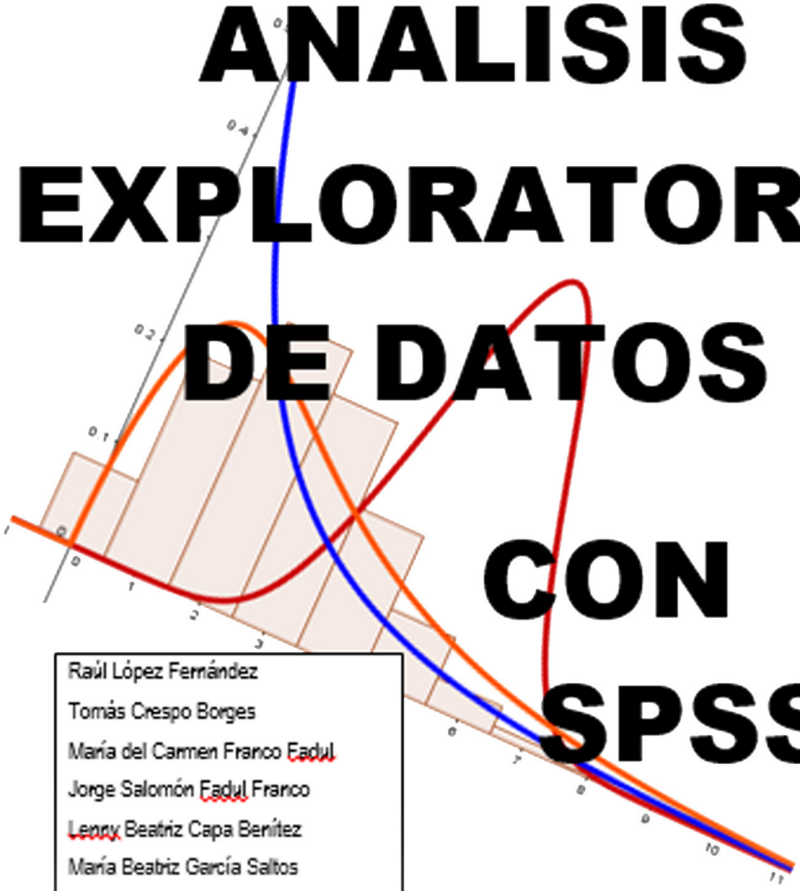
ANÁLISIS EXPLORATORIO DE DATOS CON SPSS

UMET
UNIVERSIDAD METROPOLITANA

Lenny Beatriz Capa Benítez
María Beatriz García Saltos
Erick Crespo Hurtado
Diana E. Palmero Urquiza
Raúl López Fernández
Tomás Crespo Borges
María del Carmen Franco Fadul
Jorge Salomón Fadul Franco

ANÁLISIS EXPLORATORIO DE DATOS CON SPSS



Raúl López Fernández

Tomás Crespo Borges

María del Carmen Franco Fadul

Jorge Salomón Fadul Franco

Lenny Beatriz Capa Benítez

María Beatriz García Saltos

Erick Crespo Hurtado

Diana E. Palmero Urquiza

Diseño de carátula y composición de textos: D. I. Yunisley Bruno Díaz

Corrección: MSc. Dolores Pérez Dueñas

Dirección editorial: Dr. C. Jorge Luis León González

ISBN: 978-959-257-493-9

Editorial: “Universo Sur”.

Universidad de Cienfuegos. Carretera a Rodas, Km 3 ½.

Cuatro Caminos. Cienfuegos. Cuba.

CP: 59430

E-mail: eus@ucf.edu.cu

Índice

Capítulo I. Introducción al tema. A.E.D con SPSS

1.1. ¿Qué es A.E.D.?

Una respuesta inmediata es que se trata de la abreviatura de Análisis Exploratorio de Datos (A.E.D.) o como se expresa en inglés, Exploratory Data Analysis (E.D.A.). Pero en realidad es mucho más, como su nombre lo indica, se trata de un enfoque que prioriza el análisis del dato y sobre este particular existen múltiples criterios.

Según Monterde & Perea (1991, p. 10), A.E.D es, "por una parte, una perspectiva o actitud sobre el análisis de datos, en la que se exhorta a que el investigador adopte una actitud activa en y hacia el análisis de los mismos, como un medio para sugerir nuevas hipótesis de trabajo. Por otra parte, se compone de un renovado utillaje conceptual e instrumental respecto a lo que podríamos llamar Estadística Descriptiva "clásica", con el fin de optimizar la cantidad de información que los datos recogidos puedan ofrecer al investigador, bien mediante novedosas representaciones gráficas, bien a base de reducir la influencia de las puntuaciones extremas en los estadísticos con el empleo de, los que por ello se ha convenido en llamar, "estadísticos resistentes".

Ante lo expuesto surge una pregunta ¿cómo se inserta lo que ya se conoce de estadística, aunque sea elemental en esta concepción? La respuesta no puede darse en las pocas palabras de un párrafo, pero la lectura del libro, desde el desarrollo de la teoría y la ejemplificación correspondiente puede llevar a comprender la concepción de A.E.D. y sus similitudes y diferencias con la estadística clásica. Esta es la mayor aspiración de los autores.

Cualquier lector coincidirá en que no se exagera si se dice que el objeto de la Estadística es el estudio de métodos científicos para organizar, presentar y analizar datos estadísticos (informaciones), esto es cierto, pero el problema está en cómo comenzar a organizar los datos, quien haya estudiado un curso elemental

de Estadística recordará la prioridad que se da a las tablas de frecuencia, al estudio de modelos como la distribución normal o la correlación lineal que describen de una manera simple el comportamiento de los datos. En general estos modelos, (aunque A.E.D no los desecha), representan estructuras a gran escala que resumen las relaciones entre todos los datos y actualmente, como ha expresado Silva Rodríguez (2002), *"contamos con más de 30 años de desarrollo de esas nuevas teorías, agrupadas en poderosos paquetes computarizados"*, que liberan a los investigadores de la búsqueda minuciosa de modelos, para interesarse más en el entendimiento de las estructuras subyacentes en grandes conjuntos de datos; esta es una primera idea de la concepción del A:E.D que se seguirá desarrollando a través del libro.

Desde sus orígenes, a partir de los estudios de Tukey [1] en 1977, A.E.D ha tenido como finalidad el examen de los datos previo a la aplicación de cualquier técnica estadística para alcanzar primero un entendimiento básico de los mismos y de las relaciones existentes entre las variables analizadas.

Es decir, cualquier cálculo, (promedios, desviaciones, correlaciones, etc.) debe estar precedido por un análisis, particularmente visual de los datos, dicho de otro modo, mientras la Estadística Descriptiva clásica se ocupa de recoger, ordenar y representar los datos en forma de tablas, agrupándolos por intervalo y calculando estadísticos basados principalmente en la distancia y con datos centrados en la media (promedio); el A.E.D. se preocupa primero por detectar anomalías y errores en las distribuciones univariadas de los datos, intentando descubrir en ellos patrones o modelos, pero empleando variadas técnicas gráficas y buscando estimadores no paramétricos o estimadores libres de distribución o simplemente estimadores robustos, según el término acuñado por Box[2] en 1953, tratando de llevar el estudio de la información que se tiene hacia una modelización más completa que la establecida por la Estadística Clásica, basados principalmente en el orden y centrados en la mediana.

El paquete estadístico SPSS (Statistical Package for the Social Sciences) ofrece toda una gama de posibilidades a partir de simples diálogos dinámicos que cubren tanto las exigencias de la llamada Estadística Clásica como las de A.E.D.

Por el momento vale el siguiente enunciado como un postulado comprensible para todos:

Una buena gráfica informa más que un conjunto de números disgregados.

Esta es la esencia del A.E.D., *permitir que los datos hablen y a partir de ellos encontrar los patrones y modelos que le corresponden,* con esto se logra que en muchas situaciones, el A.E.D puede preceder a una situación de inferencia formal, mientras que en otras, puede sugerir preguntas y conclusiones que se podrían confirmar con un estudio adicional, por esto el A.E.D puede ser una herramienta de utilidad en la generación de hipótesis, conjeturas y preguntas de investigación acerca de los fenómenos de los que los datos fueron obtenidos.

En la investigación relacionada con las ciencias sociales, donde influyen numerosas variables y donde los datos no son siempre numerosos, las concepciones del A.E.D. bien utilizadas se convierten en instrumentos que complementan los diseños de investigación y dan validez, confiabilidad y rigor científico a los resultados.

1.2. Algo más sobre variables y datos

De lo expresado sobre los propósitos del A.E.D. se puede inferir que se deben emplear técnicas estadísticas cuya finalidad es conseguir un entendimiento básico de los datos y de las relaciones existentes entre las variables analizadas. Ello lleva aparejado la existencia de métodos sistemáticos (generalmente sencillos) para organizar y preparar los datos, detectando los posible fallas en el diseño y recogida de los mismos, para ellos se debe dar tratamiento y evaluación de datos ausentes (missing), identificar los casos atípicos (outliers) y comprobar los supuestos que

subyacentes en la mayor parte de las técnicas multivariantes tradicionales, tales con normalidad, linealidad y homocedasticidad entre otras[1].

Desde sus primeras versiones, SPSS brinda al usuario toda la información sobre missing, outliers, normalidad, linealidad y homocedasticidad de los datos.

Los estudiosos del A.E.D. convienen en que se debe seguir las siguientes etapas con el tratamiento de los datos:

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

Estas etapas se pueden seguir en el procesamiento de datos utilizando el asistente SPSS porque a partir de ellas es posible hacer análisis más detallados de los mismos, así, las informa-

1 Posteriormente se estudiará el significado de estas exigencias de la Estadística Clásica, principalmente en el empleo del Analysis of Variance (ANOVA), técnica estadística que permite hacer la inferencia acerca de si tres o más muestras podrían venir de poblaciones que tienen la misma media (promedio); específicamente, si las diferencias entre las muestras es producto de la casualidad.

ciones (datos) necesarias para la investigación pueden ser obtenidas de fuentes primarias o secundarias, de modo que una primera clasificación de los datos puede ser:

Datos primarios: son aquellos que no han sido recopilados anteriormente por parte de personas u organismos que trabajan en la obtención y elaboración de datos y que, por consiguiente, son observados y anotados por el investigador, a partir de las fuentes directas. Ejemplos las cantidades de asistencias a clases de cada alumno controladas por el investigador, la tabulación de las encuestas, la velocidad, el grado de salinidad de distintas muestras de agua, los datos correspondientes a la evolución de los pacientes con determinada enfermedad, etc.

Datos secundarios: se trata de los que ya han sido recopilados y elaborados y que provienen principalmente de publicaciones oficiales o privadas o de entidades que elaboran estadísticas. Las fuentes de las cuales se pueden obtener los datos secundarios son muy variadas, pero hay que garantizar la confiabilidad de las mismas, algunos ejemplos pueden ser: los datos que se ofrecen en sitios Webs de diferentes Ministerios e instancias gubernamentales, datos de Organizaciones Internacionales como la UNESCO, la CEPAL, la OMS, etc.

Una vez recogidos los datos cada uno está expresado en determinadas unidades: centímetros, kilogramos, asistencias, tipo de distractor etc. y como con ellos no se ha realizado ninguna operación tales conteos, suma, cálculo de promedio, etc., se está entonces en presencia de *datos primitivos (o brutos)* los cuales *NUNCA* deben desecharse. El uso de las mayúsculas marca la intención, porque lamentablemente muchas veces estos datos son eliminados por los investigadores o no se protegen suficientemente los medios de almacenamiento y esto trae funestas consecuencias en los finales de la investigación.

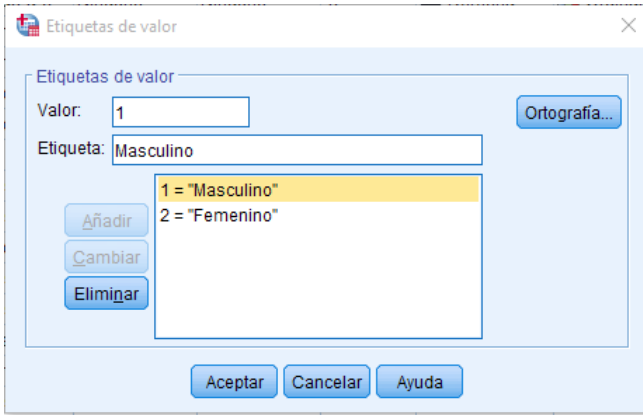
También pueden ser clasificados los datos adoptando otros criterios, como es el caso de asociarlo al tipo de variable que los producen:

Datos cualitativos: Corresponden a las mediciones de variables cualitativas, generalmente no aparecen en forma numérica, sino como categorías o atributos X. Pueden distinguirse dos tipos de estos datos, los que expresan mediciones en las que puede haber un orden subyacente (variable ordinal) y las que no admiten un orden (variable nominal).

# del alumno	Nacionalidad	Calificaciones
1	Venezolana	Bien
2	Nicaragüense	Excelente
3	Venezolana	Excelente
4	Hondureña	Bien
5	Venezolana	Regular

Ejemplos: De 5 alumnos se tiene la información que se muestra en la tabla:

En la tabla anterior la variable *nacionalidad es nominal y calificaciones* es ordinal. En ocasiones las variables cualitativas se codifican numéricamente, pero tales números no significan orden. En SPSS tales codificaciones son habituales, ejemplo:



En este caso, 1 significa Masculino y 2 Femenino, pero, aunque se expresa mediante números, la variable es cualitativa nominal.

Datos cuantitativos: Los correspondientes a las mediciones de variables cuantitativas y por lo tanto aparecen en forma numérica con el significado matemático del mismo, por ejemplo, los datos de las estaturas, peso, notas en escala de 100 puntos, entre otros.

En los datos cuantitativos se pueden diferenciar perfectamente los que están asociados a *variables cuantitativas discretas* – que frecuentemente son el resultado de contar y, por tanto, toman solo valores enteros – y los asociados a *alta variables cuantitativas continuas*, que resultan de medir y pueden contener cifras decimales. En estos últimos se deben distinguir por la escala en los que están expresados. La identificación de los tipos de datos que se desean procesar es fundamental para preparar la base de datos en SPSS.

1.3. Preparar los datos para hacerlos accesibles a cualquier técnica estadística

El trabajo de mesa

Frecuentemente los investigadores se enfrentan al problema de cómo codificar, empleado el término en el sentido amplio de sus sinónimos: recopilar, catalogar, agrupar, reunir, juntar, recoger, las mediciones u observaciones que han realizado al manipular las variables que se estudian y para ello debe retomarse lo planteado respecto a la importancia de la precisión los datos primarios y la atención que se debe dar a las respuestas de las preguntas: "¿Qué datos se necesitan? ¿Para qué se necesitan? ¿Para qué transformarlos? y ¿cómo transformarlos?".

La respuesta a ¿Qué datos se necesitan?" tiene que ser exacta y precisa, la escritura en negrita indica la obligada correspondencia con la pregunta. ¿Para qué se necesitan? Indica que, recopilar menos datos de los necesarios trae funestas consecuencias a la hora de constatar resultados, pero, para los que solicitan datos a ciegas, bajo el lema de que es mejor que sobren, se les advierte que los datos innecesarios vician la investigación en el proceso de recogida de datos e influyen principalmente en el

momento en que los encuestados brindan su información, disminuyendo la confiabilidad de los instrumentos utilizados.

¿Para qué transformarlos? Está relacionado con la pregunta ¿Para qué se necesitan los datos? ¿Qué inferencia o conclusión se necesita sacar con ellos?, pero la intención de la pregunta está más orientada a la elección del modelo estadístico, el estadígrafo, o la prueba estadística que se va a utilizar. Por no responder correctamente a esta pregunta, con frecuencia aparecen cálculos de promedios con datos enteros que dan resultado como 5,3 alumnos; pruebas chi-cuadrado con más del 20% de frecuencias esperadas inferiores a 5, porque las frecuencias observadas han sido inferiores a 10; la aplicación de pruebas que exigen normalidad de los datos aplicadas a muestras pequeñas o con datos en escala ordinal, en fin, errores estadísticos que por supuesto no dependen del asistente estadístico utilizado sino de los datos suministrado y la elección de los métodos que ha hecho el usuario; todo estos errores se pueden evitar cuando se desarrolla un buen trabajo de mesa.

1.4. Del trabajo de mesa al almacenamiento en SPSS

Introducción a la aplicación SPSS

La aplicación (o paquete estadístico) SPSS (Statistical Package for the Social Sciences), (Paquete Estadístico para las Ciencias Sociales) aunque también aparece referido como Statistical Product and Service Solutions (Producto Estadístico y Solución de Servicios) es un paquete estadístico de Análisis de Datos con más de 20 años de aplicación principalmente a la investigación de las Ciencias Sociales y Económicas. El mismo responde al funcionamiento de todo programa que lleva a cabo análisis estadísticos:

1. Pasados de los datos seleccionados para analizar a la confección de un fichero con las características de la aplicación.
2. Ejecución de un conjunto de órdenes, capaces de realizar desde un simple análisis descriptivo hasta análisis

multivariante de datos, (análisis discriminante, análisis de regresión, clúster, análisis de varianza, etc.), estudios de series temporales, tablas de frecuencias y gráficos diversos.

3. Obtener un conjunto de resultados de tipo estadístico que la aplicación ofrece como salida y que el investigador debe interpretar.

Precisando lo expresado:

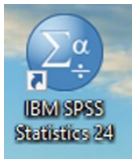
Los pasos a seguir para llevar a cabo un análisis de tipo estadístico son los siguientes:

1. Recoger la información del problema que se desee investigar y tenerla organizada generalmente en papel, preferiblemente en forma de tablas y con las especificaciones de las variables.
2. Grabar esa información en un archivo de datos correspondiente al programa que se va a usar, en el caso de SPSS en un archivo que tiene el nombre que le asigne y que por defecto se le asigna la extensión,sav.
3. Sobre tal archivo de datos se lleva a cabo el análisis con SPSS, usando diferentes procedimientos que como se ve en explicaciones posteriores se seleccionan de distintos menús.
4. Los resultados de tales análisis son volcados a un visor de resultados en el que su visualización y edición son más cómoda, y desde el que se pueden guardar en un fichero con el nombre que se desee el usuario, pero de extensión. spv.
5. El investigador interpreta los resultados y extrae las conclusiones que considere relevantes, y con eso se cierra el ciclo de A.E.D.

1.5. La ventana principal de SPSS: el editor de datos de SPSS

El paquete SPSS, desde la versión 7, es un paquete adaptado al entorno WINDOWS (Existe también PSPP que ha sido considerado un clon de código abierto que emula todas las posibilidades del SPSS), con lo cual la forma de ejecutarlo es a través

de ventanas en las que se despliegan menús, de los que se pueden elegir distintas opciones, por tanto, es a través de un entorno de tipo gráfico desde donde se resuelven los problemas, y no mediante comandos (aunque también se pueden utilizar) como antiguamente se hacía en los paquetes estadísticos más usados.

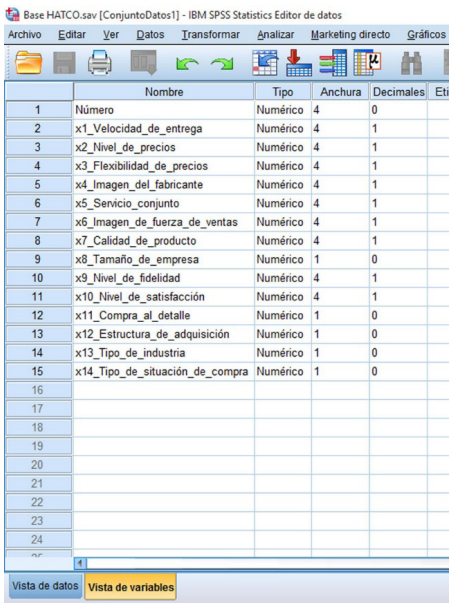


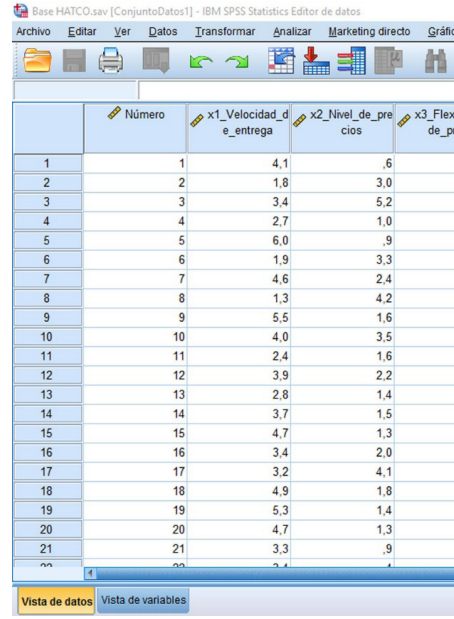
Por lo que la forma de iniciar la ejecución del programa SPSS es pinchando dos veces con el ratón (pinchar se utiliza como sinónimo de hacer clic con el botón principal del ratón, según el diccionario de la Real Academia de la Lengua Española los sinónimos de pinchar son estimular, impulsar, excitar, incitar) en el icono de SPSS que generalmente se encuentra en el escritorio en forma de enlace o en el menú de inicio dentro del apartado de programa. Una de las primeras tareas que tendrá que hacer el usuario de SPSS será localizar la posición del icono y adaptarlo a su gusto y necesidades.

Haciendo clic dos veces sobre el icono, se abre la ventana principal de SPSS que es el Editor de datos de SPSS, también la llaman ventana principal de SPSS.

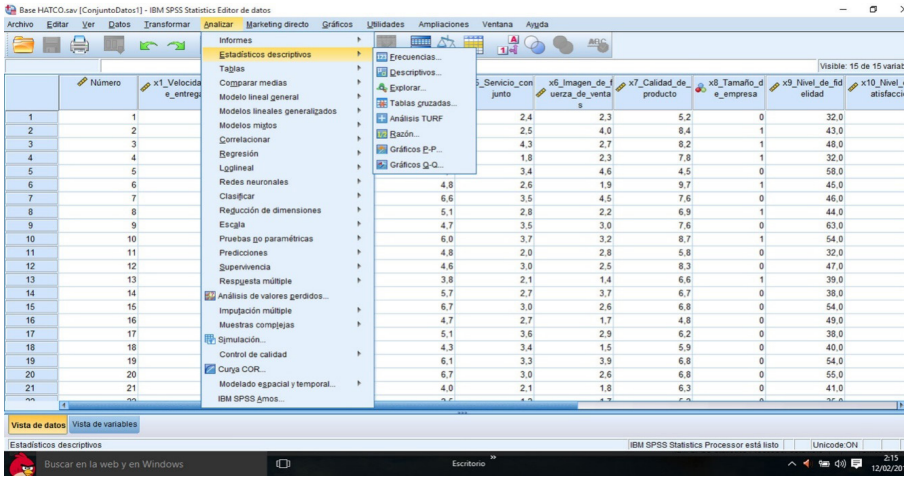
Esta ventana tiene dos versiones o vistas: vista de datos y vista de variables. En la figura adjunta se muestra a la derecha la vista de datos; en ella aparecen ya incorporados los datos de un fichero de datos llamado Base_HATCO. sav[2]. En la figura de la izquierda aparece la vista de variables con las características de todas las variables del fichero de datos. De una vista a otra se cambia pinchando con el ratón en la pestaña correspondiente en la parte inferior izquierda de la ventana.

2 Posteriormente se hará referencia detallada a Base HATCO.



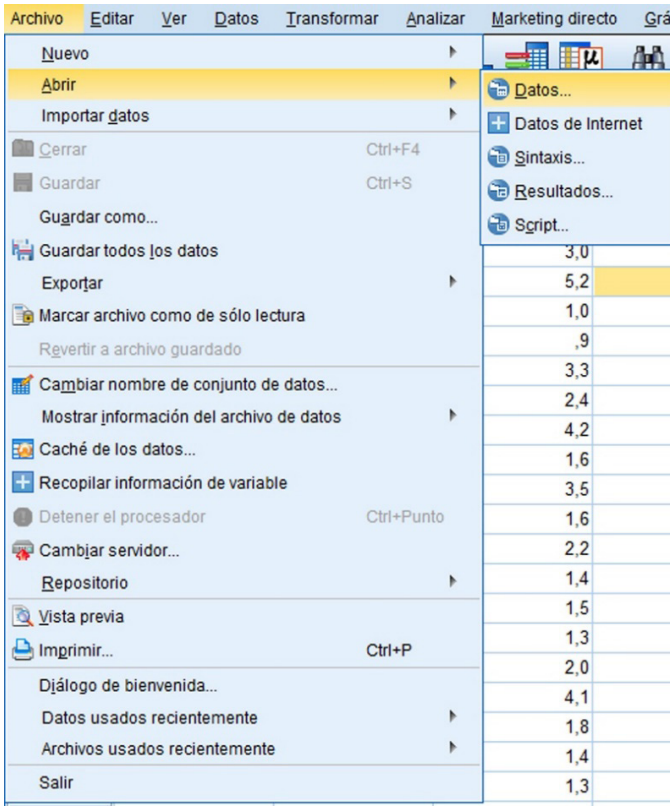


Dentro de la vista de datos se pueden distinguir varias zonas. La primera zona (parte más alta de la ventana) está formada por la barra que contiene el nombre de la ventana, con la inclusión del nombre del fichero de datos activo si existe, en este caso Base_HATCO.sav. La segunda (debajo de la anterior) es la zona de los menús con los nombres de los menús desplegables que sirven para llevar a cabo las tareas cuando se coloca el cursor sobre uno de los rótulos y se pincha con el ratón, entonces se despliega un menú, sobre el cual, se remarcan las acciones que se pueden ejecutar y de la que se escoge una; estas opciones figuran en la tabla adjunta y posteriormente se explican detalladamente.



Menú	Función
Archivo	Todas las funciones de archivos: Abrir, cerrar, guardar, importar, exportar, imprimir, etc.
Editar	Todas las funciones de la edición: cortar, copiar, eliminar, buscar, reemplazar, etc...
Ver	Controla la vista de la pantalla principal y las barras que aparecen en ella.
Datos	Contiene acciones que se pueden llevar a cabo con los datos.
Transformar	Permite realizar cualquier función conducente a crear nuevas variables a partir de otras existentes o no: transformar, recodificar, asignar rangos, etc...
Analizar	Acceso al conjunto de programas de SPSS, que van desde la generación de una tabla de frecuencias a análisis multivariantes complejos.
Marketing directo	Aparece en las últimas versiones, se relaciona con la aplicación de técnicas de marketing.
Gráficos	Acceso a las opciones de gráficos estadísticos.
Utilidades	Acceso a la descripción de las variables, crea grupos de variables y edita menús.

