisis se incluyen los casos con valor válido para la variable dependiente.

Los valores omitidos para las variables independientes se consideran un grupo más.

Considera los valores omitidos por el usuario como valores válidos.

Una vez seleccionadas las variables dependientes (item1 e item2) y la independiente o factor (nivel) y marcada la opción de ambos (gráficos y estadísticos), al dar a pegar aparecerá los siguientes comandos:

*Fichero de sintaxis*

```
EXAMINE  
VARIABLES=item1 item2 BY nivel  
/PLOT BOXPLOT STEMLEAF  
/COMPARE GROUP  
/STATISTICS DESCRIPTIVES  
/CINTERVAL 95  
/MISSING LISTWISE
```

/NOTOTAL.

Que al ejecutarse genera los resultados siguientes:

*Resultados del comando EXAMINE para la variable item1* **Comprensión lectora: Gráficos de tallo y hojas** — comprensión lectora Stem-and-Leaf Plot for NIVEL = primarios sin c.escolaridad

Frequency Stem & Leaf

5.00 0.00001

4.00 0.2233

1.00 0.4

4.00 0.6777

2.00 0.89 Stem width: 10.00 Each leaf: 1 case(s)

— comprensión lectora Stem-and-Leaf Plot for NIVEL = certificado escolaridad

Frequency Stem & Leaf

3.00 0.111

2.00 0.33

4.00 0.4445

4.00 0.6777

2.00 0.89 Stem width: 10.00 Each leaf: 1 case(s)

— comprensión lectora Stem-and-Leaf Plot for NIVEL = graduado escolar

Frequency Stem & Leaf

6.00 0.000111

4.00 0.2223

2.00 0.55

5.00 0.66667

4.00 0.8888 Stem width: 10.00 Each leaf: 1 case(s)

— comprensión lectora Stem-and-Leaf Plot for NIVEL = fp-I

Frequency Stem & Leaf

5.00 0.00000

1.00 1.0

4.00 2.0000 4.00 3.0000 3.00 4.000 2.00 5.00 2.00 6.00 3.00 7.000 6.00

8.000000 2.00 9.00 Stem width: 1.00

Each leaf: 1 case(s)

— comprensión lectora Stem-and-Leaf Plot for NIVEL = bup/cou

Frequency Stem & Leaf

5.00 0.00111  
2.00 0.33  
9.00 0.444444455  
4.00 0.6677  
2.00 0.99 Stem width: 10.00 Each leaf: 1 case(s)

— comprensión lectora Stem-and-Leaf Plot for NIVEL = fp-II

Frequency Stem & Leaf

4.00 0.0011  
2.00 0.23  
4.00 0.4445  
6.00 0.666777  
3.00 0.899 Stem width: 10.00 Each leaf: 1 case(s)

— comprensión lectora Stem-and-Leaf Plot for NIVEL = titulo de grado medio

Frequency Stem & Leaf

3.00 0.001  
4.00 0.2233  
4.00 0.4445  
1.00 0.7  
2.00 0.88 Stem width: 10.00 Each leaf: 1 case(s) — comprensión lectora  
Stem-and-Leaf Plot for NIVEL = titulo de grado superior

Frequency Stem & Leaf

2.00 1.00  
4.00 2.0000  
3.00 3.000  
1.00 4.0  
1.00 Extremes (> = 7.0)

Stem width: 1.00  
Each leaf: 1 case(s)

**Diagrama de caja**

10  
8 \*59 6  
4  
2  
0  
-2  
N= 16 15 21 32 22 19 14 11

**nivel**

**Resultados descriptivos:**

DESCRIPTIVOS

**Nivel Estadístico Error tipo**

Comprensión lectora

Título de grado superior

Media 2.7273 .5062

Intervalo de confianza para la media al 95% Límite inferior 1.5995

Media recortada al 5% Mediana

Varianza

Desv. típ.

Mínimo

Máximo

Rango

Amplitud intercuartil Asimetría

Curtosis

Límite superior 3.8551

2.5859

2.0000

2.818

1.6787

1.00

7.00

6.00

1.0000

1.764 .661 4.025 1.279

Aptitud verbal Título de grado superior

Media 5.1818 1.0516

Intervalo de confianza para la media al 95% Límite inferior 2.8388

Media recortada al 5% Mediana

Varianza

Desv. típ.

Mínimo

Máximo

Rango

Amplitud intercuartil Asimetría

Curtosis

Límite superior 7.5248

5.2576

7.0000

12.164

3.4876

.00

9.00

9.00

6.0000

-.445 .661

-1.512 1.279

Como vemos el procedimiento EXAMINE proporcionará los valores de los estadísticos descriptivos: media, mediana, moda, media ajustada al 5%, error típico de la media, varianza, mínimo, máximo, rango, rango intercuartílico, coeficientes de asimetría y curtosis y errores típicos asociados. Además de los estadísticos descriptivos (sólo se ha mostrado el nivel de titulado-superior), por cada variable dependiente se obtendrán los diagrama de tallo y hojas y los diagramas de caja.

### **9.3. La distribución de frecuencias. FRECUENCIAS**

Proporciona tablas de frecuencias y, opcionalmente, estadísticos descriptivos básicos, gráficos de barras e histogramas. Sirve como punto de partida en cualquier análisis posterior y sobre todo para localizar valores de las variables fuera del rango de medida, cuestión que aparece con frecuencia en los casos de error en la introducción de datos. Es un comando de utilidad para variables categóricas (nivel de medida nominal u ordinal) o categorizadas.

Se pueden organizar los valores de las variables en orden ascendente o descendente e incluso se pueden ordenar las categorías por sus frecuencias. Vale suprimir la tabla de frecuencias de una variable cuando ésta tiene gran cantidad de categorías.

#### *Problema-ejemplo*

Supongamos el siguiente *problema-ejemplo*.

Se pretende saber la distribución de frecuencia de la variable NIVEL, del ejemplo anterior del comando EXPLORAR, que como sabemos son los resultados de las pruebas realizadas a 150 alumnos de un instituto de enseñanza secundaria. La variable NIVEL tiene las siguientes categorías:

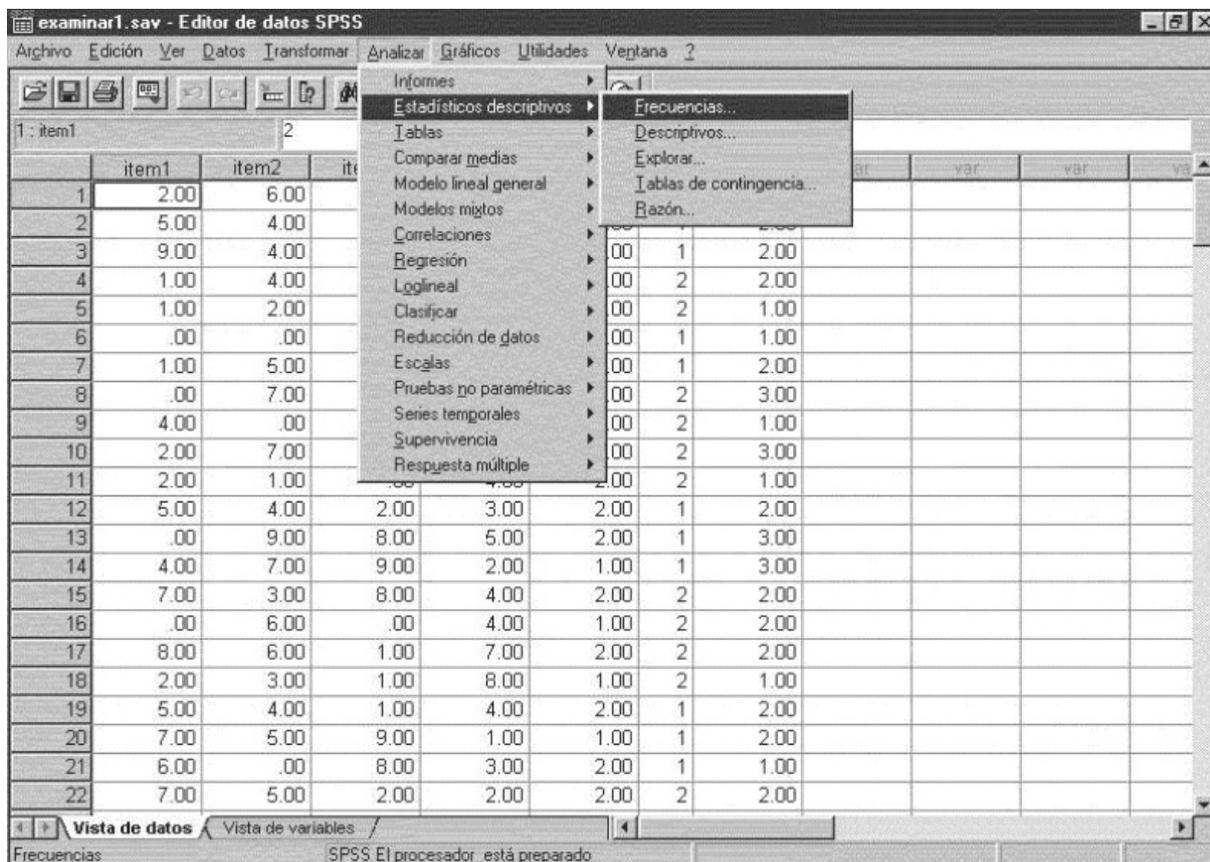
1 «Primarios sin c. escolaridad»

- 2 «Certificado escolaridad»
- 3 «Graduado escolar»
- 4 «FP-I»
- 5 «BUP/COU»
- 6 «FP-II»
- 7 «Titulado grado medio»
- 8 «Titulado grado superior»

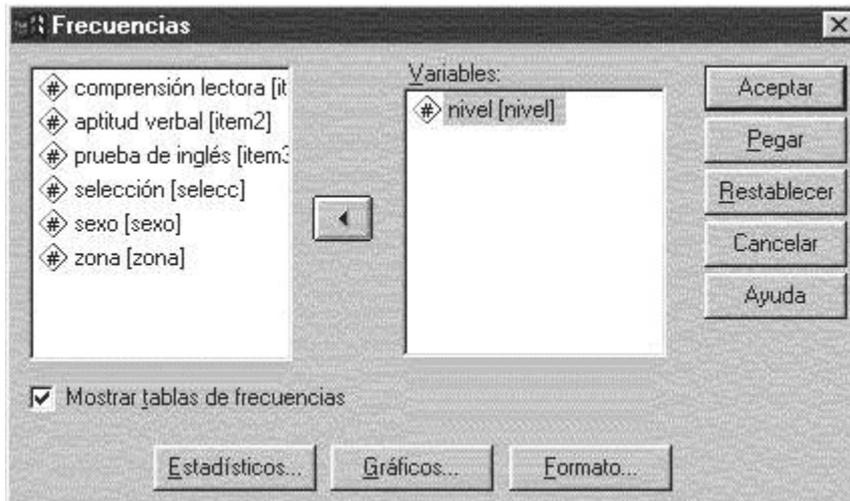
*Desarrollo del ejemplo*

Si utilizamos el SPSS para realizar el problema-ejemplo propuesto la secuencia de ventanas del menú será la siguiente: **Analizar > Estadísticos descriptivos > Frecuencias.**

Y aparecerá la siguiente ventana:



Cuando se pulse **intro** se entrará en la siguiente pantalla que permite seleccionar la variable o variables a estudiar. En nuestro caso se ha seleccionado NIVEL.



Como se observa, se puede seleccionar tres tipos de menú: estadísticos, gráfico y formato. Si seleccionamos estadísticos entraremos en el siguiente menú:



### **ESTADÍSTICOS: Valores percentiles:**

**Tendencia central:**

**Dispersión:**

**Distribución:**

Se puede seleccionar: *cuartiles* (percentiles 25, 50 y 75) dividen las observaciones en cuatro grupos de igual tamaño. *Puntos de corte para... grupos iguales*, donde proporciona los valores que dividen la muestra en n grupos con el mismo número de casos. *Percentiles*, se van añadiendo los percentiles que se consideran convenientes en el análisis.

Proporciona estadísticos descriptivos de posición: **MEAN** Media.

**MEDIAN** Mediana.

**MODE** Moda.

**SUM** Suma.

Proporciona estadísticos que miden la dispersión o variación de los datos:

**STDDEV** Desviación Típica.

**VARIANCE** Varianza.

**RANGE** Rango o amplitud.

**MINIMUM** Mínimo.

**MAXIMUM** Máximo.

**SEMEAN** Error típico de la media.

Para describir la forma y simetría de la distribución. **SKEWNESS**  
**KURTOSIS**

Coeficiente de asimetría. Coeficiente de curtosis.

Por comando podemos especificar las opciones:

**DEFAULT** Es equivalente a especificar MEAN, STDDEV, MINIMUM y MAXIMUM.

**ALL** Todos los estadísticos disponibles.

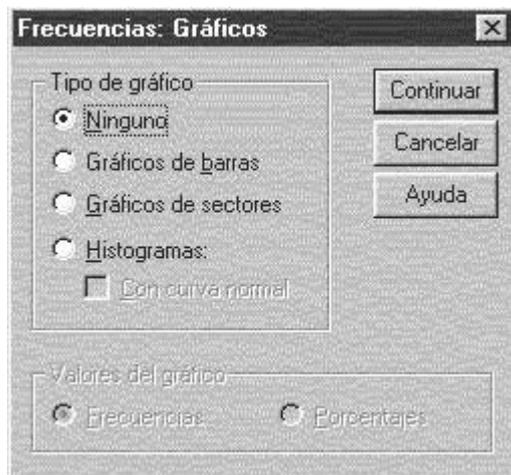
**NONE** Ninguno de los estadísticos disponibles.

**Los valores son puntos medios de grupos (grouped).** Permite indicar que los valores de las variables listadas corresponden a marcas de clase o puntos medios de intervalos de la misma longitud (valores agrupados), de tal forma que dicha información se considerará en el cálculo de los n-tiles y, en particular, de los percentiles y de la mediana. Si se especifica un único valor a continuación del nombre de una variable, dicho valor indicará la longitud de los intervalos y, si se especifica más de uno, los valores listados se considerarán como las marcas de clase (en lugar de los valores de la variable).

En el ejemplo propuesto hemos seleccionado los cuartiles, la mediana y la moda al tratarse de una variable ordinal.

**GRÁFICOS:**

Se puede seleccionar los siguientes gráficos:



**Tipo de gráfico** Permite seleccionar los gráficos de barras (**BARCHART**), histogramas (**HISTOGRAM**), con la inclusión de la curva normal superpuesta si así se desea. También se puede incluir los gráficos de sectores.

**Valores del gráfico** En los gráficos de barras o de sectores, se puede etiquetar el eje de escala con las frecuencias o los porcentajes.

Por comando se puede seleccionar:

**MINIMUM(n)** En los histogramas, n será el límite inferior del primer intervalo y, en los gráficos de barras, el mínimo valor.

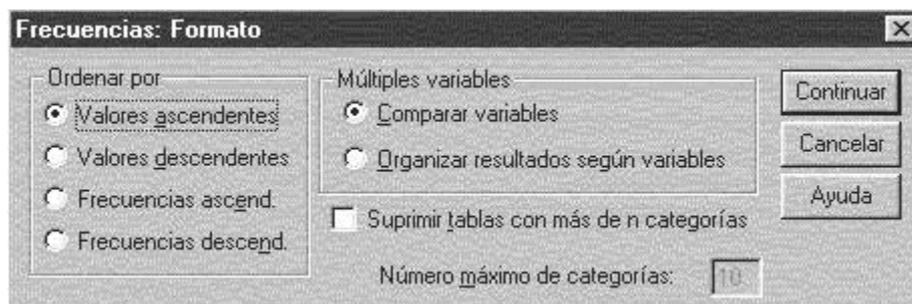
**MAXIMUM(n)** En los histogramas, n será el límite superior del último intervalo y, en los gráficos de barras, el máximo valor.

**FREQ(n)** El eje horizontal contendrá frecuencias; n es opcional e indica la máxima frecuencia que se ha de representar en un intervalo.

**PERCENT(n)** El eje horizontal contendrá porcentajes; n es opcional e indica el máximo porcentaje que se ha de representar en un intervalo.

En el ejemplo propuesto se ha seleccionado, el gráfico de barras y con valores dados en porcentajes.

Finalmente, se puede seleccionar el **FORMATO** entrando en el siguiente menú: **FORMATO**:



Permite controlar el formato de la tabla de frecuencias. **Ordenar por**

## Múltiples variables

Permite ordenar por los valores de los datos o respecto al recuento (frecuencia de aparición de la categoría de la variable).

Cuando se desea generar tablas estadísticas para múltiples variables, se puede mostrar las variables en una sola tabla (*comparar variables*) o bien en tablas independientes (*organizar resultados según variables*).

**Suprimir tablas con más de n categorías** Impide mostrar tablas con más de un número especificado de categorías.

Finalmente por comando, mediante el subcomando MISSING, se puede controlar el tratamiento de los valores omitidos. Tiene la opción:

**INCLUDE** Los valores omitidos por el usuario serán considerados como valores válidos.

Con los descriptivos y gráficos marcados en el ejemplo y la opción de formato dada por defecto, al dar a pegar aparecerán los siguientes comandos: *Fichero de sintaxis*

```
FREQUENCIES  
VARIABLES=nivel  
/NTILES= 4  
/STATISTICS=MEDIAN MODE  
/BARCHART PERCENT  
/ORDER= ANALYSIS .
```

Que al ejecutarse marcando los comandos y subcomandos tenemos los resultados siguientes:

*Resultados del comando frequencies para la variable nivel* **Resultados descriptivos:**

ESTADÍSTICOS

**Nivel**

N

Mediana Moda

Percentiles Válidos 150 Perdidos 0

4.0000

4.00

25 3.0000

50 4.0000

75 6.0000

NIVEL

Frecuencia Porcentaje Porcentaje válido acumulado

Válidos Primarios sin c. escolaridad Certificado escolaridad Graduado escolar

FP-I

BUP/COU

FP-II

Título de grado medio Título de grado superior

*Total*

16 10.7

15 10.0

21 14.0

32 21.3

22 14.7

19 12.7

14 9.3

11 7.3

150 100.0 10.7 10.7 10.0 20.7 14.0 34.7 21.3 56.0 14.7 70.7 12.7 83.3

9.3 92.7 7.3 100.0

100.0

**Gráficos:**

NIVEL 30

20 21

14

15 13 10<sub>11</sub> 10 9

7

0

**nivel**

#### **9.4. Descriptivos. La opción DESCRIPTIVOS**

Proporciona estadísticos descriptivos univariantes adecuados para variables numéricas. Además permite crear nuevas variables con los valores de variables existentes tipificados (z). Las variables se pueden ordenar alfabéticamente, según el valor de su media u otros estadísticos y conforme son referenciadas para el análisis (opción por defecto).

*Problema-ejemplo*

Supongamos el siguiente *problema-ejemplo*.

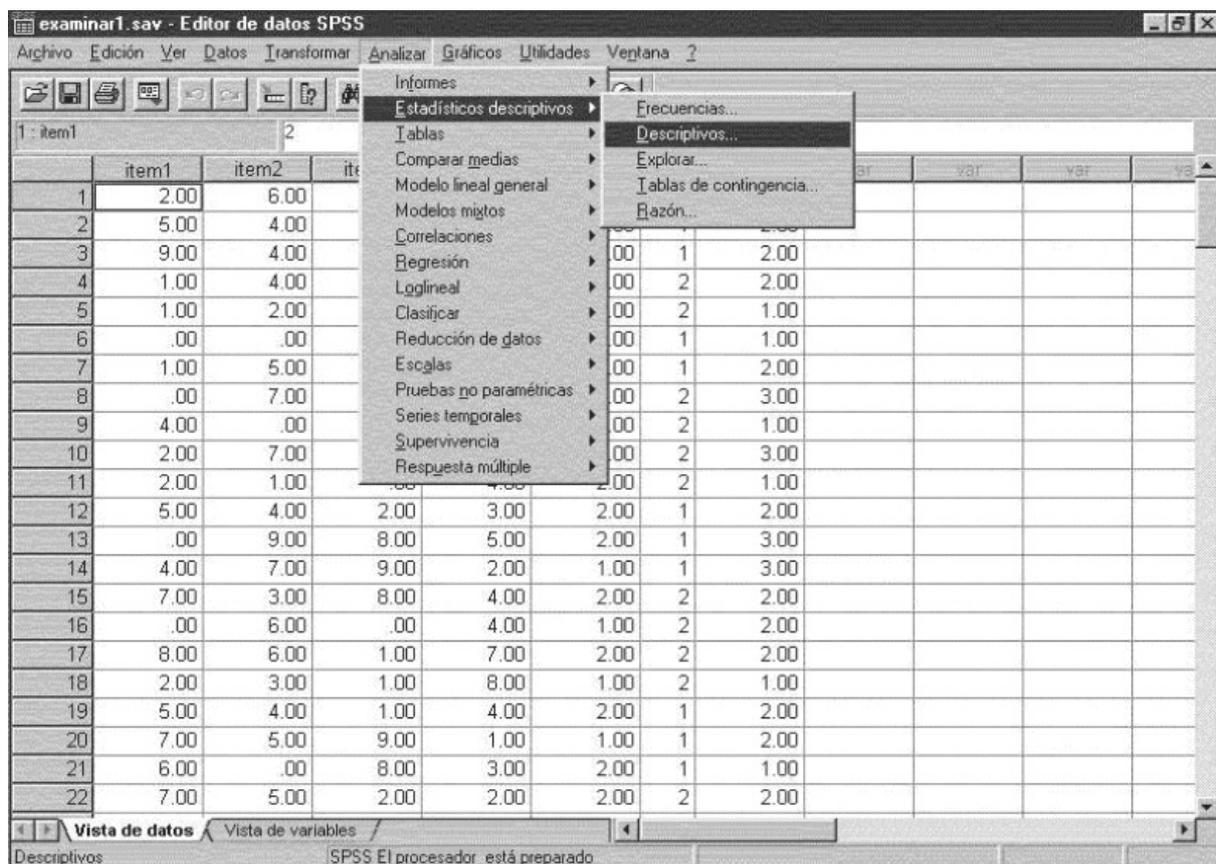
Se pretende saber los descriptivos de los ITEM1, ITEM2 y ITEM3, del ejemplo que venimos utilizando, que como sabemos son los resultados de

las pruebas realizadas a 150 alumnos de un instituto de enseñanza secundaria. Además se pide que se salve las puntuaciones z de estas tres variables

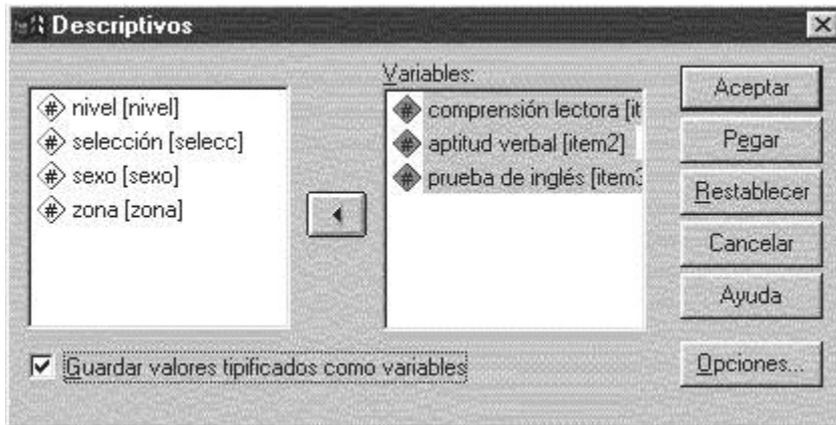
### Desarrollo del ejemplo

Si utilizamos el SPSS para realizar el problema-ejemplo propuesto la secuencia de ventanas del menú será las siguientes: **Analizar > Estadísticos descriptivos > Descriptivos**.

Y aparecerá la siguiente ventana:



Al dar la orden descriptivos se entra en el siguiente menú:



Aquí se selecciona las **variables** sobre las que se realizará el análisis. En el ejemplo propuesto se ha seleccionado: item1, item2 e item3. También se ha marcado «**Guardar valores tipificados como variables**» (puntuaciones z), conforme se había planteado en el problema.

En esta ventana tenemos acceso al menú de OPCIONES que permite seleccionar estadísticos y orden de visualización.

### OPCIONES:



Como estadísticos tenemos las siguientes opciones: Por defecto (media, desv. típica, mínimo y máximo). **Media y suma** **Dispersión**

### Distribución

### Orden de visualización

Se muestra estos dos estadísticos

Se puede seleccionar: desv. típica (STDDEV), Mínimo (MINIMUM), Máximo (MAXIMUM), Error típico de la media (SEMEAN), Varianza

(VARIANCE) y amplitud o rango (RANGE).

Se muestran los coeficientes de asimetría (SKEWNESS) y curtosis (KURTOSIS) y sus errores típicos.

Se puede seleccionar el orden de visualización de las variables:

*Lista de variables.* Según la lista de variables. *Alfabético.* Por orden alfabético.

*Medias ascendentes.* Por medias ascendentes. *Medias descendentes.* Por medias descendentes. Además, por comando, mediante el subcomando SORT se puede ordenar por:

**STDDEV**  
**MINIMUM MAXIMUM SEMEAN**  
**VARIANCE SKEWNESS RANGE**  
**SUM**

Desviación Típica.

Mínimo.

Máximo.

Error típico de la media. Varianza.

Coefficiente de asimetría. Rango.

Suma.

Y también por comando, mediante el subcomando MISSING se puede controlar el tratamiento de los valores omitidos. Con:

**INCLUDE** Los valores omitidos por el usuario serán considerados como valores válidos.

En el ejemplo propuesto se ha seleccionado el cálculo de la media, desviación típica, el coeficiente de asimetría y el de curtosis. Además, se ha pedido que se visualizara según el orden ascendente de sus medias y que se salvara las variables tipificadas (z). Al dar la opción de **pegar** aparecerá los siguientes comandos:

*Fichero de sintaxis*

```
DESCRIPTIVES  
VARIABLES=item1 item2 item3  
/SAVE  
/STATISTICS=MEAN STDDEV KURTOSIS SKEWNESS  
/SORT=MEAN (A) .
```

*Resultados del comando descriptives para las variables item1,item2 y item3*

N Estadístico

Media

Desv. típ. Asimetría

Curtosis Estadístico Estadístico

Estadístico Error típico Estadístico Error típico ESTADÍSTICOS  
DESCRIPTIVOS

Comprensión lectora

150

4.1133 2.7963

.127

.198

-1.194

.394

Aptitud Prueba N válido verbal de inglés (según lista)

150 150 150 4.2400 4.4733

2.9440 3.0360

.041 -.032

.198 .198

-1.302 -1.323

.394 .394

Además de los resultados presentados, en el fichero de datos tendremos tres nuevas variables: zitem1, zitem2 y zitem3, con los valores tipificados de las variables item1, item2 e item3. Cuando salvemos el fichero de datos quedarán definitivamente incorporadas al mismo.

## 9.5. Los comandos CORRELATIONS y NOPAR CORR

Proporciona matrices de correlaciones. Calcula el coeficiente de correlación de Pearson (comando CORRELATIONS), la r de Spearman y la tau-b de Kendall (comando NOPAR CORR) con sus niveles de significación. Es interesante, antes de realizar los cálculos, eliminar los valores atípicos porque pueden producir errores. Además conviene recordar que el coeficiente de correlación de Pearson asume que cada pareja de variables es normal bivariada.

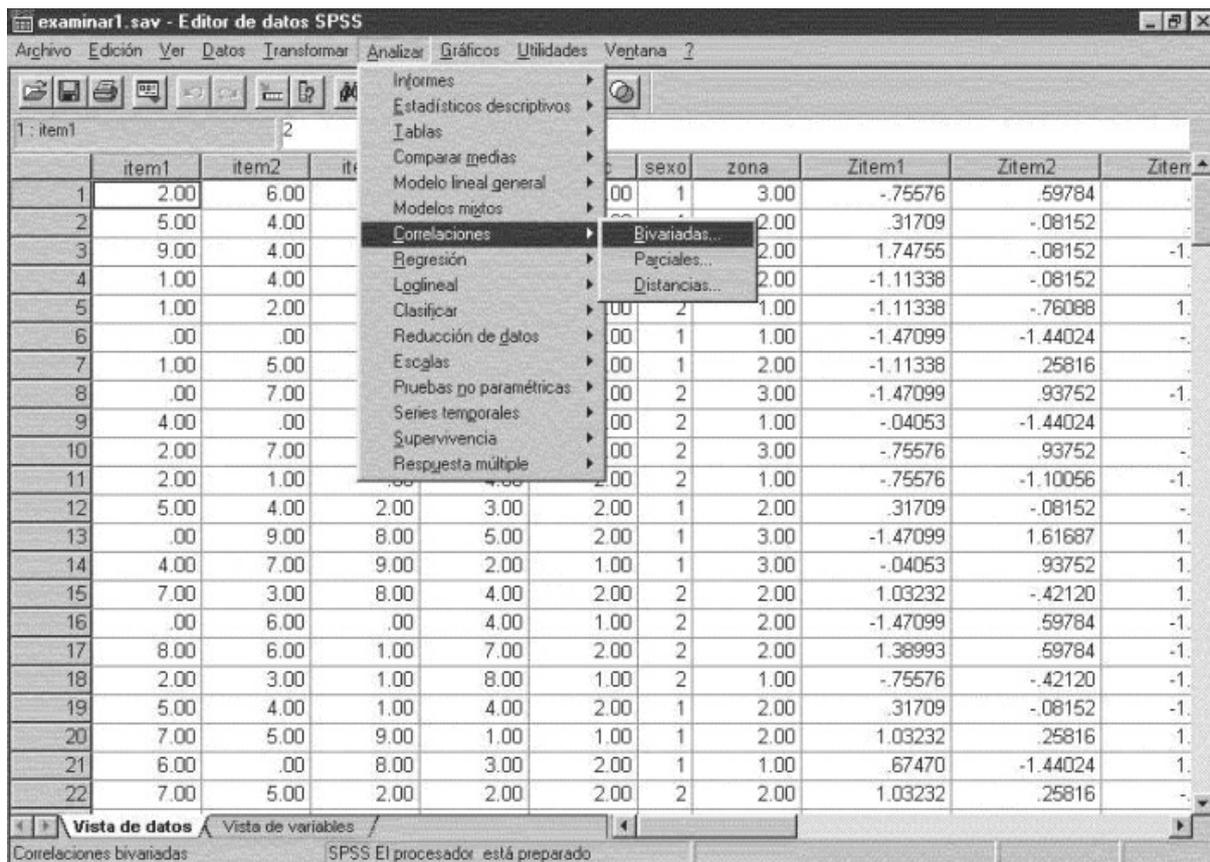
*Problema-ejemplo*

Supongamos el siguiente *problema-ejemplo*.

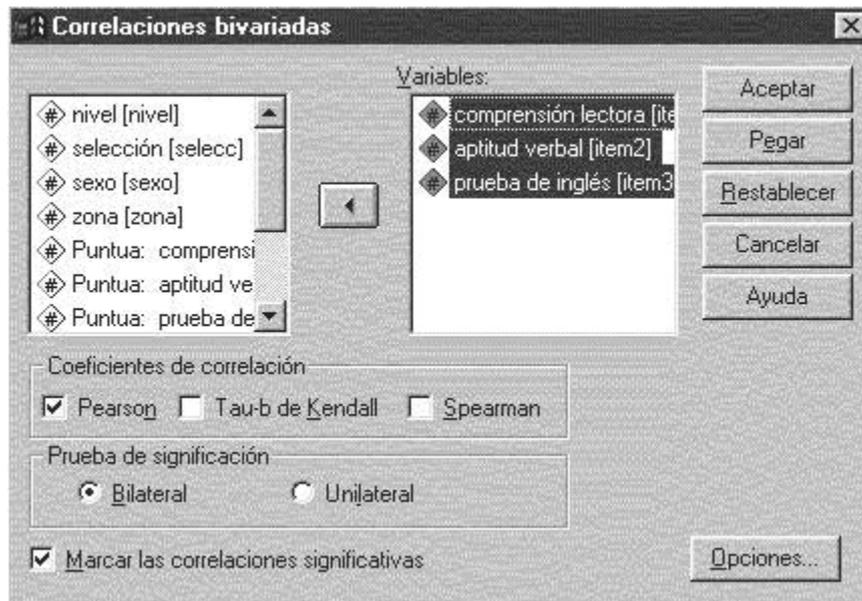
Se pretende saber la correlación entre los item1, item2 y item3, del ejemplo que se viene utilizando, que como sabemos son los resultados de las pruebas realizadas a 150 alumnos de un instituto de enseñanza secundaria.

### Desarrollo del ejemplo

Si utilizamos el SPSS para realizar el problema-ejemplo propuesto la secuencia de ventanas del menú será las siguientes: **Analizar >Correlaciones> Bivariadas**. Y aparecerá la siguiente ventana:



Donde se seleccionará las variables conforme al menú:



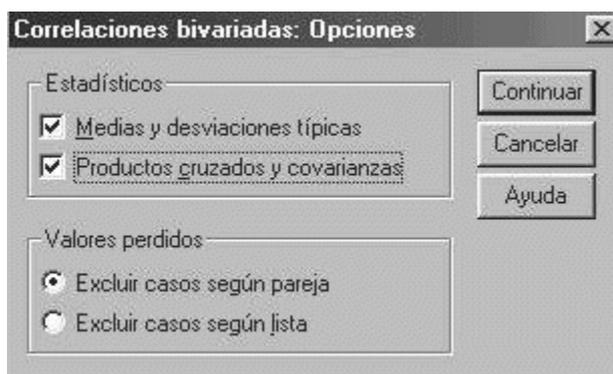
En el ejemplo propuesto se ha seleccionado los tres ítems. Además permite seleccionar:

**Coeficientes de correlación** . Indicar el coeficiente de correlación que se debe utilizar: correlación de Pearson para variables cuantitativas en escala de intervalo y suficientemente simétricas (se asume que la pareja de variables son normal bivariada), r de Spearman para variables cuantitativas ordinales o en escala de intervalo con distribución no simétrica y tau-b de Kendall para ordinales.

**Prueba de significación** . Se puede seleccionar las probabilidades unilaterales o bilaterales de la asociación entre las variables. Por defecto la opción señalada es bilateral al no conocer la dirección de dicha asociación.

**Marcar las correlaciones significativas.** Opción dada por defecto. La significación al nivel del 5% se identifica por medio de un asterisco y al 1% con dos asteriscos. También en esta pantalla se puede marcar OPCIONES con lo cual se entra en la siguiente pantalla:

**OPCIONES:**



**Estadísticos** Para la correlación de Pearson, se puede elegir las siguientes opciones:

**Medias y desviaciones típicas** Se muestran para cada variable.

**Productos cruzados y covarianzas** Se muestra para cada pareja de variables. El producto cruzado es el numerador del coeficiente de correlación de Pearson y la covarianza una medida de relación entre dos variables.

**Valores perdidos** Tiene las opciones:

**Excluir casos según pareja** Se excluyen los casos con valores perdidos en alguna de las variables de la pareja. **Excluir casos según lista** Excluye los casos que tiene algún valor perdido en cualquier variable de la lista seleccionada para el análisis. Aunque representa pérdida de información esta opción es aconsejable cuando se desea utilizar la matriz de correlaciones en procesos que requieran inversión de la misma.

Además, por comando, se puede en éste y en el comando NONPAR CORR:

— Salvar la matriz de correlaciones de Pearson (MATRIX=OUT(fichero)) para utilizarse como entrada en otros procesos, como por ejemplo en el análisis factorial.

— Obtener correlaciones de una lista de variables con todas las variables de otra segunda lista (se utiliza WITH en el subcomando VARIABLES).

En el ejemplo, se ha seleccionado el cálculo de coeficiente de correlación de Pearson, el cálculo de las medias y desviaciones típicas y los productos cruzados. Asimismo se ha excluido los casos perdidos según parejas de variables.

*Fichero de sintaxis*

```
CORRELATIONS
/VARIABLES=item1 item2 item3
/PRINT=TWOTAIL NOSIG
/STATISTICS DESCRIPTIVES XPROD
/MISSING=PAIRWISE
```

Ejecutando el fichero de sintaxis tenemos los siguientes resultados:

*Resultados del comando correlations para los item1, item2 y item3*

ESTADÍSTICOS DESCRIPTIVOS

**Media Desviación típica N** Comprensión lectora 4.1133 2.79630 150

Aptitud verbal 4.2400 2.94395 150  
Prueba de inglés 4.4733 3.03602 150

comprensión lectora  
aptitud verbal  
prueba

de inglés CORRELACIONES

**comprensión aptitud prueba lectora verbal de inglés**

Correlación de Pearson Sig. (bilateral)

Suma de cuadrados y productos cruzados Covarianza

N

1 -.086 -.035

.293 .672

1165.073 -106.080 -44.047 7.819 -.712 -.296 150 150 150

Correlación de Pearson Sig. (bilateral)

Suma de cuadrados y productos cruzados Covarianza

N

-.086 1 -.049 .293 .553

-106.080 1291.360 -65.040

-.712 8.667 -.437 150 150 150

Correlación de Pearson Sig. (bilateral)

Suma de cuadrados y productos cruzados Covarianza

N

-.035 -.049 1 .672 .553

-44.047 -65.040 1373.393

-.296 -.437 9.217 150 150 150

## **9.6. Análisis descriptivo de datos: GRÁFICOS**

### **9.6.1. Tipos de gráficos**

Además de los gráficos presentes en los comandos **Frecuencias** (diagrama de barra o histograma) y **Explorar** (diagrama de caja, de tallo y hojas e histograma), SPSS tiene un menú específico para crear un conjunto de gráficos para representar una o más variables. Para realizar dichos gráficos debemos seleccionar **Gráficos** en el menú del programa.

### **9.6.2. Los gráficos de barras, líneas, áreas y sectores**

En general, estos cuatro tipo de gráficos sirven para representar gráficamente la distribución de frecuencias de las variables (ver fig. 2.9.1 a

2.9.4).

Los datos que pueden manejar estos cuatro tipos de gráficos son: — Si seleccionamos **gráficos de barra simple, gráficos de líneas simples, gráficos de áreas simples o sectores**, tenemos las siguientes opciones:

*A. Resúmenes para grupos de casos*

Se representarán las frecuencias de las categorías de una variable.

*B. Resúmenes para distintas variables*

Se representarán las frecuencias de las categorías de dos o más variables. Cada barra, línea, área o sector, representa una de las variables.

*C. Valores individuales de los casos*

Se representarán las frecuencias de las categorías de una sola variable. Cada barra, línea, área o sector, representa un caso individual. — Si seleccionamos **gráficos de barra agrupado o apilado, gráficos de líneas múltiples o verticales y gráfico de áreas apilado**, tenemos las siguientes opciones:

*A. Resúmenes para grupos de casos*

Las frecuencias de las categorías que toma una variable se juntan con las categorías de otra variable.

*B. Resúmenes para distintas variables*

Se resumen dos o más variables dentro de categorías de otra variable.

*C. Valores individuales de los casos*

Se resumen dos o más variables por cada caso.



FIGURA 2.9.1. Gráfico de barras.



FIGURA 2.9.2. Gráfico de líneas.



FIGURA 2.9.3. Gráfico de áreas.

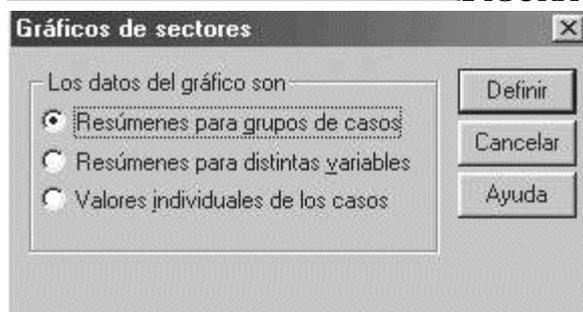


FIGURA 2.9.4. Gráfico de sectores.

**9.6.3. Diagramas de caja y barras de error** Estos dos tipos de gráfico se utilizan en la exploración de los datos (ver fig. 2.9.5 a 2.9.6).



FIGURA 2.9.5. Diagrama de caja.



FIGURA 2.9.6. Barras de error.

— Cuando seleccionamos **diagramas de caja o error simple**, tenemos las siguientes opciones:

*A. Resúmenes para grupos de casos*

Se resume una única variable numérica en categorías de otra variable. Cada cuadro muestra la mediana, los cuartiles y los valores extremos contenidos en una categoría.

*B. Resúmenes para distintas variables*

Se resumen una o más variables numéricas. Cada cuadro representa una de las variables.

— Al seleccionar **diagramas de caja o error agrupado**, tenemos las siguientes opciones:

*A. Resúmenes para grupos de casos*

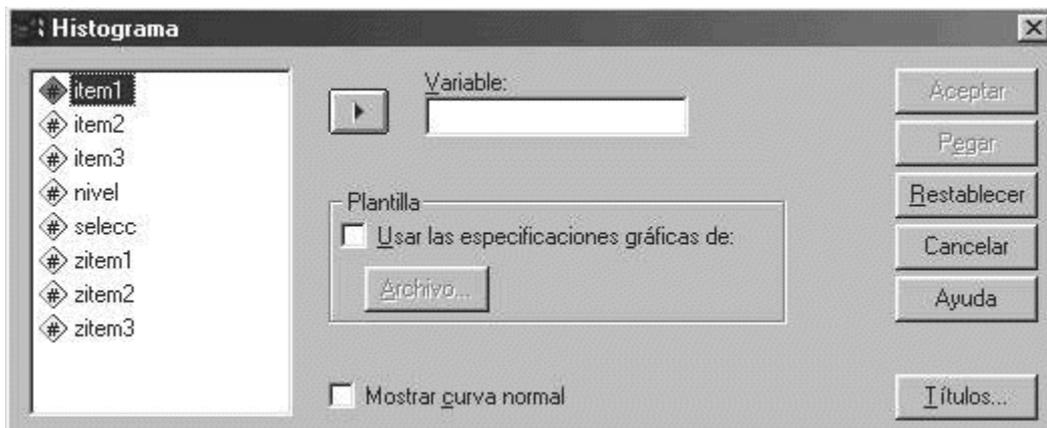
Se resume una única variable numérica en agrupaciones definidas por una variable categórica. Cada cuadro de una agrupación representa una categoría de una segunda variable categórica.

*B. Resúmenes para distintas variables*

Se resumen diferentes variables numéricas. Cada uno de los cuadros de una agrupación representa una de las variables de resumen.

**9.6.4. Histograma** El histograma servirá para mostrar el número de casos que se encuentran en cada intervalo de la variable seleccionada.

Como podemos observar en la fig. 2.9.7 se puede representar el histograma junto con la curva normal.



FIGURA

### 2.9.7. Histograma.

### 9.6.5. Diagrama de dispersión

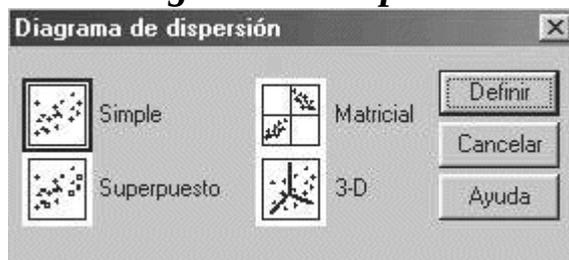


FIGURA 2.9.8. Diagrama de dispersión.

— Al seleccionar **diagrama de dispersión simple**, tenemos las siguientes opciones:

#### A. Diagrama de dispersión simple

Se representa una variable numérica en función de la otra.

B. *Diagrama de dispersión simple con variable de control opcional*  
 Pueden representarse dos variables dentro de las categorías definidas por una tercera variable (marcador). No debe confundirse con un diagrama de dispersión superpuesto.

— Mediante **diagrama de dispersión superpuesto**, se puede:

Representar dos o más pares de variables y-x. Cada par se indica mediante un marcador diferente.

— Al elegir **diagrama de dispersión matricial**, hay que tener presente que:

El número de filas y columnas es igual al número de variables de la matriz seleccionadas. Se muestran todas las combinaciones posibles de variables.

— Cuando escogemos **diagrama de dispersión 3-D**, tenemos: Se representan tres variables en tres dimensiones.

### 9.6.6. Gráficos P-P y Q-Q

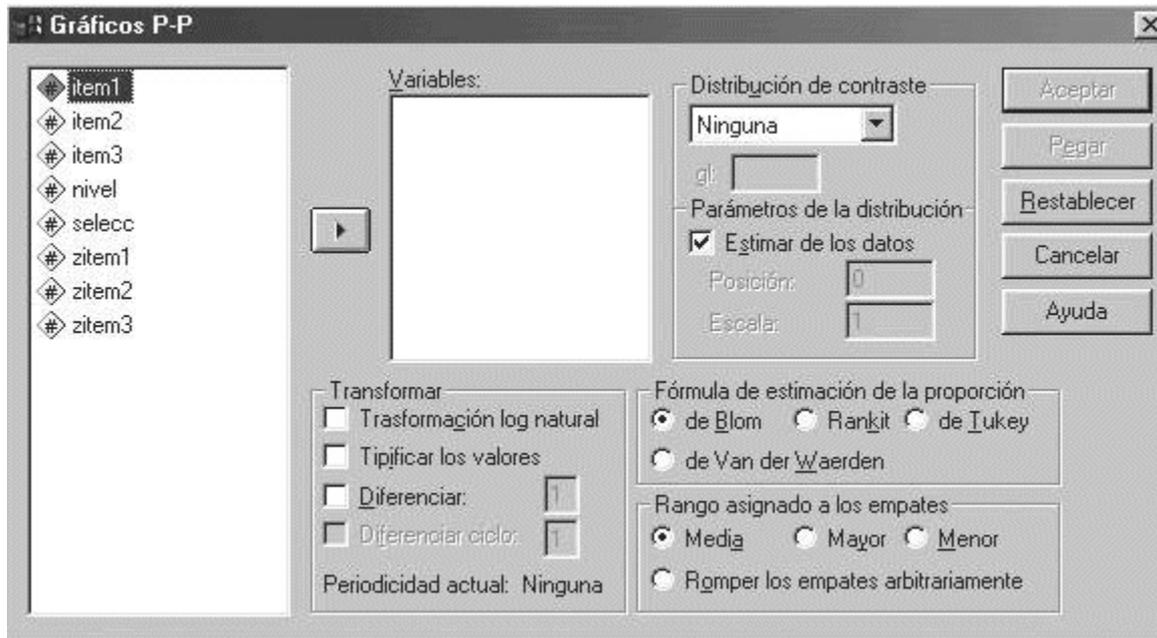


FIGURA 2.9.9. Gráfico P-P.

Crea un gráfico de las proporciones acumuladas o de los cuantiles, de una variable respecto a las/los de una distribución cualquiera de prueba. Estos tipos de gráficos suelen emplearse para determinar si la distribución de una variable coincide con una distribución dada. Si los valores empíricos de la variable seleccionada coinciden con los valores teóricos de la distribución de prueba, los puntos formarán una línea recta.

SPSS tiene entre las distribuciones de prueba disponibles: la beta, chi-cuadrado, exponencial, gamma, semi-normal, Laplace, logístico, Lognormal, normal, Pareto, t de Student, Weibull y uniforme. Según la distribución elegida, se indicará distintos grados de libertad y otros parámetros.

También se puede probar los valores empíricos de distribuciones transformadas. SPSS tiene las siguientes opciones de transformación: el logaritmo natural, tipificar valores, diferencia y diferencia estacional.

Además, se puede especificar el método utilizado para calcular las distribuciones esperadas, así como para resolver «empates», u observaciones múltiples con el mismo valor.

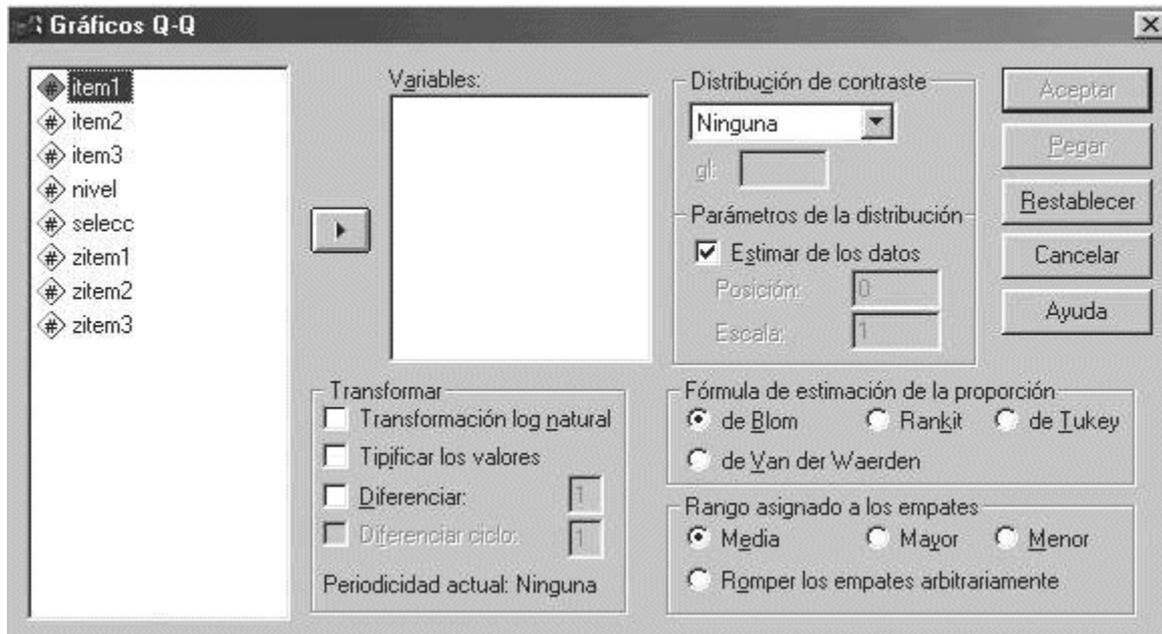


FIGURA 2.9.10. Gráfico Q-Q.

Todos los gráficos indicados tienen la posibilidad de insertarles un título, etiquetas a los ejes, un marco exterior, usar una plantilla gráfica (para crear gráficos similares a uno almacenado), manejar los casos perdidos y exportarlos en varios formatos: metaarchivo de windows (.wmf), metaarchivo cgm (.cgm), postscript (.eps), mapa de bits de windows (.bmp), jpeg (.jpg).

### 9.6.7. Desarrollo de algunos ejemplos

En este apartado se desarrollarán algunos ejemplos de gráficos. Debe quedar claro, que las posibilidades de distintas combinaciones en los gráficos hace inabordable con exhaustividad su representación; no obstante, se intentará abordar los casos más significativos.

#### 9.6.7.1. Gráficos de barras, líneas, áreas y sectores

De estos cuatro tipos de gráficos, mostraremos como ejemplo un gráfico de barras, los otros tres tipos de gráficos se realizarían de forma similar e incluso una vez confeccionado uno de los cuatro sus datos nos pueden servir para probar con otros; así por ejemplo si hemos creado un gráfico de barras, conservando los datos, podemos ver su representación mediante áreas.

#### *Problema-ejemplo*

Supongamos el siguiente *problema-ejemplo*.

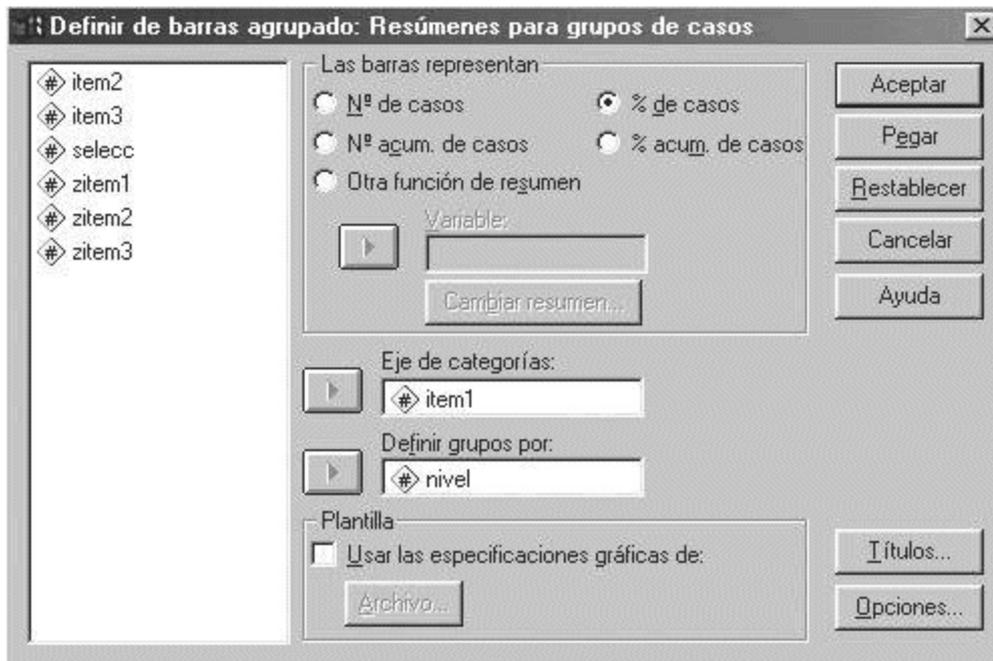
Se pretende representar, mediante un diagrama de barras, la distribución del ITEM1 con valores agrupados por NIVEL, del ejemplo

que venimos utilizando, que como sabemos son los resultados de las pruebas realizadas a 150 alumnos de un instituto de enseñanza secundaria.

### Desarrollo del ejemplo

Si utilizamos SPSS para realizar el problema-ejemplo propuesto la secuencia de ventanas del menú será la siguiente: **Gráficos > Barras > Seleccionar agrupado > Definir** (ver fig. 2.9.1).

Y aparecerá la siguiente ventana:



Como vemos en el eje de categorías está seleccionada ITEM1, los grupos se forman con la variable NIVEL y las barras serán el % de casos. Además pulsando Títulos se inserta un título al gráfico.

### Fichero de sintaxis

#### GRAPH

```
/BAR(GROUPED)=PCT BY item1 BY nivel /MISSING=REPORT  
/TITLE= 'Gráficos de barras del item1'.
```

### Gráfico de resultados

#### Gráficos de barras del item1 40

#### nivel

Primarios sin c. escolaridad 30 Certificado escolaridad

Graduado escolar 20

10

0

BUP/COU  
FP-II

Título de grado medio  
Título de grado superior

00 1.00 2.00 3.00 4.00 5.00 6.00 7.00 8.00 9.00 **comprensión lectora**

### 9.6.7.2. Diagramas de cajas y barras de error

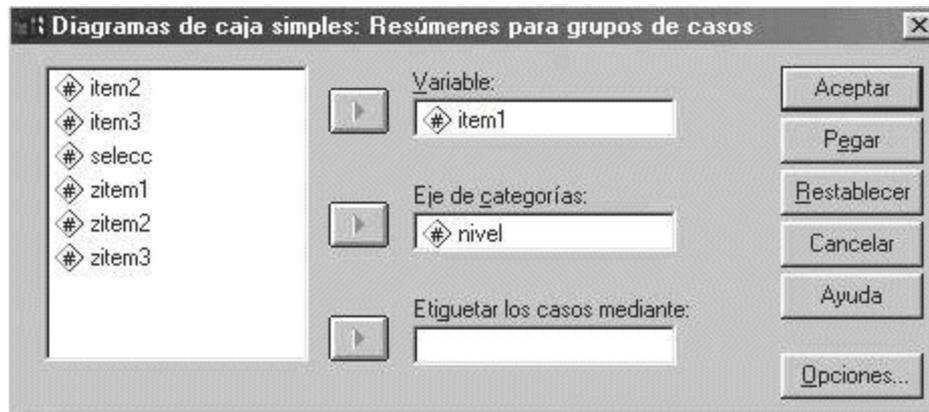
#### *Problema-ejemplo*

Supongamos el siguiente *problema-ejemplo*.

Se pretende representar, mediante un diagrama de cajas, la distribución del ITEM1 con valores agrupados por NIVEL, del ejemplo que venimos utilizando, que como sabemos son los resultados de las pruebas realizadas a 150 alumnos de un instituto de enseñanza secundaria.

#### *Desarrollo del ejemplo*

Si utilizamos SPSS para realizar el problema-ejemplo propuesto la secuencia de ventanas del menú será la siguiente: **Gráficos > Diagrama de caja > Simple > Definir** (ver fig. 2.9.5).



Seleccionaremos la variable ITEM1 en función de las categorías de la variable NIVEL.

#### *Fichero de sintaxis*

```
EXAMINE  
VARIABLES=item1 BY nivel  
/PLOT=BOXPLOT/STATISTICS=NONE/NOTOTAL  
/MISSING=REPORT.
```

#### *Gráfico de resultados*

10  
8  
\*59

6

4

2

0

-2

N= 16 15 21 32 22 19 14 11

nivel

### 9.6.7.3. Diagrama de dispersión

#### Problema-ejemplo

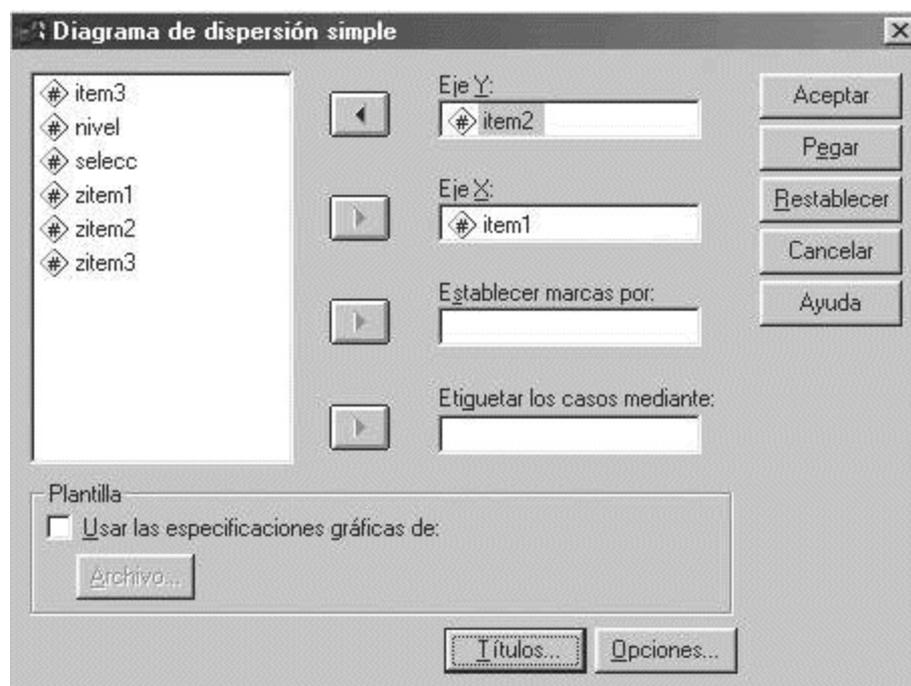
Supongamos el siguiente *problema-ejemplo*.

Se pretende representar, mediante un diagrama de dispersión, la relación entre los ITEM1 y ITEM2, del ejemplo que venimos utilizando, que como sabemos son los resultados de las pruebas realizadas a 150 alumnos de un instituto de enseñanza secundaria.

#### Desarrollo del ejemplo

Si utilizamos SPSS para realizar el problema-ejemplo propuesto la secuencia de ventanas del menú será la siguiente: **Gráficos > Diagrama de dispersión > Simple > Definir** (ver fig. 2.9.8).

Seleccionaremos el ITEM1 en el eje X y el ITEM2 en el eje Y. Además pondremos un título al gráfico: «Relación entre ITEM1 y ITEM2».



*Fichero de sintaxis*

GRAPH

/SCATTERPLOT(BIVAR)=item1 WITH item2

/MISSING=LISTWISE

/TITLE= 'Relación entre ITEM1 y ITEM2'.

*Gráfico de resultados* **Relación entre ITEM 1 y ITEM 2** 10

8

6

4

2

0

-2

-2 0246810 **comprensión lectora**

Como podemos observar en el gráfico no existe relación entre los dos ítems.

9.6.7.4. *Gráficos P-P*

*Problema-ejemplo*

Supongamos el siguiente *problema-ejemplo*.

Queremos observar, mediante un gráfico P-P, si la distribución de la variable ZITEM1, del ejemplo que venimos utilizando, que como sabemos son los resultados de las pruebas realizadas a 150 alumnos de un instituto de enseñanza secundaria, se aproxima a la normal.

*Desarrollo del ejemplo*

Si utilizamos SPSS para realizar el problema-ejemplo propuesto la secuencia de

ventanas del menú será la siguiente: **Gráficos >P-P** (ver fig. 2.9.9).

Seleccionaremos la variable ZITEM1.

*Fichero de sintaxis* PLOT

/VARIABLES=zitem1 /NOLOG

/NOSTANDARDIZE /TYPE=P-P

/FRACTION=BLOM /TIES=MEAN

/DIST=NORMAL.

*Gráficos de resultados* **Normal gráfico P-P de Punt. Comprensión lectora** 1.00

.75

.50

.25

0.00

0.00 .25 .50 .75 1.00 **Prob. acum. observada**

Como podemos observar la distribución de frecuencias se aproxima a la normal. Además lo reafirma el gráfico siguiente.

**Normal gráfico P-P de Punt. Comprensión lectora .06**

.04

.02

0.00

-.02

-.04

-.06

-.08

0.0 .2 .4 .6 .8 1.0 **Prob. acum. observada**

Donde todos los puntos están dentro de los márgenes admisibles de desviación.

### **EJERCICIOS DE AUTOCOMPROBACIÓN**

A la hora de realizar ejercicios conviene recordar que según la escala de medida, los estadísticos y gráficos más utilizados son:

Estadísticos y gráficos en los estudios descriptivos.

#### **Escala Estadísticos y gráficos**

Nominal Distribución de frecuencias Moda

Diagrama de barras

Ordinal

Intervalo o razón Mínimo

Máximo

Mediana

Cuartiles

Percentiles

Rango intercuartílico

Gráfico de caja y bigotes

Media

Rango

Varianza

Desviación típica

Coefficiente de variación

Coefficiente de asimetría

Coefficiente de curtosis

Histograma Gráfico de tallo y hojas

1. Las notas finales de 100 estudiantes de una Escuela Superior son las siguientes:

11 46 58 25 48 18 41 35 59 28  
35 2 37 68 70 31 44 84 64 82  
26 42 51 29 59 92 56 5 52 8  
1 12 21 6 32 15 67 47 61 47  
43 33 48 47 43 69 49 21 9 15  
11 22 29 14 31 46 19 49 51 71  
52 32 51 44 57 60 43 65 73 62  
3 17 39 22 40 65 30 31 16 80  
41 59 60 41 51 10 63 41 74 81  
20 36 59 38 40 43 18 60 71 44

Determinar:

a) El número de estudiantes con nota superior a 50 e inferior a 80. b) La nota del estudiante número 38 en orden a la peor puntuación.

2. Dada la siguiente tabla de frecuencias de una variable continua. Calcular su media.

**Clase  $n_i$**

10-15 3  
15-20 7  
20-25 16  
25-30 12  
30-35 9  
35-40 5  
40-45 2

3. Hállese la media, mediana, moda, desviación típica, asimetría y curtosis de los siguientes valores:

49 48 43 42 49 41 42  
43 43 44 44 51 53 54  
51 59 58 57 56 54 51  
54 53 64 62 64 63 62  
61 62 68 68 67 66 69

4. Un colegio realiza un programa de mejora de comprensión lectora mediante dos métodos A y B. De una forma descriptiva desea saber, ¿cuál de los dos métodos hace ganar más vocabulario en los alumnos? Para ello con 34 alumnos de iguales rendimientos, aplica una prueba y mide el número de palabras comprendidas mediante el método A por 18 alumnos y

el número comprendido mediante B por otros 16. Las palabras comprendidas se dan a continuación:

**A B**

256 253

274 262

**A B**

264 214

281 256

284 244

232 272

283 259

262 254

293 223

300 245

261 234

273 291

280 261

289 245

290 243

275 257

282

276

Dibujar el diagrama box-plot, localizar los casos atípicos y comparar las distribuciones dando respuesta a la pregunta formulada anteriormente.

5. Las calificaciones obtenidas, en un test de Estadística y en el examen final, por diez alumnos, fueron los siguientes:

Test (x) 6,0 7,2 4,7 7,7 6,9 3,7 7,5 8,1 9,2 5,3 Ex. final (y) 5,0 7,8 4,0 8,1 7,0 4,0 9,0 7,5 8,7 3,2

a) Calcular medias, varianzas y covarianza de x e y. b) Calcular el coeficiente de correlación lineal r.

**SOLUCIÓN A LOS EJERCICIOS DE AUTOCOMPROBACIÓN**

1. a) El fichero de sintaxis nos indica el proceso a realizar con la base de datos del ejercicio que hemos denominado p1.sav

\*\*\*\*\* primero ordenar el fichero de datos \*\*\*\*\*.

**Sort cases by**

x (A).

\*\*\*\*\* seleccionar los caso que cumplen la condición\*\*\*\*\*.

**If** (x > 50 & x<80) cont = 1 .

EXECUTE .

\*\*\*\*\* calcular su frecuencia, es decir, el número de casos  
\*\*\*\*\*.

FREQUENCIES

VARIABLES=cont

/ORDER= ANALYSIS .

El resultado será:

	Válidos	Perdidos	Total
cont			
<b>Frecuencia Porcentaje</b>	1,00	30	30,0
Sistema	70	70,0	
100	100,0		

Como se puede observar habrá 30 casos comprendidos entre  $x > 50$  y  $x < 80$  b) Como la base de datos p1.sav está ordenada se buscará el caso n° 38 y se ve que se corresponde con la nota de  $x_i = 35$ .

2. El fichero de sintaxis será:

\*\*\* primero ponderar cada valor de la variable con su frecuencia \*\*\*\*.  
WEIGHT

BY freq .

\*\*\*\*\* después calcular su distribución de frecuencia y su media \*\*\*.  
FREQUENCIES

VARIABLES=x

/STATISTICS=MEAN /GROUPEd= x

/ORDER= ANALYSIS .

Los resultados serán: ESTADÍSTICOS x N

	Media	Válidos	54	Perdidos	0
26,2037					
<b>x</b>					
<b>Frecuencia Porcentaje</b>					
<b>Porcentaje Porcentaje válido acumulado</b>					
Válidos	12,50	3	5,6	5,6	5,6
17,50	7	13,0	13,0	18,5	
22,50	16	29,6	29,6	48,1	
27,50	12	22,2	22,2	70,4	
32,50	9	16,7	16,7	87,0	

37,50 5 9,3 9,3 96,3  
42,50 2 3,7 3,7 100,0 Total 54 100,0 100,0

Como se observa la media es: 26,2

3. El fichero de sintaxis será:

```
FRECUENCIES
VARIABLES=x /FORMAT=NOTABLE
/STATISTICS=STDDEV MEAN MEDIAN MODE SKEWNESS
SESKEW KURTOSIS SEKURT
/ORDER= ANALYSIS.
```

Los resultados son: ESTADÍSTICOS x  
N

Media  
Mediana Moda  
Desv. típ. Asimetría Válidos 35 Perdidos 0

54,71  
54,00  
43(a) 8,784  
-,010  
Error típ. de asimetría ,398  
Curtosis -1,231  
Error típ. de curtosis ,778

(a) Existen varias modas. Se mostrará el menor de los valores. 4. La tabla de los cuartiles, de rango intercuartílico (IQR), media y desviación típica es la siguiente:

	<b>Q<sub>1</sub></b>	<b>Q<sub>2</sub></b>	<b>Q<sub>3</sub></b>	<b>IQR</b>	<b>Media</b>	<b>Desv.</b>
A	263,50	278,00	285,25	21,75	275,28	15,951
B	243,25	253,50	260,50	17,25	250,81	18,306

Por tanto los coeficientes de variación son:  
 $CV_A = 15,951/275,28 = 0,058$   $CV_B = 18,306/250,81 = 0,073$

Para el tipo A:

$LI = 278 - 1,5 * 21,75 = 245,375$   $LS = 278 + 1,5 * 21,75 = 310,625$

Fuera del intervalo (245,375; 310,625) se situarán los valores atípicos.

Para el tipo B:

$LI = 253,50 - 1,5 * 17,25 = 227,625$   $LS = 253,50 + 1,5 * 17,25 = 279,375$

Fuera del intervalo (227,625; 279,375) se situarán los valores atípicos.

El diagrama de caja para ambas variables es:

300 o12  
280  
260

240  
o6

220  
o3  
200

ab Gráficamente se ve que A tiene una mediana mayor que B, y sólo un valor atípico, mientras B tiene dos.

Se concluye que el método A da mejores resultados porque aprenden más palabras y tienen menor coeficiente de variación.

El fichero de sintaxis es:

```
EXAMINE  
VARIABLES=a b  
/PLOT BOXPLOT  
/COMPARE VARIABLES  
/PERCENTILES(25,50,75) HAVERAGE  
/STATISTICS DESCRIPTIVES EXTREME  
/CINTERVAL 95  
/MISSING PAIRWISE  
/NOTOTAL.
```

5. Para el cálculo de los estadísticos se plantea la siguiente tabla: a)  
Los estadísticos pedidos son:

ESTADÍSTICOS DESCRIPTIVOS

**Media Desviación típica** N x 6,630 1,6846 10  
y 6,430 2,1649 10

CORRELACIONES

**x y** Correlación de Pearson 1 ,895(\*\*) Sig. (bilateral) ,000

x

Suma de cuadrados y produc  
tos cruzados 25,541 29,371

Covarianza 2,838 3,263 N 10 10

Correlación de Pearson ,895(\*\*) 1 Sig. (bilateral) ,000

y

Suma de cuadrados y produc

tos cruzados 29,371 42,181

Covarianza 3,263 4,687 N 10 10

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

b) El coeficiente de correlación de Pearson ya ha sido calculado en la tabla anterior y su valor es: 0,896.

El fichero de sintaxis es:

```
CORRELATIONS
/VARIABLES=x y
/PRINT=TWOTAIL NOSIG
/STATISTICS DESCRIPTIVES XPROD
/MISSING=PAIRWISE .
```

### **BIBLIOGRAFÍA**

ARCE, CONSTANTINO;REAL, EULOGIO (2002): *Introducción al análisis estadístico con SPSS*. PPU. Barcelona.

CALOT, G. (1974): *Curso de Estadística Descriptiva*. Paraninfo. Madrid.

CAMACHO ROSALES, JUAN (2002): *Estadística con SPSS versión 11 para Windows*. Ra-Ma, Librería y Editorial Microinformática. Madrid.

CUADRAS, C. M.; ECHEVERRÍA,B.; MATEO, J.; SÁNCHEZ, P. (1991): *Fundamentos de Estadística*. PPU. Barcelona.

FERRÁN ARANAZ, MAGDALENA (2002): *Curso de SPSS para Windows*. McGraw-Hill/ Interamericana de España. Madrid.

LIZASOAIN HERNÁNDEZ, LUIS; JOARISTI OLARRIAGA, LUIS (2003): *Gestión y análisis de datos con SPSS*. Thomson Paraninfo. Madrid.

MARTÍN F. G. (1994): *Introducción a la Estadística Económica y Empresarial*. AC. Madrid. PARDO MERINO, ANTONIO;RUIZ DÍAZ, MIGUEL ÁNGEL (2002): *SPSS 11. Guía para el análisis de datos*. McGraw-Hill/ Interamericana de España. Madrid.

PEÑA D. (1992): *Estadística, Modelos y Métodos. Volumen I*. Alianza Universidad Textos. Madrid. PEÑA D. (1992): *Estadística, Modelos y Métodos. Volumen II*. Alianza Universidad Textos. Madrid.

P ÉREZ, CÉSAR (2001): *Técnicas estadísticas con SPSS*. Pearson Educación. Madrid. RÍOS, S. (1974): *Métodos estadísticos*. Ed. del Castillo. Madrid.

SPIEGEL, M. R. (1992): *Estadística*. McGraw-Hill. Madrid.

VISAUTA VINACUA, B. (2002): *Análisis estadístico con SPSS 11.0 para Windows*. McGraw-Hill/ Interamericana de España. Madrid.

## **UNIDAD DIDÁCTICA 3**

# MUESTREO Y ESTIMACIÓN

## Objetivos

Distinguir los conceptos básicos del muestreo estadístico. Distinguir los diferentes tipos de muestreos.

Calcular el tamaño de la muestra para los diferentes muestreos.

Resolver supuestos donde se utilicen los distintos tipos de muestreos.

Utilizar programas con SPSS para el cálculo del tamaño muestral para diferentes tipos de muestreo.

Conocer las propiedades de estimadores y reconocer las características de bondad de los mismos.

Identificar y cuantificar los riesgos en las decisiones sobre parámetros poblacionales.

Calcular los principales estimadores de la media, varianza, diferencia de medias y proporciones.

## 1. TEORÍA ELEMENTAL DEL MUESTREO. TIPOS DE MUESTREOS

### 1.1. Introducción

Al estudiar las características de un grupo de individuos u objetos, podemos o bien examinar el grupo entero, llamado *población* o *universo* o bien examinar una pequeña parte del grupo, llamada *muestra*.

Además de población y muestra existen algunos términos relacionados con el muestreo que se necesitan conocer con cierta precisión:

— **Elemento.** Es un objeto en el cual se toman las mediciones.

— **Población.** Es un conjunto de elementos acerca de los cuales se desea hacer inferencias.

— **Unidades de muestreo.** Son grupos excluyentes de elementos de la población que completan la misma.

— **Marco.** Es una lista de unidades de muestreo.

— **Muestra.** Es una colección de unidades seleccionadas de un marco o de varios.

Veamos un ejemplo que clarifique estos conceptos. En una ciudad se realiza una encuesta para determinar la actitud del ciudadano frente a una colección de fascículos antes de su lanzamiento al mercado. En concreto, se quiso saber la proporción del público favorable a su introducción en los circuitos comerciales.

En el ejemplo los **elementos** serán los potenciales encuestados, habitantes de la ciudad, de los que se toma como medida su actitud ante el lanzamiento del producto (1 = actitud favorable, 0 = actitud contraria a su implantación).

La **población** es el conjunto de habitantes de la ciudad. Para definir con precisión la población se necesita concretar los elementos que la definen y la medición que se va a realizar con los mismos.

Las **unidades de muestreo** pueden ser los habitantes de la ciudad o por ejemplo los hogares de la ciudad, consumidores de fascículos. Las unidades de muestreo deben ser disjuntas de tal manera que un hogar no pueda ser muestreado más de una vez.