Ampliaciones	Son componentes personalizados que amplían las prestaciones de IBM® SPSS Statistics.
Ventana	Acceso rápido a las ventanas de datos, de resultados, de sintaxis.
? Ayuda	Ayuda en línea sobre todo el paquete SPSS y una ayuda incorporada a la aplicación que incluye un tutorial para la toma de decisiones y la selección de la prueba adecuada.

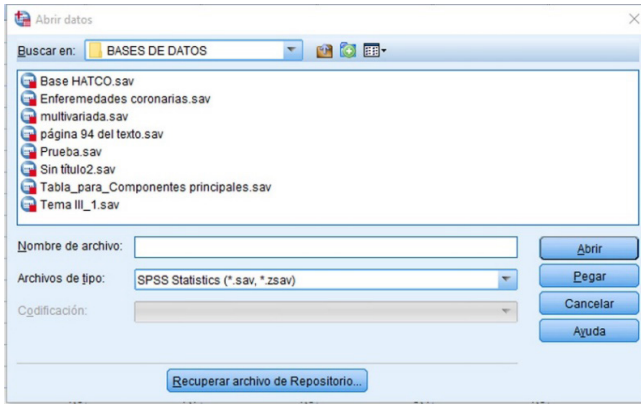


La opción archivo despliega un menú que coincide con los de casi todas las aplicaciones de Windows, por lo que solo se comentarán algunas opciones:

Abrir: permite abrir un archivo (generalmente de datos) para empezar a trabajar, esta opción es común a casi todas las aplicaciones Windows.

Abrir datos: permite abrir un archivo de datos en una base de datos distinta a la generada por SPSS, cuando se selecciona pasa a un sistema de diálogos que guía al usuario a obtener el resultado deseado.

Se adjunta el primero de esos diálogos:



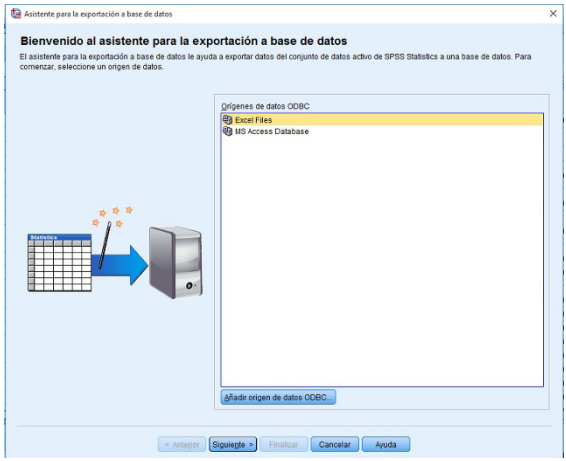
Archivo > Importar datos > Datos de texto...: Esta opción permite seleccionar un archivo de texto en el cuadro de diálogo Abrir datos. En caso necesario, se solicita seleccionar la codificación del archivo. El Asistente para la importación de texto le facilita definir cómo desea leer el archivo de datos de texto.

Guardar: permite almacenar el fichero activo en un disco. Si el fichero activo ha sido leído previamente se guardará con el mismo nombre que tenía (el fichero original que existía en el disco se perderá). Por el contrario, si el fichero ha sido creado sin que exista ninguna imagen de él en el disco, se pide que se asigne un nombre al nuevo fichero en el que se va a guardar la información. Debe quedar claro que esta opción siempre guarda un fichero de datos de SPSS, con extensión. sav.

Guardar como...: permite guardar el fichero activo con otro nombre y/o con formato de otras aplicaciones informáticas, como bases de datos u hojas de cálculo.

Guardar todos los datos: se utiliza en el caso en el que se hayan abierto varios ficheros para intercambiar datos entre ellos.

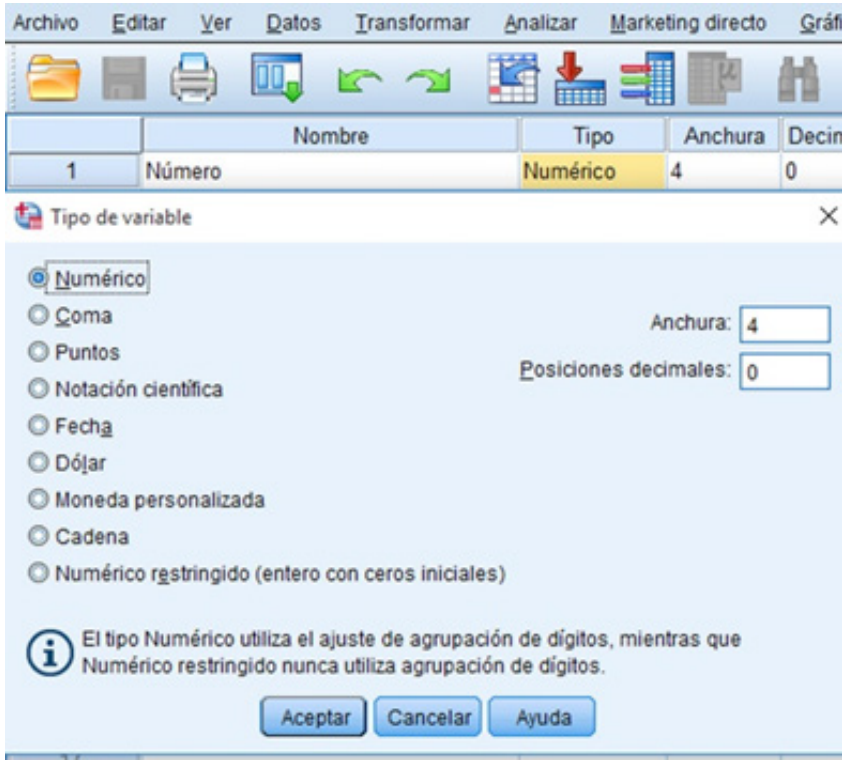
Exportar a base de datos: Convierte el fichero que se está ejecutando en un formato de base de datos conocidas y lo almacena en dicho formato. El principal cuadro de diálogo de esta opción se muestra en la figura adjunta.



1.6. La definición de las variables

Cuando se inicia el SPSS aparece la matriz de datos vacía al no existir un fichero seleccionado para trabajar con él; ante esta situación se debe crear la estructura del fichero, esto se concreta mediante la definición de las variables del nuevo fichero de datos de SPSS. A esa situación también se llega después de haber trabajado con SPSS cuando se despliega en el menú Archivo la opción Nuevo y dentro de ella la opción Datos, lo que hará que se elimine el fichero activo y se deje limpia la matriz de datos.

Antes de continuar es preciso destacar que, cada columna de la hoja de datos se corresponde con una variable y que el proceso de definir variable consiste en asignarle a cada columna un nombre y un conjunto de atributos que definen esencialmente el tipo de variable que se está definiendo y en correspondencia con esto definir su formato. SPSS reconoce los tipos de variables que se muestran en la figura y se asocian a tres tipos de medidas: Escala, Ordinal y Nominal.



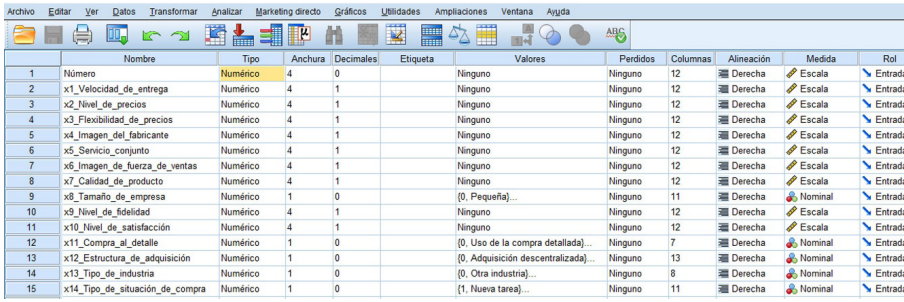
Las características de cada escala se describen a continuación:

Escala: cuando los valores de los datos son valores numéricos sobre una escala de intervalo o de razón (la edad, el peso, el número de hermanos); cuando se define una variable de tipo Numérico, Coma, Punto o Notación Científica, SPSS asigna Escala a la escala de medida de la variable.

Ordinal: estos datos representan categorías con algún orden intrínseco (bajo, medio, alto; peor, igual, mejor); las variables ordinales pueden ser cadenas (alfanuméricas) o valores numéricos que representen categorías diferentes (1=bajo, 2=medio, 3=alto); la escala ordinal corresponde a datos cualitativos ordinales.

Nominal: en esta escala los valores de los datos representan categorías sin un orden intrínseco (el grupo sanguíneo A, B, O; el tipo de trabajo de una persona); las variables nominales pueden ser cadenas (alfanuméricas) o valores numéricos que representen categorías diferentes (1= varón, 2= mujer). De la definición de la escala de medida depende, los análisis que se pueden hacer con los diferentes datos.

Aunque se pueden introducir los datos y después definir la variable, esto denota desorganización y poco rigor, por eso la manera más natural de crear las variables de un fichero es definirlas antes de introducir dato alguno. Para ello hay que situarse en la vista de variables de la ventana principal de SPSS, pinchando en la pestaña correspondiente o haciendo doble clic en la cabecera de la columna. En la vista de variables, cada fila corresponde a una variable, Para cada variable habrá que ir especificando cada una de sus características, empezando por el nombre (primera columna) como se muestra en la siguiente figura:


Archivo Editar Ver Datos Transformar Analizar Marketing directo Gráficos Utilidades Ampliaciones Ventana Ayuda

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	Número	Numérico	4	0		Ninguno	Ninguno	12	Derecha	Escala	Entrada
2	x1_Velocidad_de_entrega	Numérico	4	1		Ninguno	Ninguno	12	Derecha	Escala	Entrada
3	x2_Nivel_de_precios	Numérico	4	1		Ninguno	Ninguno	12	Derecha	Escala	Entrada
4	x3_Flexibilidad_de_precios	Numérico	4	1		Ninguno	Ninguno	12	Derecha	Escala	Entrada
5	x4_Imagen_del_fabricante	Numérico	4	1		Ninguno	Ninguno	12	Derecha	Escala	Entrada
6	x5_Servicio_conjunto	Numérico	4	1		Ninguno	Ninguno	12	Derecha	Escala	Entrada
7	x6_Imagen_de_fuerza_de_ventas	Numérico	4	1		Ninguno	Ninguno	12	Derecha	Escala	Entrada
8	x7_Calidad_de_producto	Numérico	4	1		Ninguno	Ninguno	12	Derecha	Escala	Entrada
9	x8_Tamaño_de_empresa	Numérico	1	0		{0, Pequeña}...	Ninguno	11	Derecha	Nominal	Entrada
10	x9_Nivel_de_fidelidad	Numérico	4	1		Ninguno	Ninguno	12	Derecha	Escala	Entrada
11	x10_Nivel_de_satisfacción	Numérico	4	1		Ninguno	Ninguno	12	Derecha	Escala	Entrada
12	x11_Compra_al_detalle	Numérico	1	0		{0, Uso de la compra detallada}...	Ninguno	7	Derecha	Nominal	Entrada
13	x12_Estructura_de_adquisición	Numérico	1	0		{0, Adquisición descentralizada}...	Ninguno	13	Derecha	Nominal	Entrada
14	x13_Tipo_de_industria	Numérico	1	0		{0, Otra industria}...	Ninguno	8	Derecha	Nominal	Entrada
15	x14_Tipo_de_situación_de_compra	Numérico	1	0		{1, Nueva tarea}...	Ninguno	11	Derecha	Nominal	Entrada

El nombre de las variables. Se pincha (o se hace doble clic) sobre la casilla correspondiente al nombre de la variable que se está definiendo, y se escribe el nombre que se desea, con las siguientes normas:

- Cada nombre de variable debe ser exclusivo; no se permiten duplicados.
- Los nombres de variable pueden tener una longitud de hasta 64 bytes y el primer carácter debe ser una letra o uno de

estos caracteres: @, # o $. Los caracteres posteriores puede ser cualquier combinación de letras, números, caracteres que no sean signos de puntuación y un punto (.).

- Las variables no pueden contener espacios.
- Se deben evitar los nombres de variable que terminan con un punto, ya que el punto puede interpretarse como un terminador del comando. Solo se pueden crear variables que finalicen con un punto en la sintaxis de comandos. No se puede crear variables que terminen con un punto en los cuadros de diálogo que permiten crear nuevas variables.
- Se deben evitar los nombres de variable que terminan con un carácter de subrayado, ya que tales nombres pueden entrar en conflicto con los nombres de variable creados automáticamente por comandos y procedimientos.
- Las palabras reservadas no se pueden utilizar como nombres de variable. Estas palabras son ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO y WITH.
- Los nombres de variable se pueden definir combinando de cualquier manera caracteres en mayúsculas y en minúsculas, esta distinción entre mayúsculas y minúsculas se conserva en lo que se refiere a la presentación.
- En ocasiones el nombre de la variable brinda poca información, más adelante se indicará cómo resolver este problema.

Tipo: de los tipos reconocidos por SPSS ya se habló, para ello se utiliza el cuadro de diálogo que se mostró anteriormente, precisando sobre los tipos se tiene:

Numérico: para una variable cuyos valores son números.

Coma: define una variable numérica cuyos valores se muestran con la coma de separación de miles y con un punto como separador de la parte decimal.

Punto: define una variable numérica cuyos valores se muestran

con el punto de separador de miles y con una coma como separador de la parte decimal.

Notación científica: define una variable numérica cuyos valores se muestran con una E intercalada y un exponente con signo que representa una potencia de base diez. 1,23E2.

Fecha: define una variable numérica cuyos valores se muestran en uno de los diferentes formatos de fecha-calendario u hora-reloj. Al seleccionar fecha se despliega un menú con las distintas opciones de este tipo.

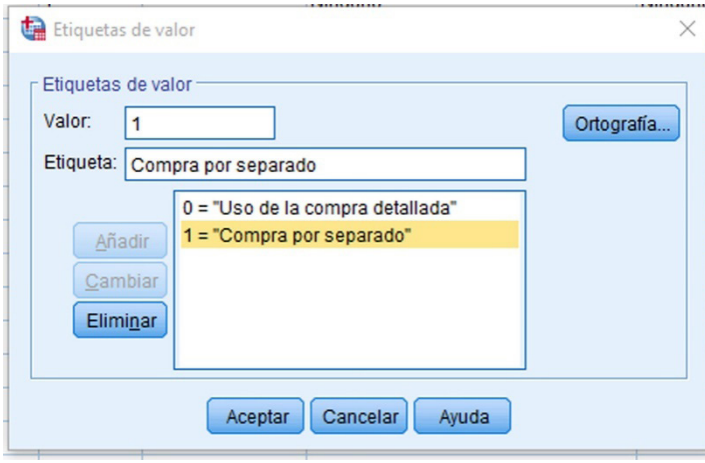
Moneda personalizada: sirve para definir una variable numérica cuyos valores se muestran en uno de los formatos de moneda personalizados que se hayan definido previamente en la pestaña Moneda.

Cadena: define una variable cuyos valores no son numéricos y, por ello, no se utilizan en los cálculos. Pueden contener cualesquiera caracteres hasta la longitud definida. Estas variables son conocidas como variables alfanuméricas.

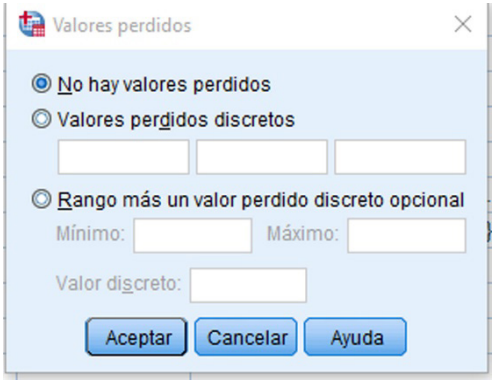
El tamaño y el formato de cada tipo se expresan en los campos que aparecen en la parte de la derecha de la ventana. Habrá que especificar el tamaño total y el número de decimales en los tipos Numérico, Coma, Punto y Notación Científica y. la anchura total que no podrá sobrepasar los 255 caracteres para el tipo Cadena.

Las etiquetas: las propias restricciones del sistema para los nombres de las variables hace que estas tengan pocos caracteres y en ocasiones es difícil de saber lo que significan. Por eso, además del nombre existe la etiqueta que identifica cada variable de una manera más precisa y permiten reconocerlas cuando se presentan los resultados. Las etiquetas pueden tener hasta 130 caracteres. Ejemplo, una variable nombrada AC60días, se explicita con la etiqueta como Asistencias a clases en 60 días, esto hace que al procesar la información la salida en una tabla sale bajo el título *Asistencias a clases en 60 días en lugar de AC60días.*

Valores: además del nombre poco explicativo de las variables y la solución que da las etiquetas, con las variables suelen estar representadas por códigos numéricos, (1=bajo, 2=medio, 3=alto), es este caso también pueden establecerse etiquetas de valor que permitan identificar a las categorías con ellas en lugar de los códigos numéricos, con lo que se hacen más explicativas. Estas etiquetas pueden ser de hasta 60 caracteres y se pueden asignar mediante el siguiente cuadro de diálogo, donde se ejemplifica la asignación de etiquetas a la variable procedencia social:

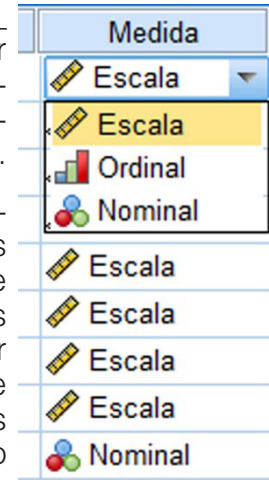


Los datos faltantes (valores perdidos): los datos faltan por distintos motivos, porque no existe, porque no la ha querido proporcionar, etc.; para cuando esto ocurra se escoge un código para representarlos, debiendo proporcionarle tal código a SPSS para que él los incluya en los análisis posteriores; a esta representación de los datos faltantes se le denomina datos faltantes del usuario, para distinguirlos de los datos faltantes del sistema (que se consiguen sin más que dejar en blanco el espacio reservado para ellos, donde, si la variable es numérica, SPSS colocará una coma (para identificarlos). La identificación de datos faltantes es crucial pues, si no se identifican, estos serán empleados con los valores que tengan, dando lugar a resultados erróneos. La pantalla de diálogo adjunta facilita la asignación de códigos para posibles datos faltantes.



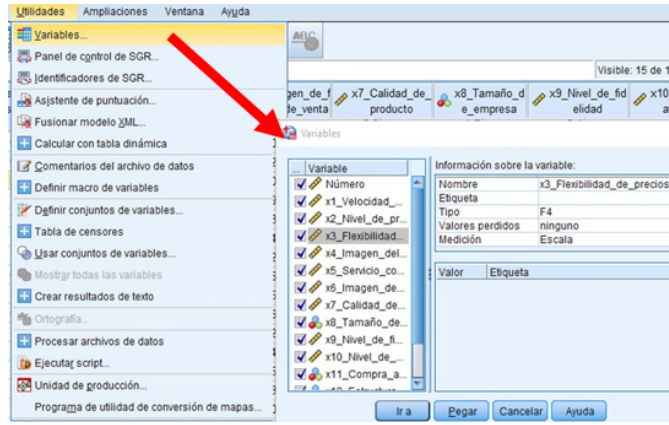
El formato de columna de las variables: los formatos que se han dado a la variable hasta el momento constituyen características internas de la variable que no se corresponden con las características de su presentación en la matriz de datos, de ahí la necesidad de las dos opciones de formato: la anchura total de la columna y la alineación que tendrá la información dispuesta en esta columna. El Ancho de la columna puede ser de hasta 256 caracteres y la alineación del texto en la columna puede ser a la izquierda, centrada o a la derecha, siendo esta última la asignada por SPSS en el caso de variable numérica y a la izquierda en el caso de variable de cadena.

La definición de la Escala de Medida de una variable: el tema ya se trató solo falta añadir que esta asignación se puede hacer mediante un menú que se despliega en la misma columna como se muestra en la imagen adjunta.



En este capítulo se ha dado una gran importancia al tratamiento de las variables, y es que la definición cuidadosa y detallada de las variables ayudará mucho en los análisis y en la interpretación de los resultados, por lo que se recomienda al usuario que emplee todo el tiempo que sea necesario en tales definiciones, tengan la seguridad de que no será tiempo perdido.

El conjunto de variables definidas, junto con las características que se les haya asignado, forman lo que se llama la estructura del fichero de datos; esto es una de las dos partes de un fichero de datos de SPSSy se puede guardar en un fichero de extensión. sav. que aparecerá sin los datos, pero donde se han guardado las variables y sus características, y se pueden ver resumida en una ventana mediante Utilidades→ Variables (imagen adjunta).



Sobre dicha estructura se puede añadir el otro componente, los datos propiamente dichos, y juntos conformarán el archivo de datos de SPSS. Para guardar el trabajo actual, efectuar Archivo→ Guardar (o usar el botón guardar) asignando el nombre deseado, por ejemplo, BaseHATCO.sav.

1.7. Bases HATCO, problema base, enfermedades coronarias y dimensiones corporales

Los ejemplos del libro están relacionados con cuatro tablas de datos tomadas de la bibliografía referida a continuación, y aparecen en los anexos que se indican:

- HATCO (Anexo 1): J. F. Hair, Jr., R. E. Anderson. R. L. Tatham, W. C. Black ANÁLISIS MULTIVARIANTE, 5.ª ed.
- PROBLEMA BASE (Anexo 2): Dra. Rosa Maria de Nascimento. “Estrategia didáctica para el uso del enfoque de problema base en el proceso de enseñanza-aprendizaje de la

estadística en la escuela superior pedagógica de Bié. Tesis de doctorado. UCP Enrique José Varona. La Habana. 2016.

- ENFERMEDADES CORONARIAS (Anexo 3): Cáceres Álvarez, Rafael. Estadística multivariante y no paramétrica con SPSS. Aplicación a las ciencias de la salud. Madrid. 1995. (EJEMPLO CORONAR).

- DIMENSIONES CORPORALES (Anexo 4): Johnson, Dalias E. Métodos Multivariados Aplicados al Análisis de Datos. Kansas State University, 2000. (Tabla 1. 2).

HATCO: es una base de datos de la Compañía Hair, Anderson y Tatham (HATCO) un enorme (aunque inexistente) distribuidor industrial. La base de datos, consiste en 100 observaciones de 14 variables separadas, es un ejemplo de un estudio de segmentación de la situación empresa a empresa, específicamente un informe sobre los clientes actuales de HATCO. Se utilizan tres tipos de datos. La primera clase es la percepción de HATCO sobre siete atributos identificados en estudios pasados como los más influyentes en la elección de distribuidor. Los encuestados, ejecutivos de compras de empresas clientes de HATCO, puntúan a HATCO sobre cada atributo.

1. X_1: Velocidad de entrega:

   Tiempo que transcurre hasta que se entrega el producto, una vez que se hubo confirmado el pedido.

2. X_2: Nivel de precio:

   Nivel de precios percibido por los clientes industriales.

3. X_3: Flexibilidad de precios:

   Disposición percibida en los representantes de HATCO para negociar el precio de todas las compras.

4. X_4: Imagen del fabricante:

   Imagen conjunta del fabricante/distribuidor.

5. X_5: Servicio conjunto:

   Nivel de servicio necesario para mantener una relación satisfactoria entre el oferente y el comprador.

6. X_6: Imagen de la fuerza de ventas:

   Imagen conjunta de la fuerza de ventas del fabricante.

7. X_7: Calidad del producto:

   Nivel de calidad percibido en un producto particular (por ejemplo, el acabado o el rendimiento).

La segunda clase de información hace referencia a los resultados de compras reales, bien sobre las evaluaciones de la satisfacción de los encuestados con HATCO, bien sobre el porcentaje de sus compras de productos a HATCO.

8. X_9: Nivel de satisfacción:

   Satisfacción del comprador con las compras anteriores realizadas a HATCO, medidas en el mismo gráfico de la escala de clasificación de las entradas X_1 a X_7.

9. X_{10}: Tamaño de la empresa:

   Tamaño de la empresa relativo respecto a otras empresas en el mismo mercado. Esta variable tiene dos categorías: 1 = grande y 0 = pequeña.

La tercera clase de información contiene características generales de las empresas clientes (por ejemplo, tamaño de la empresa, tipo de industria).

10. X_8: Nivel de fidelidad:

    Cuánto se compra a HATCO del total del producto de la empresa, medido en una escala de porcentaje de 100, que va desde 0 al 100.

11. X_{11}: Compra al detalle:

Medida por la cual un comprador particular evalúa cada compra separadamente (análisis del valor total) o en función de una compra detallada, donde se especifican precisamente las características del producto deseado. Esta variable tiene dos categorías: 1 cuando emplea la aproximación al análisis del valor total, evaluando cada compra por separado y 0 cuando hace uso de la compra detallada.

12. X_{12}: Estructura de la adquisición:

Método de adquisición/compra de productos a una compañía en particular. Esta variable tiene dos categorías: 1 = adquisición centralizada y 0 = adquisición descentralizada.

13.X_{13}: Tipo de industria:

Clasificación de la industria a la que pertenece el comprador del producto. Esta variable tiene dos categorías: 1 = industria de la clase A y 0 = otras industrias.

14.X_{14}: Tipo de situación de compra:

Tipo de situación a la que se enfrenta el comprador. Esta variable tiene tres categorías: 1 =nueva tarea, 2=re-compra similar modificada y 3 =recompra simple.

NOTA: En adelante aparecerá X1, …X12 en lugar de X1, …X12.