El **marco** puede ser una lista de todos los hogares de la ciudad obtenida del censo de la misma.

Finalmente, una **muestra** puede ser un conjunto de hogares seleccionados del marco.

La *teoría del muestreo* estudia la relación entre una población y las muestras tomadas de ella. Es de gran utilidad en muchos campos. Por ejemplo, para *estimar* magnitudes desconocidas de una población, tales como media y varianza, llamadas a menudo *parámetros* de la población o simplemente *parámetros*, a partir del conocimiento de esas magnitudes sobre muestras, que se llaman *estadísticos de la muestra* o simplemente *estadístico*.

### Muestreo Población Muestra

Inferencias a cerca de la población:

Intervalo de confianza<sub>Estadístico</sub> Análisis de datos

FIGURA 3.1.1. Intervalo de confianza.

La teoría del muestreo es también útil para determinar si las diferencias observadas entre dos muestras son debidas a variaciones fortuitas o si son realmente significativas. La respuesta implica el uso de los llamados *contrastes o tests de hipótesis y de significación*.

### Muestreo Población

Muestra

Conocimientos o Supuestos previos Análisis de datos

Hipótesis acerca de la población Confrontación de las hipótesis con los datos Estadístico

Rechazo  
de la hipótesis

### FIGURA 3.1.2. Contraste de hipótesis.

Para que las conclusiones de la teoría del muestreo y de la inferencia estadística sean válidas, las muestras deben escogerse *representativas* de la población. Una forma de obtener una muestra representativa es mediante *muestreo aleatorio*, de acuerdo con el cual, cada miembro de la población tiene la misma probabilidad de ser incluido en la muestra.

Cochran (1975) enumera cuatro ventajas que aporta el empleo del muestreo estadístico: «costo reducido, mayor rapidez, mayores posibilidades y mayor exactitud».

#### 1.2. Diseños muestrales

La selección de la muestra de un marco apropiado puede realizarse por procedimientos probabilísticos, es decir, con procedimientos que aseguren a cada una de las unidades muestrales una probabilidad de ser seleccionadas, entonces se tendrá los muestreos probabilísticos. Cuando no ocurre tal hecho se tendrá los muestreos no probabilísticos.

Los principales muestreos probabilísticos son: aleatorio simple, estratificado, por conglomerados y sistemático.

Entre los segundos están: por cuotas, el intencional o deliberado y el accidental.

Aún cuando el capítulo se dedicará a revisar el primer grupo, del segundo el muestreo por cuotas suele auxiliar al muestreo probabilístico para conservar cierta representación de grupos destacados en la investigación. En el muestreo intencional el investigador selecciona la muestra de modo directo para asegurarse la presencia de elementos necesarios en la investigación. El muestreo accidental se caracteriza por utilizar las muestras que tienen a su alcance.

**1.3. Selección de la muestra en diseños probabilísticos** Como sabemos el objetivo del muestreo es estimar *parámetros* de la población como la media, el total o la proporción de un determinado atributo. Si  $q$  es el parámetro en estudio y  $\hat{q}$  el estimador del mismo. Para la selección de  $q$  la muestra será necesario tomar dos decisiones:

a) Fijar una cota para el error de estimación:

$$\text{Error de estimación} = |q - \hat{q}| < e$$

$|<e$

a) indicador del número de veces que al repetir el muestreo y medir el parámetro, el error de estimación se mantiene menor que una cantidad fijada ( $e$ ).

mantiene menor que una cantidad fijada ( $e$ ).

a)

)

a) = 0,95 y las distribuciones de los estimadores son aproximadamente normales para tamaños muestrales razonablemente grandes. El objetivo del investigador será conseguir un diseño que produzca mínimo  $e$  con un coste reducido.

A continuación se detallarán los distintos tipos de diseños probabilísticos.

#### **1.4. Muestreo aleatorio simple**

Extraer una muestra aleatoria simple de  $n$  unidades, elegidas entre las  $N$  de la

**población, es escogerla de manera que todas las  $\binom{N}{n}$  muestras posibles tengan  $n$**

la misma probabilidad de ser elegidas. Así por ejemplo, si en un instituto de 1200 alumnos se quieren elegir 60 alumnos de forma aleatoria, habrá:

$$\binom{1200}{60} = \frac{1200!}{60! \cdot 1140!} = \frac{1200 \cdot 1199 \cdot \dots \cdot 1141}{60 \cdot 59 \cdot \dots \cdot 1} \approx 10^{10}$$

muestras diferentes que tendrán  $\frac{1}{\binom{1200}{60}} \approx 10^{-10}$  probabilidad cada muestra de ser elegida.

Para seleccionar una *muestra aleatoria simple* el primer paso es confeccionar una lista de objetos de los cuales se seleccionará la muestra. Estos objetos, según se ha comentado, son las unidades muestrales.

La manera más sencilla y más segura de obtener una muestra aleatoria de  $n$  unidades muestrales de una población grande  $N$ , es utilizar una *tabla de números aleatorios*, comprendidos entre 1 y  $N$ , y elegir  $n$ . Los números así elegidos forman la muestra. Este esquema de muestreo donde las unidades muestrales no regresan a formar parte de nuevo para la siguiente elección se llama *muestreo sin reemplazamiento*. Si vuelven a formar parte de la elección el *muestreo es con reemplazamiento*. En la mayoría de programas estadísticos de ordenador hay funciones de librería para extraer los números aleatorios. Por ejemplo SPSS tiene distintos comandos para este fin:

```
COMPUTE x = UNIFORM(10)
```

```
EXECUTE
```

Genera números aleatorios de distribución uniforme en el intervalo (0, 10).

### **1.4.1. Estimación de media, total y proporción**

Para fijar ideas se supone el siguiente ejemplo.

#### *Ejemplo 3.1.1*

Los resultados obtenidos por una muestra de 10 alumnos de una población de 100, en una prueba de matemáticas es:

**Número de ejercicios** Alumno terminados

$y_i$

**Número Sexo de ejercicios bien  $x_i$  resueltos  $z_i$**

**Número  
de ejercicios mal resueltos**

$w_i$

1 4

2 4

3 5

4 6

5 6

6 3

7 7

8 5

9 5

10 4

*Total* 49

0 2

0 2

0 3

0 2

1 3

1 1

0 3

0 2

1 2

1 2

4 22

2

2

2

4

3

2

4

3

3  
2  
27

### Estimador de la media poblacional

Si se denota por

$\mu$

la media poblacional

–

y por la media muestral, esta última es un estimador insesgado de la primera. Es decir:

$E(\bar{y}) = \mu$

–

$\bar{y}) = \mu$

En el ejemplo 3.1.1:

$\sum_{i=1}^n \hat{y}_i$

$y^{i=1} = \dots$

$n \cdot 10.49$

Además debemos estudiar la bondad del estimador que vendrá dado por la varianza:

$$V(\bar{y}) = \frac{s^2}{n} \left( \frac{N-1}{N} \right)$$

Donde:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$s^2$

es la varianza poblacional  $= \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2$

$N$

Como en la mayoría de los casos se desconoce  $s^2$ , pero se sabe que el estimador insesgado de la cuasi-varianza poblacional de la media es la cuasi-varianza muestral de la media. Es decir, se cumple:

$$E(s^2) = S^2$$

Donde:  $n$

$$N^2 \hat{\sigma}^2(\hat{\mu})$$

$\sigma^2$  es la cuasi-varianza poblacional y  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$  la cuasi

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

varianza muestral (recuérdese la equivalencia, para abreviar las notaciones, entre  $s^2 = \hat{\sigma}^2$ ).

Y que existe la relación  $Ns^2 = (N-1)S^2$  ó  $Ns^2 = (N-1)s^2$  al sustituir por su estimador. Entonces la estimación de la varianza de la media muestral será:

$$\hat{V}_y(\hat{\mu}) = s^2 \hat{N} n^{-1} \left( 1 - \frac{n}{N} \right) = n^{-1} \left( 1 - \frac{n}{N} \right) s^2$$

$$N^{-1} = \frac{1}{N}$$

donde  $f = \frac{n}{N}$  se llama fracción de muestreo y  $N^{-1}$  corrección por poblaciones finitas.

En el ejemplo 3.1.1 :

$$V_{yss}(\hat{\mu}) = 0,129$$

$$\hat{V}_y(\hat{\mu}) = 0,9$$

$n n$

$$2 \hat{\sigma}^2(\hat{\mu})$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 1,43$$

11 9

El intervalo de confianza donde, en el  $(1 - \alpha)\%$ , de las veces estará el parámetro poblacional  $\mu$  será:

$\hat{\mu}$

$$y \pm k V_y(\hat{\mu})$$

donde k: factor de desviación y le corresponde, en general, una

confianza  $P_k = 1 - \alpha$ . En la mayoría de los casos  $k = 2$ , y si la distribución del estimador es normal  $P_k = 0,954$ .

En la fórmula [2] si  $N$  se hace muy grande frente a  $n$ , en la práctica Scheaffer (1986, pág. 46) aconseja tomar esta medida cuando  $n \leq 20 N$ , entonces  $f=0$  y por tanto:

$$\hat{\sigma}^2 = s^2$$

$n$

### Estimador del total poblacional

Si se denota por  $Y$  al total poblacional de la característica  $y$ . Un estimador de  $Y$  será:

$$\hat{Y} = N\bar{y}$$

En el ejemplo 3.1.1:

$Y$   
 $\hat{Y}$

49  
 $= \diamond$

$100 \cdot 10 = 490$  ejercicios terminados La varianza será:

$$\hat{\sigma}^2 = (s^2) \left( \frac{N-n}{N} \right)$$

Donde su estimación será:

$\frac{2222}{2}$

— —

$2 \cdot 1 \hat{\sigma}^2$

0

^

$$= 0 = n \hat{\mu} \pm z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}$$

$nN$   $n$   $nNf$

Su intervalo de confianza será:  $\hat{\mu} \pm z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}$

# YkVY ()

## Estimador de la proporción poblacional

Supongamos que A es el número total de individuos que presentan un atributo. En el ejemplo 3.1.1, sea el atributo  $a_i$ , el sexo, que toma el valor 1 si el alumno es mujer y 0 si es hombre. Es decir, en general:

$a_i$   
 $= \begin{cases} 1 & \text{si el elemento tiene el atributo} \\ 0 & \text{si el elemento no tiene el atributo} \end{cases}$

$A = \sum_{i=1}^N a_i$   
 $B = N - A$ , por tanto  $Q = \frac{B}{N} = 1 - P = \frac{\sum_{i=1}^N (1 - a_i)}{N}$ . Sea  $P = \frac{A}{N}$

El estimador de P será  $\hat{p} = \frac{\sum_{i=1}^n a_i}{n}$ . Es un estimador insesgado  $E(\hat{p}) = P$   
 El estimador de A será:

$$\hat{A} = n \hat{p}$$

Es también un estimador insesgado:  $E(\hat{A}) = A$ . Realmente al ser A una variable dicotómica se cumple:

$$A = \sum_{i=1}^N a_i$$

$$\hat{p} = \frac{\sum_{i=1}^n a_i}{n}$$

$$s^2 = \frac{\sum_{i=1}^n (a_i - \hat{p})^2}{n-1}$$

$$s^2 = \frac{\sum_{i=1}^n a_i^2 - n \hat{p}^2}{n-1}$$

$$s^2 = \frac{\sum_{i=1}^n a_i - n \hat{p}^2}{n-1}$$

Por tanto todas las fórmulas de la varianza quedan así:

$$V(\hat{p}) = \frac{PQ}{n}$$

$$V(\hat{A}) = n \frac{PQ}{n-1}$$

$$V(\hat{p}) = \frac{PQ}{n}$$

$$V(\hat{A}) = n \frac{PQ}{n-1}$$

Por tanto el intervalo de confianza para p será:  $\pm k \sqrt{V(\hat{p})}$

Para A se cumplirá las fórmulas vistas para el total:

$$V(\hat{A}) = n \frac{PQ}{n-1}$$

Donde su estimación será:

$$VA = \frac{Np}{n} \sqrt{\frac{pq}{n}}$$

Por tanto su intervalo de confianza será:

$$\hat{p} \pm A k \sqrt{\hat{p}(1-\hat{p})}$$

En el ejemplo 3.1.1:

A = 4 cuando sexo = 1 y n-a = 10-4 = 6 por tanto

$$p =$$

$$a^4$$

$$n = 10 - 4 = 6$$

Y por tanto

$$Vp =$$

$$\frac{pq}{n}$$

$$=$$

$$\frac{10 \cdot 0.4 \cdot 0.6}{6} = 0.4$$

$$100 \cdot 0.4 = 40$$

$$\hat{p}$$

$$A = 100 \cdot 0.4 = 40$$

y su estimación de varianza será:  $\hat{p}(1-\hat{p}) = 0.4 \cdot 0.6 = 0.24$ . En resumen se puede plantear la siguiente tabla:

TABLA 3.1.1. Estimadores, varianzas e intervalo de confianza: media, total y proporción.

**Parámetro**

**Estimador Varianza Estimador Intervalo del parámetro del estimador de la varianza de confianza**

n

$$\hat{y}_i \quad V_y = \frac{1}{n} \sum (y_i - \bar{y})^2$$

y

$$=$$

$$= 1 = 1 \quad V_y = 1 \quad y_k \quad V_y$$

n n n

$$s^2$$

Total

$$\hat{y}_{Ny} \quad V_{Yn} \quad V_y = (\hat{y}) \quad V_Y = (\hat{y}) \pm \sqrt{(\hat{y})}$$

nNf

Proporción

$p =$

$$V(\hat{p}) = \frac{pq}{n}$$

$$V(\hat{p}) = \frac{pq}{n}$$

$$\hat{p} \pm k \sqrt{V(\hat{p})}$$

### 1.4.2. Selección del tamaño de la muestra para la estimación de la media, total y proporción

El elegir un número adecuado de unidades muestrales presenta un ahorro en tiempo, dinero y una mayor eficacia.

Para calcular el tamaño de la muestra, según se ha visto en el punto tercero del tema, el investigador necesita:

a) Fijar un error de estimación (e). Los valores más utilizados son siempre inferiores al 10% (e=0,1).

b) Fijar un valor k para un nivel de confianza. En la mayoría de los casos k=2, que en el supuesto de distribución normal del estimador, representará el 95,5% de nivel de confianza, de que los intervalos formados por

$\hat{p} \pm k \sqrt{V(\hat{p})}$  cubran al parámetro p.

#### Media poblacional

Como  $k=e$ , entonces, por lo visto anteriormente,

$$\hat{e} = \frac{e}{2}$$

$$n = \frac{N s^2}{e^2}$$

$$n = \frac{N s^2}{e^2}$$

Como en la mayoría de los casos se desconoce

s

se debe reemplazar  $V(\bar{y})$  por

su estimador

$$V(\hat{y})$$

$$=$$

-

$$s^2 \frac{N}{n} \text{ entonces: } n = \frac{N s^2}{e^2}$$

s

$$n = \frac{N s^2}{e^2}$$

$$n = \frac{N s^2}{e^2}$$

$$+ N k^2$$

Si la población es infinita o el muestreo es con reemplazamiento

$$222_s e 2 k^2$$

Vy

^()

S por tanto

$$= =fi==k nn s_{e2n n}$$

A la fórmula anterior se puede llegar mediante el límite, cuando N tiende a infinito, de la expresión [3].

En el ejemplo 3.1.1 el tamaño de muestra apropiado con e=0,05 y k=2 será:

$$n 143^a 95= 005^2, 143$$

$$2^2 + 100$$

### Total poblacional

Para estimar, con el total poblacional, el tamaño de muestra adecuado es:

^()VNy NV y^2 ^() por tanto

s

$$22n Ns^22$$

$$NN n$$

()

e

- =fi=

$$n k^2 e^2_2$$

$$k2 +Ns$$

En el cálculo del tamaño muestral, en este caso, no tiene sentido hablar de población infinita.

### Proporción

En el cálculo de p= a , sabemos que V(p) =Nn- PQ , por tanto\_n N-1 n

$$N PQ Nn PQ e^2_{N-1}$$

N

- = fi=

n

k

22

$$e 1 PQ_{k2 + 1}$$

Si la población es infinita o el muestreo es con reemplazamiento el tamaño muestral será el límite de la expresión anterior cuando N tiende a infinito.

$$PQ^2 k PQ nn. = = =_2 e2e$$

k^2

Cuando no conocemos PQ pero si se sabe el máximo valor que puede tomar

$P=Q=1$ , lo habitual es coger ese valor extremo como estimador de PQ.  
Entonces

las fórmulas anteriores quedarán así:

$N$   
Población finita:

$$n =$$

$$\frac{e^2}{k^2 + 1}$$

$$41$$

Población infinita:

$$n =$$

$$= \frac{e^2}{k^2 + 1} (N - 1)$$

$$4e^2$$

$$k^2 + 1$$

En el ejemplo 3.1.1, la muestra apropiada en las condiciones más desfavorables del muestreo

$$100 P = Q = 1, k = 2, e = 0,05 \Rightarrow n = 99 \text{ a } 80$$

En resumen se puede plantear la siguiente tabla:

TABLA 3.1.2. Cálculo del tamaño de la muestra. Muestreo aleatorio simple.

**Parámetro** Población finita Población infinita  $s^2$  Media

$$n = \frac{e^2}{k^2 + 1} \left( \frac{N}{PQ} + 1 \right)$$

$$=$$

$$N$$

$$e^2$$

$$k^2 + 1$$

$$k^2$$

Total

$$=$$

$$n = \frac{e^2}{k^2 + 1} \left( \frac{N}{PQ} + 1 \right)$$

$$= \frac{e^2}{k^2 + 1} \left( \frac{N}{PQ} + 1 \right)$$

$$= \frac{e^2}{k^2 + 1} \left( \frac{N}{PQ} + 1 \right)$$

$$= \frac{e^2}{k^2 + 1} \left( \frac{N}{PQ} + 1 \right)$$

$$= \frac{e^2}{k^2 + 1} \left( \frac{N}{PQ} + 1 \right)$$

$$N$$

$$\text{Caso } P = Q = 1, n = \frac{e^2}{k^2 + 1} (N - 1)$$

$$= \frac{e^2}{k^2 + 1} (N - 1)$$

*Nota:* En el caso de un muestreo aleatorio simple, una unidad muestral contiene solamente un elemento.

### 1.5. Muestreo estratificado

Un procedimiento para reducir los costos de sondeos de opinión pública o de los consumidores, es dividir en segmentos, llamados *estratos*, la región geográfica en la cual residen los elementos (personas) de la población. Se seleccionan muestras dentro de cada estrato; después se combina esta información para hacer inferencias acerca de toda la población.

El *muestreo aleatorio estratificado* tiene otra ventaja, aparte de la económica, no solamente puede combinar la información de las muestras de los estratos para obtener inferencias de la población, sino también utilizar la información de la muestra acerca de la característica de cada estrato. Por ejemplo nos permite ver la diferencia en la opinión entre hombres y mujeres.

La principal limitación del muestreo estratificado es que se requiere conocer de antemano los factores importantes de la población, así como sus proporciones relativas.

Para obtener mayor precisión en las estimaciones, con este tipo de muestreo, se debe perseguir que los estratos cumplan la condición de que la varianza intra-estrato sea mínima y la varianza entre-estrato máxima, con este ideal reduciremos el error muestral o conseguiremos reducir el tamaño muestral si deseamos mantener el mismo error muestral, es decir, lograremos mayor eficacia.

Para seleccionar una muestra aleatoria estratificada habrá que especificar claramente los estratos. Después se seleccionará una muestra aleatoria de cada estrato.

Por ejemplo, se supone que se va a realizar una encuesta para estudiar los hábitos de lectura en un núcleo de población donde hay tres zonas perfectamente diferenciadas: A=zona rural, B=zona semi-urbana y C=zona urbana. Para realizar el muestreo se elegirá aleatoriamente, de cada zona, las unidades muestrales pertinentes. Si N es el tamaño poblacional y L el número de estratos:

$N_1,$

$N$

$\hat{O}$

$$\sum_{j=1}^L \hat{\theta}_j = \sum_{j=1}^L \hat{\theta}_j \frac{N_j}{N} = \hat{\theta}$$

...  
 $\hat{\theta}_{j=1}$   
 $\hat{\theta}_{N_{L_c}}$

Llamaremos  $W_j$  el peso del estrato  $j$  dentro de la población

$$W_j = \frac{N_j}{N}$$

siendo

$L$

$$W_j = \frac{N_j}{N}$$

$$\sum_{j=1}^L W_j = 1$$

### 1.5.1. Estimadores de la media, el total y la proporción poblacional

#### Media poblacional

Se cumplirá:

$$1$$

$$N$$

$$1$$

$$N$$

$$2$$

$$N$$

$$L$$

$$L_j$$

$$\hat{y}_{st} = \sum_{j=1}^L W_j \hat{y}_j = \sum_{j=1}^L \frac{N_j}{N} \hat{y}_j$$

El estimador de la varianza de

$$- \hat{y}_{st}$$

$$V_y$$

$$\hat{V}(\hat{y}_{st}) = \sum_{j=1}^L W_j^2 \frac{N_j - 1}{N_j} S_{y_j}^2 + \frac{1}{N} \sum_{j=1}^L W_j^2 (N_j - 1) S_{y_j}^2$$

$$\hat{A}$$

$$j=1$$

$$V_y N n$$

$$- s^2$$

$$j j$$

$$j$$

= Como , sustituyendo:  $N_j n_j$

$$L W N n^{-2} s^2$$

$$V_y$$

$$\hat{V}_{st} = \sum_{j=1}^J \hat{A}_j N_j n_{jj}^{-1}$$

nótese que  $s^2$  es la cuasi-varianza muestral,

El intervalo de confianza de la media poblacional será:

$$\bar{y}_{st} \pm k \hat{V}_{st}$$

En el ejemplo anterior de los hábitos de lectura supongamos los siguientes datos:

### Ejemplo 3.1.2

**Zona A Zona B Zona C**  $n_1=18$   $n_2=10$   $n_3=12$

$\bar{y}_1=34$   $\bar{y}_2=25,2$   $\bar{y}_3=18,5$   $s_1^2=35,4$   $s_2^2=210,2$   $s_3^2=86,8$   $N_1=160$   
 $N_2=72$   $N_3=98$

$$N = N_1 + N_2 + N_3 = 160 + 72 + 98 = 330$$

$$\hat{V}_{st} = \sum_{j=1}^3 \hat{A}_j N_j n_{jj}^{-1}$$

$$\hat{V}_{st} = \hat{A}_1 \hat{A}_2 \hat{A}_3 = 1$$

$$\hat{A}_1 = \frac{160}{330} = 0,4848$$

$$\hat{A}_2 = \frac{72}{330} = 0,2182$$

$$\hat{A}_3 = \frac{98}{330} = 0,2970$$

$$\hat{V}_{st} = 0,4848 \cdot 0,2182 \cdot 0,2970 = 0,3118$$

$$\hat{V}_{st}$$

$$= \hat{E}$$

$$\hat{E}^2$$

$$= 0,3118^2$$

$$= 0,0972$$

$$+ \hat{E}$$

$$\hat{E}^2$$

$$= 0,0972 + 0,3118$$

$$= 0,4090$$

$$= 0,4090$$

$$= 0,4090$$

$$= 0,4090$$

$$= 0,4090$$

$$= 0,4090$$

$$= 0,4090$$

$$= 0,4090$$

$$= 0,4090$$

$$= 0,4090$$

$$= 0,4090$$

$$= 0,4090$$

$$= 0,4090$$

$$= 0,4090$$

### Total poblacional

$$\hat{V}_{st}$$

$$= \sum_{j=1}^J \hat{A}_j \hat{A}_j$$

$$j=1$$

La estimación de la varianza de  $\hat{Y}_{st}$  será:

$$L \hat{E}_{NNn}^{ij} \hat{s}^2$$

$$VY$$

$$\hat{(\hat{)}}_j$$

$$st = =_{st} j$$

$$N V_y \hat{() \hat{A}^{22} \hat{A} \hat{A} N_j \sim nj=1 \hat{E}^-$$

Por tanto el intervalo de confianza será:

$$\sim(\hat{)}$$

$$st YkVY_{st}$$

En el ejemplo 3.1.2 conocido  $_{st} y V_y \hat{() = 1,8318$  entonces:  $_{st}$

$$\hat{ = YN y_{st} = 330.27,477 = 9067,4_{st}$$

$$VY$$

$$\hat{(\hat{)}$$

$$^2$$

$$\hat{() .,$$

$$N V y_{st} = = 199480_{st}$$

**Proporción poblacional p**

$$N$$

$$N$$

$$N L$$

$$\hat{A}$$

$$p$$

$$\hat{^1 (\hat{^}) \hat{^2 = + + + = + + + = \dots \dots \sim LLjj_{st} N N p N p_{11} 2 2 N p_{L L N 1 N 2 N p w p}$$

$$j=1$$

$$\hat{^$$

El estimador de la varianza de  $p_{st}$  será:

$$V p_{st} \hat{(\hat{)}$$

$$L L \hat{^$$

$$ij ij$$

$$N N V p N V p_j$$

$$1 2 \hat{(\hat{)} 2 \hat{(\hat{)} 2 L \hat{(\hat{)} L \hat{A}^w V P 22 2 w N n^{- p q}$$

$$N V p = + + + = =_j \hat{A}_j N_j n_j^{-1} j=1 j=1$$

El intervalo de confianza será:

$$\hat{_{st} \pm p k V p \hat{(\hat{_{st} )}$$

En el ejemplo 3.1.2 supongamos que se sabe que el número de hogares donde se lee algún libro de aventuras se distribuye en los estratos de la siguiente manera:

## Núm. de hogares Estrato

### Tamaño donde se lee de la muestra algún libro de aventuras

A 18 8 0,44 B 10 4 0,4 C 12 6 0,5

La estimación de la proporción de personas que leen algún libro de aventuras en la población será:

$L$

$\sim 16 12 49$

$st$

$pwp_j = + +$

$\hat{A} 5504 165 0 5 = 0,44909_{j=1} 33044$

$L$

$\hat{(\hat{)}$

$Vp \hat{A} w Vp^2 \hat{(\hat{)}}_j \text{ pero: } st =_j$

$j=1$

$\hat{\hat{}}$

$Vp N n_{pq} 11 = 160-18 \cdot \cdot \cdot 044056 = 0,012864$

$1 11$

$N n_1 - 1 160 18-1_1$

$\hat{\hat{}}$

$\hat{(\hat{)}}^{22}$

$N$

$Vp N n_{pq} 72-10 \cdot \cdot \cdot 0406 = 0,0229632 22 n_2 - 1 = 72 10-12$

$\hat{\hat{}}$

$Vp N n$

$pq$

$33 98-12 \cdot \cdot \cdot 0505 = 0,019944$

$3 33 n_3 - 1 =$

$N 98 12-1_3$

En consecuencia

$Vp$

$st$

$= \hat{E}$

$\hat{E} \hat{A} \hat{}$

$22 2$

$\hat{(\hat{)}, \cdot \cdot \cdot, + \hat{E} \hat{E} \hat{A} \hat{ + \hat{E} \hat{E} \hat{A}^{49} \hat{ } 0 019944 = 0,005876_{165} \sim$

### 1.5.2. Cálculo del tamaño muestral para la estimación de la media total y proporción poblacional

#### Media

$LWS22$

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{j=1}^k w_j^2 S_j^2$$

$$L_e = \sum_{j=1}^k w_j^2 S_j^2$$

$$+ \frac{1}{N} \sum_{j=1}^k w_j^2 S_j^2$$

**Total**

$$L_{NS^2}$$

$$\hat{\sigma}^2$$

$$\sum_{j=1}^k w_j^2 S_j^2$$

$$L_e + \hat{\sigma}^2 NS^2$$

$$k$$

$$\sum_{j=1}^k w_j^2 S_j^2$$

**Proporción**

$$L N_j P Q_{jj} \hat{\sigma}^2 W^2$$

$$\sum_{j=1}^k N_j - 1 w_{jn} = j=1$$

$$\sum_{j=1}^k L N_j^2 e + \hat{\sigma}^2 P Q$$

$$N$$

$$j$$

$$-$$

$$1$$

$$\sum_{j=1}^k w_j^2 S_j^2$$

$$k N_{j=1}$$

$$n$$

$$j$$

es la cuasivarianza poblacional. Recordar que

$$S$$

$$2$$

$$=$$

$$N_j - 1 \sum_{j=1}^k w_j^2 S_j^2 =$$

$$N_j P Q \text{ donde } S_j^2 =$$

**1.5.3. Afijación** Se denomina afijación al reparto del tamaño muestral  $n$

entre los distintos estratos

$L$

de tal forma que se verifique  $\sum_{j=1}^L n_j = L$ . Puede ser de distintos tipos:

$j=1$

a)

*Afijación uniforme,*

donde se reparte por igual entre todos los estratos  $n_j$

=

$n_j = L/n$  b) *Afijación proporcional,* la que se hace en proporción al tamaño del estrato.

$n_j = N_j/n$ . Teniendo, en este caso, todas las unidades muestrales la misma

probabilidad de ser seleccionadas en la muestra.

c) *Afijación óptima o de mínima varianza,* donde se eligen los  $n_j$  de forma que minimicen la varianza para un  $n$  fijo:

$n_j = \frac{N_j}{\sum_{j=1}^L N_j} n$

$\hat{A}_{NSj}$

$j=1$

d) *Afijación para un coste,* en la que se eligen los  $n_j$  de forma que minimicen la varianza para un coste fijo,  $C$  que generalmente se expresa como

$L$

$C = \sum_{j=1}^L c_j n_j$

donde  $c_j$  es el coste de elegir una unidad en el estrato  $j$  y

$j=1$

$c_0$  es el coste inicial, se obtiene los tamaños:

$n_j = \frac{N_j}{\sum_{j=1}^L N_j} \frac{C - c_0}{\sum_{j=1}^L c_j}$

$n_j = \frac{N_j}{\sum_{j=1}^L N_j} \frac{C - c_0}{\sum_{j=1}^L c_j}$

$j=1$

o también

$n_j = \frac{N_j}{\sum_{j=1}^L N_j} \frac{C - c_0}{\sum_{j=1}^L c_j}$

$\sum_{j=1}^L c_j n_j = C - c_0$

$\hat{A}_{NSj}$

$WS_j = \frac{N_j}{\sum_{j=1}^L N_j} \frac{C - c_0}{\sum_{j=1}^L c_j}$

$j=1$

$j=1$

Si se diera el caso que el tamaño de  $n_j$  es mayor que el tamaño del estrato  $N_j$ , evidentemente inviable. El problema se resuelve dando al

estrato  $j$  el tamaño  $N_j$  y repartiendo el resto de unidades  $n_j - N_j$  entre los demás estratos, siempre conservando el tamaño total,  $n$ , de la muestra. Este cambio alteraría la varianza del estimador, que se debería modificar.

Es interesante destacar la influencia del tipo de afijación en el error de muestreo. Sea por ejemplo 3 estratos de tamaños:  $N_1=100$ ,  $N_2=400$  y  $N_3=500$ . Se elige una muestra de  $n=100$ . ¿Qué precisión se tiene con los datos de la siguiente tabla?. Observar el resultado para afijación proporcional y para afijación óptima.

**Estrato  $N_j$   $S_j$**

1 100 50  
 2 400 5  
 3 500 6 1000

a) En la afijación proporcional se tiene:

$$n_j = n \frac{N_j}{N} = 100 \frac{100}{1000} = 10$$

$$V_y = \sum_{j=1}^3 \left( \frac{N_j}{N} \right)^2 S_j^2 \left( \frac{n - N_j}{n} \right)$$

$V_y$

$$\hat{y}_{st} = \sum_{j=1}^3 \frac{N_j}{N} \hat{y}_j$$

$$= \sum_{j=1}^3 \frac{N_j}{N} \hat{y}_j$$

$$N_j = 100, 400, 500$$

$$\hat{y}_j$$

$$= \sum_{j=1}^3 \frac{N_j}{N} S_j^2 \left( \frac{n - N_j}{n} \right)$$

$n$

$$= \sum_{j=1}^3 \frac{N_j}{N} S_j^2 \left( \frac{n - N_j}{n} \right)$$

$$= \frac{1}{100} (100)^2 \left( \frac{100 - 100}{100} \right) + \frac{400}{1000} (5)^2 \left( \frac{100 - 400}{100} \right) + \frac{500}{1000} (6)^2 \left( \frac{100 - 500}{100} \right)$$

$$= 0 + 10 + 15 = 25$$

2

$$..278000 = 2,502 \cdot 100$$

luego el error de estimación será

$$V_{y_{st}}(\hat{\theta}) = 2,502 = 1,5818$$

=

b) En la afijación óptima

$n_j$  NS

$j$

$$n_j = n \cdot \frac{N_j}{N} = 10000 \cdot \frac{30}{10000} = 30$$

$$5000 \cdot 2000$$

$$12 \cdot 3 \cdot 100 \cdot 10000$$

$j$

$j=1$

$$V_{y_{st}}(\hat{\theta}) = L \sum_{j=1}^J \frac{N_j^2 S_j^2}{n_j} = L \sum_{j=1}^J \frac{N_j^2 S_j^2}{n_j}$$

$= \sum_{j=1}^J$

WNn

st

j

N

$$\hat{\theta} = \sum_{j=1}^J \hat{\theta}_j \frac{N_j}{N}$$

$$= \sum_{j=1}^J \hat{\theta}_j \frac{N_j}{N}$$

$$n_j = N_j \frac{n}{N}$$

j

=

como entonces L

$$\hat{\theta} = \sum_{j=1}^J \hat{\theta}_j \frac{N_j}{N}$$

$j=1$

$$\hat{\theta} = \sum_{j=1}^J \hat{\theta}_j \frac{N_j}{N}$$

$$\hat{\theta} = \sum_{j=1}^J \hat{\theta}_j \frac{N_j}{N}$$

Vy

$$V_{y_{st}}(\hat{\theta}) = \sum_{j=1}^J \frac{N_j^2 S_j^2}{n_j} = 110$$

$$V_{y_{st}}(\hat{\theta}) = \sum_{j=1}^J \frac{N_j^2 S_j^2}{n_j} = 110$$

Por tanto el error de muestreo  $V_{y_{st}}(\hat{\theta}) = 0,722 = 0,8497$ . Como se puede apreciar se consigue menor error muestral con la afijación óptima.

### 1.6. Muestreo por conglomerados

Un conglomerado es una colección de elementos que tiene una configuración parecida a la población de que procede.

Cada conglomerado es un grupo natural constituyente de la población, así

por ejemplo: las casas que componen una manzana, las personas que integran una familia.

La principal ventaja que aporta este método es su bajo coste, por la facilidad de elección de los elementos muestrales. Por contra su principal inconveniente es el posible error sistemático que se pueda incurrir al elegir por ejemplo un determinado bloque que sea poco representativo de la población.

El muestreo por conglomerados (monoetápico) en realidad es una variante del aleatorio simple o el estratificado con la única salvedad que en este caso las unidades muestrales no son los individuos sino los propios grupos (conglomerados), elementos naturales de la población. Este tipo de muestreo debe perseguir que la varianza intra-conglomerado sea elevada y la varianza inter-conglomerados baja.

Scheaffer (1986, página 197) aconseja utilizar el muestreo por conglomerados cuando:

- a) No se puede disponer o es muy costoso obtener el marco muestral que lista los elementos de la población, mientras que se puede lograr fácilmente un marco que liste los conglomerados.
- b) El costo por obtener observaciones se incrementa con la distancia que separa los elementos.

La forma de seleccionar una muestra en el muestreo por conglomerados consiste en conformar un marco que liste todos los conglomerados de la población. Después se elegirá una muestra aleatoria simple de este marco.

### ***1.6.1. Estimación de la media, total y proporción poblacional***

El muestreo por conglomerados es un muestreo aleatorio simple donde cada unidad de muestreo contiene un número de elementos.

#### **Nomenclatura**

$N$  = número de conglomerados en la población.  $n$  = número de conglomerados seleccionados en la muestra.  $m_i$  = número de elementos en el conglomerado  $i$ ,  $i = 1, \dots, N$ .

$m$

=

1

$n$

$\hat{A}m_i$ , tamaño promedio del conglomerado en la muestra.  $n \sum_{i=1}^N$

$N$

$= \hat{A}Mm_i$ , número de elementos de la población.

$i=1$

$M = \bar{M}$ , tamaño promedio del conglomerado en la población.  $N$

### Media

El estimador de la media poblacional será:

$$\hat{y} = \frac{\sum_{i=1}^n \hat{y}_i}{n}$$

El estimador de la varianza de

$\hat{y}$  será:

$$\hat{V}(\hat{y}) = \frac{N-1}{N} \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{n-1}$$

Por tanto el intervalo de confianza de la media es:

$$\hat{y} \pm k \sqrt{\hat{V}(\hat{y})}$$

### Ejemplo 3.1.3

Se realiza entrevistas en 10 hogares españoles sobre los gastos en esparcimiento (ocio) en un pueblo de 400 hogares. Los resultados se presentan en la tabla siguiente:

#### Conglomerado Número de personas Gastos en ocio $m_i$ (euros) $y_i$

1	4	120
2	6	72
3	8	320
4	10	240
5	4	82
6	6	76
7	7	94
8	5	72
9	3	60
10	2	42

Estimar la media y calcular el error de estimación. La media o gasto medio por persona será:

$$\hat{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{1178}{55} = 21,418$$

$i=1$

$n$

$\hat{A}$

$( )^2$

$V_y N n$

$ii$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = 400 - 10 \cdot 34285 = 19645 = N n M^2 n - 1_{400} 10 \cdot \hat{E} 55 \hat{\sim}^2 9 \hat{E} \hat{A} 400 \hat{\sim}$$

por tanto el error de estimación será:

$$V_{\hat{y}}() = 140,16$$

### Total poblacional

El estimador del total poblacional será:  $\hat{Y} M y$

$\hat{\sim}()$

La estimación de la varianza de es:  $Y V M y M V y^2 \hat{~}()$

Por tanto el intervalo de confianza será:  $Y k V Y \pm \hat{\sim}(\hat{~})$

### Proporción poblacional p

Sea  $a_i$  el número de elementos en el conglomerado  $i$  que posee la característica en estudio.

$n$

$\hat{A} a_i$

$\hat{=} 1 p = n$

$\hat{A} m_i$

$i=1$

El estimador de la varianza de

$\hat{^}$

$p \hat{^}$

$n$

$\hat{A}$

$($

$$a p m \hat{^})^2 V p N n^{ii}$$

$$\hat{~}(\hat{~}) \cdot \sum_{i=1}^n$$

$N n M^2 n - 1$  y su intervalo de confianza:  $\hat{^} p k V p \hat{~}(\hat{~})$

### Ejemplo 3.1.4

Supóngase el ejemplo 3.1.3 donde además se quiere saber la proporción de personas de cada hogar que son titulados superiores. La tabla se completa con los siguientes datos:

#### Conglomerado $i$

Número de personas  $m_i$  de titulados  $a_i$  ocio (euros)

$y_i$

1 4 2 120  
 2 6 2 72  
 3 8 3 320  
 4 10 4 240  
 5 4 1 82  
 6 6 2 76  
 7 7 2 94  
 8 5 3 72  
 9 3 1 60  
 10 2 1 42

Se desea saber la proporción de titulados y calcular el error de estimación.

$$\hat{p} = \frac{\sum_{i=1}^m a_i}{n} = \frac{0.3818}{1} = 0.3818$$

$\hat{A}$

$m$

55

$i$

$i=1$

$n$

$\hat{A}$

(

$apm$

$\hat{p}$

$Vp$

$ii$

2

$n$

-

1

$$= \frac{400-10}{400} \cdot \frac{2 \cdot 4264}{55} = 1.3903$$

$$NnM_{400} \cdot \hat{E} 55^{-2} 10 \hat{A} 400^{-1}$$

El error de estimación será  $Vp(\hat{p}) = 1.1791$

### **1.6.2. Tamaño de la muestra para la estimación de la media total y proporción poblacional**

Suponiendo que se ha elegido el tamaño del conglomerado y que se intenta que sea mínima la varianza entre conglomerados, entonces el número de conglomerados para la estimación de la media poblacional será:

$$e^2 Nn_{.s^2=}$$

$k$

$$22^c$$

$NnM$

$n$

$\hat{A}$

$( )^2$

$ii$

donde  $s^2 =_{i=1}$  que se obtiene de una muestra previa.  $c n - 1$

$Ns$

$2 c$

Y por tanto  $n =_{e^2 22 k^2} NM_{+sc}$

### Ejemplo 3.1.5

Supóngase que los datos de la tabla anterior se consideran como previos para

el cálculo del tamaño muestral para la estimación de la media poblacional:

400 34285

$n = 9^a 400$

$005^2, \dots \hat{E} 55^2 34285$

$4 \hat{E} \hat{A} 400^{-\sim} + 9$

Si se quiere obtener el tamaño muestral para el cálculo de total poblacional será:

$$2 M Nn Nn Msn Ns^{22}$$

$c$

$k^2$

$$e =_{2 2}$$

$$k^2 + Ns_c$$

En el caso de la proporción, se vuelve a utilizar una muestra piloto y se obtiene:

$n$

$\hat{A}$

$($

$apm^{\hat{}} )^2$

$ii$

$s^2 =_{i=1}$  con lo cual  $n$  será:  $c n - 1$

$e$

$2$

$=$

$-$

$Nn_{s2}$ , resultado similar al calculado en la estimación de la media poblacional:  $k$

$22^c$

$NnM$

$Ns$

$2 c^n = e^2 22$

$k2 NM_{+sc}$

## 1.7. Muestreo por conglomerados en dos etapas

Aún cuando por su complejidad no vamos a entrar en detalle para éste tipo de muestreo, no obstante se ha querido mostrar al lector una visión general del mismo para que conozca algún método complejo de muestreo.

Una muestra en el muestreo por conglomerados bietápico se obtiene seleccionando primero una muestra aleatoria de conglomerados y posteriormente una muestra aleatoria de los elementos de cada conglomerado elegido en la etapa precedente.

Por ejemplo para conocer la opinión de los universitarios sobre la guerra, se puede seleccionar en primer lugar de forma aleatoria las universidades, para posteriormente elegir una muestra aleatoria de los alumnos de cada universidad seleccionada anteriormente.

Scheaffer (1986, págs. 233-234) propone dos condiciones deseables a la hora de seleccionar la muestra:

a) Proximidad geográfica de los elementos dentro de cada conglomerado. b) Tamaño de conglomerado conveniente para su manejo.

Además añadimos que se necesita un tamaño de conglomerado que controle el coste total de la investigación.

Un muestreo por conglomerados en dos etapas muy utilizado es aquel donde las probabilidades de elección del conglomerado son proporcionales al tamaño del mismo.

### Ejemplo

Un investigador quiere saber la proporción de alumnos que faltan más de 2 días a clase en 6 institutos públicos. Desea elegir 3 institutos y puesto que los mismos varían en número de alumnos, su elección desea realizarla de forma proporcional a su tamaño. Además en los institutos elige un 10% de total de alumnos. Con los datos adjuntos, vamos a calcular la proporción de alumnos solicitada.

### **Instituto Número de alumnos Intervalo acumulado**

1	420	1-420
2	360	421-780
3	620	781-1400
4	480	1401-1880
5	510	1881-2390
6	210	2391-2600

Se procede de la siguiente manera:

Se elige aleatoriamente 3 números del 1 al 2600. Supongamos dan por resultado: 842, 2048 y 108. Mirando en la columna de intervalo acumulado estos números se localizan en los institutos 3, 5 y 1.

Supongamos que en estos institutos el resultado de muestrear al 10% del total arroja los siguientes datos:

#### **Instituto**

**Número de alumnos muestreados Número de alumnos que faltan más de 2 días**

1 42 20 3 62 32 5 51 22

Con estos datos la proporción por instituto será la media muestral, en consecuencia para las tres muestras:

^

$\hat{m}_1 = 20/32$

$\hat{m}_2 = 42/62 = 0,4745$

$\hat{m}_3 = 51/22$

### **1.8. Muestreo sistemático**

La simplicidad en la selección de la muestra ha hecho del muestreo sistemático uno de los procedimientos más utilizados.

Consiste en seleccionar un elemento de los primeros k elementos en el marco y después cada k-ésimo elemento.

Presenta la ventaja frente al muestreo aleatorio simple que es más fácil de llevar a cabo y además proporciona más información por unidad de costo.

Un ejemplo donde se utiliza este tipo de muestreo es en las cadenas de montaje, donde el control de calidad se realiza seleccionando, de forma secuencial, un producto de cada k fabricados.

La forma de seleccionar una muestra n de una población N de forma sistemática cada k elementos con la condición  $k \leq N/n$ , será elegir un número menor de k y luego de forma secuencial cada k elementos elegir los n elementos de la muestra.

En éste tipo de muestreo debe prestarse especial atención al reparto no secuencial de los elementos en el marco. Por ejemplo si se está haciendo un estudio en un colegio mixto y se seleccionan exclusivamente hombres, o en una encuesta de opinión sobre un producto de cosmética femenina salen seleccionadas exclusivamente mujeres mayores. Se puede evitar este problema cambiando al azar, cada cierto tiempo, el punto de partida.

A la hora del cálculo de los estimadores de la media, total y proporción existen tres posibles variantes:

a) *Población aleatoria*

Se dirá que una población es aleatoria si sus elementos están ordenados al azar. En éste caso el muestreo sistemático es equivalente al muestreo aleatorio simple y se puede aplicar las fórmulas vistas para el segundo. Por ejemplo, si se quiere elegir una muestra de investigadores en CC. de la Educación y se coge como población los socios, ordenados alfabéticamente, de AIDIPE. La población estará ordenada al azar y por tanto para estudiar, por ejemplo, el número promedio de investigaciones realizadas durante el año 2002 se podrá utilizar las mismas fórmulas vistas en el muestreo aleatorio simple.

b) *Población ordenada*

Se dirá que la población está ordenada, si los elementos dentro de la población están ordenados de acuerdo con algún esquema previo. Por ejemplo si se ordenan las calificaciones de los alumnos de un curso en una determinada materia y se quiere saber la efectividad en la misma, cogiendo una muestra de calificaciones. En este caso la población de calificaciones está ordenada y el muestreo sistemático producirá menor error de estimación que el muestreo aleatorio simple; no obstante una cota superior del mismo puede ser el error muestral obtenido para el muestreo aleatorio simple.

c) *Población periódica*

Será aquella en que los elementos de la población tienen variación cíclica. Por ejemplo se quiere saber el promedio de ventas diarias en una cadena de tiendas de ropa deportiva. La población de ventas diarias claramente es cíclica. En este caso el error muestral en la estimación de dicho promedio será superior mediante muestreo sistemático, que si se realizara un muestreo aleatorio simple de las ventas diarias.

Para evitar este problema, como ya se comentó, lo que se hace es aleatorizar varias veces el punto de arranque en el muestreo sistemático; de ésta forma se podrá utilizar la misma formulación del muestreo aleatorio

simple.

### 1.9. Muestreo por cuotas

Entre los muestreos no probabilísticos uno de los más utilizados es el muestreo por cuotas donde los elementos se eligen de acuerdo con ciertas instrucciones pero sin la intervención del azar.

Las cuotas de elementos de la muestra se eligen de acuerdo con el criterio del investigador entre las categorías de las variables independientes más relevantes: por ejemplo sexo, nivel económico, nivel de instrucción, etc. En el tamaño de las cuotas se intenta respetar la proporcionalidad con la población objeto de estudio.

Este tipo de muestreo presenta la dificultad del conocimiento del error muestral y por tanto de la representatividad de la muestra, y el desconocimiento de la idoneidad de las cuotas elegidas como variables importantes y representativas de las características del universo.

Presenta como ventaja su bajo coste.

Este tipo de muestreo no probabilístico, a veces, se combina con otro probabilístico como el de conglomerados. Por ejemplo, en principio se eligen al azar las manzanas de una ciudad y en segunda instancia se eligen las personas respetando un esquema de cuotas definido con anterioridad.

Un ejemplo de esquema de asignación por cuotas puede ser el siguiente: se desea entrevistar a un grupo de 40 personas de 3 núcleos de población: rural, urbano y semi-urbano, con representación de hombre y mujeres y de 3 intervalos de edad. El esquema será:

<b>Sexo</b>	<b>Zona</b>	<b>Edad</b>	Mujeres	18	Rural	12	20-35	10
			Hombres	22	Urbana	18	36-45	14
					Semi-urbana	10	46-60	16

*Total 40 Total 40 Total 40*

## 2. ESTIMACIÓN DE HIPÓTESIS. FUNDAMENTOS ESTADÍSTICOS

### 2.1. Inferencia estadística

La **Inferencia Estadística** es la parte de la Estadística que incluye los métodos utilizados para tomar decisiones o para obtener conclusiones sobre una característica desconocida de la población, a partir de la información contenida en una o más muestras representativas de esa población. La herramienta teórica que utiliza es la teoría de la probabilidad.

Por ejemplo: un profesor quiere experimentar un nuevo tipo de método de enseñanza y desea obtener conclusiones sobre la forma en que funcionará una vez que se aplique a gran escala. Para ello toma una muestra de 200 alumnos, y de las conclusiones que obtenga podrá inferir el funcionamiento en toda la población escolar.

La inferencia estadística aborda, a partir de una o varias muestras, problemas tales como:

— Determinación de parámetros y medición del grado de precisión obtenido. — Aceptación o rechazo de un modelo teórico predeterminado. — Discriminación entre modelos.

La Inferencia Estadística, estudia principalmente dos tipos de problemas:

a) La **Estimación**: consiste en determinar una característica desconocida de la población. Ejemplo: tasa media de aprendizaje con el citado método de enseñanza.

Puede ser:

**Puntual**: determinar el valor concreto.

**Por intervalos**: determinar un intervalo en el que esté contenido con cierto grado de probabilidad.

b) El **Contraste de hipótesis**: determinar si es aceptable, a partir de los datos muestrales, que la característica estudiada tome un valor predeterminado o pertenezca a un intervalo concreto. Ejemplo: ¿es la tasa media de aprendizaje del método mayor del 10%? ¿la tasa media de aprendizaje tiene una distribución normal?

## 2.2. Distribuciones asociadas al proceso de muestreo

Para centrarnos, veamos uno de los problemas que trata este apartado. Supongamos dos variables aleatorias independientes que son los posibles resultados que se obtienen al tirar un dado. Pensemos en el estadístico su distribución aparece en la tabla 3.2.1.

TABLA 3.2.1. Distribución media muestral.

—

$x$	1	1,5	2	2,5	3	3,5	4	4,5	5	5,5	6	Pr	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36
-----	---	-----	---	-----	---	-----	---	-----	---	-----	---	----	------	------	------	------	------	------	------	------	------	------	------

Las muestras que se escojan en un proceso de inferencia tienen que ser representativas de la población y el número de elementos si se trata de una población finita puede ser:

- con reemplazamiento .....  $n^N$
- sin reemplazamiento .....  $N$

$n()$

Donde  $N$  tamaño de la población y  $n$  tamaño de cada muestra.

–

En el muestreo con reemplazamiento, la media muestral  $\bar{X}$  verifica:

– –

$$E(\bar{X}) = \mu \quad \text{Var}(\bar{X}) = \frac{s^2}{n}$$

–

En el muestreo sin reemplazamiento, la media muestral  $\bar{X}$  verifica:

$$E(\bar{X}) = \mu \quad \text{Var}(\bar{X}) = \frac{s^2}{n} \left( \frac{N-n}{N} \right)$$

Para obtener la distribución en el muestreo de un estadístico existen distintos métodos, cuya misión es facilitar el proceso de cambio de variable aleatoria:

1. Muestreo artificial o de Montecarlo.
2. Método de la función característica.
3. Cambio de variable matemática en la integral.
4. Método geométrico.
5. Cambio de variable aleatoria.
6. Método de inducción.
7. Métodos de aproximación asintótica.
8. Métodos de cálculo aproximado.

Vamos a utilizar el método de muestreo artificial o de Montecarlo para obtener aproximadamente la distribución en el muestreo de la media:

$X_1, X_2, X_3, X_4$

$= + + +$

4

de una muestra de extensión 4 de una población  $N(20,5)$ . Por el método del muestreo artificial se han obtenido 40 muestras de cuatro valores cada una, según se muestra en la tabla 3.2.2. Para cada una de estas muestras calculamos la media:  $\bar{X} = \frac{X_1 + X_2 + X_3 + X_4}{4}$  y se obtiene los resultados de la tabla 3.2.3.

–

Si se dibuja el histograma, será una aproximación de la curva de densidad de  $X$ , que como puede verse en la fig. 3.2.1 se aproxima a la normal.

TABLA 3.2.2. Muestra artificial.

17 14 22 16 16 18 17 30 21 23

24 16 16 19 23 21 18 14 22 15  
 31 24 23 30 15 21 15 13 21 23  
 10 27 24 22 21 29 28 14 23 22  
 23 20 19 24 19 19 31 22 20 21  
 14 16 20 25 21 12 18 21 15 23  
 16 23 18 24 17 12 18 17 10 18  
 19 17 27 22 25 22 18 31 16 17  
 20 13 18 20 24 21 17 27 17 21  
 18 21 22 25 16 25 27 22 22 14  
 13 14 17 17 21 18 16 20 26 15  
 17 22 18 20 21 12 18 14 23 19  
 26 20 26 22 17 21 19 14 16 22  
 20 25 18 28 19 24 29 23 17 20  
 30 19 27 25 27 18 15 20 22 19  
 17 25 19 21 24 23 15 12 21 15

T ABLA 3.2.3. Medias de cada muestra.

20.50 20.25 21.25 21.75 18.75 22.25 19.50 17.75 21.75 20.75  
 18.00 19.00 21.00 23.75 20.50 16.25 21.25 22.75 15.25 19.75  
 17.00 17.50 18.75 20.50 20.50 19.00 19.50 20.75 22.00 17.25  
 23.25 22.25 22.50 24.00 21.75 21.50 19.50 17.25 19.00 19.00  
 10  
 8  
 6  
 4  
 2

Desv. típ. = 2.09 Media = 20.1

0 N = 40.00 15.0 16.0 17.0 18.0 19.0 20.0 21.0 22.0 23.0 24.0

**Media**

**FIGURA 3.2.1. Representación gráfica de la variable media (distribución muestral de medias).**

A partir de los valores de la tabla 3.2.3 se puede calcular la media y la desviación típica, comprobándose que la media (20,1) es aproximadamente igual a la media de la población y que la desviación típica (2,09) está cerca del valor<sup>5</sup>. La desviación típica de una distribución en el muestreo de un estadístico se suele llamar **error típico**.

Los resultados de este procedimiento confirman las expresiones de la

distribución de la media muestral que antes se habían adelantado y constituye un método útil cuando el cálculo matemático presenta dificultades.

Se presentan a continuación algunos conceptos necesarios para comprender la distribución en el muestreo de algunos estadísticos:

### A. Grados de libertad

Siguiendo a Spiegel (1992) definiremos grados de libertad de un estadístico, generalmente denotado por  $n$ , como el número  $n$  de observaciones independientes en la muestra (o sea, el tamaño de la muestra) menos el número  $k$  de parámetros de la población, que debe ser estimado a partir de observaciones muestrales  $n=n-k$

También se puede definir como el número de comparaciones lineales independientes que se pueden hacer entre las  $n$  observaciones. Así, si con 3 puntuaciones hemos de sumar 30, entonces se puede elegir 2 libremente y la tercera estará condicionada, luego el número de grados de libertad será  $2=n-1$ .

### B. Algunas distribuciones continuas

#### 1. Distribución $N(0,1)$

Una variable continua  $X$  se distribuye  $N(0,1)$  si su campo de variación es el eje

real,

•

<

¥

<+

•

y su

**función de densidad**

es

$$f(x) =$$

$$1 e^{-x^2}$$

2

$\rho$

2

siendo la **función de distribución**

$$x -t^2$$

$$F(x) = P(X \leq x) = \int_{-\infty}^x e^{-t^2} dt$$

La densidad de la variable aleatoria  $N(0,1)$  es simétrica y presenta dos asíntotas en  $\pm\infty$ . Presenta un máximo para  $x=0$ . En la fig. 3.2.2 se puede apreciar las dos funciones, donde la de densidad es la conocida campana de Gauss.

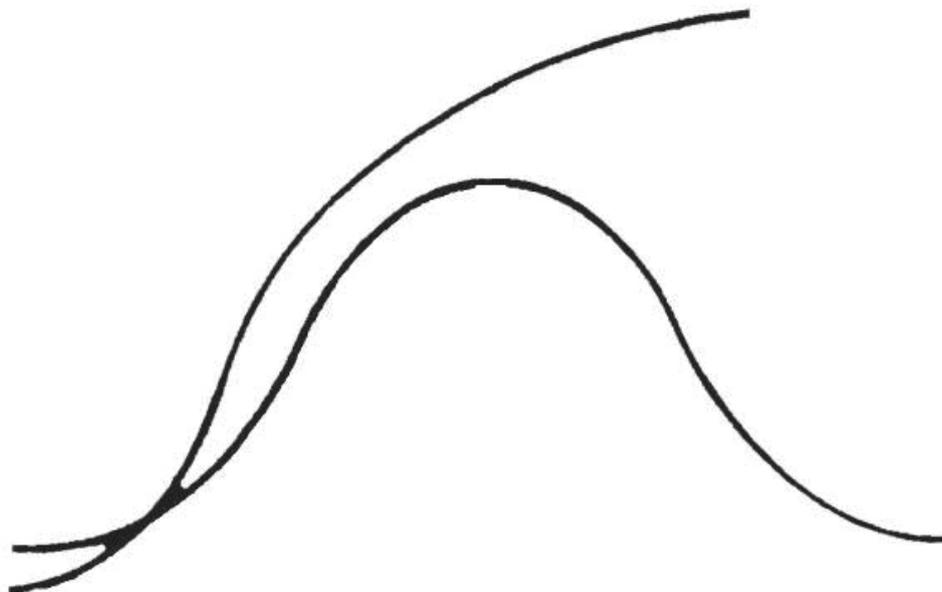
**Parámetros de la distribución Normal  $N(0,1)$**  a) Media:  $m=0$

b) Varianza:  $V(X)=s^2=1$

c) Desviación estándar:  $s=1$

1

$2p$



0

FIGURA 3.2.2. Funciones de densidad y de distribución de una  $N(0,1)$ .

**Aproximación a la distribución Normal  $N(0,1)$  de algunas distribuciones discretas.**

Veamos el esquema siguiente:

$n$  grande  $np > 10$   $N(0,1)$

$l > 20$   $l$  grande

Binomial  $(n,p)$  Poisson  $l$   $n$  grande  $p < 0,1$ ;  $np = l < 5$

Si la distribución Binomial se aproxima a la de Poisson entonces  $l = np$  y si la aproximación es hacia la distribución Normal entonces  $z = \frac{x - np}{\sqrt{npq}}$  generando una  $N(0,1)$ .

En el caso de la distribución de Poisson  $z = \frac{x - l}{\sqrt{l}}$  será una  $N(0,1)$ .

## 2. Distribución $N(m,s)$

Consideremos dos parámetros  $m$  y  $s$ , con campos de variación  $-\infty < m < \infty$  y  $s > 0$ , y definimos una nueva variable aleatoria  $Y = m + sX$ , a partir de la variable aleatoria  $X$  distribuida  $N(0,1)$ . El campo de variación de  $Y$  es, como el de  $X$ , el eje real  $(-\infty, \infty)$ ; la **distribución de probabilidad** de la nueva variable será:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

La expresión de la **función de distribución** de la variable aleatoria  $N(m,s)$  es igual a:

$$G(y) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^y e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

### Propiedades de la función de densidad

- Es simétrica respecto al eje que pasa por la abscisa  $x = m$ , o de otra forma, la media, mediana y moda coinciden.
- Es asintótica al eje de abscisas.
- Posee un máximo en  $x = m$ , de valor  $\frac{1}{\sigma\sqrt{2\pi}}$ .
- Existen puntos de inflexión, a distancia  $s$  del eje de simetría. — El área encerrada es igual a la unidad.

- Parámetros de la distribución Normal  $(N(m, s))$**
- a) Media:  $m$
  - b) Varianza:  $V(X) = s^2$
  - c) Desviación estándar:  $s$

### Manejo de tablas de la distribución $N(0,1)$

La obtención de probabilidades de sucesos relacionados con la variable

aleato

$e$

-

$x^2$

ria  $N(0,1)$  es muy complicada pues la función  $^2$  no tiene primitiva siendo preciso recurrir a procedimientos aproximados. Para evitar esta situación se ha construido tablas que proporcionan aproximaciones de las probabilidades.

A continuación presentamos diversos **ejemplos** de cálculo de dichas probabilidades:

1.  $P(Z \geq 0,56) = 0,2877$
2.  $P(Z \in [-0,24, 0,24]) = 0,4052$
3.  $P(Z \in [1,36, 1,36]) = 0,9131$
4.  $P(Z \geq -2,5) = 0,9938$
5.  $P(0,30 < Z \in [2,89, 2,89]) = 0,3802$
6.  $P(-0,70 \in Z \in [-0,15, -0,15]) = 0,1984$
7.  $P(-1,12 < Z \in [1,63, 1,63]) = 0,8170$

Para el cálculo de probabilidades en sucesos de variables  $N(m, s)$  se normaliza

la variable  $(y-m)/s$  cuya distribución es  $N(0,1)$ .

### **C. Distribuciones derivadas de la distribución normal**

#### **1. Distribución $\chi^2$ de Pearson**

Sean  $n$  variables aleatorias independientes  $X_1, \dots, X_n$  distribuidas  $N(0,1)$ . Se define la variable  $Y = X_1^2 + \dots + X_n^2$ , que recibe el nombre de  $\chi^2(n)$  siendo  $n$ , número de variables aleatorias que la integran, denominado **grados de libertad**.

Como el campo de variación de las variables normales es el intervalo  $(-\infty, \infty)$  y la variable  $\chi^2(n)$  es suma de sus cuadrados, su campo es el intervalo  $[0, \infty)$ .

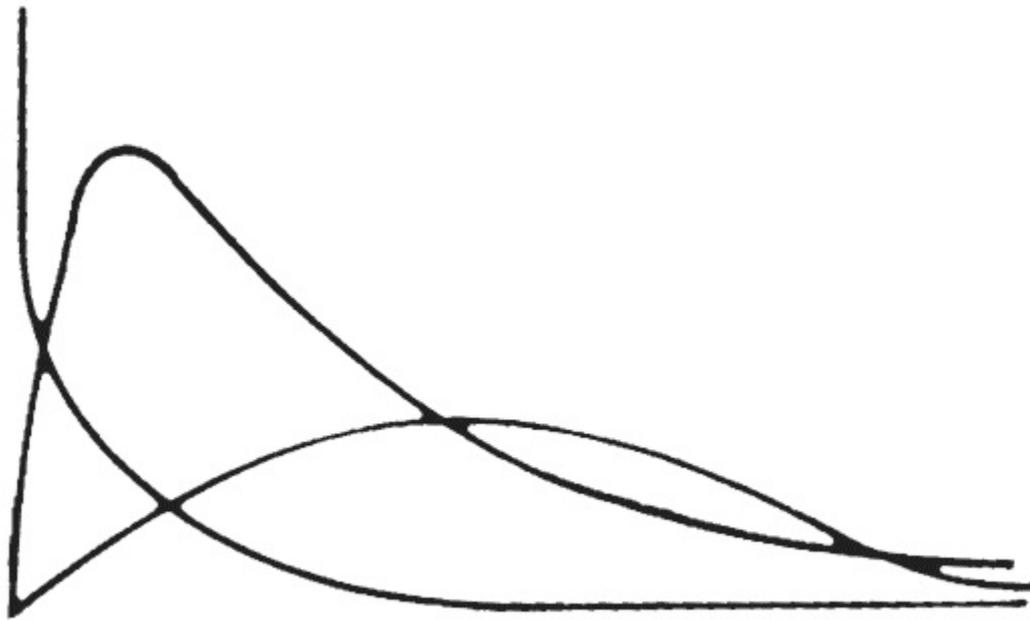
#### **Parámetros de la distribución ji-cuadrado**

a) Media:  $n$

b) Varianza:  $V(X) = 2n$

La forma de la función de densidad de probabilidad (ver fig. 3.2.3) cambia con los valores que toma  $n$  (grados de libertad).

$f(x)$



c<sup>2</sup> FIGURA 3.2.3. Funciones de densidad de la distribución ji-cuadrado para distintos grados de libertad.

### Manejo de tablas

Veamos varios ejemplos donde se utilizan la tabla de la ji-cuadrado: a)  $P(3,94 \leq \chi^2(10) \leq 15,987) = 0,85$ .

b)  $P(8,231 \leq \chi^2(19) \leq a) = 0,725$  hállese a, después de cálculos resulta  $a = 22,718$ . Para valores elevados de n (mayores que 30) la variable aleatoria  $z = \frac{\chi^2 - n}{\sqrt{2n}}$  es aproximadamente  $N(0,1)$ .

c)  $P(1027 \leq \chi^2(1000) \leq 1061) = 0,1820$  será equivalente a calcular  $P(0,61 \leq z \leq 1,36) = 0,1820$ .

d)  $P(1432 \leq \chi^2(1527) \leq a) = 0,2047$  hállese a, después de cálculos resulta  $a = 1488,74$ .

### 2. Distribución t de Student

Sean  $n+1$  variables aleatorias  $N(0,s)$  e independientes  $(y, y_1, y_2, \dots, y_n)$ .

Definimos la variable t de Student con n grados de libertad  $t(n)$ , como:

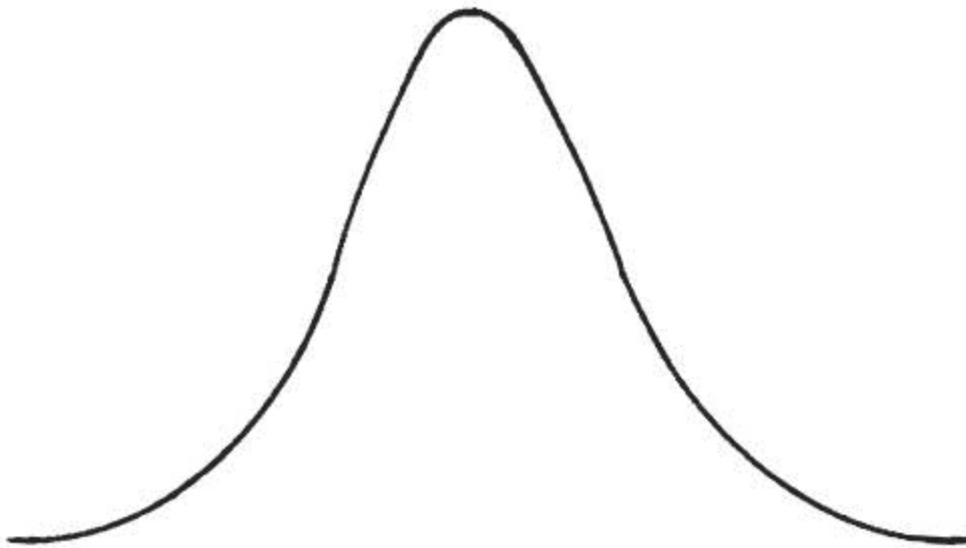
$$t(n) = \frac{y}{\sqrt{\frac{y_1^2 + y_2^2 + \dots + y_n^2}{n}}}$$

1 2 n

n

el número de grados de libertad es igual al número de variables que figuran en el denominador de  $t(n)$ . El campo de variación de la variable  $t(n)$  es el intervalo  $(-\infty, \infty)$ .

$f(x)$



0 t (n) FIGURA 3.2.4. Función de densidad de la distribución t.  
 La función de densidad de la distribución t- Student es simétrica respecto al eje de ordenadas y tiene dos asíntotas en  $\pm \infty$ .

La importancia de esta variable aleatoria reside en el hecho de no depender de la varianza de las variables que la integran, no figura en la función de densidad, y su utilidad se verá plenamente al estudiar la Inferencia Estadística y para el caso de  $s^2$  desconocida.

Si llamamos  $l_0, l_1, \dots, l_n$  a sendas variables aleatorias  $N(0,1)$  de acuerdo con la definición de la normal tenemos:

$$y_0 = sl_0, y_1 = sl_1, \dots, y_n = sl_n$$

sustituyendo en (\*) tenemos:

$$sl_0^2 + sl_1^2 + \dots + sl_n^2 = s^2(l_0^2 + l_1^2 + \dots + l_n^2)$$

luego también podemos decir que:

$$t(n) = \frac{\bar{y} - \mu}{s/\sqrt{n}} \text{ con } l_0 \in N(0,1)$$

n

**Parámetros de la distribución t-Student** a) Media: 0

b) Varianza:  $V(X) = n/(n-2)$

**Manejo de tablas**

Para el manejo de las tablas de la  $t(n)$  es importante recordar que tal distribución es simétrica.

a)  $P(t(7) \leq 1,1192) = 0,85$

b)  $P(-0,723 \leq t(22) \leq a) = 0,18$  hállese a, después de cálculos tenemos  $a = -0,256$  Cuando el número de grados de libertad es elevado (por ejemplo

mayor que

30), utilizaremos la aproximación  $N(0, \frac{1}{n-2})$

507

$$P(t(507) \geq 0,4) = 0,3449$$

505

### 3. Distribución F de Fisher-Snedecor

Consideremos  $m+n$  variables aleatorias  $N(0,s)$  e independientes  $y_1, y_2, \dots, y_m$  y  $z_1, z_2, \dots, z_n$ , la variable

$F(m,n)$

$$F(m,n) = \frac{y_1^2 + y_2^2 + \dots + y_m^2}{z_1^2 + z_2^2 + \dots + z_n^2}$$

sigue la distribución F de Fisher-Snedecor con  $m$  y  $n$  grados de libertad.

$f(x)$

$f(x)$

$$f(x) = \frac{1}{B(\frac{m}{2}, \frac{n}{2})} \left(\frac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m}{2}-1} (1 + \frac{m}{n}x)^{-\frac{m+n}{2}}$$

$f(x)$

sigue la distribución F de Fisher-Snedecor con  $m$  y  $n$  grados de libertad.

$f(x)$



$F(n)$

FIGURA 3.2.5. Función de densidad de la distribución F de Fisher-Snedecor.

Las principales características de esta función son las siguientes: no depende de la varianza de las variables integrantes, no es simétrica y su campo de variación, como producto de cuadrados, es el intervalo  $[0, \infty)$ .

Veamos que no depende de  $s$ . Por la definición de normal podemos

poner:  $y_0 = sa_0$   $y_1 = sa_1$  .....  $y_m = sa_m$

$z_0 = sb_0$   $z_1 = sb_1$  .....  $z_n = sb_n$

$122 + as$   $as222 + \dots + asm222(2 + \dots + 2)$  luego  $F(m,n) = \frac{m sa am}{22 + bs bs^2 2 + \dots + 2bs^2 = n 2(2 + \dots + 2)}$

$1 2 n m_1 sb b_n$

$n$

que nos lleva a la expresión:

$F(m,n) =$

$n^2(m)c m c^2(n)$

variable que puede expresarse como cociente de dos  $c^2$  que, como sabemos, están distribuidas  $N(0,1)$  y no dependen de  $s^2$ .

### Propiedades

Se cumple que  $F(m,n) \cdot F(n,m) = 1$  que nos lleva a  $F(n,m) = 1/F(m,n)$  También  $F(1,n) = t^2(n)$

Directamente relacionada con esta variable está la denominada  $z$  de Fisher  $z = \ln F$ , cuyo campo de variación es  $(\cdot, \cdot)$ .

### Manejo de tablas

a)  $F(5;30) = 3,70$  para  $a = 0,01$

b)  $F(5;32) = 3,65$  para  $a = 0,01$

### D. Relación entre la distribución normal, ji-cuadrado, t de Student y F de Snedecor

Existe relación entre las distribuciones mencionadas, su expresión se representa en la fig. 3.2.6.

Distribución F

$F(1,n) = t_{n2}^2$   $F(1,\cdot) = z^2$   $F(n,\cdot) = C_{n2}/n$

Distribución t de Student

Distribución Normal Distribución Ji-cuadrado

$t_{\cdot} = z$   $C_{12} = z^2$

**FIGURA 3.2.6. Relación entre la distribución normal, ji-cuadrado, t de Student y F de Snedecor.**

**E. Teorema Central del Límite** Este teorema que se aplica tanto a la media muestral

—

$\bar{x}$  como a la  $\hat{A}X_i$  esta

$i=1$

blece que cuando es grande el tamaño de la muestra  $n$ , la distribución muestral de la media tenderá aproximadamente una distribución normal con media  $m$  y una desviación estándar de  $s/\sqrt{n}$ . Si se trata de la suma tenderá a una distribución normal de media  $nm$  y desviación estándar  $s/\sqrt{n}$ . La aproximación será cada vez más exacta a medida que  $n$  se haga cada vez mayor.

### **2.2.1. Distribución en el muestreo de algunos estadísticos**

A continuación hacemos un breve repaso de las aplicaciones de los estadísticos normal,  $t$  de Student, ji-cuadrado y  $F$ , en la distribución de ciertos estadísticos en el muestreo, sobre la base de los siguientes criterios:

1. Condiciones de aplicación.
2. Forma del estadístico.
3. Distribución del estadístico.

#### **A. Media y proporción muestrales**

*Caso 1:*

1.  $\bar{x}$  es la media de una muestra de tamaño  $n$  de una población normal  $N(m,s)$  de varianza  $s^2$  conocida.
2.  $(\bar{x} - m)/(\hat{s}/\sqrt{n})$  es aproximad

3. Si la población es aproximadamente normal, la condición 2 es aproximadamente válida. Para  $n \geq 30$  la aproximación es suficientemente válida. Si la población no es aproximadamente normal, para  $n \geq 100$ , 2 es aproximadamente válida.

*Caso 2:*

1.  $\bar{x}$  es la media de una muestra de tamaño  $n$  de una población normal  $N(m,s)$  de varianza  $s^2$  desconocida, siendo  $\hat{s}^2$  la cuasi-varianza muestral.
2.  $(\bar{x} - m)/(\hat{s}/\sqrt{n})$  es una  $t_{n-1}$  [t de Student con  $(n-1)$  g.l.]

$m$

3. Se usa preferentemente en muestras pequeñas ( $n < 30$ ).

*Caso 3:*

1.  $\bar{x}$  es la media de una muestra de tamaño  $n$  de una población normal  $N(m,s)$  de varianza  $s^2$  desconocida, siendo  $\hat{s}^2$  la cuasi-varianza muestral. Si  $n \geq 30$  entonces 2.  $(\bar{x} - m)/(\hat{s}/\sqrt{n})$  es aproximadamente  $N(0,1)$

3. Si la población es aproximadamente normal y  $n \geq 30$ , 2 es aproximadamente válido. Si la población no es aproximadamente normal pero  $n \geq 100$ , es aproximadamente válido.

Caso 4:

1.  $p$  es la proporción muestral de los elementos principales (éxitos) de una binomial  $B(n,P)$  siendo  $n \geq 25$  y  $0,1 < P < 0,9$  entonces

2.  $(p-P)/\sqrt{PQ}$  es aproximadamente  $N(0,1)$ .

### B. Varianza muestral

1.

$s^2$

2

$s^2$  es la cuasi-varianza de una muestra de tamaño  $n$  de una población normal.

2.  $(n-1)$

$s^2$

2

$(n-1)s^2$  es una ji-cuadrado con  $(n-1)$  g.l.

### C. Diferencia de medias y proporciones muestrales

Supondremos tomadas dos muestras  $(n_1, \bar{x}_1, s_1^2)$  y  $(n_2, \bar{x}_2, s_2^2)$  de dos poblaciones normales independientes  $N(\mu_1, \sigma_1^2)$  y  $N(\mu_2, \sigma_2^2)$  respectivamente.

Caso 1:

1. varianzas conocidas  $s_1^2$  y  $s_2^2$ .

2.  $Z$  es  $N(0,1)$ .

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

3. Si las poblaciones son normales y  $n_1 \geq 30$  y  $n_2 \geq 30$  o siendo aproximadamente normales es  $n_1 \geq 100$  y  $n_2 \geq 100$ ,  $Z$  es válida aproximadamente. Caso 2:

1. varianzas desconocidas y  $n_1 \geq 30$  y  $n_2 \geq 30$ .

2.  $T$  es aproximadamente  $N(0,1)$ .

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

Caso 3:

1. varianzas desconocidas pero iguales ( $s_1^2 = s_2^2$ )

**$\bar{X} - \bar{Y}$  es una t de Student con  $(n_1+n_2-2)$  g.l.**

$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

+

$(0, 1)$

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

$(0, 2)$

$n_1, n_2 \geq 30$

3. Se utiliza preferentemente en muestras pequeñas ( $n_1 \geq 30, n_2 \geq 30$ ).

Si las poblaciones son aproximadamente normales y  $n_1 \geq 30$  y  $n_2 \geq 30$ ,  $\bar{X} - \bar{Y}$  es aproximadamente  $N(0,1)$ . Si las poblaciones no son aproximadamente normales pero  $n_1 \geq 100$  y  $n_2 \geq 100$ ,  $\bar{X} - \bar{Y}$  es aproximadamente  $N(0,1)$ .

Caso 4:

1. varianzas desconocidas no necesariamente iguales.

**$\bar{X} - \bar{Y}$  es una t de Student con  $m$  (g.l.) donde**

$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

$m = \frac{2n_1n_2}{n_1+n_2}$

$(0, 1)$

$n_1, n_2 \geq 30$

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

$(0, 2)$

$n_1, n_2 \geq 100$

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

$m = \frac{2n_1n_2}{n_1+n_2}$

$n_1, n_2 \geq 30$

$$\hat{\sigma}^2 = \frac{1}{n_1 + n_2 - 2} \left( \sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2 \right)$$

Caso 5:

1.  $p_1$  y  $p_2$  son las proporciones muestrales del suceso principal (éxito) de dos poblaciones binomiales independientes  $B(n_1, P_1)$  y  $B(n_2, P_2)$  respectivamente. Si se cumple  $n_1, n_2 \geq 25$  y  $0,1 < P_1, P_2 < 0,9$  entonces:

## 2.0() es aproximadamente $N(0,1)$

$$\frac{PQ}{11+22}$$

$$n_1 n_2$$

### D. Cociente de varianzas muestrales

1.  $s_1^2$  y  $s_2^2$  son las cuasi-varianzas de dos muestras (de tamaño  $n_1$  y  $n_2$ , respectivamente) de dos poblaciones normales independientes,  $N(m_1, s_1)$  y  $N(m_2, s_2)$ . Entonces:

2.

$s_1^2 / s_2^2$

$$s_1^2 / s_2^2$$

$s_1^2, s_2^2$  son las cuasi-varianzas muestrales.

$s_1^2 / s_2^2$  es una F de Snedecor con  $(n_1-1, n_2-2)$  g.l., donde  $n_1, n_2$

son los tamaños muestrales.

cuasi-varianzas muestrales.

### 2.3. Estimación estadística

La función de densidad de una variable aleatoria (v.a.)  $X$  depende de algunos parámetros, tales como la esperanza y la varianza. Los verdaderos valores de estos parámetros son desconocidos en la práctica, debiendo ser estimados a partir de una muestra aleatoria simple de  $X$ .

Sea  $f(x, q)$  la función de densidad de  $X$ ,  $q$  el parámetro desconocido. Se llama estimador de  $q$  a una v.a. función de la muestra.

$$U = g(x_1, x_2, \dots, x_n)$$

que, en algún sentido que precisaremos, alcanzan valores próximos al

valor desconocido de  $q$ .  $U$  es por lo tanto un estadístico cuya distribución dependerá de  $f$ ,  $q$  y  $n$ . Obtenida una muestra aleatoria simple  $x_1, x_2, \dots, x_n$ , entonces

$$\hat{q} = g(x_1, x_2, \dots, x_n)$$

es el valor numérico que se asigna al parámetro y recibe el nombre de estimación de  $q$ .

Es decir un **estimador** es una regla que expresa cómo calcular la **estimación**, basándose en la información de la muestra y se enuncia, en general, mediante una fórmula.

$$\hat{X}_i$$

Por ejemplo, la media muestral  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  es un **estimador puntual** de la  $\mu$

—

media poblacional  $m$ , es decir  $E[\bar{X}] = m$ , y explica exactamente cómo puede obtenerse el valor numérico de la **estimación**, una vez conocidos los valores muestrales  $X_1, X_2, \dots, X_n$ . Por otra parte, un **estimador por intervalos**, utiliza los datos de una muestra para determinar **dos** puntos que pretenden abarcar el valor real del parámetro estimado. Por ejemplo, para la estimación de un parámetro como la media  $m$  se da un intervalo de confianza que incluye al estimador:  $(\bar{X} - e, \bar{X} + e)$  y luego se exige una probabilidad de que el valor  $m$  caiga dentro de ese intervalo:

— —

$$P(\bar{X} - e < m < \bar{X} + e) = 1 - \alpha$$

### Propiedades de los estimadores

Entre los posibles estimadores de un parámetro  $q$ , debemos seleccionar aquel o aquellos que gocen de buenas propiedades. Un buen estimador es aquel que reúne ciertas propiedades: insesgado, consistente, eficiente y suficiente.

#### a) Sesgo

Se dice que  $U$  es **estimador insesgado de  $q$**  si su valor medio es  $q$   $E(U) = q$

si  $E(U) \neq q$  se dice que el estimador es sesgado, con sesgo positivo si  $E(U) > q$  y sesgo negativo si  $E(U) < q$ .

Es interesante que un estimador sea insesgado, porque tomará valores que estarán alrededor del verdadero valor del parámetro de  $q$ . Un estimador que sea sesgado tenderá a tomar valor sistemáticamente desviados de  $q$ . Por ejemplo, la media muestral es un estimador insesgado de la media

poblacional. Sin embargo la varianza muestral es un estimador sesgado de la varianza poblacional, pero este sesgo desaparece cuando  $n$  tiende a infinito, y se dice entonces que la varianza muestral es un estimador asintóticamente insesgado de la varianza poblacional.

Aunque parece una condición indispensable la carencia de sesgo en los estimadores, sin embargo esto no es así, pues existen estimadores sesgados, que presentan menor variabilidad que los insesgados y son utilizados para la estimación.

### b) Consistencia

El estadístico  $\hat{q} = g(x_1, x_2, \dots, x_n)$  es un estimador consistente de  $q$  si la probabilidad de que el valor del estadístico se aproxime al del parámetro  $q$ , se aproxima a la unidad conforme el tamaño de la muestra aumenta

$\lim_{n \rightarrow \infty} P(|\hat{q} - q| < \epsilon) = 1$  ó  $\lim_{n \rightarrow \infty} P(|\hat{q} - q| \geq \epsilon) = 0$  para todo  $\epsilon > 0$  arbitrariamente pequeño.

En otras palabras a medida que aumentamos el tamaño de la muestra la imagen de la población que nos proporciona será, en general, mejor y, por tanto, nuestras estimaciones podrán ser mejores; dicho de otra manera, parece lógico que procuremos escoger nuestros estimadores de forma que sus cualidades mejoren a medida que aumenta el tamaño de la muestra. Esta propiedad se revela importante cuando el tamaño de la muestra es grande. Por ejemplo la media muestral es un estimador consistente de la media poblacional.

### c) Eficiencia

Supongamos que  $U_1$  y  $U_2$  son dos estimadores insesgados de  $q$ . Se dice que  $U_1$  es más eficiente que  $U_2$  si

$$\text{var}(U_1) < \text{var}(U_2)$$

$U_1$  será un estimador más deseable para estimar  $q$ , porque los valores que puede tomar estarán menos dispersos alrededor del verdadero valor de  $q$  que  $U_2$ .

El cociente  $e = \text{var}(U_2) / \text{var}(U_1)$  se llama eficiencia relativa de  $U_1$  respecto de  $U_2$ . Tendremos  $0 \leq e \leq 1$ . Si  $e$  es próximo a 0,  $U_1$  es mucho mejor que  $U_2$ . Si  $e=1$ ,  $U_1$  y  $U_2$  tienen la misma eficiencia.

Un ejemplo clásico donde queda patente la eficiencia relativa es la estimación de la media poblacional ( $\mu$ ) mediante la media muestral y la

mediana muestral. Cuando la población es normal, la media y la mediana son estimadores insesgados y consistentes de  $m$ . Para comparar su eficiencia se calcula

$$\frac{s^2}{n}$$

sus varianzas. La varianza de la media es:  $\frac{s^2}{n}$  y la varianza de la mediana es:

$$\frac{1,57 s^2}{n}$$

$\frac{1,57 s^2}{n}$ . Por tanto en distribuciones iguales la varianza de la media es menor que la varianza de la mediana, y en consecuencia la media muestral es un estimador más eficiente que la mediana muestral.

### **Varianza mínima**

Fijado un tamaño muestral  $n$ , si entre todos los estimadores insesgados de  $q$  podemos encontrar uno cuya varianza sea mínima, entonces, adoptando como criterios el sesgo y la eficiencia, habremos conseguido el estimador óptimo para  $q$ .

Bajo ciertas condiciones esta varianza mínima tiene una acotación conocida

^

como *acotación de Frechet-Cramer-Rao* cuya expresión es:  $\text{var}(q) \geq 1/I(q)$ , siendo  $I(q)$  la cantidad de información:

$$I(q) = -E \left[ \frac{\partial \log f(x; q)}{\partial q} \right]^2$$

$$\text{var}(\hat{q}) \geq \frac{1}{n I(q)}$$

Un estimador es asintóticamente eficiente, o simplemente eficiente, si la varianza del estimador es igual al mínimo dado por la desigualdad de *Frechet-Cramer-Rao*.

$$\text{var}(\hat{q}) = \frac{1}{n I(q)}$$

### **d) Suficiencia**

Se considera que un estimador es suficiente cuando él sólo da toda la información muestral posible acerca del parámetro, de tal manera que cualquier otro estimador no pueda proporcionar más que escasa información adicional.

Esta propiedad tiene su trascripción operativa al calcular la estimación

por intervalos de confianza mediante el cálculo del error de estimación. En la estimación puntual no se puede establecer el error cometido al estimar el parámetro.

### **Método de cálculo de estimadores**

Existen diversos métodos de cálculo de estimadores: *método de máxima verosimilitud*, *método de los momentos*, *método de mínimos cuadrados* y *método de la mínima c*. Por limitaciones de espacio y por su importancia, sólo se verá el primero.

### **Método de máxima verosimilitud**

Este método de estimación es, sin duda, el más importante de los que se utilizan en Estadística. Fisher demostró que los estimadores obtenidos por este método son siempre consistentes y son iguales o más eficientes que los obtenidos por otro procedimiento. No poseen, en cambio, con carácter general la propiedad de ser insesgados, aunque si lo sean en muchos casos.

Veamos el siguiente ejemplo recogido de Ríos (1967). Supongamos una urna con cuatro bolas y que solo sabemos que son blancas o negras, pero sin conocer en qué proporción. Si  $p$  es la proporción de blancas, puede ser  $p=0, 1/4, 2/4, 3/4$  y  $1$ .

Si hacemos dos extracciones con devolución y obtenemos una blanca y otra negra. El problema es estimar, a partir de este resultado, la composición de la urna. La idea maestra del método de máxima verosimilitud consiste en admitir que la población base es aquella que daría máxima probabilidad al suceso considerado.

En nuestro ejemplo los posibles resultados de 1 bola blanca y 1 bola negra tendrán las siguientes probabilidades:  $0, 6/16, 8/16, 6/16$  y  $0$ , según la urna tenga de 0 a 4 bolas blancas. Luego la máxima probabilidad es  $8/16$  y admitiendo la idea del método de máxima verosimilitud la composición de la urna será 2 bolas blancas y 2 bolas negras.

Si  $f(x, q)$  es la función de densidad de la población, de la cual tenemos una muestra aleatoria simple  $x_1, x_2, \dots, x_n$  la estimación de máxima verosimilitud es la

$\hat{q} = q(x_1, x_2, \dots, x_n)$ , que hace máxima la llamada *función de verosimilitud*:  $L(x_1, x_2, \dots, x_n, q) = f(x_1, q)f(x_2, q) \dots f(x_n, q)$  donde  $L$  es una función de  $q$  y  $x_1, x_2, \dots, x_n$  se consideran fijos.

Como  $L$  es una función positiva, coinciden los máximos de  $L$  con los de su logaritmo y si es derivable, entonces el máximo se obtiene resolviendo la ecuación:  $\frac{\partial \log L}{\partial q} = 0$

^^

Toda solución  $\hat{q} = \hat{q}(x_1, x_2, \dots, x_n)$  se llama *estimador de máxima verosimilitud* del parámetro  $q$ .

Veamos un ejemplo de cálculo de un estimador de máxima verosimilitud. Supongamos una muestra de extensión  $n$  procedente de una población normal  $N(m, s)$ . El logaritmo de la función de verosimilitud será:

$$\log L = -\frac{1}{2} \sum_{i=1}^n \left[ \frac{(x_i - m)^2}{s^2} + \log s \right]$$

y si queremos obtener un estimador de  $m$  entonces:

$$\frac{\partial \log L}{\partial m} = \frac{1}{s^2} \sum_{i=1}^n (x_i - m) = 0$$

$i$

que despejando  $m$  tenemos:

$$\hat{m} = \frac{1}{n} \sum_{i=1}^n x_i$$

## 2.4. Estimadores por intervalos de confianza para parámetros de la población

Antes de comentar las expresiones de los estimadores de distintos parámetros poblacionales conviene que veamos algunos términos utilizados posteriormente.

### Intervalo de confianza, nivel de confianza y error

Sea  $x_1, x_2, \dots, x_n$  una muestra aleatoria simple de una v.a.  $X$  cuya distribución depende de un parámetro  $q$  (y posiblemente de otros parámetros). Se dice que los estadísticos

$$U = g_1(x_1, x_2, \dots, x_n) \quad V = g_2(x_1, x_2, \dots, x_n)$$

constituyen un intervalo de confianza para  $q$ , o al  $100(1-\alpha)\%$ , si se verifica:

1.  $U < V$  para toda muestra de tamaño  $n$
2.  $P(U < q < V) = 1 - \alpha$  para todos los valores de los demás parámetros.

Entonces  $(U, V)$  constituyen un *intervalo de confianza* para  $q$  con *nivel*

de confianza  $\alpha$ %)%.

El nivel de confianza es elegido por el investigador aunque en la práctica suele ser  $\alpha = 0,99$ .

$\alpha = 0,99$ .

a mayor es el grado de confianza, pero también es mayor el intervalo de confianza y más pequeña la precisión del estimador.

El sentido que debe tomar el intervalo de confianza es frecuentista, es decir, si

El sentido que debe tomar el intervalo de confianza es frecuentista, es decir, si  $\alpha = 0,95$  (por ejemplo), en una larga serie de determinaciones de intervalo  $(U,V)$ , el 95% de los casos cubrirán al verdadero valor de  $\mu$ .

El error que cometeremos en estas apreciaciones viene dado por  $\alpha$  y sus valores más utilizados son 0,05 y 0,01.

### Intervalos de confianza para las medias

—

a) Si  $\bar{x}$  es la media de una muestra de tamaño  $n$  de una población normal  $N(\mu, \sigma^2)$  de varianza  $\sigma^2$  conocida.

$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$  (1) si el muestreo es de una población infinita o finita con reposición.

$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N}{N-1}}$  (2) si el muestreo es de una población finita sin reposición de tamaño  $N$ .

En la práctica si la relación  $n/N$  es menor o igual a 0,05 se considera a la población como infinita.

En ambas fórmulas  $z_{\alpha/2}$  depende del nivel particular de confianza y sus valores son:

En ambas fórmulas  $z_{\alpha/2}$  depende del nivel particular de confianza y sus valores son:

TABLA 3.2.4. Nivel de confianza.

Nivel de confianza	99,73%	99%	95,45%	95%	90%	68,27%	$z_{\alpha/2}$	3,00	2,58
	2	1,96	1,645	1,00					

b) Si la varianza es desconocida entonces sustituiremos en las fórmulas anteriores (1) y (2) la varianza muestral  $s^2$  por su estimador insesgado la varianza muestral corregida o cuasi-varianza

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

$n - 1$

Esta sustitución es válida cuando  $n > 30$ .

Si  $\bar{x}$  es la media de una muestra de tamaño  $n$  de una población normal  $N(\mu, \sigma^2)$  de varianza  $\sigma^2$  desconocida, siendo  $s^2$  la cuasi-varianza muestral y

el tamaño de la muestra es menor que 30 entonces:

$\bar{x}_{sn}$

$\pm$

0

^/,

—  $n-1/2$  es una  $t$  de Student con  $(n-1)$  g.l. y  $s^{\wedge 2}$  es la cuasivarianza muestral.

c) Si la población no es normal, siempre que el tamaño de la muestra sea grande ( $n > 30$ , aunque algunos autores aconsejan  $n > 100$ ), entonces los intervalos se construirán de igual manera a los apartados anteriores, según sea conocida la varianza o no. Cuando el tamaño de la muestra es pequeño y la población es no normal, no se pueden construir intervalos de confianza.

### Ejemplo 1:

Se desea estimar el promedio de horas de sueño de los jóvenes de 14 – 20 años, sabiendo que el tiempo se distribuye normalmente. Después de extraer una muestra representativa de 1000 jóvenes se obtiene los siguientes resultados:

$\bar{x} = 9$  horas/diarias;  $\hat{s}^2 = 2$  horas/diarias

$\hat{s}^2 = 2$  horas/diarias

Calcular un intervalo de confianza del 95% para el promedio de horas diarias de sueño.

$z_{\alpha/2} = z_{0,025} = 1,96$

$I = 9 \pm 1,96 \sqrt{\frac{2}{1000}} = 9 \pm 1,96 \sqrt{0,002} = 9 \pm 0,124$ , luego  $I = (8,876; 9,124)_{1000} \pm \pm \sqrt{\frac{2}{1000}}$

### Ejemplo 2:

Se recogen los datos de absentismo de una muestra aleatoria de 10 jóvenes de un instituto en las fechas próximas a los exámenes. Los resultados son: 4, 3, 5, 3, 3, 6, 7, 2, 1, 2 días. Suponiendo que la distribución de los días de ausencia es normal, determinar el intervalo de confianza al 99% del promedio de ausencias.

Con los datos de la muestra se calcula:  $\bar{x} = 3,6$  días;

$\hat{\sigma}^2$

$\frac{1}{n} \sum_{i=1}^n x_i^2$

$\frac{1}{n} \sum_{i=1}^n x_i$

$\frac{1}{n} \sum_{i=1}^n x_i^2$

$\frac{1}{n} \sum_{i=1}^n x_i$

$\frac{1}{n} \sum_{i=1}^n x_i^2$  luego  $\hat{\sigma}^2 = 2,10$

$$\frac{19}{\sqrt{932436n10}}$$

En las tablas de la t de Student se calcula  $t_{0,005;9}=3,25$  por tanto el intervalo de

19  
confianza de la media será: en consecuencia

$$\pm \pm$$

$$I=(1,65;5,55).^{10}$$

### Intervalos de confianza para las proporciones

Si p es la proporción muestral de los elementos principales (éxitos) de una binomial B(n,P) siendo  $n \geq 30$  y  $0,1 < P < 0,9$  entonces las fórmulas (1) y (2) quedan

$pq \pm z_{\alpha/2} \sqrt{\frac{pq}{n}}$  si el muestreo es de una población infinita o finita con reposición.

$\pm$   
 $pq \pm z_{\alpha/2} \sqrt{\frac{pq}{N-n}}$  si el muestreo es de una población finita sin reposición de

$\pm$   
tamaño N.

Como podemos observar resultan de sustituir en (1) y (2)  $xpq$

### Ejemplo 1

Para determinar el porcentaje de fumadores en un instituto se eligió una muestra de 280 alumnos encontrándose que 80 de ellos fumaban. Establecer el intervalo de confianza del 95% de dicha proporción.

En primer lugar se utiliza como estimador de la proporción poblacional, la proporción muestral:

$$\hat{p} = \frac{80}{280} = 0,286$$

$n \hat{p} = 280 \cdot 0,286 = 80,08 > 5$  se puede hacer la aproximación

280

a la normal,  $z_{0,025}=1,96$  por tanto:

$$I = 0,286 \pm 1,96 \cdot \sqrt{0,286 \cdot 0,714} = 0,286 \pm 0,053, \text{ por tanto } I = (0,233; 0,339) \pm \pm$$

### Intervalos de confianza para la varianza

a) Si estamos ante una población normal de varianza  $s^2$  y una muestra menor ( $n < 30$ )

que sabemos que  $s^2$  se distribuye en el muestreo como una ji

cuadrado con n-1 grados de libertad, entonces se puede definir el intervalo:

$$\hat{\sigma}^2 \left( \frac{1}{2} \chi_{2, \alpha/2}^2 \leq \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \leq \frac{1}{2} \chi_{2, 1-\alpha/2}^2 \right)$$

I

ÁÁ

$$= \frac{ns}{c} \left( \frac{1}{2} \chi_{2, \alpha/2}^2 \leq \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \leq \frac{1}{2} \chi_{2, 1-\alpha/2}^2 \right)$$

$$\hat{\sigma}^2 \left( \frac{1}{2} \chi_{2, \alpha/2}^2 \leq \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \leq \frac{1}{2} \chi_{2, 1-\alpha/2}^2 \right)$$

y por tanto despejando  $\sigma^2$  tenemos:

Ê

$$\hat{\sigma}^2 \left( \frac{1}{2} \chi_{2, \alpha/2}^2 \leq \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \leq \frac{1}{2} \chi_{2, 1-\alpha/2}^2 \right)$$

$$\hat{\sigma}^2 \left( \frac{1}{2} \chi_{2, \alpha/2}^2 \leq \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \leq \frac{1}{2} \chi_{2, 1-\alpha/2}^2 \right)$$

$$\hat{\sigma}^2 \left( \frac{1}{2} \chi_{2, \alpha/2}^2 \leq \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \leq \frac{1}{2} \chi_{2, 1-\alpha/2}^2 \right) = a$$

Ë

c

/( )

21

mn

a a

b) Cuando la muestra es mayor o igual a 100 los valores de ji-cuadrado en los límites de confianza se calculan de forma aproximada mediante las fórmulas:

0

2

C

2

a

(0);()

-

=<sub>2</sub>

0

2

C

2

(/);( )21 =2

$\hat{I} = (\hat{\sigma}^2)^{211}; (\hat{\sigma}^2)^{\wedge}$

ÁÁ- ~

$\hat{I} = C C; ( ) /; ( ) -$

### Ejemplo 1

En una muestra de 20 personas de una ciudad, se sabe que la altura tiene una media de 170 cm y una desviación típica de 10 cm. Se quiere saber a un nivel de confianza del 95% el intervalo de confianza de la varianza poblacional. Resolver el mismo supuesto cuando n=200.

$22 \cdot 2^{20}$

$= = = ; \hat{\sigma}^2, 105 \cdot 26$

19 100

$2 = 2 = 32 \cdot 85_{( / ); ( ) n-C C}$

$21 \cdot 0,025; 19$

$2 = 2 = 890_{(- / ); ( ) n-C C} 0,975; 19$

Luego el intervalo de confianza para la varianza será:

$\hat{I} = (19 \cdot 105,26, 19 \cdot 105 \cdot 26) = (60,88; 224,71) \text{OI} = \hat{\sigma}^2 \cdot 32,85 \cdot 890^{-}$

y para la desviación típica: (7,80;14,99).

Si el tamaño de la muestra es n=200 entonces se aproximará: 22

$$2 = 0 = 0 = 162 \ 27c12 \ 1 \ 2 \ 2$$

$$c_2 = 0 = 2 \cdot 200 - 1 + 0$$

$$s_2 = 240 \cdot 57 / (21 - 2)$$

$$\hat{E}I = (\hat{\mu}_1 - \hat{\mu}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 199.1005 \pm 1.99 \sqrt{\frac{1005^2}{21} + \frac{16227^2}{10}}$$

Y el intervalo de confianza de la desviación típica será: (9,12; 11,10).

### **Intervalo de confianza para las diferencia de medias y proporciones muestrales**

a) Supondremos tomadas dos muestras  $(n_1, x_1, s_1^2)$  y  $(n_2, x_2, s_2^2)$  de dos poblaciones normales independientes  $N(\mu_1, \sigma_1^2)$  y  $N(\mu_2, \sigma_2^2)$  respectivamente, con varianzas conocidas  $\sigma_1^2$  y  $\sigma_2^2$ , entonces el intervalo de confianza de la diferencia de medias será:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

12

b) Si las poblaciones son normales y  $n_1 \geq 30$  y  $n_2 \geq 30$  o siendo aproximadamente normales es  $n_1 \geq 100$  y  $n_2 \geq 100$ , la expresión también es válida aproximadamente.

c) Si las varianzas son desconocidas y  $n_1 \geq 30$  y  $n_2 \geq 30$  entonces el intervalo de confianza quedará de la siguiente forma:

$s_1^2$  y  $s_2^2$  son las cuasi-varianzas muestrales.

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, n_1+n_2-2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

d) Si las varianzas son desconocidas pero iguales ( $s_1^2 = s_2^2$ ) entonces

el intervalo será.

$$\frac{\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}{1}$$

xx t nn

± +

1111

11<sup>22</sup> donde t<sub>a/2</sub> es una t de <sub>nn-2</sub> 12

Student con (n<sub>1</sub>+n<sub>2</sub>-2) g.l. esta expresión se utiliza preferentemente en muestras pequeñas (n<sub>1</sub> £ 30, n<sub>2</sub> £ 30).

e) Si las varianzas desconocidas no son necesariamente iguales entonces:

0

xx t<sup>s2~s2</sup>

12 2

a

± +<sup>1 2</sup> donde t<sub>a/2</sub> es una t de Student con m (g.l.) donde

n<sub>1</sub> n<sub>2</sub>

Ê<sup>s2~s2</sup> ~<sup>2</sup>

ËÁ

1+ 2

n<sub>2</sub><sup>~</sup> - <sup>~2</sup> y <sup>2</sup> son las cuasi-varianzas. m = n<sub>1</sub>

ËÁ ^

2 ËÁ ~<sup>2 1 2</sup>

~

n

1

+ +<sub>2111</sub>

Si p<sub>1</sub> y p<sub>2</sub> son las proporciones muestrales del suceso principal (éxito) de dos poblaciones binomiales independientes B(n<sub>1</sub>,P<sub>1</sub>) y B(n<sub>2</sub>,P<sub>2</sub>) respectivamente, las fórmulas anteriores son válidas con la salvedad de sustituir xpq

Si en lugar de querer obtener la diferencia de medias queremos obtener la suma de medias tendríamos las mismas fórmulas anteriores solamente que sustituyendo

$$\bar{x}_1 - \bar{x}_2 \text{ por } \bar{x}_1 + \bar{x}_2$$

$\times 2$ .

En todas las fórmulas que hemos presentado en los diferentes cálculos de los intervalos de confianza, de forma general, cuando hablemos del *error probable de la estimación* vendrá dado por  $z_{\alpha/2} \sqrt{\text{Var}(q)}$  o  $t_{\alpha/2} \sqrt{\text{Var}(q)}$ , es decir, el segundo término de las fórmulas.

### Ejemplo 1

Se quiere comparar el promedio de problemas matemáticos resueltos semanalmente por dos clases similares de alumnos. Se toma dos muestras de 8 y 9 alumnos respectivamente y se observa el número de problemas resueltos:

Clase A: 8, 9, 9, 10, 10, 11, 11, 12

Clase B: 9, 8, 8, 10, 10, 11, 10, 10, 11

Suponiendo que la distribución de resueltos en ambas clases es normal y de

varianzas iguales, calcular el intervalo de confianza del 95% para la diferencia de promedios entre ambas clases.

Clase A:  $\bar{x}_s n = 8_{11} 1$

Clase B:  $\bar{x}_s n = 9_{22} 2$

En las tablas obtenemos:  $t_{0,025,(8+9-2)} = 2,13$

$I = 10 - 9,67$

$($

$2,13 \pm 1 (0,131)^{22}$

$\pm +$

$+ (0) - .,9 \pm 1 \pm 109$

9

892

$= 0,33 \pm 1,23 \pm$

8

luego  $I = (-0,90; 1,56)$ .

### Ejemplo 2:

El promedio de horas de vida de las cobayas, procedentes de dos camadas, utilizadas en una prueba del laboratorio de biología de un instituto es el siguiente:

Camada A: 800, 720, 740, 710, 750, 760

Camada B: 412, 420, 390, 410, 400

Construir el intervalo de confianza del 95% para la diferencia de vida media de

cada camada, en el supuesto de que la distribución de horas de vida es normal y

que la variabilidad es distinta para cada grupo en la población de cobayas.

Camada A:  $x_{s1} = 6$

Camada B:  $x_{s2} = 5$

Por ser la distribución de horas de vida de las cobayas normal y las varianzas distintas calculamos el número de grado de libertad mediante la fórmula:

$$\hat{E}_{s_1}^2 \sim \hat{E}_{s_2}^2 \quad \hat{E}_{32,042}^2 \sim \hat{E}_{22}^2$$

$$m = n_1 n_2 \frac{\hat{E}_{s_1}^2}{\hat{E}_{s_2}^2} + 1161$$

$$\hat{E}_{22}^2$$

$$\frac{\hat{E}_{s_1}^2}{\hat{E}_{s_2}^2} = 1 + 2 \sim -2 = 5$$

$$\hat{E}_{32,042}^2 \sim \hat{E}_{22}^2 = \hat{E}_{s_1}^2 \sim \hat{E}_{s_2}^2$$

$$\hat{E}_{s_1}^2$$

$$1 \sim \hat{E}_{s_1}^2 \sim \hat{E}_{s_2}^2 + 1161$$

$$5 \sim$$

$$n$$

$$n_1 + n_2 = 11 + 7 = 18$$

Luego la  $t_{0,025;7} = 2,36$

Por tanto el intervalo de confianza será:

$$32,042 \sim 22$$

$$I = (746,67 - 406,4)$$

$$11,61$$

$$\pm \pm = 340,27 \sim 33,214 \sim 6 \sim 5$$

luego  $I = (307,056; 373,484)$ .

**Intervalo de confianza para las diferencia de medias, datos relacionados** Si la distribución es normal y la diferencia  $d = x - y$  se pueden dar los siguientes casos:

La desviación típica ( $s_d$ ) es conocida:

$$I = dz \cdot s_{d0,2n}$$

Si la desviación típica ( $s_d$ ) es desconocida,  $n > 30$ .

$I =$

$$dz \hat{s}_d \frac{1}{\sqrt{n}}$$

Si la desviación típica ( $s_d$ ) es desconocida,  $n < 30$ .  $\hat{s}_d$

$$I = dt_{\pm} \frac{1}{\sqrt{n}}$$

### Ejemplo 1

Para probar la eficacia de un seminario de resolución de problemas matemáticos realizado en una Escuela de Magisterio, se les pasó una prueba antes y después de realizar el seminario a una muestra de 12 alumnos. Calcular el límite de confianza del 95% para el cambio producido en los errores de resolución, en el supuesto de que las diferencias se distribuyan normalmente.

Errores/alumno 1 2 3 4 5 6 7 8 9 10 11 12

Antes 12 13 15 10 11 14 15 7 8 7 12 12

Después 10 11 14 8 11 12 12 5 9 7 10 12

Para la resolución del ejercicio se suele construir la siguiente tabla:

**Antes Después  $d_i$  ( $d_i - \bar{d}$ )<sup>2</sup>**

12	10	2	0.5625
13	11	2	0.5625
15	14	1	0.0625
10	8	2	0.5625
11	11	0	1.5625
14	12	2	0.5625
15	12	3	3.0625
7	5	2	0.5625
8	9	-1	5.0625
7	7	0	1.5625
12	10	2	0.5625
12	12	0	1.5625

Total 15 16,25

$$t_{\alpha/2, (n-1)} = 2,20 \quad \hat{s}_d = \sqrt{\frac{1}{11} \sum d_i^2} =$$

11

$$I = 125 \pm 2,20 \cdot 2,20 = 125 \pm 4,84$$

$\pm = \pm =$

12

**Intervalo de confianza para la razón de varianzas de dos poblaciones normales**

Si tenemos dos muestras independientes de tamaños  $n_1$  y  $n_2$ , procedentes de poblaciones distribuidas normalmente con varianzas  $s_1^2$  y  $s_2^2$ . La variable aleatoria:

$$F = \frac{s_1^2/n_1}{s_2^2/n_2}$$

a) de la razón de varianzas será:

$$I = \left( \frac{s_1^2/n_1}{s_2^2/n_2} \right)_{\alpha/2; 2n_1-2, 2n_2-2}$$

### Ejemplo 1

Dos clase de 1.º de Secundaria fueron sometidas a una prueba de ortografía, la primera clase, formada por 31 alumnos obtuvo una cuasi-varianza de 124 errores y la segunda clase de 28 alumnos, consiguió 128 errores. Construir el intervalo de confianza de la razón de varianzas del 90%, en el supuesto de normalidad de población y muestras independientes.

$F_{0,05;30.27}=1,88$ ;  $F_{0,95;30.27}=0,53$  conviene recordar que

$$F_{\alpha/2; n_1-2, n_2-2} = \frac{1}{F_{1-\alpha/2; n_2-2, n_1-2}}$$

Por tanto  $I = (0,5153; 1,8278)_{188 \ 053}$

### EJERCICIOS DE AUTOCOMPROBACIÓN

1. En un muestreo aleatorio simple con tamaño poblacional de 100, un

error del 5% y un nivel de confianza del 95,5%. Se desea calcular el tamaño de la muestra en las condiciones más desfavorables del muestreo  $p=q=0,50$ . Construir el programa SPSS para resolver el supuesto y mostrar como salida los datos de entrada y el tamaño de la muestra para una población finita.

2. Se realiza un muestreo estratificado con el siguiente reparto por estrato:

$j \quad N_j \quad S_j$

1 30 9

2 45 4

3 75 6

Para una afijación óptima con  $e=2$ ,  $k=1,96$  cual será el tamaño de la muestra apropiado y el reparto de la misma por estrato. Calcular también el reparto de la muestra para afijación proporcional. Construir el programa SPSS para resolver el supuesto.

3. Una población consiste en los números 6, 4, 8 y 10. Se consideran todas las posibles muestras de tamaño 2 de esa población con reposición. Hallar: la media y la desviación típica de la población y la media y la desviación típica de la distribución en el muestreo de las medias.

4. Con los datos del problema anterior resolver si el muestreo es sin reemplazamiento.

5. Resolver con ayuda del ordenador el cálculo de las siguientes probabilidades realizadas anteriormente con el empleo de las tablas estadísticas de la distribución normal:

$P(Z \leq -0,24)$

$P(Z \geq -2,5)$

$P(0,30 < Z \leq 2,89)$

6. Ídem al problema anterior pero con la distribución t de Student.

$P(t(7) \leq 1,1192)$

$P(t(507) \geq 0,4)$

7. Resolver con ayuda del ordenador el cálculo del valor de F tal que  $P[F > F(5,30)] = 0,01$ .

8. Resolver con ayuda del ordenador el cálculo de las siguientes probabilidades realizadas anteriormente con el empleo de las tablas estadísticas de la distribución ji-cuadrado:

$P(3,94 \leq \chi^2(10) \leq 15,987)$

$P(1027 \leq c^2(1000) \leq 1061)$

## SOLUCIÓN A LOS EJERCICIOS DE AUTOCOMPROBACIÓN 1.

El programa SPSS pedido será:

\*\*\* MUESTREO ALEATORIO SIMPLE\*\*\*\*\*.

\*\*\*Tamaño de una muestra para un error y un nivel de confianza

\*\* El programa calcula el tamaño de la muestra en función del

\*\* nivel de confianza (nc) y del error (e), para una población.

\*\* Como salida se obtiene el tamaño corregido por poblaciones finitas (n\_mue) además de las entradas. Como regla general la proporción (p) de un atributo a estimar es desconocida.

\*\*\*Por eso se asume un valor por defecto de  $p = .50$ , que representa la situación más conservadora (donde el tamaño es mayor).

\*\*

NEW FILE.

INPUT PROGRAM.

LOOP #X=1 TO 1.

END CASE.

END LOOP.

END FILE.

END INPUT PROGRAM.

\* Introduzca un nivel de confianza (95,5% por defecto).

COMPUTE nc = 95.5 .

VARIABLE LABELS nc 'nivel de confianza' .

\* Introduzca un error (en porcentaje).

COMPUTE e = 5 .

VARIABLE LABELS e 'error muestral' .

\* Introduzca el tamaño de la población (si se desconoce, n\_pob = 0).

COMPUTE n\_pob = 100.

VARIABLE LABELS n\_pob 'tamaño población' .

\* Introduzca un p del atributo (por defecto,  $p = 0.50$ ).

COMPUTE p= 0.50.

VARIABLE LABELS p 'p atributo' .

EXECUTE .

RECODE n\_pob(0=SYSMIS).

COMPUTE e1=e/100.

COMPUTE k=ABS(IDF.NORMAL((1-nc/100)/2,0,1)).

COMPUTE n\_inf=k\*k\*p\*(1-p)/(e1\*e1).

```
COMPUTE n_mue=n_inf*n_pob/((n_pob-1)+n_inf).
COMPUTE n_inf=TRUNC(n_inf)+1.
COMPUTE n_mue=TRUNC(n_mue)+1.
EXECUTE.
FORMATS n_inf(F8.0) n_mue(F8.0) n_pob(F12.0) nc(F8.2) e(F8.2)
p(F8.2).
```

\*\*\*\*\*Salida.

```
LIST VARIABLES=n_mue e nc p n_pob.
```

La salida será la siguiente:

—

```
n_mue e nc p n_pob 81 5,00 95,50 ,50 100 Number of cases read: 1 Number of cases listed: 1
```

2. El programa SPSS para realizar el ejercicio es:

```
*** MUESTREO ESTRATIFICADO*****.
```

```
***Tamaño de una muestra para un error y un nivel de confianza
```

```
** El programa calcula el tamaño de la muestra en función del
```

```
** nivel de confianza (nc) y del error (e), para una población.
```

```
** Como salida se obtiene el tamaño (n_mue) y el reparto de la muestra  
por estrato según la afijación proporcional y la óptima.
```

```
NEW FILE.
```

```
INPUT PROGRAM.
```

```
LOOP #X=1 TO 1.
```

```
END CASE.
```

```
END LOOP.
```

```
END FILE.
```

```
END INPUT PROGRAM.
```

\* Introduzca un nivel de confianza (95,5% por defecto).

```
COMPUTE nc = 95 .
```

```
VARIABLE LABELS nc 'nivel de confianza' .
```

```
COMPUTE k=ABS(IDF.NORMAL((1-nc/100)/2,0,1)).
```

\* Introduzca un error (en porcentaje).

```
COMPUTE e = 2 .
```

```
VARIABLE LABELS e 'error muestral' .
```

```
compute e1=e/100.
```

```
execute.
```

```
matrix.
```

```
get datos
```

```
/file=*
```

```

compute k=datos(2).
compute e1=datos(4).

**** Introducir datos de tamaño y desviación típica de cada
estrato****. print /title '***Datos tamaño y desviación típica de cada
estrato***'.
compute x={30,9;45,4;75,6}.
print x/title ' '.
compute px=x(:,1)*x(:,2).
compute n=csum(x(:,1)).
compute px1=csum(px)/n.
compute px2=x(:,1)*x(:,2)*x(:,2).
compute px3=csum(px2)/n.
compute c=e1**2/k**2.
compute n_mue=rnd(px1**2/(c+px3/n)).
print/title '***La muestra total será***'.
print n_mue/title ' '.
*****.
*****Afijación óptima, reparto de la muestra*****.
print/title '***Reparto de la muestra por estrato afijación óptima*** '.
compute y=rnd(n_mue*px/csum(px)).
print y/title ' '.
***** Afijación proporcional*****.
print/title '***Reparto de la muestra por estrato afijación proporcional***
'. compute y1=rnd(n_mue*x(:,1)/n).
print y1/title ' '.
end matrix.

```

Cuya salida es:

```

Run MATRIX procedure:
***Datos tamaño y desviación típica de cada estrato***
    30 9
    45 4
    75 6
***La muestra total será***
138
***Reparto de la muestra por estrato afijación óptima***
    41
    28
    69
***Reparto de la muestra por estrato afijación proporcional***
    28
    41

```

———— END MATRIX ———

3. La media poblacional será:

$$m = \frac{648}{10} = 64.8$$

4

La desviación típica:

s

$$s = \sqrt{\frac{2222}{4}} = 23.5$$

$$s = \sqrt{\frac{2222}{4}} = 23.5$$

$$s = \sqrt{\frac{2222}{4}} = 23.5$$

Al ser el muestreo con reemplazamiento el número de muestras posibles será:  $4^2 = 16$

(6,6) (6,4) (6,8) (6,10) (4,4) (4,6) (4,8) (4,10) (8,8) (8,6) (8,4) (8,10)

(10,10) (10,6) (10,4) (10,8)

La media de cada muestra será:

6 5 7 8

4 5 6 7

8 7 6 9

10 8 7 9

La media de la distribución muestral de las medias será:

$\bar{X}$

$$\bar{X} = \frac{65784567876910879112}{16} = 64.8$$

16 16

V $\bar{X}$

$\sigma^2$

$$\sigma^2 = \frac{2222}{16} = 138.875$$

16

V $\bar{X}$

38 39 8 9

$$\sigma^2 = \frac{2222}{16} = 138.875$$

16

Como vemos en el muestreo con reemplazamiento (o en población infinita):

$$s^2 = 5$$

$$V\bar{X} = \frac{s^2}{n} = \frac{5}{16} = 0.3125$$

$n = 25$

4. Si el muestreo es sin reemplazamiento el número de muestras posibles será:

$$\hat{E}4^4 = 4! = 24$$

$$\hat{A}_2 = \frac{25!}{22!} = 25 \cdot 24 = 600$$

(6,4) (6,8) (6,10)

(4,8) (4,10) (8,10)

La media de cada muestra será:

57.8

67.9

La media de la distribución de medias será:

$$\bar{x} = \frac{57.8 + 67.9 + \dots + 67.9 + 57.8}{7} = 67.9$$

6

La desviación típica de la distribución de medias será:

$$s_x = \sqrt{\frac{22 \cdot 22 + 41 \cdot 14}{6}}$$

$$s_x = \sqrt{\frac{22 \cdot 22 + 41 \cdot 14}{6}} = \sqrt{\frac{167}{6}} = 5.25$$

$$s_x = \sqrt{\frac{167}{6}} = 5.25$$

6.6

$$s_x = \sqrt{\frac{167}{6}} = 5.25$$

Se puede comprobar como:

$$VX$$

( )

=

$$s^2 = \frac{N-1}{N} \hat{E}4^2 = \frac{24}{25} \cdot 52.5 = 50.4$$

ÉÁ

-

N

-

1

~ =

2

ÉÁ

-

$$41 \cdot \frac{1}{25} = 1.64$$

-

5. El programa de SPSS que realiza el cálculo se presenta a continuación. Debe destacarse que la introducción del comando BEGIN

## DATA

–END DATA se realiza para construir un fichero de datos con un solo registro donde introducir los valores de la variables creada con COMPUTE

```
data list free/x.  
begin data  
1  
end data.  
execute.  
format x (f8.4).
```

```
***** probabilidad de z menor o igual que a *****.  
***** por ejemplo prob. z menor o igual que -0.24 *****.  
COMPUTE pz1 = CDFNORM(-0.24) .  
format pz1 (f8.4).
```

```
***** probabilidad de z mayor o igual que a *****. ***** por  
ejemplo prob. z mayor o igual que -2.5 *****. COMPUTE pz2 =  
1-CDFNORM(-2.5) .  
format pz2 (f8.4).
```

```
***** probabilidad de z entre dos valores a, b *****.  
***** por ejemplo prob. z entre 0.30 y 2.89 *****.
```

```
COMPUTE pz3 = CDFNORM(2.89)-CDFNORM(0.30) . format pz3  
(f8.4).  
execute.
```

6. De forma similar al ejercicio anterior, el programa que realiza el cálculo será:

```
data list free/x.  
begin data  
1  
end data.  
execute.  
format x (f8.4).
```

```
***** probabilidad de t menor o igual que a *****.  
***** por ejemplo prob. t(7) menor o igual que 1.1192 *****.  
COMPUTE pt1 = CDF.T(1.1192,7) .  
format pt1 (f8.4).
```

```
***** probabilidad de t mayor o igual que a *****.  
***** por ejemplo prob. t(507) mayor o igual que 0.4 *****.
```

```
COMPUTE pt2 = 1-CDF.T(0.4,507) .  
format pt2 (f8.4).
```

```
execute.
```

7. El programa que realiza el cálculo propuesto es:

```
data list free/x.  
begin data  
1  
end data.  
execute.
```

```
format x (f8.4).
```

```
***** Valor de F tal que  $P[F > F(5,30)] = 0.01$  ***** . COMPUTE if1  
= IDF.F(0.99,5,30) .  
format if1 (f8.4).
```

```
execute.
```

8. El programa que realiza el cálculo es:

```
data list free/x.  
begin data  
1  
end data.  
execute.
```

```
format x (f8.4).
```

```
***** probabilidad de chi-cuadrado entre dos valores a, b  
***** . ***** por ejemplo prob. chi(10) entre 3.94 y 15.987  
***** .
```

```
COMPUTE pchi1 = CDF.CHISQ(15.987,10)-CDF.CHISQ(3.94,10) .  
format pchi1 (f8.4).  
execute.
```

```
***** probabilidad de chi-cuadrado entre dos valores a, b  
***** . ***** por ejemplo prob. chi(1000) entre 1027 y 1061  
***** .
```

```
COMPUTE pchi2 = CDF.CHISQ(1061,1000)-  
CDF.CHISQ(1027,1000) . format pchi2 (f8.4).  
execute.
```

## **BIBLIOGRAFÍA**

Arce, Constantino; Real, Eulogio (2002): *Introducción al análisis*

*estadístico con SPSS*. PPU. Barcelona.

Calot, G. (1974): *Curso de Estadística Descriptiva*. Paraninfo. Madrid.

Camacho Rosales, Juan (2002): *Estadística con SPSS versión 11 para Windows*. Ra-Ma, Librería y Editorial Microinformática. Madrid.

Cuadras, C. M.; Echeverría, B.; Mateo, J.; Sánchez, P. (1991): *Fundamentos de Estadística*. PPU. Barcelona.

Ferrán Aranaz, Magdalena (2002): *Curso de SPSS para Windows*. McGraw-Hill/ Interamericana de España. Madrid.

Labrousse, C. (1976): *Estadística. Ejercicios resueltos (Tomo II)*. Madrid:

Paraninfo. Lizasoain Hernández, Luis; Joaristi Olarriaga, Luis (2003): *Gestión y análisis de datos con SPSS*. Thomson Paraninfo. Madrid.

Martín F. G. (1994): *Introducción a la Estadística Económica y Empresarial*. AC. Madrid. Martín, M. F.; Fernández, R.; Seisdedos, A.

(1985): *Estadística inferencial. Manual de prácticas para las ciencias de la conducta*. Salamanca: Universidad pontificia. Pardo Merino, Antonio;

Ruiz Díaz, Miguel Ángel (2002): *SPSS 11. Guía para el análisis de datos*. McGraw-Hill/ Interamericana de España. Madrid.

Peña D. (1992): *Estadística, Modelos y Métodos. Volumen I*. Alianza Universidad Textos. Madrid.

Peña D. (1992): *Estadística, Modelos y Métodos. Volumen II*. Alianza Universidad Textos. Madrid.

Pérez, César (2001): *Técnicas estadísticas con SPSS*. Pearson Educación. Madrid. Ríos, S. (1974): *Métodos estadísticos*. Ed. del Castillo. Madrid.

Spiegel, M. R. (1992): *Estadística*. McGraw-Hill. Madrid.

Visauta Vinacua, B. (2002): *Análisis estadístico con SPSS 11.0 para Windows*. McGraw-Hill/ Interamericana de España. Madrid.

## **UNIDAD DIDÁCTICA 4**

### **ANÁLISIS INFERENCIAL DE DATOS**

#### **Objetivos**

Conocer los fundamentos de los tests estadísticos y su incidencia en las decisiones sobre los parámetros poblacionales.

Diferenciar las pruebas de una sola cola frente a las de dos. Distinguir los condicionantes paramétricos de las pruebas estadísticas. Valorar la utilización de pruebas paramétricas.

Utilizar con soltura las principales pruebas paramétricas.

Diferenciar el contexto de utilización de las pruebas no paramétricas.

Distinguir, según las condiciones de los problemas, la prueba más adecuada para resolverlo.

Aplicar el contraste no paramétrico más pertinente.

Utilizar los comandos SPSS para resolver pruebas paramétricas y pruebas no paramétricas.

## **1. DECISIÓN ESTADÍSTICA. PRUEBAS PARAMÉTRICAS**

### **1.1. Introducción**

Hay ocasiones en que tenemos que tomar decisiones relativas a una población sobre la base de los conocimientos que tenemos de la muestra. Por ejemplo, si los resultados de un determinado medicamento afectan positivamente la salud de los enfermos, si un método pedagógico es más eficaz que otro, si un determinado dentífrico es preferido por la mayoría de la población, etc.

A la hora de tomar una decisión, tenemos que enunciar una hipótesis (según el diccionario de la Real Academia «suposición de una cosa, sea posible o imposible, para sacar de ella una consecuencia»), es decir, tenemos que hacer unas suposiciones y ver si estas se cumplen o se rechazan. Para realizar este proceso necesitamos el empleo de unas pruebas estadísticas denominadas también test o dóxicas para contrastar la veracidad o falsedad de las hipótesis enunciadas.

Los test o pruebas estadísticas se clasifican en dos grandes grupos: pruebas paramétricas, las más estudiadas y de uso más frecuente, y las pruebas no paramétricas, de uso más reciente y aún no muy desarrolladas.

En las primeras se considera que la variable que estamos tratando tiene una función de distribución determinada, bien totalmente o con algunos parámetros desconocidos. En las segundas no se hace ninguna consideración sobre la función de distribución.

### **1.2. Hipótesis estadística**

En el apartado anterior se ha discutido la estimación de parámetros, con el conocimiento de estadísticos de las muestras. En este apartado se hablará del contraste de hipótesis centrándose en las pruebas paramétricas. Tanto la estimación por intervalos como el contraste de hipótesis tratan el mismo problema: el conocimiento de la población a través del estudio de la muestra y proporcionan información similar (Ríos 1977, página 359):

#### **Intervalo de confianza Test de hipótesis**

Deja cubrir el parámetro  
Cubre valores erróneos

Extensión de muestra necesaria para limitar la longitud del intervalo de confianza

Error de tipo I  
Error de tipo II

Extensión de muestra necesaria para que la potencia sea mayor que un cierto valor

Una hipótesis estadística es una proposición sobre la distribución de una o más variables aleatorias, en referencia a la forma que adopta dicha distribución o los parámetros que definen la población. Así, por ejemplo, se puede decir que la variable aleatoria se distribuye como una ji-cuadrado o que los parámetros poblacionales son: media igual 20 y desviación típica igual a 3.

Lo habitual en el planteamiento de hipótesis estadísticas es dar por supuesto ciertas condiciones de la distribución de la variable aleatoria y centrarse en el conocimiento de los parámetros de la población.

La formulación de una hipótesis estadística se puede realizar mediante una **hipótesis simple** ( $m=5$ ) o mediante una **hipótesis compuesta**, formada con varias simples ( $m\pi 5$ ,  $m>5$  o  $m<5$ ).

### **1.2.1. Hipótesis nula**

En la nomenclatura la **hipótesis  $H_0$**  se denomina **hipótesis nula** y será aquella formulación teórica que trata de «proteger» a priori a la distribución que se pretende contrastar, teniendo dos planteamientos:

$H_0$ : «el estadístico es compatible o pertenece a una distribución de parámetro X». Por ejemplo:  $H_0$ :  $m=100$ .

$H_0$ : «no hay diferencia significativa entre los estadísticos». Por ejemplo  $H_0$ :  $m_1 - m_2 = 0$ .

El término nula indica que la diferencia puede ser explicada por variaciones aleatorias de la muestra y no que la diferencia tenga que ser igual a cero.

### **1.2.2. Hipótesis alternativa**

Frente a la hipótesis nula, se plantean las **hipótesis alternativas** expresadas mediante  $H_1$ . En los dos casos anteriores las hipótesis

alternativas serán: Frente a  $H_0: m=100$  se plantean las siguientes hipótesis alternativas:  $H_1: m > 100$ ;  $H_1: m < 100$ ;  $H_1: m \neq 100$ ;  $H_1: m = 104$  ....

Frente a  $H_0: m_1 - m_2 = 0$ ;  $H_1: m_1 - m_2 \neq 0$ ;  $H_1: m_1 - m_2 < 0$ ;  $H_1: m_1 - m_2 > 0$  ....

En realidad la hipótesis alternativa es la que queremos probar y su formulación es imprescindible aún cuando  $H_0$  no se rechace.

### 1.3. Formulación de hipótesis. Hipótesis simple frente alternativa simple

Sea  $X$  una variable aleatoria cuya distribución de probabilidad  $F(X)$  depende de un parámetro  $q$ . La densidad de probabilidad es:  $f(x_i, q)$ . El parámetro  $q$  desconocido puede tomar los dos valores  $q_0$  y  $q_1$ . El contraste consiste en elegir entre dos hipótesis simples:

$H_0$  (**hipótesis nula**):  $q = q_0$   $H_1$  (**hipótesis alternativa**):  $q = q_1$

La decisión estadística denomina test o contraste de hipótesis, a una regla que permite decidir cuál de las dos hipótesis es aceptada, partiendo de una muestra aleatoria  $x = (x_1, \dots, x_n)$  de  $x$ .

#### 1.3.1. Regiones críticas y de aceptación

La aceptación o rechazo de la hipótesis nula, supone la división previa de la distribución muestral en dos regiones: crítica ( $W$ ) o de rechazo de la hipótesis nula y no crítica o de aceptación de dicha hipótesis. Se llama **región crítica  $W$  de nivel de significación  $\alpha$**  a un subconjunto del espacio muestral  $R^n$ , tal que la probabilidad de que la muestra  $x$  pertenezca a  $W$ , cuando es cierta  $H_0$ , es igual a  $\alpha$ , es decir,

$$P(x \in W | H_0 \text{ es cierta}) = \alpha$$

La probabilidad de que  $x$  no pertenezca a  $W$  cuando es cierta  $H_0$  es  $P(x \notin W | H_0) = 1 - \alpha$

El **nivel de significación  $\alpha$**  es una probabilidad pequeña (0,1; 0,05 ó 0,01) que normalmente elige el experimentador. Si suponemos que una hipótesis particular es cierta pero vemos que los resultados hallados en una muestra aleatoria difieren substancialmente de los esperados desde tal hipótesis (es decir, no debidos al azar), entonces se dirá que las diferencias observadas son significativas a nivel de significación  $\alpha$ .

La regla de decisión para resolver entre  $H_0$  y  $H_1$  es la siguiente: — Si se presenta el suceso ( $x \in W$ ), se rechaza  $H_0$  y se acepta  $H_1$ , es decir que si la muestra pertenece a la región crítica se rechaza  $H_0$  y se acepta la

alternativa.

— Si se presenta el suceso ( $x \in W$ ), se acepta  $H_0$  y se rechaza  $H_1$ . En otras palabras, si la muestra no pertenece a la región crítica, es decir pertenece a la región de aceptación, se acepta  $H_0$  y se rechaza  $H_1$ .

Se justifica este criterio porque si  $H_0$  es cierta, el suceso ( $x \in W$ ) tiene solamente una probabilidad  $a$  de presentarse. Como es razonable suponer que no se presentará bajo  $H_0$ , cuando realmente se presenta ( $x \in W$ ) (la muestra pertenece a la región crítica  $W$ ), existen razones para inferir que la hipótesis verdadera es  $H_1$ .

### 1.3.2. Error tipo I, error tipo II y potencia de una prueba

El rechazar la hipótesis nula  $H_0$  siendo cierta, se llama **error de tipo I**. La probabilidad de cometer este error es el nivel de significación  $\alpha$ .

La probabilidad de equivocarse cuando se acepta  $H_0$ , siendo cierta  $H_1$ , es  $P(x \in W | H_1 \text{ cierta}) = b$

y es la probabilidad de cometer el llamado **error tipo II**. Si la hipótesis cierta es  $H_1$ , la probabilidad de que la muestra pertenezca a  $W$  (es decir, aceptar  $H_1$  siendo cierto  $H_1$ ) es:

$$P(x \in W | H_1) = 1 - b$$

que recibe el nombre de **potencia del test**. Decimos que una prueba es más potente que otra cuando, con un  $\alpha$  es mayor en una prueba que en otra. Asimismo dentro de una prueba es más potente la decisión unilateral que la bilateral.

TABLA 4.1.1. Hipótesis y decisión estadística.

#### Decisión

$H_0$   $H_1$

$H_0$

Hipótesis  $H_1$  verdaderas

sin error  $\alpha$  (error  $-I$ )

$b$  (error  $-II$ ) sin error

Potencia  $(1 - b)$

#### Ejemplo

Se supone que la altura de una población de jóvenes es 172 cm con una desviación típica de 10 cm. Se toma una muestra de 144 alumnos de un instituto dando como media 175 cm. Además se sabe que la altura media

de los alumnos del instituto es 177 cm. ¿cuál es la probabilidad de cometer error tipo I y tipo II si se establecen las siguientes hipótesis  $H_0=172$ , frente  $H_1=177$ , sabiendo que la muestra escogida tiene como media 175 cm?

a) Se calcula el error típico de la distribución muestral de la media:

$$s_x = \frac{s}{\sqrt{n}} = \frac{10}{\sqrt{50}} = 1.414$$

b) Se toma como punto crítico ( $X_c$ ) el valor de la media muestral, en nuestro caso  $X_c=175$ . Si se supone la distribución normal, se puede hacer el siguiente planteamiento gráfico:

a= error tipo I

$$m_0 = 172 \quad m_1 = 175$$

b= error tipo II

$$m_0 = 175 \quad m_1 = 177$$

Se calcula los valores de z y su probabilidad asociada:

$$z_{H0} = \frac{175 - 172}{1.414} = 2.121 \quad P(Z > 2.121) = 0.0168$$

0.0168,,

$$z_{H1} = \frac{175 - 177}{1.414} = -1.414 \quad P(Z < -1.414) = 0.0786$$

Si la muestra fuera menor, por ejemplo  $n=36$  entonces los valores de los errores tipo I y II serian mayores:

$$z_{H0} = \frac{175 - 172}{1.155} = 2.601 \quad P(Z > 2.601) = 0.0047$$

0.0047,,

$$z_{H1} = \frac{175 - 177}{1.155} = -1.732 \quad P(Z < -1.732) = 0.0418$$

#### 1.4. Hipótesis simple frente alternativa compuesta

En muchos contrastes normalmente existe más de una hipótesis alternativa frente a una hipótesis nula  $H_0$ . A menudo el problema se puede formular de modo que la densidad de la v.a. X es  $f(x,q)$  dependiendo de un parámetro q. El contraste de hipótesis se formula entonces así:

$H_0: q=q_0$  (un valor conocido)

$H_1: q \neq q_0$

y se dice que  $H_1$  es una alternativa bilateral.

Los test de hipótesis simple frente alternativa simple generan regiones críticas de una sola extremidad y son llamadas **pruebas de una sola cola o unilaterales**

Región de rechazo de rechazo

1- a 1- a

Zona de aceptación a a Zona de aceptación

FIGURA 4.1.1. Pruebas unilaterales derecha e izquierda. Regiones de aceptación y rechazo.

y los test de hipótesis simple frente alternativa compuesta tienen regiones críticas de dos extremidades y por eso son denominadas **pruebas de dos colas o bilaterales**.

Región Región de rechazo  $1-a$  de rechazo

$a/2$  Zona de aceptación  $a/2$

FIGURA 4.1.2. Prueba bilateral. Regiones de aceptación y rechazo.

En todas las pruebas fijaremos un nivel de significación  $\alpha$  y hay que tener presente que si el estadístico calculado cae en la región de rechazo entonces se refutará la hipótesis nula.

### 1.5. Potencia de una prueba de hipótesis

Según hemos visto  $b$  representa la probabilidad de aceptar  $H_0$  cuando es falsa, cuando es falsa,  $b$  es la probabilidad de rechazar dicha hipótesis cuando es falsa. Esta probabilidad es conocida como potencia de una prueba.

El objetivo es conseguir pruebas potentes con valores bajos del error tipo I ( $\alpha$ ) y del error tipo II ( $\beta$ ) sin embargo éstos y la potencia tienen unos condicionantes:

1. El valor de  $\alpha$  se fija al escoger la región de rechazo.
2. El valor de  $\beta$  dependerá de la hipótesis alternativa que se escoja.
3. Para un tamaño muestral fijo, al aumentar la región de rechazo (y por lo tanto  $\alpha$ ),  $\beta$  disminuye. Si  $\alpha$  decrece,  $\beta$  aumentará.
4. Al aumentar el tamaño muestral  $n$ ,  $\alpha$  y  $\beta$  decrecerán a la vez.

Como vemos para conseguir el objetivo de máxima potencia y menor error la solución está en aumentar el tamaño de la muestra. Así mismo, dentro de una prueba la decisión unilateral es más potente que la bilateral. Veamos un ejemplo:

Supongamos una muestra aleatoria  $n=36$ . Se plantea  $H_0: \mu=10$  frente la alternativa  $H_1: \mu=12$ , se elige  $\alpha=0,05$  y se supone una población normal con  $\sigma^2=16$ .

16

En primer lugar calculemos el error típico:  $s_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{36}} = \frac{4}{6} = 0,66$

a) Si se elige una región crítica unilateral derecha entonces  $z_a=1,645$  por tanto:

$$1,645 = \frac{X_c - 10}{\sqrt{0,66}} \Rightarrow X_c = 11,086$$

Con este punto crítico ( $X_c$ ), si suponemos  $H_1$  como verdadera entonces  $X < X_c = 11,086$ , con lo que si aceptamos  $H_0$  cometemos un error tipo II. Para el valor de  $H_1: m=12$  tendremos que averiguar qué  $z$  corresponde a  $X_c$  y calcular la probabilidad de valores menores que esa  $z$ .

$$z = \frac{11,086 - 12}{\sqrt{0,66}} = -1,385 \Rightarrow P(z < -1,385) = 0,08379 \text{ y por tanto } b = 0,08379,$$

de ahí que la potencia  $= 1 - b = 1 - 0,08379 = 0,91621$ .

b) Si se elige una región crítica unilateral izquierda, siguiendo un procedimiento similar al anterior.

$$-1,645 = \frac{X_c - 10}{\sqrt{0,66}} \Rightarrow X_c = 8,914$$

066

Suponiendo  $H_1$  cierta

$$z = \frac{8,914 - 12}{\sqrt{0,66}} = -4,675 \Rightarrow P(z > -4,675) = 0,9997 \text{ y por tanto } b = 0,9997, \text{ de ahí}$$

que la potencia  $= 1 - b = 1 - 0,9997 = 0,0003$

c) Finalmente al elegir una región crítica bilateral habrá dos puntos críticos:

$$-1,96 = \frac{X_{c1} - 10}{\sqrt{0,66}} \Rightarrow X_{c1} = 8,706$$

$$1,96 = \frac{X_{c2} - 10}{\sqrt{0,66}} \Rightarrow X_{c2} = 11,294$$

Suponiendo  $H_1$  cierta  $z = \frac{8,706 - 12}{\sqrt{0,66}} = -4,991$

$$z = \frac{11,294 - 12}{\sqrt{0,66}} = -1,070$$

La  $P(-4,991 < z < -1,070) = 0,14228$ , por tanto  $b = 0,14228$ , de ahí que la potencia  $= 1 - b = 1 - 0,14228 = 0,85772$ .

En consecuencia es evidente que la región crítica que da mayor potencia es la unilateral derecha, por supuesto porque 12 es mayor que 10, o lo que es lo mismo  $H_1 > H_0$ .

### 1.5.1. Función de potencia y curvas características operativas

El caso anterior se ha dado una hipótesis nula ( $H_0: m=10$ ) y una alternativa simple ( $H_1: m=12$ ) pero también se podría dar una alternativa compuesta  $H_1: m > 10; m < 10$  o  $m \neq 10$ . En cualquiera de los casos de la alternativa compuesta, no hay un riesgo  $b$  único sino un conjunto de

valores en función de los distintos valores que vaya tomando  $m$  en la hipótesis  $H_1$ ; entonces se habla de una **función de potencia** en lugar de hablar de una potencia única.

Veamos un ejemplo. Supongamos  $H_0:m=200$  frente  $H_1:m\neq 200$  con  $\alpha=0,05$ ,  $n=625$ , y que la población se distribuye normalmente con  $\sigma=100$ . Con estos datos se podrían dar infinitos valores de  $m$ , sin embargo para concretar se ha dado once datos, cinco por encima y cinco por debajo del valor 200, además del  $m=200$ . Los valores de los datos ( $X$ ),  $Z_1, Z_2$ ,  $\beta$  aparecen a continuación:

$X$	$Z_1$	$Z_2$	$\beta$	$1-\beta$
224.00	-7.96	-4.04	.0000	1.0000
220.00	-6.96	-3.04	.0012	.9988
212.00	-4.96	-1.04	.1492	.8508
208.00	-3.96	-.04	.4840	.5160
204.00	-2.96	.96	.8299	.1701
200.00	-1.96	1.96	.9500	.0500
196.00	-.96	2.96	.8299	.1701

$X$	$Z_1$	$Z_2$	$\beta$	$1-\beta$
192.00	.04	3.96	.4840	.5160
186.00	1.54	5.46	.0618	.9382
182.00	2.54	6.46	.0055	.9945
176.00	4.04	7.96	.0000	1.0000

Como se observa en la tabla cuanto más lejos está  $H_1$  de  $H_0$  mayor es la potencia. Esta cuestión debe destacarse como uno de los factores que influye en el aumento de la potencia de una prueba.

La gráfica de la función de potencia aparece a continuación:

FIGURA 4.1.3. Curva de potencia de una prueba bilateral.

La forma en V de la curva de potencia es propia de las pruebas bilaterales, conforme más estrecha sea mejor discrimina la prueba entre  $H_0$  y los distintos valores de  $H_1$ .

Las pruebas unilateral izquierda y unilateral derecha tienen unas curvas de potencia mostradas en las fig. 4.1.4 y fig. 4.1.5.

FIGURA 4.1.4. Curva de potencia de una prueba unilateral izquierda.

FIGURA 4.1.5. Curva de potencia de una prueba unilateral derecha.

En la práctica estas curvas se utilizan poco, en su lugar se emplea la

curva característica operativa donde se representa X en función de b. Las gráficas serán entonces (ver fig. 4.1.6 a 4.1.8).

FIGURA 4.1.6. Curva característica operativa de una prueba bilateral.

FIGURA 4.1.7. Curva característica operativa de una prueba unilateral izquierda.

FIGURA 4.1.8. Curva característica operativa de una prueba unilateral derecha.

### 1.5.2. Cálculo de n para un a y b dados

Existe relación entre los valores de a y b y el tamaño de la muestra (n). Aunque los valores de a y b son inversos entre sí, lo que indudablemente se necesita es que ambos tengan sus valores lo más bajos posibles. Para concretar su cálculo supongamos el siguiente ejemplo: sea una población normal con  $s^2 = 625$  y se plantean las hipótesis  $H_0 = 40$ ;  $H_1 = 36$  con los datos  $a = 0,05$  y  $b = 0,10$ .

Al ser  $H_1 < H_0$  se plantea una región crítica unilateral izquierda, que para  $a = 0,05$  nos lleva a  $z = -1,645$ , por tanto el valor crítico será:

$$X_c = 1,645 \sqrt{25} + 40$$

n

y por los datos de  $b = 0,10$  nos lleva a  $z = 1,28$  y por tanto el valor crítico será:

$$1,28 \sqrt{25}$$

$$X_c = +36 + n$$

La representación gráfica de las dos zonas será:

$$a = 0,05$$

$$40 \quad z = -1,645$$

$$b = 0,10$$

$$36 \quad z = 1,28$$

Igualando las dos expresiones tenemos:

$$1,645 \sqrt{25}$$

$$1,28 \sqrt{25}$$

„- + = + 36 después de despejar n se tiene:  $n = 334$ . nn

El procedimiento de cálculo que se ha realizado para una prueba de alternativa simple, se puede ejecutar, sin mayor problema, para una prueba de alternativa compuesta.

## 1.6. Principales pruebas paramétricas

En la selección de la prueba estadística hay que tener presente no sólo

la naturaleza de la población, sino la aleatoriedad de las muestras, la independencia de los datos y la relación o no de las muestras. Estas cuestiones son algunas de las condiciones que exigen las pruebas denominadas paramétricas, en concreto se exige:

a) Independencia de los datos:

— Cualquier sujeto tiene las mismas posibilidades de ser elegido en la muestra (aleatoriedad).

— La puntuación de un sujeto no influye en la asignada a otro. b) Normalidad:

— Las poblaciones de las que se extraen los sujetos de las muestras deben estar distribuidas normalmente para el parámetro a estimar. — Esta condición es habitual asumir su cumplimiento cuando la muestra es grande.

c) Homocedasticidad:

— Cuando hay varios grupos, se supone que proceden de la misma población o poblaciones con igual varianza.

— El incumplimiento de ésta condición afecta a los contrastes de varios grupos. d) Medida de intervalo:

— Las variables deben medirse en una escala de intervalo o casi-intervalo, es decir, en una escala ordinal multicategórica.

e) Linealidad (sólo en la prueba F):

— La relación atribuida a los efectos de las interacciones entre fila y columna o a ambos, debe ser aditiva y no multiplicativa para evitar su influencia sobre las varianzas.

El principio de Neyman-Pearson y la construcción de la región crítica por medio de la razón de las funciones de verosimilitud asociadas a las hipótesis  $H_0$  y  $H_1$ , que no veremos en el texto, son elementos importantes a la hora de construir test paramétricos.

Para contrastar una hipótesis estadística se pueden seguir los siguientes pasos:

1. **Formular** la hipótesis nula  $H_0$  y la hipótesis alternativa o de investigación.

2. **Fijar** el nivel de significación ( $\alpha$ ).

3. **Comprobar** las características de las variables y plantear las suposiciones necesarias. Cumplimiento o no de las condiciones paramétricas.

4. **Elegir** un estadístico para contrastar la hipótesis.
5. **Estudiar** las características de la distribución muestral del estadístico.
6. **Determinar** la región crítica o de rechazo de  $H_0$  y la de aceptación. Viene determinada por el nivel de significación ( $\alpha$ ) dado y por la dirección de  $H_1$ .
7. **Decidir** sobre la aceptación o rechazo de  $H_0$ . Si el valor calculado en la muestra cae dentro de la zona de aceptación, se acepta la hipótesis nula y si no se rechazará.

Veamos, para finalizar este punto, las principales pruebas paramétricas según la siguiente secuencia:

1. Formulación de la hipótesis nula.
2. Enunciado de la hipótesis alternativa, presentando la exposición para una y/o dos colas.
3. Estadístico utilizado en la prueba.
4. Región de rechazo.

### **1.6.1. Muestras grandes: media y proporción muestrales**

#### **Media:**

1. Hipótesis nula:  $H_0: \mu = \mu_0$
2. Hipótesis alternativa:  
*Prueba de una extremidad (o cola)*  $H_a: \mu > \mu_0$  o bien  $H_a: \mu < \mu_0$   
*Prueba de dos extremidades (o colas)*  $H_a: \mu \neq \mu_0$

3. Estadístico de la prueba:

$$z =$$

$$\frac{\bar{x} - \mu_0}{\frac{s_x}{\sqrt{n}}}$$

Si se desconoce  $s$  (lo que normalmente sucede), sustitúyase por la desviación estándar muestral.

4. Región de rechazo:

*Prueba de una extremidad.*

$z > z_\alpha$  (o bien  $z < -z_\alpha$ , cuando la hipótesis alternativa es  $H_a: \mu < \mu_0$ ). *Prueba de dos extremidades.*

$z > z_{\alpha/2}$  o bien  $z < -z_{\alpha/2}$

$$Z_{\alpha/2} - Z_{\alpha/2} Z_{\alpha/2}$$

Suposiciones: Muestra aleatoria con  $n > 30$

### Proporción:

1. Hipótesis nula:  $H_0: p = p_0$
2. Hipótesis alternativa:

*Prueba de una extremidad (o cola)*  $H_a: p > p_0$  o bien  $H_a: p < p_0$

*Prueba de dos extremidades (o colas)*  $H_a: p \neq p_0$

3.