*PROBLEMA BASE:*Tiene 12 variables controlada a 64 alumnos de un aula:

Notación	Significado	Notación	Significado	Notación	Significado
#	Número del alumno en el listado oficial	AC	Asistencias a 60 días de clases.	CI	Cociente de Inteligencia[3]
Sexo	M: Masculino, F: femenino	NPIS	Nota promedio (Inicio Semestre)	ISF	Índice de satisfacción con la escuela

E	Edad	NPA	Nota promedio actual	ISF	Índice de satisfacción con la familia[4]
Pf	Asignatura de preferencia: C_e: Ciencias exactas; C_s: Ciencias Sociales; C_h: Ciencias Humanísticas; C_n : Ciencias Naturales	PS	Procedencia social: CA: clase alta; CM: clase media; CB: Clase baja	DC	Disciplina y conducta: MB: Muy Buena; B: Buena; R: Regular; M: Mala; MM: Muy mala

ENFERMEDADES CORONARIAS:

Notación	Significado	Notación	Significado
X1	paciente #	X2	Edad
X3	Sexo 1: MASCULINO; 2: FEMENINO	X4	Clase Social 1: ALTA; 2: MEDIA; 3 : BAJA
X5	Colesterolemia Basal	X6	Colesterolemia HDL Basal
X7	Trigliceridemia Basal	X8	Tensión arterial sistólica
X9	Tensión arterial diastólica	X10	Enfermedad coronaria 1: SI; 2: NO
X11	Fuma 1: SI; 2: NO	X12	Sedentarismo 1: SI; 2: NO
X13	Peso	X14	Talla
X15	Nivel de estudios 1: PRIMARIO; 2: MEDI0; 3 : SUPERIOR	X16	Antecedentes cardiacos Familiares 1: SI; 2: NO

DIMENSIONES CORPORALES:

Estatura	Longitud brazo	Ancho mano	Longitud interior pierna
Estatura sentado	Longitud antebrazo	Longitud muslo	Longitud pie

Capítulo II. Iniciando el trabajo con E.A.D. utilizando SPSS

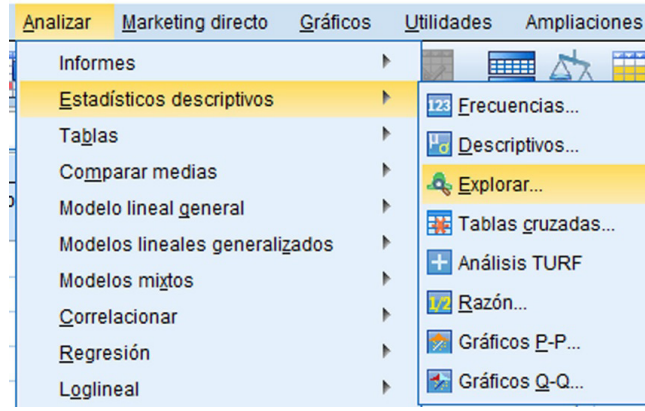
2.1. Examen gráfico y numérico de las variables

Este apartado está orientado a realizar un examen gráfico de la naturaleza de las variables individuales a analizar y desarrollar un análisis descriptivo numérico para cuantificar los aspectos más significativos de los datos.

SPSS posee varias opciones que permiten hacer en forma aislada exámenes gráficos y numéricos de las variables, pero la opción con resultados más completos es Explorar qué se obtiene de la secuencia de menú:

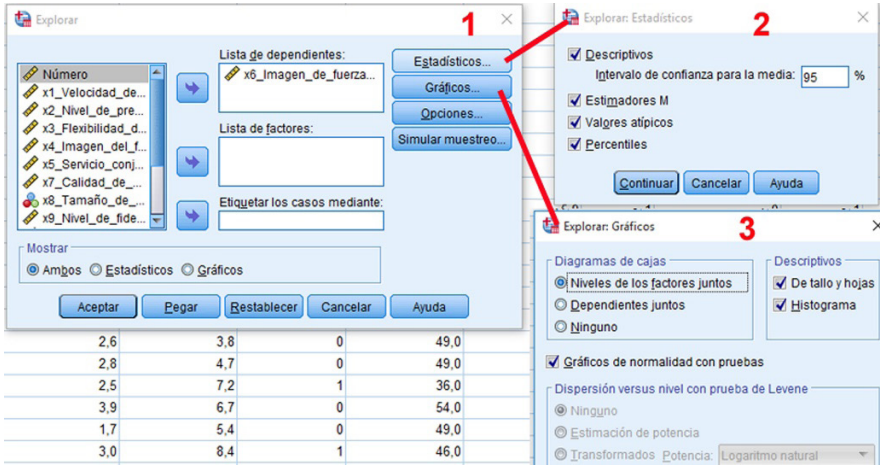
Analizar → Estadísticos descriptivos → Explorar

Explorar permite obtener las principales informaciones de las características numéricas de los datos correspondientes a la variable estudiada; para ejemplificar se tomará de la base de datos HATCO los datos correspondientes a la variable X_6: (Imagen de la fuerza de ventas), información referida a la imagen conjunta de la fuerza de ventas del fabricante.



La imagen adjunta muestra el inicio de las acciones para procesar la información.

Seleccionada la opción de Explorar sigue a los siguientes cuadros de diálogo:



En el menú 1 se destaca la lista de variables (la lista de factores se comentará posteriormente); el acceso a los submenús Estadísticos y Gráficos (Opciones y Simular muestreo son más específicos y pueden quedar para estudios posteriores); la selección de las opciones de cómo mostrar los resultados; una opción es solo los estadísticos, otra opción es solo los gráficos y una tercera es que se muestren ambos, estadísticos y gráficos.

El submenú 2 (Estadísticos) permite seleccionar los análisis estadísticos que se van a realizar:

- *Estadísticos descriptivos*; media, moda, mediana, desviaciones y permite fijar el nivel de confianza que se desea obtener en el intervalo de confianza para la media. El valor de k por defecto es 95, pero es posible introducir cualquier otro valor entre 1 y 99,99.
- *Estimadores M*; son estimadores de tendencia central basados en el método de máxima verosimilitud (de ahí el nombre de estimadores M). Un estimador M no es más que una media ponderada en la que los pesos asignados a los casos dependen de la distancia de cada caso al centro de la distribución: los casos

centrales reciben un peso de 1 y los demás valores reciben un peso tanto menor cuanto más alejados se encuentran del centro.

Análogo a lo que ocurre con la media truncada, los estimadores M son menos sensibles que la media aritmética a la presencia de valores extremos por eso su principal aplicación es en distribuciones muy asimétricas.

Existen varios estimadores M que difieren entre sí por la forma concreta de asignar pesos a los casos. El procedimiento Explorar incluye cuatro de esos estimadores: Huber, Andrew, Hampel y Tukey.

- *Valores atípicos*; son observaciones con una combinación única de características identificables que les diferencian claramente de las otras observaciones; ellos no pueden ser caracterizados categóricamente como benéficos o problemáticos, sino que deben ser contemplados en el contexto del análisis y deben ser evaluados por los tipos de información que pueden proporcionar.

  Cuando son benéficos, los casos o valores atípicos, aunque diferentes a la mayor parte de la muestra, pueden ser indicativos de las características de segmentos de la población que se llegarían a descubrir en el curso normal del análisis. Por el contrario, los casos atípicos problemáticos no son representativos de la población y están en contra de los objetivos del análisis. Los casos atípicos problemáticos pueden distorsionar seriamente los test estadísticos. Debido a la variabilidad en la evaluación de los casos atípicos, se necesita que el investigador examine los datos en busca de su presencia con el fin de averiguar el tipo de influencia que ejercen.

- *Percentiles*; el percentil es una medida de tendencia central que indica, una vez ordenados los datos de menor a mayor, el valor de la variable por debajo del cual se encuentra un porcentaje dado de observaciones en un grupo de ellas. Por

ejemplo, el percentil 20º es el valor bajo el cual se encuentran el 20 por ciento de las observaciones.

Se representan con la letra P. Para el percentil i-ésimo, donde la i toma valores del 1 al 99. El i % de la muestra son valores menores que él y el 100-i % restante son mayores.Aparecen citados en la literatura científica por primera vez por Francis Galton[5] en 1885. De particular interés son:

- P25 = Q1.
- P50 = Q2 = mediana.
- P75 = Q3.

El submenú 3 permite seleccionar los gráficos:

Diagramas de cajas; en la gráfica de cajas los datos correspondientes a cada variable numérica se representan con una caja, tiras que salen de ellas y límites, con lo que se representa:

- La caja:
  - La altura de la caja representa la amplitud intercuartil (AI), en ella está representado el 50% de la muestra.
  - El borde superior de la caja es el percentil 75.
  - El borde inferior el percentil 25.
  - La línea central de la caja es el percentil 50 o mediana.
- Los límites:
  - El limite después de cada tira es la puntuación entre el extremo de la caja y como máximo 1.5 AI's.
- Los datos más alejados (*, 0) se denominan casos extremos.
  - Con una 0 se marcan los casos entre 1.5 y 3 AI's del extremo de la caja.

- Con un asterisco se marcan los casos que están a más de 3 AI's del extremo de la caja

Gráficos de tallos y hojas; esta opción permite obtener gráficos en modo texto que son similares a los histogramas, pero que proporcionan información más precisa que éstos porque no solo representan cuántos dado corresponden a cada categoría, también indican cuáles son esos datos como se muestra en la siguiente gráfica correspondiente a la variable Asistencias a clases de la base de datos PROBLEMA BASE:

Asistencias a clases en 60 días Gráfico de tallo y hojas

Frecuencia	Stem &	Hoja
6,00	3 .	000133
4,00	3 .	5579
10,00	4 .	0001123334
12,00	4 .	555556699999
2,00	5 .	13
6,00	5 .	558899

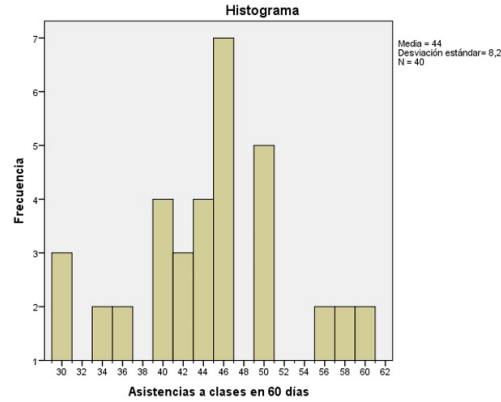
Ancho del tallo: 10
Cada hoja: 1 caso(s)

Al igual que en un histograma, las longitudes de las líneas reflejan el número de casos que pertenecen a cada intervalo, además, cada caso (o grupo de casos) está representado por un número que coincide con el valor de ese caso en la variable.

En un diagrama de tallo y hojas cada valor se descompone en dos partes: el primer o primeros dígitos (el tallo o stem) y el dígito que sigue a los utilizados en el tallo (las hojas o leaf). Por ejemplo, los valores correspondientes a 35 asistencias se han descompuesto en un tallo de 3 y una hoja de 5; (un número como 12.300 puede descomponerse en un tallo de 12 y una hoja de 3).

Histograma; es una gráfica conocida y se construye agrupando

los datos en intervalos de la misma amplitud y levantando barras de altura proporcional al número de casos de cada intervalo, aunque estas opciones pueden controlarse utilizando el editor de gráficos. La figura muestra un histograma de la variable anteriormente representada en el diagrama de tallo y hojas, por lo que se puede comparar ambos diagramas y observar las coincidencias y diferencias existentes entre ellos.



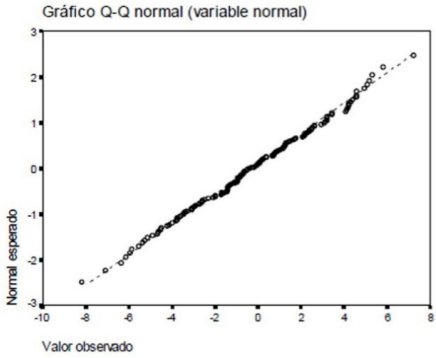
Gráficos de normalidad; muchos procedimientos estadísticos se sustentan en dos supuestos básicos:

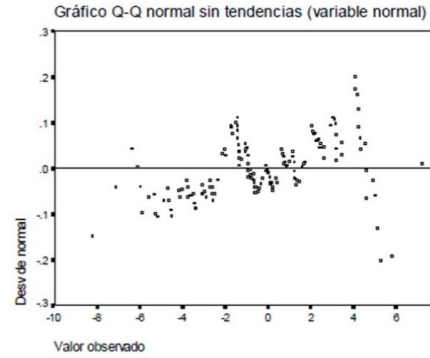
- Normalidad: las muestras con las que se trabajan proceden de poblaciones normalmente distribuidas.
- Homocedasticidad u homogeneidad de varianzas: todas esas poblaciones normales poseen la misma varianza.

Esa es la causa por la que en el menú Explorar aparece esta opción en submenú 3 (Gráficos), la cual permite contrastar estos supuestos, mediante dos gráficos de normalidad (Q-Q normal y Q-Q normal sin tendencia) junto con dos pruebas de significación: Kolmogorov-Smirnov[6] (Kolmogorov[7], 1933; Smirnov, 1948; Lillieffors, 1967) y Shapiro-Wilk[8] (Shapiro & Wilk, 1965).

Un gráfico Q-Q ("Q" viene de cuartil) es un método gráfico para el diagnóstico de diferencias entre la distribución de probabilidad de una población de la que se ha extraído una muestra aleatoria y una distribución usada para la comparación (en este caso interesa la distribución normal).

Una muestra de puntuaciones aleatorias tomadas de una distribución normal genera dos gráficos como los que se muestran:





En ellos se observa que:

1. Los puntos del diagrama Q-Q normal se ajustan a la diagonal.
2. Los puntos del diagrama Q-Q normal sin tendencia se distribuyen aleatoriamente sin mostrar una pauta o patrón claramente definido.

Cuando estas dos condiciones se cumplen se puede afirmar que los datos de la muestra estudiada proceden de una población normalmente distribuida, si falla cualquiera de las dos condiciones antes referidas, se puede concluir que los datos no proceden de una población normalmente distribuida.

Como ejemplo de los resultados que se obtienen al aplicar la opción Explorar en la secuencia Analizar → Estadísticos descriptivos → Explorar se mostrarán los alcanzados con la variable x6_Imagen_de_fuerza_de_ventas de la base HATCO.

Para diferenciar los resultados que devuelve el sistema de los comentarios de los autores, estos últimos tendrán al inicio del párrafo el carácter ✍

Descriptivos			Estadístico	Error estándar
x6_lmagen_de_fuerza_de_ventas	Media		2,635	,0815
	95% de intervalo de confianza para la media	Límite inferior	2,473	
		Límite superior	2,797	
	Media recortada al 5%		2,624	
	Mediana		2,550	
	Varianza		,664	
	Desviación estándar		,8148	
	Mínimo		,0	
	Máximo		4,6	
	Rango		4,6	
	Rango intercuartil		,8	
	Asimetría		,169	,241
	Curtosis		,716	,478

✍La tabla anterior es un resumen de los estadísticos descriptivos.

Estimadores M				
	Estimador M de Huber[a]	Biponderado de Tukey[b]	Estimador M de Hampel[c]	Onda de Andrews[d]
x6_lmagen_de_fuerza_de_ventas	2,585	2,545	2,587	2,544
a. La constante de ponderación es 1,339.				
b. La constante de ponderación es 4,685.				
c. Las constantes de ponderación son 1,700, 3,400 y 8,500				
d. La constante de ponderación es 1,340*pi.				

✍Compare los Estimadores M y los percentiles.

Percentiles									
		Percentiles							
		5	10	25	50	75	90	95	
Promedio ponderado(Definición 1)	x6_Imagen_de_fuerza_de_ventas	1,400	1,610		2,200	2,550	3,000	3,900	4,000
Bisagras de Tukey	x6_Imagen_de_fuerza_de_ventas		2,200		2,550	3,000			

Valores extremos					
			Número del caso	Valor	
x6_Imagen_de_fuerza_de_ventas	Mayor	1	5	4,6	
		2	42	4,6	
		3	7	4,5	
		4	82	4,4	
		5	2	4,0[a]	
	Menor	1	100	,0	
		2	35	1,1	
		3	43	1,3	
		4	92	1,4	
		5	50	1,4[b]	

a. Solo se muestra una lista parcial de casos con el valor 4,0 en la tabla de extremos superiores.

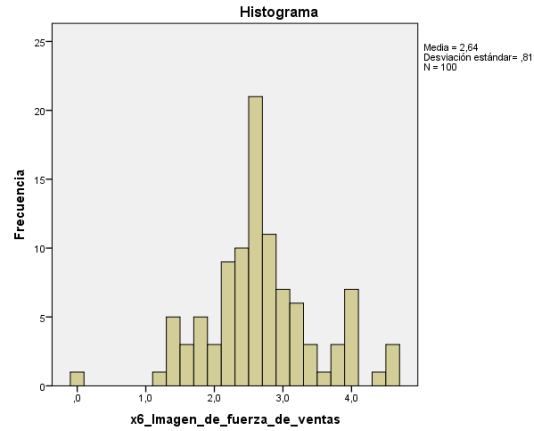
b. Solo se muestra una lista parcial de casos con el valor 1,4 en la tabla de extremos inferiores.

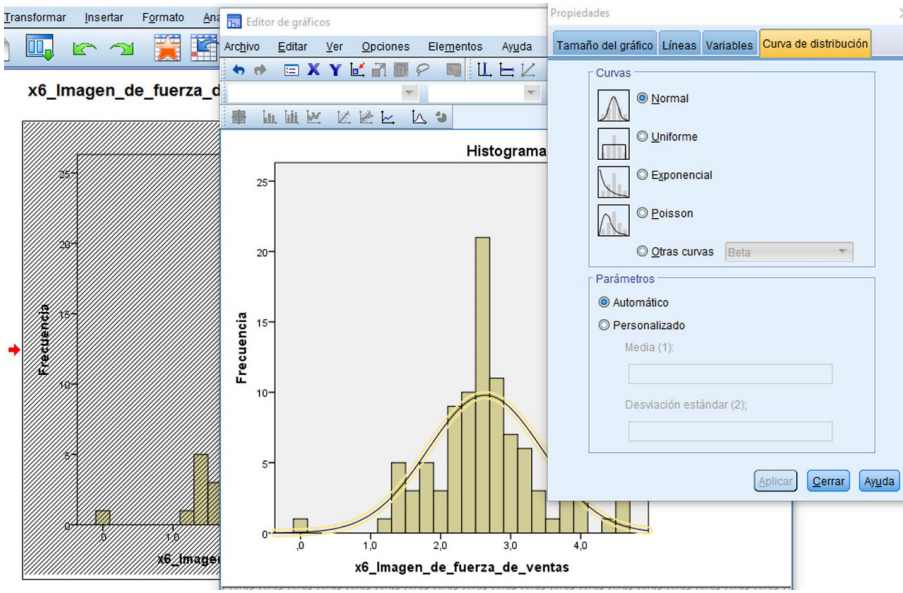
✍La tabla muestra los valores extremos inferiores y superiores, lo indicado en a y b expresa que hay más casos que tienen los valores señalados.

Pruebas de normalidad						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
x6_Imagen_de_fuerza_de_ventas	,118	100	,002	,969	100	,017
a. Corrección de significación de Lilliefors						

✍La tabla ofrece los estadísticos de Kolmogorov-Smirnov y de Shapiro-Wilk acompañados de sus correspondientes niveles críticos (Sig. = Significación).Ambos permiten contrastar la hipótesis nula de que los datos muestrales proceden de poblaciones normales. Se rechaza la hipótesis de normalidad cuando el nivel crítico (Sig.) sea menor que el nivel de significación establecido (generalmente 0,05). En el ejemplo, los estadísticos tienen asociados niveles críticos menores que 0,05, y de esta relación se debe inferir que la muestra x6_Imagen_de_fuerza_de_ventas no procede de una población con distribución normal.

El histograma muestra la distribución que siguen los datos; por defecto SSPSS divide la muestra en intervalos de igual longitud y determina la media y la desviación estándar, pero el editor del SPSS permite al usuario realizar otros ajustes al gráfico como los que se indican a continuación:





Pulsando doble clic sobre la imagen aparece el cuadro de diálogo en el que se puede seleccionar diferentes opciones, una de ella es la de superponer curvas de distribuciones entre ellas, la distribución normal y como se puede observar hay poca coincidencia entre el histograma de la muestra y la curva normal.

x6_Imagen_de_fuerza_de_ventas Gráfico de tallo y hojas

Frecuencia	Stem & Hoja
1,00	Extremes (=<,0)
1,00	1 . 1
1,00	1 . 3
6,00	1 . 444455
5,00	1 . 67777
4,00	1 . 8999
5,00	2 . 11111
11,00	2 . 22223333333

16,00 2 . 4445555555555555

15,00 2 . 666666667777777

7,00 2 . 8888999

6,00 3 . 000011

4,00 3 . 2222

3,00 3 . 444

2,00 3 . 67

5,00 3 . 88999

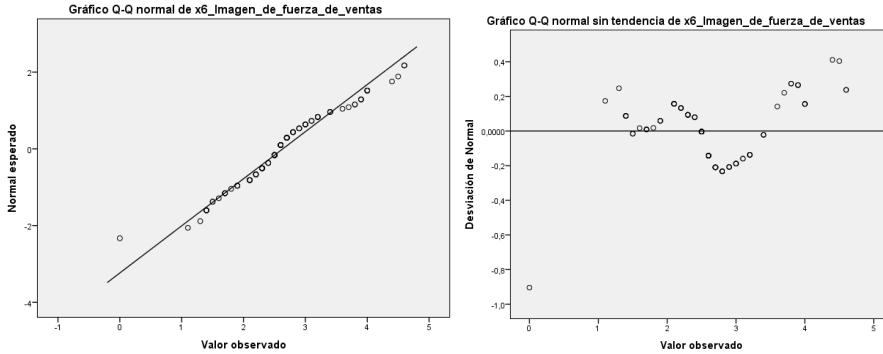
4,00 4 . 0000

4,00 Extremos (>=4,4)

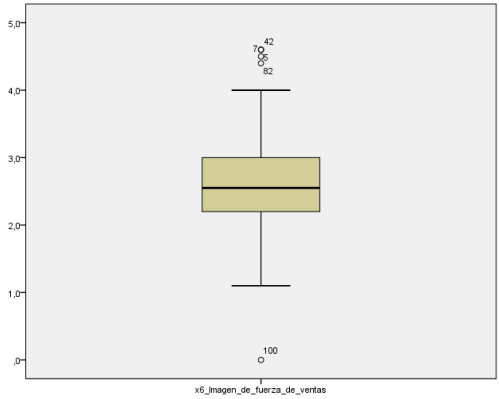
Ancho del tallo: 1,0

Cada hoja: 1 caso(s)

Observe la similitud entre el gráfico de tallo y hojas y el histograma.

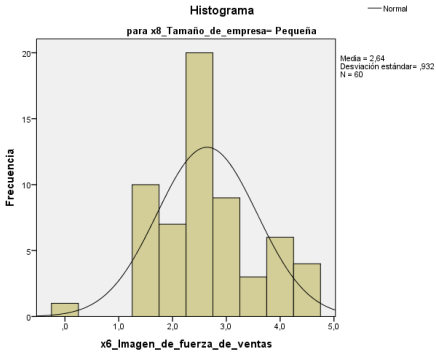


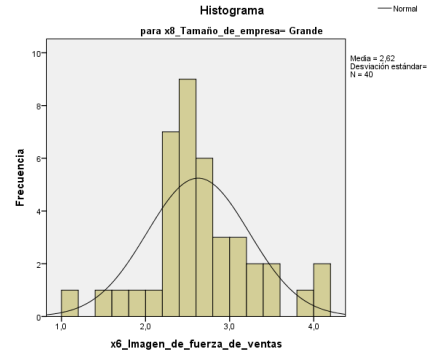
Aunque el gráfico Q-Q normal se distribuye alrededor de la diagonal, el gráfico Q-Q normal sin tendencias no sigue una distribución aleatoria, los puntos se agrupan siguiendo cierta regularidad como si se tratara de una curva, esto corrobora desde la interpretación del gráfico lo que ya se demostró con las pruebas de normalidad de Kolmogorov-Smirnov y Shapiro-Wilk.



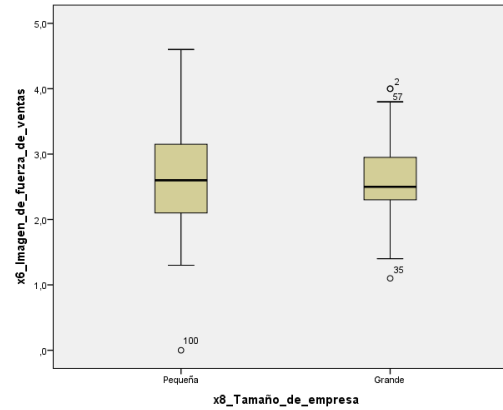
En el gráfico de caja y bigotes adjunto se destacan los casos extremos, casos atípicos, o outliers, en este caso todos son del tipo O, que están situados entre entre 1.5 y 3 AI's del extremo de la caja. Obsérvese que junto al correspondiente símbolo que identifica al dato aparece un número correspondiente al número de orden del dato.

Cuando se comentó el que se identificó como “menú 1” de la caja de diálogo “Analizar” se expresó que: “la lista de factores se comentará posteriormente” y es que en la celda de ese nombre se puede colocar una variable nominal u ordinal que clasifica la muestra; en este caso puede ser la variable x8_Tamaño_de_empresa que divide a muestra en dos subgrupos: empresas pequeñas y empresas grandes, la presencia de esta variable hace en los resultados todos estén referidos a estos dos grupo, algunos ejemplos de tales resultados son:





En los gráficos anteriores se muestran dos histogramas con sus respectivas distribuciones normales correspondientes a las submuestras para fábricas grandes y pequeñas e igual sucede con los gráficos de caja y bigotes. Obsérvese que ahora se sabe que el dato número 100 es un outlier para las fábricas pequeñas y que los datos con números 35, 2 y 57 lo son para las fábricas grandes, pero ninguno de estos lo es cuando se analizan los datos en conjunto. Este tipo de análisis de los datos de la muestra en su conjunto total y por submuestras resulta de gran importancia en un análisis de datos.

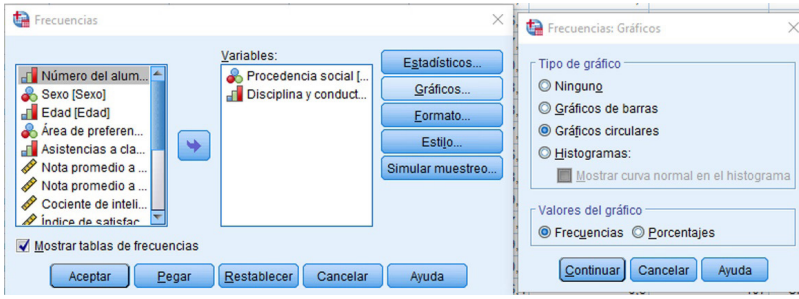


2.2. ¿Cómo agrupar los datos almacenados con SPSS?

La agrupación de los datos es una necesidad del A.E.D., ello genera las distribuciones de frecuencias. Se llama distribución de frecuencias a una tabla en la cual se agrupan en clases los valores posibles para una variable y se registra la frecuencia absoluta correspondiente a cada una, o sea, el número de valores observados que corresponde a cada clase.

De la frecuencia absoluta se obtiene la frecuencia relativa mediante el cociente entre cada frecuencia absoluta y el total de datos. En tanto que la frecuencia porcentual, se obtiene convirtiendo la frecuencia relativa en porcentaje.

En SPSS es posible obtener las tablas de frecuencias siguiendo la secuencia de menú: Analizar → Estadísticos descriptivos → Frecuencias que conduce al siguiente cuadro de diálogo:

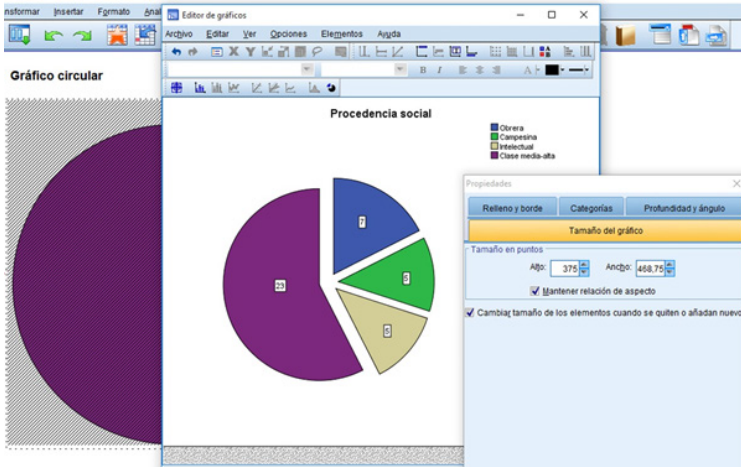


Si se observan las características de las variables tomadas en este ejemplo de la base de datos PROBLEMA BASE puede observarse que ambas son categóricas, por lo que permiten realizar una clasificación de los datos como se muestran en los resultados:

Procedencia social					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Obrera	7	17,5	17,5	17,5
	Campesina	5	12,5	12,5	30,0
	Intelectual	5	12,5	12,5	42,5
	Clase media-alta	23	57,5	57,5	100,0
	Total	40	100,0	100,0	

Disciplina y conducta en la escuela					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Muy mala	4	10,0	10,0	10,0
	Mala	7	17,5	17,5	27,5
	Regular	8	20,0	20,0	47,5
	Buena	5	12,5	12,5	60,0
	Muy buena	16	40,0	40,0	100,0
	Total	40	100,0	100,0	

En cuanto a los gráficos seleccionados la secuencia de edición muestra el resultado del procesamiento de la información:

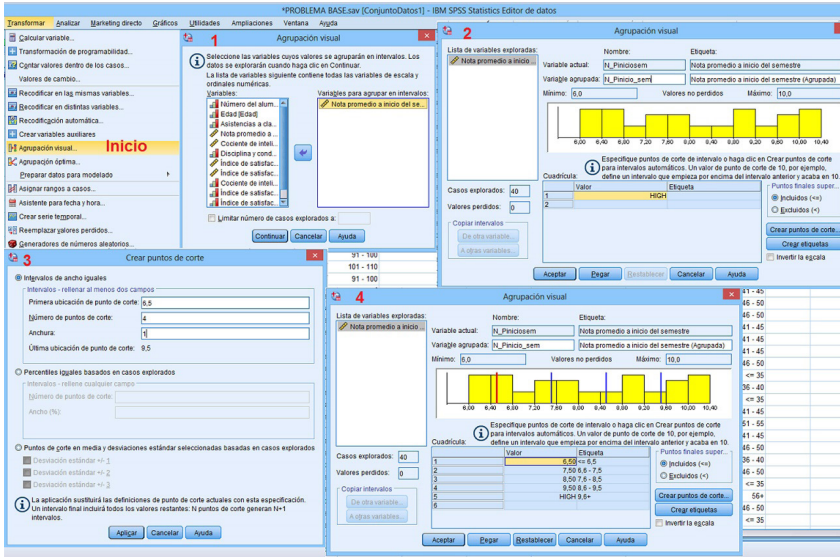


Para las variables de escala el empleo del comando Frecuencias no permite hacer una verdadera agrupación de los datos como se puede observar en el siguiente fragmento de tabla.

Nota promedio a inicio del semestre

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	6,0	1	2,5	2,5	2,5
	6,2	1	2,5	2,5	5,0
	6,3	3	7,5	7,5	12,5
	6,4	2	5,0	5,0	17,5
	6,5	2	5,0	5,0	22,5
	6,6	1	2,5	2,5	25,0
	6,8	2	5,0	5,0	30,0
	7,2	1	2,5	2,5	32,5
	7,3	2	5,0	5,0	37,5
	7,5	1	2,5	2,5	40,0
	7,6	1	2,5	2,5	42,5
	7,7	1	2,5	2,5	45,0
	7,8	1	2,5	2,5	47,5
	7,9	1	2,5	2,5	50,0

Una opción para resolver este problema puede ser el uso del menú Transformar, siguiendo el camino Transformar →Agrupación visual como se muestra en siguiente composición de la secuencia de diálogos.



La opción “Agrupación visual” tiene la particularidad de generar una nueva variable con los datos agrupados, es por ello que se necesita realizar varias consultas al usuario a partir de cuadros de diálogos que se desarrollan después de seleccionar la opción marcada como Inicio y a continuación aparecen los cuadros de diálogos numerados con las siguientes funciones:

Diálogo 1. Se caracteriza por definir las variables que se agruparán en intervalos.

Diálogo 2. Este cuadro tiene varias funciones a las que se accede por los botones y entradas de textos:

a. En *Variable actual* aparece el nombre de la variable seleccionada.

b. En *Variable agrupada* aparece el espacio para que el usuario escriba el nombre que va a dar a la nueva variable.

c. *Bajo Etiquetas* aparece el nombre de la etiqueta actual y el que implícitamente el sistema da a la variable de datos agrupados.

d. Aparece el valor Mínimo y Máximo del conjunto de datos.

e. Debajo aparece el histograma de los datos correspondientes a la variable.

f. El botón Crear puntos de cortes da acceso al siguiente cuadro de diálogo.

Diálogo 3. El objetivo de este cuadro es construir los intervalos en los que quedarán dividido los datos y tiene tres opciones:

a. La primera opción es los intervalos con anchos iguales; para esta opción se necesita:

   i. La ubicación del primer punto de corte; en este caso 6,5.

   ii. Si el usuario selecciona el número de puntos de cortes, SPSS calcula el ancho del intervalo.

   iii. Si por el contrario el usuario escribe la anchura del intervalo, SPSS calcula automáticamente el número de intervalos.

   iv. Tanto para ii como para iii, SPSS devuelve la Última ubicación de punto de corte.

b. La segunda opción ubica los puntos de cortes por los percentiles y tiene un comportamiento análogo a la relación número de puntos de cortes, ancho del intervalo.

   i. Si se da el número de puntos de cortes, SPSS calcula el ancho %".

   ii. Si se da "ancho %" SPSS calcula el número de puntos de cortes.

c. La última opción toma como puntos de cortes la media y las desviaciones estándares. La selección es más sencilla, basta

seleccionar si se desea hacer cuatro intervalos de amplitud una desviación estándar, hasta intervalos de amplitud 3 desviaciones.

d. El botón aplicar de este menú conduce al cuadro de diálogo 4.

Diálogo 4. Este diálogo coincide con el Dialogo 2 con las siguientes particularidades:

a. El histograma tiene ahora los puntos de cortes.

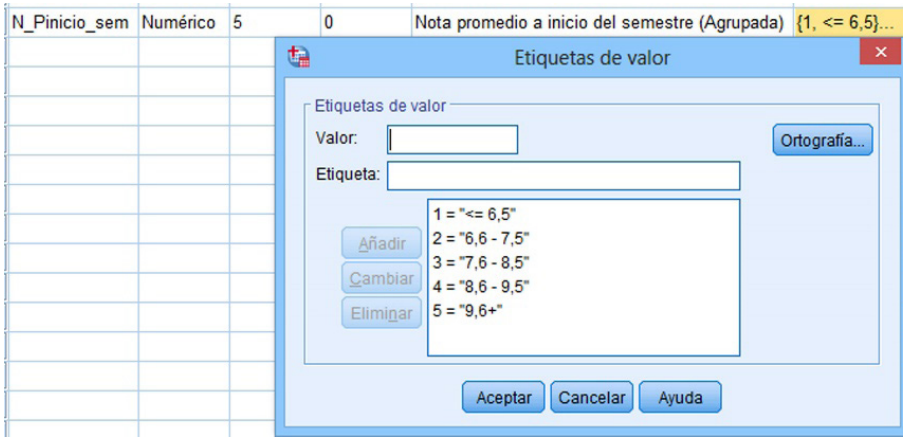
b. Aparecen en la tabla inferior los valores de los puntos de cortes.

c. El botón *Crear etiquetas* permite realizar esa acción que aparece también en la tabla cuando el botón se activa.

d. Permite determinar si se incluyen (<=) los extremos de los intervalos o se excluye (<).

e. Permite invertir la escala activando la pestaña correspondiente.

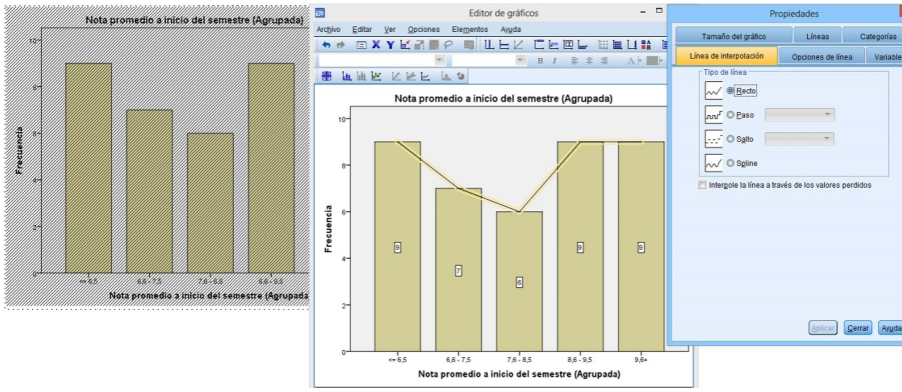
f. Tras pulsar *Aceptar* se crea una nueva variable que aparece en la base de datos, se puede buscar en ella o en la vista de variables, que en el caso del ejemplo tomado de presenta del siguiente modo:



Ahora se está en condiciones de determinar la frecuencia de la nueva variable en forma análoga a cómo se hizo con las varia-

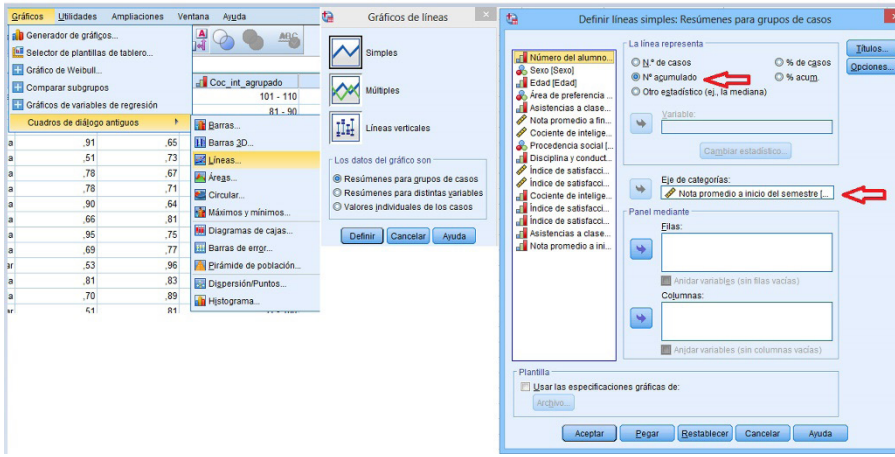
bles “Procedencia social” y “Disciplina y conducta en la escuela”, obteniéndose los siguientes resultados:

Nota promedio a inicio del semestre (Agrupada)					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	<= 6,5	9	22,5	22,5	22,5
	6,6 - 7,5	7	17,5	17,5	40,0
	7,6 - 8,5	6	15,0	15,0	55,0
	8,6 - 9,5	9	22,5	22,5	77,5
	9,6+	9	22,5	22,5	100,0
	Total	40	100,0	100,0	

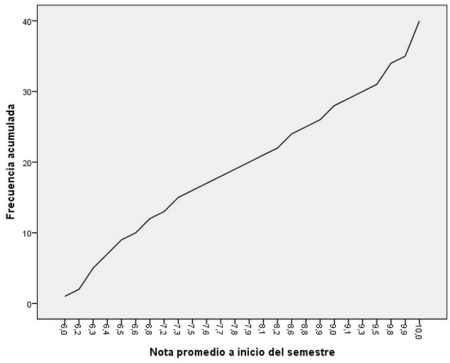


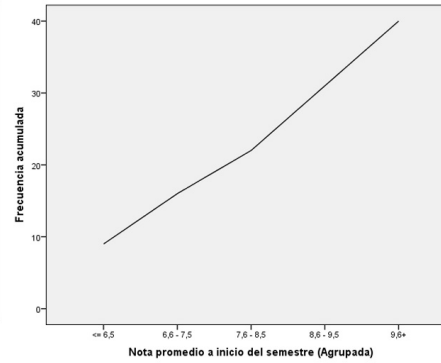
En este caso, para presentar los distintos tipos de gráficos, se muestra la frecuencia en un gráfico de barras con línea de interpolación.

Además de los gráficos de frecuencia y frecuencia relativa, son importantes los de frecuencia acumulada y relativa acumulada; una forma de lograr estos gráficos es mediante la siguiente secuencia:



Para las variables *Nota promedio a inicio del semestre* de la base de datos PROBLEMA BASE y la variable construida Nota promedio a inicio del semestre (agrupada) se obtienen los siguientes gráficos:





2.3. ¿Cómo resumir numéricamente los datos almacenados con SPSS?

En el apartado dedicado al análisis de frecuencias se presentó la necesidad de agrupar las clases por intervalos, esta necesidad de agrupar la información es frecuente en el Análisis de Datos y en el caso que se estudiará en este epígrafe se trata de resumir la información con un solo número. Este número que generalmen-

te se puede situar hacia el centro de la distribución de datos se denomina medida o parámetro de tendencia central o de centralización. Cuando se hace referencia únicamente a la posición de estos parámetros dentro de la distribución, independientemente de que esté más o menos centrada, se habla de estas medidas como medidas de posición y en este caso se incluyen también los cuartiles entre estas medidas, aunque Q_2 coincide con la mediana que es una medida de tendencia central.

Entre las medidas de tendencia central se tienen:

a. Media aritmética.

b. Mediana.

c. Moda.

Aunque estas se extienden también a:

a. Media geométrica.

b. Media armónica.

c. Media recortada al 5%.

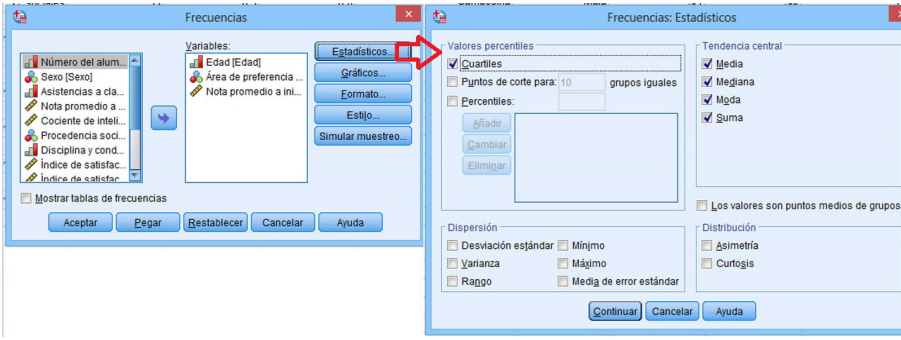
Media Aritmética: cuando se tiene una serie de datos referidos a una magnitud medible, se puede determinar uno que los represente a todos. Se tomará aquel que pueda centralizar más los datos que se posee. Se advierte que la media solo puede ser calculada a variables cuantitativas, por ende, solo a aquellas variables que hayas señalado en el SPSS como variables de escala.

Mediana: la mediana es otra medida de centralización y se define como el valor que ocupa el punto central cuando la serie numérica está dada en orden creciente o decreciente, por eso, lo primero que se tiene que hacer para determinar la mediana es ordenar los datos según sus valores. Cuando el número de términos es par, se toma como mediana la semisuma de los dos valores centrales y cuando es impar, se toma el valor del centro. La mediana se calcula a variables medidas al menos en una escala ordinal, señaladas en SPSS como de tipo escala u ordinal.

La Moda: de una distribución cualitativa es la modalidad que más veces se repite y de una cuantitativa el número que más se redunda. Se puede aplicar a cualquier tipo de variables.

El SPSS facilita el cálculo de estas medidas de tendencia central en diferentes menús, ya se hizo mención a ellas en la opción *Explorar* del menú Estadísticos descriptivos pero el estudio más completo es dentro de la secuencia ya estudiada de

Analizar → Estadísticos descriptivos → Frecuencias



En la gráfica se muestra esta secuencia con la selección de las medidas de tendencias centrales para las variables *Edad,* **Áreas de preferencia** *y Nota Promedio* al Inicio del Semestre de la base de datos PROBLEMA_BASE

Los resultados que devuelve SPSS son los siguientes:

Estadísticos				
		Edad	Área de preferencia	Nota promedio a inicio del semestre
N	Válido	40	40	40
	Perdidos	0	0	0
Media		16,60	2,58	8,095
Mediana		17,00	3,00	8,000
Moda		17	4	10,0

Suma		664	103	323,8
Percentiles	25	16,00	1,00	6,650
	50	17,00	3,00	8,000
	75	17,00	4,00	9,450

Seguramente que el lector que posea conocimientos, aunque sean elementales de estadística, se ha dado cuenta que no tiene sentido una edad media de 16,60 años y mucho menos una media de 2,58 del área de preferencia. Esto indica que aunque el SPSS efectúa los cálculos, es el usuario quien determina el cálculo que debe hacerse; o seleccionar el resultado que debe presentarse, en este caso se tienen variables nominales, ordinales y de escala y cada una admite distintas medidas de tendencia central.

Estadísticos				
		Edad	Área de preferencia	Nota promedio a inicio del semestre
N	Válido	40	40	40
	Perdidos	0	0	0
Media				8,095
Mediana		17,00		8,000
Moda		17	4 (C_humanísticas)*	10,0
Percentiles	25	16,00		6,650
	50	17,00		8,000
	75	17,00		9,450

* Para la variable Área de preferencias a 4 la corresponde C_humanísticas

La media geométrica es un parámetro de centralización que se utiliza para datos exponenciales o del tipo de crecimiento de

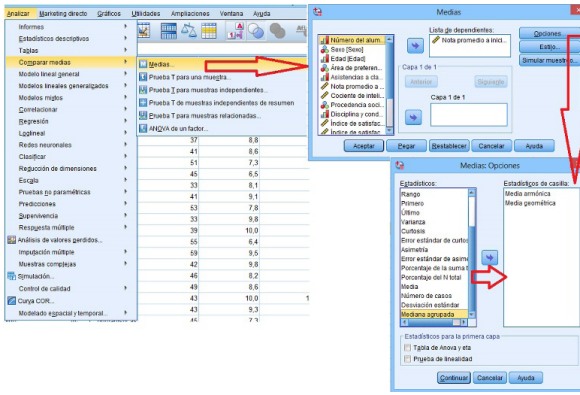
poblaciones. Se calcula multiplicando los datos entre sí y aplicando después la raíz de orden n.

$$\overline{X_G} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n}$$

Se utiliza con mucha menor frecuencia que la media aritmética. También es recomendada para datos de progresión geométrica, para promediar razones, interés compuesto y números índices.

La media armónica (designada usualmente mediante H) de una cantidad finita de datos numéricos es igual al recíproco, o inverso, de la media aritmética de los recíprocos de dichos valores y es recomendada para promediar velocidades.

$$\overline{X_H} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots + \frac{1}{x_n}}$$



En SPSS ambas medias se localizan según la secuencia:

Analizar → Comparar medias→ Medias... (ver imagen ajunta):

El resultado se da en la siguiente tabla:

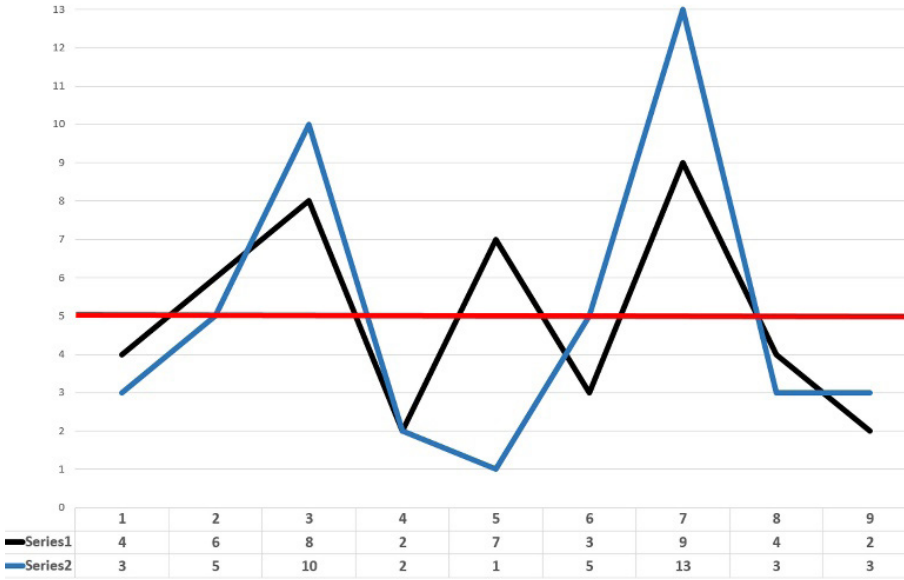
Informe	
Nota promedio a inicio del semestre	
Media armónica	Media geométrica
7,861	7,978

Media recortada al 5%: mediante esta opción el cálculo de la media se hace entre los datos que quedan al eliminar el 5% de los valores superiores y los inferiores, de modo que la media no se ve afectada por los valores extremos, esta media es calculada en Explorar y fue comentada en el epígrafe correspondiente.

Descriptivos				
			Estadístico	Error estándar
Nota promedio a inicio del semestre	Media		8,095	,2191
	95% de intervalo de confianza para la media	Límite inferior	7,652	
		Límite superior	8,538	
	Media recortada al 5%		8,100	

2.4. ¿Cómo determinar la dispersión de los datos almacenados con SPSS?

Aunque las medidas de tendencia central permiten agrupar la información de los datos, por muy concentrados que estén alrededor de tales medidas siempre existen datos que están más o menos alejados de las mismas; por otro lado, es posible que dos o más conjuntos de datos tengan la misma media pero distinta dispersión alrededor de estas como se muestra en el gráfico con dos series de datos que teniendo media igual a 5 la diferencia de dispersión de los datos es evidente y por ellos es necesario determinar cuán disperso está cada uno.



	1	2	3	4	5	6	7	8	9
Series1	4	6	8	2	7	3	9	4	2
Series2	3	5	10	2	1	5	13	3	3

Existen diferentes medidas de dispersión o variabilidad como son:

Recorrido o rango: Es la medida de variación más simple y ya ha sido tratada directa o indirectamente anteriormente. El rango de la muestra está dado por:

$$R = X \text{ mayor} - X \text{ menor}$$

Recorrido intercuartílico: a esta medida ya se hizo referencia al hablar de los valores atípicos y los gráficos de caja y bigotes. Se define por la diferencia existente entre el tercer y el primer cuartil

$$RI = Q3 - Q1$$

Desviación media: esta medida de dispersión hace referencia a un promedio, cosa que no hacen las anteriores; puede entenderse como la media de las desviaciones de los datos de la variable respecto al promedio utilizado; no obstante, para evitar que las desviaciones positivas queden compensadas por las negativas y que esta desviación media resulte igual a 0, (que nos haría

pensar que no hay dispersión) se utiliza el valor absoluto de la desviación de los datos respecto del promedio.

La desviación media respecto de la media se define mediante la expresión:

$$D_{\bar{x}} = \sum_{i=1}^{k} |x_i - \bar{x}| \frac{f_i}{n}$$

También se puede utilizar la desviación media respecto de la mediana como:

$$D_{M_e} = \sum_{i=1}^{k} |x_i - M_e| \frac{f_i}{n}$$

Varianza: se define como la media de los cuadrados de las desviaciones de los valores de la variable respecto de la media aritmética, es decir:

La varianza de una población es la media o promedio del cuadrado de las desviaciones de los elementos respecto a la media poblacional y se representa por. El símbolo se emplea para representar la varianza de las muestras. En una población de tamaño N, de media , la varianza de la población se estima por:

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

En los problemas reales, la media de la población no se conoce, y generalmente se tiene que trabajar con la varianza de una muestra. La varianza de una muestra se estima por la fórmula dada para datos agrupados o no agrupados cuando :

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

Desviación típica o estándar: Se define como la raíz cuadrada positiva de la varianza, es decir:

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

Obsérvese que todos estas estas medidas de dispersión se sustentan en el concepto de distancia, así, el rango y el recorrido intercuartílico son respectivamente la distancia entre el valor máximo y mínimo y entre el primero y tercer cuartil, mientras las desviaciones medias y varianzas son promedios de distancias.

Coeficiente de variación de Karl Pearson: este coeficiente expresa la relación entre el tamaño de la media y la variabilidad de la variable y se define como el cociente entre la desviación típica y el valor absoluto de la media aritmética.

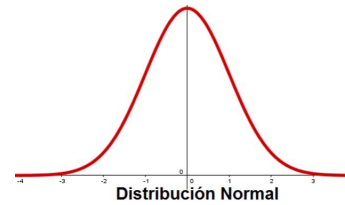
$$CV = \frac{s}{|\bar{x}|}$$

Generalmente el coeficiente de variación se expresa en porciento, para ello la fórmula anterior se multiplica por 100.

$$CV = \frac{s}{|\bar{x}|} * 100$$

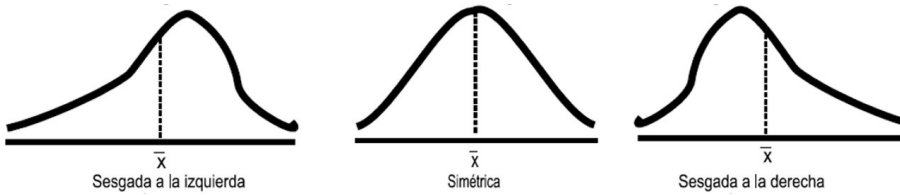
Este coeficiente ilustra con mayor precisión la dispersión de los datos, a mayor valor del coeficiente de variación mayor heterogeneidad de los valores de la variable; y a menor C.V., mayor homogeneidad en los valores de la variable.