*Estadístico de la prueba:*

$$\frac{\hat{p} - p_0}{\sqrt{\hat{p}\hat{q}}}$$

$$\text{donde } \hat{p} = \frac{x}{n}$$

$n$

4. Región de rechazo:

*Prueba de una extremidad*

$z > z_{\alpha}$  (o bien  $z < -z_{\alpha}$ , cuando la hipótesis alternativa es  $H_a: p < p_0$ ) *Prueba de dos extremidades*

$z > z_{\alpha/2}$  o bien  $z < -z_{\alpha/2}$

$$Z_{\alpha/2} - Z_{\alpha/2} Z_{\alpha/2}$$

Suposiciones: El muestreo satisface los supuestos de un experimento binomial y  $n$  suficientemente grande para que la distribución muestral de  $x$  (y por consiguiente de

$\hat{p}$ ) tenga una distribución aproximadamente normal.

### Diferencia de medias:

1. *Hipótesis nula:*  $H_0: (m_1 - m_2) = D_0$  donde  $D_0$  es alguna diferencia especificada que se quiera probar. En muchos casos se deseará probar la hipótesis de que no hay una diferencia entre  $m_1$  y  $m_2$ , es decir,  $D_0 = 0$ .

2. *Hipótesis alternativa:*  
*Prueba de una extremidad (o cola)*

$H_a: (\mu_1 - \mu_2) > D_0$  o bien  $H_a: (\mu_1 - \mu_2) < D_0$

Prueba de dos extremidades (o colas)  $H_a: \mu_1 - \mu_2 \neq D_0$

( )

3. Estadístico de la prueba: 
$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Si se desconoce  $s_1^2$  y  $s_2^2$  (lo que normalmente sucede), sustitúyase por las cuasi-varianzas muestrales

$\hat{s}_1^2$  y  $\hat{s}_2^2$

4. Región de rechazo:

Prueba de una extremidad

$z > z_{\alpha}$  (o bien  $z < -z_{\alpha}$ , cuando la hipótesis alternativa es  $H_a: \mu_1 - \mu_2 < D_0$ )

Prueba de dos extremidades

$z > z_{\alpha/2}$  o bien  $z < -z_{\alpha/2}$

Suposiciones: Se seleccionaron las muestras aleatorias e independientes de dos poblaciones y  $n_1 > 30$  y  $n_2 > 30$ .

**Diferencia de proporciones:**

1. Hipótesis nula:  $H_0: (p_1 - p_2) = D_0$  donde  $D_0$  es alguna diferencia especificada que se quiera probar. En muchos casos se deseará probar la hipótesis de que no hay una diferencia entre  $p_1$  y  $p_2$  - es decir,  $D_0 = 0$ .

2. Hipótesis alternativa:

Prueba de una extremidad (o cola)

$H_a: (p_1 - p_2) > D_0$  o bien  $H_a: (p_1 - p_2) < D_0$

Prueba de dos extremidades (o colas)

$H_a: p_1 - p_2 \neq D_0$

3. Estadístico de la prueba:

( )

$$= \frac{z \sqrt{p_1 p_2 D}}{\sqrt{\frac{x_1^2}{n_1} + \frac{x_2^2}{n_2} - p_1 p_2}}$$

Como en general se desconoce  $p_1$  y  $p_2$  tendremos que aproximar sus valores para poder calcular la desviación estándar de que aparecen en el denominador de  $z$ . Existen aproximaciones para dos casos:

*Caso I:* Si suponemos que  $p_1$  es igual a  $p_2$ , o sea,  $D_0=0$  entonces  $p_1=p_2=p$ , y el mejor estimador de  $p$  se obtiene acumulando ponderativamente los datos de ambas muestras. De modo que, si  $x_1$  y  $x_2$  son los números de éxitos obtenidos de las dos muestras, entonces la estimación ponderada de  $p$  es

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

El estadístico de la prueba es:  $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}}}$

$$= \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}}}$$

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}}}$$

*Caso II:* Si suponemos que  $D_0$  no es igual a cero, es decir  $D_0 \neq 0$  entonces los mejores estimadores para  $p_1$

$$\hat{p}_1 = \frac{x_1}{n_1}$$

son  $\hat{p}_1$  y  $\hat{p}_2$  respectivamente

El estadístico de la prueba será:

$$z = \frac{\hat{p}_1 - \hat{p}_2 - D_0}{\sqrt{\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2}}$$

$$= \frac{\hat{p}_1 - \hat{p}_2 - D_0}{\sqrt{\hat{p}_1 \hat{q}_1 + \hat{p}_2 \hat{q}_2}}$$

$p_1 + 22$

$n_1 n_2$

4. Región de rechazo:

*Prueba de una extremidad*

$z > z_{\alpha}$  (o bien  $z < -z_{\alpha}$ , cuando la hipótesis alternativa es  $H_a: (p_1 - p_2) < D_0$ )

*Prueba de dos extremidades*

$z > z_{\alpha/2}$  o bien  $z < -z_{\alpha/2}$

$\alpha$   $\alpha/2$   $\alpha/2$

$z_{\alpha} - z_{\alpha/2} z_{\alpha/2}$

*Suposiciones:* Se seleccionaron las muestras aleatorias e independientes de dos poblaciones binomiales y  $n_1$  y  $n_2$  suficientemente grandes para que las distribuciones muestrales de  $p_1$

1

y  $p_2$

2

(y por lo tanto, de

—

$p_1$ ) sea aproximadamente normales.

### 1.6.2. Muestras pequeñas: media, varianza y correlaciones muestrales

**Media:**

1. Hipótesis nula:  $H_0: m = m_0$

2. Hipótesis alternativa: *Prueba de una extremidad (o cola)*  $H_a: m > m_0$  o bien  $H_a: m < m_0$  *Prueba de dos extremidades (o colas)*  $H_a: m \neq m_0$

3. Estadístico de la prueba:

$t$

=

$\frac{\bar{x} - m_0}{s/\sqrt{n}}$

4. Región de rechazo:

*Prueba de una extremidad*  $t > t_{\alpha}$  (o bien  $t < -t_{\alpha}$  cuando la hipótesis alternativa es  $H_a: m < m_0$ ) *Prueba de dos extremidades*

$t > t_{\alpha/2}$  o bien  $t < -t_{\alpha/2}$

$\alpha$   $\alpha/2$   $\alpha$

$t_{\alpha} - t_{\alpha/2} t_{\alpha/2}$

$t_{\alpha} - t_{\alpha/2} t_{\alpha/2}$

Los valores de  $t$ ,  $t_a$  y  $t_{a/2}$  se basan en  $(n-1)$  grados de libertad.  
*Suposiciones:* La muestra se seleccionó aleatoriamente de una población distribuida normalmente

**Diferencia de medias-muestras independientes:**

1. *Hipótesis nula:*  $H: (m_1 - m_2) = D_0$  donde  $D_0$  es alguna diferencia especificada que se quiera probar. En muchos casos se deseará probar la hipótesis de que no hay una diferencia entre  $m_1$  y  $m_2$ , es decir,  $D_0 = 0$ .

2. *Hipótesis alternativa:*

*Prueba de una extremidad (o cola)*

$H_a: (m_1 - m_2) > D_0$  o bien  $H_a: (m_1 - m_2) < D_0$

*Prueba de dos extremidades (o colas)*

$H_a: m_1 - m_2 \neq D_0$

3. Estadístico de la prueba:

$t = \frac{\bar{x}_1 - \bar{x}_2 - D_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

donde

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

$\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i$

$\bar{x}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} x_i$

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

$\hat{\sigma}_1^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2$

$\hat{\sigma}_2^2 = \frac{1}{n_2} \sum_{i=1}^{n_2} (x_i - \bar{x}_2)^2$

donde

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

$s_1 = \sqrt{\frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2}$

$s_2 = \sqrt{\frac{1}{n_2} \sum_{i=1}^{n_2} (x_i - \bar{x}_2)^2}$

$n = n_1 + n_2$

4. Región de rechazo:

*Prueba de una extremidad*

$t > t_a$  (o bien  $t < -t_a$ , cuando la hipótesis alternativa es  $H_a: m_1 - m_2 < D_0$ )

*Prueba de dos extremidades*  $t > t_{a/2}$  o bien  $t < -t_{a/2}$

donde  $t_a = t_{a/2}$

$t_{a/2}$

$t_a = t_{a/2}$

Los valores de  $t$ ,  $t_a$  y  $t_{a/2}$  se basan en  $(n_1 + n_2 - 2)$  grados de libertad.

*Suposiciones:* Se seleccionaron las muestras aleatorias e independientes de dos poblaciones normalmente distribuidas. Las

varianzas poblacionales  $s_1^2$  y  $s_2^2$  son iguales.

Si las varianzas no son iguales entonces:

$t = \frac{\bar{x}_1 - \bar{x}_2 - D_0}{\sqrt{\hat{s}^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$  donde  $t$  se distribuye según una  $t$  de Student con  $m$  gra

$n_1 + n_2 - 2$

dos de libertad.

$\hat{s}^2 = \frac{\hat{s}_1^2 + \hat{s}_2^2}{2}$

Con  $m = n_1 + n_2 - 2$

$\hat{s}_1^2$  y  $\hat{s}_2^2$

$n_1$

$n_2$

$\hat{s}_1^2$  y  $\hat{s}_2^2$

$n_1$

$n_2$

$n_1 + n_2 - 2$

$\hat{s}_1^2$  y  $\hat{s}_2^2$

$n_1 + n_2 - 2$

$n_1 + n_2 - 2$

### Diferencia de medias -muestras relacionadas:

1. Hipótesis nula:  $H_0: (\mu_1 - \mu_2) = \mu_d = 0$

2. Hipótesis alternativa:

*Prueba de una extremidad (o cola)*  $H_a: \mu_d > 0$  o bien  $H_a: \mu_d < 0$

*Prueba de dos extremidades (o colas)*  $H_a: \mu_d \neq 0$

3. Estadístico de la prueba:  $t = \frac{\bar{d} - D_0}{\sqrt{\frac{ss_d}{n}}}$

$n$  donde  $n$  = número de diferencias por parejas

$$n \hat{\sigma}_{\bar{d}}$$

$s_d$

$d$

$=$

$i=1$

$n - 1$

4. Región de rechazo:

*Prueba de una extremidad*

$t > t_{\alpha}$  (o bien  $t < -t_{\alpha}$ , cuando la hipótesis alternativa es  $H_a: \mu < 0$ ) *Prueba de*

dos extremidades

$t > t_{a/2}$  o bien  $t < -t_{a/2}$

$a/2$

$t_{a/2}$

$t_{a/2}$

Los valores críticos de  $t$ ,  $t_a$  y  $t_{a/2}$  se basan en  $(n-1)$  grados de libertad.  
*Suposiciones:* Se seleccionan aleatoriamente las  $n$  diferencias por parejas de una población normalmente distribuida.

### **Varianza poblacional .**

1. Hipótesis nula:  $H_0: s^2 = s_0^2$

2. Hipótesis alternativa:

*Prueba de una extremidad (o cola)*

$H_a: s^2 > s_0^2$  o bien  $H_a: s^2 < s_0^2$

*Prueba de dos extremidades (o colas)*

$a$

:

$s$

$2\pi s^2$

$0$   $(ns^2)$

3. Estadístico de la prueba:

$2$

$c = s_0^2$

4. Región de rechazo:

*Prueba de una extremidad*

$c^2 > c_a^2$  (o bien  $c^2 < c_{c_a}$ , cuando la hipótesis alternativa es  $H_a: s^2 < s_0^2$ )

donde  $a$

$c^2$  y  $c_{c_a}$  son los valores de la cola superior e inferior respectivamente de  $a$

$c^2$  que ponen  $a$  en las áreas de las colas.

*Prueba de dos extremidades*

$c^2 > c_a^2$  o sea  $c^2 < c_{c_{a/2}}$ , donde  $c^2$  y  $c_{c_{a/2}}$  son los valores de la cola superior e inferior respectivamente de  $a/2$

que ponen  $a/2$  en las áreas de las colas.

$a/2$   $a/2$

$c$

$2 c_2 c_2^a 1-a/2^{a/2}$

Los valores críticos de  $c^2$  se basan en  $(n-1)$  grados de libertad.

*Suposiciones:* La muestra se seleccionó aleatoriamente de una población normalmente distribuida.

### **Igualdad de dos varianzas poblacionales .**

1. Hipótesis nula:  $H_0: s_1^2 = s_2^2$

2. Hipótesis alternativa:

*Prueba de una extremidad (o cola) H*

$a$

:

$s$

$12$

$>$

$s$

$2^2$

$2$  o bien  $H_a: s_1^2 < s_2^2$  *Prueba de dos extremidades (o colas) H*

$a$

:

$s$

$12$

$\pi$

$s$

$2$

$2$

3. Estadístico de la prueba:

*Prueba de una extremidad:*

$s$

$2s^2$

$F^1$  (o bien,  $F_{s_2}$  para  $H_a: s_1^2 < s_2^2$ )  $s$

$2 = =^2$

$2 1$

*Prueba de dos extremidades:*

$s$

$2$

$F$

$=$

$1$

$s$

$s_2$  donde  $s_1$  es la mayor varianza muestral

4. Región de rechazo:

*Prueba de una extremidad*  $F > F_{\alpha}$

*Prueba de dos extremidades*  $F > F_{\alpha/2}$

Cuando

$F$

=

$s$

$s_2$

$F_{\alpha}$  y  $F_{\alpha/2}$ , se basan en  $n_1 = n_1 - 1$  y  $n_2 = n_2 - 1$   $s_2$

$s_2$

grados de libertad.

$\alpha$

$F_{\alpha}$   $F_{\alpha/2}$

*Suposiciones:* Las muestras se seleccionaron aleatoria e independientemente de poblaciones con distribución normal.

**Prueba para el coeficiente de correlación poblacional simple  $r$**

1. Hipótesis nula:  $H_0: r = 0$

2. Hipótesis alternativa:

*Prueba de una extremidad (o cola)*

$H_a: r > 0$  o bien  $H_a: r < 0$

*Prueba de dos extremidades (o colas)*  $H_a: r \neq 0$

3. Estadístico de la prueba:  $t = r \sqrt{\frac{n-2}{1-r^2}}$

4. Región de rechazo:

*Prueba de una extremidad*

$t > t_{\alpha}$  (o bien  $t < -t_{\alpha}$ , cuando la hipótesis alternativa es  $H_a: r < 0$ ) *Prueba de dos extremidades*

$t > t_{\alpha/2}$  o bien  $t < -t_{\alpha/2}$

$\alpha$

$\alpha/2$

$t_{\alpha}$   $t_{\alpha/2}$   $t_{\alpha/2}$

Los valores críticos de  $t$ ,  $t_{\alpha}$  y  $t_{\alpha/2}$  se basan en  $(n-2)$  grados de libertad.

*Suposiciones:* Las suposiciones son las relacionadas con el modelo de regresión lineal y se verán en la sección dedicada a dicho modelo.

**1.7. Pruebas para comparación de medias. Los comandos T-TEST Y**

## MEANS

### El comando T-TEST

Permite aplicar la prueba t de Student para una muestra, para dos muestras relacionadas o para dos muestras independientes.

#### 1.7.1. Prueba T para una muestra

Se utilizará para comparar la media de una variable con un valor conocido o que queremos inferir.

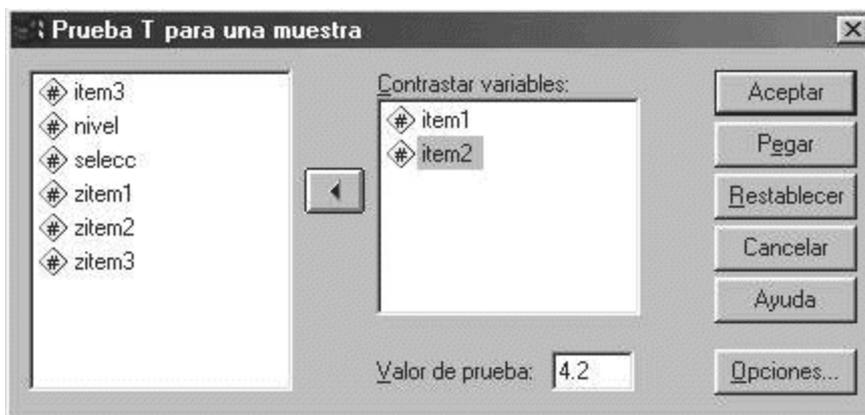
Supongamos el siguiente problema-ejemplo.

#### Problema-ejemplo

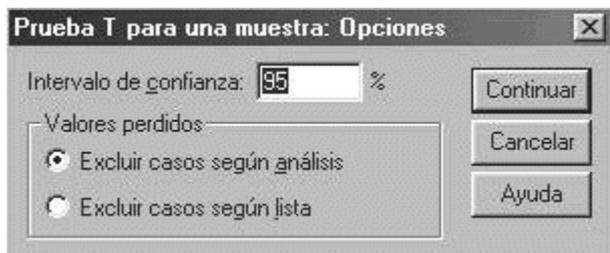
Queremos calcular el estadístico t, para contrastar la hipótesis de que la media es igual a 4,2, en cada una de las variables item1 y item2, del ejemplo práctico que venimos manejando (fichero examinar.sav).

#### Desarrollo del ejemplo

Seleccionaremos en el menú del SPSS, **Analizar>Comparar medias>Prueba T para una muestra** que nos llevará a la siguiente ventana:



Aquí seleccionaremos las variables item1 y item2 y escribiremos el valor que queremos contrastar (4,2). Después pulsaremos **opciones** para indicar el intervalo de confianza y el tratamiento de los datos perdidos.



*Fichero de sintaxis*

T-TEST

```
/TESTVAL=4.2  
/MISSING=ANALYSIS  
/VARIABLES=item1 item2  
/CRITERIA=CIN (.95) .
```

### **Resultados prueba T para una sola muestra ESTADÍSTICOS PARA UNA MUESTRA N Media**

#### **Desviación típ. Error tip. de la media**

Comprensión lectora	150	Aptitud verbal	150	4.1133	2.7963	4.2400
	2.9440		.2283		.2404	

#### **PRUEBA PARA UNA MUESTRA**

**Valor de prueba = 4.2** Intervalo de confianza para la diferencia

t gl Sig. Diferencia (bilateral) de medias Inferior Superior

Comprensión -.380 149 lectora

.705 -8.6667E-02

-.5378 .3645

aptitud verbal .166 149 ,868 4.000E-02

-.4350 .5150

Como podemos observar, se plantea la hipótesis  $H_0: m=4,2$  frente a la alternativa  $H_1: m \neq 4,2$ , nos da la significación que es mayor de 0,05 luego aceptamos  $H_0$ . Incluso nos da el intervalo de confianza de la media poblacional.

**1.7.2. Prueba T para muestras independientes** Compara la igualdad de medias para dos grupos independientes.

#### **Problema-ejemplo**

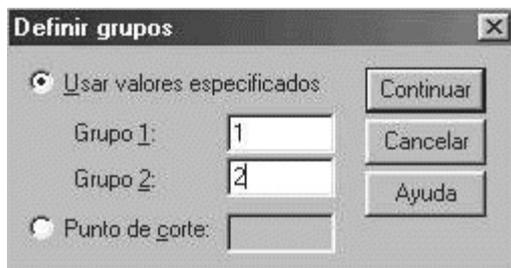
Supongamos que queremos saber si existe diferencia significativa en la variable item1 en función de la variable sexo del alumno, en los datos del ejemplo que venimos tratando (fichero examinar.sav).

#### **Desarrollo del ejemplo**

Será necesario indicar al SPSS: **Analizar>Comparar medias>Prueba T para muestras independientes** que nos llevará a la siguiente ventana:



Aquí señalaremos las variables para contrastar, en nuestro caso item1 y la variable de agrupación que será sexo. En esta variable habrá que definir los grupos (en nuestro caso 1=hombres y 2=mujeres).



También como opción alternativa está el punto de corte que nos permitirá dividir una variable en función de un valor de corte, de tal forma que los valores menores que ese valor formarán un grupo y aquellos mayores o iguales que este valor formará otro grupo, así podremos contrastar la igualdad de medias en los dos grupos formados.

Después podemos pulsar **opciones** que nos permite seleccionar el intervalo de confianza y el tratamiento de valores perdidos.



*Fichero de sintaxis*

T-TEST

/ROUPS=sexo(1 2) /MISSING=ANALYSIS /VARIABLES=item1

/CRITERIA=CIN(.95) .

## Resultados de la prueba T para muestras independientes ESTADÍSTICOS DEL GRUPO

sexo

N

**Media**

**Desviación típ. Error tip. de la media**

Comprensión lectora hombre 76 4.1447 2.7699 mujer 74 4.0811  
2.8417 .3177 .3303

PRUEBA DE MUESTRAS INDEPENDIENTES

**Comprensión lectora**

**Se han asumido varianzas iguales**

**No se han asumido  
varianzas iguales**

Prueba de Levene para la igualdad de varianzas F  
Sig. .100 .752

Prueba T para la igualdad de medias

t

gl

sig. (bilateral)

Diferencia de medias Error típ. de la diferencia Intervalo de confianza para  
la diferencia

Inferior Superior .139

148

.890

6.366E-02 .4582

-.8418 .9691

.139

147.594 .890

6.366E-02 .4583

-.8421 .9694

Como los condicionantes para aplicar la prueba son igualdad de varianzas y normalidad, aunque respecto a esta última, la prueba t es muy robusta para desviaciones de la normalidad, sólo será interesante comprobar que las distribuciones son simétricas, y que no contiene valores

atípicos.

En este ejemplo observamos que se puede asumir que las varianzas son iguales y además, según el contraste, vemos no existe diferencia significativa para los grupos formados.

### 1.7.3. Prueba T para muestras relacionadas

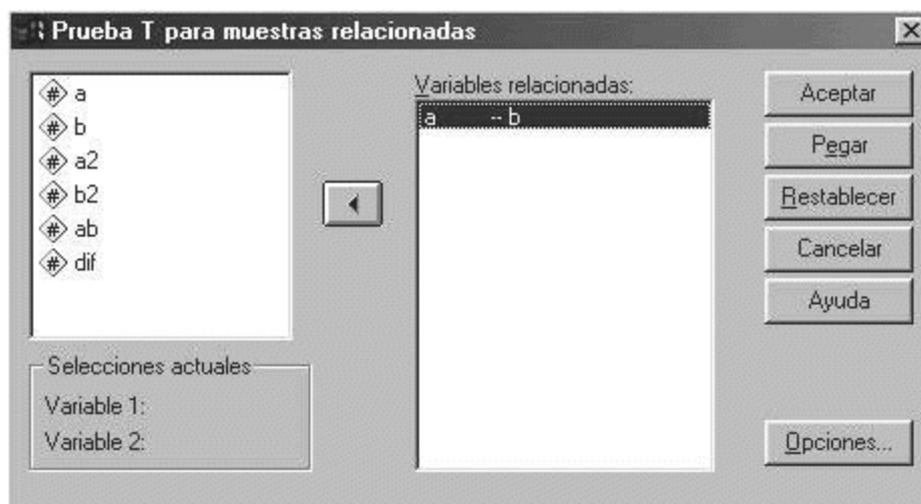
Compara la igualdad de medias para dos grupos relacionados.

#### Problema-ejemplo

Supongamos que tenemos los resultados del pretest y postest de una prueba de comprensión lectora realizada a un grupo de 34 alumnos de una escuela pública (archivo t-testrel.sav). Queremos saber si existe diferencia significativa de medias entre los resultados de estos alumnos, antes y después del test.

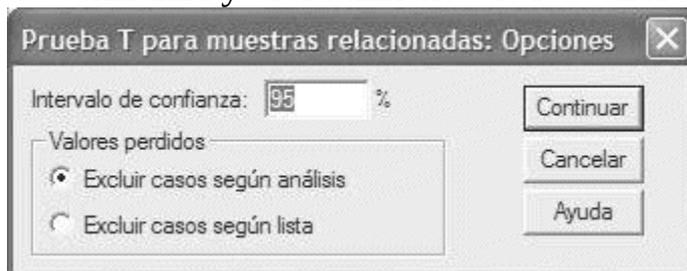
#### Desarrollo del ejemplo

Será necesario indicar al SPSS: **Analizar>Comparar medias>Prueba T para muestras relacionadas** que nos llevará a la siguiente ventana:



Iremos

seleccionando las variables por pares. En nuestro caso hemos elegido a y b que se corresponde con la variable bajo estudio, antes y después de la prueba. Si pulsamos el menú opciones se entrará a configurar el intervalo de confianza y el tratamiento de los valores perdidos.



*Fichero de sintaxis*

T-TEST

PAIRS= a WITH b (PAIRED) /CRITERIA=CIN(.95)  
/MISSING=ANALYSIS.

**Resultado de la prueba T para muestras relacionadas**  
ESTADÍSTICOS DEL GRUPO

**Media**

**N**

**Desviación típ.**

**Error tip. de la media**

Par 1 A 60.0000 34 8.9341 1.5322

B 54.0000 34 10.7280 1.8398

CORRELACIONES DE MUESTRAS RELACIONADAS

**N Correlación Sig.**

Par 1 A y B 34 .903 .000

PRUEBA DE MUESTRAS RELACIONADAS

**Par 1**

**A-B**

Diferencias relacionadas

t

gl

Sig. (bilateral) Media

Desviación típ.

Error típ. de la media

Intervalo de confianza para la diferencia

6.0000

4.6775 .8022

Inferior 4.3680 Superior 7.6320

7.480 33 .000

Se contrasta la hipótesis de que no hay diferencia entre la media antes y después de la prueba. Según vemos en los resultados, si hay diferencia significativa, es mayor en A respecto a B. También se calcula la correlación entre las notas y su significación estadística, bajo la hipótesis de  $r=0$ .

**1.7.4. El comando MEANS (Medias)**

Calcula estadísticos descriptivos básicos para una o más variables dependientes en los grupos de casos definidos por los valores de una o más variables independientes. Se puede obtener el análisis de la varianza de un factor, la eta y una prueba de linealidad.

Las variables dependientes son cuantitativas y las independientes categóricas. Los valores de las variables categóricas pueden ser numéricos o de cadena corta.

### **Problema-ejemplo**

Supongamos el siguiente **problema-ejemplo**.

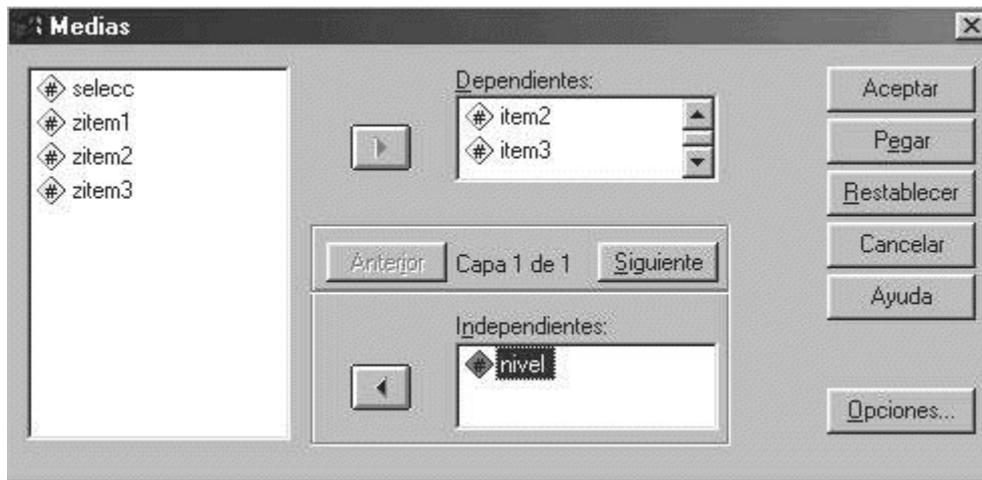
Se pretende saber las medias de tres ítems: ITEM1 «Prueba de comprensión lectora», ITEM2 «Prueba de aptitud verbal» e ITEM3 «Valoración en una prueba de ingles», resultados de las pruebas realizadas a 150 alumnos de un instituto de enseñanza secundaria, medidos en una escala de 0 a 10 puntos en función de la variable NIVEL:

1. «Primarios sin c. escolaridad».
2. «Certificado escolaridad».
3. «Graduado escolar».
4. «FP-I».
5. «BUP/COU».
6. «FP-II».
7. «Titulado grado medio».
8. «Titulado grado superior».

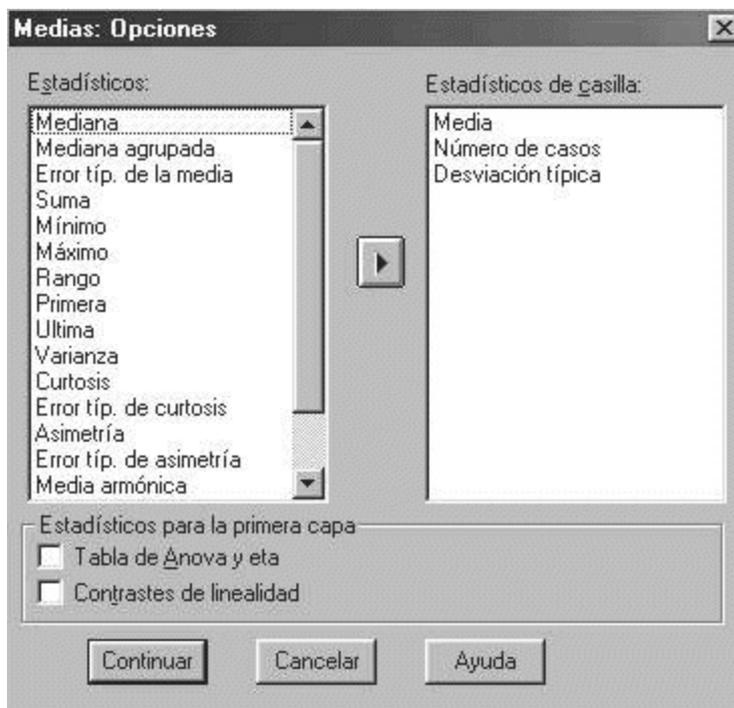
### **Desarrollo del ejemplo**

Si utilizamos el SPSS para realizar el problema-ejemplo propuesto la secuencia de ventanas del menú será la siguiente: **Analizar >Comparar medias> Medias**.

Que nos lleva a la siguiente ventana:



Seleccionamos las variables dependientes: ITEM1, ITEM2 y ITEM3 y la variable independiente o factor NIVEL. Después pulsaremos **Opciones** para seleccionar los estadísticos.



Se puede seleccionar uno o varios estadísticos de los indicados a continuación para las variables dentro de cada categoría de la variable independiente o factor: suma, número de casos, media, mediana, mediana agrupada, error típico de la media, mínimo, máximo, rango, valor de la variable de la primera categoría de agrupación, valor de la variable de la última categoría de la variable de agrupación, desviación típica, varianza, curtosis, error típico de la curtosis, asimetría, error típico de la asimetría.

Además, permite obtener un análisis de la varianza para cada una de las variables dependientes considerando únicamente la primera lista de

variables independientes. También se puede obtener las medidas de asociación eta y eta cuadrado. Eta cuadrado es la proporción de la varianza en la variable dependiente que se explica por la diferencia entre los grupos. Resulta de dividir la suma de cuadrado entre grupos, por la suma de cuadrados total.

Además se puede observar la linealidad del modelo cuando ordenamos las categorías de la variable independiente. Como indicador de la medida de ajuste tenemos R y R<sup>2</sup>.

*Fichero de sintaxis*

```
TMEANS  
TABLES=item1 item2 item3 BY nivel /CELLS MEAN COUNT  
STDDEV .
```

### **Resultados del comando MEANS INFORME**

#### **Nivel**

Primarios sin c. escolaridad

Certificado escolaridad

Graduado escolar

FP-I

BUP/COU

FP-II

Título de Grado Medio

Título de Grado Superior

*Total*

#### **Comprensión Aptitud Prueba lectora verbal de inglés**

Media 3.6785 4.2500 3.9375

N161616

Desv. típ. 3.1983 3.2762 3.2958

Media 4.6667 2.9333 5.6667

N151515

Desv. típ. 2.6095 2.7894 2.9439

Media 4.0476 4.7143 4.5238

N212121

Desv. típ. 2.9913 2.2835 3.1878

Media 4.4375 3.6563 3.9375

N323232

Desv. típ. 3.0367 2.8240 2.9939

Media 4.1364 3.9545 5.8182  
N222222

Desv. típ. 2.5690 2.9677 3.3329

Media 4.6842 4.2632 3.4211  
N191919

Desv. típ. 2.9069 3.1419 2.5015

Media 3.6429 5.9286 3.8571 N141414 Desv. típ. 2.6489 2.6736  
2.8516

Media 2.7273 5.1818 5.0000  
N111111

Desv. típ. 1.6787 3.4876 2.2804

Media 4.1133 4.2400 4.4733  
N 150 150 150

Desv. típ. 2.7963 2.9440 3.0360

Como podemos observar en la tabla denominada **informe** aparecen los estadísticos seleccionados (media, n y desviación típica). Si hubiéramos marcado el análisis de la varianza de los item1, item2 y item3 en función del factor NIVEL aparecería otra tabla con su estudio, asimismo si se selecciona el contraste de linealidad, se presentaría otra tabla dedicada a informar sobre el grado de bondad de ajuste al modelo lineal y la proporción de la varianza explicada en la variable dependiente por la diferencia entre grupos del factor.

## **2. DECISIÓN ESTADÍSTICA. PRUEBAS NO PARAMÉTRICAS**

### **2.1. Introducción**

Todos los estadísticos y las razones críticas de z, t o F de Fisher parten de unos supuestos denominados paramétricos: normalidad de la distribución base, la medición de los datos por lo menos en una escala de intervalo, la igualdad de varianzas entre las poblaciones, la independencia de las observaciones realizadas de modo que la selección de un dato cualquiera no afecte a las probabilidades de selección de otro distinto, etc.

En la práctica aparecen situaciones en las que tales requisitos no se cumplen, como el caso de distribuciones claramente asimétricas o muestras muy pequeñas. En estas ocasiones existen otros métodos, denominados no paramétricos, que no suponen nada sobre la distribución poblacional básica, a lo sumo la continuidad de la función de distribución o la continuidad de la función de densidad.

Estos métodos aportan unas ventajas:

— Capacidad para trabajar con datos que sean mediciones cuantitativas ordinales o incluso nominales.

— En general no necesitan que se cumplan supuestos previos para su aplicación.

— Son sencillos de aplicar.

— Queda como única posibilidad cuando el tamaño de la muestra es pequeño. Tienen unos inconvenientes:

— Para la misma potencia de la prueba, los test paramétricos necesitan menor tamaño muestral que los test no paramétricos.

— Cuando el tamaño de la muestra es elevado obtenemos los mismos resultados con las pruebas paramétricas que con las no paramétricas.

Podemos establecer una clasificación de las pruebas no paramétricas atendiendo a la organización de los datos y si son utilizadas preferentemente con medidas nominales u ordinales. Esta clasificación aparece en la tabla 4.2.1.

TABLA 4.2.1. Clasificación de las pruebas no paramétricas.

**Organización de los datos Utilizada con medidas nominales**

**Utilizada con medidas ordinales**

Una muestra Ji-cuadrado Binomial

Rachas

Kolmogorov-Smirnov

Dos muestras relacionadas McNemar Signos

Wilcoxon

K-muestras relacionadas Q-Cochran Análisis de la varianza de Friedman

Dos muestras independientes Ji-cuadrado

Prueba exacta de Fisher Mediana

U de Mann-Whitney Kolmogorov-Smirnov Rachas

K-muestras

independientes Ji-cuadrado Mediana

Análisis de la varianza de Kruskal-Wallis

## **2.2. Pruebas no paramétricas a partir de una sola muestra**

Cuando tenemos una única muestra las propiedades más importantes que podemos estudiar son las siguientes:

1. Comprobar el supuesto de que la muestra proviene de una distribución conocida (bondad de ajuste).

2. Ver si la muestra es aleatoria para saber si los resultados se pueden

extender a toda la población o a un único extracto o por el contrario no se puede extender. 3. Comprobar la simetría de la muestra.

### **2.2.1. De bondad de ajuste**

#### **2.2.1.1. Kolmogorov- Smirnov**

El método de Kolmogorov-Smirnov se basa en la comparación entre las frecuencias acumuladas de la distribución empírica de la muestra y de la distribución hipotética teórica, fijándose en el punto en el que las dos distribuciones presentan mayor divergencia. Sólo se puede aplicar para mediciones ordinales o de intervalo. Se calcula  $D_n$ =máxima desviación  $|F_n(x)- F(x)|$  donde:

$F_n$ = frecuencia relativa acumulada (observada)

$F$ = frecuencia relativa acumulada (teórica)

Una vez calculado  $D_n$ , construimos el contraste  $H_0$  (no existe discrepancia entre la distribución teórica y la observada) de la siguiente forma: fijado el nivel de significación buscamos en las tablas de valores críticos del test de Kolmogorov-Smirnov el valor  $D$  (que depende tanto del nivel de significación  $\alpha$  como del tamaño muestral  $n$ ) tal que  $P(D>D_\alpha)=\alpha$ .

Si  $D_n>D$  ( $R.C_\alpha\{D>cte\}$  ) entonces **se rechaza** la hipótesis nula sobre igualdad de distribuciones, en caso contrario se acepta.

La prueba de Kolmogorov-Smirnov tiene una ventaja en relación con la prueba de  $\chi^2$ , que se verá a continuación, no se ve afectada por los agrupamientos de intervalos y se puede aplicar en muestras pequeñas.

Para observar la bondad de ajuste cuando se trata de una distribución Normal puede que algunas veces interese estimar su media y su varianza. En este caso utilizaremos el contraste de **Kolmogorov-Smirnov-Lilliefors**.

Este contraste se define de la misma forma que el contraste de KolmogorovSmirnov pero ahora no se supone el valor de la media y de la varianza sino que se estiman mediante sus valores muestrales.

### **Ejemplo 1**

Se supone que los niños menores de 10 años prefieren regalos en los que se desarrolla más actividad. Para confirmar esta hipótesis se reúnen 15 niños a los que se les propone que elijan un regalo entre cuatro posibles: cuentos, pinceles-pinturas, juguetes mecánicos y juguetes de «guerra», obteniéndose los siguientes resultados:

Número de niños cuentos 2  
 pinceles-pinturas 1  
 juguetes mecánicos 3  
 juguetes de guerra 9

Los cuentos, por ser el regalo que menor actividad hace desarrollar recibe el puesto o rango 1; los pinceles-pinturas el rango 2, los juguetes mecánicos el rango 3 y, por fin, los juguetes de «guerra» el rango 4.

Si los regalos son elegidos independientemente de la actividad que hacen desarrollar serán escogidos por igual número de niños ( $P_x=1/4$ )

La hipótesis nula sostiene que la muestra puede provenir de una población en la que los regalos se eligen por igual y sin preferencias (distribución uniforme).

1.º	2.º	3.º	4.º	Número de niños que eligen cada rango	2	1	3	9	$F_n(x)$	=
				frecuencias relativas acumuladas (observadas)	2/15	3/15	6/15	15/15	$F(x)$	=
				frecuencia relativa acumulada(teórica)	1/4	2/4	3/4	4/4	$ F_n(x)-F(x) $	7/60
					18/60	21/60	0			

En este caso  $D_n$ =máxima desviación=21/60=0,35

Si buscamos en las tablas de Kolmogorov-Smirnov  $D(N=15, \alpha=0,05)$  tenemos  $D=0,338$  luego como  $D_n > D$  se rechaza la hipótesis nula. Luego se admite que los niños menores de 10 años prefieren aquellos regalos que les supongan acción y actividad.

### Ejemplo 2: Contraste Kolmogorov- Smirnov- Lillefors

Se contrasta la hipótesis que los datos siguientes provienen de una distribución Normal.

20,22,24,30,31,32,38

La hipótesis planteada es:  $H_0:N(m,s)$  frente  $H_1: no N(m,s)$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{20+22+\dots+38}{7} = 28,14$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 = \frac{1}{7} [(20-28,14)^2 + (22-28,14)^2 + \dots + (38-28,14)^2]$$

$$s_{n-1} = \frac{2814}{6} = 40,83; \hat{\sigma} = 6,39$$

Además  $x_{(i)} = (20, 22, 24, 30, 31, 32, 38)$

$\hat{F}(x) = 0$  si  $x < x_{(1)}$

$\hat{F}(x) = \frac{i}{n}$  si  $x_{(i)} \leq x < x_{(i+1)}$

$\hat{F}(x) = 1$  si  $x > x_{(n)}$

$F(x) = P\{z \leq (x-28,14)/6,39\}$  así  $F(20) = P\{z \leq (20-28,14)/6,39\} = 0,1020$  Así:

x	20	22	24	30	31	32	38	$F_n(x)$	1/7	2/7	3/7	4/7	5/7	6/7	7/7	$F(x)$
$ F_n(x) - F(x) $	0,1685	0,2578	0,6141	0,6736	0,7258	0,9383		0,04	0,1165	0,1702	0,0431	0,0404	0,1312	0,0617		

$$D = \max\{D_n\} = 0,1702$$

$RC_{\alpha}\{D > c\}$ ;  $RC_{0,05}\{D > 0,38\}$  como  $0,1702 < 0,38$  aceptamos  $H_0$  y por tanto se trata de una distribución Normal.

### 2.2.1.2. Contraste de $\chi^2$ de bondad de ajuste

Como el contraste anterior tiene la misión de observar si una variable tiene una distribución de probabilidad dada. Por tanto, la hipótesis nula  $H_0$  hará referencia a que las observaciones muestrales constituyen un conjunto de  $n$  valores procedentes de una variable con distribución de probabilidad dada  $F(x)$ .

$$\chi^2 = \sum_{i=1}^n \frac{(o_i - e_i)^2}{e_i}$$

Se calcula el estadístico  $\chi^2 = \sum_{i=1}^n \frac{(o_i - e_i)^2}{e_i}$  [1] donde  $\sum_{i=1}^n e_i = n$

$o_i$  = frecuencia observada y  $e_i$  = frecuencia esperada

Si el acuerdo entre las frecuencias observadas y esperadas es grande la diferencia  $(o_i - e_i)$  será pequeña y entonces  $\chi^2$  será pequeño. Si la diferencia es grande será grande.

La distribución muestral de  $\chi^2$  se aproxima muy bien con la distribución ji-cuadrada, si las frecuencias esperadas son al menos iguales a 5 y mejora para valores más grandes.

El número de grados de libertad  $n$  viene dado por:

El número de grados de libertad  $n$  viene dado por:

1.  $n = k - 1$  si las frecuencias esperadas se pueden calcular sin tener que estimar los parámetros de la población a partir de estadísticos muestrales.

Nótese que hemos restado 1 ya que si por ejemplo los datos de 50 casos están clasificados en dos categorías, tan pronto como sepamos que 35 casos pertenecen a una categoría, sabremos que 15 deben estar en la otra, de ahí que haya  $k-1$  grados de libertad.

2.  $n=k-1-m$  si las frecuencias esperadas se pueden calcular sólo estimando  $m$  parámetros de la población a partir de estadísticos de la muestra.

Este procedimiento se emplea para datos numéricos:

— Si la población es discreta y finita.

— Si la población es discreta pero infinita o es continua. En este caso lo que hacemos es formar categorías equiprobables o de forma que en cada categoría haya como mínimo 3 datos y que el número de clases sea mayor o igual que 5 y además que las categorías sean mutuamente exhaustivas y excluyentes.

El método para la obtención de  $\chi^2$  será el siguiente:

1. Se clasifican las frecuencias observadas en un número  $k$  de categorías. La suma de las frecuencias debe ser  $N$ , es decir, el número de observaciones independientes.

2. A partir de  $H_0$ , se determinan las frecuencias esperadas (las  $e_i$ ) para cada una de las  $k$ -celdillas. Cuando  $k > 2$ , si más del veinte por ciento de las  $e_i$  son menores que 5, habrá que combinar las categorías adyacentes cuando sea razonable hacerlo, reduciendo de ese modo el valor de  $k$  e incrementando los valores de algunas de las  $e_i$ . Cuando  $k=2$ , la prueba  $\chi^2$  para el caso de una muestra no puede usarse apropiadamente a menos que cada frecuencia esperada sea 5 o más. Si después de agruparse las categorías las  $e_i$  continúan siendo menores que 5 se aconseja el uso de la prueba **binomial** en lugar de la  $\chi^2$  para determinar la probabilidad asociada con la ocurrencia de las frecuencias observadas conforme  $H_0$ . Si  $N < 50$  no se aconseja utilizar esta prueba.

3. Con la fórmula [1] se calcula el valor de  $\chi^2$ .

4. Se determina el valor de los g.l. según hemos expuesto anteriormente.

5. Con los criterios expuestos sobre los g.l., fijado el nivel de significación  $\alpha$ , calculamos  $P[\chi^2 > \chi^2_{\alpha}$

$c$

$\chi^2 > \chi^2_{\alpha}$

$\alpha$ ] =  $\alpha$  empleando las tablas de la  $\chi^2$ .

La regla de decisión sobre la aceptación o rechazo del test se regirá por el siguiente criterio: si  $\chi^2_{n-1} > \chi^2_{k-1, \alpha}$  (  $RC_{\alpha}\{\chi^2_{n-1} > \chi^2_{k-1, \alpha}\}$  ) calculada entonces rechazamos la hipótesis nula y en caso contrario se acepta.

En contraste con la prueba de Kolmogorov-Smirnov esta prueba se puede aplicar cuando se desconocen parámetros poblacionales y para todo tipo de medidas.

### Ejemplo

Se sabe que el 30% de las pequeñas barras de hierro se romperán cuando se las someta a una carga de 3000 Kg. En una muestra aleatoria de 50 barras se encontró que 21 de ellas se rompían cuando se les sometía a esa carga. Investíguese si esta muestra proviene de la misma población.  $\alpha=0,01$ .

Se trata de una prueba de bondad de ajuste. Datos categóricos luego aplicaremos la prueba de  $\chi^2$ .

Hay dos categorías: se rompe (1.º) y no se rompe (2.º).

$H_0: P\{\text{rompe}\}=0,3 \quad H_1: p_1=0,3$

$P\{\text{no se rompe}\}=0,7 \quad p_2=0,7$

Calculemos el valor crítico de  $P[\chi^2_{n-1} > \chi^2_{k-1, \alpha}] = \alpha$  con  $\chi^2$  de g.l=2-1=1  $\chi^2_{n-1} = 7,879$ . Como  $\chi^2 = 3,429 < 7,879$  entonces nada se opone para aceptar la hipótesis  $H_0$ .

$\frac{o_i - e_i}{e_i} \quad \frac{(o_i - e_i)^2}{e_i}$  se rompe 21 50.0,3=15 36 36/15= 2,4 no se rompe 29 50.0,7=35 36 36/35=1,029  $\chi^2$  3,429

### 2.2.2. La prueba binomial

Existen ocasiones que se realizan investigaciones estadísticas sobre poblaciones a las que se considera divididas en dos categorías: por ejemplo, masculino-femenino, aprobados-suspensos, afiliados a un partido-no afiliados, etc. Naturalmente, en toda población dividida en dos categorías al conocer la proporción de casos en una categoría p se conoce automáticamente la proporción de casos en la otra categoría q ya que siempre  $p+q=1$ .

La prueba binomial se basa en la distribución binomial que es la distribución muestral de las proporciones (o frecuencias) observadas en muestras tomadas al azar de una población dividida en dos categorías.

## Método

La probabilidad de obtener  $x$  objetos en una categoría y  $N-x$  objetos en la otra es dada por:

$P(x) = \binom{N}{x} p^x q^{N-x}$  [2] donde  $p$ =proporción de casos esperados en una categoría

y  $q=1-p$ = proporción de casos esperados en la otra categoría, con

$$\binom{N}{x} = \frac{N!}{x!(N-x)!}$$

$N! / (x!(N-x)!)$

Por ejemplo, suponiendo un dado perfecto, la probabilidad de que salga seis en dos de cinco tiradas viene dada por la fórmula [2] cuya expresión numérica será la siguiente:

$$P(2) = \binom{5}{2} p^2 q^3 = 10 \cdot 0,4^2 \cdot 0,6^3 = 0,2592$$

También si deseamos conocer la probabilidad de obtener dos o menos veces el seis, cuando tiramos cinco veces un dado perfecto. Tenemos:

$$P(x \leq 2) = P(0) + P(1) + P(2) = 0,07776 + 0,2592 + 0,2592 = 0,59616$$

## Muestras pequeñas

Cuando se observa que el menor de  $Np$  ó  $Nq$  es menor que 10 consideraremos  $x$

**que  $N$  es pequeña, entonces emplearemos la fórmula:  $P(x) = \sum_{i=0}^x \binom{N}{i} p^i q^{N-i}$  si  $N < 10$**

$i=0$

do  $N$  el número de casos en la muestra,  $p$  la proporción en la población de casos favorables,  $q$  la proporción de casos desfavorables y  $x$  el número de casos favorables en la muestra.

## Ejemplo

Se sabe que en una determinada ciudad por cada 100 hombres que nacen, nacen 105 mujeres. Se pregunta por la probabilidad de que una familia con 4 hijos tenga o todos los hijos varones o todos los hijos hembras.

Sabemos que  $N=4$ ,  $p$ ={proporción de hombres}=100/205=0,49  $q$ =

{proporción de mujeres}=105/205=0,51;  $x=\{4 \text{ y } 0\}$

$$p(4) = \hat{E}^{40} 0,49^{40} = 0,0576$$

$$p(0) = \hat{E}^{00} 0,49^{00} = 0,271$$

$p$

$(0)$

$$= \hat{E}$$

$$\hat{E}^{\wedge}$$

$$0,49^{04}$$

$$(,).(,.) = 0,271$$

luego  $p(x) = p(4) + p(0) = 0,0576 + 0,271 = 0,329$  aproximadamente el 33%

### Muestras grandes

Si el menor de  $Np$  o  $Nq$  es mayor que 10 consideramos que  $N$  es grande, entonces la distribución binomial se aproxima con la distribución normal, pudiéndose solucionar las cuestiones según la distribución muestral de proporciones o frecuencias que sigue la distribución normal.

La prueba binomial, según esto, utilizará la razón crítica  $z$ :

— con proporciones

$$z = \frac{p - \hat{p}}{\sqrt{\hat{p}\hat{q}}}$$

$p, q, N$  — con frecuencias

$$z = \frac{f_m - m_x}{\sqrt{m_x N p q}}$$

Es importante tener en cuenta que cuando se trabaja con frecuencias se suele hacer la corrección por falta de continuidad ya que  $f_m$  de la distribución binomial es un dato discontinuo y la distribución normal exige datos continuos. Si  $f_m > m_x$  se toma  $f_m - 0,5$ . Si  $f_m < m_x$  se toma  $f_m + 0,5$ .

Entonces

$$z = \frac{f_m \pm 0,5 - m_x}{\sqrt{m_x N p q}}$$

### Ejemplo

Un comerciante al por mayor sabe que normalmente el 5% de las cajas de productos agrícolas que le envían durante el mes no sirven para la venta al público. ¿Qué probabilidad existe de que en un determinado mes, de 3400 cajas recibidas a lo más 150 cajas estén en mal estado?.

Tenemos  $f_m = 150$   $p = 0,05$  ;  $m_x = Np = 3400 \cdot 0,05 = 170$

$z = \frac{f_m - m_x}{\sqrt{m_x N p q}} = \frac{150 - 170}{\sqrt{170 \cdot 0,05 \cdot 0,95}} = -1,57$  por tanto  $P[z < -1,57] = 0,0582$  luego será el 5,82%.

$$3400 \cdot 0,05 \cdot 0,95$$

### 2.2.3. De aleatoriedad

#### 2.2.3.1. Prueba de rachas

Esta prueba tiene por objeto comprobar el carácter aleatorio de una muestra atendiendo a la obtención original de los valores u observaciones realizadas. Para ver la aleatoriedad de unos datos se analiza el número de rachas (sucesiones de símbolos idénticos) que se producen en la obtención de esos datos. Tanto si hay pocas rachas, como si su número es importante este será un motivo para afirmar que la muestra no es aleatoria y rechazar la hipótesis.

Como metodología para resolver el problema cogeremos los datos de forma dicotómica o dicotomizada (p y q) y contaremos las ocurrencias del hecho p (que denominaremos  $n_1$ ) y las ocurrencias del hecho q (por  $n_2$ ). Luego contaremos las rachas (r) número de ocurrencias de cambios de p a q en la serie objeto de estudio.

Se toma la decisión de acuerdo con el siguiente criterio:

— Si  $n_1$  y  $n_2$  tiene un valor igual o inferior a 20 se determinan los valores de rachas mínimos ( $r_1$ ) y rachas máximo ( $r_2$ ) mirando en las tablas correspondientes para un nivel de significación de por ejemplo  $\alpha=0,05$  se cumplirá  $P(r \leq r_1)=0,025$  y  $P(r < r_2)=0,975$  y entonces si el valor r cumple  $r_1 < r < r_2$  se acepta de hipótesis de aleatoriedad.

— Si  $n_1$  o  $n_2$  son mayores que 20 habrá que hacer una transformación de r en z de la siguiente forma:

$$z = \frac{r - \frac{1}{2}}{\sqrt{\frac{2n_1 n_2}{n_1 + n_2 + 1}}}$$

$$z = \frac{r - \frac{1}{2}}{\sqrt{\frac{2n_1 n_2}{n_1 + n_2 + 1}}}$$

12

0

## ***nn nn()***

Si  $z < z(\alpha)$  aceptamos  $H_0$  de aleatoriedad

Si  $z > z(\alpha)$  se rechaza  $H_0$  donde  $z(\alpha)$  se obtiene de la  $N(0,1)$

Si empleamos la prueba de dos colas la región de rechazo será  $z > z(\alpha/2)$  o bien

$z < -z(\alpha/2)$ .

Cuando los datos son numéricos una forma de aplicar el test de rachas es proceder como sigue:

- *Test sobre- y bajo mediana para la aleatoriedad de datos numéricos.*

Para determinar si unos datos numéricos son aleatorios, se colocan primero en el mismo orden que fueron cogidos, hallamos la mediana y sustituimos cada entrada por las letras p y q según que este valor esté sobre o bajo mediana. Si un valor coincide con la mediana se suprime. La muestra será aleatoria según lo sea la secuencia de p y q obtenida.

### **Ejemplo 1**

Supongamos los resultados de extraer 15 letras A y B según la siguiente secuencia: A,B,B,A,A,B,A,A,A,B,A,B,A,B,B. Queremos saber si la extracción ha sido aleatoria.

Datos:

$n_1 = 8$  letras A

$n_2 = 7$  letras B

$r = 10$  rachas

$\alpha = 0,05$

Cálculos:

Miramos en las tablas  $P(r \leq r_1) = 0,025$  y  $P(r < r_2) = 0,975$  para  $n_1$  y  $n_2$  dados llegamos a que  $r_1 = 5$  y  $r_2 = 12$  luego como  $r_1 < r < r_2$  aceptamos la hipótesis de aleatoriedad en la extracción.

### **Ejemplo 2**

La sucesión de 40 extracciones de A y B es la siguiente:

B,A,A,B,B,B,A,A,B,B,B,B,B,A,A,B,B,B,A,B,A,B,B,B,A,A,B,B,B,A,|  
B,B,B. Queremos saber si la extracción ha sido aleatoria.

Datos:

$n_1=13$  letras A

$n_2=27$  letras B

$r=19$  rachas

$\alpha=0,05$

Como  $N=40$  tenemos que emplear la aproximación a la normal, con la fórmula Cálculos:

$r$

$-$

$\hat{E} 2nn \wedge$

$\hat{E} \hat{A}$

$12 + 1 - z =$

$nn$

12

22

**12 12 1 2 nn nn()**

12

**( )**

2

**nn nn()**

Sustituyendo por los valores tenemos  $z=(19-18,55)/2,73=0,16$

La región de rechazo para una prueba de cola inferior con  $\alpha=0,05$  es  $z < -z_{0,05}$  o bien  $z < -1,645$ . Como el  $z$  obtenido (0,16) es mayor que -1,645 entonces nada se opone en aceptar la hipótesis  $H_0$ .

### **2.3. Pruebas bimuestrales (muestras relacionadas)**

Las pruebas bimuestrales para muestras relacionadas se usan cuando el investigador desea saber la diferencia entre dos tratamientos: por ejemplo comparar dos métodos de enseñanza, haciendo que cada sujeto sirva como su propio control.

En estas comparaciones de dos grupos, algunas veces se observan diferencias significativas que no son resultado del tratamiento. Por

ejemplo, para comparar dos métodos de enseñanza, un investigador elige un grupo al azar para aplicarle un método de enseñanza y otro grupo diferente para el otro método. El resultado puede no ser fiable porque en un grupo existan alumnos más capacitados que en el otro.

Una manera de resolver este problema es usar dos muestras relacionadas y esto se puede lograr si a cada individuo se le aplican los dos tratamientos o si se forman parejas lo más semejante posibles, con respecto a cualquier variable externa que pueda influir en el resultado de la investigación.

Siempre que sea posible es preferible aplicar a cada individuo los dos tratamientos, debido a que nuestra capacidad para formar parejas se ve limitada por la ignorancia de las variables que determinan su conducta.

La prueba paramétrica usual para analizar datos provenientes de dos muestras relacionadas o de parejas de datos igualados es aplicar una prueba de t a los puntajes de las diferencias, pero la prueba t exige que estas diferencias estén distribuidas normalmente y las medidas tienen que estar dadas por lo menos en una escala de intervalo.

Con estos condicionantes algunas veces no se puede aplicar la prueba t porque: a) los datos no se ajusten a la escala de medida indicada, b) las diferencias entre las puntuaciones no son datos numéricos o c) la distribución de las diferencias no es normal o independiente.

En estos casos el investigador puede escoger una de las pruebas estadísticas no paramétricas para dos muestras relacionadas. Estas pruebas tienen la ventaja que no requieren que las parejas provengan de una misma población.

### ***2.3.1. La prueba de Mc.Nemar para la significación de los cambios*** **Objetivo**

Es llamada prueba de «antes y después» y resulta adecuada para observar estos cambios. Los mismos sujetos constituyen el grupo experimental y control de la prueba. Se utiliza para detectar cambios de actitudes, por ejemplo: la efectividad de una visita personal para la venta de un producto, la construcción de un edificio para conseguir votos un partido político. En la prueba se presentan los datos de forma dicotomizada en dos categorías, según el esquema siguiente:

Después  
– + Antes + A B  
– C D

Como se puede observar en la tabla las filas representan las distintas respuestas antes de utilizar ningún método y las columnas las respuestas después de emplear el método que queremos estudiar.

En esta tabla vamos a designar por + y - los dos tipos de respuestas que se pueden dar, luego los cambios producidos van a aparecer en las casillas A y D en A serán los cambios de + a - y en D los cambios de - a +. Si no existen cambios entonces todos los casos estarán en las casillas B ó C.

**Hipótesis**

La hipótesis nula se plantea en los siguientes términos: no existe diferencia en las situaciones antes y después del tratamiento o estudio.

**Estadístico**

Puesto que A+D representa el número total de personas que cambiaron, se espera, según  $H_0$  que  $1/2(A+D)$  casos cambiarán en una dirección y  $1/2(A+D)$  cambiarán en la otra.

El estadístico que utilizaremos es:

$$2k() \text{ como } E_j=(A+D)/2 \text{ operando tenemos } c^2 = \hat{A}E_{ii}=1$$

$$\frac{c}{2} = \hat{A} AAD$$

$$() []$$

22

$$DAD()[] \text{ que después de cálculo resulta: } AD AD()/2^+ AD()/2$$

c

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$\hat{\chi}_{AD}^2$  [3] que se distribuye como una  $\chi^2$  con 1 g.l.  
 $AD$

Si hacemos la corrección por continuidad (se utiliza una distribución continua ji-cuadrado para aproximar una distribución discreta), entonces la fórmula [3] queda de la siguiente forma:

$c$

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$AD$

Si las frecuencias esperadas  $((A+D)/2)$ , es muy pequeña (menor que 5), deberá usarse la prueba Binomial antes de la prueba de McNemar. En el caso binomial  $N=A+D$ , y  $x$  es la menor de las dos frecuencias observadas  $A$  o  $D$ , y  $p=0,5$ .

### Decisión

Si  $\chi^2 < \chi^2$  crítico, se acepta  $H_0$

Si  $\chi^2 > \chi^2$  crítico, se rechaza  $H_0$ . El cambio es significativo.

### Ejemplo

Los 25 alumnos de una clase de 1.º de Bachillerato son sometidos a un entrenamiento en habilidades numéricas obteniéndose los siguientes resultados antes y después de la aplicación de esta metodología de apoyo. Se desea contrastar si el entrenamiento mejora estas habilidades.

Después

– + Antes + 14 4

– 3 4

Donde + indica clasificación mayor que la mediana y - menor que la mediana respectivamente en las puntuaciones de la prueba de habilidades numéricas.

$$\chi^2 = \frac{(2-2)^2}{2} + \frac{(1-1)^2}{1} + \frac{(4-4)^2}{4} + \frac{(1-1)^2}{1} = 0 + 0 + 0 + 0 = 0$$

AD

+

Cuya región de rechazo será  $P[ c^2 > c^2 ]=a$  que dará  $c^2 = 3,84$  como  $4,5 > 3,84_{1,a}$

rechazamos la hipótesis  $H_0$  sobre que no hay variación en el grupo de alumnos con el entrenamiento en habilidades numéricas.

Si hubiéramos aplicado la prueba binomial a los datos anteriores  $(A+D)/2$  entonces:  $(14+4)/2=9 > 5$  no sería necesario utilizar esta prueba, pero si aún así se hubiera hecho tendríamos  $N=A+D=18$  y  $x=4$  para  $P=Q=0,5$  que nos lleva a una significación  $p=0,015$  aproximadamente un valor casi similar al que se obtiene aplicando McNemar.

### 2.3.2. La prueba de los signos Objetivo

Es útil para muestras relacionadas cuando el experimentador desea establecer que ambas condiciones son diferentes. El único supuesto que se exige es la continuidad de la variable considerada y naturalmente que las muestras tengan la misma extensión para que se puedan comparar por pares. Las muestras pueden provenir de poblaciones distintas con la única condición que el experimentador haya logrado formar pares de semejantes características.

La prueba de los signos es de sencilla aplicación pero debido a que no posee información cuantitativa o si la posee la pierde ya que usa solamente los signos + o - (de ahí su denominación) y no las cantidades, es de eficacia limitada.

#### Hipótesis

Se halla para cada par de datos el signo de la diferencia, por ejemplo entre la muestra A y la muestra B, sean tales datos mediciones cuantitativas o no. Si alguna diferencia es cero ésta se suprime.

Supongamos que tenemos pares de muestras:  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  de poblaciones continuas cuyas funciones de densidad son  $f_1(x)$  y  $f_2(x)$  respectivamente.

Consideramos la hipótesis nula  $H_0: f_1(x)=f_2(x)$ , es decir, una vez calculado el número de diferencias positivas y negativas se verifica la hipótesis de que una muestra no es superior a la otra (igualdad de frecuencias relativas poblacionales):

$$p(X_A > X_B) = p(X_A < X_B) = 1/2$$

Entonces la variable  $z_i = 1$  si  $y_i - x_i > 0$  signo + ;  $0$  si  $y_i - x_i < 0$  signo -  
 $z_i \sim B(1; 0,5)$

será una binomial correspondiente a un solo experimento y  $p=0,5$ ; por lo que la

suma  $\sum_{i=1}^n z_i$  será una binomial correspondiente a  $n$  pruebas y  $p=1/2$ .  
 $Z \sim B(n, 1/2)$ .

Si la hipótesis alternativa  $H_1$  se define como  $f_1(x) = f_2(x-c)$  con  $c > 0$  entonces las  $x_i$  serán mayores que las  $y_i$  lo cual indica que debemos tomar como región crítica la rama de la derecha de la distribución binomial. Análogamente, si la hipótesis alternativa fuese la correspondiente a  $c < 0$ , tomaríamos como región crítica la rama izquierda.

En el caso de tomar como hipótesis alternativa  $c \neq 0$  tomaríamos la región crítica bilateral.

### Estadístico

Para **muestras pequeñas (N ≤ 25)** se aplica la distribución binomial con  $p=q=1/2$ , hallando la probabilidad asociada a la ocurrencia de  $S$  signos + (o  $S$  signos -) en  $N$  número de parejas igualadas cuyo puntaje de diferencia tiene un signo (no se contabiliza el caso de igualdad de valores) y siendo  $x$  el número menor de signos(+ o -).

$$p(x) = \binom{N}{x} 0,5^N$$

Para **muestras grandes (N > 25)** hacemos las siguientes consideraciones:

$$\hat{p} = \frac{x}{N} \approx p, \quad \hat{q} = \frac{N-x}{N} \approx q$$

1

$z = \frac{x - Np}{\sqrt{Npq}}$

$$z = \frac{x - Np}{\sqrt{Npq}}$$

haciendo la corrección por continuidad:

1

$$z = \frac{x - Np}{\sqrt{Npq}}$$

$$z = \frac{x - Np}{\sqrt{Npq}}$$

donde  $x+0,5$  se usa cuando  $x < 1/2 N$  y  $x-0,5$  cuando  $x > 1/2 N$

### Decisión

Si  $N$  es 25 o menos utilizamos la binomial para el cálculo del valor  $p$  asociado con el valor observado  $x$ , luego si  $p < 0,05$  (2p en la prueba de dos colas) rechazamos la hipótesis nula.

Si  $N > 25$  utilizamos la aproximación normal calculando  $z$  según la fórmula [5], luego obtenemos  $Q$  según  $P[Q > z_a] = a$  si  $z < Q$  aceptamos la hipótesis  $H_0$  (prueba de una cola).

### Ejemplo

Sean 12 adolescentes elegidos al azar entre miembros de un club de cine-forum. Interesa comprobar si después de la proyección de una película violenta, los adolescentes muestran una mayor agresividad que perdure incluso varios días. Realizados test, antes de la película y tres días después, resultan los siguientes datos:

#### Adolescentes Antes (A) Después (B) B-A

1 14 19 +  
2 16 18 +  
3 23 22 -  
4 26 27 +  
5 24 28 +  
6 28 35 +  
7 27 30 +  
8 18 17 -  
9 15 17 + 10 22 28 + 11 20 30 + 12 25 24 -

En este ejemplo tenemos, 9 signos positivos y 3 negativos al calcular las diferencias.

Si no hubiera diferencia tendríamos 6 signos positivos y otros tantos negativos. Como  $N < 25$  aplicamos la distribución binomial:

3

**$p(x) = \binom{12}{x} 0,5^x 0,5^{12-x}$  si efectuamos cálculos tendremos:  $p(x) = 0,079$  como**

$x=0$

$p > 0,05$  no puede afirmarse al nivel del 5% que existe diferencia significativa y de ahí que se acepte la hipótesis de que no existe ningún

influjo entre las películas violentas sobre la agresividad de los adolescentes de un club de cine-forum.

### **2.3.3. Prueba de pares igualados de Wilcoxon Objetivo**

La prueba T de Wilcoxon es muy similar a la prueba de los signos pero más potente en la medida en que utiliza la información adicional del tamaño de las diferencias entre los datos.

Se utiliza para datos cuantitativos continuos o discretos. Y como mínimo se exige que las variables estén dadas en escala ordinal.

#### **Hipótesis**

La hipótesis nula se plantea en el sentido de que no existe diferencia significativa entre los dos grupos de individuos (igualdad de frecuencias relativas poblacionales).

#### **Estadístico**

Con las diferencias entre los valores de cada par de sujetos, se establece una ordenación de esas diferencias desde la más pequeña a la mayor (independientemente del sentido o signo de la diferencia). Si el valor absoluto de esa diferencia coincide para dos o más datos se asignan a todos la media de los lugares que les correspondería. Si una diferencia es cero, no se toma en consideración. La suma de los rangos obtenidos para las diferencias positivas y negativas nos aporta dos valores,  $T_p$  y  $T_n$  de tal modo que se cumpla:

$T_p + T_n = N(N+1)/2$  (Suma de una progresión aritmética de razón 1) Si no hubiera diferencia significativa entre las dos muestras  $T_p$  y  $T_n$  tendrían que ser iguales con un valor cercano al promedio de los N rangos:

**$T$   $NN +()$  luego tomamos como T de Wilcoxon el menor de los valores en= 4**

tre  $T_p$  y  $T_n$  .

#### **Decisión**

a) Si  $n \leq 25$  donde n es el número de pares de sujetos cuya diferencia es distinta de cero, se observa en la tabla de Wilcoxon el valor crítico de T. Si  $T > T$  crítico, aceptamos  $H_0$ . No hay diferencia significativa. Si  $T \leq T$  crítico, rechazamos  $H_0$  (Prueba de una sola cola).

b) Si  $n > 25$  hacemos normalización del valor de T mediante la ecuación:

$$T_{nn}^{+}(0)$$

$$z = 4$$

$$nn + ()12 1+()$$

24

si  $|z| \geq z(\alpha)$  rechazamos  $H_0$  (prueba de dos colas).

### Ejemplo

Podemos utilizar el mismo ejemplo que hemos presentado en el caso del test de signos:

Sean 12 adolescentes elegidos al azar entre miembros de un club de cine-forum. Interesa comprobar si después de la proyección de una película violenta, los adolescentes muestran una mayor agresividad que perdure incluso varios días. Realizados test, antes de la película y tres días después, resultan los siguientes datos:

#### Adolescentes Antes (A) Después (B)

1	14	19
2	16	18
3	23	22
4	26	27
5	24	28
6	28	35
7	27	30
8	18	17
9	15	17
10	22	28
11	20	30
12	25	24

Para resolver el ejercicio planteamos la siguiente tabla:

#### Adolescentes Antes (A) Después (B) Diferencia B-A Rango de diferen. $T_p$ $T_n$

1	14	19	5	9	9
2	16	18	2	5,5	5,5
3	23	22	-1	2,5	2,5
4	26	27	1	2,5	2,5

5 24 28 4 8 8  
 6 28 35 7 11 11  
 7 27 30 3 7 7  
 8 18 17 -1 2,5 2,5  
 9 15 17 2 5,5 5,5 10 22 28 6 10 10 11 20 30 10 12 12 12 25 24 -1 2,5 2,5

La suma de las  $T_p$  es 70,5 y las  $T_n=7,5$  luego se toma  $T=T_n=7,5$ . Mirando en las tablas de Wilcoxon para  $N=12$  a un nivel de  $\alpha=0,01$  tenemos que  $T(\alpha)=10$  luego como  $7,5 < 10$  rechazamos la hipótesis de igualdad, en consecuencia los adolescentes después de la película son más agresivos que antes.

La prueba T de Wilcoxon, como se ve, mejora los resultados de la prueba de los signos apreciando la diferencia significativa donde antes no había sido detectada. Si hubiéramos supuesto para T la aproximación a la normal tendríamos como valor de z:

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{7,5 - \frac{12(12+1)}{4}}{\sqrt{\frac{12(12+1)(24+1)}{24}}} = -2,47$$

12 13 25  
 24 24

$P(z < -2,47) = 0,0068$ , valor que se aproxima al que pudiéramos obtener con la tabla de Wilcoxon para  $P(T > 7,5)$ , luego por los dos procedimientos debemos rechazar la hipótesis al nivel de  $\alpha=0,01$

#### 2.4. Pruebas bimuestrales (muestras independientes)

Frecuentemente, la naturaleza de la variable dependiente impide usar a los sujetos como su propio control, como sucede cuando la variable dependiente es el tiempo empleado para resolver un problema poco familiar, pues esto solo sucede una vez.

Cuando el uso de dos muestras relacionadas no es práctico ni adecuado, pueden usarse muestras independientes. En este diseño, las dos muestras pueden obtenerse con la ayuda de dos métodos:

- a) Tomando al azar de dos poblaciones.
- b) Asignando al azar ambos tratamientos a miembros de alguna muestra de orígenes arbitrarios.

En cualquier caso no es necesario que las dos muestras tengan el mismo tamaño. Es importante que los datos de las dos muestras a

contrastar sean independientes.

Un ejemplo del método de asignación al azar puede presentarse en un estudio de la efectividad docente de dos instructores en la enseñanza del mismo curso, al asignar al azar la mitad de los estudiantes a un instructor y la otra mitad al otro.

Las técnicas paramétricas usuales para analizar datos de dos muestras independientes consisten en aplicar una prueba t a las medias de los dos grupos. La prueba t supone que los puntajes (que se suman al calcular las medias) son observaciones independientes de poblaciones distribuidas normalmente con varianzas iguales. Debido a que usar las medias y otros estadísticos obtenidos por operaciones aritméticas, requiere que las observaciones se midan por lo menos en una escala de intervalo.

En todas las pruebas que vamos a examinar en este apartado queremos ver si las dos muestras proceden de la misma población. Sirven para examinar la significación de la diferencia entre dos muestras independientes.

— Cuando se desea saber si las dos muestras representan poblaciones que difieren en su media  $\mu$  Test U de Mann-Witney.

— Cuando se desea contrastar si las dos muestras son de poblaciones que difieren en cualquier aspecto: media, varianza,.....  $\chi^2$ , Kolmog., W-W). Emplearemos  $\chi^2$  cuando los datos aparezcan clasificados por grupos. El test de Kolmogorov o el de W-W para datos continuos, bien entendido que la prueba de Kolmogorov es la más potente.

#### **2.4.1. La prueba de la probabilidad exacta de Fisher**

##### **Objetivo**

Esta prueba es útil para analizar datos discretos (nominales u ordinales) cuando las dos muestras independientes son pequeñas. Requiere que los datos estén clasificados dicotómicamente, es decir, que los datos de cada muestra pertenezcan a una o a otra de las categorías de la variable analizada; por ejemplo acuerdo-desacuerdo, si-no, positivo-negativo, etc. Los puntajes se representarán en una tabla 2x2.

##### **Hipótesis**

La hipótesis nula plantea que los grupos no difieren en la proporción correspondiente a las clasificaciones (independencia  $n_{ij} = n_i \cdot n_j / N$ ).

##### **Estadístico**

La probabilidad exacta de observar un conjunto particular de frecuencias en una tabla 2x2, cuando los totales marginales se consideran fijos, está dada por la distribución hipergeométrica

$$\frac{\binom{A}{a} \binom{B}{b} \binom{C}{c} \binom{D}{d}}{\binom{A+B}{a+b} \binom{C+D}{c+d}} \text{ que desarrollando queda como: } \frac{\binom{A}{a} \binom{B}{b} \binom{C}{c} \binom{D}{d}}{\binom{A+B}{a+b} \binom{C+D}{c+d}}$$

$$P = \frac{\binom{A}{a} \binom{B}{b} \binom{C}{c} \binom{D}{d}}{\binom{A+B}{a+b} \binom{C+D}{c+d}}$$

$N \ A \ B \ C \ D$  ! ! ! ! !

Esta probabilidad determina la ocurrencia de una distribución dada; pero si se quiere conocer la probabilidad de ocurrencia de tal distribución u otra más extrema habrá que calcular las distintas probabilidades para cada forma de la distribución hasta llegar a la más extrema.

**Decisión**

Se rechaza  $H_0$  si  $pp = <_i a$

$\hat{A}$

$i$

**Ejemplo**

Se quiere comprobar si en el ambiente universitario de una Facultad de CC. de la Educación los líderes «carismáticos» se «queman» significativamente más que los líderes «no-carismáticos». Estudiados 14 líderes resultó la siguiente distribución:

«Quemados»	«No quemados»	Lid. «carismáticos»	6	2	8
Lid. «no carismáticos»	1	5	6		
	7	7	14		

La probabilidad exacta de ocurrencia de tal distribución será:

$$P = \frac{\binom{A}{a} \binom{B}{b} \binom{C}{c} \binom{D}{d}}{\binom{A+B}{a+b} \binom{C+D}{c+d}}$$

$N \ A \ B \ C \ D$  ! ! ! ! !

$\hat{A}$

+

0

+

0

!!!

0

8677!!!!

= =fi  
14 6 2 1 5 !!! !, %

Concluiríamos, a nivel de confianza del 5% que tal distribución no puede darse al azar. Ahora bien, es necesario calcular la probabilidad asociada a una distribución todavía más extrema que la dada (se logra aumentando el valor mayor de la tabla en el sentido de la hipótesis alternativa). Esta sería la siguiente,

«**Quemados**» «**No quemados**» Lid. «carismáticos» 7 1 8

Lid. «no carismáticos» 0 6 6

7 7 14

8677

p

= =fi

147610

!!!! !!!!!, %

Otra distribución más extrema que esta última ya no puede darse. Por consiguiente, la probabilidad exacta de ocurrencia de una distribución empírica como la primitiva o más extrema será:

$$p=4,90+0,233=5,133\%$$

Existe una modificación del test de Fisher introducida por Tocher que consiste en sumar tan sólo las probabilidades de las configuraciones que sugieren mayor asociación que la dada ( $p_0$ ) y compararlo con el nivel de significación  $\alpha$ , de tal manera que:

1. Si  $p_1+p_2+\dots+p_n > a$ , aceptamos  $H_0$ .
2. Si  $p_1+p_2+\dots+p_n < a$  y  $p_0+p_1+p_2+\dots+p_n < a$ , rechazamos  $H_0$ .
3. Si  $p_1+p_2+\dots+p_n < a$  y  $p_0+p_1+p_2+\dots+p_n > a$ , Tocher recomienda el cálculo del siguiente cociente:

$T_{pp}$   
 .....

$$T_{pp} = \frac{p_0}{p_0 + p_1 + \dots + p_n}$$

$p_0$

Una vez calculado el valor de  $T$  acudimos a una tabla de números aleatorios y tomamos al azar un número  $n$  comprendido entre 0 y 1. El criterio de aceptación o rechazo de la hipótesis nula  $H_0$  será entonces el siguiente:

— si  $n < T$  rechazamos  $H_0$  — si  $n > T$  aceptamos  $H_0$

#### 2.4.2. Contraste de Kolmogorov

##### Objetivo

Este contraste ya se ha visto para el caso unimuestral, en el caso de dos muestras se aplica de la misma forma pero ahora no examinamos el grado de ajuste entre las frecuencias acumuladas de una distribución teórica específica y la frecuencia acumulada de la función de distribución empírica sino entre las frecuencias acumuladas de las dos muestras.

Si las dos muestras provienen de la misma población cabe esperar que sus distribuciones respectivas de frecuencias acumuladas sean semejantes entre si ya que únicamente presentan diferencias casuales debidas al azar. Si por el contrario las diferencias son grandes, habrá que concluir que no pueden provenir de la misma población.

**Hipótesis** La hipótesis nula se plantea en los siguientes términos:  $H_0: F_A = F_B$  (no hay diferencia entre las funciones de distribución). Frente a la hipótesis alternativa:  $H_1: F_A \neq F_B$  **Estadístico** Para cada una de las dos muestras se construyen los mismos intervalos, pues si los intervalos son distintos entonces no se pueden comparar.

Si denotamos por  $F_{bi}$  la frecuencia acumulada hasta el intervalo  $i$ -ésimo de la muestra B y  $F_{ai}$  la frecuencia acumulada hasta el intervalo  $i$ -

ésimo de la muestra A entonces obtenemos:

$$D_i = \frac{F_{ai} - F_{bi}}{n \cdot m}$$

La prueba de Kolmogorov-Smirnov examina la mayor diferencia  $D_i$  encontrada  $D = \max \{D_i\}$

**Decisión** Si  $n=m=N$  y cuando  $N \geq 40$  utilizaremos la tabla de Kolmogorov para hallar la región de rechazo ( $RC = \{D > cte\}$ ).

Si  $n$  y  $m > 40$  entonces utilizaremos otra tabla de Kolmogorov o bien

calcular el estadístico

$$c = 22 \cdot \sqrt{\frac{1}{nm}}$$

con  $g.l = 2$  no siendo necesario que  $n=m$ ,  $tam_{nm}$

+

bién esta última fórmula la podemos aplicar con muestras pequeñas y si se rechaza la hipótesis nula podemos confiar en la decisión. La región de rechazo para la ji-cuadrado será:  $RC_a = \{c^2 > cte\}$ .

Cuando decidimos hacer el contraste de Kolmogorov hay que tener en cuenta que cuanto más grande hagamos los intervalos más información perdemos, por tanto, lo idóneo es usar una cantidad elevada de intervalos.

Esta prueba es más poderosa que las de  $c^2$  y la mediana.

### Ejemplo

Se realiza el siguiente experimento: tomemos un grupo de niños de 1.º ESO ( $m=10$ ) y los hacemos memorizar una serie de objetos. Pasada una hora les decimos que nos comuniquen todos los objetos que recuerdan. Contamos el número de fallos que cometen. Repetimos la misma operación con ( $n=10$ ) niños de 2.º ESO. Queremos probar que la distribución de los errores (proporción de fallos) es mayor en los de 1.º ESO que en los de 2.º ESO.

Datos:

	1.º ESO	2.º ESO
39,1	35,2	
41,2	39,2	
45,2	40,9	

46,2 38,1  
 48,4 34,4  
 48,7 29,1  
 55 41,8  
 40,6 24,3  
 52,1 32,4  
 47,2 32,6

Sea:  $F_A$ :proporción de fallos de 1.º ESO

$F_B$ :proporción de fallos de 2.º ESO

Las dos muestras por definición son iguales. Propongamos las hipótesis:

$H_0:F_A=F_B$

$H_1:F_A>F_B$

Tomando intervalos de amplitud 3 unidades tenemos:

24-27	28-31	32-35	36-39	40-43	44-47	48-51	52-55	$F_{ai}/m$	0/10	0/10		
0/10	0/10	3/10	5/10	8/10	10/10	$F_{bi}/n$	1/10	2/10	5/10	7/10	10/10	10/10
10/10	10/10	$ F_{ai}/m-F_{bi}/n $	1/10	2/10	5/10	7/10	7/10	5/10	2/10	0		

Calculamos  $D=\max\{D_i\}=7/10=0,7$

para  $\alpha=0,05$   $n=10$   $P\{D>D_\alpha\}=$  nos lleva a la  $R.C_{0,05}=\{D>0,6\}$  como  $0,7>0,6$  rechazamos  $H_0$ .

### 2.4.3. Contraste de $\chi^2$

#### Objetivo

El objetivo de esta prueba es determinar la significación de las diferencias entre dos grupos independientes. Los datos tienen que estar dados en categorías discretas y básicamente el método utilizado es contar los casos de cada grupo en cada categoría y comparar la proporción de casos en las distintas categorías de un grupo con la del otro.

#### Hipótesis

La hipótesis nula se plantea en los siguientes términos los dos conjuntos de características son independientes.

#### Estadístico

Sea  $O_{ij}$  n.º de casos observados en la fila  $i$  y columna  $j$  y  $E_{ij}$  n.º de casos esperados en la fila  $i$  columna  $j$

$r \times k$

$$\hat{A}\hat{A}d=()$$

$$\sum_{i=1}^r \sum_{j=1}^k E_{ij}$$

sigue una  $\chi^2$  con  $(r-1) \cdot (k-1)$  grados de libertad donde:  
 $r$  n.º de filas / n.º de categorías  
 $k$  n.º de columnas / n.º de grupos

OO

Las frecuencias esperadas

E

$E_{ij}$

=

$\frac{r_{i \cdot} \cdot c_{\cdot j}}{N}$

N

Es necesario recordar que la prueba  $\chi^2$  está sometida a una serie de condiciones según se trate de tablas 2x2, es decir, de 1 grado de libertad o de F filas y K columnas (F ó K > 2), es decir, de más de 1 grado de libertad, cuando las frecuencias teóricas son pequeñas:

#### a) Tabla 2 x 2

1. Cuando  $N < 20$  se usa en todos los casos la prueba de la probabilidad exacta de Fisher.
2. Cuando N está entre 20 y 40 la prueba  $\chi^2$  puede usarse siempre que las frecuencias esperadas sean mayor que 5. Si alguna frecuencia es menor que 5 debe usarse la prueba de la probabilidad exacta de Fisher.
3. Cuando  $N > 40$  se usa la  $\chi^2$  corregida por la falta de continuidad.

$$NAB \cdot BC_j \cdot N_j^2$$

c

$$2 = - 2$$

()()()()

#### b) Tabla F x K

1. Puede usarse  $\chi^2$  cuando ninguna celdilla tenga una frecuencia teórica menor que 1.

2. Cuando varias celdillas (el 20% de ellas o más) tenga una frecuencia teórica menor que 5 conviene combinar categorías de modo que la frecuencia teórica para todas o casi todas las celdillas sea mayor que 5. Hay que tener presente que la prueba  $\chi^2$  es insensible al orden, de ahí que haya que tener precaución al agrupar categorías si se necesita sacar conclusiones sobre este aspecto.

### Decisión

La región crítica será:  $\{d \geq c_{2(r-1)(k-1), \alpha}\}$ , así una vez calculado  $d$  vamos a la tabla de  $\chi^2$ : Si  $d \geq c_{2(r-1)(k-1), \alpha}$  se rechaza  $H_0$   
 Si  $d < c_{2(r-1)(k-1), \alpha}$  se acepta  $H_0$

### Ejemplo

Deseamos probar que las personas altas y bajas difieren con respecto a sus cualidades como dirigentes. Tomemos dos muestras la primera entre las personas bajas con  $m=43$  y la segunda entre las altas con  $n=52$ . Los clasificamos en tres categorías: líder, adepto e inclasificable. Obtenemos la siguiente tabla de datos:

<b>Bajo</b>	<b>Alto</b>	Líder	12	32
Adepto	22	14		
Inclasif.	9	6		
<i>Total</i>	43	52		

$H_0$ : proporción de líderes en las personas altas=proporción de líderes en las personas bajas.

$H_1$ : Estos dos grupos {personas altas y personas bajas} difieren en sus cualidades como dirigentes.

Calculemos en primer lugar las frecuencias esperadas:

$$N=43+52=95$$

$$O_{1.} = 12+32=44$$

$$O_{2.} = 22+14=36$$

$$O_{3.} = 9+6=15$$

$$O_{.1} = 12+22+9=43$$

$$O_{.2} = 32+14+6=52$$

luego por ejemplo  $E_{11} = 44 \cdot 43 / 95 = 19,92$  y procediendo de forma similar calculamos las restantes frecuencias esperadas. Entonces:

$$r \cdot k d = ()^{-2}$$

**$ij$  como sigue una con  $(r-1) \cdot (k-1)$  grados de libertad.**

$\hat{A} \hat{A} ij$

$$i=1 \quad j=1 \quad E_{ij}$$

tenemos:  $d=10,67$

La región crítica será R.C. =  $\{d \geq 5,991\}$  como  $10,67 > 5,991$  fi rechazamos

H<sub>0</sub>

### 2.4.3.1. El procedimiento Tablas de Contingencia

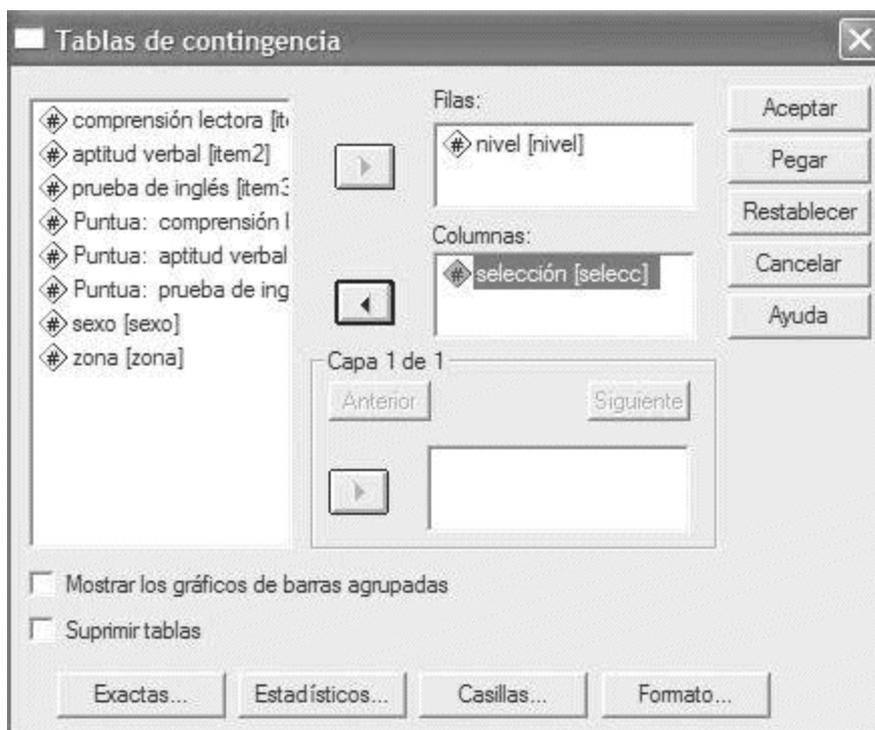
El procedimiento Tablas de Contingencia proporciona tablas de clasificación múltiple, estadísticos y medidas de asociación entre las variables. Los estadísticos y las medidas de asociación sólo se calculan para dos vías de clasificación, si existe un tercer factor o capa, se calcularán aquellos por categoría del tercer factor.

#### Problema-ejemplo

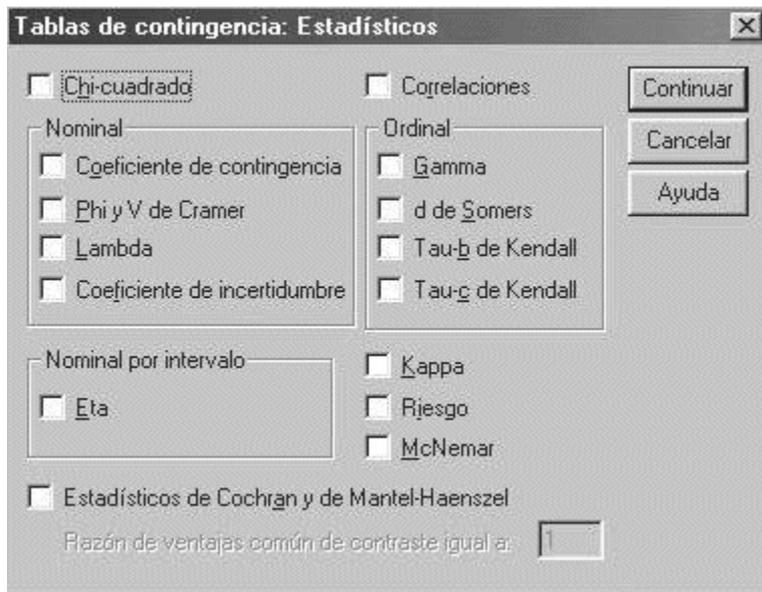
Supongamos que queremos saber si existe relación entre la variable nivel y la variable selección en una prueba, con el fichero de datos que venimos utilizando en el desarrollo de los ejemplos.

#### Desarrollo del ejemplo

Para ejecutar el procedimiento Tablas de Contingencia, seleccionaremos en el menú: **Analizar>Estadísticos descriptivos>Tablas de contingencia.**



En primer lugar seleccionaremos las variables. Hemos elegido, nivel en las filas y selección en las columnas. Estas variables pueden ser numéricas o de cadena. Cuando pulsemos el botón **estadísticos** podremos seleccionar:



— **Chi-cuadrado.** Como medida de relación general. Calcula el estadístico jic cuadrado, el ji-cuadrado de la razón de verosimilitud. La prueba exacta de Fisher para tablas 2x2 y el ji-cuadrado corregido de Yates.

— **Correlaciones.** Da el coeficiente de correlación de Spearman, rho. Sólo debemos utilizarlos en datos numéricos. Cuando las variables sean continuas el coeficiente de correlación que obtenemos es el de Pearson, como medida de asociación lineal.

— Para **variables nominales**, utilizar los siguientes estadísticos: Phi para tablas 2x2 y V de Cramer para el resto de las tablas, coeficiente de contingencia (cc), lambdas simétrica y asimétricas (LAMBDA) y coeficientes de incertidumbre simétrico y asimétricos (uc).

— Para **variables ordinales**, utilizaremos los siguientes estadísticos: Gamma(GAMMA), Taub de Kendall (BTAU), Tauc Kendall (CTAU), D de Somers, versiones simétrica y asimétricas (D) —servirá para predecir las categorías de columnas a partir de las categorías de fila—.

— Para **variables nominal por intervalo**, es decir cuando una variable es categórica y la otra cuantitativa: Eta (ETA).

— **KAPPA.** Coeficiente Kappa de Cohen para tablas con igual números de categorías en filas y en columnas.

— **Riesgo.** Riesgo relativo, para tablas con dos filas y dos columnas.

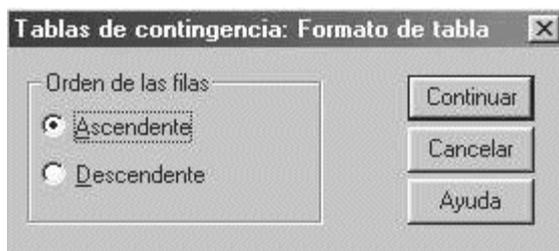
— **McNemar.** Como sabemos es la prueba no-paramétrica para dos muestras relacionadas, ligada a los diseños del tipo antes-después.

En las **casillas** podemos seleccionar:



**Frecuencias** (observadas y esperadas —si hay independencia entre las filas y las columnas—), **porcentajes** (de fila, de columna, total) y los **residuos** —diferencia entre las frecuencias observadas y las esperadas— (no tipificados, tipificados y tipificados corregidos).

En **formato** podemos:



Ordenar de forma ascendente o descendente las categorías de las variables. Independiente de las anteriores facilidades, mediante el comando CROSSTABS podemos realizar las siguientes funciones:

Mediante el subcomando VARIABLES especificar la lista de variables para el análisis, así como, a continuación y entre paréntesis, el mínimo y el máximo valor a considerar en cada una de ellas. Los valores mínimo y máximo deben ser enteros.

Con el subcomando MISSING nos permite controlar el tratamiento de los valores omitidos por el usuario y tiene las siguientes opciones:

**TABLE** En cada tabla se considerarán los casos con valores válidos para todas las variables de la tabla.

**INCLUDE** Los valores omitidos por el usuario serán tratados como válidos. **REPORT** Los valores omitidos por el usuario serán considerados como un valor válido más en la construcción de la tabla, pero no en el cálculo de los estadísticos.

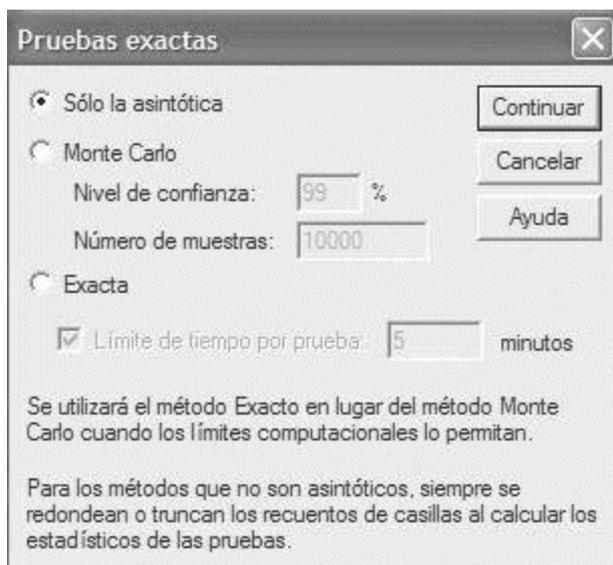
Con el subcomando WRITE podemos escribir las frecuencias observadas en cada celda de las tablas de contingencia en el fichero de resultados especificado en SET (por defecto, en el SPSS.PRC), de tal forma que puedan servir como entrada para otros procedimientos. Tiene las siguientes opciones:

NONE La información relativa a todas las celdas no se escribirá en el fichero de resultados.

ALL La información relativa a todas las celdas se escribirá en el fichero de resultados.

CELLS La información relativa a las celdas no vacías se escribirá en el fichero de resultados.

Además si se dispone del **módulo de pruebas exactas**, se habilitan ciertas opciones para tablas con pocas frecuencias esperas en las casillas, sobre todo es útil para muestras pequeñas.



*Fichero de sintaxis*

**CROSSTABS**

```
/TABLES=nivel BY selecc  
/FORMAT= AVALUE TABLES  
/STATISTIC=CHISQ GAMMA  
/CELLS= COUNT COLUMN.
```

*Ejecución del comando crosstabs*

Tabla de contingencia nivel\* selección.

**Selección**

**Seleccionado No seleccionado Total**

Nivel Primarios sin c. escolaridad Recuento 8 8 16 % de selección 10.5%

10.8% 10.7%

Certificado escolaridad Recuento 8 7 15 % de selección 10.5% 9.5%  
10.0%

Graduado escolar Recuento 11 10 21 % de selección 14.5% 13.5% 14.0%

FP-I Recuento 15 17 32 % de selección 19.7% 23.0% 21.3%

BUP/COU Recuento 10 12 22 % de selección 13.2% 16.2% 14.7%

FP-II Recuento 11 8 10 % de selección 14.5% 10.8% 12.7%

Título de grado medio Recuento 8 6 14 % de selección 10.5% 8.1% 9.3%

Título de grado superior Recuento 5 6 11 % de selección 6.6% 8.1% 7.3%

*Total* Recuento 76 74 150 % de selección 100.0% 100.0% 100.0%

Pruebas de chi-cuadrado.

Valor

gl

Sig. asint. (bilateral)

Chi-cuadrado de Pearson 1.245<sup>a</sup> 7 .990

Razón de verosimilitud 1.248<sup>a</sup> 7 .990

Asociación lineal por lineal .009<sup>a</sup> 1 .924

N de casos válidos 150

Medidas simétricas.

<sup>a</sup> 0. Casillas (.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 5.43.

Ordinal por ordinal

Gamma N de casos válidos Valor

-0.11 150

Error típ. asint.<sup>a</sup> .108

T aproximada<sup>b</sup>

-1.05

Sig. aproximada .917

<sup>a</sup> No asumiendo la hipótesis nula.

<sup>b</sup> Empleando el error típico asintótico basado en la hipótesis nula.

Como vemos en las tablas de las medidas de asociación, es independiente nivel de selección.

#### **2.4.4. Test de la mediana**

##### **Objetivo**

La prueba de la mediana es un test para probar si dos muestras independientes difieren en sus tendencias centrales (medianas). Más exactamente, la prueba de la mediana dará información acerca de la probabilidad de que dos grupos independientes se hayan tomado de

poblaciones con igual mediana.

### Hipótesis

La hipótesis nula supone que provienen de poblaciones con igual mediana; la hipótesis alternativa puede ser unilateral (que la mediana de una muestra sea mayor que la de la otra) o bilateral (que las medianas son distintas).

### Estadístico

Al aplicar la prueba de la mediana, se empieza por determinar el valor de la mediana para las dos muestras combinadas.

El siguiente paso es dicotomizar los conjuntos de valores de la mediana combinada y se distribuyen los datos en una tabla de 2x2.

**Muestra I Muestra II** + A B  
– C D

donde + son los valores por encima de la mediana y – son los valores por debajo o iguales que la mediana.

Ahora bien, si las muestras I y II tenemos muestras procedentes de poblaciones con igual mediana, cerca de la mitad de los valores de cada muestra deberán estar por encima de la mediana combinada y la otra mitad por debajo. Es decir, esperamos que las frecuencias de A y C sean aproximadamente iguales, y las frecuencias B y D también lo sean.

Se cumple que:

$$\hat{E}AC \hat{E}BD$$

ËÁ

$$+ \sim \hat{E}A +$$

$$Pab(, ) = \frac{A B}{n} \text{ donde } n = n_1 + n_2$$

$$\hat{E}A B$$

$n_1$ : número de individuos de la muestra I.

$n_2$ : número de individuos de la muestra II.

Cuando  $n_1 + n_2$  es menor que 20, se usa la prueba de Fisher.

Si el número total de casos es suficientemente grande ( $n_1 + n_2 > 20$ ), puede usarse el contraste de  $\chi^2$  con un grado de libertad, corregida por continuidad, teniendo en cuenta las observaciones que hicimos anteriormente para tablas 2x2.

### Decisión

En cualquier caso si la probabilidad (p) asociada al valor calculado es  $p < \alpha$

se rechaza la hipótesis nula.

### Ejemplo

Sean dos muestras de  $n_1=16$  casos y  $n_2=13$  casos cuyos datos aparecen en la tabla siguiente y que se refieren al grado de aceptación personal que tiene el gerente de una empresa química por parte de las mujeres y por parte de los hombres que trabajan en ella.

Mujeres	Hombres
6	7
10	15
19	26
21	27
20	8
14	30

Mujeres	Hombres
29	26
11	23
12	9
24	25
13	28
17	17
18	15
23	
22	
16	

Calculamos la mediana  $Me=18$  Formamos el siguiente cuadro:

Mujeres	Hombres	Superior a la mediana	7	7	14
Igual o inferior a la mediana	9	6	15		
	16	13	29		

$$c^2 = \frac{NABBC_1}{N_1 N_2} = \frac{2 \cdot 29 \cdot 7 \cdot 7}{9 \cdot 6 \cdot 15 \cdot 16} = 0.00016$$

Como  $c^2 < c_a^2$  entonces se acepta la hipótesis nula. Los dos grupos no son significativamente distintos.

### 2.4.5. Prueba U de Mann-Whitney

#### Objetivo

Esta prueba es, ciertamente, muy útil cuando las medidas de una determinada variable se han tomado de forma ordinal. Incluso cuando los datos se han tomado de forma cuantitativa pero por alguna causa se

sospecha que las condiciones de normalidad, e igualdad de varianzas entre las dos poblaciones de las que provienen las muestras no se cumplen.

### Hipótesis

La hipótesis alternativa bilateral establece en forma sencilla que las distribuciones no son la misma para las dos muestras. Pero la hipótesis alternativa sólo implica un desplazamiento en la tendencia central de una distribución respecto a la otra y no sugiere la diferencia en la forma o en la dispersión.

### Estadístico

La prueba de Mann-Whitney es en realidad un cálculo del «desorden de clasificaciones» entre las dos muestras, es decir, cuántas veces los datos de un grupo están precedidos por los datos de otra muestra.

En la práctica, para obtener  $U_a$  o  $U_b$  es necesario calcular de antemano  $R_a$  y  $R_b$ .  $R_a$  es la suma de los rangos del grupo A.  $R_b$  es la suma de los rangos del grupo B.

$U_a = \frac{R_a(R_a + 1)}{2} - R_a R_b$   
 $U_b = \frac{R_b(R_b + 1)}{2} - R_a R_b$   
**Se cumplirá:  $R_a + R_b = n(n+1)/2$**

$n = n_1 + n_2$

Si las muestras provienen de poblaciones que tienen la misma distribución, se espera que los rangos en cada muestra se encuentren lo suficientemente dispersos, cuando se observa en que orden se encuentran las observaciones. Si las poblaciones tienen distinta distribución, entonces se espera que los rangos estén agrupados.

Con estos valores calculados,  $U_a$  y  $U_b$  vienen dados de la siguiente forma:

$$U_a = \frac{R_a(R_a + 1)}{2} - R_a R_b$$

$n = n_1 + n_2$

$$U_b = \frac{R_b(R_b + 1)}{2} - R_a R_b$$

bb 2

que cumple:  $U_a + U_b = n.m$  donde  $n$  es el tamaño del grupo A y  $m$  el tamaño del grupo b. Para determinar si las dos muestras provienen de la misma población se elige el  $\min(U_a, U_b)$  y la comparamos con la  $U$  de las tablas de Mann-Whitney.

### Decisión

Si  $\min(U_a, U_b) < U$ , siendo  $U$  el valor de las tablas para  $n, m$  a un determinado nivel de confianza entonces rechazamos la hipótesis nula afirmando que existe diferencia significativa entre las dos muestras.

Si  $n$  y  $m$  aumentan de tamaño (mayor que 20) la distribución muestral de  $U$  sigue una Normal con:

$n.m$   
/  
2  
 $m s$   
y  
=

$$.nm \ n \ m + \ +()$$

12

### Ejemplo

Se dispone de dos tipos de cobayas I y II que se las adiestra para recorrer un determinado laberinto. Se nos pide contrastar si las cobayas de tipo I alcanzan la salida con un número medio de errores igual al que da las cobayas de tipo II. Fijemos  $\alpha = 0,05$ . Tomamos una muestra de  $m = 8$  cobayas de tipo I y  $n = 9$  cobayas de tipo II y se quiere contrastar si las dos muestras son idénticas.

**I** 17 12 15 10 14 11 8 13 **II** 13 16 9 11 9 18 6 10 7

Se quiere contrastar la hipótesis  $H_0$ : provienen de la misma población ( $\mu_1 = \mu_2$ ) frente  $H_1$ :  $\mu_1 > \mu_2$  (test de una cola).

**$X_i$   $R_i$**

17 16  
12 10  
15 14  
10 6,5

14 13  
 11 8,5  
 8 3  
 13 11,5

13 11,5 16 15 9 4,5 11 8,5 9 4,5 18 17 6 1 10 6,5 7 2

$$R_A = 16 + 10 + \dots + 12 = 82,5 \quad R_B = 11,5 + 15 + \dots + 2 = 70,5$$

$$U_A = m \cdot n + m(m+1)/2 - R_A = 8 \cdot 9 + 8 \cdot (8+1)/2 - 82,5 = 25,5 \quad U_B = m \cdot n + n(n+1)/2 -$$

$$R_B = 8 \cdot 9 + 9 \cdot (9+1)/2 - 70,5 = 46,5$$

$$U = \min\{U_A, U_B\} = \min\{25,5; 46,5\} = 25,5$$

Como la hipótesis alternativa que queremos contrastar es  $H_1: \mu_1 > \mu_2 \Rightarrow$  test de una sola cola.

Calculamos el valor crítico en la tabla U de Mann-Witney para  $\alpha = 0,05$   $m = 8$   $n = 9$  tenemos  $U_c = 18$  como  $U > U_c$  aceptamos la hipótesis nula.

Si hubiéramos aproximado por la normal tendríamos:

$$E[U] = m \cdot n / 2 = 8 \cdot 9 / 2 = 36$$

$$V[U] =$$

$$nm \cdot n \cdot m + \dots = 10,39 \cdot 8 \cdot 9 = 10,39 \cdot 72 = 748,08$$

12

$$P\{z < (25,5 - 36) / 10,39\} = P\{z < -1,01\} = 0,16$$

como  $\alpha < 0,16$  aceptamos la hipótesis nula.

Cuando existen bastantes pares igualados es conveniente hacer la corrección

$nm$  <sup>33</sup>.

**por pares igualados quedan  $V[U] = \dots$**   
**112 Â Â 12**

siendo t el valor de la frecuencia de cada par igualado.

#### 2.4.6. Test de Wald-Wolfowitz o de rachas Objetivo

Otro método para comparar las distribuciones de dos poblaciones con base en muestras aleatorias independientes es la prueba de tendencias de Wald-Wolfowitz. Considera una distribución continua y requiere que la

medida de la variable sea por lo menos en una escala ordinal.

### Hipótesis

Para esta prueba la hipótesis nula es que las dos muestras aleatorias provienen de poblaciones con distribuciones idénticas, pero a diferencia de la prueba de U de Mann-Whitney, la hipótesis alternativa no es solo una diferencia de medias sino mucho más amplia. En este contraste la hipótesis alternativa es que las distribuciones difieren en algún aspecto como la tendencia central, en la dispersión o la asimetría. O sea la prueba de Wald-Wolfowitz descubre cualquier clase de diferencia.

### Estadístico

En este contraste las observaciones de las dos muestras se ordenan y se calcula el número de rachas que hemos formado: asignando el signo + a aquellos rangos cuya observación pertenece a la primera muestra y el signo - a los rangos cuya observación corresponde a la segunda muestra.

Si hay pares ligados se deben construir las combinaciones posibles de parejas y calcular las rachas (R) para las diferentes posibilidades. Si hay muchos pares ligados la prueba no se puede aplicar.

Si la hipótesis nula es que tienen la misma distribución entonces las observaciones de las dos muestras se deben encontrar bien mezcladas, produciendo de esta forma un gran número de rachas. Pero si las distribuciones difieren en algún aspecto entonces las observaciones aparecerán agrupadas.

Si  $n, m \leq 20$  utilizamos las tablas de rachas.

Para valores  $n, m > 20$  se cumple que:

$$E[R] = \frac{nm}{n+m}$$

$$V[R] = \frac{nm}{n+m} \left( \frac{n-1}{n+m} + \frac{m-1}{n+m} \right)$$

### Decisión

Si  $n, m \leq 20$  la región crítica será:  $RC = \{R < cte\}$

Si  $n, m > 20$  entonces  $z = \frac{R - E[R]}{\sqrt{V[R]}}$  se distribuye según una normal, luego la región crítica será  $RC = \{|z| > cte\}$ .

Para valores próximos a  $n, m > 20$  se efectúa ajuste por distribución continua.

$$() - 0,05 = z \text{ RER VR}()$$

### Ejemplo

Se desea estudiar si hay diferencias sexuales en la cantidad de agresión exhibida por los niños en el juego.

Para ello se observa el comportamiento de 12 niños y 12 niñas durante el desarrollo de una sesión de juego, puntuando el grado de agresión de cada individuo. Los datos obtenidos son los siguientes:

Niños(+)	Niñas(+)
86	55
69	40
72	22
65	58
113	16
65	7
118	9
45	16
141	26
104	36
41	20
50	15

$H_0$ : la agresión es la misma en ambos sexos.

$H_1$ : los niños y las niñas muestran diferencias en el grado de agresión.

Como la hipótesis  $H_1$  es cualquier clase de diferencia entre los dos grupos (y por

supuesto estos son independientes) se escoge la prueba de rachas de Wald-Wolfowitz.

Ordenamos los datos:

7 9 15 16 16 20 22 26 36 40 41 45 50 55 58 65 65 69 72 86 104 113 118 141

----- + + + -- + + + + + + + + +

$r=4$  rachas

La región crítica será  $\{r < cte\}$

Si miramos en la tabla para  $n=m=12$  y  $\alpha=0,05$   $\{r < 7\}$

Como  $4 < 7$  rechazamos  $H_0$  existe diferencia en función del sexo.

### 2.5. Pruebas para k-muestras relacionadas

A veces, las circunstancias requieren que diseñemos un experimento de

modo que más de dos muestras o condiciones puedan estudiarse simultáneamente. Cuando tres o más muestras o condiciones van a ser comparadas en un experimento, es necesario usar una prueba estadística que indique si hay una diferencia *total* entre las  $k$  muestras o condiciones, antes de escoger un par para probar la significación de la diferencia entre ellas.

Si quisiéramos usar una prueba estadística de dos muestras para probar diferencias entre, digamos, 5 grupos, necesitaríamos calcular, para comparar cada par de muestras, 10 pruebas estadísticas  $\hat{\alpha} \approx 10$ . Pero al hacer 10 pruebas estadísticas de la misma hipótesis, se incrementa la probabilidad de error tipo I. Se puede demostrar que el nivel de significación *efectivo* de tal procedimiento llega a ser  $\alpha = 0,40$ .

Solamente, cuando una prueba total —una prueba de  $k$  muestras— nos permite rechazar la hipótesis de nulidad, se justifica el uso subsiguiente de un procedimiento para probar las diferencias significativas entre cualquier par de las muestras  $k$ .

La técnica paramétrica para probar si varias muestras proceden de poblaciones idénticas es el análisis de varianza o prueba F. Las suposiciones asociadas con el modelo estadístico que fundamenta la prueba F son éstas: que los puntajes u observaciones sean sacados independientemente de poblaciones distribuidas normalmente; que las poblaciones tengan todas la misma varianza, y que las medidas en las poblaciones distribuidas normalmente sean combinaciones lineales de «efectos» debidos a renglones y columnas, es decir, que los efectos sean aditivos. Además, la prueba F requiere por lo menos medidas de intervalo de las variables involucradas.

Si encontramos inadecuadas semejantes suposiciones para los datos deberemos utilizar las pruebas no paramétricas.

### **2.5.1. Prueba Q de Cochran**

#### **Objetivo**

Esta prueba se puede considerar como una extensión de la prueba de McNemar. La prueba Q pretende analizar la posible diferencia significativa de los cambios producidos en tres o más situaciones distintas.

Las variables tienen que ser nominales dicotómicas o dicotomizadas (pierde potencia la prueba si se utiliza con variables continuas

dicotomizadas) y los sujetos pueden estar relacionados o se relacionan. Esta prueba se puede utilizar por ejemplo, para saber la dificultad que tienen distintos ítems de un test por medio del análisis de datos compuestos de información «pasa-falla» en k ítems para N individuos. En este diseño, los k grupos se consideran «igualados» porque cada persona contesta todos los k ítems.

Por otra parte, podríamos también analizar solamente un ítem, para comparar las respuestas de los N sujetos en k diferentes condiciones.

### **Hipótesis**

La hipótesis nula se plantea en la forma siguiente: se supone no existe diferencia significativa entre las características que tienen los sujetos en los k-grupos.

### **Estadístico**

Se colocan los resultados de la experimentación en un cuadro con tantas filas como sujetos y tantas columnas como grupos existan.

Se calcula el estadístico Q que sigue una distribución  $\chi^2$  con k-1 g.l. mediante la fórmula:

$$Q = \frac{k \sum_{j=1}^k \hat{A}_j^2 - \frac{(\sum_{j=1}^k \hat{A}_j)^2}{N}}{kL} \quad \text{donde: } N = \sum_{i=1}^k L_i$$

k=número de grupos

A= suma de los valores 1 de cada grupo (suma de columnas)

L= suma de los valores 1 de cada sujeto o grupo de sujetos iguales (suma en las filas)

### **Decisión**

Si  $Q < \chi^2_{\text{crítico}}$  se acepta  $H_0$

Si  $Q \geq \chi^2_{\text{crítico}}$  se rechaza  $H_0$  y por tanto hay diferencia significativa.

### **Ejemplo**

Supongamos 3 grupos de 18 amas de casa de iguales características. Cada uno de estos grupos es sometido a una entrevista con diferente estilo. Queremos saber si las diferencias brutas entre los tres estilos de entrevistas

influyeron en el número de respuestas «si» dadas a un ítem particular por los tres grupos igualados. Los datos aparecen en la tabla siguiente:

Grupos	I	II	III	$L_i$	$L_i^2$
1	0	0	0	0	0
2	1	1	0	2	4
3	0	1	0	1	1
4	0	0	0	0	0
5	1	0	0	1	1
6	1	1	0	2	4
7	1	1	0	2	4
8	0	1	0	1	1
9	1	0	0	1	1
10	0	0	0	0	0
11	1	1	1	3	9
12	1	1	1	3	9
13	1	1	0	2	4
14	1	1	0	2	4
15	1	1	0	2	4
16	1	1	1	3	9
17	1	1	0	2	4
18	1	1	0	2	4
$A_i$	13	13	3	29	63
$A_i^2$	169	169	9		

$$Q = \sum_{j=1}^k \frac{A_j^2}{N} - \frac{(\sum_{i=1}^N L_i)^2}{N} = \frac{13^2 + 13^2 + 3^2}{329} - \frac{(29)^2}{329} = 16,7$$

Si hacemos el cálculo de  $P(Q \geq 16,7)$  da  $p < 0,001$  lo que nos lleva a pensar en que para una significación de  $\alpha = 0,05$  se rechace la hipótesis nula de igualdad entre los tres grupos.

### 2.5.2. Análisis de la varianza de Friedman

#### Objetivo

Es comparar k-muestras relacionadas para ver si provienen de la misma población. Se emplea con variables, por lo menos, en una escala ordinal.

Funciona de forma similar a como lo hace el análisis de la varianza (prueba F de Snedecor) con datos de intervalo continuos. Los tamaños de los grupos son iguales o se pueden igualar.

#### Hipótesis

La hipótesis nula se plantea en la idea de que no existe diferencia

significativa entre las k-muestras relacionadas.

### **Estadístico**

Se establece un cuadro de resultados formado por n filas y k columnas (n sujetos y k grupos).

Se ordena cada fila (es decir, cada sujeto o grupo de sujetos igualados) en rangos, asignando el puesto 1.º al valor más pequeño y el último al mayor. En caso de empates se reparte el puesto o rango medio entre los valores que forman ese empate.

Se calcula la suma de rangos para cada columna o grupo (R).

Se aplica el estadístico:

$$\frac{12^k}{c^2 \hat{A} 1 31) \text{ donde: } r_j() \cdot j=1$$

n=número de filas

k= número de columnas

R= suma de rangos en la columna j

Si la hipótesis de nulidad (que todas las muestras —columnas— proceden de la misma población) es en efecto verdadera, la distribución de los rangos de cada columna será obra del azar y los rangos en los diferentes grupos 1, 2, 3, ..., k deberán aparecer en todas las columnas con frecuencia casi igual.

### **Decisión**

Si  $k=3$  y  $1 < n < 10$  o  $k=4$  y  $1 < n < 5$

Las tablas de Friedman nos proporcionan un valor p para un nivel de significación. Entonces:

Si  $p \leq \alpha$  se rechaza  $H_0$ . Hay diferencia significativa entre los grupos. Si  $p > \alpha$  se acepta  $H_0$ .

Si no se cumple las condiciones anteriores entonces se calcula  $c^2$  (ji-cuadrado) crítico según el valor de a y k-1 g.l.

Si  $c_{r2} < c^2$  crítico, aceptamos  $H_0$ .

Si  $c_{r2} \geq c^2$  crítico, rechazamos  $H_0$ .

### **Ejemplo**

Supongamos que 18 conjuntos de ratas formados cada uno por tres ratas de la misma camada son sometidos en su aprendizaje a tres diferentes métodos de motivación. Las puntuaciones que se adjudican a cada rata se obtienen al computar los errores cometidos por cada una de ellas a lo largo del recorrido a realizar y son los siguientes:

### **Grupos I II III**

1 3 5 4  
2 3 4 2  
3 1 4 2  
4 2 5 6  
5 7 1 3  
6 4 5 2  
7 3 2 1  
8 2 8 6  
9 7 3 4  
10 6 3 5  
11 5 6 4

### **Grupos I II III**

12 2 4 1  
13 4 2 1  
14 3 5 2  
15 3 3 1  
16 5 3 2  
17 5 4 2  
18 3 4 1

Se trata, de contrastar la hipótesis de igualdad de los tres grupos. Evidentemente, son tres muestras compuestas por 18 sujetos y relacionadas. Primero ordenamos por separado cada uno de los 18 grupos de ratas según los

errores cometidos en los tres tipos de motivación, calculando los rangos por grupo de ratas. Los resultados aparecen en la tabla siguiente:

### **Grupos I II III**

1 1 3 2  
2 2 3 1  
3 1 3 2  
4 1 2 3  
5 3 1 2  
6 2 3 1  
7 3 2 1  
8 1 3 2  
9 3 1 2  
10 3 1 2  
11 2 3 1

12 2 3 1  
 13 3 2 1  
 14 2 3 1  
 15 2,5 2,5 1  
 16 3 2 1  
 17 3 2 1  
 18 2 3 1  
 39,5 42,5 26

Si hacemos la suma de rangos por tipo de motivación los resultados permiten calcular la suma de los  $R_j$ . Veamos los resultados:

$$c = 1222 + \dots$$

El valor de  $c^2$  para  $k-1=3-1=2$  g.l y un nivel de significación del 5% es 5,99. Luego como  $8,56 > 5,99$  entonces rechazamos la hipótesis nula y se admite que las puntuaciones de las ratas, es decir, los errores que cometen dependen del tipo de motivación a que son sometidas en el aprendizaje.

## 2.6. Pruebas para k-muestras independientes

Vamos a estudiar la significación de diferencias entre tres o más grupos o muestras independientes. Se trata de técnicas para probar la hipótesis de nulidad de que  $k$  muestras independientes se recogieron de la misma población o de  $k$  poblaciones idénticas.

La técnica paramétrica usual para probar si varias muestras independientes proceden de la misma población es el análisis de la varianza de una clasificación o prueba F. Las suposiciones asociadas con el modelo estadístico en que se basa la prueba F piden observaciones independientes tomadas de poblaciones distribuidas normalmente, todas las cuales tienen la misma varianza. El requisito de medida de la prueba F es, por lo menos, una medida de intervalo de la variable estudiada.

Cuando estas suposiciones no se cumplen, o necesitamos una generalización de nuestros resultados, tenemos las pruebas no-paramétricas para ayudarnos en tal labor.

### 2.6.1. Extensión de la prueba de la mediana

#### Objetivo

La extensión de la prueba de la mediana determina si  $k$  grupos independientes (no necesariamente de igual tamaño) han sido recogidos de la misma población o de poblaciones con mediana iguales. Es útil cuando

la variable en estudio ha sido medida por lo menos en una escala ordinal.

### **Hipótesis**

La hipótesis nula supone que provienen de poblaciones con igual mediana; la hipótesis alternativa puede ser unilateral (que la mediana de una muestra sea mayor que la de las otras) o bilateral (que las medianas son distintas).

### **Estadístico**

Los pasos necesarios para usar esta prueba son los siguientes:

1. Se determina la mediana común de los puntajes de los k grupos.

2. Se convierten en signos positivos todos los puntajes que estén por encima de la mediana y en signo menos todos los que sean menores o iguales que la mediana. Se colocan las frecuencias en una tabla de  $2 \times K$ .

### **Muestra I Muestra II Muestra J Muestra K**

+ $O_{11} O_{12} \dots O_{1j} \dots O_{1k}$

-  $O_{21} O_{22} \dots O_{2j} \dots O_{2k}$

3. Con los datos de esta tabla se calcula el estadístico: Sea  $O_{ij}$  n.º de casos observados en la fila i y columna j y  $E_{ij}$  n.º de casos esperados en la fila i columna j  
donde  $i=1,2$

# $d=0$ sigue una $\chi^2$ con $(2-1) \cdot (k-1)$ grados de libertad $\sum_{i=1}^k \sum_{j=1}^2 E_{ij}$

k n.º de columnas n.º de grupos

OO

Las frecuencias esperadas

E

ij

=

ij

N

## Decisión

En cualquier caso si la probabilidad (p) asociada al valor calculado es  $p < \alpha$  se rechaza la hipótesis nula.

## Ejemplo

Supongamos que un investigador de la educación desea estudiar la influencia que el nivel de educación tiene en el grado de interés de las madres en la instrucción escolar de sus hijos. Para ello, toma el grado escolar más alto que cada madre alcanzó como índice de su nivel de educación. Además como índice del grado de interés en la instrucción de su hijo toma el número de visitas voluntarias que cada madre hace a la escuela durante un año escolar: a representaciones de la clase, a reuniones de padres, a entrevistas de iniciativa propia con maestros y administradores, etc. Sacando cada décimo nombre de la lista de nombres de los 440 niños inscritos en la escuela, obtiene los nombres de 44 madres, quienes constituyen su muestra. Su hipótesis es que el número de visitas de las madres variará de acuerdo con el número de años de instrucción que las madres completaron en la escuela.

**Escuela primaria 10.º Secundaria Estudios (8.º grado)**  
**grado**

**12.º grado**

**universitarios incompletos Con títulos universitarios Estudios de post-graduados**

2 4 2 9

0 3 4 4

4 0 1 2  
 3 7 6 3  
 8 1 3  
 0 2 0  
 5 0 2  
 2 3 5  
 1 5 1  
 7 1 2  
 1 6  
 5  
 1

Calculamos la mediana  $Me=2,5$  Formamos el siguiente cuadro:

2  
 4  
 5  
 2  
 2 6

Escuela **10.º Secundaria** Estudios **primaria** grado 12.º grado **universitarios**

**(8.º grado) incompletos Con títulos universitarios Estudios**

**de postTotal graduados**

Superior a la mediana 5 4 7 3 2 1 22

Igual o inferior a la mediana

5 7 6 1 2 1 22

10 11 13 4 4 2 44

$$c^2 = 0 + 0,409 + 0,0385 + 0,2 + 0 + 0,409 + 0,0385 + 0,2 = 1,295 \quad \text{Como } c^2 < c^2_a$$

entonces se acepta la hipótesis nula. El número de visitas escolares hechas por las madres es independiente de su nivel educativo.

### **2.6.2. Análisis de la varianza de Kruskal- Wallis Objetivo**

Esta prueba pretende ver si existe diferencia entre varias muestras que provienen de la misma población. Se exige, como mínimo, una medición ordinal de la variable. La prueba supone que la variable en estudio tiene como base una distribución continua.

### **Hipótesis**

La hipótesis nula se plantea en los términos de que no existe diferencia significativa entre k-muestras independientes, es decir provienen de la misma población.

## Estadístico

Se ordena todos los valores de los  $k$  grupos en una única serie desde 1.º hasta en  $N$ , independientemente del grupo al que pertenezcan esos valores, asignando el rango 1.º al valor más pequeño y respetando, aunque se ordenen en conjunto, la pertenencia de cada valor a su grupo correspondiente. Los empates se resuelven asignando un rango medio a todos los valores empatados.

- Se calcula  $R$  que es la suma de rangos de cada uno de los  $k$  grupos.
- Se calcula el valor de  $H$  mediante la ecuación:

$$H = \frac{12k R^2}{N(N+1)} - \frac{\sum_{j=1}^k n_j^3}{N^3}$$

donde  $N$  es el número total de sujetos y  $n_j$  el número de sujetos de cada grupo. Si hay más del 25% de empates conviene introducir una corrección en la fórmula de  $H$ .

$$H_c = \frac{H}{1 - \frac{\sum T}{N^3}}$$

donde  $T = \sum t^3$  y  $t$  es el número de sujetos empatados en cada grupo de puntuaciones repetidas. Es decir, habrá tantos valores  $T$  como distintas puntuaciones empatadas haya.

## Decisión

Si  $k=3$  y  $n_1, n_2$  y  $n_3 \geq 5$  se consulta la tabla de Kruskal-Wallis y se toma el valor  $p$  asociado al valor  $H$  calculado.

Si  $p \leq \alpha$  rechazamos  $H_0$ .

Si  $p > \alpha$  aceptamos  $H_0$ .

En cualquier otro caso se calcula el valor de  $c^2$  crítico para  $k-1$  g.l. Si  $H \geq c^2$  crítico, rechazamos  $H_0$ .

## Ejemplo

Supongamos que un profesor de EGB quiere comprobar si tres métodos distintos de dirección de grupo originan diferentes rendimientos

escolares entre los alumnos. Para ello, elige al azar tres grupos de 9, 8 y 7 alumnos a los cuales los somete respectivamente a la dirección de grupo «autoritario», «directivo-democrática» y «dejar-hacer». Los resultados escolares son los siguientes para cada grupo:

**Autoritaria Directivo-democrática Dejar-hacer**

33 38 22  
 15 50 36  
 17 39 37  
 19 23 21  
 26 35 16  
 32 41 20  
 18 40 25  
 28 47  
 29

Planteamos la hipótesis nula de que no hay diferencia significativa entre los tres grupos.

*Resolución:*

Ordenando conjuntamente todos los datos y calculando los rangos tenemos:

**Autoritaria Directivo-democrática Dejar-hacer** 15 19 8 1 24 17 3 20  
 18 5 9 7

11 16 2 14 22 6 4 21 10 12 23

13

$R_1=78$   $R_2=154$   $R_3=68$

Si no existiera diferencia significativa entre las muestras, es decir, si no fueran significativamente distintos los rendimientos escolares el rango promedio de las muestras tendería a coincidir o, al menos, sus diferencias serían explicables por efecto del azar. Si los rangos promedio, por el contrario, se diferencian bastante, quiere decir que las muestras son significativamente diferentes o que no provienen por elección al azar de una misma población.

El valor de H será:

$12^k R^2$  12 78 154 22 268

H NN

+

()

$\hat{A}^j - () () [] - ()$ , <sup>11 02</sup> 1 31

$j=1 \dots k$

Una  $H=11,02 > 9,21$  que es el  $\chi^2$  al nivel de confianza del 1% para 3-1 grados de libertad. Por consiguiente, existe diferencia significativa entre las tres muestras, o lo que es lo mismo, los tres modos de dirección de grupo originan rendimientos escolares significativamente distintos.

La prueba U de Mann-Whitney es la apropiada para comprobar posteriormente si tal diferencia es producida por los resultados de un solo grupo o de varios.

## **2.7. Correlación no paramétrica Introducción**

En muchas investigaciones necesitamos conocer si dos conjuntos de puntajes están relacionados o conocer el grado de su relación.

En algunos casos el fin último de la investigación es detectar la existencia de esa relación para probar la confiabilidad de nuestras observaciones.

Hay que hacer distinción entre el problema de la **existencia** de una correlación y el **grado** de asociación entre los conjuntos de puntajes. El coeficiente de correlación es un indicador del grado de asociación en tanto que la significación nos indica que las variables en estudio posiblemente estén relacionadas en la población.

En el caso paramétrico la medida de asociación más utilizada es el coeficiente de correlación r momento-producto de Pearson, que exige una escala de intervalos en las variables y los puntajes deben provenir de una población normal bivariada.

Las medidas no paramétricas que vamos a estudiar exigen en alguno de los casos que las variables tengan una base continua.

### **2.7.1. El coeficiente de contingencia: C Objetivo**

El coeficiente de contingencia es una medida del grado de asociación o relación entre dos conjuntos de atributos. Es especialmente útil cuando la información acerca de los atributos consiste en una serie no ordenada de frecuencias. No necesita la existencia de una base continua de las diferentes categorías usadas para medir uno o ambos conjuntos de atributos. El coeficiente de contingencia, calculado a partir de una tabla de contingencia, tendrá el mismo valor al margen del orden de las categorías en las filas y en las columnas.

### **Hipótesis**

La hipótesis nula se plantea en la forma siguiente: se supone no existe correlación en la población.

### **Estadístico**

Se colocan los datos formándose una tabla de contingencia de  $r$ (filas) y  $k$ (columnas).

$$C = \frac{\sum_{i=1}^r \sum_{j=1}^k E_{ij}^2}{N}$$

$$C =$$

$$c$$

$$N$$

$$+$$

$$c$$

**2 donde  $\hat{A}_{ij} - ij$  sigue una  $\chi^2$  con  $(r-1).(k-1)$  grados de libertad.**

$$i=1 \quad j=1 \quad E_{ij}$$

$r$  n.º de filas  $k$  n.º de categorías.  $k$  n.º de columnas  $j$  n.º de grupos.

$\sum_{i=1}^r \sum_{j=1}^k$

Y las frecuencias esperadas

$$E_{ij}$$

$$ij$$

$$=$$

$$\dots ij$$

$$N$$

### Decisión

Para probar si el valor observado de  $C$  indica que hay una asociación entre las dos variables de la población muestreada, se determina la probabilidad asociada conforme  $H_0$  de un valor tan grande como la  $\chi^2$  observada con  $(r-1).(k-1)$  en la tabla. Si la probabilidad es igual o menor que  $\alpha$ , se rechaza  $H_0$  y se acepta  $H_1$ .

### Limitaciones

1. El coeficiente de contingencia no vale 1 si la correlación es perfecta, aunque si es igual a cero cuando es nula.

2. El valor máximo ( $C_{\text{máx}}$ ) de correlación es función del número de categorías. Si  $k=r$  tenemos que  $\sqrt{k-1}$  es el valor que puede tener. De ahí que sólo se-  $k$

rán comparables dos coeficientes de correlación si proceden de tablas de igual tamaño.

3. Se necesita que las tablas de contingencias tengan unas frecuencias

esperadas en más del 80% de las celdas, mayor que 5, y ninguna menor que 1. 4. No es directamente comparable con ninguna medida de correlación, por ejemplo, la  $r$  de Pearson, la  $r_s$  de Spearman, o la  $r$  de Kendall.

### Observaciones y otros estadísticos

Para tablas cuadradas se puede calcular  $C_a = C/C_{máx}$  que varia en todos los casos entre 0 y 1.

2

También se puede calcular la  $V$  de Cramer que se define como  $V = \frac{c}{\sqrt{Nm}}$  donde:

$m = \min(r-1, c-1)$  que alcanza el valor 1 si la asociación es perfecta, independiente de la dimensión de la tabla.

### Ejemplo

Tenemos una tabla de contingencia que recoge la asociación entre la posición social (frecuencia de una serie ordenada) y los cursos de capacitación (frecuencia de una serie no ordenada) a la que pertenecen un grupo de jóvenes de un Colegio Privado. Los datos aparecen en la siguiente tabla:

	I	II	III	IV	V	Total
Primaria	23	40	16	2	81	
EGB	11	75	107	14	207	
BUP	1	31	60	10	102	
Total	35	146	183	26	390	

$c$

$^2 69,2,$

Si calculamos el valor de  $C$  tendremos  $C = \frac{c}{\sqrt{N + c^2}} = \frac{69,2}{\sqrt{390 + 69,2^2}} = 0,39$

Como  $P(c^2 \geq 69,2) < 0,001$  nos lleva a rechazar la hipótesis nula y concluir que el status de clase social y la elección de curso están relacionadas.

La  $V$  de Cramer será igual a:

$V =$

$\frac{c^2}{Nm} = \frac{69,2}{390} = 0,29$

$\frac{69,2}{390} = 0,29$

### 2.7.2. El coeficiente de correlación de rangos de Spearman: $r_s$ .

#### Objetivo

Tiene por objeto medir el coeficiente de correlación entre dos variables medidas en una escala ordinal, de manera que los objetos o individuos en estudio puedan colocarse en dos series ordenadas.

#### Fundamento

Sean  $d_i = X_i - Y_i$  la disparidad entre dos conjuntos de rangos. Si la relación entre los rangos fuera perfecta  $d_i = 0$ . Cuando tratamos de observar la discrepancia total las  $d_i$  positivas serían entorpecidas por las  $d_i$  negativas de ahí que se tome  $d_i^2$ .

–

Supongamos  $x = X -$

$X$

e  $y = Y -$

$Y$

– la expresión general para el coeficiente de correlación es:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

22  $r_s$  Cuando  $X$  e  $Y$  son los rangos de las muestras tenemos  $r = r_s$  luego tenemos:

$$r_s = \frac{\sum d^2}{\sum x^2 + \sum y^2}$$

$r_s = 1 -$

$\frac{\sum d^2}{\sum x^2 + \sum y^2}$

6

Como  $d = x - y = (X - X) - (Y - Y) = X - Y$  puesto que  $X = Y$  al tratarse de rangos quedando finalmente la fórmula:

$\sum$

6

$\sum$

$d^2$

2

$\sum$

$r_s = 1 - \frac{\sum d^2}{\sum x^2 + \sum y^2}$

Si existen muchas observaciones ligadas esta fórmula debe ser sustituida por la siguiente:

22 2

$$r_s = \frac{\sum xy + \sum d^2}{\sum x^2 + \sum y^2}$$

3x  $\sum xy + \sum d^2$  /  $\sum x^2 + \sum y^2$  donde  $\sum d^2 = \sum T$  con  $T$

= (t12 xy

12

### Método

Una vez calculados los rangos de las dos muestras se determinan el valor de  $d_i$  para cada uno de los sujetos, substrayendo su rango Y de su rango X. Se eleva al cuadrado este valor para determinar las  $d_i^2$  de cada sujeto. Se suman las  $d_i^2$  de los N casos para determinar  $S_{d_i^2}$ .

Si la proporción de ligas en las observaciones de X o de Y es grande, se usa la fórmula [2] para calcular  $r_s$ . En otros casos, se usa la fórmula [1].

Si los sujetos constituyen una muestra al azar de alguna población, se puede probar si el valor observado de  $r_s$  indica una asociación entre las variables X e Y en la población. El método para hacerlo depende del tamaño de N:

a) Para N de 4 a 10, los valores críticos vienen dados en la tabla de Spearman.

b) Para  $N \geq 10$ , la significación se determina mediante la aproximación a la t de Student por la fórmula:

$t_r$

$N - 2$

$= \frac{t_r}{\sqrt{1 - r_s^2}}$  calculándose el valor crítico con este estadístico,  $t_{\alpha, N-2}$

1

—

r

2

s

### Ejemplo

Supongamos que la dirección de una empresa ha clasificado a ocho gerentes en relación con sus habilidades administrativas, y que a todos ellos se les ha hecho una prueba psicológica para la cual se supone que la calificación está correlacionada con la potencialidad para la administración. Con los datos que presentamos en la tabla siguiente podemos suponer alguna relación entre ambas notas.

### Gerente

### Rango para las Calificaciones habilidades administrativas de la prueba

2 4 72  
 3 2 69  
 4 6 70  
 5 1 93  
 6 3 82  
 7 8 67  
 8 5 80

Las dos variables de interés son el rango y las calificaciones de la prueba. La primera variable ya se encuentra en forma de rangos, y las calificaciones de la prueba se pueden ordenar de manera similar, según aparece en la tabla siguiente:

**Gerente**  
**Rango para las habilidades administrativas**

1 7  
 2 4  
 3 2  
 4 6  
 5 1  
 6 3  
 7 8  
 8 5

**Calificaciones** de la prueba

1 6 36  
 5 -1 1  
 3 -1 1  
 4 2 4  
 8 -7 49  
 7 -4 16  
 2 6 36  
 6 -1 1

*N*  
 6  
 $\hat{A}$   
*d*  
 2  
 $i - ()$   
*r*  
*s*



grupo de ligas de la variable X.

$y_{T=1} \hat{A}tt -()$  y  $t$  es el número de observaciones ligadas en cada grupo de  $2$

ligas de la variable Y.

5. Si los  $N$  sujetos constituyen una muestra aleatoria de alguna población, se

puede examinar el grado en que el valor observado de  $t$  indica la existencia de una asociación entre las variables X e Y de esa población. El método para

obtener tal resultado, depende del tamaño de  $N$ :

a) Para  $N \leq 10$  utilizaremos la tabla de rango de Kendall que contiene la probabilidad asociada —de una cola— de un valor tan grande como el de una  $S$  observada.

b) Para  $N > 10$  aproximaremos a la normal mediante la fórmula:

media =  $mt = 0,22(N)^{0,5}$  desviación estándar =  $(t/91)$

$tm_{tz=st}$

### Decisión

Si la  $p$  obtenida por el método adecuado es igual o menor que  $\alpha$  entonces rechazamos  $H_0$  y confirmamos  $H_1$ .

### Comparación de $t$ y $r_s$

Aún cuando los valores que obtenemos con  $t$  y  $r_s$  son distintos y por tanto no comparables, sin embargo la prueba de significación que hagamos respecto a la población nos debe llevar en ambos casos a resultados similares.

### Ejemplo

Estamos interesados en analizar el grado de asociación entre el nivel de industrialización de los países de la CEE (medido por el porcentaje de empleo en la industria con respecto al total, factor A), y la parte de gastos totales familiares que dedican a transportes y comunicaciones, factor B. Los datos aparecen a continuación:

<b>País</b>	<b>Factor A</b>	<b>Factor B</b>	<b><math>X_i</math></b>	<b><math>Y_i</math></b>
Alemania	40,7	13,9	10	6
Francia	32,1	13,7	5	4
Italia	34,6	13,6	8	3
Holanda	28,7	10,7	3	1
Bélgica	32,7	12,2	6	2
Luxemburgo	32,0	16,5	4	7

Reino Unido 35,0 16,9 9 8

Dinamarca 26,6 17,5 2 9

Grecia 26,0 17,9 1 10

España 33,0 13,8 7 5

Si calculamos el valor de  $t=2.(-7)/10.9=-0,15$  que indica una débil asociación negativa entre los dos factores considerados. Si queremos contrastar la hipótesis nula (no hay asociación) en la población, aún cuando  $N=10$  podemos hacer la aproximación a la normal  $z=-0,625$  que nos lleva a no rechazar la hipótesis de independencia entre factores.

#### **2.7.4. El coeficiente de correlación parcial de rango de Kendall: $t_{xy.z}$**

##### **Función**

Cuando se observa una correlación entre dos variables, siempre hay la posibilidad de que esta correlación sea debida a la asociación entre cada una de las dos variables y una tercera. Por ejemplo, entre un grupo de niños de diversas edades que van a la escuela, es posible encontrar una alta correlación entre la amplitud del vocabulario y la estatura. Esta correlación puede no reflejar una relación genuina o directa entre estas dos variables, sino más bien resulta del hecho de que tanto la extensión del vocabulario como la estatura están asociadas con una tercera variable, la edad.

Estadísticamente, este problema puede ser examinado con los métodos de correlación parcial. En la correlación parcial, los efectos de variación por una tercera variable sobre la relación entre las variables X e Y son eliminados. En otras palabras, la correlación entre X e Y se encuentra al tener la tercera variable, Z como un valor constante.

Como en el apartado anterior los datos de las muestras han de ser medido como mínimo en una escala ordinal.

##### **Método**

1. Sean X e Y las dos variables cuya relación va a determinarse y Z la variable cuyo efecto sobre X e Y va a ser parcializado o mantenido constante. 2. Se ordenan las observaciones en la variable X de 1 a N. Se ordenan las observaciones en las variables Y y Z.

3. Se calculan  $t_{xy}$ ,  $t_{zy}$  y  $t_{xz}$  atendiendo a las correcciones oportunas si existen pares ligados.

4. Se calcula el valor de  $t_{xy.z}$  con los valores anteriores mediante la fórmula:

$t_{xy.z}$

$t_{xz}$

$t_{xy}$

(  
11

—

22 )

$t_{xy.z}$

Se demuestra que

$t$

$t_{xy.z}$

=

$c^2_N$  lo que de alguna forma nos indica que  $t_{xy.z}$  mide el grado en que X e Y concuerdan independientemente de su relación con Z.

### Ejemplo

Supongamos tres variables indicadoras de búsqueda de posición social, autoritarismo y conformidad (condescendencia) de 12 individuos, dadas por sus rangos según muestra la tabla siguiente. Queremos saber si existe relación entre búsqueda de posición social (X) y autoritarismo (Y) independiente de la variable Z (condescendencia).

### Sujeto

**Búsqueda de Autoritarismo Y Conformidad posición social X (condescendencia) Z**

1	3	2	1,5
2	4	6	1,5
3	2	5	3,5
4	1	1	3,5
5	8	10	5
6	11	9	6
7	10	8	7
8	6	3	8
9	7	4	9
10	12	12	10,5
11	5	7	10,5
12	9	11	12

Empleando la fórmula anterior tenemos:

$$t_{xy.z} = \frac{c^2_N}{2} = \frac{0,67036039}{2} = 0,335180195$$

$$t_{xy.z} = \frac{0,335180195}{\sqrt{1 - 0,335180195}} = 0,507076835$$

Como no podemos hacer ninguna prueba de significación para este coeficiente de correlación parcial, sólo nos queda observar si  $t_{xy}$  es aproximadamente igual a  $t_{xy.z}$  para entonces hacer pruebas con el primero e inferior resultados aproximados para el segundo.

La correlación parcial con SPSS se realiza con dos comandos distintos

según el tipo de nivel de medida. En el caso de variables ordinales el cálculo de la  $t$ , de Kendall se efectúa mediante el comando NOPAR CORR y en el caso de variables de escala se realiza mediante PARTIAL CORR. Para seleccionar estos comandos habrá que pasar por los siguientes pasos:

**Analizar fiCorrelacionesfiBivariadas** en el caso de la  $t$ , de Kendall y **AnalizarfiCorrelacionesfiParciales** para el coeficiente de correlación parcial de Pearson.

### 2.7.5. El coeficiente de concordancia de Kendall: W Función

En este punto vamos a considerar una medida de la relación entre varias ordenaciones de  $N$  objetos o individuos.

Cuando tenemos  $k$  ordenaciones, podemos determinar la asociación entre ellas usando el coeficiente de concordancia de Kendall. Mientras que  $r_s$  y  $t$  expresa el grado de asociación entre dos variables medidas o transformadas en rangos,  $W$  expresa el grado de asociación entre  $k$  variables semejantes. Tal medida puede ser particularmente útil en estudios de confiabilidad entre jueces o entre pruebas y también tiene aplicaciones en estudios de agrupamientos de variables.

#### Estadístico

Si queremos comparar  $k$  conjuntos de rangos, parece razonable la determinación de las  $r_s$  (o las  $t$ ) entre todas las parejas posibles de ordenaciones para calcular el promedio de estos coeficientes y determinar la asociación total. Siguiendo tal

procedimiento, necesitaríamos calcular  $\hat{r}_{sav}$  coeficientes de correlación de rango.

A menos que  $k$  fuera pequeño este procedimiento sería largo.

Si hacemos los cálculos mediante  $W$  es mucho más fácil, y  $W$  mantiene una relación lineal con el promedio de  $r_s$  tomado de todos los grupos. Si denotamos el valor promedio de los coeficientes de correlación de rango de Spearman entre los  $\hat{r}_{sav}$  pares posibles de ordenaciones como  $r_{sav}$ , se cumple que:

$$\begin{aligned} r_{sav} &= \frac{kW - 1}{k - 1} \end{aligned}$$

Para calcular  $W$ , coeficiente de concordancia de Kendall es necesario realizar los siguientes pasos:

1. Sea N=número de entidades que van a ordenarse y sea k=el número de jueces que asignarán los rangos. Ordénese los rangos observados en una tabla de kxN.

2. Para cada entidad, se determina  $R_j$ , la suma de los rangos asignados a esa entidad por los k jueces.

3. Obtenemos s, suma de desviaciones al cuadrado, mediante el cálculo de la media de las  $R_j$ , después se hace las diferencias de las  $R_j$  a esta media y se eleva al cuadrado tal diferencia, para finalmente hacer la suma de estas diferencias al cuadrado.

4. Si la proporción de ligas de los k conjuntos de rangos es grande, utilizamos la fórmula:

s

$$W = \frac{1}{N(N-1)} \left( \sum_{j=1}^k R_j^2 - \frac{(\sum_{j=1}^k R_j)^2}{N} \right)$$

$T = \sum_{t=1}^T \hat{A}_{tt}^3$

donde siendo t=número de observaciones en un grupo ligado por un rango dado.

S indica sumar todos los grupos de ligas dentro de cualquiera de las ordenaciones.

El efectuar esta corrección por ligas representa aumentar el valor de W. Si el número de ligas es pequeño o si no existe la fórmula que emplearemos es la siguiente:

$$s^W = \frac{1}{N(N-1)} \left( \sum_{j=1}^k R_j^2 - \frac{(\sum_{j=1}^k R_j)^2}{N} \right)$$

### Significación

Si queremos estudiar si el estadístico W es significativamente distinto de 0 tenemos que atender al tamaño de N:

a) Si N es menor o igual a 7 emplearemos la tabla de Kendall que contiene los valores críticos a los niveles 0,05 y 0,01.

b) Si  $N > 7$  la fórmula:  $c^2 = k(N-1)W$  puede usarse para calcular un valor de  $c^2$  cuya significación, para  $gl = N-1$ , podemos encontrar en la tabla de la ji-cuadrado.

### Interpretación de W

Un alto valor de W, es decir un alto grado de acuerdo acerca de un orden no significa necesariamente que el orden escogido es el «objetivo», ya que esta coincidencia puede tener algún trasfondo no deseado. En este

sentido Kendall nos indica que cuando la  $W$  resulta significativa una estimación «mejor» está asociada con el cuadrado mínimo.

### Ejemplo

Supongamos que a tres ejecutivos se les pide entrevistar a seis solicitantes de empleo, con instrucciones de que han de separar los órdenes de las calificaciones que resulten. Los datos aparecen en la tabla siguiente:

#### Solicitante

**a b c d e f**

Ejecutivo X 1 6 3 2 5 4

Ejecutivo Y 1 5 6 4 2 3

Ejecutivo Z 6 3 2 5 4 1

$R_j$  8 14 11 11 11 8

Los cálculos para obtener  $W$  serán los siguientes:  $s=(8-10,5)^2+(14-10,5)^2+(11-10,5)^2+(11-10,5)^2+(11-10,5)^2+(8-10,5)^2=25,5$

luego  $W = \frac{1}{12} \frac{25,5}{23} = 0,9375$

$s^2 = 25,5^2 = 650,25$  también podemos encontrar  $r_{sav} = 0,16$

$kW = 12 \cdot 0,9375 = 11,25$

$r_{sav} = -0,026$

$k = 12$

La diferencia entre  $r_{sav}$  y  $W$  es que el primero está comprendido entre  $-1$  y  $1$  y el segundo entre  $0$  y  $1$ . Esta relación lineal entre ambos coeficientes nos permite observar los desacuerdos entre los  $k$  jueces.

Si queremos ver la significación de  $W$  como  $N=6$  debemos emplear la tabla de Kendall que nos da un valor crítico para  $s$  de  $103,9$  como  $25,5 < 103,9 \Rightarrow$  que  $H_0$  se acepta. No hay diferencia significativa entre jueces.

También si hubiéramos calculado el valor de ji-cuadrado tendríamos  $\chi^2 = 3(6-1) \cdot 0,16 = 2,4$  que como se distribuye según una ji-cuadrado con  $N-1$  grados de libertad nos da un valor crítico de  $11,07$  luego como  $2,4 < 11,07$  aceptamos  $H_0$ .