Estadígrafos de la distribución: estos estadígrafos son medidas que expresan la forma de la distribución al compararla con la distribución normal (a la cual se ha hecho referencia) y permiten conocer el comportamiento de la

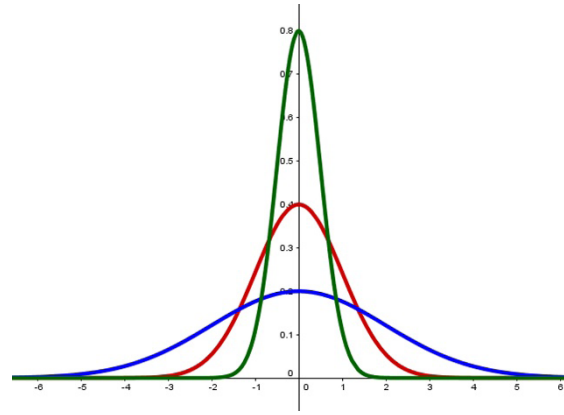


distribución sin ver la gráfica, constatar con las gráficas tal comportamiento, o apreciarlo cuando el mismo es poco perceptible visualmente.

Simetría: medida de la asimetría de una distribución. La distribución normal es simétrica por lo que tiene un valor de asimetría 0. Una distribución que tenga una asimetría positiva significativa tiene una cola derecha larga. Una distribución que tenga una asimetría negativa significativa tiene una cola izquierda larga. Un valor de asimetría mayor que 1, en valor absoluto, indica generalmente una distribución que difiere de manera significativa de la distribución normal. Como regla aproximada, un valor de la asimetría mayor que el doble de su error estándar se asume que indica una desviación de la simetría.



Curtosis: es una medida del grado en que las observaciones se agrupan en torno a un punto central. Para una distribución normal, el valor del estadístico de curtosis es 0. Una curtosis positiva indica que, con respecto a una distribución normal, las observaciones se concentran más en el centro de la distribución y presentan colas más estrechas hasta los valores extremos de la distribución, en cuyo punto las colas de la distribución leptocúrtica son más gruesas con respecto a una distribución normal. Una curtosis



negativa indica que, con respecto a una distribución normal, las observaciones se concentran menos y presentan colas más gruesas hasta los valores extremos de la distribución, en cuyo punto las colas de la distribución platicúrtica son más estrechas con respecto a una distribución normal.

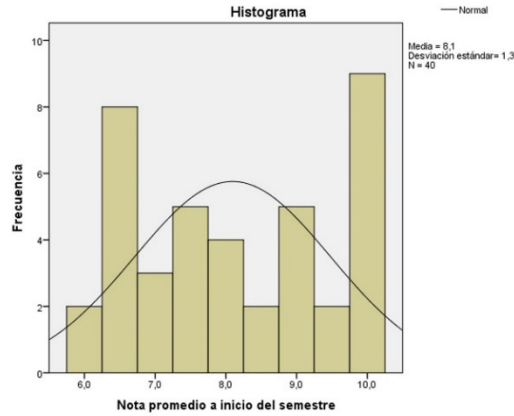
Al tratar la opción *Explorar* del menú Estadísticos descriptivos se presentaron estas medidas en la tabla descriptivos.

Descriptivos		
Nota promedio a inicio del semestre	Estadís-tico	Error están-dar
Media	8,095	,2191
Varianza	1,920	
Desviación estándar	1,3856	
Rango	4,0	
Rango intercuartil	2,8	
Asimetría	,025	,374
Curtosis	-1,493	,733

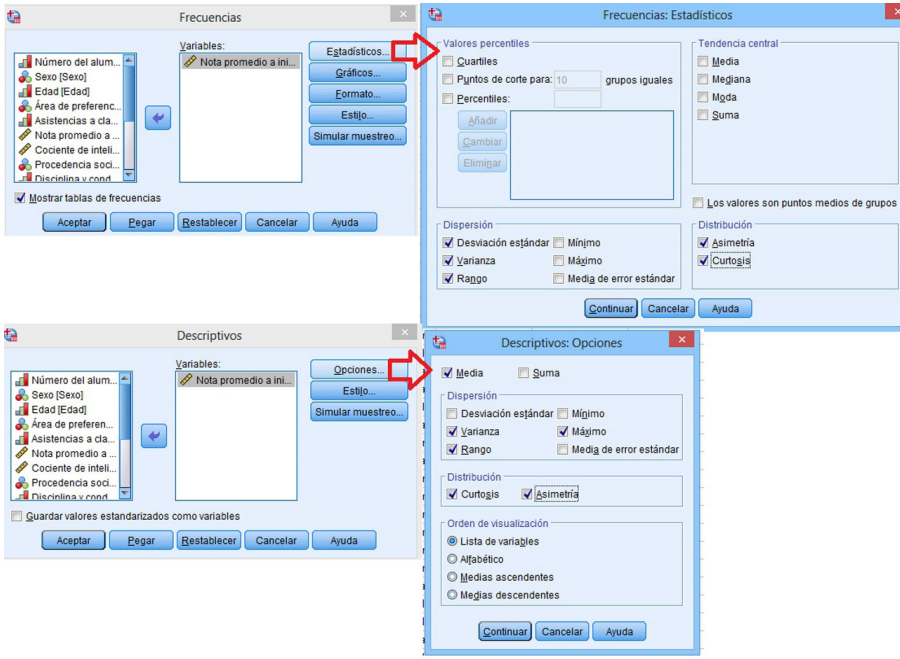
Con estos datos es posible tener una idea de cuán alejados están los datos de la media, así, el rango indica que entre el mayor y el menor valor solo hay 4 unidades, y que la distancia intercuartílica es menor está próxima a 3 lo que indica una concentración de los datos, pero el coeficiente variacional da una información más ilustrativa:

$$CV = \frac{s}{|\bar{x}|} * 100;\ CV = \frac{1,3856}{|8,095|} * 100 = 17,12\%$$

Lo anterior indica que los datos tienen solo una variación aproximada de un 17% respecto a la media; además, la asimetría positiva pero menor que 1 por tanto está ligeramente sesgada a la derecha y es platicúrtica por ser la curtosis negativa lo que corrobora el gráfico.



Desde los submenús *Descriptivos y Frecuencias* del menú *Estadísticos descriptivos* es posible acceder a estas opciones como se muestra en la figura.



Capítulo III. Etapas del A.E.D

3.1. Etapas

No es propósito de los autores de este libro ocupar demasiado espacio del mismo desarrollando teoría sobre A.E.D. y estadística, pero existen aspectos elementales de ambas que requieren un tratamiento mínimo con el propósito de orientar el trabajo a realizar y explicar los algoritmos que se desarrollan en la aplicación del SPSS para la solución de diversos problemas; unos de esos aspectos que merece ser referenciado son las etapas para la resolución de problemas de estadística bajo las concepciones del A.E.D.

En el primer epígrafe del libro se hizo mención a etapas para el *tratamiento de los datos* y en ellas se particularizó en el tratamiento específico y detallado que se debe seguir con los datos en su preparación e identificación para posteriormente desarrollar cualquier procesamiento estadístico, ahora se establecerá la clasificación en tres etapas más generales para enfrentar el Análisis Exploratorio de Datos a partir de la referida preparación de los datos:

Primera etapa

La descripción (representación externa, agrupación en categorías, etc.) de los datos correspondientes a cada indicador seguido de una caracterización consistente en identificar los valores de significativo interés para la investigación que se desarrolla, así como las particularidades y relaciones que existen entre las categorías en que se han agrupado los datos. Atendiendo a las particularidades de cada tipo de variable es preciso contar la cantidad de elementos correspondientes a cada clase (frecuencia absoluta)y en base ella graficar la relación de cantidad correspondiente a cada clase, *visualizando* las clases con mayor, igual o menor cantidad de datos; el análisis de estas relaciones puede ser significativos para la investigación que se realiza y basado en ellas hacer observaciones y valoraciones de tipo cualitativo que expliquen el comportamiento de los datos desde los presupuestos teóricos y prácticos del objeto de estudio de la investigación.

Otro elemento descriptivo de los datos es la frecuencia relativa expresada generalmente en forma porcentual y las *frecuencias acumulativas*. Finalmente se relaciona con esta etapa la determinación de los *estadígrafos de tendencia central y de dispersión, así como la simetría y el apuntamiento.*

El lector comprenderá que el estudio realizado hasta el momento se corresponde con esta primera etapa.

Segunda etapa

En esta etapa se establecen *relaciones estadísticas* entre los datos que se han analizado en forma aislada. Esto lleva a las *tablas de contingencia* y la determinación de los coeficientes de correlación entre variables y los correspondientes *estudios de regresión.*

Tanto los estudios descriptivos desarrollados en la etapa anterior como los de esta segunda etapa son propicios para hacer inferencias generalmente empíricas sobre el comportamiento de los datos en las poblaciones de donde han sido extraídas las muestras estudiadas, las que se verificarán o rechazarán en la tercera etapa.

Tercera etapa

Selección de los modelos estadísticos apropiados para demostrar las inferencias realizadas en caso que esto sea necesario, especialmente, los relacionados con pruebas de hipótesis tanto desde las posiciones de la estadística paramétrica como de la no paramétrica.

En adelante se explicará cómo abordar la solución de problemas relacionados con la segunda etapa utilizando el SPSS y en las ocasiones que sean necesarias se incursionará en acciones de la tercera etapa, porque en realidad esta división en etapa responde más a un problema organizativo y metodológico de la exposición en el texto que a una situación que se corresponda exactamente con el proceso de investigación donde estas etapas se solapan.

3.2. Segunda etapa: Tabla de contingencia y prueba χ^2 de Pearson

Una forma mediante la cual es posible establecer relaciones entre variables es mediante el empleo de *Tablas de contingencia*[9], en las cuales los niveles de un criterio de clasificación de un conjunto forman las filas y los de otro criterio forman las columnas. Las celdas que se encuentran en la intersección de las filas y las columnas contienen conteos o frecuencias de sujetos que se han clasificado en forma cruzada con base en los dos criterios; en general tales tablas tienen la siguiente forma:

		Criterio 2			
		Nivel_1	Nivel_2	Nivel_3	Total filas
Criterio 1	Nivel_1	Cant_1		Cant_3	
	Nivel_2	…	…	…	
	Nivel_3	…	…	…	
	Nivel_4	…	…	Cant_12	
	Total columnas				Total general

Un ejemplo de tales tablas es el siguiente tomado de PROBLEMA_BASE:

Tabla cruzada Sexo*Área de preferencia						
Recuento						
C_exactas		Área de preferencia				Total
		C_naturales	C_sociales	C_humanísticas		
Sexo	Masculino	3	5	2	11	21
	Femenino	10	1	4	4	19
Total		13	6	6	15	40

Indiscutiblemente que la organización de los datos en estas tablas permite hacer inferencias sobre los resultados que los

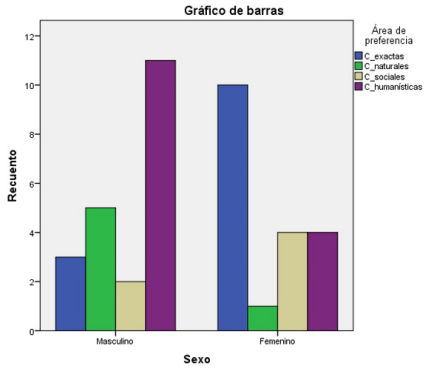
mismos expresan; para el caso del ejemplo anterior la pregunta que "salta a la vista" es si existe relación entre el sexo y las preferencias de los estudiantes por uno u otro grupo de disciplinas científicas. Además de las frecuencias absolutas las tablas pueden reflejar las frecuencias relativas, pero con mayor amplitud, pudiéndose establecer la relación entre la frecuencia absoluta de cada celda con el total de la fila a la que pertenece, con el total de la correspondiente columna o con el total general como se muestra en la siguiente tabla.

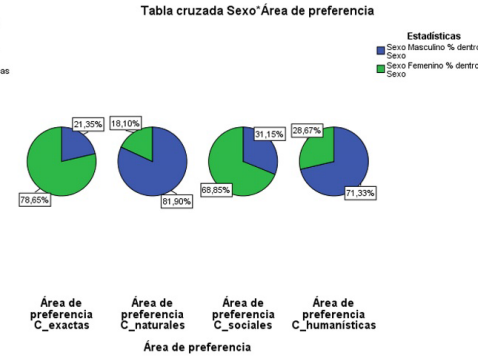
Tabla cruzada Sexo*Área de preferencia							
C_exactas C_naturales			Área de preferencia				Total
			C_sociales	C_humanísticas			
Sexo	Masculino	% dentro de Sexo	14,3%	23,8%	9,5%	52,4%	100,0%
		% dentro de Área de preferencia	23,1%	83,3%	33,3%	73,3%	52,5%
		% del total	7,5%	12,5%	5,0%	27,5%	52,5%
	Femenino	% dentro de Sexo	52,6%	5,3%	21,1%	21,1%	100,0%
		% dentro de Área de preferencia	76,9%	16,7%	66,7%	26,7%	47,5%
		% del total	25,0%	2,5%	10,0%	10,0%	47,5%

Total	% dentro de Sexo	32,5%	15,0%	15,0%	37,5%	100,0%
% dentro de Área de prefe-rencia	100,0%	100,0%	100,0%	100,0%	100,0%	
% del total	32,5%	15,0%	15,0%	37,5%	100,0%	

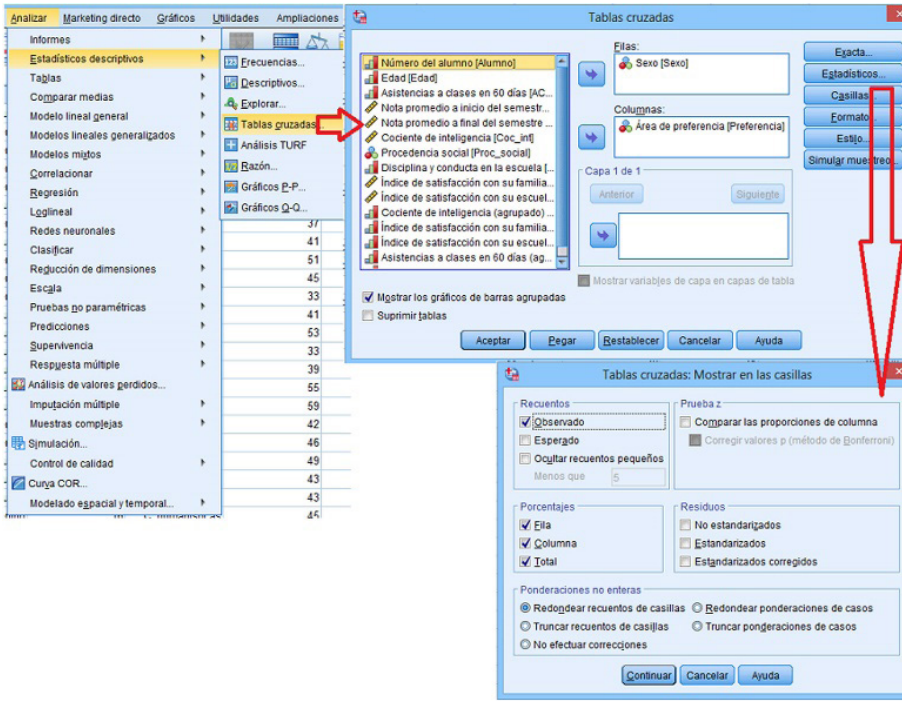
Los gráficos basados en tablas de contingencias se hacen más complejos, lo que contribuye a que el investigador haga inferencias respecto a la existencia o no de dependencia entre los datos a la asociación o correlación entre las variables, la explicación del comportamiento de los datos pareados y finalmente, la generalización y objetivación de los resultados alcanzados en la investigación.

Dos gráficos asociados a las variables seleccionadas de PROBLEMA_BASE son:



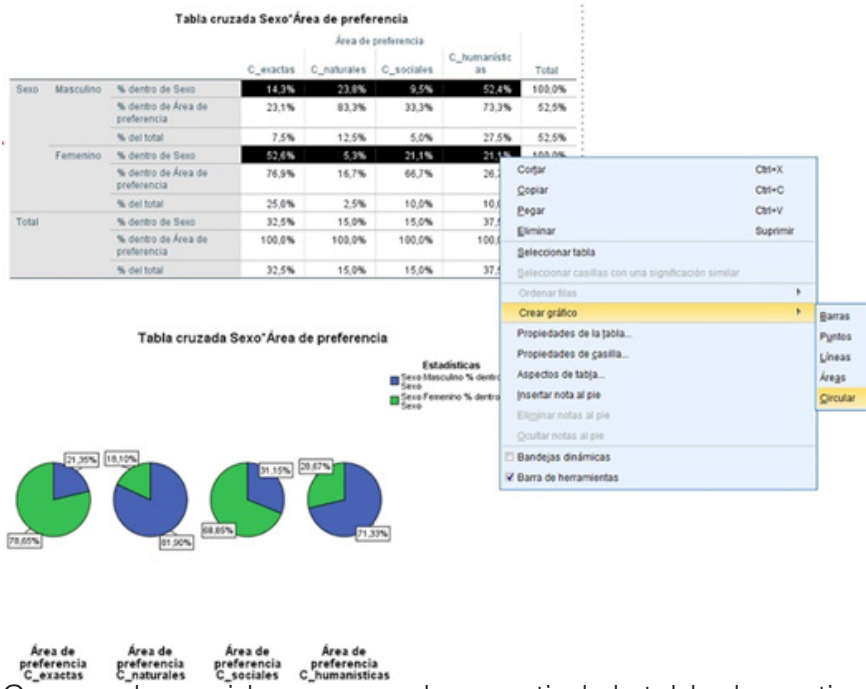


Las tablas de contingencias se obtienen en SPSS siguiendo la secuencia de menú: Analizar → Estadísticos descriptivos → Tablas cruzadas, según se muestra en la siguiente figura:



De esta forma se obtiene la tabla de frecuencias observadas y el gráfico de barra que se mostró anteriormente.

Observe que en la figura están activadas las opciones de *Porcentajes* por *Fila, Columna y Total*. Ello permite obtener la tabla con porcientos antes mostradas. A partir de esta tabla y mediante un proceso de edición análogo al explicado en epígrafes anteriores (la siguiente figura muestra el proceso) es posible obtener gráficos como el de sectores que se ha mostrado anteriormente y que nuevamente aparece en la figura adjunta como resultado de la secuencia de pasos donde se ilustra la vía para obtenerlo.



Como se ha venido expresando, a partir de la tabla de contingencia y de los gráficos que la acompañan y otros que se pueden generar, el investigador puede inferir que es posible que haya relación entre el sexo y las preferencias por una u otra rama de la ciencia. ¿Es cierta o es falsa esta inferencia?

La respuesta corresponde a la tercera etapa definida para E.A.D. Para hacer esa comprobación la estadística se vale de la prueba Chi-cuadrado, Ji-cuadrado, o prueba χ^2 de Pearson[10], esta es una prueba considerada no paramétrica[11] que mide la discrepancia entre una distribución observada y otra teórica (bondad de ajuste), indicando en qué medida las diferencias existentes entre ambas, de haberlas, se deben al azar en el contraste de hipótesis[12]. También se utiliza para probar la independencia de dos variables entre sí, mediante la presentación de los datos en tablas de contingencia como la que se estudia.

Para dar un acercamiento elemental desde el punto de vista de la

estadística a la solución que da la prueba χ^2 de Pearson para determinar si existe dependencia entre dos variables, para el caso del ejemplo que se ha seguido, entre el sexo y las preferencias por distintas ramas de la ciencia, es necesario recordar que en el primer cuadro de diálogo presentado en este epígrafe, se muestra activado el "Recuento Observado", pero también es posible obtener el "Recuento Esperado", para comprender mejor la diferencia entre ambos se presentarán a continuación en la misma tabla diferenciados por un sombreado en la frecuencia esperada:

*Tabla cruzada Sexo*Área de preferencia*							
C_exactas C_naturales			Área de preferencia				Total
			C_sociales	C_humanísticas			
Sexo	Masculino	Recuento	3	5	2	11	21
		Recuento esperado	6,8	3,2	3,2	7,9	21,0
	Femenino	Recuento	10	1	4	4	19
		Recuento esperado	6,2	2,9	2,9	7,1	19,0
Total		Recuento	13	6	6	15	40
Recuento esperado		13,0	6,0	6,0	15,0	40,0	

Analizando el *Recuento esperado* se puede observar que las sumas de las filas y las columnas se mantienen, además, cada celda es el resultado de multiplicar la suma de la fila correspondiente de la frecuencia observa por la suma de la columna y el resultado dividirlo entre la cantidad de individuos de la muestra (40).

Así, el valor 6,8 de la primera celda del Recuento esperado es igual a:

$$\frac{(Suma\ de\ la\ primera\ columna) * (Suma\ de\ la\ primera\ fila)}{Total\ de\ individuos} = \frac{13 * 21}{40} = 6{,}8$$

¿Qué representa en realidad el *Recuento esperado*?

La respuesta a esta pregunta es la clave de la prueba χ^2, porque la tabla del Recuento esperado o frecuencia esperada *se distribuye proporcionalmente* a la cantidad de individuos que hay en cada clase en que se divide la muestra, por tanto *representa cómo debían ser los valores de la tabla si no existiera relación de dependencia entre cada una de las agrupaciones de los datos.*

Evidentemente si ambas tablas coincidieran la diferencia sería 0 y por tanto se podría asegurar que no hay dependencia entre las variables, para el caso ejemplo no habría dependencia entre el sexo y las preferencias por determinadas ramas de la ciencia; como es lógico, esta coincidencia no es frecuente, por tanto es problema es otro, se trata de determinar *cuándo hay diferencia significativa* entre ambas tablas de contingencias, la observada y la esperada y de eso se encarga el SPSS mediante la prueba χ^2.

Es conveniente precisar lo que se entiende por significativo y por diferencia significativa en estadística: se dice que un resultado o efecto es estadísticamente significativo cuando es *improbable que haya sido debido al azar*, por tanto, decir que una diferencia es *estadísticamente significativa* significa que hay evidencias estadísticas de tal diferencia existe y que no es producto de la casualidad o el azar, lo que por supuesto no significa que tal diferencia sea grande, importante o radicalmente diferente.

El procedimiento a seguir tiene como base el procedimiento para determinar tablas de contingencias (tablas cruzadas) ya estudiado, añadiendo ahora las marcas de verificación (tick o check en inglés) relacionadas con:

- Chi-cuadrado[xiii].
- Coeficiente de contingencia.
- Phi y V de Cramer.

Esta selección devuelve la siguiente tabla:

Pruebas de chi-cuadrado			
	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	10,295[a]	3	,016
Razón de verosimilitud	10,864	3	,012
Asociación lineal por lineal	4,735	1	,030
N de casos válidos	40		
a. 4 casillas (50,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 2,85.			

En este caso la comprobación de independencia hace hipótesis de que las variables Sexo y Área de preferencia no están relacionadas, es decir, que las proporciones de columnas son las mismas en todas las columnas y que cualquier discrepancia observada se debe a la variación atribuible al azar. El estadístico de chi-cuadrado mide la discrepancia general entre los recuentos de casillas observados y los recuentos que esperaría si las proporciones de columna fuesen las mismas entre columnas. Un estadístico de chi-cuadrado mayor indica una discrepancia mayor entre los recuentos de casillas observados y esperados; lo que prueba aún más que las proporciones de columna no son iguales, que la hipótesis de independencia no es correcta y, por tanto, que Sexo y Área de preferencia están relacionados.

El estadístico de chi-cuadrado calculado tiene un valor de 10.295. Para determinar si sirve como prueba suficiente para reflejar la hipótesis de independencia, se calcula el valor de significación del estadístico. El valor de significación es la probabilidad de que una variable aleatoria dibujada a partir de una distribución chi-cuadrado con 3 grados de libertad[xiv] sea mayor que 7,8147279[xv]. Puesto que este valor es inferior al nivel alfa especificado en la pestaña Estadísticos de prueba, ***pudiera rechazarse***[xvi] la hipótesis de independencia (recuerde que la hipótesis que asume Chi-cuadrado es que las variables no están relacionadas) en el nivel 0,05. Así, para esta muestra las variables Sexo y Área de preferencia están relacionadas.

En el párrafo anterior se dijo que ***pudiera rechazarse*** porque se cumplen algunas condiciones de la prueba Chi-Cuadrado de independencias, aunque esta prueba no requiere supuestos sobre la forma de la distribución como ocurre con cualquier prueba no paramétrica, se asume que:

Los datos son una muestra aleatoria.

1. Las frecuencias esperadas para cada categoría deberán ser mayor o a lo sumo igual a 1 (1 como mínimo).
2. No más de un 20% de las categorías deberán tener frecuencias esperadas menores que 5.

Por estos dos últimos supuestos, SPSS advierte, como en el caso ejemplo que:

> 4 casillas (50,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 2,85.

Por lo tanto, cualquier análisis que se haga sin que se cumplan estos supuestos no es válido, como ocurre en el caso que se estudia. Ante dificultades como estas, en los textos aparecen diferentes variantes, pero la más efectiva es la de aumentar la muestra tal como se hace a continuación y se muestran los resultados en tabla:

Tabla cruzada Sexo*Área de preferencia

			Área de preferencia				Total
C_exactas C_naturales			C_sociales	C_humanísticas			
Sexo	Masculino	Recuento	7	13	5	29	54
		Recuento esperado	16,2	8,6	8,6	20,5	54,0
	Femenino	Recuento	23	3	11	9	46
		Recuento esperado	13,8	7,4	7,4	17,5	46,0

Total	Recuento	30	16	16	38	100
Recuento esperado	30,0	16,0	16,0	38,0	100,0	

Pruebas de chi-cuadrado			
	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	27,093a	3	,000
Razón de verosimilitud	28,472	3	,000
Asociación lineal por lineal	12,668	1	,000
N de casos válidos	100		
a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 7,36.			

Ahora sí es posible *rechazar* la hipótesis de independencia por las razones dadas en el párrafo comentado.