## 2.8. Pruebas no paramétricas. El comando NPAR TEST

### El comando NPAR TEST

Permite aplicar diversas pruebas estadísticas para una muestra, así como pruebas no paramétricas para muestras relacionadas o para muestras independientes. Los ejemplos propuestos para la comprensión de las

pruebas, se basan en el fichero de datos que se viene utilizando (fichero examinar.sav).

### 2.8.1. Pruebas para una muestra

#### Prueba de rachas

Contrasta la aleatoriedad de los valores de una muestra. Una racha es una secuencia de valores de la variable. Si una muestra tiene muchas o pocas rachas indican una tendencia y por tanto no será aleatoria.

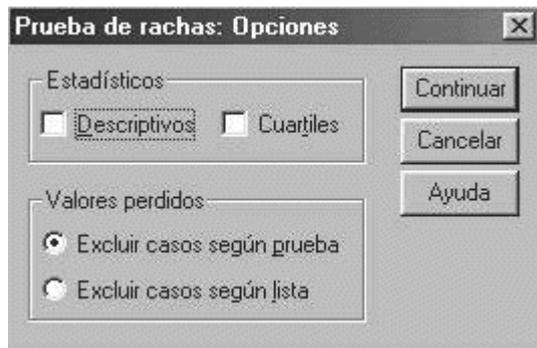
Para realizar la prueba dar: **Analizar>Pruebas no paramétricas>Rachas**



Se aplicará la prueba a cada una de las variables seleccionada.

Se puede personalizar el punto de corte (v), para cada variable, la primera categoría estará formada por los valores menores que v y la segunda por el resto; el punto de corte puede ser la media (MEAN), la mediana (MEDIAN), la moda (MODE).

Si se pulsa el botón de **opciones** podemos seleccionar los estadísticos descriptivos y los cuartiles, así como el tratamiento de los valores omitidos.



Una vez seleccionada la variable, el punto o criterio de corte, los estadísticos y los valores perdidos, tenemos, el siguiente fichero de sintaxis:

```
NPART TESTS  
/RUNS(MEDIAN)=item2  
/MISSING ANALYSIS.
```

Que ejecutado nos lleva a la siguiente salida:  
**PRUEBA DE RACHAS Aptitud verbal**

Valor de prueba<sup>a</sup> 4.0000 Casos < Valor de prueba 63 Casos >= Valor de prueba 87 Casos en total 150 Números de rachas 72 Z  $-.350$  Sig. asintót. (bilateral)  $.726$

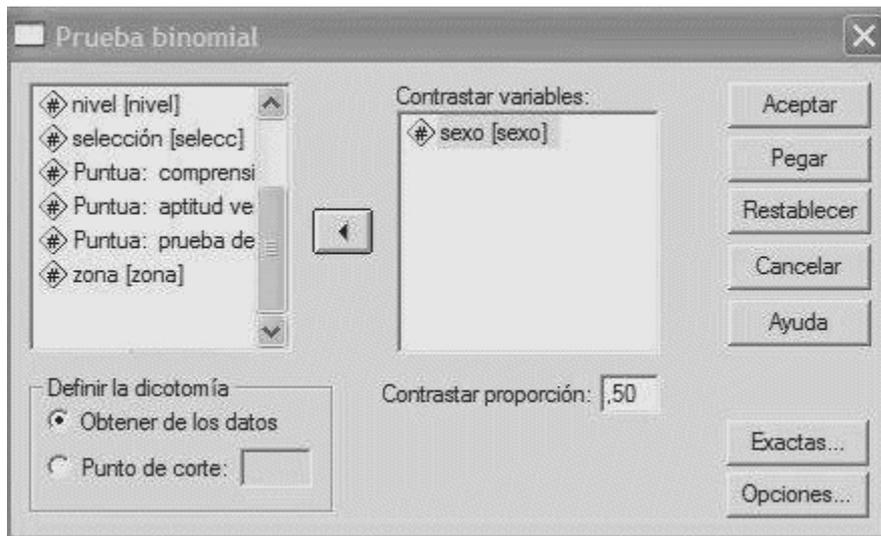
<sup>a</sup> Mediana

Como podemos observar, utilizando como punto de corte la mediana, tenemos 63 por debajo de este valor y 87 iguales o superiores. Vemos que estas diferencias no son estadísticamente significativas y por tanto podemos afirmar que la muestra de valores de la variable item2, es aleatoria.

### **Prueba de la binomial**

La prueba binomial nos sirve para contrastar las frecuencias observadas de una variable dicotómica con las frecuencias esperadas de una distribución binomial de parámetro (p) especificado. Por defecto este parámetro es 0,5.

Para realizar la prueba dar: **Analizar>Pruebas no paramétricas>Binomial.**



Podemos definir la dicotomía: obtener de los datos o fijar un punto de corte y dar el valor de p.

También podemos dar **opciones** y podemos seleccionar los estadísticos descriptivos y los cuartiles, así como el tratamiento de los valores omitidos. Una vez seleccionada la variable, definida la dicotomía, los estadísticos y los valores perdidos, tenemos, el siguiente fichero de sintaxis:

```

NPAR TEST
/BINOMIAL (.50)= sexo
/MISSING ANALYSIS.

```

Que ejecutado nos lleva a la siguiente salida:  
PRUEBA BINOMIAL

### Sexo

Grupo 1	Grupo 2	Total	Categoría	hombre	mujer	Prop. de prueba
N 76	74	150	Proporción observada	.51	.49	1.00
Sig. asintót. (bilateral) .935 <sup>a</sup>						

<sup>a</sup> Mediana

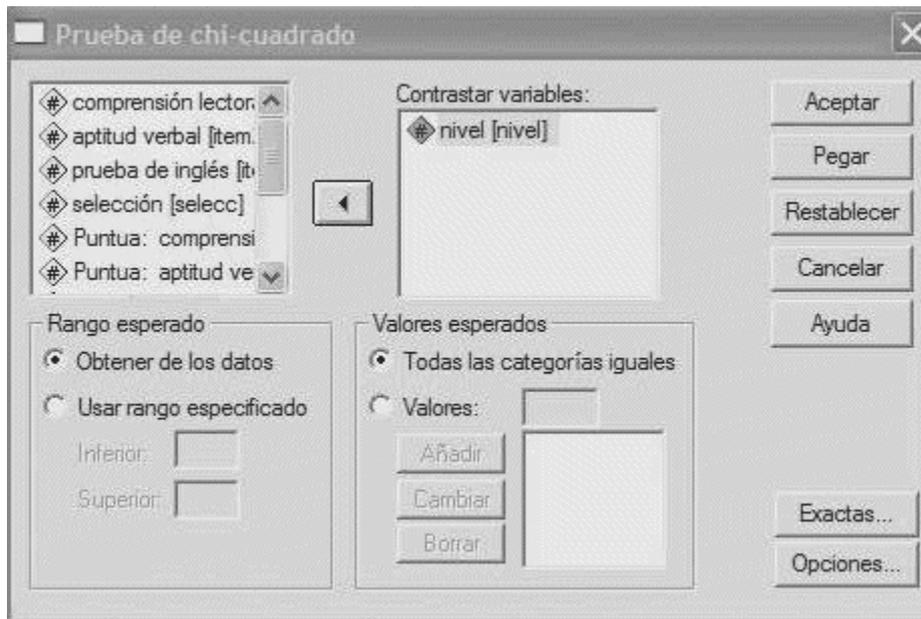
Como podemos observar, en la variable sexo hay una proporción de 0,51 hombres y 0,49 mujeres, que en la hipótesis de  $p=0,5$  vemos que no es significativa.

### Prueba jic cuadrado para una muestra

La prueba ji-cuadrado tabula una variable en categorías y calcula este estadístico. Esta prueba de bondad de ajuste compara las frecuencias observadas y esperadas en cada categoría.

Para realizar la prueba dar: **Analizar>Pruebas no paramétricas>Chi-**

**cuadrado.**



Se puede definir el rango esperado cuando se quiere concretar distintas categorías dentro de un rango específico o se puede obtener estas categorías de los datos. También se pueden especificar los valores esperados o que todas las categorías tengan valores esperados iguales.

Si se pulsa el botón de **opciones** podemos seleccionar los estadísticos descriptivos y los cuartiles, así como el tratamiento de los valores omitidos. Una vez realizadas las operaciones que hemos comentado, se puede **pegar** el siguiente fichero de sintaxis:

```
NPART TEST
/CHISQUARE=nivel
/EXPECTED=EQUAL
/MISSING ANALYSIS.
```

Que ejecutado nos lleva a la siguiente salida:

```
NIVEL N observado N esperado Residual Primarios sin c.
escolaridad 16 18.8 -2.8 Certificado escolaridad
Graduado escolar
fp-I
bup/cou
fp-II
Título de grado medio
Título de grado superior
Total
```

15 18.8 -3.8 21 18.8 2.3 32 18.8 13.3 22 18.8 3.3 19 18.8 .3 14 18.8 -4.8  
11 18.8 -7.8

150

## ESTADÍSTICOS DE CONTRASTE

**Nivel** Chi-cuadrado<sup>a</sup> 15.760 gl 7 Sig. asintót. .027

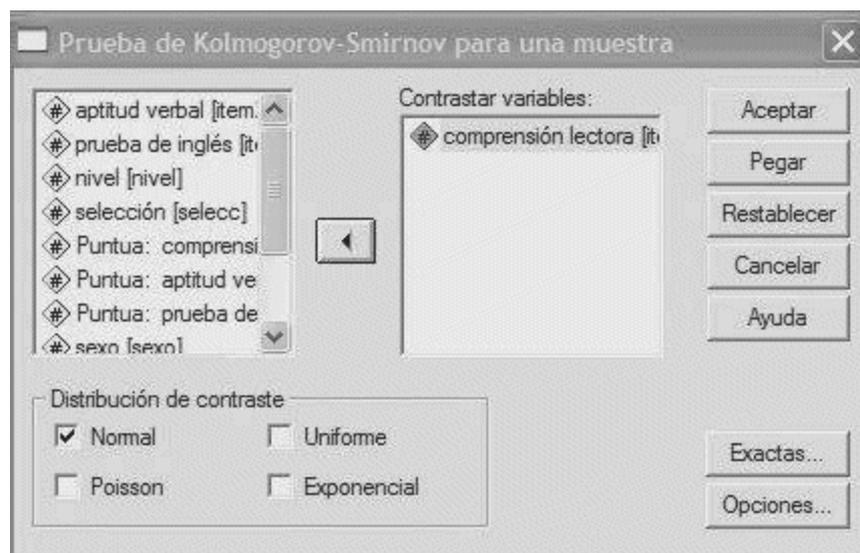
<sup>a</sup> 0 casillas (.0%) tienen frecuencias esperadas menores que 5. La frecuencia de casilla esperada mínimas es 18.8.

Como vemos hay diferencia significativa entre la distribución de frecuencias observada y esperada.

### **Prueba de KolmogorovSmirnov para una muestra**

Está prueba compara la distribución acumulada de frecuencias observadas con

la esperada, que puede ser la normal, uniforme, exponencial o de poisson. Para realizar la prueba dar: **Analizar>Pruebas no paramétricas>Kolmogorov-Smirnov.**



Si se pulsa el botón de **opciones** podemos seleccionar los estadísticos descriptivos y los cuartiles, así como el tratamiento de los valores omitidos. Una vez seleccionada la distribución esperada, se puede **pegar** el siguiente fichero de sintaxis:

NPARTESTS

/K-S(NORMAL)= item1

/MISSING ANALYSIS.

Que tras su ejecución nos lleva a los siguientes resultados:

**PRUEBA DE KOLMOGOROV-SMIRNOV PARA UNA MUESTRA**

N

Parámetros normales<sup>a,b</sup>

Diferencias más extremas

Z de Kolmogorov-Smirnov Sig. asintót. (bilateral)

### Comprensión lectora 150

Media 4.1133 Desviación típica 2.7963

Absoluta .116 Positiva .108 Negativa -.116

1.417 .036

<sup>a</sup> La distribución de contraste es la Normal. <sup>b</sup> Se han calculado a partir de los datos.

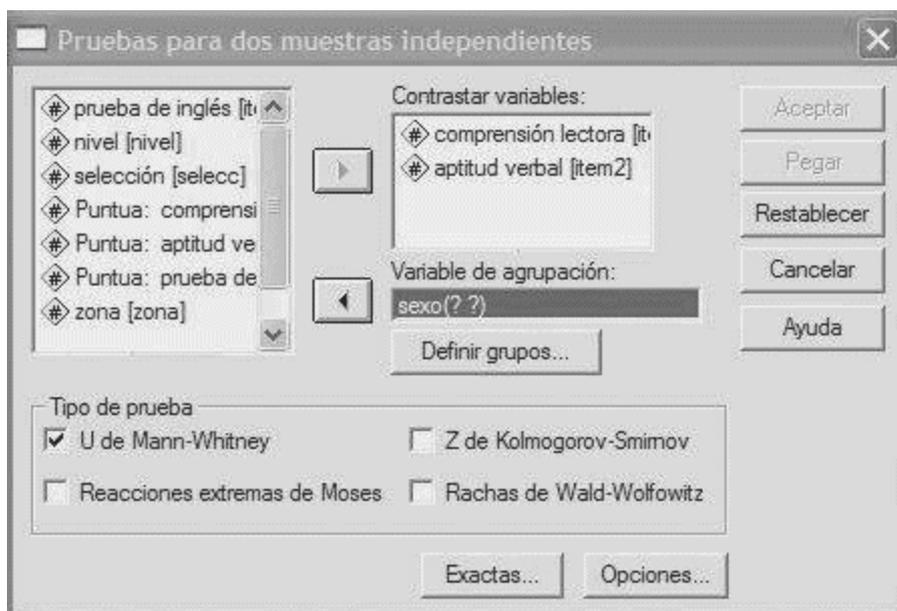
Existe una ligera discrepancia entre los datos empíricos de la distribución del item1 y la distribución normal, como vemos el contraste resulta ligeramente significativo.

### 2.8.2. Pruebas para dos muestras independientes

En general las pruebas que se incluyen en este apartado comparan dos muestras independientes de una variable. Las pruebas que podemos utilizar son: la U de Mann Whitney, Kolmogorov Smirnov, rachas de Wald Wolfowitz y reacciones extremas de Moses.

Para una exposición detallada en los aspectos estadísticos, consúltese el apartado dedicado a las pruebas no-paramétricas en este manual.

En general, las cuatro pruebas necesitan que las muestras sean independientes y aleatorias. Además, la prueba U de Mann-Whitney, quizás la más conocida, exige igualdad de forma para las dos muestras.



Para realizar la prueba dar: **Analizar>Pruebas no paramétricas>2-muestras independientes** .

Necesitamos definir las variables dependientes que queremos estudiar y la variable de agrupación o factor, en nuestro ejemplo hemos seleccionados para las primeras item1 y item2 y entre las segundas, la variable sexo.

Después será necesario seleccionar las dos categorías del factor que vamos a analizar.



Finalmente seleccionaremos la prueba:

— U de Mann-Whitney si queremos contrastar la hipótesis de igualdad de rangos entre las dos muestras.

— La prueba z de Kolmogorov-Smirnov y la prueba de rachas de WaldWolfowitz, si queremos detectar diferencias entre las posiciones y formas de las distribuciones.

— La prueba de reacciones extremas de Moses presupone que el factor afectará a algunos sujetos en una dirección y a otros sujetos en la dirección opuesta.

Como en las pruebas anteriores al pulsar el botón de **opciones** podemos seleccionar los estadísticos descriptivos y los cuartiles, así como el tratamiento de los valores omitidos.

Una vez realizadas las operaciones que hemos detallado, se puede **pegar** el siguiente fichero de sintaxis:

```
NPARTESTS
/M-W= item1 item2 BY sexo(1 2)
/MISSING ANALYSIS.
```

Que ejecutado da los siguientes resultados:

### **Prueba de Mann-Whitney RANGOS**

Sexo	N	Rango promedio	Suma de rangos	Comprensión lectora	Hombre				
76	76.07	5781.00	Mujer	74	74.92	5544.00	Total	150	
Aptitud verbal	Hombre	76	71.84	5459.50	Mujer	74	79.26	5865.50	Total

U de Mann-Whitney W de Wilcoxon

Z

Sig. asintót. (bilateral) ESTADÍSTICOS DE CONTRASTE<sup>a</sup>

### Comprensión lectora Aptitud verbal

2769.000 2533.500

5544.000 5459.500

-.163 -1.053 .871 .292

<sup>a</sup> Variable de agrupación: sexo.

Como vemos la última tabla presenta la U de Mann-Whitney y la W de Wilcoxon de suma de rangos. El resultado para el item1 y item2 en el factor sexo, no resulta significativo.

### 2.8.3. Pruebas para dos muestras relacionadas

Las pruebas para dos muestras relacionadas comparan las distribuciones de dos variables. Las pruebas que podemos utilizar son: la W de Wilcoxon, signos y Mc Nemar.

Para una exposición detallada en los aspectos estadísticos, consúltese el apartado dedicado a las pruebas no-paramétricas en este manual.

Para realizar la prueba dar: **Analizar>Pruebas no paramétricas>2-muestras relacionadas.**



Seleccionaremos la prueba en función del tipo de dato: para datos continuos u ordinales, utilizar la prueba de signos o la de Wilcoxon. Para datos de sujetos recogidos antes-después de la ocurrencia de un evento,

utilizar la prueba de McNemar.

Como en las pruebas anteriores al pulsar el botón de **opciones** podemos seleccionar los estadísticos descriptivos y los cuartiles, así como el tratamiento de los valores omitidos.

Una vez seleccionada la prueba, se puede **pegar** el siguiente fichero de sintaxis:

```
NPART TEST
```

```
/WILCOXON=item1 WITH item2 (PAIRED)
```

Que ejecutado nos da el siguiente resultado:

```
RANGOS
```

```
Aptitud verbal-Comprensión lectora
```

```
Rangos negativos Rangos positivos Empates Total
```

```
N 15a 15b 15c 150
```

```
Rango promedio 67.33 68.64
```

```
Suma de rangos 4444.00 4736.00
```

<sup>a</sup> Aptitud verbal < comprensión lectora. <sup>b</sup> Aptitud verbal > comprensión lectora. <sup>c</sup> Comprensión lectora = aptitud verbal.

```
ESTADÍSTICOS DE CONTRASTEb
```

```
Aptitud verbalComprensión lectora
```

```
Z -.322a
```

```
Sig. asintót. (bilateral) .748
```

<sup>a</sup> Basado en los rangos negativos.

<sup>b</sup> Prueba de los rangos con signo de Wilcoxon.

Por fines didácticos, hemos supuesto que aptitud verbal (pretest) y comprensión lectora(postest), son los resultados de una prueba, realizada con el mismo grupo de sujetos. Como podemos observar no existe diferencia significativa en los resultados.

#### **2.8.4. Pruebas para varias muestras independientes**

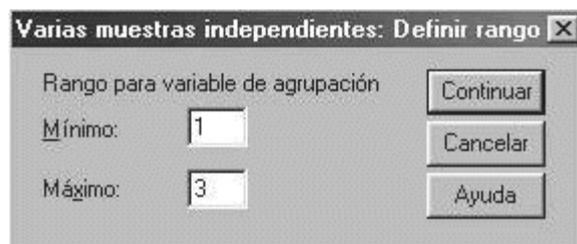
Compara varios grupos de casos en una variable. Las pruebas que utilizaremos son: de la mediana y análisis de la varianza de Kruskal Wallis para k muestras independientes.

Para una exposición detallada en los aspectos estadísticos, consúltese el apartado dedicado a las pruebas no-paramétricas en este manual.

Para realizar la prueba dar: **Analizar>Pruebas no paramétricas>k-muestras independientes.**



Habr  que seleccionar las variables dependientes que queremos contrastar y la independiente, variable de agrupaci3n o factor. En este  ltimo ser  necesario definir el rango, es decir, la categor a m nima y la m xima.



La prueba H de Kruskal-Wallis puede considerarse una extensi3n de la U de Mann-Whitney, es la versi3n no param trica del an lisis de la varianza de un factor.

La prueba de la Mediana es m s general que la anterior, pero menos potente, que la primera, detecta diferencia entre las distribuciones.

Como en otras pruebas al pulsar el bot3n de **opciones** podemos seleccionar los estad sticos descriptivos y los cuartiles, as  como el tratamiento de los valores omitidos.

Una vez seleccionada la prueba, se puede **pegar** el siguiente fichero de sintaxis:

NPARTESTS

/K-W=item1 item2 BY zona(1 3)

Que ejecutado dar  los siguientes resultados:

RANGOS

**Zona N Rango promedio**

Comprensi3n lectora Rural 55 77.12

Semirural 44 77.69 Urbana 51 71.86 Total 150

Aptitud verbal Rural 55 28.29

Semirural 44 77.77 Urbana 51 124.45 Total 150

ESTADÍSTICOS DE CONTRASTE<sup>a, b</sup>

**Comprensión lectora Aptitud verbal**

Chi-cuadrado .552 131.361

gl 2 2

Sig. asintót. (bilateral) .759 .000

<sup>a</sup> Prueba de Kruskal-Wallis. <sup>b</sup> Variable de agrupación: zona.

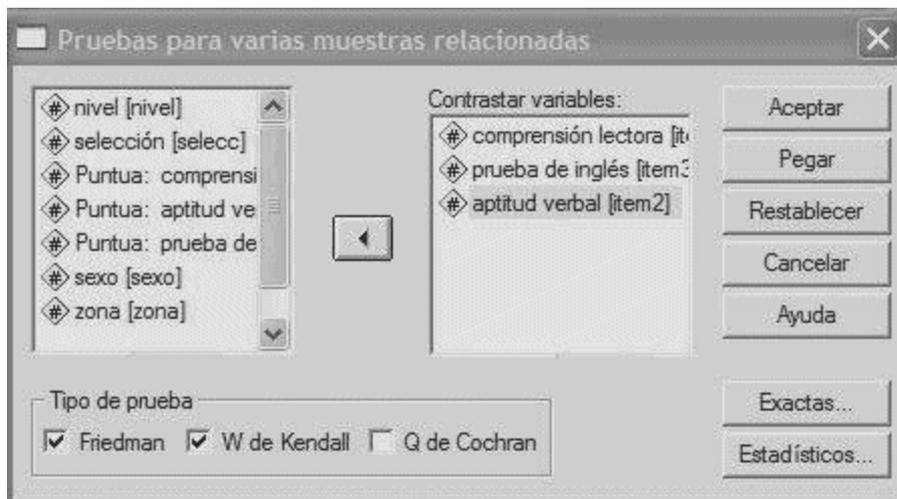
Vemos que existe diferencia significativa en aptitud verbal. Claramente resultan superiores los valores en la zona urbana respecto a las otras zonas.

### **2.8.5. Pruebas para varias muestras relacionadas**

Las pruebas para K muestras relacionadas comparan las distribuciones de dos o más variables. Las pruebas que podemos utilizar son: la Q de Cochran, F de Friedman y la W de Kendall.

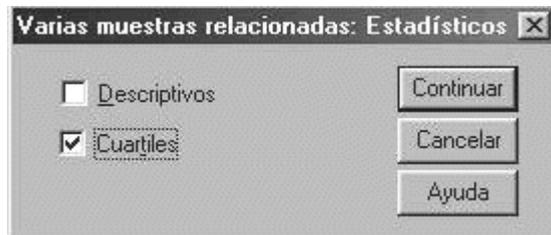
Para una exposición detallada en los aspectos estadísticos, consúltese el apartado dedicado a las pruebas no-paramétricas en este manual.

Para realizar la prueba dar: **Analizar>Pruebas no paramétricas>k-muestras relacionadas.**



Seleccionaremos la prueba **Friedman**, como equivalente no paramétrico de un análisis de la varianza de dos vías con una observación por casilla, en diseños de medidas repetidas. Se prueba si las k variables relacionadas proceden de la misma población. La **W** de **Kendall**, se puede interpretar como el coeficiente de concordancia, es decir, una medida de acuerdo entre jueces; es una normalización de la prueba de Friedman.

La prueba Q de **Cochran** es similar a la prueba de Friedman, pero para variables dicotómicas. Es una extensión de la prueba de McNemar. En estas pruebas se pueden calcular ciertos **estadísticos**:



Una vez seleccionada la prueba, se puede **pegar** el siguiente fichero de sintaxis:

```
NPART TESTS
/FRIEDMAN = item1 item2 item3
/KENDALL = item1 item2 item3
/STATISTICS QUARTILES
/MISSING LISTWISE.
```

Que ejecutado nos da el siguiente resultado:

ESTADÍSTICOS DESCRIPTIVOS

### Percentiles

**N 25 50 (Mediana) 75**

Comprensión lectora 150 2.0000 4.0000 7.0000

Aptitud verbal 150 1.7500 4.0000 7.0000

Prueba de inglés 150 2.0000 5.0000 7.0000

### Prueba de Friedman RANGOS

#### Rango promedio

Comprensión lectora 1.93

Aptitud verbal 2.03

Prueba de inglés 2.04

### ESTADÍSTICOS DE CONTRASTE<sup>a</sup>

N 150

Chi-cuadrado 1.228

gl 2

Sig. asintót. .541

<sup>a</sup> Prueba de Friedman.

### Prueba W de Kendall

#### Rango promedio

Comprensión lectora 1.93

Aptitud verbal 2.03

## Prueba de inglés 2.04 ESTADÍSTICOS DE CONTRASTE

N 150

W de Kendall<sup>a</sup> .004

Chi-cuadrado 1.228

gl 2

Sig. asintót. .541

<sup>a</sup> Coeficiente de concordancia de Kendall.

Para los dos estadísticos, Friedman y Kendall, la diferencia no es significativa.

### EJERCICIOS DE AUTOCOMPROBACIÓN

1. El tiempo medio de permanencia en los estudios de una muestra de 100 médicos es de 7,2 años con una desviación típica de 0,3 años. Si  $m$  representa el tiempo medio de permanencia en los estudios de toda la Facultad, contrastar la hipótesis  $m=6,8$  frente a la alternativa  $m \neq 6,8$ , usando el nivel de significación  $\alpha=0,05$ .

2. Un instituto de enseñanza secundaria sostiene que sus alumnos aprueban la selectividad en el 90% de los casos. En una muestra de 180 alumnos, aprobaron 150. Determinar si la afirmación del instituto es cierta.

3. Un profesor está interesado en analizar la actitud de sus alumnos hacia la Estadística. Por su experiencia personal piensa que, si ofrece explicaciones amplias sobre el uso, la utilidad y la importancia de la materia, conseguirá que estos alumnos mejoren su actitud, aunque no está seguro. Para probar su hipótesis de que la actitud podría cambiar, divide la muestra total ( $n=36$ ) en dos grupos: uno de control ( $n=16$ ) que seguirá con sus clases normales, y otro experimental ( $n=20$ ) donde se ofrecerán esas explicaciones complementarias.

Tras un semestre de prueba en práctica de la experiencia, y basándose en su observación personal, otorga unas puntuaciones en actitud a cada uno de los sujetos de la muestra (en una escala de 0-80). Dada la amplitud de la escala y a la rigurosidad del observador, se supone distribución normal. Al nivel del  $\alpha=0,05$ , ¿existe diferencia en las valoraciones de los dos grupos?

Grupo A: 5- 7- 9- 14- 19- 25- 27- 35- 39- 40- 42- 47- 56- 60- 65- 72.

Grupo B: 5- 7- 9- 14- 19- 25- 27- 35- 39- 40- 42- 47- 56- 60- 65- 72. 75- 78- 80.

4. Una compañía que produce lapiceros escolares, afirma que tiene una varianza de diámetro no mayor de 0,0002 cm. Una muestra aleatoria de 10 lápices dio una varianza muestral de  $s^2=0,0003$ . Si se supone que las medidas del diámetro se distribuyen de forma normal, ¿hay evidencia para

refutar lo que afirma la compañía?  $\alpha=0,05$

5. Se quiere contrastar la hipótesis de que los datos siguientes provienen de una distribución normal.

20, 22, 24, 30, 31, 32, 38.

- Se trata del ejemplo 2: Contraste Kolmogorov-Smirnov de la parte teórica.

6. Se sabe que el 30% de las pequeñas barras de hierro se romperán cuando se las someta a una carga de 3000 Kg. En una muestra aleatoria de 50 barras se encontró que 21 de ellas se rompían cuando se les sometía a esa carga. Investíguese si esta muestra proviene de la misma población.  $\alpha=0,01$ .

- Se trata del ejemplo de la prueba ji-cuadrado de adherencia de ajuste de la parte teórica.

7. Supongamos los resultados de extraer 15 letras A y B según la siguiente secuencia: A,B,B,A,A,B,A,A,A,B,A,B,A,B,B. Queremos saber si la extracción ha sido aleatoria.

- Se trata del ejemplo 1 de la prueba de rachas de la parte teórica.

8. Los 25 alumnos de una clase de 6° de Primaria son sometidos a un entrenamiento en habilidades numéricas obteniéndose los siguientes resultados antes y después de la aplicación de esta metodología de apoyo. Se desea contrastar si el entrenamiento mejora estas habilidades.

Después

– + Antes + 14 4

– 3 4

Donde + indica clasificación mayor que la mediana y – menor que la mediana respectivamente en las puntuaciones de la prueba de habilidades numéricas.

- Es el ejemplo de la prueba de McNemar de la parte teórica.

9. Sean 12 adolescentes elegidos al azar entre miembros de un club de cineforum. Interesa comprobar si después de la proyección de una película violenta, los adolescentes muestran una mayor agresividad que perdure incluso varios días. Realizados test, antes de la película y tres días después, resultan los siguientes datos:

**Adolescentes Antes (A) Después (B) B-A**

1 14 19 +

2 16 18 +

3 23 22 –

4 26 27 +  
 5 24 28 +  
 6 28 35 +  
 7 27 30 +  
 8 18 17 –  
 9 15 17 + 10 22 28 + 11 20 30 + 12 25 24 –

- Es el ejemplo de la prueba de los signos de la parte teórica. 10. Se utiliza el mismo enunciado que el ejercicio anterior, pero en esta ocasión se pide utilizar la prueba de Wilcoxon.

- Es el ejemplo de la prueba de Wilcoxon de la parte teórica.

11. Se quiere comprobar si en el ambiente universitario de una Facultad de CC. de la Educación los líderes «carismáticos» se «queman» significativamente más que los líderes «no-carismáticos». Estudiados 14 líderes resultó la siguiente distribución:

«Quemados» «No quemados» Lid. «carismáticos» 6 2 8  
 Lid. «no carismáticos» 1 5 6  
 7 7 14

- Es el ejemplo de la prueba de Fisher de la parte teórica.

12. Se realiza el siguiente experimento: tomemos un grupo de niños de 1.º ESO (m=10) y los hacemos memorizar una serie de objetos. Pasada una hora les decimos que nos comuniquen todos los objetos que recuerdan. Contamos el número de fallos que cometen. Repetimos la misma operación con (n=10) niños de 2.º ESO. Queremos probar que la distribución de los errores (proporción de fallos) es mayor en los de 1.º ESO que en los de 2.º ESO.

Datos:

1.º ESO	2.º ESO
39,1	35,2
41,2	39,2
45,2	40,9
46,2	38,1
48,4	34,4
48,7	29,1
55	41,8
40,6	24,3
52,1	32,4
47,2	32,6

- Es el ejemplo de la prueba de Kolmogorov de la parte teórica.

13. Deseamos probar que las personas altas y bajas difieren con respecto a sus cualidades como dirigentes. Tomemos dos muestras la primera entre las personas bajas con  $m=43$  y la segunda entre las altas con  $n=52$ . Los clasificamos en tres categorías: líder, adepto e inclasificable. Obtenemos la siguiente tabla de datos:

<b>Bajo</b>	<b>Alto</b>	Líder	12	32
Adepto	22	14		
Inclasif.	9	6		
<i>Total</i>	43	52		

- Es el ejemplo de la prueba de ji-cuadrado de la parte teórica.

14. Sean dos muestras de  $n_1=16$  casos y  $n_2=13$  casos cuyos datos aparecen en la tabla siguiente y que se refieren al grado de aceptación personal que tiene el gerente de una empresa química por parte de las mujeres y por parte de los hombres que trabajan en ella.

	<b>Mujeres</b>	<b>Hombres</b>
6	7	
10	15	
19	26	
21	27	
20	8	
14	30	
29	26	
11	23	
12	9	
24	25	
13	28	
17	17	
18	15	
23		
22		
16		

- Es el ejemplo de la prueba de la mediana de la parte teórica. 15. Se dispone de dos tipos de cobayas I y II que se las adiestra para recorrer un determinado laberinto. Se nos pide contrastar si las cobayas de tipo I alcanzan la salida con un número medio de errores igual al que da las cobayas de tipo II. Fijemos  $\alpha=0,05$ . Tomamos una muestra de  $m=8$  cobayas de tipo I y  $n=9$  cobayas de tipo II y se quiere contrastar si las dos

muestras son idénticas.

**I** 17 12 15 10 14 11 8 13 **II** 13 16 9 11 9 18 6 10 7

Se quiere contrastar la hipótesis  $H_0$ : provienen de la misma población ( $\mu_1 = \mu_2$ ) frente  $H_1$ :  $\mu_1 > \mu_2$  (test de una cola).

- Es el ejemplo de la prueba U de Mann-Whitney de la parte teórica.

16. Se desea estudiar si hay diferencias sexuales en la cantidad de agresión exhibida por los niños en el juego.

Para ello se observa el comportamiento de 12 niños y 12 niñas durante el desarrollo de una sesión de juego, puntuando el grado de agresión de cada individuo.

Los datos obtenidos son los siguientes:

**Niños(+)** **Niñas(+)**

86 55  
69 40  
72 22  
65 58  
113 16  
65 7 118 9  
45 16  
141 26  
104 36  
41 20  
50 15

- Es el ejemplo de la prueba de Wald-Wolfowitz de la parte teórica.

17. Supongamos 3 grupos de 18 amas de casa de iguales características. Cada uno de estos grupos es sometido a una entrevista con diferente estilo. Queremos saber si las diferencias brutas entre los tres estilos de entrevistas influyeron en el número de respuestas «si» dadas a un ítem particular por los tres grupos igualados. Los datos aparecen en la tabla siguiente:

**Grupos I II III**

1 0 0 0  
2 1 1 0  
3 0 1 0  
4 0 0 0  
5 1 0 0  
6 1 1 0

7 1 1 0  
8 0 1 0  
9 1 0 0  
10 0 0 0  
11 1 1 1  
12 1 1 1  
13 1 1 0  
14 1 1 0  
15 1 1 0  
16 1 1 1  
17 1 1 0  
18 1 1 0

- Es el ejemplo de la prueba Q de Cochran de la parte teórica.

18. Supongamos que 18 conjuntos de ratas formados cada uno por tres ratas de la misma camada son sometidos en su aprendizaje a tres diferentes métodos de motivación. Las puntuaciones que se adjudican a cada rata se obtienen al computar los errores cometidos por cada una de ellas a lo largo del recorrido a realizar y son los siguientes:

**Grupos I II III**

1 3 5 4  
2 3 4 2  
3 1 4 2  
4 2 5 6

**Grupos I II III**

5 7 1 3  
6 4 5 2  
7 3 2 1  
8 2 8 6  
9 7 3 4  
10 6 3 5  
11 5 6 4  
12 2 4 1  
13 4 2 1  
14 3 5 2  
15 3 3 1  
16 5 3 2  
17 5 4 2  
18 3 4 1

Se trata, de contrastar la hipótesis de igualdad de los tres grupos.

- Es el ejemplo de la prueba de Friedman de la parte teórica.

19. Supongamos que un profesor de EGB quiere comprobar si tres métodos distintos de dirección de grupo originan diferentes rendimientos escolares entre los alumnos. Para ello, elige al azar tres grupos de 9,8 y 7 alumnos a los cuales los somete respectivamente a la dirección de grupo «autoritario», «directivo-democrática» y «dejar-hacer». Los resultados escolares son los siguientes para cada grupo:

**Autoritaria Directivo-democrática Dejar-hacer**

33 38 22

15 50 36

17 39 37

19 23 21

26 35 16

32 41 20

18 40 25

28 47

29

Planteamos la hipótesis nula de que no hay diferencia significativa entre los tres grupos.

- Es el ejemplo de la prueba de Kruskal-Wallis de la parte teórica. 20. Supongamos que la dirección de una empresa ha clasificado a ocho gerentes en relación con sus habilidades administrativas, y que a todos ellos se les ha hecho una prueba psicológica para la cual se supone que la calificación está correlacionada con la potencialidad para la administración. Con los datos que presentamos en la tabla siguiente podemos suponer alguna relación entre ambas notas.

**Gerente**

**Rango para las habilidades administrativas**

1 7

2 4

3 2

4 6

5 1

6 3

7 8

8 5

## Calificaciones de la prueba

44

72

69

70

93

82

67

80

- Es el ejemplo de correlación no paramétrica rangos de Spearman de la parte teórica.

21. Estamos interesados en analizar el grado de asociación entre el nivel de industrialización de los países de la CEE (medido por el porcentaje de empleo en la industria con respecto al total, factor A), y la parte de gastos totales familiares que dedican a transportes y comunicaciones, factor B. Los datos aparecen a continuación:

<b>País</b>	<b>Factor A</b>	<b>Factor B</b>
Alemania	40,7	13,9
Francia	32,1	13,7
Italia	34,6	13,6
Holanda	28,7	10,7
Bélgica	32,7	12,2
Luxemburgo	32,0	16,5
Reino Unido	35,0	16,9
Dinamarca	26,6	17,5
Grecia	26,0	17,9
España	33,0	13,8

- Es el ejemplo de correlación no paramétrica rangos de Kendall de la parte teórica.

22. Supongamos tres variables indicadoras de búsqueda de posición social, autoritarismo y conformidad (condescendencia) de 12 individuos, dadas por sus rangos según muestra la tabla siguiente. Queremos saber si existe relación entre búsqueda de posición social (X) y autoritarismo (Y) independiente de la variable Z (condescendencia).

### Sujeto

**Búsqueda de Autoritarismo Y Conformidad posición social X  
(condescendencia) Z**

1 3 2 1,5  
 2 4 6 1,5  
 3 2 5 3,5  
 4 1 1 3,5  
 5 8 10 5  
 6 11 9 6  
 7 10 8 7  
 8 6 3 8  
 9 7 4 9 10 12 12 10,5 11 5 7 10,5 12 9 11 12

- Es el ejemplo de correlación parcial no paramétrica rangos de Kendall de la parte teórica.

23. Supongamos que a tres ejecutivos se les pide entrevistar a seis solicitantes de empleo, con instrucciones de que han de separar los órdenes de las calificaciones que resulten. Los datos aparecen en la tabla siguiente:

**Solicitante a b c d e f** Ejecutivo X 1 6 3 2 5 4 Ejecutivo Y 1 5 6 4 2 3  
 Ejecutivo Z 6 3 2 5 4 1

- Es el ejemplo de correlación no paramétrica coeficiente de concordancia W de Kendall de la parte teórica.

### **SOLUCIÓN A LOS EJERCICIOS DE AUTOCOMPROBACIÓN**

1. La solución con SPSS pasa por la creación de un fichero mediante simulación con distribución normal de media 7,2 y desviación 0,3. Después se contrasta la hipótesis:

$H_0: \mu = 6,8$

$H_1: \mu \neq 6,8$

Para lo cual se utiliza el comando t-test. La sintaxis de los dos ficheros SPSS de procedimientos es:

\*Nota:

- \* Para utilizarlo basta con que cambies los parámetros referentes a: \* -  
Numero de vectores a generar.
- \* -Numero de casos a generar.
- \* -Parámetros de media y desviación típica que deseas en cada vector.

NEW FILE.

INPUT PROGRAM.

/\* NUMERO DE VECTORES A GENERAR (ej:1) [Modificable].

VECTOR X(1).

/\* NUMERO DE CASOS A GENERAR (ej:100) [Modificable].

LOOP #J=1 TO 100.

```

LOOP #I=1 TO 1.
compute X(#I)=normal(1)+0. /* GENERACION DE VECTORES
NORMALIZADAS.
END LOOP.
END CASE.
END LOOP.
END FILE.
END INPUT PROGRAM.
* ASIGNACION DEL CENTRO Y DISPERSION DESEADOS
[Modificable].media=7,2; desviación=0,3.
COMPUTE x1=(x1*0.3)+7.2.
EXECUTE.
+++++
T-TEST
/TESTVAL = 6.8
/MISSING = ANALYSIS
/VARIABLES = X1
/CRITERIA = CI(.95) .

```

Los resultados de aplicar el comando t-test de una sola muestra son:  
**PRUEBA PARA UNA MUESTRA Valor de prueba = 6.8** 95% Intervalo  
de confianza para la diferencia t  
gl  
Sig. Diferencia (bilateral) de medias

X1 12,517 99 ,000 ,38938 Inferior Superior ,3277 ,4511

Como la diferencia es significativa entonces se rechaza la hipótesis nula. 2. Para tratar el problema con SPSS hay que calcular la probabilidad de error tipo I o valor  $\alpha$ . Se plantea las hipótesis:

$H_0$ :  $p=0,9$  aprueban en el 90% de los casos.

$H_1$ :  $p<0,9$  la afirmación es falsa y lo hacen en menor proporción. Donde  $p$  representa la proporción de aprobados en la selectividad. Para contrastar la hipótesis se supone  $\alpha=0,05$ .

De la población no se tiene datos pero al ser la muestra grande se supone aproximadamente normal, y por tanto cumple las condiciones paramétricas.

Con la aproximación a la normal  $m=np=180 \times 0,9=162$  y  $s_p = npq = 4,02$ . En consecuencia la hipótesis anterior será similar a la siguiente:  
 $H_0:m=162$

$H_1: m < 162$

Se calcula  $z = \frac{x - 150}{\sqrt{402}} = 2.99$ . Por el tipo de hipótesis alternativa  $sp$

(unilateral) será necesario encontrar  $P[z < -2.99]$ . Se aplicará el programa siguiente:

```
data list free/x.
begin data
1
end data.
execute.

format x (f8.4).

***** probabilidad de z menor o igual que a *****.
***** por ejemplo prob. z menor o igual que -2,99 *****.
COMPUTE pz1 = CDFNORM(-2.99) .
format pz1 (f8.4).
```

execute.

da 0,0014, valor menor que el  $\alpha = 0,05$  fijado, y por tanto se rechaza la hipótesis nula, es decir, no aprueban en el 90% de los casos.

3. Las hipótesis son:

$H_0: m_A = m_B$

$H_1: m_A < m_B$

Luego se trata de un contraste de una cola.  $\alpha = 0,05$ .

Se supone normalidad. Además se desconocen las varianzas de las distribuciones. Por otra parte, será necesario introducir los datos de los grupos (se crea el fichero de datos t-test2.sav) y después se ejecuta el comando t-test:

```
T-TEST
GROUPS = g('a' 'b')
/MISSING = ANALYSIS
/VARIABLES = x
/CRITERIA = CI(.95) .
```

Dando como resultado:

ESTADÍSTICOS DE GRUPO

**g**

**N**

## Media

### Desviación Error tip. de típ. la media

x a 16 35,13 21,276 5,319 b 20 54,25 20,113 4,497

## PRUEBA DE MUESTRAS INDEPENDIENTES

### Prueba de

Levene para **Prueba T para la igualdad de medias** la igualdad  
**de varianzas**

F Sig. t gl Sig.

Diferencia de

(bilateral) medias

Error 95% Intervalo típ. de de confianza para la dife la diferencia

rencia

Inferior Superior Se han

asumido x varianzas iguales ,098 ,756 -2,763 34 ,009 -19,125 6,921 -33,190  
-5,060

No se

han

asumido varianzas iguales

-2,746 31,431 ,010 -19,125 6,965 -33,323 -4,927

4. Los resultados son los siguientes:

1) Se plantean las hipótesis:

$$H_0: s^2=0,0002$$

$$H_1: s^2>0,0002$$

2)  $\alpha=0,05$

3) Se supone normalidad según afirma el enunciado del problema.

4) El estadístico utilizado es:

$$F = \frac{s_1^2}{s_2^2} = \frac{0,098}{0,0002} = 490$$

$$F_{(13, 5)} = 16,919$$

5) Este estadístico se distribución según una ji-cuadrado con n-1 grados de libertad.

6) Como  $c^2_{(n-1)}=16,919$ , según se puede recoger del programa siguiente,

entonces la región crítica (RC) a un nivel de significación  $\alpha=0,05$  es:  $RC = \{c^2 > 16,919\}$ .

```
data list free/x.  
begin data  
1  
end data.  
execute.
```

```
format x (f8.4).
```

```
***** Valor de  $c^2$  tal que  $P[c^2 > c^2(9)] = 0.05$  *****.  
COMPUTE if1 = IDF.CHI(0.95,9) .  
format if1 (f8.4).
```

```
execute.
```

7) Como  $c^2 = 13,5$  no pertenece a la región crítica entonces se acepta la hipótesis nula, no hay evidencia para refutar lo que afirma el proveedor de lapiceros.

5. El programa SPSS para realizar el ejercicio es el siguiente:

\* PRUEBAS UNIMUESTRALES (TEST DE K-S).

```
data list free/x .  
begin data  
20  
22  
24  
30  
31  
32  
38  
end data.  
execute.
```

```
NPARTESTS  
/K-S(NORMAL)= x  
/MISSING ANALYSIS.
```

Los resultados son los siguientes:

PRUEBA DE KOLMOGOROV-SMIRNOV PARA UNA MUESTRA x N

Parámetros normales (a,b)

Diferencias más extremas

Z de Kolmogorov-Smirnov Sig. asintót. (bilateral)

7 Media 28,1429 Desviación típica 6,38823

Absoluta ,186 Positiva ,170 Negativa -,186

,492 ,969

a: La distribución de contraste es la Normal. b: Se han calculado a partir de los datos.

Nótese que las diferencias extremas difieren ligeramente del resultado obtenido de forma manual. Esta diferencia se debe a que el algoritmo que utiliza SPSS para calcular diferencias utiliza además de las diferencias entre  $F_n(x) - F(x)$ , las diferencias  $D_i = F_n(x_{i-1}) - F(x_i)$

6. El programa SPSS para realizar el ejercicio es el siguiente:

\* PRUEBAS UNIMUESTRALES (TEST DE CHI-CUADRADO).

data list free/x peso .

begin data

1 21

2 29

end data.

execute.

WEIGHT

BY peso .

NPARTEST

/CHISQUARE=x (1,2)

/EXPECTED=15 35

/MISSING ANALYSIS.

Y los resultados:

Categoría

1 1,00

2 2,00

Total

FRECUENCIAS

x

N observado N esperado Residual 21 15,0 6,0 29 35,0 -6,0 50

ESTADÍSTICOS DE CONTRASTE

x Chi-cuadrado(a) 3,429 gl 1

Sig. asintót. ,064

a: 0 casillas (,0%) tienen frecuencias esperadas menores que 5. La frecuencia de casilla esperada mínima es 15,0.

Resultados que coinciden con los obtenidos anteriormente de forma

manual.

7. El programa SPSS para realizar el ejercicio es el siguiente:

\* PRUEBAS UNIMUESTRALES (TEST DE RACHAS).  
data list free/x (A2).

begin data

A B B A A B A A A B A B A B B

end data.

execute.

RECODE  
x (CONVERT)  
(‘A’=1) (‘B’=2) INTO rx .

EXECUTE .

NPART TESTS  
/RUNS(1.5)=rx  
/MISSING ANALYSIS.

Nótese la necesaria conversión de caracteres en números (comando RECODE) para realizar la prueba de rachas.

Los resultados son:

PRUEBA DE RACHAS **rx**

Valor de prueba (a) 1,5000 Casos en total 15 Número de rachas 10 Z  
,556

a: Especificado por el usuario.

8. El programa SPSS para resolver el ejercicio es:

\* PRUEBAS BIMUESTRALES- MUESTRAS RELACIONADAS  
(PRUEBA DE MACNEMAR). data list free/x z peso.

begin data

1 0 14

1 1 4

0 0 3

0 1 4

end data.

execute.

WEIGHT  
BY peso .

NPART TEST  
/MCNEMAR= x WITH z (PAIRED)

/MISSING ANALYSIS.

Cuyo resultado es:

X y Z

**z**  
x 0 1  
0 3 4  
1 14 4

ESTADÍSTICOS DE CONTRASTE(b) x y z N 25 Sig. exacta  
(bilateral) ,031(a) a: Se ha usado la distribución binomial. b: Prueba de McNemar.

Como se puede observar da resultados significativos a un nivel del 5%. 9.  
El programa SPSS para realizar el ejercicio es:

\* PRUEBAS BIMUESTRALES- MUESTRAS RELACIONADAS  
(PRUEBA DE LOS SIGNOS).

data list free/x z .

begin data

14 19

16 18

23 22

26 27

24 28

28 35

27 30

18 17

15 17

22 28

20 30

25 24

end data.

execute.

    NPAR TEST

/SIGN= x WITH z (PAIRED)

/MISSING ANALYSIS.

Dando como resultado:

FRECUENCIAS

    N z - x Diferencias negativas(a) 3

Diferencias positivas(b) 9

Empates(c) 0

Total 12

a:  $z < x$   
b:  $z > x$   
c:  $z = x$

## ESTADÍSTICOS DE CONTRASTE (b)

**z - x**

Sig. exacta (bilateral) ,146(a)

a: Se ha usado la distribución binomial.

b: Prueba de los signos.

No existe por tanto diferencia significativa entre antes y después de ver la película.

10. El programa SPSS es:

\* PRUEBAS BIMUESTRALES- MUESTRAS RELACIONADAS (PRUEBA DE WILCOXON).

```
data list free/x z .
```

```
begin data
```

```
14 19
```

```
16 18
```

```
23 22
```

```
26 27
```

```
24 28
```

```
28 35
```

```
27 30
```

```
18 17
```

```
15 17
```

```
22 28
```

```
20 30
```

```
25 24
```

```
end data.
```

```
execute.
```

```
NPART TEST
```

```
/WILCOXON=x WITH z (PAIRED)
```

```
/MISSING ANALYSIS.
```

Con el siguiente resultado:

RANGOS

N

**Rango Suma promedio de rangos**

z - x Rangos negativos 3(a) 2,50 7,50 Rangos positivos 9(b) 7,83 70,50

Empates 0(c)

Total 12

a:  $z < x$

b:  $z > x$

c:  $z = x$

## ESTADÍSTICOS DE CONTRASTE (b)

**z - x** Z -2,2828(a) Sig. asintót. (bilateral) ,013

a: Basado en los rangos negativos.

b: Prueba de los rangos con signo de Wilcoxon.

Que coincide plenamente con el obtenido en la parte teórica. 11. El programa SPSS es:

\* PRUEBAS BIMUESTRALES- MUESTRAS INDEPENDIENTES (PRUEBA DE FISHER). data list free/x z peso.

begin data

1 1 6

1 0 2

0 1 1

0 0 5

end data.

execute.

WEIGHT

BY peso .

CROSSTABS

/TABLES=x BY z

/FORMAT= AVALUE TABLES

/STATISTIC=CHISQ

/CELLS= COUNT

/COUNT ROUND CELL .

Cuyo resultado es:

x

Total TABLA DE CONTINGENCIAS X \* Z

,00 1,00

**z**

**,00 1,00 Total 5**

2

7

1 6 6 8 7 14

## PRUEBAS DE CHI-CUADRADO Sig. asintótica Valor gl

(bilateral) Sig. exacta Sig. exacta (bilateral) (unilateral)

Chi-cuadrado de Pearson

Corrección por continuidad(a) 4,667(b) 1 ,031

2,625 1 ,105

Razón de verosimilitud 5,004 1 ,025

Estadístico exacto

de Fisher

Asociación lineal<sub>4,333 1 ,037</sub>por lineal

N de casos válidos 14

a: Calculado sólo para una tabla de 2x2. b: 4 casillas (100,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 3,00.

,103 ,051

Será necesario fijarse en el estadístico exacto de Fisher y ver que no hay diferencia significativa al nivel del 5%. Además hay que tener presente la utilización del procedimiento tablas de contingencia que muestra el estadístico de Fisher en tablas de 2x2 con un 25% o más de casillas con frecuencia esperada menor que 5.

12. El programa SPSS para resolver el ejercicio es:

\* PRUEBAS BIMUESTRALES- MUESTRAS INDEPENDIENTES (PRUEBA DE KOLMOGOROV).

data list free/x z.

begin data

39,1 1

41,2 1

45,2 1

46,2 1

48,4 1

48,7 1

55,0 1

40,6 1

52,1 1

47,2 1

35,2 0

39,2 0

40,9 0

38,1 0

34,4 0  
29,1 0  
41,8 0  
24,3 0  
32,4 0  
32,6 0  
end data.  
execute.

NPARTESTS  
/K-S= x BY z(1 0)  
/MISSING ANALYSIS.

El resultado es el siguiente:  
ESTADÍSTICOS DE CONTRASTE(a)  
Diferencias más extremas  
Z de Kolmogorov-Smirnov Sig. asintót. (bilateral)

x Absoluta ,700  
Positiva ,000  
Negativa -,700  
1,565  
,015

a: Variable de agrupación: z

Como se puede ver hay diferencia significativa.  
13. El programa SPSS es:

\* PRUEBAS BIMUESTRALES- MUESTRAS INDEPENDIENTES  
(PRUEBA DE JI-CUADRADO).

data list free/x z peso.

begin data

1 0 12

2 0 22

3 0 9

1 1 32

2 1 14

3 1 6

end data.

execute.

WEIGHT  
BY peso .

```

CROSSTABS
/TABLES=x BY z
/FORMAT= AVALUE TABLES
/STATISTIC=CHISQ
/CELLS= COUNT
/COUNT ROUND CELL.

```

Cuyo resultado es:

TABLA DE CONTINGENCIAS X \* Z z

```

,00 1,00 Total x 1,00 12 32 44
2,00 22 14 36
3,00 9 6 15 Total 43 52 95

```

PRUEBAS DE CHI-CUADRADO

**Valor gl Sig. asintótica (bilateral)** Chi-cuadrado de Pearson 10,712(a)  
 2 ,005 Corrección por continuidad  
 Razón de verosimilitud 10,976 2 ,004 Asociación lineal por lineal 8,166 1  
 ,004 N de casos válidos 95

a: 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 6,79.

Donde el resultado del estadístico de ji-cuadrado es ligeramente superior al calculado en la teoría por el tema de manejo de decimales. Su diferencia es significativa, es decir, no son independientes la altura de una persona y sus cualidades como dirigentes.

14. El programa SPSS que realiza el ejercicio es:

\* PRUEBAS VARIAS MUESTRAS- MUESTRAS  
 INDEPENDIENTES (PRUEBA DE LA MEDIANA).

```
data list free/x z .
```

```
begin data
```

```
6 1
```

```
10 1
```

```
19 1
```

```
21 1
```

```
20 1
```

```
14 1
```

```
29 1
```

```
11 1
```

```
12 1
```

```
24 1
```

13 1  
17 1  
18 1  
23 1  
22 1  
16 1  
7 0  
15 0  
26 0  
27 0  
8 0  
30 0  
26 0  
23 0  
9 0  
25 0  
28 0  
17 0  
15 0  
end data.  
execute.

NPARTESTS  
/MEDIAN=x BY z(0 1) /MISSING ANALYSIS.

Cuyo resultado es:

FRECUENCIAS z  
**,00 1,00 x >**

Mediana 7 7  
<=

Mediana 7 7

ESTADÍSTICOS DE CONTRASTE(a)

x N 29 Mediana 18,0000

Sig. exacta ,715 a: Variable de agrupación: z  
15. El programa SPSS será:

\* PRUEBAS BIMUESTRALES- MUESTRAS INDEPENDIENTES  
(PRUEBA U DE MANNWHITNEY).

data list free/x z .

begin data

17 1

12 1

15 1

10 1

14 1

11 1

8 1

13 1

13 0

16 0

9 0

11 0

9 0

18 0

6 0

10 0

7 0

end data.

execute.

NPART TESTS

/M-W= x BY z(1 0) /MISSING ANALYSIS.

Que da por resultado: RANGOS

**Z**

**N**

**Rango Suma promedio de rangos**

x ,00 9 7,83 70,50 1,00 8 10,31 82,50 *Total* 17

ESTADÍSTICOS DE CONTRASTE(b)

x U de Mann-Whitney 25,500

W de Wilcoxon 70,500

Z -1,013

Sig. asintót. (bilateral) ,311 Sig. exacta [2\*(Sig. unilateral)] ,321(a)

a: No corregidos para los empates.

b: Variable de agrupación: z

16. El programa SPSS del ejercicio es:

\* PRUEBAS BIMUESTRALES- MUESTRAS INDEPENDIENTES  
(PRUEBA DE WALDWOLFOWITZ).

```
data list free/x z .
begin data
86 1
69 1
72 1
65 1
113 1
65 1
118 1
45 1
141 1
104 1
41 1
50 1
55 0
40 0
22 0
58 0
16 0
7 0
9 0
16 0
26 0
36 0
20 0
15 0
end data.
execute.
```

#### NPART TESTS

/W-W= x BY z(1 0) /MISSING ANALYSIS.

Cuyo resultado es:

ESTADÍSTICOS DE CONTRASTE(b,c)

**Número<sub>z</sub> Sig. exacta de rachas (unilateral)**

$\chi^2$  Número exacto 4(a) --3,548 ,000 de rachas

a: No se han hallado empates intra-grupo. b: Prueba de Wald-Wolfowitz.  
c: Variable de agrupación: z.

17. El programa SPSS del ejercicio es:

\* PRUEBAS k-MUESTRAS RELACIONADAS (PRUEBA Q DE

COCHRAN). data list free/x z v.

begin data

0 0 0

1 1 0

0 1 0

0 0 0

1 0 0

1 1 0

1 1 0

0 1 0

1 0 0

0 0 0

1 1 1

1 1 1

1 1 0

1 1 0

1 1 0

1 1 1

1 1 0

1 1 0

end data.

execute.

NPART TESTS

/COCHRAN = x z v

/MISSING LISTWISE.

Con el siguiente resultado:

ESTADÍSTICOS DE CONTRASTE

N 18 Q de Cochran 16,667(a)

gl 2 Sig. asintót. ,000

a: 0 se trata como un éxito.

Como se puede observar altamente significativo.

18. El programa SPSS es:

\* PRUEBAS k-MUESTRAS RELACIONADAS (PRUEBA DE FRIEDMAN). data list free/x z v.

begin data

3 5 4

3 4 2

1 4 2

2 5 6  
7 1 3  
4 5 2  
3 2 1  
2 8 6  
7 3 4  
6 3 5  
5 6 4  
2 4 1  
4 2 1  
3 5 2  
3 3 1  
5 3 2  
5 4 2  
3 4 1

#### NPARTIAL TESTS

/FRIEDMAN = x z v /MISSING LISTWISE.

Con el resultado:

ESTADÍSTICOS DE CONTRASTE(a)

N 18 Chi-cuadrado 8,704

gl 2 Sig. asintót. ,013 a: Prueba de Friedman.

19. El programa SPSS es:

\* PRUEBAS k-MUESTRAS INDEPENDIENTES (PRUEBA DE KRUSKALWALLIS).

data list free/x z .

begin data

33 1

15 1

17 1

19 1

26 1

32 1

18 1

28 1

29 1

38 2

50 2

39 2

23 2  
35 2  
41 2  
40 2  
47 2  
22 3  
36 3  
37 3  
21 3  
16 3  
20 3  
25 3

#### NPART TESTS

/K-W=x BY z(1 3) /MISSING ANALYSIS.

Con el siguiente resultado:

RANGOS

**z**

**N**

**Rango promedio**

x 1,00 9 8,67  
2,00 8 19,25  
3,00 9 9,71  
*Total* 24

ESTADÍSTICOS DE CONTRASTE(a,b) x Chi-cuadrado 11,021 gl 2  
Sig. asintót. ,004 a: Prueba de Kruskal-Wallis. b: Variable de agrupación: z.

Coincide plenamente con el obtenido en la parte teórica.

20. El programa SPSS es:

\* PRUEBAS CORRELACIÓN DE RANGOS DE SPEARMAN. data  
list free/x z .

begin data

7 44  
4 72  
2 69  
6 70  
1 93  
3 82  
8 67  
5 80

```
NONPAR CORR
/VARIABLES=x z
/PRINT=SPEARMAN TWOTAIL NOSIG /MISSING=PAIRWISE .
```

Con el siguiente resultado:

CORRELACIONES

**x z**

Rho de Spearman x Coeficiente de correlación 1,000  $-.714^{(*)}$  Sig. (bilateral)  $.047$  N 8 8

z Coeficiente de correlación  $-.714^{(*)}$  1,000 Sig. (bilateral)  $.047$  . N 8 8

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

21. El programa SPSS es:

\* PRUEBAS CORRELACIÓN DE RANGOS DE KENDALL. data  
list free/x z .

begin data

40,7 13,9

32,1 13,7

34,6 13,6

28,7 10,7

32,7 12,2

32,0 16,5

35,0 16,9

26,6 17,5

26,0 17,9

33,0 13,8

end data.

execute.

```
NONPAR CORR
/VARIABLES=x z
/PRINT=KENDALL TWOTAIL NOSIG
/MISSING=PAIRWISE .
```

Cuyo resultado es:

CORRELACIONES

**x z**

Tau\_b de Kendall x Coeficiente de correlación 1,000  $-.156$  Sig. (bilateral)  $.531$  N 10 10

z Coeficiente de correlación  $-.156$  1,000 Sig. (bilateral)  $.531$  . N 10 10

Como se puede observar no es significativo, por tanto las variables no están correlacionadas.

22. El programa SPSS es:

```
* PRUEBAS CORRELACIÓN PARCIAL DE RANGOS DE  
KENDALL. data list free/x y z.
```

```
begin data
```

```
3 2 1,5
```

```
4 6 1,5
```

```
2 5 3,5
```

```
1 1 3,5
```

```
8 10 5
```

```
11 9 6
```

```
10 8 7
```

```
6 3 8
```

```
7 4 9
```

```
12 12 10,5
```

```
5 7 10,5
```

```
9 11 12
```

```
end data.
```

```
execute.
```

```
NONPAR CORR
```

```
/VARIABLES=x y
```

```
/PRINT=KENDALL TWOTAIL NOSIG
```

```
/MISSING=PAIRWISE .
```

```
NONPAR CORR
```

```
/VARIABLES=z y
```

```
/PRINT=KENDALL TWOTAIL NOSIG
```

```
/MISSING=PAIRWISE .
```

```
NONPAR CORR
```

```
/VARIABLES=x z
```

```
/PRINT=KENDALL TWOTAIL NOSIG /MISSING=PAIRWISE .
```

Con los resultados:

Tau\_b de Kendall x

y

CORRELACIONES x y

Coeficiente de correlación 1,000 ,667(\*\*) Sig. (bilateral) ,00310 N 12  
12 Coeficiente de correlación ,667(\*\*) 1,000 Sig. (bilateral) ,00310 . N 12

12

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

Tau\_b de Kendall z

y CORRELACIONES

z y Coeficiente de correlación 1,000 ,357

Sig. (bilateral) . ,112

N 12 12 Coeficiente de correlación ,357 1,000

Sig. (bilateral) ,112 . N 12 12

Tau\_b de Kendall x

z CORRELACIONES

x z Coeficiente de correlación 1,000 ,388

Sig. (bilateral) . ,084

N 12 12 Coeficiente de correlación ,388 1,000

Sig. (bilateral) ,084 . N 12 12

Que coinciden con el texto. Después será necesario aplicar la fórmula presentada en la parte teórica.

23. El programa SPSS es:

\* PRUEBAS COEFICIENTE DE CONCORDANCIA DE KENDALL. data list free/a b c d e f.

begin data

1 6 3 2 5 4

1 5 6 4 2 3

6 3 2 5 4 1

end data.

execute.

NPAR TESTS

/KENDALL = a b c d e f

/MISSING LISTWISE.

Con los siguientes resultados:

ESTADÍSTICOS DE CONTRASTE

N 3 W de Kendall (a) ,162 Chi-cuadrado 2,429 gl 5 Sig. asintót. ,787 a:

Coeficiente de concordancia de Kendall.

## **BIBLIOGRAFÍA**

ARCE, CONSTANTINO; REAL, EULOGIO (2002): *Introducción al análisis estadístico con SPSS*. PPU. Barcelona.

CALOT, G. (1974): *Curso de Estadística Descriptiva*. Paraninfo. Madrid.

CAMACHO ROSALES, JUAN (2002): *Estadística con SPSS versión 11 para*

Windows. Ra-Ma, Librería y Editorial Microinformática. Madrid.

CUADRAS, C. M.; ECHEVERRÍA, B.; MATEO, J.; SÁNCHEZ, P. (1991): *Fundamentos de Estadística*. PPU. Barcelona.

FERRÁN ARANAZ, MAGDALENA (2002): *Curso de SPSS para Windows*. McGraw-Hill/ Interamericana de España. Madrid.

Labrousse, C.(1976). *Estadística. Ejercicios resueltos (Tomo II)*. Madrid: Paraninfo.

LIZASOAIN HERNÁNDEZ, LUIS; JOARISTI OLARRIAGA, LUIS (2003): *Gestión y análisis de datos con SPSS*. Thomson Paraninfo. Madrid.

MARTÍN F. G. (1994): *Introducción a la Estadística Económica y Empresarial*. AC. Madrid.

MARTÍN, M. F.; FERNÁNDEZ, R.; SEISDEDOS, A. (1985): *Estadística inferencial. Manual de prácticas para las ciencias de la conducta*. Salamanca: Universidad pontificia.

PARDO MERINO, ANTONIO; RUIZ DÍAZ, MIGUEL ÁNGEL (2002): *SPSS 11. Guía para el análisis de datos*. McGraw-Hill/ Interamericana de España. Madrid.

PEÑA D. (1992): *Estadística, Modelos y Métodos. Volumen I*. Alianza Universidad Textos. Madrid.

PEÑA D. (1992): *Estadística, Modelos y Métodos. Volumen II*. Alianza Universidad Textos. Madrid.

P ÉREZ, CÉSAR (2001): *Técnicas estadísticas con SPSS*. Pearson Educación. Madrid.

RÍOS, S. (1974): *Métodos estadísticos*. Ed. del Castillo. Madrid.

SPIEGEL, M. R. (1992): *Estadística*. McGraw-Hill. Madrid.

VISAUTA VINACUA, B. (2002): *Análisis estadístico con SPSS 11.0 para Windows*. McGraw-Hill/ Interamericana de España. Madrid.

## **UNIDAD DIDÁCTICA 5**

### **HACIA UN ESTUDIO DEL MODELO**

#### **Objetivos**

Diferenciar los principales conceptos relacionados con el análisis de regresión: método de selección de las variables, condiciones de aplicación del modelo y evaluación del mismo.

Saber interpretar los resultados del modelo de regresión, efectuando una valoración del proceso y del ajuste final obtenido. Aplicar los conocimientos de análisis de regresión para resolver supuestos de investigación educativa utilizando el programa SPSS. Interpretar correctamente las salidas de ordenador de SPSS en los supuestos

anteriores.

Comprender el significado, los procedimientos y los elementos básicos de los diseños experimentales.

Interpretar el significado de la variabilidad y la interacción en el diseño experimental.

Comprender la importancia de los diseños factoriales en la experimentación y resolver supuestos de diseños factoriales utilizando el análisis de la varianza.

## 1. EL MODELO DE REGRESIÓN. EL COMANDO REGRESION

### 1.1. El modelo de regresión

#### 1.1.1. Introducción

La regresión lineal estudia la relación existente entre una o más variables, denominadas independientes y otra, denominada dependiente, con propósitos tanto descriptivos como predictivos.

Podemos plantear una relación, en principio lineal, entre una variable Y dependiente que trata de ser explicada por k variables independientes y un término de perturbación aleatoria e. De esta forma para cada observación se tendrá:

$$y_i = b_0 + b_1 x_{i1} + \dots + b_k x_{ik} + e_i \quad i=1, \dots, n \quad [1]$$

donde:

$b_0, \dots, b_k$  son parámetros desconocidos a estimar, y

$e_i \quad i=1, \dots, n$  son variables error, independientes y con distribución  $N(0, s^2)$

De forma matricial  $\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$  donde  $\mathbf{X}$  es una matriz con la primera columna unitaria.

El análisis de regresión es una de las técnicas más utilizada en investigación, sus posibilidades son innumerables como lo demuestran las continuas referencias en publicaciones. Sus aplicaciones se pueden agrupar en dos grandes apartados: **predicción** y **explicación**. Estos dos usos no son mutuamente excluyentes y existirán investigaciones donde se utilicen con ambas finalidades.

En la **predicción**, la combinación lineal de las variables independientes se dirige a maximizar la estimación de la variable dependiente, y es un predictor del poder explicativo de la variable dependiente por las variables independientes. Se deben conseguir niveles adecuados de explicación de la variable dependiente para justificar el modelo de regresión. También, la

faceta predictiva del análisis de regresión, sirve para evaluar el conjunto de variables independientes como predictoras de la variable dependiente.

La vertiente **explicativa** del análisis de regresión, se utiliza para dar una visión de la importancia relativa de cada variable independiente valorando su magnitud y signo. Además, se puede trabajar para determinar el tipo de relación existente (lineal, cuadrática, logarítmica, exponencial, potencial, etc) con la variable dependiente.

### **Regresión lineal simple**

En el caso particular de una única variable independiente **X**, se habla de regresión lineal simple. La correspondiente función de regresión será del tipo:  $f(X, b_0, b_1) = b_0 + b_1 X$

$$y_i = b_0 + b_1 x_i + e_i \quad i=1, \dots, n \quad [2]$$

Es de destacar la semejanza entre el modelo [1] y el análisis de la varianza modelo factorial con un solo factor, siendo la única diferencia la relativa a que mientras en [2] la variable **X** puede tomar cualquier valor, en el modelo de análisis de la varianza sólo puede tomar los valores 1,0, según se encuentre presente o no el nivel considerado.

### **Regresión lineal múltiple**

En el caso de más de una variable independiente, se habla de regresión lineal múltiple. Su modelo matemático se presentó en [1].

#### **1.1.2. Procedimiento**

La consecución del modelo de regresión exige el siguiente procedimiento: a) Elegir un método de selección de variables.

b) Determinar si hay observaciones que desvirtúen el modelo y analizar las condiciones de aplicación. Es decir, evaluar el modelo.

c) Evaluación de la significación en el modelo.

d) Interpretar los resultados efectuando una valoración del proceso y del ajuste final obtenido.

#### **A) Selección de las variables**

Existen diversos criterios: unos, emanados del problema de investigación, y con claros tintes teóricos, y otros por criterios empíricos.

En el primer caso se puede dar **errores de especificación**, tomando variables o incluyendo otras irrelevantes para la investigación. La inclusión de variables irrelevantes afecta a la parsimonia del modelo, y la falta de variables relevantes influye en el poder explicativo del mismo.

Además del error de especificación, las variables pueden tener errores de medida, que en el caso de las independientes, influyan en las predicciones

de la dependiente. Los errores de medida se pueden evaluar mediante ***análisis causal***.

Cuando se utilice variables ficticias, los coeficientes del modelo de regresión representarán las diferencias entre la media del grupo y la del grupo de referencia (el de valor nulo).

### **Métodos de selección de variables**

Entre los procedimientos alternativos a calcular todas las posibles ecuaciones de regresión, en función de todas las combinaciones posibles de las variables independientes, destacan los métodos de construcción por pasos:

a) *Método Backward*: la ecuación comienza con todas las variables incluidas; en cada paso se eliminará una variable.

b) *Método Forward*: en cada paso se introduce una variable.

c) *Método Stepwise*: en cada paso puede eliminarse o introducirse una variable. Dado que una variable puede entrar y salir de la ecuación en más de una ocasión, es conveniente establecer un límite para el número de pasos. En general, se considera el doble del número de variables independientes. En este procedimiento por pasos se debe tener en cuenta la influencia de la multicolinealidad entre las variables independientes. El investigador debe plantear un modelo teórico con la inclusión de las variables más relevantes y los signos de las mismas.

A la hora de calcular los coeficientes, para asegurar que la tasa de error conjunto a lo largo de todos los tests de significación sea razonable, deben emplearse umbrales muy conservadores (0,01) al añadir o quitar variables (Hair 1999, página 173).

**B) Supuestos y limitaciones para la construcción de la ecuación de regresión** Identificar el cumplimiento de los condicionantes del modelo, debe considerarse como paso previo y de validación del análisis de regresión.

#### **Identificación de observaciones influyentes**

Hair (1999, página 177) las clasifica en tres grupos: datos atípicos, puntos de apalancamiento e influyentes.

Estos puntos «distintos» se basan en alguna de las siguientes condiciones (Hair 1999, página 178):

— Un error en la entrada de observaciones o datos.

— Una observación válida, aunque excepcional, explicable por una situación extraordinaria.

— Una observación excepcional sin una explicación plausible. — Una observación ordinaria en sus características individuales pero excepcional en su combinación de características.

Estas observaciones influyentes deben ser aisladas antes de comenzar la aplicación del método para evitar defectos en las predicciones realizadas con el mismo. Los casos atípicos han sido muy estudiados, de forma que se han desarrollado métodos de regresión robustos para minimizar su impacto.

Los datos relevantes (de gran peso o importancia en el modelo), son identificados cuando se emplea el SPSS mediante el «Dfajuste» . Se calcula el valor de la predicción para un elemento, cuando el mismo está vinculado a la muestra y cuando no está incluido en ella, de tal forma la diferencia viene representada por el valor de «Dfajuste» o su valor tipificado «Dfajuste tipificado». Si esta diferencia es grande la observación  $(x_i, y_i)$  tendrá mucha importancia en el modelo de regresión, en caso contrario será menor su influencia. También se puede valorar los casos atípicos a través de los residuos estandarizados cuya distribución es  $N(0,1)$  y por tanto, valores mayores a 2 o 3, según criterio del investigador, serán considerados datos atípicos.

### **Comprobación de las hipótesis del modelo**

El modelo de regresión debe: a) estar **bien especificado**; b) las variables medidas sin error sistemático; y c) los errores en la predicción cumplir unas determinadas condiciones (ser independientes con distribución  $N(0, s^2)$ ).

El estar bien definido exige tener unas **variables independientes relevantes**, o de otra manera, el modelo de regresión debe cumplir el principio de parsimonia, es decir, la conformación del modelo con el menor número posible de variables independientes. Para valorar la aportación de cada variable independiente al modelo habrá que observar si el incremento del coeficiente de determinación ( $R^2$ ) es significativo.

La existencia de errores sistemáticos de medida, en general, dificulta la creación de cualquier modelo predictivo.