Con lo expresado se sabe cómo determinar si existe o no asociación entre los datos correspondientes a dos atributos de una muestra, pero no se sabe la magnitud de esa asociación, este problema es consecuencia de que si bien el estadístico está acotado inferiormente por cero, no tiene límite superior, por eso es necesario buscar coeficientes que estén acotados inferior y superiormente, lo que permite medir con más precisión el nivel de asociación entre los atributos.

Existen varios de tales coeficientes, pero el SPSS ofrece dos fundamentales:
Coeficiente de contingencia C de Karl Pearson: se define según la fórmula

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

Este coeficiente toma valores entre 0 y 1. Para el ejemplo se tiene:

$$C = \sqrt{\frac{27{,}093}{27{,}093 + 100}} = 0{,}4617$$

Coeficiente V de Cramer: este coeficiente toma valor 1 cuando existe asociación perfecta sin importar el número de filas y columnas. Es un coeficiente conocido y confiable, que aparece en la mayoría de los asistentes estadísticos. Su expresión matemática es:

$$V = \sqrt{\frac{\chi^2}{mN}}$$

Con m = min(r-1, f-1).

Para el ejemplo se tiene

$$V = \sqrt{\frac{27{,}093}{1*100}} = 0{,}5205.$$

SPSS devuelve el siguiente resultado:

Medidas simétricas			
		Valor	Significación aproximada
Nominal por Nominal	Phi	,521	,000
	V de Cramer	,521	,000
	Coeficiente de contingencia	,462	,000
N de casos válidos		100	

Estos coeficientes indican que existe asociación entre las variables, pero esta es una asociación media porque su diferencia con 0,5 es de solo 0,021.

3.3. Segunda etapa: Correlación y regresión

Las investigaciones estadísticas tienen como objetivos, por un lado, determinar la fuerza de asociación o correlación entre va-

riables y por el otro la generalización y objetivación de los resultados a través de una muestra para hacer inferencia hacia una población. Del cumplimiento al segundo objetivo se encarga la estadística inferencial relacionada con la tercera etapa del A.E.D. que será analizada en otro apartado, pero el primero se puede lograr desde la segunda etapa de A.E.D. mediante el estudio del grado de asociación o correlación entre variables para tratar de hacer inferencia causal, con el propósito de explicar por qué las cosas son así y no de otra manera; por eso se considera que dos variables cuantitativas están correlacionadas cuando los valores de una de ellas varían sistemáticamente con respecto a los valores homónimos (semejantes, parecidos, análogos, equivalentes) de la otra: si tenemos dos variables (A y B) existe correlación si al aumentar los valores de A lo hacen también los de B y viceversa, pero la correlación va más allá de determinar si dos o más variables están correlacionadas o no, la correlación determina la fuerza o intensidad de las relaciones entre las variables, por ejemplo, es un hecho dado por la experiencia y explicado por la ciencia que: la asistencia a clases incide sobre los resultados de los alumnos, al igual que la satisfacción de los alumnos por las condiciones de la escuela, el cociente de inteligencia es un factor que incide en el aprendizaje y por tanto en los resultados académicos; la gestión de venta incide en las ganancias de una empresa; la tensión sistólica basal está íntimamente relacionada con la edad, por mencionar solo algunas de las relaciones entre variables en distintas esferas, pero lo que no se sabe es cuan fuerte son esas relaciones ni con cuánta intensidad se dan y eso solo es posible determinarlo con los estudios de correlación.

Aunque hasta el momento se ha dado poca información sobre la correlación, para evitar errores conceptuales es preciso advertir que la correlación entre dos variables no implica, por sí misma, ninguna relación de causalidad, ejemplo, es posible que en la base PROBLEMA_BASE el alumno número 25 haya asistido a la mitad de las clases en los 60 días estudiados y sin embargo obtuvo el máximo de las calificaciones, mientras el 29 solo faltó 5 días y obtuvo calificaciones mínimas, esto se da dentro un

análisis puntual, pero también es posible que en la referida base de datos se pueda encontrar una correlación estadística entre el número de orden en el listado de los alumnos y los cocientes de inteligencia de estos alumno, pero se sabe que cociente de inteligencia en nada depende del número de orden en el listado de los alumnos de un aula el cual cambia con solo cambiar de grupo.

Por su parte, el análisis de regresión se refiere a la naturaleza de las relaciones entre las variables; la regresión es un conjunto de técnicas que son usadas para establecer una relación entre una variable cuantitativa llamada variable dependiente y una o más variables independientes, llamadas predictoras. Estas deben ser por lo general cuantitativas, aunque usar predictoras que sea cualitativas es permisible.

No se puede hablar de regresión sin mencionar a Francis Galton y Karl Pearson.

El trabajo de Pearson se centró en describir los rasgos físicos de los descendientes (variable A) a partir de los de sus padres (variable B). Estudiando la altura de padres e hijos a partir de más de mil registros de grupos familiares llegó a la conclusión de que los padres muy altos tenían una tendencia a tener hijos que heredaban parte de esta altura, pero que revelaban también una tendencia a regresar (de ahí el término regresión) a la media. Galton generalizó esta tendencia bajo la "ley de la regresión universal": «Cada peculiaridad en un hombre es compartida por sus descendientes, pero en media, en un grado menor».

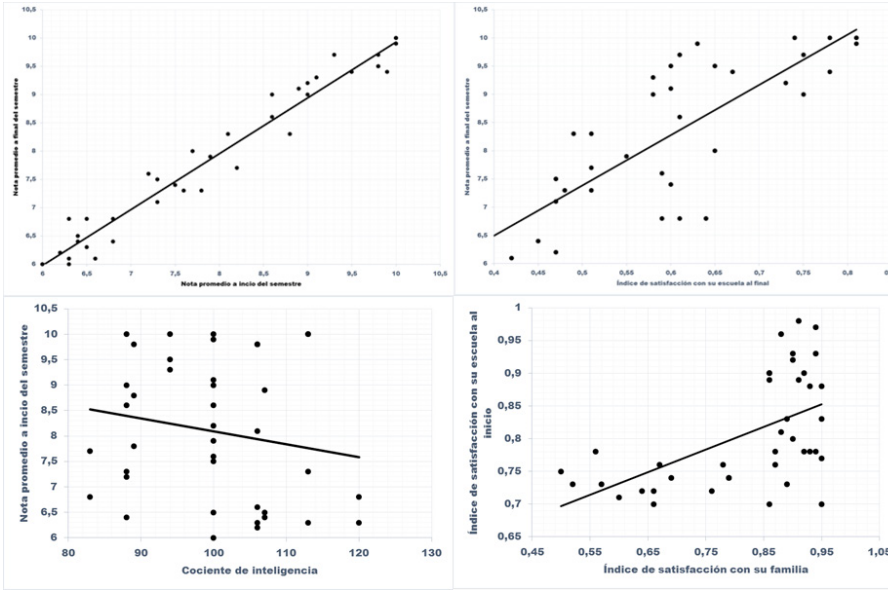
El análisis de regresión es ampliamente utilizado para la predicción y previsión, donde su uso tiene superposición sustancial en el campo de aprendizaje automático, también se utiliza para comprender cuáles de las variables independientes están relacionadas con la variable dependiente, y explorar las formas de estas relaciones.

Cuando la investigación de las relaciones está limitada solamente a dos variables, se denominan esos métodos analíticos como

de análisis de regresión simple y análisis de correlación simple. Cuando se consideran más de dos variables, entonces se necesitan técnicas de análisis de regresión múltiple y análisis de correlación múltiple.

3.4. Los coeficientes de correlación

En las gráficas se muestran las relaciones entre pares de variables de la base PROBLEMA BASE y su comportamiento alrededor de una recta que mejor se aproxima a los datos.



En el primer caso la nube de puntos se encuentra bien cerca de la recta; en el segundo gráfico si bien la nube de puntos está alrededor de la recta las distancias a la misma es mayor que en el primer gráfico; en el tercer gráfico, la dispersión es mucho mayor, no se muestra linealidad en el comportamiento de los puntos, la recta tiene otra posición, finalmente en el cuarto se pueden distinguir tres grupos de puntos más o menos definidos, uno primero en la parte superior de la recta, el segundo en la parte inferior y el tercero más distante en la parte superior.

Este análisis visual permite llegar a conclusiones como:

1. En el primer gráfico las variables: *nota promedio al inicio del semestre y nota promedio al final del semestre* están muy relacionadas.
2. En el segundo *el índice de satisfacción por la escuela al final y la nota promedio al final del semestre* muestran cierta relación, aunque no tanta como en el primer caso.
3. En el tercer gráfico no hay relación, la nube de puntos es un verdadero caos, incluso hay momentos que mientras la variable *cociente de inteligencia* aumenta, la nota promedio a inicio del semestre disminuye.
4. El cuarto caso, los datos de agrupan alrededor de la recta, hasta la mitad aproximadamente se comportan parecidos al gráfico dos, pero al final el comportamiento es anárquico.

Pero estas conclusiones son imprecisas y la estadística no puede trabajar sobre indefiniciones, por eso, para resolver este problema existen los coeficientes de correlación, los que informalmente se puede definir como índices que permiten medir el grado de relación de dos variables, siempre y cuando ambas sean cuantitativas; como ya se sabe por el epígrafe anterior, a la pregunta "¿Existe relación entre el sexo y las preferencias por las asignaturas?" no se puede responder desde los estudios de la correlación y la regresión porque las variables involucradas son nominales; la estadística puede resolverlo mediante la prueba chi-cuadrado y los coeficientes asociados a la misma.

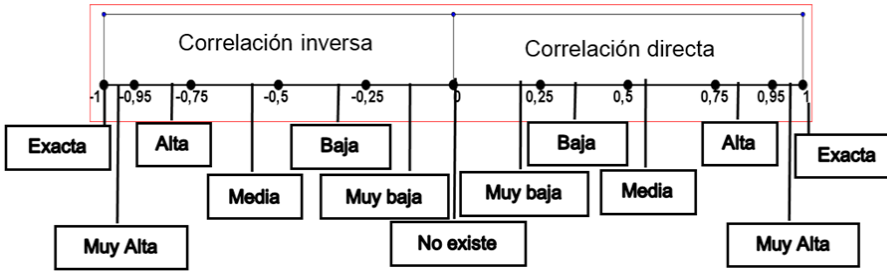
Existen numerosos coeficientes de correlación, pero la relación más sencilla y de amplia aplicación en la Estadística es la correlación lineal.

Los coeficientes de correlación (R) tienen las siguientes propiedades:

1. R siempre es un número real del intervalo [-1; 1], o sea, $-1 \leq R \leq 1$.
2. Si R = 1, existe dependencia lineal directa exacta entre X e Y.

3. Si R = -1, existe dependencia lineal inversa exacta entre X e Y.
4. Si R = 0, no existe dependencia lineal entre X e Y.
5. Cuanto más se aproxime R a -1 o a 1, más dependencia lineal existe entre X e Y; y cuanto más se aproxime R a 0, menos dependencia lineal existe entre X e Y.
6. Si R > 0, al aumentar la variable X, aumenta la variable Y.
7. Si R < 0, al aumentar la variable X, disminuye la variable Y.

La literatura refiere lo expresado en 5, pero siguiendo una Lógica Polivalente[xvii] Lukasiewicz[xviii] – Tarski[xix] es posible emplear la siguiente escala,



3.5. El coeficiente de correlación de Pearson

Para iniciar este estudio es preciso indicar que hasta el momento se han estudiado fundamentalmente indicadores para una sola variable, ahora se hace necesario considerar los de dos variables, una variable independiente X y una dependiente Y, relacionado con ella se tiene:

$\bar{X}$ media de X; $\bar{Y}$ media de Y y S_x y S_y sus desviaciones estándar.

Además se tiene la covarianza entre ambas variables (una varianza conjunta) que se calcula utilizando cualquiera de las siguientes fórmulas:

$$S_{XY} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N} = \frac{\sum_{i=1}^{N} X_i Y_i}{N} - \bar{X}\bar{Y} = \overline{XY} - \bar{X}\bar{Y}$$

En base a estos conceptos se define el coeficiente de correlación de Pearson por la fórmula:

$$R = \frac{S_{XY}}{S_x S_y}$$

Para las variables

X: Nota promedio a inicio del semestre.

Y: Nota promedio a final del semestre

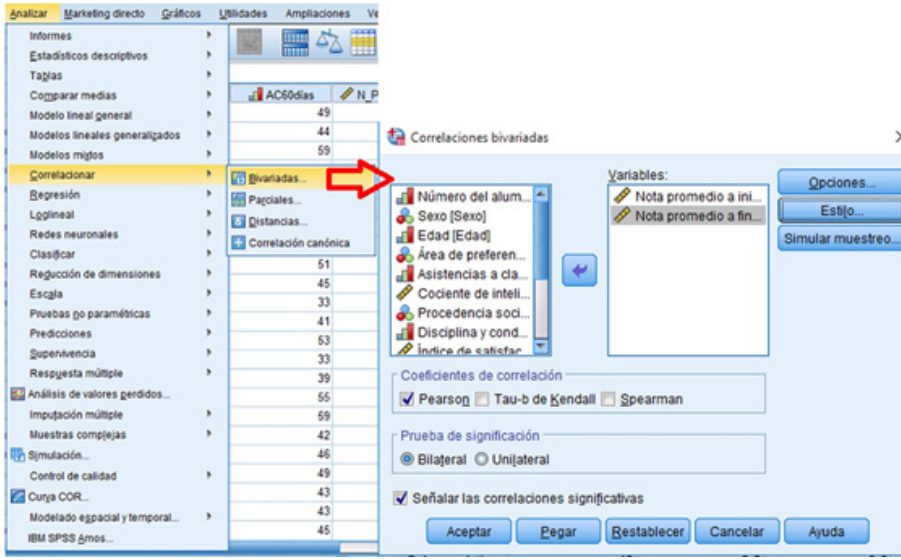
Se tiene:

Estadística Descriptiva			
	N	Media	Desviación estándar
Nota promedio a inicio del semestre	40	8,095	1,36820137
Nota promedio a final del semestre	40	8,050	1,37749773
Nota_al_inicio_X_Nota_al_final	40	67,0123	22,0996344
Casos válidos	40		

Con estos datos el cálculo de R se expresa del siguiente modo:

$$R = \frac{S_{XY}}{S_x S_y} = \frac{67{,}0123 - 8{,}095 * 8{,}050}{1{,}36820137 * 1{,}377497731} = \frac{1{,}8475}{1{,}8846943} = 0{,}980$$

Coeficiente que según la escala se corresponde con una ALTA CORRELACIÓN mediante SPSS el procedimiento es sencillo según se muestra en la siguiente imagen:



El resultado es la siguiente tabla de doble entrada, que como tal expresa una lectura de filas por columnas.

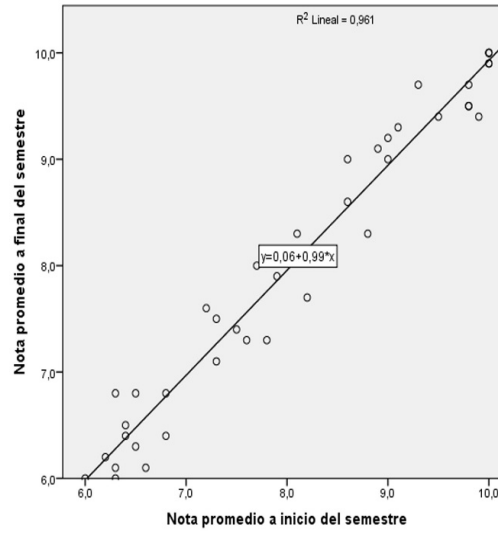
Correlaciones			
		Nota promedio a inicio del semestre	Nota promedio a final del semestre
Nota promedio a inicio del semestre	Correlación de Pearson	1	,980**
	Sig. (bilateral)		,000
	N	40	40
Nota promedio a final del semestre	Correlación de Pearson	,980**	1
	Sig. (bilateral)	,000	
	N	40	40
**. La correlación es significativa en el nivel 0,01 (bilateral).			

En las celdas correspondientes a la primera fila y primera columna y segunda fila segunda columna la correlación es 1, evidente se trata de una variable consigo misma. En las otras dos celdas están las correlaciones que interesan, que como se puede ver coinciden con el valor anteriormente calculado.

El cálculo del coeficiente de correlación de Pearson tiene las siguientes exigencias:

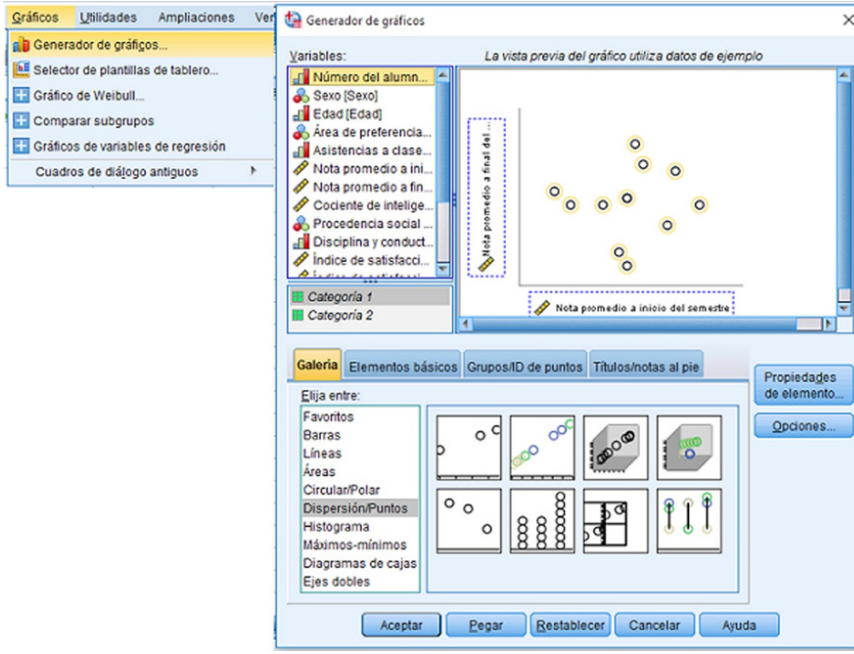
1. Para calcular el coeficiente de correlación de Pearson *ambas variables deben estar al menos en escala por intervalos.* En caso que esta condición no se cumpla es posible la determinación de otros coeficientes de correlación algunos de los cuales se mostrarán posteriormente.

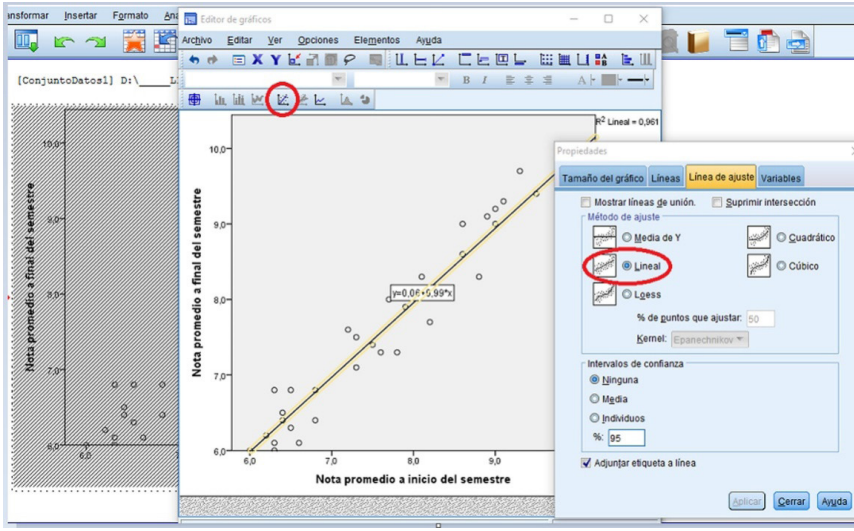
2. Una correlación estadísticamente significativa, por ejemplo, $p < .05$, quiere decir que si no hay elación en la población (es decir, si se da esa condición importante de ausencia de relación) la probabilidad de obtener un coeficiente de esa magnitud por puro azar es inferior al 5%. Para los casos estudiados la correlación es significativa al nivel 0,01 (bilateral). Esta indicación expresa que la correlación obtenida es confiable en un 99%.



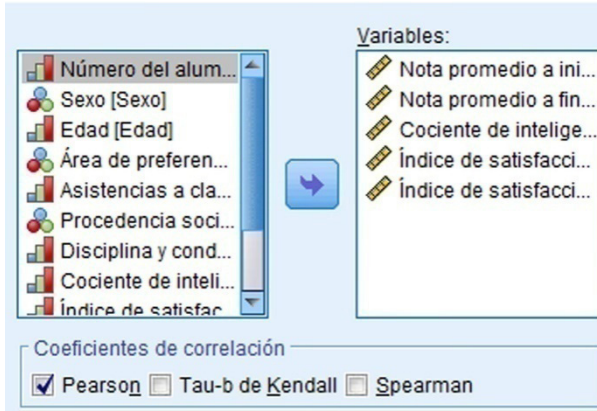
El gráfico correspondiente al ejemplo es el que se adjunta, al cual se ha añadido la recta de ajuste, la acumulación de la nube de puntos alrededor de la recta de ajuste ilustra la alta correlación que existe entre las dos variables estudiadas.

La vía de acceder a la construcción y la edición de este gráfico se muestra en las siguientes imágenes.





Es posible obtener una matriz con todas las correlaciones de las variables que pertenecen a la muestra a partir de seleccionar más de dos variables en el menú inicial como se muestra en la figura adjunta y el resultado en la tabla siguiente:



Correlaciones

		Nota promedio a inicio del semestre	Nota promedio a final del semestre	Cociente de inteligencia	Índice de satisfacción con su familia	Índice de satisfacción con su escuela
Nota promedio a inicio del semestre	Correlación de Pearson	1	,980**	-,179	-,381*	,016
	Sig.		,000	,270	,015	,924
Nota promedio a final del semestre	Correlación de Pearson	,980**	1	-,211	-,347*	,050
	Sig.	,000		,191	,028	,761
Cociente de inteligencia	Correlación de Pearson	-,179	-,211	1	,110	,123
	Sig.	,270	,191		,501	,450

Índice de satisfacción con su familia	Correlación de Pearson	-,381*	-,347*	,110	1	,054
	Sig.	,015	,028	,501		,742
Índice de satisfacción con su escuela	Correlación de Pearson	,016	,050	,123	,054	1
	Sig.	,924	,761	,450	,742	
**. La correlación es significativa en el nivel 0,01 (bilateral).						
*. La correlación es significativa en el nivel 0,05 (bilateral).						

Observe que hay correlaciones que no son significativas, variables que tienen una correlación con significación en el nivel 0,01 y en el nivel 0.05. Hay correlaciones inversas (negativas) y por supuesto en la diagonal principal todos los valores son 1 al comprarse cada variable consigo misma.

3.6. Los coeficientes de correlación de Spearman[xx] y de Kendall

Al tratar el coeficiente de correlación de Pearson se advirtió que solo se puede utilizar cuando *ambas variables estén al menos en escala de intervalo*, pero para el caso en que ambas variables estén al menos en escala ordinal, existe el coeficiente de correlación de Spearman, ρ (ro), el cual es una medida de la correlación propia de la estadística no paramétrica.

Para calcular ρ, los datos son ordenados (preferiblemente en orden decreciente) y reemplazados los valores por su respectivo orden o rango (1, 2, 3, …N) de modo que rango(X) es el número de orden que ocupó el dato X y rango(Y) el número de orden que corresponde al dato Y. El estadístico ρ viene dado por la expresión:

$$\rho = 1 - \frac{6\sum D^2}{N^3 - N}$$

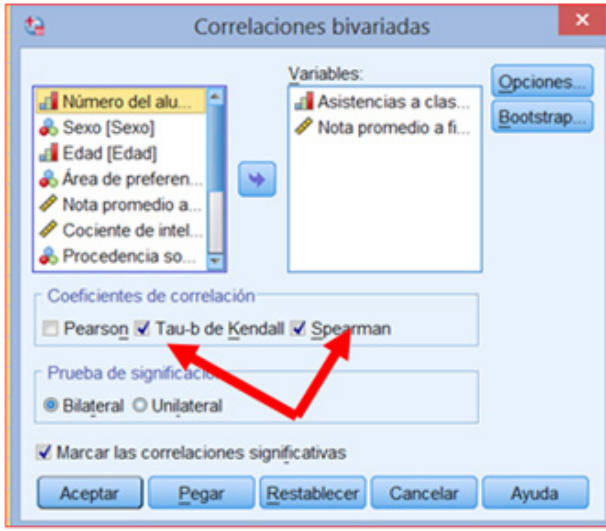
D: diferencia (rango(X) - rango (Y)); N: número de parejas.

La tabla muestra el cálculo de coeficiente de correlación de Spearman para las variables Asistencias a clases en 60 días y Nota promedio a final del semestre.

X: Asistencias	Y: Notas al final	Rango (X)	Rango(Y)	[Rango(X)-Rango (Y)]2
49	9,1	9	14	25
44	6,4	21	33	144
59	6	1	39	1444
45	6,1	16	37	441
30	7,3	38	26	144
35	10	33	1	1024
37	8,3	32	18	196
41	8,6	26	17	81
51	7,5	8	24	256
45	6,3	17	35	324
33	8,3	35	19	256
41	9,3	27	12	225
53	7,3	7	27	400
33	9,5	36	8	784
39	9,9	31	4	729
55	6,5	5	32	729
59	9,4	2	10	64
42	9,5	25	9	256
46	7,7	14	22	64
49	9	10	15	25
43	10	22	2	400
43	9,7	23	6	289
45	7,1	18	28	100
49	6	11	40	841

30	10	39	3	1296
40	7,9	28	21	49
35	6,1	34	38	16
45	6,4	19	34	225
55	6,8	6	29	529
43	9	24	16	64
49	8	12	20	64
40	6,8	29	30	1
46	9,2	15	13	4
30	6,8	40	31	81
58	9,9	3	5	4
49	9,7	13	7	36
31	7,4	37	25	144
58	6,2	4	36	1024
40	9,4	30	11	361
45	7,6	20	23	9
		Numerador	Suma	13148
			6*suma	78888
		Denominador	N3-N	63960
			Fracción	1,23339587
		Coeficiente de Spearman	1-Fracción	-0,23339587

Análogo al coeficiente de correlación de Spearman existe el coeficiente de correlación de Kendall, que también se rigen por criterios análogos al explicado. Para calcularlos con SPSS al seleccionar correlación aparece el cuadro de diálogo que se adjunta y en él se seleccionan los métodos correspondientes, resulta la siguiente tabla:



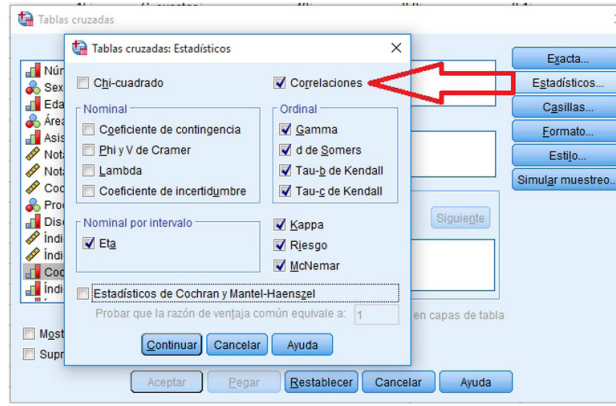
Correlaciones				
			Asistencias a clases en 60 días	Nota promedio a final del semestre
Tau_b de Kendall	Asistencias a clases en 60 días	Coeficiente de correlación	1,000	-,167
		Sig. (bilateral)	.	,140
		N	40	40
	Nota promedio a final del semestre	Coeficiente de correlación	-,167	1,000
		Sig. (bilateral)	,140	.
		N	40	40

Rho de Spearman	Asistencias a clases en 60 días	Coeficiente de correlación	1,000	-,236
		Sig. (bilateral)	.	,143
		N	40	40
	Nota promedio a final del semestre	Coeficiente de correlación	-,236	1,000
		Sig. (bilateral)	,143	.
		N	40	40

Se ha destacado el resultado del coeficiente de correlación de Spearman por la coincidencia con el cálculo manual a partir de la fórmula. El resultado evidencia una baja correlación inversa entre asistencias y resultados académicos.