Respecto a los residuos y la definición del modelo, se cumplirá: a) *Linealidad*

Cada variable independiente tiene una relación lineal con la dependiente; o de otra forma, para cada variable independiente la linealidad indica que el coeficiente de regresión es constante a lo largo de

los valores de la variable independiente (regresión lineal simple). O de forma equivalente  $E(e_i)=0$  para  $i=1,\dots,n$

La comprobación de la linealidad de cada variable independiente se puede hacer por:

— Los residuos no deben presentar ningún patrón sistemático respecto de las predicciones o respecto de cada una de las variables independientes, se observará mediante el gráfico de **residuos estandarizados**.

— La correlación parcial entre la variable dependiente y cada una de las independientes debe ser alta. También los **gráficos de regresión parcial** deben presentar una forma lineal.

#### b) *Homocedasticidad*

Las varianzas de las distribuciones de Y ligadas a los distintos valores de las variables independientes deben ser iguales.  $\text{Var}(Y/x_{i1},x_{i2},\dots,x_{ik})=s^2$  o de forma equivalente  $\text{Var}(e_i)=s^2$ , para  $i=1,\dots,n$ :

— Los residuos no deben presentar ningún patrón sistemático respecto de las predicciones o respecto de cada una de las variables independientes.

— Se puede emplear el test de Levene. Si hay heterocedasticidad se puede utilizar transformaciones en las variables o el método de mínimos cuadrados ponderados.

#### c) *Independencia*

El valor observado en una variable para un individuo no debe estar influenciado en ningún sentido por los valores de esta variable observados en otros individuos, es decir, cada variable predictor es independiente. En el supuesto de normalidad, equivale a  $\text{Cov}(Y_i,Y_j)=0$  si  $i \neq j$ . y para los residuos, con el mismo supuesto de normalidad, será  $\text{Cov}(e_i,e_j)=0$  si  $i \neq j$ . Estas condiciones se traducen:

— Los residuos no deben presentar ningún patrón sistemático respecto a la secuencia de casos.

— Los residuos deben estar incorrelados; el estadístico de Durbin-Watson, D, debe tener valores próximos a 2, si D es menor que 1,5 existe autocorrelación. Si D se aproxima a 4 los residuos estarán negativamente autocorrelados y, si se aproxima a 0, estarán positivamente autocorrelados.

#### d) *Normalidad*

Se cumple  $Y/x_{i1},x_{i2},\dots,x_{ik}$  es  $N(b_0+b_1x_{i1}+\dots+b_kx_{ik}; s^2)$ , o de forma

equivalente, que la distribución de los residuos sea normal,  $N(0, s^2)$ : — Los residuos observados y los esperados, bajo hipótesis de distribución normal, deben coincidir.

— Para su comprobación se puede utilizar métodos gráficos como el diagrama P-P, o métodos analíticos, como la prueba de KolmogorovSmirnov.

### e) *Multicolinealidad*

El término multicolinealidad influye en la definición del modelo y se utiliza para describir la situación en que un gran número de variables independientes están altamente interrelacionadas. Las variables que sean aproximadamente una combinación lineal de otras se denominan multicolineales.

Si una variable es una combinación lineal perfecta de otras variables independientes, la matriz de correlaciones será singular (matriz singular es aquella cuyo determinante es igual a 0), lo que se traducirá a la hora de calcular la ecuación de regresión, en que no existirá una única solución mínimo-cuadrática insesgada de cálculo de sus coeficientes.

Una matriz de correlaciones con coeficientes muy altos es un indicio de probable multicolinealidad; sin embargo, puede haber multicolinealidad aunque los coeficientes sean relativamente bajos.

Uno de los procedimientos más utilizado para detectar la interdependencia entre variables es el criterio de la tolerancia.

La **tolerancia** de una variable  $X_i$  con las restantes variables independientes se define como:

$$\text{Tol}_i = 1 - R_i^2$$

donde  $R_i^2$  es el cuadrado del coeficiente de correlación múltiple entre  $X_i$  y las variables  $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_k$

— Si  $\text{Tol}_i = 0$  la variable  $X_i$  es casi una combinación lineal de las restantes variables y,

— Si  $\text{Tol}_i = 1$  la variable  $X_i$  puede reducir la parte de variación de  $Y$  no explicada por las restantes variables.

En el método de selección de variables por pasos, la variable seleccionada debe tener una tolerancia mínima con las variables incluidas en la ecuación para poder entrar en el siguiente paso. Por otro lado, al entrar la variable, ninguna variable en la ecuación debería superar esa mínima tolerancia con las restantes.

Para solucionar los problemas de multicolinealidad se puede: a) aumentar el tamaño muestral, b) a partir de las variables relacionadas construir otra como combinación lineal de las anteriores y c) utilizar un procedimiento jerárquico para introducir las variables y controlar la tolerancia de las mismas.

## C) Evaluación de la significación del modelo de regresión

### 1. Estimación de los parámetros

Calcular la ecuación de regresión supone deducir la ecuación del plano que mejor se ajusta a la nube de puntos (Etxeberria 1999, página 54).

^

Sea  $\hat{B}$  un estimador del vector de parámetros  $\mathbf{B}$ . Se define el vector de predicciones como:

^^

$\hat{Y} = \hat{X}\hat{B}$  El vector de residuos es

^

$e = Y - \hat{Y}$

Uno de los criterios para obtener los coeficientes de regresión  $B_0, B_1, \dots, B_k$ , estimaciones de los parámetros desconocidos  $b_0, b_1, \dots, b_k$ , es el de mínimos cuadrados, que consiste en minimizar la suma de los cuadrados de los residuos.

Si en el modelo de regresión se calcula  $[X'X]$  y es una matriz no singular, es decir, si su determinante  $|X'X|$  es distinto de cero, se puede calcular la inversa  $[X'X]^{-1}$  y entonces la matriz de los coeficientes será:

^

$\hat{B} = [X'X]^{-1}X'Y$

Los  $b_i$  son los *coeficientes de regresión parciales*, y así, por ejemplo,  $b_2$  nos da la variación de  $y$ , inducida por una variación de  $X_2$ , suponiendo que las demás variables permanecen constantes.

### 2. Propiedades de los estimadores

#### Estimador de los coeficientes del modelo lineal

Como hemos visto el estimador de  $B$  por el método de mínimos cuadrados es:

^

$\hat{B} = [X'X]^{-1}X'Y$

^

Es un estimador insesgado con  $\text{Var}(\hat{B}) = s^2[X'X]^{-1}$

**El estimador de la varianza**

Una hipótesis del modelo es la homocedasticidad, por tanto,  $\text{Var}(e_i) = s^2$  para  $i=1, \dots, n$ . El parámetro  $s^2$  habitualmente es desconocido y por tanto es necesario estimarlo. El estimador de este parámetro es la **varianza residual** definida como «el cociente entre la suma de residuos al cuadrado ( $SC_{\text{res}}$ ) y el número de grados de libertad del modelo ( $gl$ )»

$$s^2 = \frac{SC_{\text{res}}}{n - k} = \frac{\sum_{i=1}^n (e_i)^2}{n - k}$$

Si se utiliza la hipótesis de normalidad se obtiene la relación siguiente de la distribución de  $S_e^2$

$$\frac{SC_{\text{res}}}{s^2} \sim \chi^2_{n-k}$$

Obteniéndose como intervalo de confianza de  $s^2$  el siguiente:  $\frac{SC_{\text{res}}}{n-k} \left[ \frac{1}{F_{\alpha/2, n-k, \infty}}, \frac{1}{F_{1-\alpha/2, n-k, \infty}} \right]$

$$\frac{SC_{\text{res}}}{n-k} \left[ \frac{1}{F_{\alpha/2, n-k, \infty}}, \frac{1}{F_{1-\alpha/2, n-k, \infty}} \right]$$

### 3. El análisis de la varianza

A continuación se verá la descomposición de la variabilidad de la variable  $Y$  cuando se ajusta a un modelo de regresión múltiple.

Se puede comprobar la descomposición de cada observación muestral en:

$$y_i = \hat{y}_i + e_i$$

$$y = \hat{y} + e$$

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$$

o de forma matricial:

$$SC_{\text{res}} = [y - X\hat{\beta}]' [y - X\hat{\beta}]$$

$$SC_{\text{res}} = [y - X\hat{\beta}]' [y - X\hat{\beta}]$$

$$SC_{\text{res}} = [y - X\hat{\beta}]' [y - X\hat{\beta}]$$

$$-2 \\ y ] \\ SC \\ total \\ = [y'y - n \\ -2 \\ y ]$$

La descomposición de la suma de cuadrados nos lleva a la siguiente tabla:

TABLA 5.1.1. Tabla de análisis de la varianza para el modelo de regresión.

**Fuente Suma Grados de variación de cuadrados de Libertad Media de cuadrados Estadístico o varianzas F**

Regresión  $SC_{reg}$   $k$

$$2 = SMC_{reg} = SC_{reg} S^2 MC_{regR}$$

$$R^2 k S^2 = MC_{rese}$$

Residual  $SC_{res}$   $n - (k + 1)$   $e = SC_{res} 2 = SMC_{res}$   $n - k - 1$

Total  $SC_{total}$   $n - 1$

$$2 = SMC_{total} = SC_{total}$$

$T$   $n - 1$

**Contraste múltiple:  $H_0: B_1 = \dots = B_k = 0$  frente a  $H_1: \exists i: B_i \neq 0$**

La hipótesis nula significa que las variables independientes no mejoran la predicción de Y sobre

—

$y y = \bar{y}$ . La **tabla anterior de análisis de la varianza** a través de  $F_{k, n - (k + 1)}$  permite estudiar la significación en el contraste múltiple. Si resulta significativo algún  $B_i$  es distinto de cero.

**Contraste simple:  $H_0: B_i = 0$  frente  $H_1: B_i \neq 0$**

La hipótesis nula significa que la variable  $X_i$  no mejora la predicción de Y sobre la regresión obtenida con las  $k - 1$  variables restantes.

El estadístico de contraste  $t = \frac{B_i}{s_{B_i}}$  donde  $s_{B_i}$  se distribuye bajo  $H_0$  como una  $s_{B_i}$

t de Student con  $n - (k + 1)$  grados de libertad. Si el p-valor asociado es menor que  $\alpha$ , se rechazará la hipótesis nula al nivel de significación  $\alpha$ .

Existe otro procedimiento de realizar esta prueba que presenta la mejora respecto al anterior en permitir ejecutar contraste de varias variables a la vez. Para ello si se quiere contrastar la influencia de la variable  $X_i$  se ajusta el modelo de regresión completo con las  $k$  variables

independientes y se calcula la  $SC_{reg}(k)$ . Después se realiza el mismo proceso pero con las  $k-1$  variables, todas menos la  $X_i$  y se calcula  $SC_{reg}(k-x_i)$ . Se define la suma de cuadrados incremental debida a  $X_i$  como:

$$DSC_{reg}(i) = SC_{reg}(k) - SC_{reg}(k-x_i)$$

Se plantea la hipótesis anterior  $H_0: B_i=0$  frente  $H_1: B_i \neq 0$  y se utiliza como estadístico:

$$F = \frac{DSC_{reg}(i)}{SC_{reg}(k-x_i)} \sim F_{1, n-(k+1)}$$

que se distribuye según  $F_{1, n-(k+1)}$ . Con este procedimiento se obtiene los mismos resultados que con el contraste t, pero además tiene la ventaja que se puede utilizar para un conjunto  $l \leq k$   $\{x_{j1}, x_{j2}, \dots, x_{jl}\}$  de variables independiente, dando:

$$DSC_{reg}(l) = SC_{reg}(k) - SC_{reg}(k-l)$$

que se distribuye según una F con  $l, n-(k+1)$  grados de libertad.

#### 4. Análisis de la asociación entre las variables

Al ajustar un modelo de regresión múltiple a una nube de puntos es importante disponer de medidas que permitan medir la bondad del ajuste. Esto se consigue con los coeficientes de correlación múltiple. Como sabemos la correlación mide el grado o fuerza de relación existente entre variables.

##### El coeficiente de correlación simple (o de Pearson)

El coeficiente de correlación simple  $r$ , mide el grado de asociación lineal entre las variables X e Y  $r_{xy}$  es tal que:  $-1 \leq r_{xy} \leq 1$

- Si  $r_{xy}=1$  la asociación será lineal positiva
- Si  $r_{xy}=-1$  la asociación será lineal negativa y,
- Si  $r_{xy}=0$  no existirá asociación lineal

El estimador muestral del  $r_{xy}$  es el coeficiente de correlación muestral  $r_{xy}$

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

SS<sub>XY</sub>

Donde el numerador es la covarianza muestral entre las variables X e Y;  $S_X, S_Y$  son las desviaciones típicas muestrales de X e Y respectivamente.

### El coeficiente de determinación

En general cuando se ajusta un modelo estadístico a una nube de observaciones, una medida de la bondad de ajuste es el coeficiente de determinación, definido como:

$$R^2 = \frac{\sum_{i=1}^n \hat{A}_i^2}{\sum_{i=1}^n A_i^2} = \frac{SC_{reg}}{SC_{tot}}$$

$R^2$  representa el porcentaje de variabilidad de Y que explica el modelo de regresión. Si el modelo de ajuste es el lineal múltiple, R se denomina **coeficiente de correlación múltiple**.

Además, se puede establecer una relación con la F de la tabla 5.1.1. Como:

$$F = \frac{R^2 / k}{(1 - R^2) / (n - k - 1)}$$

$R^2$  y F se pueden relacionar F y R mediante:  $F = \frac{R^2 / k}{(1 - R^2) / (n - k - 1)}$

Cuando n es pequeño,  $R^2$  es muy sensible a los valores de n y k, entonces los programas de ordenador dan el  $R^2$  (ajustado) que modula la influencia del tamaño muestral en su valor:

$$R^2_{ajustado} = \frac{R^2(n-1)}{n-k-1}$$

## El coeficiente de correlación múltiple

El coeficiente de correlación múltiple  $r_{Y.1..k}$ , es una medida del grado de asociación lineal entre  $Y$  y el conjunto de variables independientes  $X_1, \dots, X_k$ , y es tal que:  $0 \leq r_{Y.1..k} \leq 1$

— Si  $r_{Y.1..k} = 1$  el ajuste del plano de regresión a la población es casi perfecto y — Si  $r_{Y.1..k} = 0$  el plano de regresión no mejora la predicción de  $Y$  sobre la predicción con la media muestral de  $Y$ .

El estimador muestral del  $r_{Y.1..k}$ , es el **coeficiente de correlación múltiple** muestral,  $R$ .

Todos los cálculos necesarios para el análisis del grado de asociación lineal se suelen disponer en una tabla como la siguiente:

### Fuente de variación Suma de cuadrados Varianza Correlación

Debida a la regresión

$$b' X' y - ny$$

—

2

S

$R^2$

R

2

=

S

$2R$

S

$2y$

Debida al error

$$y'y - b' X'y$$

S

$e^2$

$$1 - R^2 = \frac{e^2}{S_y^2}$$

$$\text{Total } y'y - ny^2 = S_y^2$$

Con todo lo anterior, el coeficiente de correlación múltiple será:

$R =$

$$\frac{1 - \frac{e^2}{S_y^2}}{1 - \frac{yy - b' X'y}{ny^2}}$$

$$\frac{S_y^2 - \frac{yy - b' X'y}{ny^2}}{S_y^2 - \frac{yy - ny^2}{ny^2}}$$

### El coeficiente de correlación parcial

Puede interesar estudiar el grado de asociación existente entre dos

variables (por ejemplo  $Y$  y  $X_1$ ) una vez que se ha eliminado la influencia que las restantes independientes ejercen sobre ella. Este problema viene resuelto mediante la determinación del **coeficiente de correlación parcial**, que representaremos como

$r_{y1.2,3,4,\dots,k}$

Una de las expresiones más utilizada es:

$$r_{y1.2,3,4,\dots,k}^2 = \frac{adj\ s_{12}^2}{SS_{11}}$$

Donde  $adj\ s_{12}$

representa el adjunto del elemento  $s_{12}$  en la matriz de covarianzas.

### 1. Predicción en el modelo de regresión lineal múltiple

Uno de los fines primordiales que se persigue al ajustar una función a una nube de puntos es el de poder extrapolar, esto es, dado el valor de la variable/s «independiente/s» exterior al recorrido que presenta la nube de puntos, calcular el correspondiente valor teórico de la variable «dependiente».

El ajuste será más preciso conforme el valor de la variable independiente esté próximo a los valores primitivos.

#### D) Interpretación de resultados

Para interpretar los resultados del análisis de regresión múltiple será necesario:

##### Evaluar el coeficiente de regresión

Para ver la influencia de cada variable en el modelo. Se utiliza los coeficientes beta con los datos estandarizados.

##### Evaluación de la multicolinealidad

- Valorar el grado de multicolinealidad.
- Determinar su impacto en los resultados.

Según hemos comentado, para evaluar la colinealidad de parejas o de múltiples variables se utiliza el valor de la tolerancia o su inverso el factor de influencia de la varianza (VIF).

La multicolinealidad hace inestable los coeficientes de la ecuación de regresión aumentando la variación de los mismos y en consecuencia los intervalos de confianza.

Además de interpretar los resultados, el análisis de regresión exige la validación de resultados como observación del poder de generalización de los mismos.

### Validación de resultados

— En primer lugar será necesario tener en cuenta el valor de  $R^2$ . — Se puede coger una muestra adicional o dividir la muestra.

— Se puede utilizar el estadístico «PRESS» que es una medida parecida al  $R^2$  pero para  $n-1$  modelos de regresión. Es un procedimiento similar a las técnicas de «bootstrapping» de remuestreo.

— Comparación de los modelos de regresión. Se utilizará distinto número de predictores y/o distinto ajuste (lineal, cuadrático, cúbico, etc.). Será necesario utilizar el  $R^2$  ajustado para evitar la influencia del tamaño muestral.

#### 1.1.3. Variables de intervención

En ocasiones se desea incluir en la ecuación de regresión, variables categóricas. Para ello es necesario crear las denominadas variables de intervención.

Si la variable independiente es nominal dicotómica, bastará con crear una variable con el valor 0 para una categoría y 1 para la otra e incluir esta variable en la ecuación como una más.

Si la variable independiente es nominal con más de dos categorías, será necesario crear más de una variable. Por ejemplo, si la variable tiene cuatro categorías, A, B, C y D, será necesario crear tres variables de la siguiente forma:

TABLA 5.1.2. Ejemplo de variables de intervención en el modelo de regresión.

$X_i$	$I_1$	$I_2$	$I_3$
A	1	0	0
B	0	1	0
C	0	0	1
D	0	0	0

Las variables  $I_1$ ,  $I_2$  y  $I_3$  se incluirán en la ecuación de regresión junto con las restantes variables independientes.

**1.2. El comando REGRESSION** Permite realizar análisis de regresión, tanto simple como múltiple, proporcionando diversos métodos y criterios para la construcción de cada ecuación de regresión.

#### Problema-ejemplo

El ejemplo propuesto recoge los resultados (simulados) de 200

alumnos en una prueba de aptitud musical con seis variables X1( tono), X2(intensidad), X3(ritmo), X4(tiempo), X5(timbre), X6(memoria tonal). La escala de medición es de 0 a 100 para cada variable. Además se almacenó la valoración en una prueba de entonación vocal (Y) en una escala de 0 a 100.

Se propone realizar un análisis de regresión lineal con Y como variable dependiente.

## **Desarrollo del ejemplo**

### **1.2.1. Regresión lineal simple**

Para comenzar pensemos en un modelo de regresión lineal simple con Y como variable dependiente y X2 (intensidad) como independiente. Para observar el tipo de relación se dibuja un diagrama de dispersión:

FIGURA 5.1.1. Diagrama de dispersión Y-X2.

Como se observa en el gráfico la nube de puntos parece ajustarse a una línea recta, para encontrar la expresión de dicha función mediante SPSS habrá que seleccionar: **Analizar > Regresión >Lineal** y se accederá al cuadro de diálogo de la fig. 5.1.2.

Se seleccionará como variable **dependiente** Y (entonación vocal) y como **independiente** X2 (intensidad).

Como en otros procedimientos de SPSS, cuando se pulse pegar se añade al fichero de sintaxis. En este caso dicho fichero tomará la expresión:

```
**** Diagrama de dispersión *****.
GRAPH
/SCATTERPLOT(BIVAR)=X2 WITH Y

/MISSING=LISTWISE.
***** Análisis de regresión*****.
REGRESSION

/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT y
/METHOD=ENTER x2 .
```

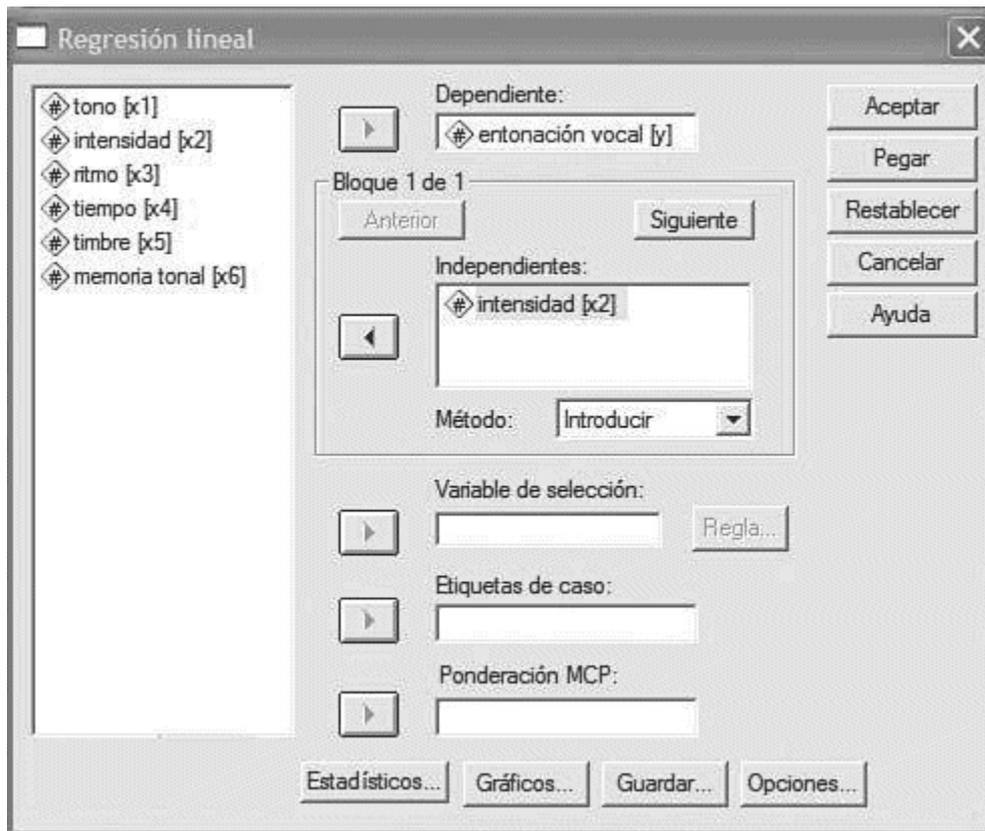


FIGURA 5.1.2. Cuadro de diálogo. Regresión lineal.

Los resultados de ejecutar el procedimiento análisis de regresión, con las opciones dadas por defecto por el programa serán:

TABLLA 5.1.3. Resumen del modelo.

**Modelo**

**R**

**R cuadrado**

**R cuadrado Error típ. de corregida la estimación**

1 ,351(a) ,123 ,118 7,31940

a: Variables predictoras: (Constante), intensidad

Donde R es el *coeficiente de correlación* y R cuadrado el *coeficiente de determinación*. El valor de éste último permite afirmar que de la variación de la variable dependiente el 12,3% se puede explicar por la variable independiente X2. El error típico de estimación es  $S_{e\prime}$ , es decir, la raíz cuadrada de la media cuadrática residual, según se puede comprobar en la tabla 5.1.4.

TABLLA 5.1.4. ANOVA(b).

**Modelo** Suma de gl Media F Sig. **cuadrados cuadrática**

1 Regresión Residual Total 1485,862 1 1485,862 27,735 ,000(a)

10607,587 198 53,574

12093,449 199

a: Variables predictoras: (Constante), intensidad b: Variable dependiente: entonación vocal

En la tabla ANOVA del modelo se contrasta la hipótesis R poblacional igual a cero, lo que en el modelo de regresión simple representa que la pendiente de la recta es igual a cero. Como se rechaza la hipótesis nula esto significa que las variables X2 e Y están relacionadas.

Finalmente, como resultado se muestra el estudio de los coeficientes de la recta de regresión. La tabla 5.1.5 presenta dicha salida:

TABLLA 5.1.5. Coeficientes (a).

**Coeficientes Coeficientes no estandarizados estandarizados**

**Modelo B Error típ. Beta t Sig.** 1 (Constante) 59,291 2,487 23,839 ,000  
intensidad ,332 ,063 ,351 5,266 ,000

a: Variable dependiente: entonación vocal

Los *coeficientes no estandarizados* permiten construir la recta de regresión:  $Y=59,291+0,332 X2$

Los *coeficientes de regresión estandarizados* (b) toman la expresión:  $b_1=b_1(S_x/S_y)$ . En el caso de la regresión lineal simple dicho coeficiente coincide con el coeficiente de correlación (R). En regresión lineal múltiple será un indicador de la importancia de cada variable en el modelo.

En la tabla 5.1.5 también se realiza un estudio de la significación de cada coeficiente de regresión. Se plantea la hipótesis nula, donde cada coeficiente es igual a cero, frente a la alternativa, distinto de cero. El valor de t, como se ha dicho anteriormente, será:

t

=

$B_i S_{Bi}$

donde t se distribuye como una t de Student de n-(k+1) grados de libertad. En el ejemplo los dos coeficientes resultan significativos.

**1.2.2. Regresión lineal múltiple**

Con los datos del problema-ejemplo, se supone Y como variable dependiente y las restantes seis variables X1 a X6 como independientes. Para realizar el análisis de regresión será necesario seleccionar: **Analizar > Regresión >Lineal** y se accederá al cuadro de diálogo de la fig. 5.1.2. En este caso se seleccionan como independientes y dependiente las variables indicadas. Además se utilizará el método de inclusión por pasos de las variables independientes, por lo que al seleccionar opciones en el cuadro

de diálogo de la fig. 5.1.2, se entrará en el cuadro de diálogo de la fig. 5.1.3.

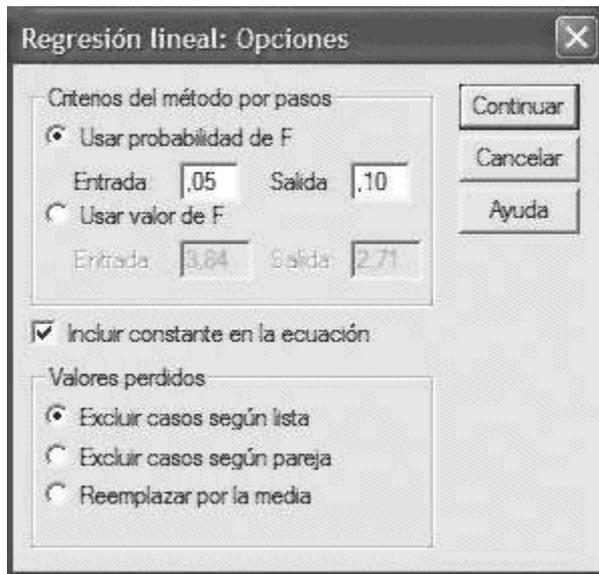


FIGURA 5.1.3. Cuadro de diálogo. Regresión lineal: Opciones.

donde se pueden seleccionar valores del estadístico F para inclusión de las variables en el modelo, el tratamiento de valores perdidos y la inclusión o no de la constante en el modelo (valor  $b_0$ ).

Con las opciones por defecto del programa SPSS y con la selección del método de inclusión por pasos para las variables independientes, el fichero de sintaxis será:

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA /CRITERIA=PIN(.05)
POUT(.10)
/NOORIGIN
/DEPENDENT y
/METHOD=STEPWISE x1 x2 x3 x4 x5 x6 .
```

Ejecutado dará las siguientes salidas:

TABLA 5.1.6. Resumen del modelo.

**Modelo**

**R**

**R cuadrado**

**R cuadrado corregida**

**Error típ. de la estimación**

1 ,745(a) ,555 ,553 5,21225

2 ,866(b) ,751 ,748 3,91256  
 3 ,947(c) ,897 ,895 2,52276  
 4 ,986(d) ,973 ,972 1,29931  
 5 ,990(e) ,981 ,980 1,09428  
 6 ,992(f) ,984 ,983 1,01506

- a: Variables predictoras: (Constante), memoria tonal.
- b: Variables predictoras: (Constante), memoria tonal, timbre.
- c: Variables predictoras: (Constante), memoria tonal, timbre, intensidad.
- d: Variables predictoras: (Constante), memoria tonal, timbre, intensidad, tono.
- e: Variables predictoras: (Constante), memoria tonal, timbre, intensidad, tono, ritmo.
- f: Variables predictoras: (Constante), memoria tonal, timbre, intensidad, tono, ritmo, tiempo.

Como se puede ver el orden de entrada de las variables en el modelo es: memoria tonal, timbre, intensidad, tono, ritmo y tiempo. Es el modelo 6, formado por éstas variables, quien mejor explica la variación total. El grado de explicación del modelo 6 es del 98,40% (R cuadrado).

Respecto a los **estadísticos** pulsando el botón correspondiente tenemos:

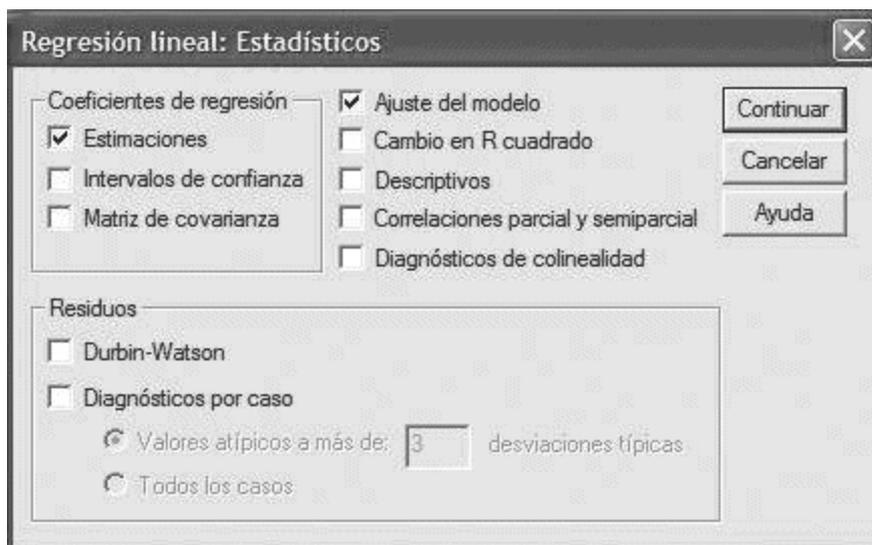


FIGURA 5.1.4.

Cuadro de diálogo. Regresión lineal. Estadísticos.

TABLA 5.1.7. ANOVA del modelo(g)  
**Modelo** Suma de gl Media F Sig. **cuadrados cuadrática**

1 Regresión Residual Total  
 6714,276 1 6714,276 247,143 ,000(a) 5379,173 198 27,168  
 12093,449 199

2 Regresión Residual Total  
 9077,750 2<sub>4538,875</sub> 296,501 ,000(b)

3015,699 197 15,308<sup>1</sup>12093,449 199

### 3 Regresión Residual Total

10846,046 3 3615,349 568,067 ,000(c)

1247,403 196 6,364

12093,449 199

### 4 Regresión Residual Total

11764,248 4 2941,062<sub>1742,116</sub> ,000(d)

329,201 195 1,688

12093,449 199

### 5 Regresión Residual Total

11861,144 5 2372,229 1981,072 ,000(e)

232,305 194 1,197

12093,449 199

### 6 Regresión Residual Total

11894,593 6 1982,432 1924,051 ,000(f)

198,856 193 1,030

12093,449 199

a: Variables predictoras: (Constante), memoria tonal.

b: Variables predictoras: (Constante), memoria tonal, timbre.

c: Variables predictoras: (Constante), memoria tonal, timbre, intensidad.

d: Variables predictoras: (Constante), memoria tonal, timbre, intensidad, tono. e: Variables predictoras: (Constante), memoria tonal, timbre, intensidad, tono, ritmo. f: Variables predictoras: (Constante), memoria tonal, timbre, intensidad, tono, ritmo, tiempo. g: Variable dependiente: entonación vocal.

Si se contrasta la hipótesis  $R=0$  del coeficiente de correlación múltiple por la tabla 5.1.7 se puede afirmar que hay diferencias significativas, y por tanto existe relación entre las variables X1 a X6 con Y.

Respecto a los coeficientes de regresión, la tabla 5.1.8, presentada a continuación, permite expresar la relación entre las variables independientes respecto a la dependiente.

TABLA 5.1.8. Coeficientes(a)

**Coeficientes no estandarizados estandarizados Modelo B Error típ. Beta t Sig.**

1 (Constante) memoria tonal

53,432 1,244 42,968 ,000

,976 ,062 ,745 15,721 ,000

2 (Constante) memoria tonal timbre 30,796 2,047 15,045 ,000

,958 ,047 ,731 20,544 ,000 ,671 ,054 ,442 12,426 ,000

3 (Constante) memoria tonal timbre intensidad 16,918 1,561 10,841 ,000

,991 ,030 ,756 32,882 ,000

,649 ,035 ,428 18,616 ,000

,363 ,022 ,383 16,669 ,000

4 (Constante) memoria tonal timbre intensidad tono

6,787 ,914 7,428 ,000 1,029 ,016 ,785 65,932 ,000

,636 ,018 ,420 35,460 ,000

,382 ,011 ,403 33,928 ,000

,305 ,013 ,278 23,321 ,000

5 (Constante) memoria tonal timbre intensidad tono

ritmo

2,427 ,909 2,669 ,008 1,019 ,013 ,778 77,253 ,000

,624 ,015 ,411 41,070 ,000

,383 ,009 ,405 40,465 ,000

,315 ,011 ,287 28,475 ,000

,184 ,020 ,091 8,996 ,000

6 (Constante) memoria tonal timbre intensidad tono

ritmo

tiempo

-1,872 1,132 -1,654 ,100

1,014 ,012 ,774 82,672 ,000

,630 ,014 ,415 44,591 ,000

,390 ,009 ,412 43,994 ,000

,315 ,010 ,287 30,697 ,000

,193 ,019 ,095 10,142 ,000

,118 ,021 ,054 5,698 ,000

a: Variable dependiente: entonación vocal.

Los coeficientes estandarizados, basados en las puntuaciones tipificadas de las variables, son directamente comparables entre si. Dan pistas sobre la importancia relativa de cada variable en el modelo. Cuanto mayor sean (en valor absoluto) mayor importancia tiene la variable en el

modelo.

La ecuación de regresión tomará la expresión:

$$Y = -1,872 + 1,014 X_6 + 0,63 X_5 + 0,390 X_2 + 0,315 X_1 + 0,193 X_3 + 0,118 X_4$$

#### 1.2.2.1. Estadísticos adicionales

Además de todos los resultados dados por defecto por el programa SPSS, al entrar en el cuadro de diálogo de la fig. 5.1.2 y pulsar **Estadísticos** se pueden seleccionar (ver fig. 5.1.4):

a) Estadísticos relacionados con los coeficientes de regresión:

- Estimaciones (dado por defecto).
- Intervalos de confianza.
- Matriz de covarianzas.

b) Otros estadísticos:

- Ajuste del modelo (dado por defecto).
- Cambio en R cuadrado.
- Descriptivos.
- Correlaciones parciales y semiparciales.

c) Estadísticos relacionados con los supuestos del modelo:

- Diagnóstico de colinealidad.
- Residuos:

- Durbin-Watson.

- Diagnóstico por caso:

Valores atípicos a más de 3 desviaciones típicas.

Todos los casos.

Como ya se han visto las salidas dadas por defecto, a continuación se detallarán los estadísticos incluidos en los apartados a) y b) anteriores. Posteriormente, en otro apartado, se entrará en los indicados en el punto c).

TABLA 5.1.9. Estadísticos descriptivos.

<b>Media</b>	<b>Desviación típ.</b>	<b>N</b>	entonación vocal	72,1023	7,79558	200
			tono	29,86	7,102	200
			intensidad	38,60	8,232	200
			ritmo	25,03	3,842	20
			tiempo	31,10	3,532	200
			timbre	34,26	5,140	200
			memoria tonal	19,13	5,951	200

En los estadísticos descriptivos se incluyen la media, desviación típica

y el número de casos. Los valores, para el ejemplo, se muestran en la tabla 5.1.9. En la tabla 5.1.10 se muestran las correlaciones, su significación y el número de casos incluido en el análisis.

TABLA 5.1.10. Correlaciones.

entonación      tono intensidad ritmo tiempo timbre memoria vocal tonal

Correlación de Pearson entonación vocal

tono

intensidad ritmo

tiempo

timbre

memoria tonal

1,000 ,184 ,351 ,168 ,006 ,465 ,745

,184 1,000 -,063 -,109 ,009 ,023 -,099

,351 -,063 1,000 -,016 -,139 ,036 -,064

,168 -,109 -,016 1,000 -,082 ,093 ,098

,006 ,009 -,139 -,082 1,000 -,091 ,068

,465 ,465 ,023 ,093 -,091 1,000 ,031

,745 ,745 -0,99 ,098 ,068 ,031 1,000

Sig.

(unilateral) entonación vocal

tono

intensidad ritmo

tiempo

timbre

memoria tonal

. ,005 ,000 ,009 ,467 ,000 ,000

,005 . ,189 ,063 ,451 ,375 ,081

,000 ,189 . ,409 ,025 ,305 ,184

,009 ,063 ,409 . ,123 ,095 ,083

,467 ,451 ,025 ,123 . ,100 ,170

,000 ,375 ,305 ,095 ,100 . ,330

,000 ,081 ,184 ,083 ,170 ,330 .

N entonación

vocal

tono

intensidad ritmo

tiempo

timbre

memoria tonal

200 200 200 200 200 200 200 200 200 200 200 200 200 200 200 200 200 200  
200 200 200 200 200 200 200 200 200 200 200 200 200 200 200 200 200 200  
200 200 200 200 200 200 200 200

200 200 200 200 200 200 200

En el resumen del modelo (tabla 5.1.11) se muestra las variaciones de R cuadrado, con su equivalencia en valores de F, para los diferentes modelos. Recuérdese que en el ejemplo el modelo 6 era aquel que incluía todas las variables independientes. Se observa una disminución inapreciable para los modelos 5 y 6, reafirmando el poco poder explicativo que aportan las variables X3 (ritmo) y X4 (tiempo).

TABLA 5.1.11. Resumen del modelo.

### Modelo Estadísticos de cambio

#### Cambio en R cuadrado Cambio en F gl1 gl2 Sig. del cambio en F

1	,555	247,143	1	198	,000
2	,195	154,394	1	197	,000
3	,146	277,846	1	196	,000
4	,076	543,890	1	195	,000
5	,008	80,919	1	194	,000
6	,003	32,464	1	193	,000

a: Variables predictoras: (Constante), memoria tonal.

b: Variables predictoras: (Constante), memoria tonal, timbre.

c: Variables predictoras: (Constante), memoria tonal, timbre, intensidad.

d: Variables predictoras: (Constante), memoria tonal, timbre, intensidad, tono. e: Variables predictoras: (Constante), memoria tonal, timbre, intensidad, tono, ritmo. f: Variables predictoras: (Constante), memoria tonal, timbre, intensidad, tono, ritmo, tiempo.

La tabla 5.1.12 nos habla de los coeficientes de regresión, su intervalo de confianza, y de las correlaciones. El intervalo de confianza resulta de sumar y restar al coeficiente de regresión 1,96 por el error típico. Así por ejemplo, para la variable timbre en el modelo 6 se obtendrá  $0,630 \pm 1,96 \times 0,014 = (0,658; 0,602)$ .

### Modelo

TABLA 5.1.12. Coeficientes(a).

#### Intervalo de confianza para B al 95% Correlaciones

#### Límite inferior Límite superior Orden cero Parcial Semiparcial

1 (Constante) memoria tonal 50,979 ,854 55,884 1,099 ,745 ,745 ,745

2 (Constante) memoria tonal timbre

26,759

,866 ,564

34,832

1,050

,777

,745 ,465 ,826 ,731 ,663 ,442

**Modelo Intervalo de confianza para B al 95% Correlaciones**

**Límite inferior Límite superior Orden cero Parcial Semiparcial**

3 (Constante) memoria tonal timbre  
intensidad 13,840

,931

,580

,320

19,995 1,050 ,717 ,406 ,745 ,920 ,754 ,465 ,799 ,427 ,351 ,766 ,382

4 (Constante) memoria tonal timbre  
intensidad tono

4,985

,998

,601

,359

,279

8,588 1,060 ,672 ,404 ,331 ,745 ,978 ,779 ,465 ,930 ,419 ,351 ,925 ,401  
,184 ,858 ,276

5 (Constante) memoria tonal timbre  
intensidad tono

ritmo

,634 ,993 ,594 ,365 ,293 ,144 4,221 1,045

,654

,402

,337

,225

,745 ,984 ,769 ,465 ,947 ,409 ,351 ,946 ,403 ,184 ,898 ,283 ,168 ,543 ,090

6 (Constante) memoria tonal timbre  
intensidad tono

ritmo

tiempo

-4,104 ,990  
,602  
,373  
,295  
,156  
,077  
,360  
1,038  
,658  
,408  
,335  
,231  
,159  
,745 ,986 ,763 ,465 ,955 ,412 ,351 ,954 ,406 ,184 ,911 ,283 ,168 ,590 ,094  
,006 ,379 ,053

a: Variable dependiente: entonación vocal.

La correlación de orden cero indica la correlación Pearson entre cada variable independiente con la dependiente. La correlación parcial indica el grado de relación entre la variable independiente y la dependiente aislando el efecto del resto de las variables en ambas. La correlación semiparcial indica el grado de relación existente entre la variable dependiente y la parte de cada variable independiente que no está explicada por el resto de variables independientes.

Las correlaciones de los coeficientes de regresión dadas en la tabla 5.1.13, muestran como los mismos no son independientes, sino que en algunos casos tienen una correlación apreciable (tiempo e intensidad (0,134)).

TABLA 5.1.13. Correlaciones de los coeficientes(a).

**Modelo**

1 Correlaciones Covarianzas 2 Correlaciones

Covarianzas

3 Correlaciones

Covarianzas

4 Correlaciones

Covarianzas

5 Correlaciones

Covarianzas

6 Correlaciones

Covarianzas memoria tonal memoria tonal

memoria tonal timbre  
memoria tonal timbre

memoria tonal timbre  
intensidad  
memoria tonal timbre  
intensidad

memoria tonal timbre  
intensidad  
tono  
memoria tonal timbre  
intensidad  
tono

memoria tonal timbre  
intensidad  
tono  
ritmo  
memoria tonal timbre  
intensidad  
tono  
ritmo

memoria tonal timbre  
intensidad  
tono  
ritmo  
tiempo  
memoria tonal timbre  
intensidad  
tono  
ritmo  
tiempo

**memoria timbre intensidad tono ritmo tiempo<sub>tonal</sub>**

1,000  
,004

1,000 -,031  
-,031 1,000  
,002 -7,874E-05

-7,874E-05 ,003

1,000 -,034 ,065  
-,034 1,000 -,038  
,065 -,038 1,000  
,001 -3,537E-05 4,286E-05  
-3,537E-05 ,001 -2,917E-05  
4,286E-05 -2,917E-05 ,000

1,000 -,037 ,072 ,104  
-,037 1,000 -,040 -,029  
,072 -,040 1,000 ,071  
,104 -,029 ,071 1,000  
,000 -1,022E-05 1,266E-05 2,131E-05  
-1,022E-05 ,000 -8,145E-06 -6,740E-06  
1,266E-05 -8,145E-06 ,000 1,037E-05  
2,131E-05 -6,740E-06 1,037E-05 ,000

1,000 -,028 ,070 ,095 -,084  
-,028 1,000 -,042 -,038 -,094  
,070 -,042 1,000 ,072 ,021  
,095 -,038 ,072 1,000 ,104  
-,084 -,094 ,021 ,104 1,000  
,000 -5,666E-06 8,760E-06 1,384E-05 -2,270E-05  
-5,666E-06 ,000 -6,063E-06 -6,424E-06 -2,930E-05  
8,760E-06 -6,063E-06 8,978E-05 7,582E-06 4,082E-06  
1,384E-05 -6,424E-06 7,582E-06 ,000 2,353E-05  
-2,270E-05 -2,930E-05 4,082E-06 2,353E-05 ,000

1,000 -,034 ,060 ,095 -,089 -,071  
-,034 1,000 -,031 -,038 -,087 ,081  
,060 -,031 1,000 ,072 ,032 ,134  
,095 -,038 ,072 1,000 ,103 ,000  
-,089 -,087 ,032 ,103 1,000 ,084  
-,071 ,081 ,134 ,000 ,084 1,000  
,000 -5,868E-06 6,512E-06 1,191E-05 -2,092E-05 -1,797E-05  
-5,868E-06 ,000 -3,857E-06 -5,529E-06 -2,337E-05 2,382E-05  
6,512E-06 -3,857E-06 7,865E-05 6,522E-06 5,413E-06 2,462E-05  
1,191E-05 -5,529E-06 6,522E-06 ,000 2,025E-05 -3,638E-08  
-2,092E-05 -2,337E-05 5,413E-06 2,025E-05 ,000 3,328E-05  
-1,797E-05 2,382E-05 2,462E-05 -3,638E-08 3,328E-05 ,000

a: Variable dependiente: entonación vocal.

### 1.2.2.2. Estudio de los supuestos del modelo de regresión

Mediante SPSS se puede estudiar el cumplimiento de los supuestos del modelo de regresión lineal múltiple: linealidad, independencia, homocedasticidad, normalidad y multicolinealidad. Para ensayar la mayoría de estos supuestos se necesita el conocimiento de los **residuos**.

Se entiende por **residuo**, en regresión lineal, la diferencia entre los valores observados en la variable dependiente  $y_i$  y sus correspondientes valores pronosticados  $y_i^*$ . No debe confundirse el residuo ( $e_i$ ) con el **error de predicción**, que representa la diferencia entre el verdadero valor de  $y_i$  en la población y su correspondiente valor estimado mediante la ecuación de regresión.

Hay diferentes tipos de residuos, empleados en paquetes estadísticos como SPSS:

a) **Residuos brutos**. Dependen de la unidad de medición de la variable dependiente y son los definidos anteriormente:  $e_i = y_i - y_i^*$ .

b) **Residuos estandarizados** (ZRESID —en el programa SPSS—).

$$e_i \text{ZRESID} = \frac{e_i}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n e_i^2}}$$

$\hat{\sigma}_e$

$e$

$2$

$i$

$i=1$

$$nk - 1$$

Sus valores oscilan desde  $-3$  a  $+3$  aproximadamente. Naturalmente tienen media 0 y desviación típica 1.

c) **Residuos estudentizados** (SRESID —en el programa SPSS—). Su expresión es:

$$\text{SRESID} = \frac{e_i}{\sqrt{1 - d_{ii}}}$$

$$e_i \text{ donde } e_i \text{ es el residuo bruto, } \hat{\sigma}_e \text{ la desviación típica estimada}$$

de la desviación típica del residuo bruto y  $d_{ii}$  la distancia entre el punto  $i$  y el punto medio. Los residuos estudentizados se distribuyen según una  $t$  de Student

con  $n-k-1$  grados de libertad.

d) **Residuos eliminado estudentizado** (SDRESID —en el programa SPSS—). Con una expresión similar a la anterior, sólo que en esta ocasión no se incluye el  $i$ ésimo residuo ( $-i$ ) al calcular la desviación típica estimada.

$$e_i \text{SDRESID}_{(-i)} = \frac{e_i}{\sqrt{1 - d_{ii}}}$$

s

1-d

0iii

Se distribuye según una  $t$  de Student con  $n-k-2$  grados de libertad. Son muy interesantes para detectar puntos de influencia (casos con gran peso en la ecuación de regresión).

**Linealidad.** La primera aproximación al estudio del cumplimiento del primer supuesto se puede realizar mediante el examen gráfico de los diagramas de dispersión de la variable dependiente y cada una de las independientes. La aproximación de la nube de puntos a una recta dará idea del grado de cumplimiento del supuesto. También aportará información en este sentido la representación gráfica (mediante un gráfico de dispersión) de los residuos tipificados en función de los valores pronosticados tipificados para la variable dependiente. Valores de sujetos superiores a 2; 2,5 o 3 desviaciones típicas serán indicadores de casos perturbadores, influyentes o extraños.

Para realizar el examen gráfico de los diagramas de dispersión se seleccionará en el cuadro de diálogo de la fig. 5.1.2 tal opción entrando en el cuadro de la fig. 5.1.5 aquí se marcará **Generar todos los gráficos parciales**. El resultado de tal selección se muestra en las fig. 5.1.6 al 5.1.11.

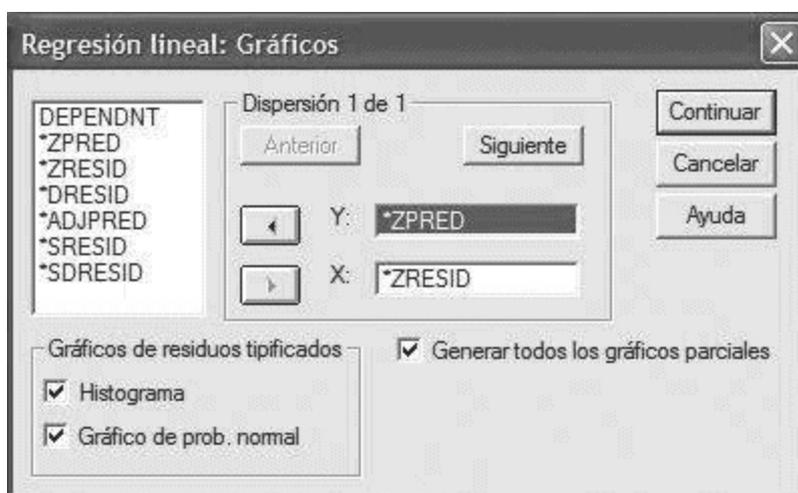


FIGURA 5.1.5. Cuadro de

diálogo. Regresión lineal. Gráficos. FIGURA 5.1.6. Gráfico de regresión parcial: tono. FIGURA 5.1.7. Gráfico de regresión parcial: intensidad. FIGURA 5.1.8. Gráfico de regresión parcial: ritmo. FIGURA 5.1.9. Gráfico de regresión parcial: tiempo. FIGURA 5.1.10. Gráfico de regresión parcial: timbre. FIGURA 5.1.11. Gráfico de regresión parcial: memoria tonal.

Como se observa en los seis gráficos anteriores, la linealidad resulta patente en todas las variables independientes excepto en tiempo y ritmo.

La otra forma de comprobar la linealidad, el gráfico de los valores pronosticados en función de los residuos tipificados para la variable dependiente, será: FIGURA 5.1.12. Gráfico de dispersión: residuo tipificado-valor pronosticado tipificado.

Como se puede observar en el gráfico precedente no existe ninguna tendencia en la distribución de los residuos, pues los mismos se distribuyen de forma aleatoria en torno a una recta horizontal que pasa por el origen.

**Independencia.** El supuesto de independencia de los residuos se muestra especialmente relevante cuando los datos han sido tomados de forma temporal. El estadístico de Durbin-Watson proporciona información sobre el grado de cumplimiento:

$$DW = \frac{\sum_{i=2}^n \hat{e}_i^2 - 2 \sum_{i=1}^{n-1} \hat{e}_i \hat{e}_{i+1}}{\sum_{i=1}^n \hat{e}_i^2}$$

Para calcular dicho estadístico será necesario seleccionar el mismo en el cuadro de diálogo de la fig. 5.1.2.

Los valores del estadístico de Durbin-Watson oscilan entre 0 y 4, y valores entre 1,5 y 2,5 (según algunos autores) indica independencia de los residuos.

TABLA 5.1.14. Resumen del modelo(g).

**Modelo R R cuadrado R cuadrado Error típ. de Durbin-Watson corregida la estimación**

1	,745(a)	,555	,553	5,21225
2	,866(b)	,751	,748	3,91256
3	,947(c)	,897	,895	2,52276

4 ,986(d) ,973 ,972 1,29931  
5 ,990(e) ,981 ,980 1,09428  
6 ,992(f) ,984 ,983 1,01506 1,979

a: Variables predictoras: (Constante), memoria tonal.  
b: Variables predictoras: (Constante), memoria tonal, timbre.  
c: Variables predictoras: (Constante), memoria tonal, timbre, intensidad.  
d: Variables predictoras: (Constante), memoria tonal, timbre, intensidad, tono. e: Variables predictoras: (Constante), memoria tonal, timbre, intensidad, tono, ritmo. f: Variables predictoras: (Constante), memoria tonal, timbre, intensidad, tono, ritmo, tiempo. g: Variable dependiente: entonación vocal.

En el ejemplo propuesto el valor obtenido (1,979) del estadístico de DurbinWatson nos permite afirmar la independencia de los datos.

**Homocedasticidad.** Para estudiar la homocedasticidad y la normalidad nos podemos ayudar de representaciones gráficas de los residuos, de hecho para la linealidad ya las hemos utilizado. Para realizar tales gráficas se seleccionará en el cuadro de diálogo de la fig. 5.1.5.

La homocedasticidad indicará la igualdad de variación de los residuos en todo el rango de valores pronosticados. O de forma semejante que los residuos no presenten ninguna pauta de relación entre los valores pronosticados y los residuos, ambos tipificados. En la fig. 5.1.5 se ha seleccionado justo esta opción.

Los resultados para el ejemplo propuesto se muestran en fig. 5.1.13. Según se puede observar no existe ninguna pauta de relación, los residuos se distribuyen aleatoriamente en una franja de dos líneas paralelas horizontales que parten de +2 y -2.

FIGURA 5.1.13. Gráfico de dispersión: valor pronosticado tipificado-residuo tipificado.

Cuando por el gráfico de dispersión anterior se detecta una falta de homocedasticidad se suelen emplear las transformaciones de la variable dependiente, que afectan también a la normalidad de la distribución de los residuos, de hecho la heterocedasticidad se da con mayor frecuencia cuando:

- a) Se incumple el supuesto de normalidad.
- b) Algunas de las variables son asimétricas mientras que otras no lo son.

El gráfico de dispersión también es útil para detectar el tipo de relación entre las variables, porque su forma da idea del tipo de relación que las ligan.

Además de la forma gráfica, para observar la homocedasticidad también se puede utilizar el test de Levene, donde se compara por separado la variabilidad de la variable dependiente en los distintos valores de las variables independientes. Para cada caso, se calcula la diferencia entre el valor de dicho caso y la media de su casilla y se lleva a cabo un análisis de varianza de un factor sobre estas diferencias. En el ejemplo no se puede realizar dicha prueba porque se necesita que las variables independientes sean categóricas y no continuas como sucede en este caso.

**Normalidad.** Para cumplir esta condición, según se ha expuesto anteriormente, se debe dar que la distribución de los residuos tipificados sea normal. El cuadro de diálogo de la fig. 5.1.5 tiene dos representaciones gráficas para esta opción: *el histograma y el gráfico de probabilidad normal*.

El histograma de los residuos tipificados presenta dicho gráfico con una curva normal superpuesta. En el ejemplo los resultados se muestran en la fig. 5.1.14. Como se puede apreciar existe un buen ajuste a la curva normal.

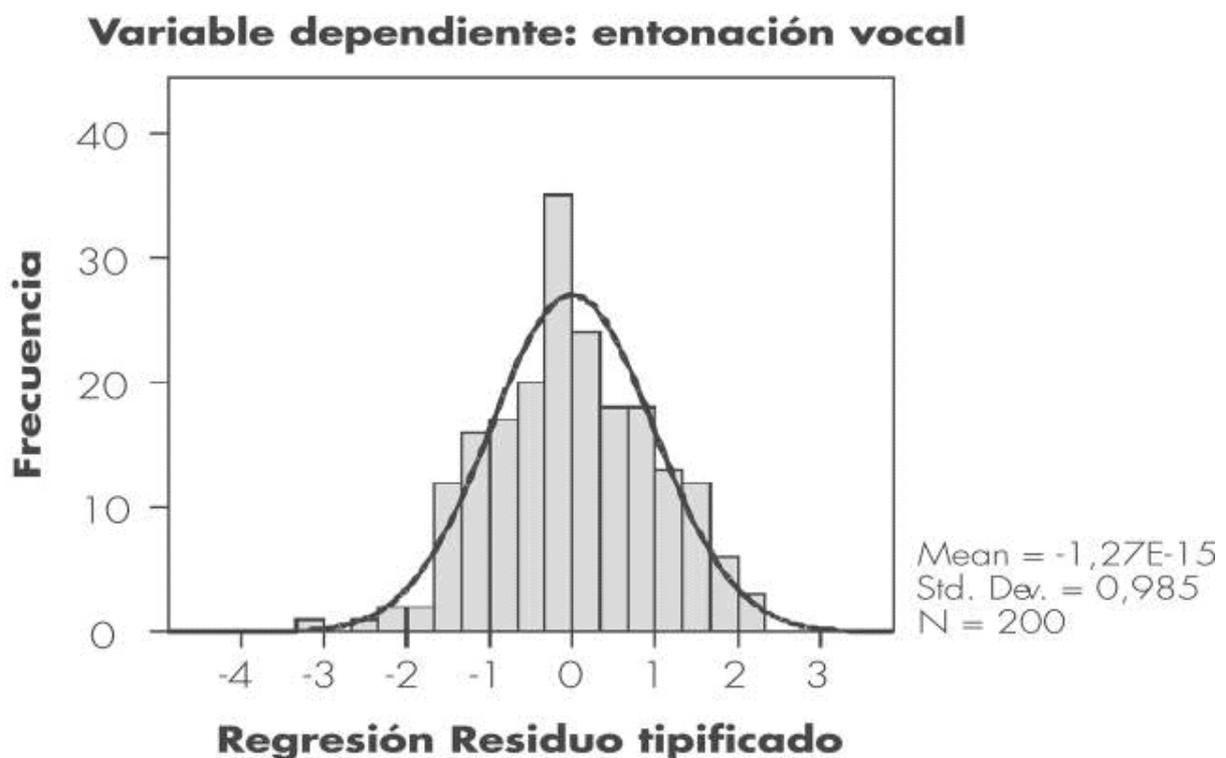


FIGURA 5.1.14. Histograma de residuos tipificados. Comprobación de normalidad.

El gráfico de probabilidad normal (ver fig. 5.1.15) presenta en el eje de abscisas la probabilidad acumulada que corresponde a cada residuo tipificado. En el eje de ordenadas se representan las probabilidades teóricas acumuladas. El gráfico resultante, si los residuos tipificados tienen una distribución normal, será una línea recta bisectriz del primer cuadrante. Cuando la línea de puntos cae por debajo de la diagonal la distribución es *platicúrtica* y con *asimetría negativa*, si tiene un arco sencillo, es decir, es achatada o plana y tiene mayor acumulación de punto en valores superiores a la media. Si la línea de puntos cae por encima de la diagonal, la distribución es *leptocúrtica* y con *asimetría positiva*, si tiene un arco sencillo, o sea, es puntiaguda y con valores inferiores a la media. Como se observa en la fig. 5.1.15 la gráfica del ejemplo confirma la distribución normal de los residuos tipificados.

FIGURA 5.1.15. Gráfico de prob. normal. Comprobación de normalidad.

**Colinealidad.** Como se ha comentado la colinealidad se da cuando existe relación entre las variables independientes. El mayor problema de su presencia, cuando la colinealidad es perfecta, es la imposibilidad de calcular los coeficientes de regresión al ser la matriz de correlaciones singular y en consecuencia no existir una solución única mínimo-cuadrática. Cuando la colinealidad es parcial aumentan el valor de los residuos tipificados y estos generan unos coeficientes de regresión inestables ante variaciones en los datos, de tal forma que una simple variación en un dato puede ocasionar el cambio de signo de algún coeficiente. Entonces el coeficiente de correlación múltiple R, no sufre variación.

Resulta difícil evaluar cual es el grado de colinealidad permitida para que no se produzcan los hechos anteriores. Algunos indicadores son:

a) Valores de tolerancia pequeños (0,01 por ejemplo). Indicador de que esa variable comparte el 99% de su varianza con el resto de las variables independientes y por tanto resulta superflua.

b) Coeficientes de correlación estimados superiores a 0,9 en valor absoluto. c) Coeficientes de regresión parcial estandarizados (b) con valores superiores a 1 o inferiores a -1.

d) F del coeficiente de correlación múltiple significativo, pero no los coeficientes de regresión individuales de las variables muy colineales. Además SPSS en el cuadro de diálogo de la fig. 5.1.2 se puede seleccionar la opción diagnósticos de colinealidad. Los resultados para el ejemplo son:

TABLA 5.1.15. Coeficientes(a).

**Modelo**   **Coefficientes**   **no**   **estandarizados**   **Coefficientes<sub>t</sub>**   **Sig.**  
**Estadísticos de**

**estandarizados colinealidad**

**B**

1 (Constante) memoria tonal

2 (Constante) memoria tonal timbre

53,432

,976

30,796

,958

,671

3 (Constante) memoria tonal timbre  
intensidad 16,918

,991

,649

,363

4 (Constante) memoria tonal timbre  
intensidad tono

6,787 1,029

,636

,382

,305

5 (Constante) memoria tonal timbre  
intensidad tono

ritmo

2,427 1,019

,624

,383

,315

,184

**Error típ.**

1,244 ,062

2,047

,047

,054

1,561  
,030  
,035  
,022

,914  
,016  
,018  
,011  
,013

,909  
,013  
,015  
,009  
,011  
,020

**Beta Tolerancia FIV**

42,968 ,000  
,745 15,721 ,000 1,000 1,000

15,045 ,000  
,731 20,544 ,000 ,999 1,001 ,442 12,426 ,000 ,999 1,001

10,841 ,000  
,756 32,882 ,000 ,995 1,005 ,428 18,616 ,000 ,998 1,002 ,383 16,669 ,000  
,994 1,006

7,428 ,000  
,785 65,932 ,000 ,984 1,016 ,420 35,460 ,000 ,997 1,003 ,403 33,928 ,000  
,989 1,011 ,278 23,321 ,000 ,985 1,016

2,669 ,008  
,778 77,253 ,000 ,977 1,024 ,411 41,070 ,000 ,988 1,012 ,405 40,465 ,000  
,989 1,011 ,287 28,475 ,000 ,974 1,027 ,091 8,996 ,000 ,971 1,029

6 (Constante) memoria tonal timbre  
intensidad tono  
ritmo  
tiempo  
-1,872 1,014

,630  
,390

,315  
 ,193  
 ,118  
 1,132 ,012 ,014 ,009 ,010 ,019 ,021  
 -1,654 ,100  
 ,774 82,672 ,000 ,972 1,029 ,415 44,591 ,000 ,981 1,019 ,412 43,994 ,000  
 ,971 1,029 ,287 30,697 ,000 ,974 1,027 ,095 10,142 ,000 ,965 1,037 ,054  
 5,698 ,000 ,962 1,039

a: Variable dependiente: entonación vocal.

Según se puede ver en el paso 6 del procedimiento, los valores de la tolerancia son próximos a 1, valor extremo. Un valor de tolerancia inferior a 0,2 es indicador de multicolinealidad, en consecuencia como los valores son mayores de 0,9 se supone ausencia de colinealidad.

En la tabla 5.1.15 también aparece el FIV (factor de inflación de la varianza) definido como:  $FIV_i = 1/Tol_i$ , en consecuencia para que no exista multicolinealidad interesa FIV próximos a 1 y será indicador de colinealidad valores elevados. Cuanto mayor es el FIV de una variable mayor es la varianza del correspondiente coeficiente de regresión.

TABLA 5.1.16. Diagnósticos de colinealidad(a).

**Proporciones de la varianza**

**Modelo** **Dimensión** **Autovalor** **Índice de** **(Constante)** **memoria** **timbre**  
**intensidad** **tono** **ritmo** **tiempo** **condición tonal**

11 1,955 1,000 ,02 ,02  
 2 ,045 6,596 ,98 ,98  
 21 2,927 1,000 ,00 ,01 ,00  
 2 ,063 6,829 ,03 ,93 ,08  
 3 ,010 16,724 ,97 ,06 ,92  
 31 3,884 1,000 ,00 ,01 ,00 ,00  
 2 ,076 7,143 ,00 ,82 ,01 ,12  
 3 ,031 11,151 ,02 ,10 ,29 ,68  
 4 ,009 20,814 ,97 ,08 ,69 ,20  
 41 4,830 1,000 ,00 ,00 ,00 ,00 ,00  
 2 ,084 7,596 ,00 ,78 ,00 ,03 ,11  
 3 ,051 9,718 ,00 ,03 ,00 ,39 ,51  
 4 ,027 13,277 ,01 ,09 ,47 ,37 ,20  
 5 ,008 25,101 ,98 ,10 ,52 ,21 ,18  
 51 5,806 1,000 ,00 ,00 ,00 ,00 ,00 ,00  
 2 ,084 8,325 ,00 ,78 ,00 ,03 ,10 ,00  
 3 ,052 10,581 ,00 ,04 ,00 ,31 ,56 ,01  
 4 ,033 13,253 ,00 ,12 ,10 ,49 ,13 ,22  
 5 ,020 17,085 ,00 ,00 ,64 ,01 ,02 ,45  
 6 ,006 31,565 ,99 ,05 ,25 ,17 ,18 ,32

61 6,789 1,000 ,00 ,00 ,00 ,00 ,00 ,00 ,00  
 2 ,084 8,989 ,00 ,79 ,00 ,02 ,09 ,00 ,00  
 3 ,052 11,441 ,00 ,04 ,00 ,31 ,56 ,01 ,00  
 4 ,034 14,103 ,00 ,13 ,07 ,50 ,17 ,16 ,02  
 5 ,020 18,468 ,00 ,00 ,68 ,01 ,01 ,39 ,00  
 6 ,018 19,281 ,01 ,02 ,09 ,01 ,07 ,25 ,43  
 7 ,003 45,990 ,99 ,01 ,15 ,15 ,09 ,19 ,54

a: Variable dependiente: entonación vocal

Otro procedimiento de detección de multicolinealidad se refiere al estudio de los **autovalores** de la matriz de productos cruzados de las variables independientes. Si algún autovalor se aproxima a 0 será indicador de colinealidad.

El **índice de condición** se obtiene del cociente entre el autovalor mayor y cada uno de los autovalores. Valores mayores que 15 pero menores que 30 existe colinealidad aunque es moderada, pero valores superiores a 30 indican un severo problema de colinealidad.

Las **proporciones de varianza** recogen las proporciones de varianza del coeficiente de regresión parcial que está explicada por cada dimensión o factor. Cuando no hay presencia de colinealidad cada dimensión explica gran cantidad de varianza de un sólo coeficiente, a excepción del coeficiente ( $B_0$  –constante) que siempre se asocia con otro coeficiente.

Por los resultados de la tabla 5.1.16 los índices de condición son inferiores a 30 excepto el último. Si para la dimensión que tiene ese índice se estudia la proporción de varianza de los coeficientes de regresión parcial, se observa que sólo la constante viene representada por esa dimensión.

### **1.2.3. Estudio de casos influyentes y extraños**

En el cuadro de diálogo de la fig. 5.1.2, cuando se selecciona **guardar**, se accede a diferentes opciones que se van a comentar (ver fig. 5.1.16) a continuación:

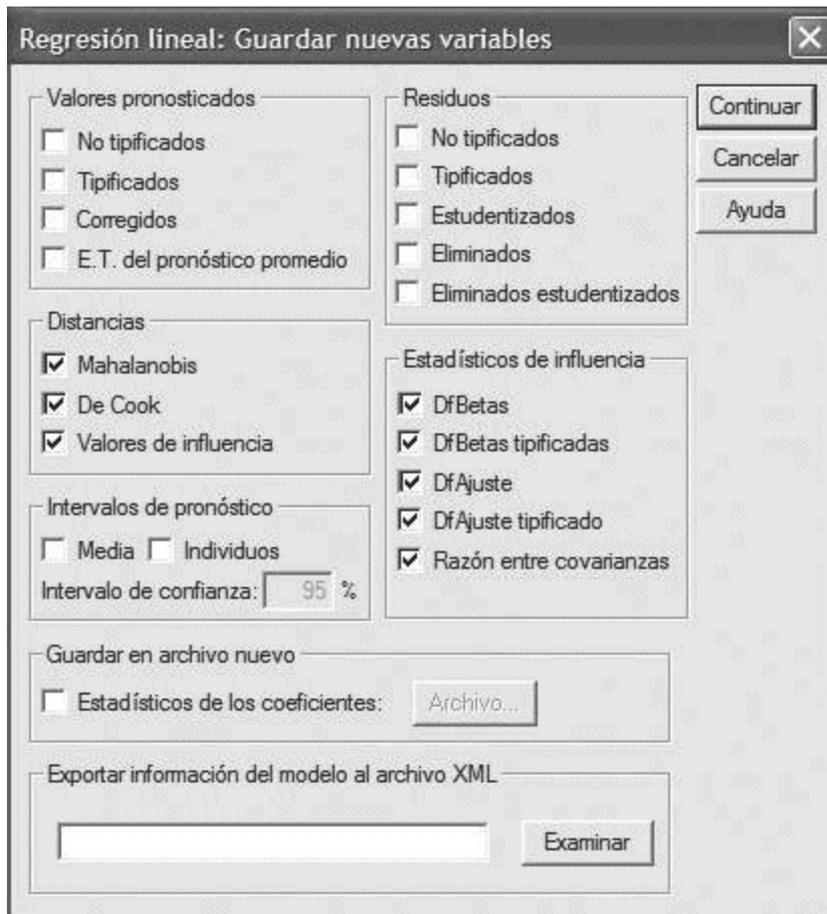


FIGURA 5.1.16.

### Regresión lineal: Guardar variables.

**Valores pronosticados.** Son valores que el modelo de regresión pronostica para cada caso.

**Distancias.** Son medidas para identificar casos con valores poco usuales como combinación de las variables independientes y casos que puedan tener una gran influencia en el modelo.

Las distancias que SPSS calcula son:

a) **Distancia de Mahalanobis.** Mide la distancia de cada caso a los promedios de valores de las variables independientes. Se obtiene multiplicando por  $n-1$  el **valor de influencia** de cada caso.

b) **Distancia de Cook.** Mide la cantidad en que varían las estimaciones de los coeficientes de regresión, si la observación concreta se elimina del análisis. Si su valor es elevado será indicador de influencia del caso en el modelo. Casos con distancias de Cook superiores a 1 deben ser considerados como influyentes.

c) **Valores de influencia.** Son indicadores del grado de influencia de cada caso. SPSS considera que valores menores de 0,2 no deben

considerarse como influyentes, valores entre 0,2 y 0,5 comportan riesgo de influencia, y finalmente, valores superiores a 0,5 deben evitarse.

**Intervalos de pronóstico.** Son los límites superior e inferior para los intervalos de predicción individual y promedio.

**Residuos.** Como ya se ha expuesto, el residuo representa la diferencia entre el valor actual de la variable dependiente y el valor pronosticado por la ecuación de regresión. SPSS guarda como nuevas variables los residuos: no tipificados, tipificados, estudentizados, eliminados y eliminados estudentizados.

**Estadísticos de influencia.** Todos los estadísticos que presenta el programa (ver fig. 5.1.16) intentan precisar la posible presencia de casos influyentes. a) **DfBetas.** Representa el cambio en los coeficientes de regresión estandarizados (Beta) cuando el caso es eliminado.

b) **DfBetas tipificados.** Resultarán de dividir DfBetas entre su error típico. Valores superiores a  $2$  son sospechosos de mayor influencia en el modelo.<sub>n</sub>

c) **DfAjuste.** Mide el cambio que se produce en el pronóstico de un caso cuando éste es excluido del modelo de regresión.

d) **DfAjuste tipificado.** Resultarán de dividir DfAjuste entre su error típico. Valores superiores a  $2^k$  son sospechosos de mayor influencia en el

modelo.

e) **Razón entre covarianzas(RV).** Es la razón entre el determinante de la matriz de covarianza con un caso particular excluido y el determinante de la matriz de covarianza con todos los casos incluidos. Un caso será influyente si  $RV-1$  es mayor que  $3+k/n$ .

En el problema-ejemplo si se selecciona todas las distancias y los estadísticos de influencia, en el fichero de datos se añadirán las siguientes variables:

mah\_1 Distancia de Mahalanobis  
coo\_1 Distancia de Cook  
lev\_1 Valores de influencia  
cov\_1 Razón de covarianza  
dff\_1 DfAjuste  
sdf\_1 DfAjuste tipificado  
dfb0\_1 DfBeta para la constante  
dfb1\_1 DfBeta para la variable X1  
dfb2\_1 DfBeta para la variable X2  
dfb3\_1 DfBeta para la variable X3

dfb4\_1 DfBeta para la variable X4  
 dfb5\_1 DfBeta para la variable X5  
 dfb6\_1 DfBeta para la variable X6  
 sdb0\_1 DfBeta tipificado para la constante sdb1\_1 DfBeta tipificado para la variable X1 sdb2\_1 DfBeta tipificado para la variable X2 sdb3\_1 DfBeta tipificado para la variable X3 sdb4\_1 DfBeta tipificado para la variable X4 sdb5\_1 DfBeta tipificado para la variable X5 sdb6\_1 DfBeta tipificado para la variable X6

Además de añadir las variables al fichero de datos, el programa SPSS presenta un resumen de los valores de los residuos, las distancias y los casos influyentes.

TABLA 5.1.17. Estadísticos sobre los residuos(a).

	Mínimo	Máximo	Media	Desviación	N <sub>típ.</sub>
Valor pronosticado	53,5007	92,6671	72,1023	7,73122	200
Valor pronosticado tip.	-2,406	2,660	,000	1,000	200
Error típico del valor pronosticado	,094	,308	,185	,044	200
Valor pronosticado corregido	53,6099	92,7008	72,1022	7,73026	200
Residuo bruto	-3,18074	2,24343	,00000	,99964	200
Residuo tip.	-3,134	2,210	,000	,985	200
Residuo estud.	-3,177	2,245	,000	1,002	200
Residuo eliminado	-3,26866	2,31524	,00013	1,03568	200
Residuo eliminado estud.	-3,255	2,269	,000	1,007	200
Dist. de Mahalanobis	,717	17,337	5,970	3,303	200
Distancia de Cook	,000	,040	,005	,007	200
Valor de influencia centrado	,004	,087	,030	,017	200

a: Variable dependiente: entonación vocal

Los valores máximos de la distancia de Cook y los valores de influencia centrado permiten afirmar que no hay ningún caso especialmente influyente.

Se debe destacar que los casos influyentes no producen residuos elevados por lo que no causan problemas de ajuste, sin embargo conviene caracterizarlos por su especial incidencia en la generalización del modelo.

### 1.3. La elaboración de pronósticos

Cuando el objetivo del análisis de regresión es obtener valores de la variable dependiente desconocidos a partir de otros de las independientes, se puede seleccionar en el cuadro de diálogo de la fig. 5.1.16 diferentes opciones de *valores pronosticados*:

a) **No tipificados.** Pronósticos directamente obtenidos de la ecuación de regresión de las puntuaciones directas de las variables independientes. En SPSS toma la expresión: **pre\_%**

b) **Tipificados.** Pronósticos tipificados. Su expresión es: **zpr\_%** c) **Corregidos.** Pronóstico de la ecuación de regresión obtenida por exclusión del caso. Su expresión es: **adj\_%**

d) **E.T del pronóstico promedio.** Estimación de la desviación típica del valor promedio de la variable dependiente para los casos que tienen valores iguales en las variables independientes (SPSS, ayuda). Su expresión es: **sep\_%**

e) **Intervalos de pronóstico.** Intervalos de confianza para la **media** basados en los e.t. del pronóstico medio. También intervalos de confianza para los **individuos** basados en los e.t. de los pronósticos individuales. Se generarán cuatro variables:

1. **Lmci\_%:** limite inferior del intervalo de confianza para el pronóstico medio.
2. **Umci\_%:** limite superior del intervalo de confianza para el pronóstico medio.
3. **Lici\_%:** limite inferior del intervalo de confianza para el pronóstico individual.
4. **Uici\_%:** limite superior del intervalo de confianza para el pronóstico individual.

En todos los casos el símbolo % en el nombre de variable indicará el número de orden de realización del cálculo. Así por ejemplo la primera vez que se calcule los valores no tipificados tomará la expresión pre\_1, si se ejecuta de nuevo el procedimiento su expresión será: pre\_2.

## 2. ANÁLISIS DE LA VARIANZA

### 2.1. El significado de la experimentación

Cuando se intenta comprender los procesos comerciales, industriales o de servicios, muchas veces se piensa que el mejor procedimiento es recoger gran cantidad de datos sobre el funcionamiento del proceso y luego analizarlos para sacar conclusiones.

Aunque este procedimiento no suele prodigarse, sin embargo, es cierto que su uso ayudaría enormemente a comprender el proceso. No obstante, también es verdad, que si el proceso de recogida de datos va más allá y piensa en interpelar al sistema intentando controlar la validez de la información, estaremos **experimentando**, es decir, variando

deliberadamente las condiciones habituales de trabajo de un sistema, para encontrar mejores maneras de proceder y a la vez ahondar en el conocimiento del comportamiento del mismo.

El objetivo fundamental de la experimentación es estudiar la posible relación de causalidad existente entre dos o más variables. Este estudio representa la elección de las variables que pueden determinar a otras, pero sobre todo, el establecer una relación de causa/efecto implica el cumplimiento de tres condiciones: a) debe existir una covariación (correlación) entre la variable causal —independiente— y el efecto —dependiente—, b) temporalmente, la causa debe preceder al efecto, y c) no existen variables extrañas, ajenas al modelo, que puedan explicar el efecto.

La última condición es la que mediatiza la consecución de la relación causal. Para conseguir este objetivo, fundamentalmente, se proponen algunas estrategias: la manipulación, el control y la aleatorización.

La utilización de la experimentación en la industria se concreta en dos áreas: el diseño o mejora de productos y la explotación o mejora de procesos. En los servicios, su actuación se centra fundamentalmente en la mejora de los últimos. Supóngase el siguiente ejemplo de un experimento.

### **Ejemplo**

En un instituto politécnico se quiere saber si cuatro tipos de métodos de estudio tienen la misma eficacia. Para conseguir el objetivo se realiza el siguiente diseño experimental: se cogen cuatro grupos, de tres alumnos cada uno, empleando en cada grupo uno de los cuatro tipos de métodos de estudio. La asignación del método a cada grupo de alumnos se realiza al azar.

En este ejemplo utilizando una técnica de aleatorización, se pretende saber que tipo de método de estudio resulta más eficaz, es decir, se experimenta en lugar de tomar datos sin un plan previamente establecido.

En suma el experimentar significa modificar de forma intencionada las condiciones normales de trabajo para localizar mejores maneras de proceder y conseguir, a su vez, un conocimiento más profundo sobre el comportamiento de productos y procesos.

## **2.2. Elementos básicos del diseño de experimentos**

### **Variabilidad de los resultados**

El objetivo de un experimento es medir el efecto de las variables explicativas o de las variables independientes, sobre una variable

dependiente, mientras se controlan otras variables que pueden confundir el resultado.

Supóngase que una variable respuesta  $y$  se va a medir como resultado de dos variables independientes  $x_1$  y  $x_2$ . Las fluctuaciones que se produzcan en esas variables, y la intervención de otras variables que el experimentador desconoce, harán que el modelo se exprese como:

$$y=f(x_1, x_2)+e \text{ donde } e \text{ representará el error}$$

Evidentemente cuanto menor sea el error experimental más perfecto será el modelo, en el sentido de poder obtener con él unas estimaciones más precisas.

Indudablemente si en este modelo se introduce otras variables que influyen en el resultado disminuirá el error. Este será el objetivo prioritario del diseño de experimentos, minimizar el error controlando la variabilidad de las variables que intervienen y evitando la presencia o controlando otras variables que no intervienen en el modelo.

Como estrategias para minimizar el error, el experimentador utilizará fundamentalmente dos: la **repetición** y la **aleatorización**.

### **Interacción**

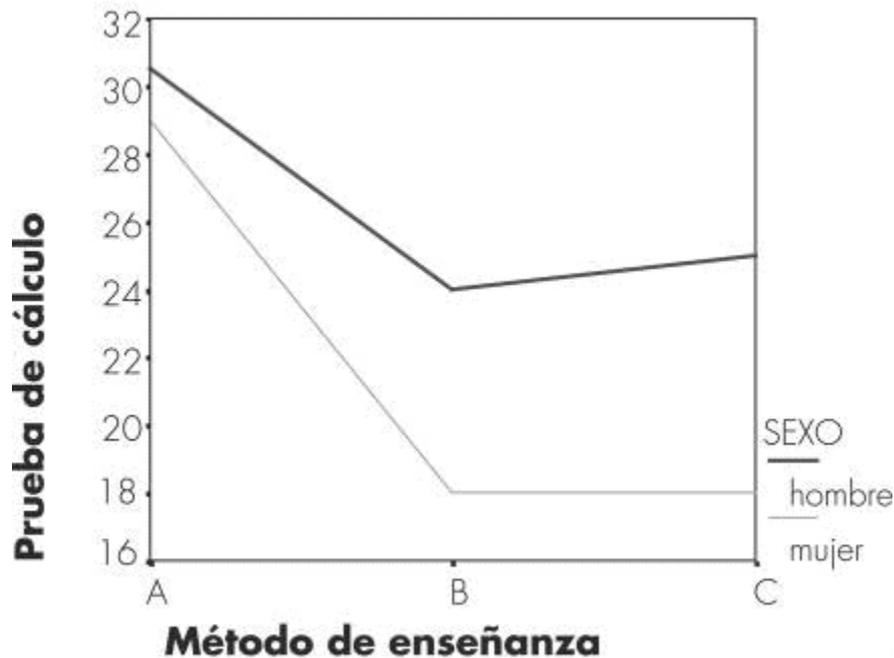
Se entiende por interacción el efecto que produce la combinación de dos o más variables independientes sobre la dependiente. Supongamos, por ejemplo que se tiene dos variables independientes: métodos de enseñanza y sexo de los alumnos, y como dependiente los resultados en una prueba de cálculo:

#### **Métodos de enseñanza Métodos de enseñanza A B**

Media  
resultados: 30,5 24 Hombres  
C A B C

Media  
25 resultados: 29 18 18  
Mujeres

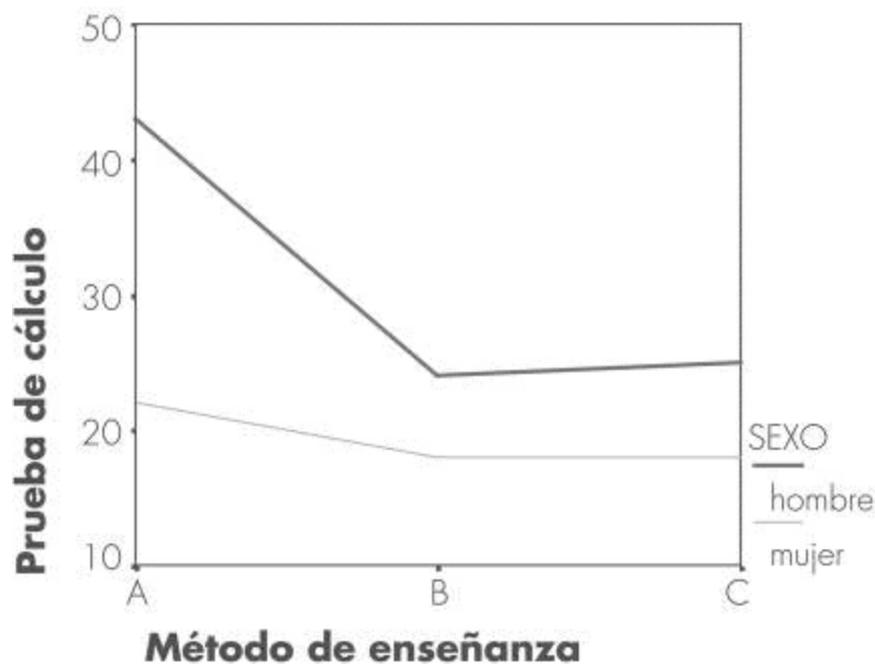
La representación gráfica de los resultados será:



Si las gráficas de hombre y mujer son aproximadamente paralelas no hay interacción, pero si no son paralelas se evidencia interacción.

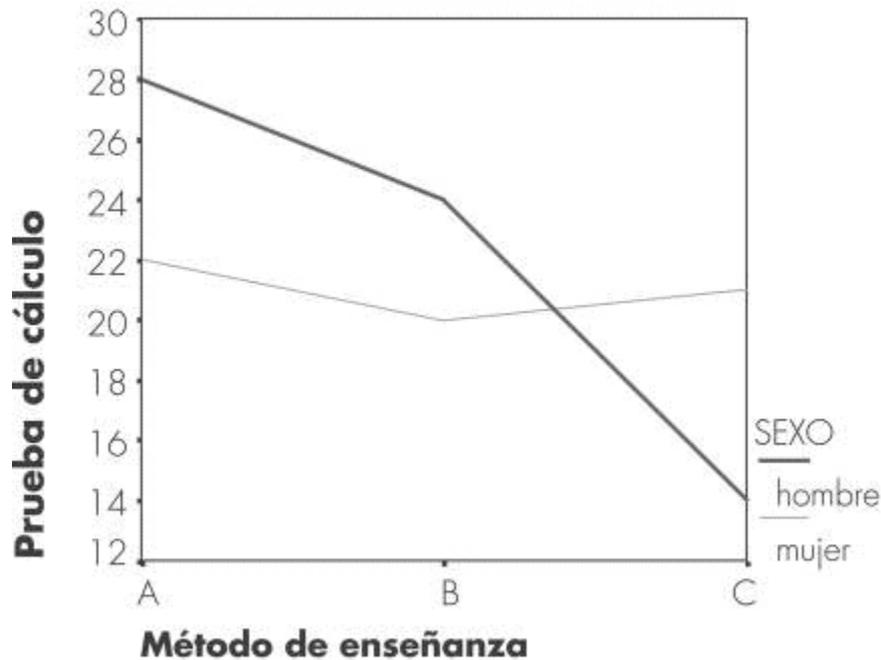
Puede presentarse dos tipos de interacción:

a) Si los hombres superan en los tres métodos pero sobre todo en uno de ellos, el gráfico será similar al siguiente:



Este tipo de interacción se denomina ordinal y se caracteriza por la ausencia de cruce entre las gráficas de hombre y mujer.

b) Los hombres superan a las mujeres en algunos métodos pero no en otros. Se produce cruce entre las gráficas. Es la llamada interacción no ordinal:



También se puede dar el caso que el efecto de una variable independiente sea constante para todos los niveles de la segunda variable independiente, en tal caso se dirá que no hay interacción entre ambas variables.

### 2.3. Los diseños factoriales

En primer lugar conviene definir, para fijar ideas, ciertos términos que ya han sido mencionados:

— **Factor.** Es cada una de las variables que pueden influir en la variable respuesta o resultado y, por tanto, se incluirán en el plan de experimentación. — **Nivel.** Es cada uno de los valores que toma el factor en un experimento.

— **Respuesta.** Es la denominación que toma la variable resultado o dependiente. Los estudios experimentales se pueden hacer sobre varias variables dependientes, sin embargo, por razones didácticas, sólo se estudiará el caso de una sola variable respuesta.

Para establecer una formulación matemática del modelo factorial supóngase un factor A que se presenta con «i» niveles, dentro de cada nivel analizamos una serie de observaciones; así  $x_{ij}$  es la observación j-ésima del i-ésimo nivel de tratamiento del factor A. Este valor se puede descomponer en dos partes:  $m_i$ ,  $e_{ij}$ .

$m_i$  es la parte del resultado  $x_{ij}$  debida al i-ésimo nivel y común a todos los elementos sometidos a ese nivel del factor A;  $e_{ij}$  será el error

experimental debido a la intervención de otros factores no estudiados o controlados por el experimentador. Si se supone que el modelo de variación es lineal, el modelo matemático apropiado será:

$x_{ij}=m_i + e_{ij}$  Por extensión, si se supone que existen más factores, por ejemplo, dos A y B el modelo que se puede plantear es el siguiente:

$$x_{ijk}=m_i+m_j+e_{ijk}$$

Sin embargo, en este modelo, no se ha tenido en cuenta el efecto de la **interacción**, es decir, el efecto producido por A influido por los niveles que toma B, de otra forma, la respuesta al nivel 2 de A no será igual si B está en el nivel 1 o en el 5, por ejemplo. En el ejemplo del método de enseñanza y el género del alumno, no dará los mismos resultados dependiendo del sexo (hombre, mujer). Esta influencia se verá reflejada en una distinta formulación matemática del modelo, que tomará la siguiente expresión:

$$x_{ijk}=m_i+m_j+m_{ij}+ e_{ijk} \text{ donde } m_{ij} \text{ refleja el efecto de la interacción.}$$

En el planteamiento del modelo la influencia de los niveles puede operar de varias maneras. Si para obtener los resultados sólo intervienen unos determinados niveles, por ejemplo, dos o tres niveles por factor y siempre los mismos en cualquiera de las observaciones que se realicen, estaremos ante un modelo de efectos fijos o **modelo fijo** y los factores se llamarán fijos. Si de los diferentes niveles que pueden tomar los factores se coge una muestra de los mismos, estaremos ante un modelo de efectos aleatorios o **modelo aleatorio** y los factores se denominarán aleatorios.

Cuando en un modelo hay más de un factor, uno o varios pueden ser fijos y el resto aleatorios, en tal caso el **modelo** se denomina **mixto**.

Son muchos los ejemplos de diseños factoriales, no obstante, para fijar ideas, a continuación se proponen tres ejemplos, uno de cada tipo.

### **Ejemplo de diseño de efectos fijos**

En una empresa de fabricación de telas, se está estudiando el rendimiento producido en los paños según tres modelos distintos de elaboración; para ello se toma veinticuatro telas similares en grupos de ocho y se las somete a los tres procesos distintos de fabricación. Como variable respuesta se estudia el número telas producidas por cada uno de los tres métodos.

Este ejemplo se puede expresar mediante el siguiente modelo matemático:  $x_{ij}=m_i+e_{ij}$  donde  $x_{ij}$  representa el rendimiento del j-ésimo

pañó sometido al  $i$ -ésimo nivel del factor (tipo de proceso de fabricación),  $m_i$  será la media del nivel  $i$ -ésimo del factor, en el ejemplo  $i=1,2,3$ , y finalmente el error  $e_{ij}$ .

### **Ejemplo de diseño de efectos aleatorios**

El mismo ejemplo anterior, con algunas modificaciones, puede servir en el modelo de efecto aleatorio. Supóngase que los tres procesos de fabricación, son sólo una muestra de los muchos métodos de elaboración que existen.

El modelo matemático será similar pero se supondrá que los  $i$ -niveles del factor son una muestra aleatoria simple de una población de infinitos niveles que se suponen con distribución normal (media 0, desviación  $s$ ).

### **Ejemplo de diseño de efectos mixtos**

Surge por extensión de los anteriores, donde unos niveles son fijos y otros son una muestra aleatoria de todos los niveles posibles. Supóngase el caso que hay diez métodos de fabricación y tres se toman fijos y siete son una muestra de los «infinitos» (supuesto) posibles.

## **2.4. Análisis de la varianza factorial**

Cuando en el diseño de experimentos se toman varias muestras de los distintos niveles de un factor, se está formando grupos para cada uno de los niveles.

El análisis de la varianza resuelve el problema de comparar varios grupos que tienen sus propias particularidades estimándose, en virtud de los resultados, la diferencia que existe entre dichos grupos y la significación o no de dicha diferencia. En concreto, el análisis de la varianza considera los datos de los grupos como un conjunto y a través de las pruebas estadísticas oportunas decide si los grupos provienen o no de la misma población (o de poblaciones distintas con la misma varianza) y, por consiguiente, si los grupos tienen medias significativamente distintas o no.

Por otra parte, el análisis de la varianza no es otra cosa que el estudio de la variación total de un conjunto de datos respecto a la media general, dividiendo esta variación en varios componentes que dependen de las particularidades de cada grupo.

Puede parecer extraño que un procedimiento que compara las medias se llame análisis de la varianza. Sin embargo, este nombre se deriva del hecho que para probar la diferencia de medias, estamos comparando realmente (es decir, analizando) las variaciones.

Por tanto, el procedimiento de comparar más de dos medias poblacionales se conoce como análisis de la varianza (ANOVA).

Alguien se podría preguntar por qué no se utiliza los contrastes de «t» múltiples. La respuesta es sencilla, aumentan el porcentaje del error tipo I global. El ANOVA evita este aumento del error tipo I.

Para aplicar el análisis de la varianza se debe comprobar tres supuestos teóricos:

a) **Independencia de las observaciones.** se refiere al hecho de que los datos de los distintos grupos han sido elegidos al azar de una población normal.

b) **Igualdad de varianza (homocedasticidad).** se supone que los distintos grupos tienen una varianza común, de modo que no existe diferencia significativa entre las varianzas de los distintos grupos.

c) Los **errores** que interviene en las observaciones están distribuidos normalmente.

En las fórmulas que vamos a emplear el término varianza será equivalente a la suma de los cuadrados dividida por los grados de libertad  $SC_{gl}$

### ***2.4.1. Análisis de la varianza con un factor. Modelo de efectos fijos***

#### **1. Formulación y objetivos**

Sea un experimento con un factor A de tratamientos y J niveles en dicho factor, en cada uno de los cuales se dispone de  $n_j$   $j=1, \dots, J$  observaciones independientes, seleccionadas de poblaciones normales con varianza común  $s^2$ . El análisis de la varianza de un factor trata de contrastar la hipótesis nula de que las poblaciones normales de las cuales se tomaron las muestras tienen media común, frente a la alternativa de que al menos dos de ellas difieren.

Para centrar los objetivos de esta técnica, supongamos el siguiente ejemplo: un profesor de enseñanza secundaria desea saber la influencia que tiene la utilización de tres métodos distintos de enseñanza, en los resultados de una determinada materia.

En este ejemplo se plantea un modelo de un solo factor, los métodos de enseñanza, con efectos fijos (se escoge exclusivamente, tres métodos determinados).

El modelo que se postula para explicar los datos en este análisis es el modelo lineal, donde la variable analizada la hacemos depender de la influencia de un único factor de tal manera que el resto de las causas de

variación son englobadas en el denominado error experimental, de carácter aleatorio.

Supongamos:

$$x_{ij} = m_j + e_{ij} \quad [1] \text{ con } E(x_{ij}) = m_j.$$

Si descomponemos  $m_j$  en dos partes, una común a todas las observaciones,  $m$ , y otra exclusiva de cada nivel  $a_j$ , entonces  $m_j = m + a_j$  que sustituyendo en [1] nos lleva a:

$$x_{ij} = m + a_j + e_{ij} \quad i=1, \dots, n_j \quad j=1, \dots, J$$

donde  $x_{ij}$  se puede descomponer en:

$x_{ij} =$  constante común para todos los valores de la variable + efecto del tratamiento + error

$x_{ij}$  representa la puntuación  $i$  en el grupo  $j$ .

$m$  constante común para todos los valores de la variable.  $a_j$  media de la variable para el nivel  $j$  del factor.

$e_{ij}$  error del modelo lineal.

También podemos hacer la siguiente descomposición:

$$x_{ij} - m = (a_j - m) + (x_{ij} - a_j) \quad [2]$$

desviación del puntaje sobre la media en general = desviación del promedio del grupo sobre la media general (efecto del tratamiento) + desviación individual del puntaje sobre la media del grupo (el error).

Si elevamos la expresión [2] al cuadrado y efectuamos las sumas tenemos:  $SC_{total} = SC_{entre} + SC_{intra(error)}$  donde:

$SC_{total}$  es la suma de las desviaciones al cuadrado de cada puntuación respecto a la media total:

$\sum_{j=1}^J \sum_{i=1}^{n_j}$

$$\sum_{j=1}^J \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2$$

siendo  $j=1, \dots, J$  y  $N$  el número total de puntuaciones  $\sum_{j=1}^J n_j = N$

$$\sum_{j=1}^J N_j = N$$

$SC_{entre}$  es la **suma de cuadrados intergrupo**, desviaciones de la media de las puntuaciones en cada grupo respecto a la media total, equivale a la varianza de la media de los grupos tratadas estas como datos individuales:

$$SC_{entre} = \sum_{j=1}^J n_j (\bar{X}_{.j} - \bar{X})^2$$

$SC_{intra}$  es la **suma de cuadrados intragrupo**, desviaciones de las puntuaciones en cada grupo respecto a la media del grupo, refleja la dispersión de los datos dentro de cada grupo:

$$J n_j$$

$$SC_{intra} = \sum_{j=1}^J \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{.j})^2$$

$$j=1, i=1$$

Si planteamos la hipótesis nula:  $H_0: m_1 = m_2 = \dots = m_j$  o  $H_0: a_j = 0$ , se conforma la siguiente tabla 5.2.1 del análisis de la varianza:

TABLA 5.2.1. Análisis de la varianza de un factor. Efectos fijos.

Fuente de variación	Suma de Grados	Media de Estadístico F	cuadrados de libertad
Entre grupos	$SC_{entre}$	$J-1$	$MC_{entre} = \frac{SC_{entre}}{J-1}$
Intra grupos	$SC_{intra}$	$N-J$	$MC_{intra} = \frac{SC_{intra}}{N-J}$
Total	$SC_{total}$	$N-1$	

Entre grupos  $SC_{entre}$   $J-1$   $MC_{entre} = \frac{SC_{entre}}{J-1}$

Intra grupos  $SC_{intra}$

$N-J$

$MC_{intra} = \frac{SC_{intra}}{N-J}$

$MC_{entre} = \frac{SC_{entre}}{J-1}$

$MC_{total} = \frac{SC_{total}}{N-1}$

$MC_{total} = \frac{SC_{total}}{N-1}$

$N-J$

Total  $SC_{total}$   $N-1$

Con la hipótesis nula de homogeneidad de varianzas, el estadístico F se distribuye con una F de Snedecor con  $J-1$  y  $N-J$  grados de libertad. Si el p-valor asociado a F es menor que  $\alpha$ , se rechazará la hipótesis nula al nivel de significación  $\alpha$ .

## 2. Supuestos

Para aplicar el análisis de la varianza nos hemos apoyado en tres supuestos: independencia de las observaciones, igualdad de varianza

(homocedasticidad) y normalidad. Veamos a continuación, que puede ocurrir si alguno de estos tres supuestos no se cumple.

### Supuesto de independencia

Cuando se viola la hipótesis de independencia de las observaciones, las consecuencias son importantes. En el caso de dos tratamientos, existen técnicas para realizar el tratamiento en el caso que los  $e_{ij}$  no sean independientes. En el caso de más de dos tratamientos, en ocasiones el modelo puede plantearse como un problema de medidas repetidas.

### Supuesto de homocedasticidad

El efecto de varianzas heterogéneas depende de las condiciones del análisis:

— Cuando los tamaños de las muestras son iguales, la presencia de varianzas heterogéneas tiene un efecto mínimo sobre el nivel de significación del contraste F.

— Cuando los tamaños muestrales son desiguales, y se han seleccionado menos individuos de las poblaciones con mayor varianza, la probabilidad de cometer un error de tipo I, rechazar  $H_0$  siendo verdadera, es mayor que  $\alpha$ .

— Cuando los tamaños muestrales son desiguales, y se han seleccionado más individuos de las poblaciones con mayor varianza, la probabilidad de cometer un error de tipo I es menor que  $\alpha$ .

Hay diversos contrastes para observar la homogeneidad de varianzas:  $F_{ma}$  de Hartley, la C de Cochran, la prueba de Bartlett, pero sobre todo la prueba de Levene, en la que nos vamos a detener.

Esta prueba es insensible a la violación de la suposición de que las J muestras proceden de poblaciones normales; además se puede aplicar al caso de tamaños muestrales desiguales.

La expresión del estadístico de contraste es:

$$F_{Levene} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (\hat{A}_{ij} - \bar{A}_i)^2}{(k-1) \sum_{i=1}^k \sum_{j=1}^{n_i} (\hat{A}_{ij} - \bar{A})^2}$$

∧∧

0 donde:

2

$\sum_{i=1}^j \sum_{j=1}^i$

0

$\sum_{ij}$

$Dx x||$

$\sum_{ij} i$  son las desviaciones en valor absoluto de las j observaciones respecto a la media del grupo i.

$n_i$

$\hat{A}D_{ij}$

$D_i = \sum_{j=1}^{n_i} \dots$  la media de las desviaciones en valor absoluto de cada grupo ( $n_i$ ).  $n_i$

∧∧ $D_{ij}$

$D_t = \sum_{i=1}^j \sum_{j=1}^i$  la media de las desviaciones en valor absoluto respecto al total (N).  $N$

Esta  $F_{Levene}$  se distribuye según una F de Snedecor con (k-1), (N-k) grados de libertad.

En los casos en que se presenta heterocedasticidad, es muy común que la varianza cambie con la media. Esto es, siendo  $m_j$  la media en el grupo j-ésimo y  $s_j$  la desviación típica:  $s_j = T(m)$  para alguna función T. La pregunta que ahora nos hacemos es la siguiente: ¿cómo encontraremos una función T de modo que los datos transformados T(x) tengan varianza constante? En general, para estabilizar la varianza, puede utilizarse una transformación potencial del tipo Box y Cox.

### Supuesto de normalidad

La no normalidad tiene muy pocos efectos sobre el nivel de significación del contraste F, es prácticamente irrelevante en lo referente a la probabilidad de cometer un error de tipo I

### 3. Procedimiento

Los pasos que tenemos que hacer en el análisis de la varianza serán: a) Se postula el modelo lineal

$$x_{ij} = m_j + e_{ij} \quad i=1,\dots,n_j \quad j=1,\dots,J$$

b) Se estiman los parámetros  $(m_1, \dots, m_J, s^2)$ :

$$\hat{m}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}$$

$$s^2 = MC_{intra}$$

c) Se plantea la hipótesis nula frente a la alternativa:

$$H_0: m_1 = m_2 = \dots = m_J \quad \text{frente a} \quad H_1: \exists i, j : m_i \neq m_j$$

d) Escoger un nivel de significación  $\alpha$ , probabilidad de rechazar la hipótesis nula siendo ésta verdadera. Calcular el estadístico  $F = MC_{inter} / MC_{intra}$  que bajo  $H_0$  se distribuye como una F de Snedecor con  $J-1$  y  $N-J$  grados de libertad. Si el p-valor asociado a F es menor que  $\alpha$ , se rechazará la hipótesis nula al nivel de significación  $\alpha$ .

#### 4. Pruebas a posteriori (post-hoc)

Con el análisis de la varianza se llega a la conclusión de rechazar o no la hipótesis nula de igualdad de medias. En caso de rechazarla, se sabe que por lo menos existe una diferencia significativa entre algún par de medias, pero no se sabe entre que par o pares se encuentran las diferencias. El objetivo de los *métodos de comparación múltiples* es aislar las comparaciones entre las medias que, por ser significativas, influyen en la decisión. El contraste t no es adecuado en esta situación, dado que está diseñado para dos muestras aleatorias y no es válido para determinar la significación de las diferencias entre valores altos y bajos de las medias de toda una serie de J medias, puesto que en tal caso no tiene en cuenta la magnitud de J. Si J es grande, al tomar J muestras al azar de una misma población normal, en la mayoría de los casos, el contraste t indicará diferencias significativas entre la media mayor y la menor.

Los métodos de comparación de Tukey y Scheffé, aunque son procedimientos que toman formas muy generales, abarcan casi todos los problemas que se pueden solucionar mediante estos contrastes y han dado buenos resultados.

El método de Tukey se emplea cuando los tamaños muestrales son iguales y existe homocedasticidad.

El método de Scheffe no necesita que los tamaños muestrales sean iguales. Pero es menos potente que el anterior. En este caso su expresión es la siguiente:

$dk F MC nn_{12}$

donde  $d$  es la diferencia mínima entre dos medias para que sean significativas,  $n_1$  y  $n_2$  los tamaños de las muestras y  $k$  el número de grupos.

### **Ejemplo**

En una institución educativa se quiere saber si los resultados en matemáticas de sus alumnos se ven modificados por el tipo de método didáctico de los profesores. Para conseguir el objetivo marcado se elige al azar un grupo de 21 alumnos de similares conocimientos en la materia y se le asigna al azar a cada profesor. Sabiendo que el profesor A utiliza el método tradicional de exposición inicial del tema y práctica de problemas con los alumnos; el profesor B hace un resumen del tema y después continua con ejercicios y problemas, y finalmente el profesor C realiza desde el primer momento ejercicios y problemas con sus alumnos y sólo cuando surge una duda la explica a todos sus alumnos.

Con los datos medidos en una prueba de rendimiento de matemáticas, realizada tras la experiencia, y que se recoge en la tabla adjunta. Se desea saber, a un nivel de significación del 5%, si existe diferencia significativa entre los distintos métodos de enseñanza.

Profesor A 68 72 77 52 42 60 58

Profesor B 71 78 56 59 63 71 74

Profesor C 78 81 82 83 70 65 78

Aplicando el análisis de la varianza a los datos del problema tendremos la tabla de análisis de la varianza presentada a continuación:

ANOVA X

Inter-grupos Intra-grupos Total

**Suma de cuadrados** **gl** **Media cuadrática** **F** **Sig.** 844.670 2 422.335  
4.903 .020 1550.530 18 86.140  
2395.200 20

Como la significación  $p=0,020$  es menor que  $\alpha=0,05$  entonces hay diferencia significativa entre los métodos de enseñanza. Al dar diferencia significativa entre las medias de los grupos podemos intentar escudriñar los grupos entre los que existe dicha diferencia. Para conseguir este objetivo podemos utilizar la prueba de Tukey o la de Scheffé. Utilizaremos la última por su mayor poder de generalización.

Las medias de los grupos son:

Grupo A=61,28

Grupo B=67,42

Grupo C=76,71

$d = 2.86, 14.3, 55 = \neq, 13 \ 22$

Luego sólo resulta significativa la diferencia entre la media del grupo A y el C donde ésta es mayor que d.

### 2.4.2. Análisis de la varianza de dos factores

Para fijar ideas, supóngase que se va a estudiar un problema en el cual se piensa existen 2 factores, A con a niveles y B con b niveles. Cada replicación del experimento contiene ab niveles. Se supone que existe n repeticiones del experimento y que  $x_{ijk}$  representa la observación bajo el i-ésimo nivel del factor A y el j-ésimo nivel del factor B en la k-ésima replicación. El orden en el que las abn observaciones se toman se selecciona aleatoriamente. Las observaciones se pueden poner mediante el siguiente modelo lineal:

$$x_{ijk} = m + a_i + b_j + (ab)_{ij} + e_{ijk}$$

donde m es el efecto medio,  $a_i$  es el efecto debido al i-ésimo del factor A,  $b_j$  es el efecto del j-ésimo nivel del factor B,  $(ab)_{ij}$  es el efecto de la interacción entre  $a_i$  y  $b_j$  y  $e_{ijk}$  es la componente aleatoria (error). Los efectos de los factores se suponen fijos y los efectos de los niveles se suponen como desviaciones a la media, por lo que se verificará:

$$a \ b \ a \ b$$

$$\hat{A} \ \hat{A} \ \hat{A} \ \hat{A} \ 0 \ 000ij$$

$$i = = = 11 \ 1 \ 1$$

El total de observaciones será:  $abn = N$  Se deberá contrastar la hipótesis:

$$H_0: a_1 = a_2 = a_3 = \dots = a_a = 0$$

$$H_1: \text{al menos un } a_i \neq 0$$

ó

$$H_0: b_1 = b_2 = b_3 = \dots = b_b = 0$$

$$H_1: \text{al menos un } b_j \neq 0$$

También se estará interesado en determinar sí:

$$H_0: (ab)_{ij} = 0$$

$H_1$ : al menos un  $(ab)_{ij} \neq 0$

Se puede descomponer la suma de cuadrados total  
 (( ))

$$\hat{A} \hat{A}$$

$$\hat{A}_{ijk}$$

-

2

... en

$$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n$$

suma de los siguientes cuadrados: en la suma de cuadrados respecto al factor A

$$\sum_{a,b}$$

$$\hat{A}_i$$

i

..

$$(( ))$$

2

... , en la suma de cuadrados respecto al factor B  $(\sum_{i=1}^a \hat{A}_{.j.})^2$ ,

$$\sum_{i=1}^a \sum_{j=1}^b$$

en la suma de cuadrados respecto a la interacción y

$$(\sum_{i,j} \hat{A}_{ij.} - \frac{(\sum_{i,j} \hat{A}_{ij.})^2}{n})^2$$

..

$$\sum_{i=1}^a \sum_{j=1}^b$$

en una suma de cuadrados de errores no explicado por el modelo  
 (diferencia del

$$\sum_{a,b,n}$$

total menos los factores y la interacción) .

$$\hat{A} \hat{A}$$

$$\hat{A}_{ijk}$$

ijk

$$()$$

2

ij.

$$i=1 \ j=1 \ k=1$$

Simbólicamente se escribirá así la ecuación:

$$SC_T = SC_A + SC_B + SC_{AB} + SC_E$$

indudablemente este modelo necesita que  $n \geq 2$  para poder calcular  $SC_E$

Estas fórmulas se pueden poner de más sencilla como:

$$a \ b \ n \ T \ 2$$

SC

T

=

$$\hat{A} \ \hat{A}$$

$\hat{A}$

x

2

$ijk \ T_N$  con T suma de todos los elementos.

$$i=1 \ j=1 \ k=1$$

$$a \ T_{i..} \ T^2$$

$$SC_A = \hat{A}^2$$

$i=1 \ n \ b \ N$  con  $T_{i..}$  suma de todos los elementos de la fila i (nivel i del factor A).

$$b \ T \ 2 \ T \ 2$$

$SC_B = \hat{A} \ T_{.j}$  con  $T_{.j}$  suma de todos los elementos de la columna j (nivel j del factor B).

del  $j=1 \ n \ a$

$$a \ b \ T \ 2 \ T \ 2$$

**$SC_{AB} = \hat{A} \hat{A}^{ij}$  con  $T_{ij}$  suma de los elementos de cada celda**

$$i=1 \ j=1 \ n \ N \ AB$$

(nivel i del factor A y nivel j del factor B).

$$SC_E = SC_T - SC_A - SC_B - SC_{AB}$$

El número de **grados de libertad** asociado con cada suma de cuadrados es:

**Efectos Grados de libertad** A a-1

B b-1

AB (a-1)(b-1) Error ab(n-1) Total abn-1

Cada suma de cuadrados dividida por sus grados de libertad nos dará el cuadrado medio (CM).

Si se supone en el modelo que  $e_{ijk}$  son normales e independientes con varianza constante  $s^2$ , entonces cada cociente ( $CM/CM_E$ ) se distribuye como una F-Snedecor con los grados de libertad del numerador y del denominador, y la región crítica del test, desde la hipótesis nula, serán aquellos valores del cociente que superen el correspondiente valor de  $F_a$  (F teórica). A continuación se expresa en una tabla, denominada de análisis de la varianza, todos los pasos a realizar para contrastar las hipótesis nulas planteadas anteriormente:

Tabla de análisis de la varianza de dos factores.

**F.V. G.L.**

Factor A a-1

Factor B b-1

Interacción AB (a-1)(b-1)

Error N-ab

**C.M. F (empírica) F (teórica)**

$SCA CM^A F(a-1, N-ab)_{a-1} CM E$

$SCB CM^B F(b-1, N-ab)_{b-1} CM E$

$SCAB CM^{AB} F((a-1)(b-1), N-ab)_{()()} CM E$

$SC_E$

$Nab$

Total N-1

Como ilustración veamos un ejemplo.

**Ejemplo**

Se piensa que la máxima producción de cierto producto petroquímico es función de la utilización de diferentes catalizadores y de la temperatura en que se produce la reacción. Se realizan 4 replicaciones de un diseño factorial con 3 catalizadores y 3 temperaturas, los resultados son:

**Temperatura**

**Catalizador 50 65 80  $X_{i.}$**

1 130 155 34 40 70 20 998 74 180 80 75 82 58

2 150 188 136 122 70 25 1300 159 126 106 115 58 45

3 138 110 174 120 104 96 1501 168 160 150 139 82 60  $x_{.j}$  1738 1291 770

3799

Al agrupar las réplicas se obtiene la siguiente tabla:

**Temperatura**

<b>Catalizador</b>	<b>50</b>	<b>65</b>	<b>80</b>	$\sum X_{i..}$	1 539	229	230	998	2 623	479	198	1300	3 576
	583	342	1501	$\sum x_{.j}$	1738	1291	770	3799					

El análisis de la varianza asociado, desde la hipótesis nula de la no influencia de cada uno de los factores, da el siguiente resultado:

**Fuente G.L. SC de variación**

Tipo de catalizador	2	10684
Temperatura	2	39119
Interacción	4	9613
Error	27	18231
Total	35	77647

**C.M.**

**F (empírica)**

**F (teórica)  $\alpha=0,05$**

	7,91	3,35
	19559,5	28,97
	3,35	
	2403,25	3,56
	2,73	
	675,22	

Las sumas de cuadrados (SC) de la tabla se obtienen aplicando las fórmulas anteriores. No obstante, para reforzar conocimientos, se van a desarrollar a continuación:

SC

total

= 130

$$22 \cdot 2 \cdot 2 + \dots + 104^2 \cdot 2 \cdot 2 + 3799^2 + 155 + 74 + 180 + 82 + 60_{36} = 77647$$

998

1300

1501

22 2 2

3799

$$SC_A = + + - = 10684 \quad 12 \quad 12 \quad 12 \quad 36 \quad BSC \text{ (temperatura)} = 1738 \cdot 22 \cdot 2 \cdot 2 + 1291 \cdot 770$$

$$3799 \cdot 3911912 \cdot 12 \cdot 12 + + - =$$

$$36 \quad SC_{AB} \text{ (interacción catalizador x temperatura)} =$$

539

623  
 576  
 229  
 479  
 583  
 230  
 198  
 342 2 2 22 2222 2 2  
 3799  
 = + + + + + + + + ()  
 4 4 4 4 4 4 4 4 36  
 - - = 9613  
 $SC_E = 77647 - 10684 - 39119 - 9613 = 18231$   
 $F_A = \frac{5342}{675 \cdot 22} = 7,91$   
 $F_B = \frac{19559}{5 \cdot 675 \cdot 22} = 28,97$   
 $F_{AB} = \frac{2403}{25 \cdot 675 \cdot 22} = 3,56$

Los valores de F(teórica) se recogerán de la tabla de F con los grados de libertad indicados en la tabla de análisis de la varianza y con el nivel de significación  $\alpha=0,05$  (en el ejemplo).

Como  $F(\text{empírica}) > F(\text{teórica})$  se puede concluir que los factores influyen en el modelo y además existe interacción entre ellos.

Para interpretar los resultados se puede añadir una gráfica de las respuestas medias de cada combinación de niveles de los diferentes tratamientos. La presencia de la interacción viene indicada mediante la falta de paralelismo de las líneas.

## 2.5. El comando ONEWAY

Permite realizar análisis de la varianza para una variable dependiente cuantitativa en los grupos establecidos por los valores de una variable independiente o factor. El análisis de la varianza se utiliza para contrastar la hipótesis de que varias medias son iguales. Si resulta significativa tal diferencia, necesitamos localizar las discrepancias, para ello utilizaremos las pruebas post-hoc.

### Problema-ejemplo

Queremos saber si existe diferencia significativa entre las medias de las puntuaciones de una prueba de inglés (item3) en función de la zona de procedencia 1=rural, 2=semi-urbana y 3=urbana (fichero examinar.sav). Además, si existe diferencia significativa, deseamos que se realicen

pruebas post-hoc.

## Desarrollo del ejemplo

Utilizando el SPSS para realizar el problema-ejemplo propuesto, la secuencia de ventanas del menú será las siguientes: **Analizar > Comparar medias > Anova de un factor.**



Seleccionaremos el item3 (prueba de inglés), como variable dependiente, y zona, como factor. Si pulsamos **contrastes** podemos realizar pruebas a priori (antes de realizar el experimento) con el estadístico t.



La idea es realizar una partición de la suma de cuadrados entre-grupos en componentes de tendencia o especificar contrastes a priori. Si seleccionamos **Polinómico** se puede contrastar la tendencia de la variable dependiente por los niveles ordenados del factor. Por ejemplo se puede contrastar una tendencia lineal (creciente o decreciente) en la prueba de inglés, a través de los niveles ordenados de mayor orden percibido. Con **orden** podemos especificar el orden del polinomio (lineal, cuadrático o cúbico, 4.º o 5.º).

Si seleccionamos coeficientes, vamos a realizar contrastes a priori por el estadístico t. Habrá que introducir un coeficiente por cada categoría del

factor. El orden de introducción de los coeficientes es importante porque corresponde al orden ascendente de las categorías del factor. Es aconsejable que la suma de coeficientes sea 0.

Hemos seleccionado el contraste polinómico lineal.

Por otra parte cuando seleccionemos los **contrastes post-hoc**, podremos elegir entre un amplio surtido de pruebas, que se clasifican de forma genérica en dos grandes grupos:

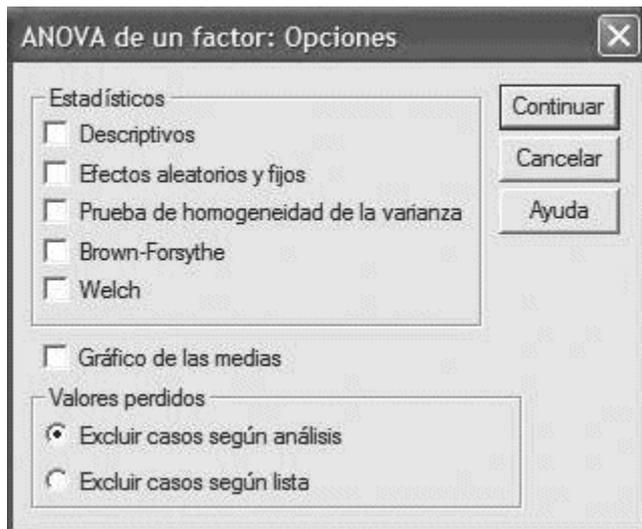


a) Aquellas que asumen las varianzas iguales:

DMS (Prueba de la mínima diferencia significativa), DUNCAN (Prueba del rango múltiple de Duncan), SNK (Prueba de Student-Neuman-Keuls), TUKEY (Método de Tukey), SCHEFFE (Método de Scheffé), etc.

b) Las pruebas de comparaciones múltiples que no asumen varianzas iguales son T2 de Tamhane, T3 de Dunnett, etc.

Finalmente si seleccionamos opciones podemos seleccionar estadísticos descriptivos, calcular el estadístico de Levene para contrastar la igualdad de las varianzas de grupo, hacer los gráficos de las medias y controlar el tratamiento de los valores perdidos.



### **Fichero de sintaxis**

La sintaxis de los subcomandos seleccionados será:

```

ONEWAY
item3 BY zona
/POLYNOMIAL= 1
/MISSING ANALYSIS
/POSTHOC = SCHEFFE T2 ALPHA(.05).

```

### **Resultados del comando ONEWAY**

#### **ANOVA de un factor**

ANOVA

#### **Prueba de inglés**

Suma de gl Media F Sig. **cuadrados cuadrática**

Inter-grupos (Combinados)

Término lineal No ponderado Ponderado Desviación 14,764 2 7,382 ,799  
,452

5,658 1 5,658 ,612 ,435  
5,953 1 5,953 ,644 ,424  
8,811 1 8,811 ,953 ,330

Intra-grupos 1358,630 147 9,242

Total 1373,393 149

Como se puede observar no hay diferencia significativa para las distintas zonas.

### **Pruebas post hoc COMPARACIONES MÚLTIPLES**

**Variable dependiente: prueba de inglés**

**(I) zona (J) zona Intervalo de confianza al 95% Diferencia de**

**Error<sub>Sig.</sub> Límite Límite medias (I-J) típico inferior superior Scheffé**

rural

semirural

urbana rural

semirural urbana

rural

semirural urbana

rural

semirural urbana

,76364 ,61490 ,464 -,7569 2,2842 ,46239 ,59099 ,737 -,9991 1,9238

-,76364 ,61490 ,464 -2,2842 ,7569

-,30125 ,62552 ,891 -1,8481 1,2456

-,46239 ,59099 ,737 -1,9238 ,9991 ,30125 ,62552 ,891 -1,2456 1,8481

Tamhane rural

semirural

urbana rural

semirural urbana

rural

semirural urbana

rural

semirural urbana

,76364 ,63621 ,549 -,7832 2,3104 ,46239 ,58303 ,814 -,9526 1,8774

-,76364 ,63621 ,549 -2,3104 ,7832

-,30125 ,61025 ,946 -1,7868 1,1843

-,46239 ,58303 ,814 -1,8774 ,9526 ,30125 ,61025 ,946 -1,1843 1,7868

**Subconjuntos homogéneos**

**Prueba de inglés**

**Subconjunto para alfa = .05**

**zona N 1**

Scheffé (a,b) semirural 44 4,0909

urbana 51 4,3922

rural 55 4,8545

Sig. ,459

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

a: Usa el tamaño muestral de la media armónica = 49,573.

b: Los tamaños de los grupos no son iguales. Se utilizará la media armónica de los tamaños de los grupos. Los niveles de error de tipo I no están garantizados.

La prueba a posteriori se ha realizado por motivos didácticos, al no resultar significativa la diferencia de medias de la variable item3 en función del factor zona.

## 2.6. El comando UNIANOVA

### Problema-ejemplo

Se desea saber si en los resultados de la prueba de inglés (item3), del ejemplo que venimos tratando, influyen las acciones de la zona de donde procede el alumno (factor fijo) y alguna categoría elegida al azar de la variable nivel (factor aleatorio).

**Desarrollo del ejemplo** Con el programa SPSS los cálculos que podemos realizar son muy amplios y pueden tener la siguiente secuencia de operaciones:

1. Elegir en los menús: **Analizar >Modelo lineal general > Univariante.**

2. Seleccionar una variable dependiente (en nuestro caso item3). Seleccionar los factores fijos (en nuestro caso zona), los factores aleatorios (en nuestro caso nivel) y los factores covariantes (en nuestro caso no tenemos) . Si queremos una variable de ponderación, utilizar ponderación MCP.

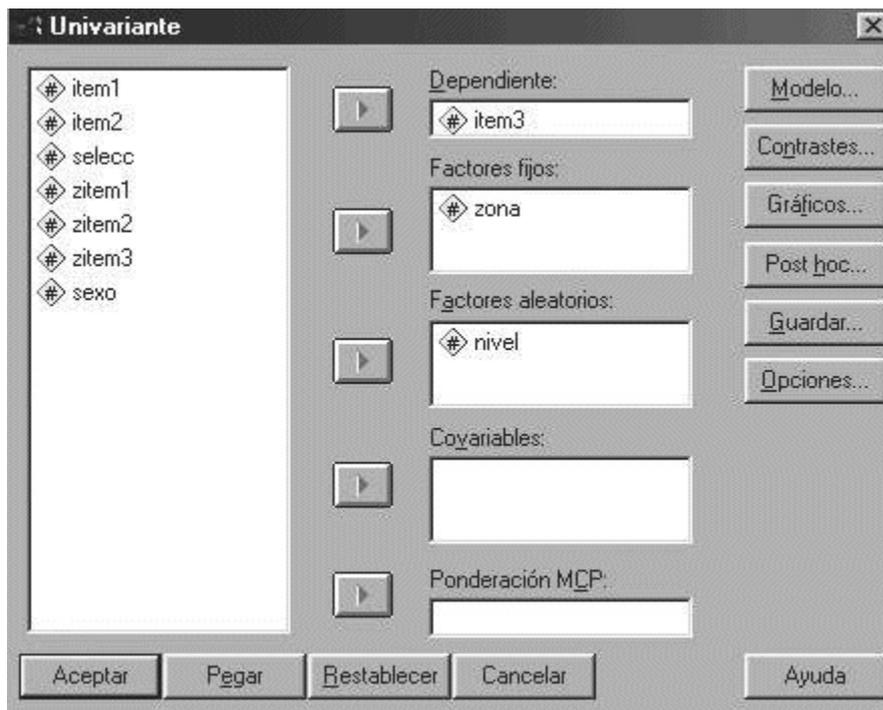
3. Seleccionar los contrastes para observar las diferencias entre los niveles de un factor.

4. Se puede seleccionar entre una gama amplia de pruebas post-hoc. Un grupo asumiendo homogeneidad de varianzas(Tukey, Scheffe, Duncan, DMS, Bonferroni, etc) y otras sin asumirla ( $T_2$  de Tamhane, $T_3$  de Dunnett, etc).

5. En opciones se pueden seleccionar un conjunto muy amplio de estadísticos: descriptivos, prueba de homogeneidad, falta de ajuste, diagrama de dispersión x nivel, gráfico de los residuos, etc.

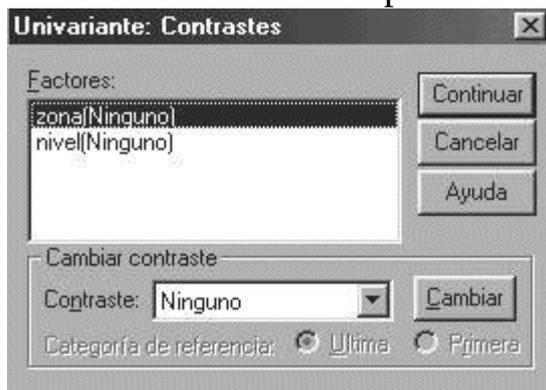
6. Especificar el modelo para seleccionar los efectos principales de factores, covariables y las interacciones entre factores. Indicar el modelo de suma de cuadrados (I, II, III y IV)( para más información ver manual del programa SPSS pág. 5).

El realizar los puntos 1 y 2 nos llevan a la siguiente ventana:



El botón **contrastes**

está relacionado con el punto 3 de la secuencia de operaciones:



Los puntos 4 y 5 se relacionan con el botón **post-hoc**.

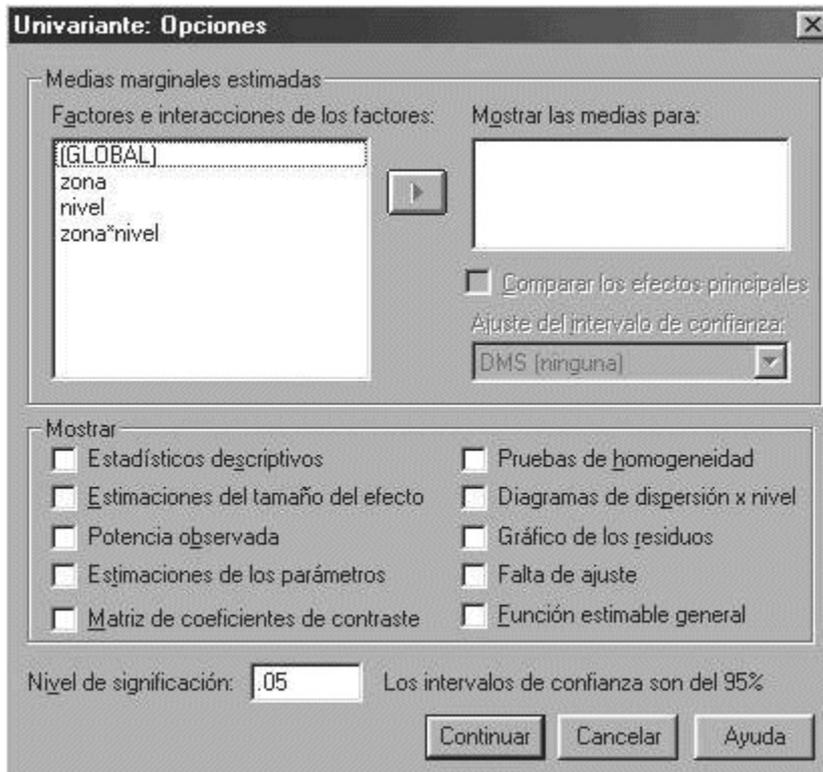
En este caso hemos seleccionado las pruebas que aparecen marcadas en la figura (DMS, para igualdad de varianzas y  $T_2$  de Tamhane, cuando se supone que no hay igualdad de varianzas).



El botón **modelo** está relacionado con el punto 6 de la secuencia.



Además, pulsando **opciones** accedemos a una variedad de opciones cuyo estudio detallado está fuera de los objetivos de esta publicación.



## Fichero de sintaxis

### UNIANOVA

```

item3 BY zona nivel
/RANDOM = nivel
/METHOD = SSTYPE(3)
/INTERCEPT = INCLUDE
/POSTHOC = zona ( LSD T2 )
/CRITERIA = ALPHA(.05)
/DESIGN = zona nivel zona*nivel .

```

### Resultado del comando unianova

PRUEBAS DE LOS EFECTOS INTER-SUJETOS **Variable dependiente: prueba de inglés**

**Fuente Suma de**

**cuadrados**  $gl^{Media}$  **F** **Significación** **tipo III** **cuadrática**

Intersección Hipótesis Error 2078,814 1 2078.814 169,874 ,000 107,975  
8.823 12,237 (a)

Zona Hipótesis Error 4,083 2 2,042 ,196 ,824 217,485 20,860 10,426(b)

Nivel Hipótesis Error

Zona \* nivel Hipótesis Error 90,081 7 10,657(c) 1,208 ,351 180,125  
16,902 10,893

152,497 14 10,893 1,245 ,252 1102,740 126 8,752(d)

a: ,847 MS(nivel) + ,153 MS(Error). b: ,782 MS(zona \* nivel) + ,218 MS(Error). c: ,890 MS(zona \* nivel) + ,110 MS (Error). d: MS(Error).

**MEDIA CUADRÁTICA ESPERADA (a, b)**

**Componente de la varianza**

**Fuente**

Intersección zona  
 nivel  
 zona \* nivel Error

**Var (nivel)**

**Var** Var (Error) **Término (zona \* nivel) cuadrático** 12,592 4,197 1,000

Intercept, zona

,000 4,356 1,000 zona 14,873 4,958 1,000  
 ,000 5,571 1,000  
 ,000 ,000 1,000

a: Para cada fuente, la media cuadrática esperada es igual a la suma de los coeficientes de las casillas por las componentes de la varianza, más un término cuadrático que incluye los efectos de la casilla Término cuadrático.

b: Las medias cuadráticas esperadas se basan en la suma de cuadrados tipo III.

**Pruebas post hoc COMPARACIONES MÚLTIPLES**

**Variable dependiente: prueba de inglés**

**(I) zona (J) zona**

**Diferencia entre Error** Sig.medias (I-J) típico

**Intervalo de confianza al 95%**

**Límite Límite inferior superior** DMS rural  
 semirural

urbana rural  
 semirural urbana  
 rural  
 semirural urbana  
 rural  
 semirural urbana  
 ,7636 ,59836 ,204 ,4624 ,57509 ,423  
 -,7636 ,59836 ,204  
 -,4205 1,9478  
 -,6757 1,6005

-1,9478 ,4205  
     -,3012 ,60870 ,622  
 -,4624 ,57509 ,423 ,3012 ,60870 ,622  
 -1,5058 ,9033  
 -1,6005 ,6757  
 -,9033 1,5058  
 Tamhane rural  
 semirural  
     urbana rural  
 semirural urbana  
 rural  
 semirural urbana  
 rural  
 semirural urbana  
 ,7636 ,63621 ,549 ,4624 ,58303 ,814  
 -,7636 ,63621 ,549  
 -,7832 2,3104  
 -,9526 1,8774  
 -2,3104 ,7832  
     -,3012 ,61025 ,946  
 -,4624 ,58303 ,814 ,3012 ,61025 ,946  
 -1,7868 1,1843  
 -1,8774 ,9526  
 -1,1843 1,7868

Basado en las medias observadas.

Como podemos observar no influyen los factores zona y nivel, e incluso la interacción zona\*nivel. Como consecuencia de este resultado la prueba post-hoc, que se ha puesto por motivos didáctico, tampoco da significativa.

### **EJERCICIOS DE AUTOCOMPROBACIÓN**

1. La siguiente tabla muestra la edad (x) y la presión sanguínea (y) de 12 personas.

$x_i$  56 42 72 36 63 47 55 49 38 42 68 60     $y_i$  147 125 160 118 149 128 150  
 145 115 140 152 155

Obtégase:

a) El coeficiente de correlación lineal de la muestra.

- b) La estimación de la recta de regresión de  $y$  sobre  $x$ .  
 c) El intervalo de confianza al 95% de la pendiente y la ordenada en el origen de la recta.

2. Para acceder a la carrera de Educación se plantea una prueba selectiva que se supera con 34 puntos sobre 100. Elegida una muestra de 10 alumnos se toma los datos relativos a la prueba selectiva ( $x$ ), número de días perdidos ( $y$ ) y la calificación final del curso de acceso ( $z$ ). Se desea obtener la ecuación de regresión lineal múltiple de  $z$  sobre  $x$  e  $y$ .

$x_i$  50 35 55 35 90 90 60 60 55 65  
 $y_i$  2 1 1 5 1 2 3 2 1 2  
 $z_i$  53 61 68 11 79 54 48 71 53 57

3. Tres profesores de Estadística atienden respectivamente a tres grupos de 10, 12 y 9 estudiantes a los que han de calificar al final de curso. Si las calificaciones dadas, en una escala de 0 a 10 puntos, son las expresadas en la tabla adjunta. ¿Existe diferencia significativa entre las medias de las calificaciones obtenidas por los tres grupos?  $\alpha=0,05$

**Profesor A** 6 4 5 7 9 7 4 2 3 5 **Profesor B** 5 7 4 8 2 5 6 1 2 6 9 4  
**Profesor C** 8 3 7 4 4 5 7 6 7

4. Se comparó la velocidad de escritura en palabras por minuto de 24 estudiantes de cuatro academias elegidos al azar de una población donde la variable está distribuida normalmente. Verificar si podemos suponer que los métodos aplicados por cada academia son esencialmente iguales. Si existe diferencia al nivel del 5% comparar academias.

**academia A** 50 50 55 60 45 55  
**academia B** 55 60 65 55 70 65  
**academia C** 50 65 75 55 60 65  
**academia D** 70 80 65 70 75 60

### SOLUCIÓN A LOS EJERCICIOS DE AUTOCOMPROBACIÓN

1. El fichero de sintaxis para resolver el problema propuesto es el siguiente:

\* ANÁLISIS DE REGRESIÓN.  
 data list free/x y .  
 begin data  
 56 147  
 42 125  
 72 160

36 118  
63 149  
47 128  
55 150  
49 145  
38 115  
42 140  
68 152  
60 155  
end data.  
execute.

```
REGRESSION  
/MISSING LISTWISE  
/STATISTICS COEFF OUTS CI R ANOVA /CRITERIA=PIN(.05)  
POUT(.10)  
/NOORIGIN  
/DEPENDENT y  
/METHOD=ENTER x.
```

El resultado será:

**RESUMEN DEL MODELO**

**Modelo R R cuadrado R cuadrado Error típ. corregida de la estimación**

1 ,896(a) ,803 ,783 7,01760

a: Variables predictoras: (Constante), x

Como se puede observar al tratarse de una regresión lineal simple el coeficiente de correlación será  $R=0,896$ .

**COEFICIENTES(a)**

**Coefficientes no Coeficientes estandarizados estandarizados**

**Intervalo de confianza para B al 95%**

**Modelo B Error típ. Beta t Sig. Límite Límite inferior superior**

1 constante 80,778 9,544 8,464 ,000 59,513 102,043

x 1,138 ,178 ,896 6,386 ,000 ,741 1,535

a: Variable dependiente: y.

La recta de regresión será:  $y=80,778+1,138 x$

El intervalo de confianza al 95% para la pendiente será: [0,741;1,535] y para la ordenada en el origen es: [59,513;102,043].

2. El fichero de sintaxis será:

\* ANÁLISIS DE REGRESIÓN.

data list free/x y z.

begin data

50 2 53

35 1 61

55 1 68

35 5 11

90 1 79

90 2 54

60 3 48

60 2 71

55 1 53

65 2 57

end data.

execute.

REGRESSION

    /MISSING LISTWISE

/STATISTICS COEFF OUTS R ANOVA /CRITERIA=PIN(.05)

POUT(.10) /NOORIGIN

/DEPENDENT z

/METHOD=ENTER x y.

Los resultados serán:

RESUMEN DEL MODELO

**Modelo**

**R**

**R cuadrado**

**R cuadrado Error típ. corregida de la estimación**

1 ,907(a) ,822 ,771 8,77493

a: Variables predictoras: (Constante), y, x.

COEFICIENTES(a)

**Coefficientes no estandarizados estandarizados**

**Modelo B Error típ. Beta t Sig.**

1 constante 63,854 12,617 5,061 ,001

x ,250 ,164 ,257 1,526 ,171

y -11,608 2,484 -,789 -4,673 ,002

a: Variable dependiente: y.

Por tanto la recta de regresión es:  $z=63,854+0,250 x-11,608 y$

3. El fichero de sintaxis será:

\* ANÁLISIS DE LA VARIANZA. data list free/x f.

begin data

6,00 1  
4,00 1  
5,00 1  
7,00 1  
9,00 1  
7,00 1  
4,00 1  
2,00 1  
3,00 1  
5,00 1  
5,00 2  
7,00 2  
4,00 2  
8,00 2  
2,00 2  
5,00 2  
6,00 2  
1,00 2  
2,00 2  
6,00 2  
9,00 2  
4,00 2  
8,00 3  
3,00 3  
7,00 3  
4,00 3  
4,00 3  
5,00 3  
7,00 3  
6,00 3  
7,00 3  
end data.  
execute.

ONEWAY  
x BY f  
/MISSING ANALYSIS.

El resultado es:

ANOVA x  
Suma de gl Media F Sig. **cuadrados cuadrática**

Inter-grupos 2,903 2 1,451 ,311 ,735

Intra-grupos 130,517 28 4,661

Total 133,419 30

Como se puede observa no hay diferencia significativa entre las medias de las calificaciones obtenidas por los alumnos de los tres profesores. 4. El fichero de sintaxis para realizar el ejercicio es:

\* ANÁLISIS DE LA VARIANZA.

data list free/x f.

begin data

50 1

50 1

55 1

60 1

45 1

55 1

55 2

60 2

65 2

55 2

70 2

65 2

50 3

65 3

75 3

55 3

60 3

65 3

70 4

80 4

65 4

70 4

75 4

60 4

end data.

execute.

ONeway

x BY f

/MISSING ANALYSIS

/POSTHOC = SCHEFFE T2 ALPHA(.05).

Cuyo resultado es:

ANOVA x

Suma de gl Media F Sig. **cuadrados cuadrática**

Inter-grupos 919,792 3 306,597 6,426 ,003

Intra-grupos 954,167 20 47,708

Total 1873,958 23

Resultado, como vemos, significativo, por tanto, se puede plantear la diferencia entre los grupos:

COMPARACIONES MÚLTIPLES **Variable dependiente: x**

**Intervalo de confianza al 95%**

**(I) f (J) f<sup>Diferencia de</sup> Error típico<sub>medias (I-J)</sub>**

**Sig.**

**Límite inferior**

**Límite superior** Scheffé 1,00 1,00

2,00 -9,16667 3,98783

3,00 -9,16667 3,98783

4,00 -17,50000(\*) 3,98783

,187 -21,3248 2,9914 ,187 -21,3248 2,9914 ,003 -29,6581 -5,3419

2,00 1,00 9,16667 3,98783

2,00

3,00 ,00000 3,98783

4,00 -8,33333 3,98783

,187 -2,9914 21,3248

1,000 -12,1581 12,1581 ,257 -20,4914 3,8248

3,00 1,00 9,16667 3,98783

2,00 ,00000 3,98783

3,00

4,00 -8,33333 3,98783

,187 -2,9914 21,3248 1,000 12,1581

-12,1581

,257 -20,4914 3,8248

4,00 1,00 17,50000(\*) 3,98783

2,00 8,33333 3,98783

3,00 8,33333 3,98783

4,00

Tamhane<sub>1,00</sub> 1,00

2,00 -9,16667 3,27024  
3,00 -9,16667 4,16667  
4,00 -17,50000(\*) 3,59398  
,003 5,3419 29,6581 ,257 -3,8248 20,4914 ,257 -3,8248 20,4914

,109 -19,8927 1,5593  
,303 -23,5069 5,1736  
,005 -29,4664 -5,5336

2,00 1,00 9,16667 3,27024

2,00  
3,00 ,00000 4,34613  
4,00 -8,33333 3,80058 ,109 -1,5593 19,8927  
1,000 -14,6098 14,6098 ,282 -20,8083 4,1417  
3,00

4,00 1,00 9,16667 4,16667 2,00 ,00000 4,34613 3,00  
4,00 -8,33333 4,59468

1,00 17,50000(\*) 3,59398  
2,00 8,33333 3,80058  
3,00 8,33333 4,59468  
4,00

\* La diferencia entre las medias es significativa al nivel .05.  
,303 -5,1736 23,5069 1,000 -14,6098 14,6098  
,473 -23,4882 6,8215

,005 5,5336 29,4664  
,282 -4,1417 20,8083  
,473 -6,8215 23,4882

Las pruebas de Scheffé suponiendo igualdad de varianzas y la de Tamhane sin asumir tal condición, arrojan el mismo resultado hay diferencias entre las academias 1 y 4.

### **BIBLIOGRAFÍA**

AFIFI, A. A. AND CLARK, V. (1996): *Computer-aided Multivariate Analysis*, 3/e.. London: Chapman and Hall.

ALLISON, P. D. (1999): *Multiple regression*. California: Sage.

AMÓN, J. (1991): *Introducción al análisis multivariante (cálculo matricial)*. Barce-lona: PPU.

ARCE, CONSTANTINO; REAL, EULOGIO (2002): *Introducción al análisis*

*estadístico con SPSS*. PPU. Barcelona.

BISQUERRA, R. (1989): *Introducción Conceptual al Análisis Multivariable*.

Barce-lona: PPU. BISQUERRA, R. (1987): *Introducción a la estadística aplicada a la investigación educativa*. Barcelona: PPU.

CALOT, G. (1974): *Curso de Estadística Descriptiva*. Paraninfo. Madrid.

CAMACHO ROSALES, JUAN (2002): *Estadística con SPSS versión 11 para Windows*. Ra-Ma, Librería y Editorial Microinformática. Madrid.

CEA, M. A. (2002): *Análisis multivariable. Teoría y práctica en la investigación social*. Madrid: Síntesis.

C OCHRAN, W. G. (1971): *Técnicas de Muestreo*. México: CECSA.

CUADRAS, C. M. (1991): *Métodos de Análisis Multivariante*. Barcelona: Eunibar.

CUADRAS, C. M.; ECHEVERRÍA, B.; MATEO, J.; SÁNCHEZ, P. (1991): *Fundamentos de Estadística*. Barcelona: PPU.

DOMÉNECH, J. M. y RIBA, M. D. (1985): *Métodos estadísticos. Modelo lineal de regresión*. Barcelona: Herder.

ETXEBERRIA, J. (1999): *Regresión múltiple*. Madrid: La muralla.

FERRÁN ARANAZ, MAGDALENA (2002): *Curso de SPSS para Windows*. McGraw-Hill/ Interamericana de España. Madrid.

F OX, J. (1991): *Regression diagnostics*. Newbury Park: Sage.

GUILLÉN, M. F. (1992): *Análisis de regresión múltiple*. Madrid: CIS,

*Cuaderno metodológico nº 4*. HAIR, ANDERSON, TATHAM, BLACK (1999):

*Análisis multivariante. 5/e*. Madrid: Prentice Hall. HARDY, M. A. (1993):

*Regression with dummy variables*. Newbury Park: Sage. HOPE, K. (1972):

*Métodos de Análisis Multivariate*. Madrid: Inst. Est. Polit.

JOHNSON R. Y WICHERN D. (1988): *Applied multivariate statistical analysis. 2/e*. New Jersey: Prentice-Hall international.

K ENDALL, M. G. (1975): *Multivariate Analysis*. Londres: Griffin.

KIELLERER, H. (1960): *La Estadística en la vida económica y social*.

Editorial: Alianza. LEBART, L. Y OTROS (1985): *Tratamiento estadístico de datos*. Barcelona: Boixerau editores.

LIZASOAIN HERNÁNDEZ, LUIS; JOARISTI OLARRIAGA, LUIS (2003): *Gestión y análisis de datos con SPSS*. Thomson Paraninfo. Madrid.

MARTÍN F. G. (1994): *Introducción a la Estadística Económica y Empresarial*. Madrid: AC.

MARTÍN PLIEGO, F. J. (1994): *Introducción a la estadística económica y empresarial (Teoría y práctica)*. Madrid: Editorial AC.

MARTÍN, M. F.; FERNÁNDEZ, R.; SEISDEDOS, A. (1985): *ticas para las ciencias de la conducta*. Salamanca: Universidad pontificia.

MILES, J. Y SHEVLIN, M. (2000): *Applying regression and correlation*. Newbury Park: Sage. MIRAS J. (1985): Elementos de muestreo para poblaciones finitas. Madrid: INE.

PARDO MERINO, ANTONIO; RUIZ DÍAZ, MIGUEL ÁNGEL (2002): *SPSS 11. Guía para el análisis de datos*. McGraw-Hill/ Interamericana de España. Madrid.

PEÑA D. (1992): *Estadística, Modelos y Métodos. Volumen I*. Alianza Universidad Textos. Madrid.

PÉREZ, CÉSAR (2001): *Técnicas estadísticas con SPSS*. Pearson Educación. Madrid. RÍOS, S. (1977): *Métodos Estadísticos*. Madrid: Ed. del Castillo.

RUIZ-MAYA, L. (1986): *Métodos estadísticos de investigación (Introducción al análisis de la varianza)*. Madrid: INE.

SÁNCHEZ CARRIÓN, J. J. (1984): *Introducción a las técnicas de análisis multivariable aplicadas a las ciencias sociales*. Madrid: Centro de Investigaciones Sociológicas. SÁNCHEZ CRESPO J. L. (1971): Principios elementales de muestreo y estimación de proporciones. Madrid: INE.

SÁNCHEZ CRESPO J. L. (1976): Muestreo de poblaciones finitas aplicado al diseño de encuestas. Madrid: INE.

SÁNCHEZ CRESPO J. L. y DE PRADA J. (1990): Ejercicios y problemas resueltos de muestreo en poblaciones finitas. Madrid: INE.

SCHEAFFER, R.; MENDENHALL, W.; OTT, L. (1987): Elementos de muestreo. México: Grupo Editorial Iberoamérica.

SPIEGEL M. R. (1991): *Estadística*. Madrid: McGraw Hill.

VISAUTA VINACUA, B. (2002): *Análisis estadístico con SPSS 11.0 para Windows*. McGraw-Hill/ Interamericana de España. Madrid.

El texto nace de una reflexión sobre las necesidades instrumentales de los estudiantes o profesionales para abordar el proceso de tratamiento de datos. Sabemos que la estadística es un *instrumento básico en la investigación*, y constituye *el lenguaje de expresión y formalización de la actividad investigadora*. Pero, indudablemente, la estadística necesita la ayuda de la Informática para realizar su labor. Este apoyo se traduce en programas estadísticos de tratamiento de datos, siendo el SPSS uno de los más utilizados en el ámbito de las ciencias sociales y de la salud.

El libro aborda los conocimientos básicos de estadística descriptiva e inferencial, introduce al lector en el diseño de experimentos, con el análisis

de la varianza, y completa los procesos de decisión estadística con una breve introducción al muestreo. Todo ese conocimiento estadístico se introduce de forma gradual con el apoyo de casos prácticos realizados con el programa SPSS.

**Juan Antonio Gil Pascual** , profesor titular de Métodos de Investigación en Educación en el Departamento M.I.D.E. de la Facultad de Educación de la UNED. Matemático (Estadística e Investigación Operativa) y doctor en Educación. Con dilatada docencia en todos los niveles educativos y como experto en estudios de opinión en el Área de Recursos Humanos de Telefónica. Especialista en análisis y tratamiento de datos de carácter multivariante y profundo conocedor de paquetes estadísticos (SPSS,R). Autor de publicaciones y ponencias de aplicaciones de metodologías estadísticas al ámbito social. Colaborador en numerosos proyectos de investigación como analista de datos. Premio Nacional de Investigación Educativa (1998).

ISBN: 978-84-362-5264-4 84216

## **Editorial**

9 788436 252644

0184216EP01A02