Al igual que con el coeficiente de correlación de Pearson es posible obtener una matriz con todas las correlaciones entre variables en la escala propia para empleas.

Desde la opción de Tablas Cruzadas es posible acceder al cálculo de correlaciones de variables de tipo ordinal como se muestra en la figura adjunta.



La explicación de las opciones de este cuadro de diálogo se sintetiza a continuación («») tomando lo que se expresa en la ayuda del SPSS (los subrayados son sugerencias de los autores):

«*Ordinal*. Para las tablas en las que tanto las filas como las columnas contienen valores ordenados, seleccione Gamma (orden cero para tablas bidimensionales y condicional para tablas cuyo factor de clasificación va de 3 a 10), Tau-b de Kendall y Tau-c de Kendall. Para pronosticar las categorías de columna de las categorías de fila, seleccione d de Somers.

- *Gamma*. Medida de asociación simétrica entre dos variables ordinales cuyo valor siempre está comprendido entre -1 y 1. Los valores próximos a 1, en valor absoluto, indican una fuerte relación entre las dos variables. Los valores próximos a cero indican que hay poca o ninguna relación entre las dos variables. Para las tablas bidimensionales, se muestran las gammas de orden cero. Para las tablas de tres o más factores de clasificación, se muestran las gammas condicionales.

- *d de Somers*. Medida de asociación entre dos variables ordinales que toma un valor comprendido entre -1 y 1. Los valores próximos a 1, en valor absoluto, indican una fuerte relación entre las dos variables. Los valores próximos a cero indican que hay poca o ninguna relación entre las dos variables. La d de Somers es una extensión asimétrica de gamma que difiere solo en la inclusión del número de pares no empatados en la variable independiente. También se calcula una versión no simétrica de este estadístico.

- *Tau-b de Kendall*. Medida no paramétrica de la correlación para variables ordinales o de rangos que tiene en consideración los empates. El signo del coeficiente indica la dirección de la relación y su valor absoluto indica la fuerza de la relación. Los valores mayores indican que la relación es más estrecha. Los valores posibles van de -1 a 1, pero un valor de -1 o +1 solo se puede obtener a partir de tablas cuadradas.

- *Tau-c de Kendall*. Medida no paramétrica de asociación para variables ordinales que ignora los empates. El signo del coeficiente indica la dirección de la relación y su valor absoluto indica la fuerza de la relación. Los valores mayores indican

que la relación es más estrecha. Los valores posibles van de -1 a 1, pero un valor de -1 o +1 solo se puede obtener a partir de tablas cuadradas.

Nominal por intervalo. Cuando una variable es categórica y la otra es cuantitativa, seleccione Eta. La variable categórica debe codificarse numéricamente.

- *Eta*. Medida de asociación cuyo valor siempre está comprendido entre 0 y 1. El valor 0 indica que no hay asociación entre las variables de fila y de columna. Los valores cercanos a 1 indican que hay gran relación entre las variables. Eta resulta apropiada para una variable dependiente medida en una escala de intervalo (por ejemplo, ingresos) y una variable independiente con un número limitado de categorías (por ejemplo, género). Se calculan dos valores eta: uno trata la variable de las filas como una variable de intervalo; el otro trata la variable de las columnas como una variable de intervalo.

- *Kappa*. La kappa de Cohen mide el acuerdo entre las evaluaciones de dos jueces cuando ambos están valorando el mismo objeto. Un valor igual a 1 indica un acuerdo perfecto. Un valor igual a 0 indica que el acuerdo no es mejor que el que se obtendría por azar. Kappa se basa en una tabla cuadrada en la que los valores de filas y columnas representan la misma escala. Cualquier casilla que tenga valores observados para una variable, pero no para la otra se le asigna un recuento de 0. No se calcula Kappa si el tipo de almacenamiento de datos (cadena o numérico) no es el mismo para las dos variables. Para una variable de cadena, ambas variables deben tener la misma longitud definida.

- *Riesgo*. Para tablas 2x2, una medida del grado de asociación entre la presencia de un factor y la ocurrencia de un evento. Si el intervalo de confianza para el estadístico incluye un valor de 1, no se podrá asumir que el factor está asociado con el evento. Cuando la ocurrencia del factor es poco común, se puede utilizar la razón de las ventajas como estimación o riesgo relativo.

- *McNemar*. Prueba no paramétrica para dos variables dicotómicas relacionadas. Contrasta los cambios de respuesta utilizando una distribución chi-cuadrado. Es útil para detectar cambios en las respuestas causadas por la intervención experimental en los diseños del tipo "antes-después". Para las tablas cuadradas de mayor orden se informa de la prueba de simetría de McNemar-Bowker.

- *Estadísticos de Cochran y Mantel-Haenszel*. Los estadísticos de Cochran y de Mantel-Haenszel se pueden utilizar para comprobar la independencia entre una variable de factor dicotómica y una variable de respuesta dicotómica, condicionada por los patrones en las covariables, que vienen definidos por la variable o variables de las capas (variables de control). Tenga en cuenta que mientras que otros estadísticos se calculan capa por capa, los estadísticos de Cochran y Mantel-Haenszel se calculan una sola vez para todas las capas.

3.7. Regresión

Desde el inicio del tema, junto al gráfico de dispersión de los puntos se ha mostrado una línea recta y con relación a ella se han hecho varios comentarios, esta es la llamada recta de regresión. Aunque la correlación, informa sobre la intensidad de una relación lineal, no dice cuál es la relación numérica exacta y ella se expresa mediante la ecuación de la recta de regresión.

El análisis de regresión permite obtener una ecuación que produce valores de Y para valores dados X. Uno de los principales objetivos del análisis de regresión es hacer predicciones y aunque generalmente no se predice el valor exacto de Y, se acepta por lo general si las predicciones están razonablemente cercanas a los valores reales. El estadístico busca una ecuación que le permita expresar la relación entre los datos de las variables. La ecuación que se elige es aquella que se ajusta mejor al diagrama de dispersión.

Existen varias relaciones que reciben el nombre de ecuación de predicción, dentro de éstas están:

$Y = a + bX$	Lineal	(1)
$Y = a + bX + X^2$	Cuadrática	(2)
$Y = ab^X + c$	Exponencial	(3)
$Y = a \log_{10} X + b$	Logarítmica	(4)

Cada ecuación corresponde al modelo que mejor se ajuste a la distribución de los datos en el gráfico de dispersión. Ahora bien, dado un conjunto de datos binarios (X, Y), ¿cómo se obtiene la recta de mejor ajuste? Si parece apropiada una relación definida por una recta, la recta de mejor ajuste se encuentra utilizando el método de los mínimos cuadrados[xxi].

La ecuación (1) de la recta del mejor ajuste está determinada por su pendiente "b" y su ordenada en el origen "a". Los valores de las constantes que satisfacen el criterio de mínimos cuadrados se obtienen con las siguientes fórmulas:

Pendiente:

$$b = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sum X^2 - \frac{(\sum X)^2}{N}}$$

Ordenada en el origen

$$a = \frac{1}{N}\left(\sum Y - b \sum X\right)$$

Ecuación de regresión:

$$Y = a + bX$$

La pendiente b, indica el cambio predicho en Y como resultado de que X se incremente en una unidad.

La ordenada en el origen es el valor de Y en el que la recta de mejor ajuste corta al eje vertical, es decir, donde X=0. Sin embargo, al interpretar "a" primero debe considerarse si X= 0 es un

valor razonable, antes de concluir que Y= a. Cuando se hagan predicciones acerca del valor de Y con base en un valor X, se debe estar seguro de que el valor de X está dentro del dominio de los valores.

De los datos de las variables “Índice de satisfacción con su escuela al final” y “Nota promedio a final del semestre” (base PROBLEMA_BASE) se obtiene la siguiente tabla que permite calcular los parámetros a y b de la recta de regresión.

SUMA X	SUMA Y	SUMA XY	SUMA X^2
22,98	322	190,932	13,8696
b=	8,90217049	a=	2,93570305

Recta de regresión

$$Y = 2{,}93570305 + 8{,}90217049\,X$$

¿Qué aporta esta ecuación?

De los datos de ambas variables se tienen las siguientes medias.

MEDIAS DE:	
Índice de satisfacción con su escuela al final	Nota promedio a final del semestre
0,5745	8,05

Si se evalúa en la ecuación de regresión la media de la variable Índice de satisfacción con su escuela al final se obtendrá la media de Nota promedio a final del semestre.Si se sustituye en la ecuación de la recta de regresión X por 0,5745 es posible comprobar que y=8,05. Es decir, la ecuación de regresión establece la relación que existe entre la media de los datos de la variable independiente y la media de los datos de la variable dependiente, ahora se está en condiciones de poder “predecir” para las condiciones de ese grupo cómo debe comportarse la media de los resultados en dependencia del comportamiento de la media del índice de satisfacción con su escuela.

Concretamente, si a partir de un trabajo del claustro y la dirección de la escuela se logra que el promedio del índice de satisfacción con la escuela alcance un valor próximo a 0,7, entonces la nota promedio debe alcanzar un valor cercano a 9,17.

Ahora es necesario preguntase, ¿Qué confiabilidad tiene esta predicción?, porque en el caso que nos ocupa, donde los factores subjetivos juegan un rol determinante un error es comprensible, pero con ecuaciones como estas se determina desde la posibilidad de éxito de una vacuna hasta las predicciones para decisiones económicas de un país.

En matemática se demuestra que R^2 (el cuadrado del coeficiente de correlación) es la proporción de error que se elimina de S^2_Y (varianza de Y) cuando se pronostica Y mediante la recta de regresión mínimo cuadrática, en vez de pronosticarlo mediante la media; bajo estas consideraciones se puede concluir que cuanto más se aproxime R^2 a 1, mayor es la bondad del ajuste lineal; y mientras más lo haga a cero, menor será esta.

En el caso que nos ocupa $R^2 \approx 0{,}700569$, de manera que la bondad de ajuste lineal es media siguiendo la escala convenida. Por tanto, la predicción o pronóstico utilizando la ecuación tiene una efectividad media.

Observe que para el par de variables del ejemplo escogido para este estudio el valor de R es 0,837 que lo hace clasificar como una *alta correlación*, pero es el valor de R^2 quien significa que se ha eliminado solo un 70% de error al pronosticar Y mediante la recta de regresión mínimo cuadrática, en lugar de hacerlo utilizando la media.

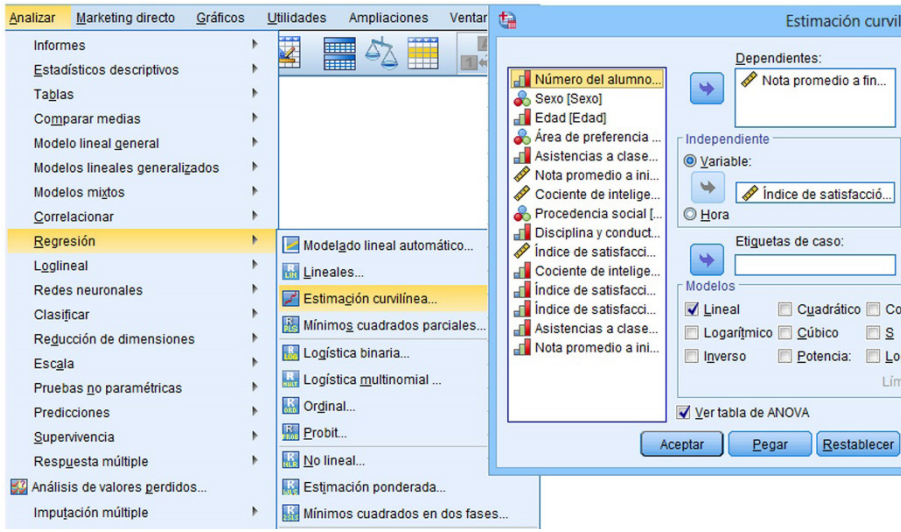
Note que, aunque se proporciona una escala que puede ayudar a decidir primero si la correlación es B y ahora si lo es el ajuste a la recta de regresión el problema no se puede decidir categóricamente y menos confiar ciegamente en los resultados de un número que solo indica una tendencia de los datos empíricos. Este problema es realmente extremadamente complejo y está determinado, entre otros factores, por la homogeneidad de las varianzas y la presencia de datos atípicos.

Como conclusión y recomendación final es que solo el uso combinado del gráfico, del valor de R^2, y de la experiencia derivada del estudio particular que se desarrolla con las variables que se estudian pueden ser criterios de la confiabilidad predictiva de la ecuación de regresión.

Otros autores, partiendo de la recta Y=a+bX proponen formular el modelo de regresión lineal en la forma Y=a+bX+e

siendo e el error típico de estimación, es decir, la diferencia entre la puntuación predicha por el modelo y la observada.

A continuación, se ilustrará el proceso de cálculo de la ecuación de regresión con SPSS y los resultados que ofrece la aplicación para las variables que se han venido estudiando:



Resumen del modelo			
R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
,835	,697	,689	,778
La variable independiente es Índice de satisfacción con su escuela.			

En la tabla anterior se han destacado dos valores principales, R^2 (R cuadrado) y el Error estándar de la estimación por la importancia que tiene para la ecuación de regresión; del primero ya se habló, él expresa que se ha eliminado un 70% de error al pronosticar y el segundo expresa el error estándar de la estimación (al que llamaremos Se) es la desviación típica de los residuos, es decir, la desviación típica de las distancias existentes entre las puntuaciones en la variable dependiente (Yi) y los pronósticos efectuados con la recta de regresión, aunque no exactamente, pues la suma de las distancias al cuadrado están divididas por n-2:

$$Errorestándardelaestimación = S_e = \sqrt{\sum {(Y_i - \hat{Y}_i)^2}/{n-2}}$$

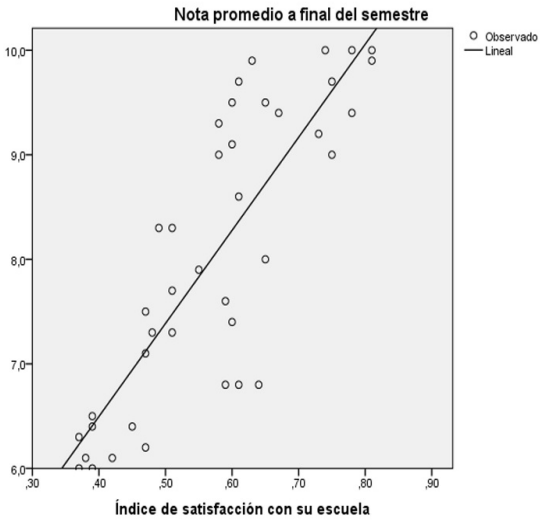
En realidad, este error típico es la raíz cuadrada de la media cuadrática residual y representa una medida de la parte de variabilidad de la variable dependiente que no es explicada por la recta de regresión. En general, cuanto mejor es el ajuste, más pequeño es este error típico. Para el caso que se ejemplifica el valor 0,778 indica que no se está en presencia de un buen ajuste.

Coeficientes					
	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
	B	Error estándar	Beta		
Índice de satisfacción con su escuela	8,902	,952	,835	9,350	,000
(Constante)	2,936	,561		5,237	,000

De esta tabla también se destacan los valores que conformarán la ecuación, observe que no existe diferencia con los datos calculados manualmente.

La ecuación de regresión es

$$Y = 2,936 + 8,902X$$

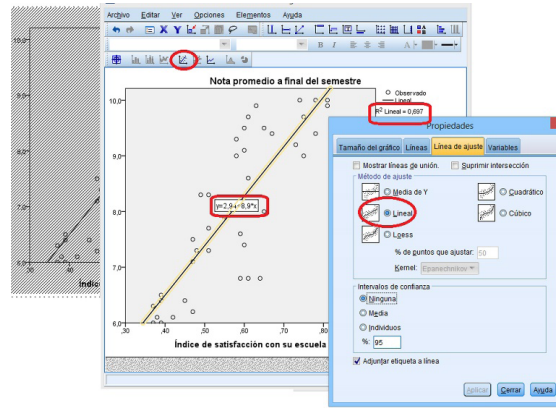


Los coeficientes Beta (coeficientes de regresión parcial estandarizados) son los coeficientes que definen la ecuación de regresión cuando ésta se obtiene tras estandarizar las variables originales, es decir, tras convertir las puntuaciones directas en típicas:

$$\beta_1 = B_1 \left({}^{S_x}\!/_{S_y} \right)$$

En el análisis de regresión simple, el coeficiente de regresión estandarizado correspondiente a la única variable independiente presente en la ecuación coincide exactamente con el coeficiente de correlación de Pearson. En regresión múltiple, según veremos, los coeficientes de regresión estandarizados permiten valorar la importancia relativa de cada variable independiente dentro de la ecuación.

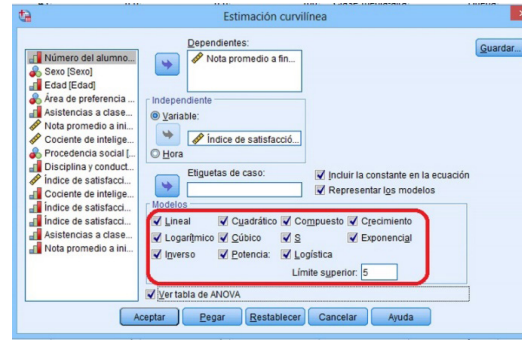
El gráfico de dispersión se adjunta a los resultados del análisis de regresión.



Otras particularidades para obtener la ecuación de regresión son las siguientes:

1. Desde el gráfico de dispersión es posible obtener la ecuación de regresión y el valor de R^2 como puede observarse en la siguiente figura.

2. Desde el mismo cuadro de diálogo que se utilizó para obtener la ecuación de la recta de regresión es posible obtener otras curvas de regresión que puede que se ajusten mejor al conjunto de datos que se estudia; la siguiente figura ilustra tal proceso.

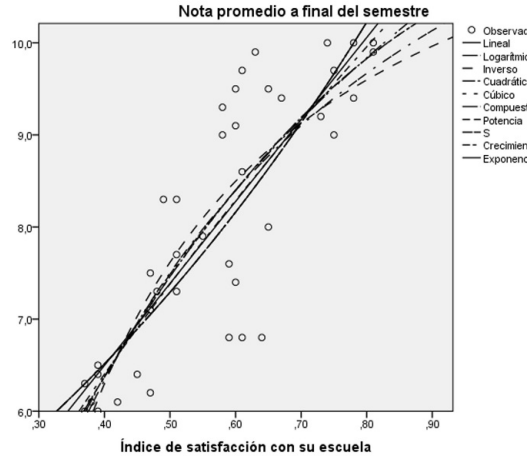


A partir de este cuadro de diálogo se obtiene para las variables con las que se desarrolla la ejemplificación los parámetros de las curvas de ajustes que se seleccionen:

Resumen de modelo y estimaciones de parámetro									
Variable dependiente: Nota promedio a final del semestre									
Ecuación	Resumen del modelo					Estimaciones de parámetro			
	R cuadrado	F	gl1	gl2	Sig.	Constante	b1	b2	b3

Lineal	,697	87,431	1	38	,000	2,936	8,902		
Logarít-mico	,706	91,070	1	38	,000	10,953	5,002		
Inverso	,695	86,727	1	38	,000	12,897	-2,640		
Cuadrá-tico	,706	44,336	2	37	,000	,526	17,611	-7,478	
Cúbico	,706	44,336	2	37	,000	,526	17,611	-7,478	,000
Com-puesto	,695	86,761	1	38	,000	4,151	3,086		
Potencia	,713	94,210	1	38	,000	11,477	,637		
S	,711	93,395	1	38	,000	2,692	-,338		
Creci-miento	,695	86,761	1	38	,000	1,423	1,127		
Expo-nencial	,695	86,761	1	38	,000	4,151	1,127		
La variable independiente es Índice de satisfacción con su escuela.									

De esta tabla se puede inferir cuál es la curva que más se ajusta al conjunto de datos en dependencia del valor que tome R2 en este caso la potencia. La aplicación también devuelve la representación de todas las curvas de ajustes.



3. Se advierte que los coeficientes de correlación no significan "causalidad", es decir, aunque indican asociación entre dos variables, esto no implica que una variable sea la causa de la otra.

Lo planteado en este tema respecto a la correlación se puede resumir en la siguiente tabla:

Coeficiente de correlación	Escala que exige
Lineal de Pearson	Al menos de intervalo
De rangos de Spearman	Al menos Ordinal
De rangos de Kendall	Al menos Ordinal
Otro coeficiente que expresa relación entre variables.	
V de Cramer (Varía entre 0 y 1)	Nominal

Mediante los gráficos de dispersión se dan nuevos elementos al Análisis Exploratorio de Datos, ellos permiten tomar decisiones respecto a curva de regresión más apropiada para representar el comportamiento de los datos que se estudian.

Capítulo IV. Selección de los modelos estadísticos apropiados para demostrar las inferencias realizadas

4.1. Introducción al tema

En las investigaciones cuantitativas se ponen de manifiesto dos métodos íntimamente relacionados:

- *El método experimental*: consiste en observar los hechos o fenómenos en condiciones predeterminadas, para establecer luego las leyes que lo rigen.
- *El método estadístico*: consiste en observar un fenómeno tal como él ocurre, y en tomar una serie de valores que permiten su análisis y de ellos deducir la relación en los distintos valores y las leyes o causas que los han originado

Si bien el experimento requiere una secuencia completa de pasos, tomados de antemano para asegurar que los datos obtenidos sean apropiados y permitan un análisis objetivo que conduzca a deducciones válidas con respecto al problema establecido el método estadístico requiere del análisis de los datos para cumplir con las funciones esenciales de la Estadística:

- Determinar la fuerza de asociación o correlación entre variables. Inferencia causal, que dé cuentas sobre por qué las cosas son así y no de otra manera.
- La generalización y objetivación de los resultados a través de una muestra para hacer inferencia hacia una población.

En epígrafes anteriores se trató la correlación que indica la fuerza y la dirección de una relación lineal entre dos variables aleatorias y según el tipo de variable se pueden utilizar diferentes coeficientes de correlación:

La segunda función de la estadística, es la inferencia será el tópico a tratar en este epígrafe.

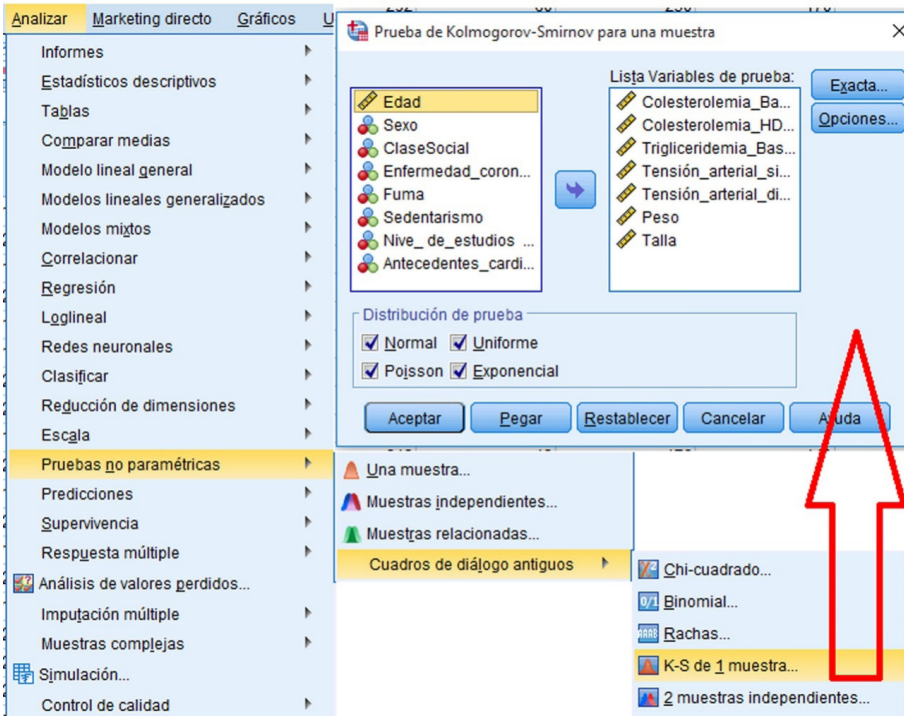
La inferencia estadística tiene como objetivo fundamental, a partir de la observación de una muestra tomada de forma aleatoria, extraer conclusiones válidas que puedan ser generalizables a la población de origen. Para su estudio se asume como nivel de partida que el lector tiene conocimientos elementales de probabilidades y sobre los elemente teóricos de las dócimas de hipótesis, esto da la posibilidad de poder utilizar una terminología comprensible a todo el que utilice este texto y no permite alejar el objetivo con que ha sido concebido.

Entre los criterios establecidos por teoría para comenzar cualquier estudio estadístico está el supuesto de que el conjunto de los datos sigue una distribución normal, y aunque en ocasiones los investigadores soslayan esta condición, ella es fundamental en cualquier estudio estadístico. Como ya se precisó al estudiar la opción Explorar del menú Estadísticos descriptivos, existen dos gráficos que por su comportamiento es posible determinar si los datos siguen o no una distribución normal, son los diagramas Q-Q normal y Q-Q normal sin tendencia y dos pruebas, la Shapiro-Wilk y la Kolmogorov-Smirnov (Lilliefors), la primera se emplea cuando el tamaño de la muestra es pequeño (n menor de 50) y la segunda para el caso de muestras grandes (n mayor de 50).

Prueba de Shapiro-Wilk: contrasta la hipótesis de que una muestra (pequeña) procede de una población normal.

Kolmogorov-Smirnov (Lilliefors): la prueba de Lilliefors es una modificación de la prueba de Kolmogorov-Smirnov, contrasta la normalidad cuando las medias y las varianzas no son conocidas, sino que deben ser estimadas a partir de los datos.

Aunque ya se explicó cómo determinar si lo datos de una muestra siguen o no una distribución normal utilizando la opción explorar, ahora se ejemplificará desde las opciones que se muestra desde el menú de Pruebas no paramétricas, tomando es este caso una base de datos de 70 pacientes hipertensos (Anexo 3).



En este caso no solo se indaga si lo datos siguen una distribución normal, también se incluyen otras distribuciones como la Uniforme, la distribución de Poisson[22] y la distribución Exponencial; en todos los casos la hipótesis nula es que los datos siguen la distribución seleccionada y por tanto si el valor de probabilidad asociado al estadígrafo es de p <= 0,05, es indicador de que no hay diferencia significativa, es decir, tomamos la hipótesis alternativa que es donde se plantea la diferencia, lo que se traduce, en los datos, que no siguen la distribución seleccionada.

Esta opción devuelve los siguientes resultados:

Prueba de Kolmogorov-Smirnov para una muestra

		Colesterolemia_Basal	Colesterolemia_HDL_Basal	Trigliceridemia_Basal	Tensión_arterial_sistólica	Tensión_arterial_diastólica	Peso	Talla
N		70	70	67	70	70	68	70
Parámetros normalesa,b	Media	241,06	42,27	138,97	144,40	81,74	70,01	166,57
	Desviación estándar	54,278	7,776	37,865	39,742	11,228	12,048	8,732
Máximas diferencias extremas	Absoluta	,171	,092	,164	,215	,154	,103	,090
	Positivo	,165	,092	,164	,187	,154	,103	,090
	Negativo	-,171	-,059	-,125	-,215	-,110	-,082	-,070
Estadístico de prueba		,171	,092	,164	,215	,154	,103	,090
Sig. asintótica (bilateral)		,000c	,200c,d	,000c	,000c	,000c	,068c	,200c,d

a. La distribución de prueba es normal.

b. Se calcula a partir de datos.

c. Corrección de significación de Lilliefors.
d. Esto es un límite inferior de la significación verdadera.

Se ha marcado en rojo los valores asintóticos que indican que los datos correspondientes siguen una distribución normal.

Prueba de Kolmogorov-Smirnov de una muestra 4								
		Colesterolemia_Basal	Colesterolemia_HDL_Basal	Trigliceridemia_Basal	Tensión_arterial_sistólica	Tensión_arterial_diastólica	Peso	Talla
N		70	70	67	70	70	68	70
Parámetro exponencial.a,b	Media	241,06	42,27	138,97	144,40	81,74	70,01	166,57
Máximas diferencias extremas	Absoluta	,517	,491	,491	,535	,548	,510	,594
	Positivo	,256	,242	,172	,263	,277	,247	,320
	Negativo	-,517	-,491	-,491	-,535	-,548	-,510	-,594
Z de Kolmogorov-Smirnov		4,328	4,109	4,019	4,474	4,589	4,209	4,967
Sig. asintótica (bilateral)		,000	,000	,000	,000	,000	,000	,000
a. La distribución de prueba es exponencial.								

b. Se calcula a partir de datos.

Como en todos los casos el valor de probabilidad asociado al estadígrafo p = 0,000, es indicador de que no hay diferencia significativa, es decir, se debe tomar la hipótesis alternativa, los datos no siguen una distribución exponencial.

Prueba de Kolmogorov-Smirnov de una muestra 2								
		Colesterolemia_Basal	Colesterolemia_HDL_Basal	Trigliceridemia_Basal	Tensión_arterial_sistólica	Tensión_arterial_diastólica	Peso	Talla
N		70	70	67	70	70	68	70
Parámetros uniformesa,b	Mínimo	175	26	87	110	65	50	150
	Máximo	590	60	245	430	105	98	190
Máximas diferencias extremas	Absoluta	,648	,182	,372	,753	,211	,225	,261
	Positivo	,648	,182	,372	,753	,211	,225	,261
	Negativo	-,014	-,123	-,055	-,014	-,039	-,039	-,082
Z de Kolmogorov-Smirnov		5,425	1,519	3,041	6,297	1,763	1,859	2,181

Sig. asintótica (bilateral)	,000	,020	,000	,000	,004	,002	,000

a. La distribución de prueba es uniforme.

b. Se calcula a partir de datos.

Los que no siguen una distribución uniforme por razones ya expresada.

Prueba de Kolmogorov-Smirnov de una muestra 3								
		Colesterolemia_ Basal	Colesterolemia_ HDL_ Basal	Trigliceridemia_ Basal	Tensión_ arterial_sistólica	Tensión_arterial_ diastólica	Peso	Talla
N		70	70	67	70	70	68	70
Parámetro de Poissona,b	Media	241,06	42,27	138,97	144,40	81,74	70,01	166,57
Máximas diferencias extremas	Absoluta	,343	,086	,317	,320	,181	,138	,128
	Positivo	,343	,086	,317	,320	,181	,138	,128
	Negativo	-,197	-,081	-,177	-,151	-,147	-,134	-,126
Z de Kolmogorov-Smirnov		2,870	,718	2,596	2,679	1,511	1,141	1,072
Sig. asintótica (bilateral)		,000	,681	,000	,000	,021	,148	,201

a. La distribución de prueba es Poisson.

b. Se calcula a partir de datos.

Las marcas en rojo de los valores asintóticos indican que los datos correspondientes siguen una distribución de Poisson.

4.2. ¿Cómo desarrollar el análisis inferencial?

Cuando ya se han identificado las características de los datos es posible comenzar a proyectar un procesamiento inferencial formulando hipótesis, pero ¿qué es una hipótesis?

En la literatura se pueden varias definiciones, pero desde el punto de vista formal, una hipótesis es una conjetura o suposición que se expresa en forma de un enunciado afirmativo. Las hipótesis constituyen uno de los métodos fundamentales de la investigación científica.

Pero cualquiera sea la definición que se elija, se puede asegurar con absoluta certeza que toda hipótesis que se formule requiere ser sometida a verificación, para determinar si puede ser aceptada o si, por el contrario, debe ser rechazada. De modo particular se someten a tales verificaciones las hipótesis estadísticas.

Las hipótesis estadísticas: son suposiciones sobre la población o las poblaciones que se investigan. Estas suposiciones pueden estar referidas a:

- La comparación de cada parámetro de la población con un valor específico, o también, comparación de dos o más universos, en cuanto a un parámetro determinado.
- La distribución de probabilidad que sigue cada una de las variables que se investigan en esas poblaciones.
- La independencia entre dos o más variables que se miden en un universo.

Ejemplo de hipótesis estadística:

a. Un profesor de Matemática sostiene que, después de haber aplicado un nuevo método de enseñanza, las calificaciones de sus alumnos tienen una dispersión diferente de 0.8 puntos.

b. Un especialista en mercadotecnia plantea que después de haber aplicado determinada la estrategia de mercadeo los resultados de venta de una empresa A es superior a los de otra empresa B donde tal estrategia no se ha empleado.

c. El mismo especialista sostiene que las ventas diarias durante un mes de las mercancías promocionadas en la empresa A se distribuyen normalmente.

d. Un economista plantea que las ventas por departamentos en ambas empresas dependen de los años de experiencias en el ramo de los empleados, es decir, los años de experiencia y las ventas por departamento son variables dependientes.

En cada uno de estos ejemplos, se han realizado conjeturas sobre una o más poblaciones, por lo que se está ante hipótesis estadísticas. En los incisos a y b, la hipótesis está referida a parámetros del universo: en el caso a, se compara la varianza de una población con un valor específico (0.8); mientras que, en b, se comparan dos poblaciones en cuanto a la media de estas. Por su parte, en el inciso c, se ha formulado una hipótesis referida a la distribución de probabilidad de una variable que se investiga, y en el inciso c, es evidente que se refiere a la independencia de dos variables.

Pero para realizar el estudio de las hipótesis estadísticas, estas se plantean de modo simbólico, mediante la utilización de los operadores aritméticos igual a (=), desigual a (≠), menor que (<), mayor que (>), y en ocasiones, menor o igual a (≤) y mayor o igual a (≥), entre otros. De acuerdo con el operador que se utilice, las hipótesis estadísticas se clasifican en nulas y alternativas.

Hipótesis nula: es una hipótesis estadística que contiene la igualdad. Generalmente, en su formulación simbólica se emplea el operador igual a (=), aunque en ocasiones también se utilizan los operadores menores o igual a (≤) o mayor o igual a (≥).

Cuando se utiliza el signo de igualdad (=), la hipótesis nula se dice que es sim-ple; en cambio, cuando se emplea el operador menor o igual a (≤) o el mayor o igual a (≥), la hipótesis nula se denomina compuesta.

Para denotar la hipótesis nula se utiliza el símbolo H0. Esta hipótesis, por lo general, se formula con el objetivo de no aceptarla como verdadera.

Para un parámetro Θ cualquiera, que se compara con un valor determinado Θ0, simbólicamente escribiremos: H0: Θ= Θ0.

Dado que la diferencia, según la formulación anterior, entre Θ y Θ0 es cero "nula", es decir, Θ - Θ0= 0, es por lo que a la hipótesis H0 se le denomina nula o de nulidad.

Hipótesis alternativa: es una hipótesis estadística distinta de la de nulidad que se ha planteado contiene la desigualdad. Generalmente, en su formulación simbólica se emplean los operadores desiguales a (≠), menor que (<) y mayor que (>).

Para denotar la hipótesis nula se emplea el símbolo H1. Esta hipótesis, por lo general, se formula con el objetivo de aceptarla como verdadera. A H1 también se le denomina la hipótesis del investigador, ya que en ella se plasma la pre-dicción científica de este.

Las hipótesis nula y alternativa son "complementarias", si se tiene en cuenta que, cuando H0 no se acepta como verdadera, entonces H1 lo será y viceversa.

Para la hipótesis de nulidad planteada, existen tres alternativas: si para un parámetro Θ cualquiera, que se compara con un valor determinado Θ0, se tiene la hipótesis nula: H0: Θ=Θ0, entonces para esta, se puede plantear las alternativas siguientes:

I. H1: Θ≠Θ0

II. H1: Θ>Θ0

III. H1: Θ<Θ0

En la práctica, de estas tres alternativas, el investigador trabaja con una y solo con una de ellas: con aquella que contiene su predicción científica.

Cuando en la hipótesis alternativa se emplea el signo ≠, esta se denomina bilateral (caso I); mientras que, si los signos son > o < la hipótesis se denomina unilateral: específicamente si el operador es >, la hipótesis se llama unilateral a la derecha (caso II) y si es el <, se clasifica de unilateral a la izquierda (caso III).

Prueba de una hipótesis estadística: después de haber planteado las hipótesis nula y la alternativa, el interés estará en determinar cuál de ellas es la que no se rechaza, es decir, la que se toma como verdadera: se podrá pensar en que ello se responde estudiando todos los elementos del universo, y si por ejemplo, se somete a verificación la varianza de la población, se calcula esta para ese universo y se compara con el valor hipotético; si ambos resultados son iguales, se aceptará la hipótesis nula, y en caso contrario, se rechazará. Sin embargo, este procedimiento no es aplicable en la práctica, por la sencilla razón de que, por lo general, no es posible trabajar con los datos del universo.

Para tomar la decisión sobre cuál hipótesis no será rechazada, lo que se hace es estudiar una muestra aleatoria del universo que se investiga, y sobre la base de esa muestra, llegar a adoptar una decisión. Para esto, se toma en cuenta la distribución muestral del estadígrafo, y sobre la base de una probabilidad prefijada por el investigador, se toma la decisión.

Prueba de hipótesis: es el procedimiento que se utiliza, para comprobar si una hipótesis de nulidad planteada se rechaza o no, sobre la base de la información poblacional contenida en la muestra. A las pruebas de hipótesis también se les denominan dócimas de hipótesis o tests.

Errores de decisión: Al realizar una prueba de hipótesis, la decisión de rechazar o no a H0, se toma sobre la base de la muestra aleatoria que se ha seleccionado, como en dicha muestra no están incluidos todos los elementos del universo, existe la posibilidad de que se adopte una decisión equivocada; en realidad, se puede cometer uno de los dos errores siguientes:

Error de tipo I: consiste en rechazar una hipótesis de nulidad que es verdadera, o lo que es lo mismo, tomar como falsa una hipótesis nula que es cierta.

Error de tipo II: consiste en no rechazar una hipótesis de nulidad que es falsa, o lo que es lo mismo, tomar como verdadera una hipótesis nula que es falsa.

En resumen, ante una hipótesis nula planteada, se dará una de las situaciones que se recogen en el siguiente cuadro:

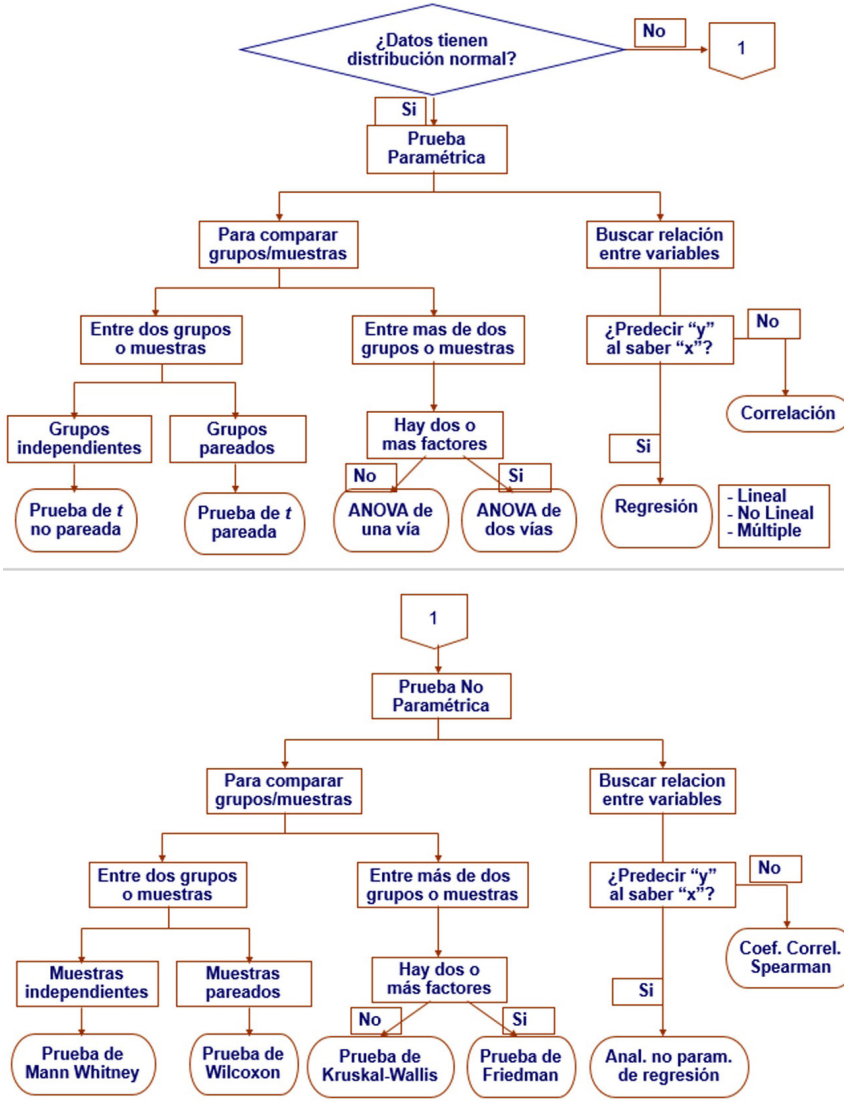
	Decisión	
Hipótesis H0	No rechazar H0	Rechazar H0
Verdadera	Correcta 1-α	Error de tipo I α
Falsa	Error de tipo II β	Correcta 1-β

4.3. Procedimiento que por lo común se sigue, en una prueba de hipótesis

1. Formulación de la hipótesis de nulidad (Ho).
2. Elección de una prueba estadística (con su modelo estadístico asociado) para probar Ho. De las pruebas capaces de usarse en un diseño de investigación dado, hay que escoger aquella cuyo método se aproxima más a las condiciones de la investigación (en término de los supuestos que califican el uso de la prueba) y cuyos requisitos de medición satisfacen las medidas usadas en la investigación.
3. Especificación del nivel de significancia () y del tamaño de la muestra (N).
4. Encuentro (o suposición) de la distribución muestra) de la prueba estadística conforme a Ho.
5. Sobre la base de los incisos 1, 2, 3 y 4, definición de la región de rechazo.
6. Cálculo del valor de la prueba estadística con los datos obtenidos de la(s) muestra(s). Si el valor desciende a la región de rechazo Ho debe rechazarse; si el valor cae fuera de la

región de rechazo, Ho no puede rechazarse al nivel de significación escogido.

El siguiente algoritmo resume los métodos y pruebas estadísticas a utilizar:



4.4. Pruebas paramétricas

En estadística, un parámetro es un número que resume la gran cantidad de datos que pueden derivarse del estudio de una variable estadística. La necesidad de realizar tales resúmenes se trató en el epígrafe titulado ¿Cómo resumir numéricamente los datos almacenados con SPSS?

El cálculo de estos números está bien definido, generalmente mediante fórmulas matemáticas obtenidas aplicando sus leyes a partir de datos de la población y son la base de uno de los propósitos esenciales de la estadística: crear un modelo de la realidad que permita tener una idea global de la población, compararla con otras, comprobar su ajuste a un modelo ideal, realizar estimaciones sobre datos desconocidos de la misma y, en definitiva, tomar decisiones y es a estas tareas a lo que contribuyen de modo esencial los parámetros estadísticos.

A estos modelos se refieren los estudios estadísticos cuando se habla de una distribución normal de parámetros μ y σ o de otras distribuciones de probabilidad definidas mediante parámetros como la media, la varianza, la curtosis; ellos permiten desarrollar las pruebas de hipótesis que se desarrollarán en este epígrafe.

4.5. Para probar la Media contra un valor hipotético

Esta prueba consiste en determinar si la media de la muestra que se está utilizando en el estudio puede ser comparada contra un valor que se conoce, puede ser por bueno como unidad de medida en ese entorno, por diversas razones.

Ejemplo: se desea saber si los datos que se estudian en un hospital difieren significativamente de las normas de la OPS o la OMS.

El algoritmo a seguir para localizar la opción deseada es:

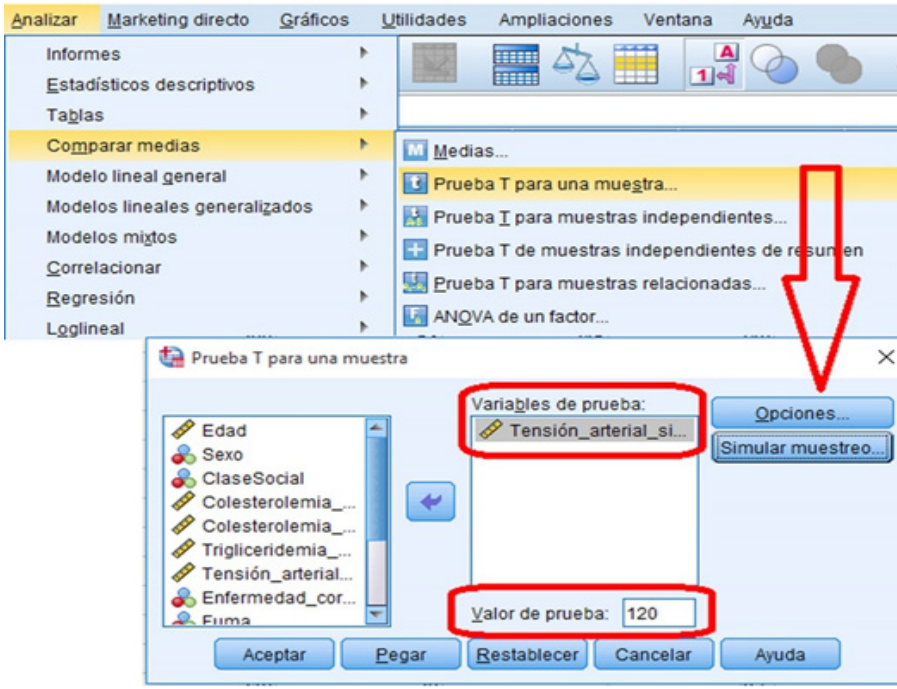
Analizar → Comparación de Media → Prueba t para una muestra.

Problema ejemplo: Comprobar si la Tensión sistólica basal de la

base de datos del ejemplo anterior difiere del valor que históricamente se ha dicho que es bueno para que una persona sea considerada sana bien, es decir, 120 mm/hg.

En este caso la hipótesis H_0 es que la media de Tensión sistólica de la muestra no difiere de 120, es decir mientras Se advierte que esta igualdad de expresa que está significativamente "cercana" a 120 con un error α= 0,05.

En la imagen adjunta se muestra el proceso a seguir y el resultado se da en las siguientes tablas:



Estadísticas de muestra única				
	N	Media	Desviación estándar	Media de error estándar
Tensiónarterialsistólica	70	144,40	39,742	4,750

Prueba de muestra única						
	Valor de prueba = 120					
	t	gl	Sig. (bilateral)	Diferencia de medias	95% de intervalo de confianza de la diferencia	
					Inferior	Superior
Tensiónarterialsistólica	5,137	69	,000	24,400	14,92	33,88

El investigador puede llegar al resultado a través de dos vías:

1. Observando que el valor probabilístico asociado al estadígrafo, en este caso p = 0,000 donde se acepta la hipótesis alternativa que es donde se platea que hay diferencia entre los valores de la muestra y el valor de prueba.
2. Constatando si el valor correspondiente a la diferencia de medias está dentro del intervalo que se forma entre el límite inferior y el límite superior, para el ejemplo se puede apreciar que 24,400 pertenece al intervalo [14,92 ; 33,88] y se llega a la misma conclusión, como es lógico, de aceptar la hipótesis alternativa y plantear que hay diferencia significativa, es de suma importancia aclarar con qué nivel de confianza, aquí es del 95 %, lo que indica dice que de cada 100 muestras de tamaño 70 que se tomen, en la población objeto de estudio, en al menos 95 de ellas se obtendrá el mismo resultado.

4.6. Prueba para dos muestras relacionadas

El caso más clásico de dos muestras relacionadas es cuando a un mismo sujeto se le hace una medición antes y otra después, producto de una intervención.

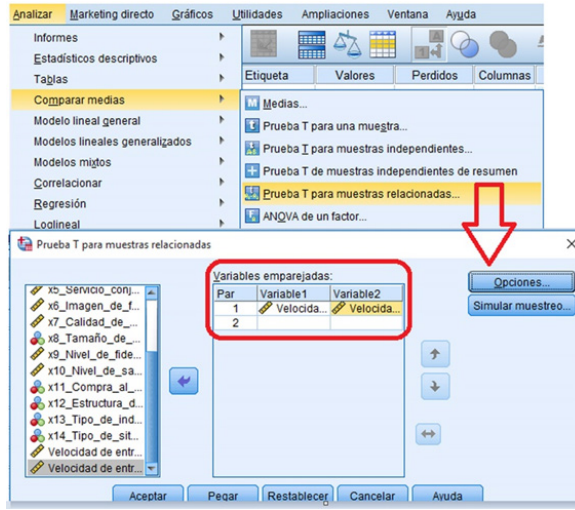
Problema ejemplo: En una empresa se realizan entrega a domicilio de los productos que en la misma se procesan; la dirección decide hacer cambios organizativos en el sistema de despacho y transportación y para comprobar si se disminuye el tiempo de entrega, desarrolla un control aleatorio de 100 servicios antes de modificar el sistema y al mes de instaurada la modificación vuelve a comprobar a 100 servicios tomados también al azar. Haga la prueba estadística correspondiente que permita determinar si hubo cambio favorable o no.

En este caso H_0 plantea que no hay cambios, mientras H_1 apuesta porque si lo hay y que son favorable. Como la media será utilizado como estadígrafo de prueba se tiene que:

$$H_0: \bar{x}_{antes} = \bar{x}_{después}$$

$$H_1: \bar{x}_{antes} > \bar{x}_{después}$$

El esquema a seguir se muestra en la siguiente imagen adjunta.



Como se puede apreciar en la imagen, en el cuadro de Prueba T para muestras relacionadas aparecen las dos variables que se van a comparar, es bueno destacar que hasta que no se marquen las dos variables no se activa el botón de aceptar y en opciones se define el nivel de confianza con que se desea trabajar, y por último se pulsa aceptar para obtener los siguientes resultados:

Estadísticas de muestras emparejadas

		Media	N	Desviación estándar	Media de error estándar
Par 1	Velocidad de entrega antes	3,5150	100	1,32073	,13207
	Velocidad de entrega después	3,3920	100	1,40509	,14051

Correlaciones de muestras emparejadas

		N	Correlación	Sig.
Par 1	Velocidad de entrega antes & Velocidad de entrega después	100	,923	,000

Prueba de muestras emparejadas

	Diferencias emparejadas					t	gl	Sig. (bilateral)
Media	Desviación estándar	Media de error estándar	95% de intervalo de confianza de la diferencia					
			Inferior	Superior				

Par 1	Velocidad de entrega antes - Velocidad de entrega después	,12300	,54008	,05401	,01584	,23016	2,277	99	,025

En este resultado la probabilidad asociada al estadígrafo es de p = 0,025 < 0,05, por lo que se puede decir que existen diferencias significativas, para un 95% de confianza, entre la "Velocidad de entrega antes" y la "Velocidad de entrega después"; obsérvese que este estadígrafo está referido a la diferencia de las medias (0,12300) y como esta diferencia es positiva y se ha tomado.

Velocidad de entrega antes - Velocidad de entrega después

Esto indica que la velocidad de entrega después es inferior y por tanto se satisfacen las expectativas de los directivos.

4.7. Prueba para dos muestras no relacionadas

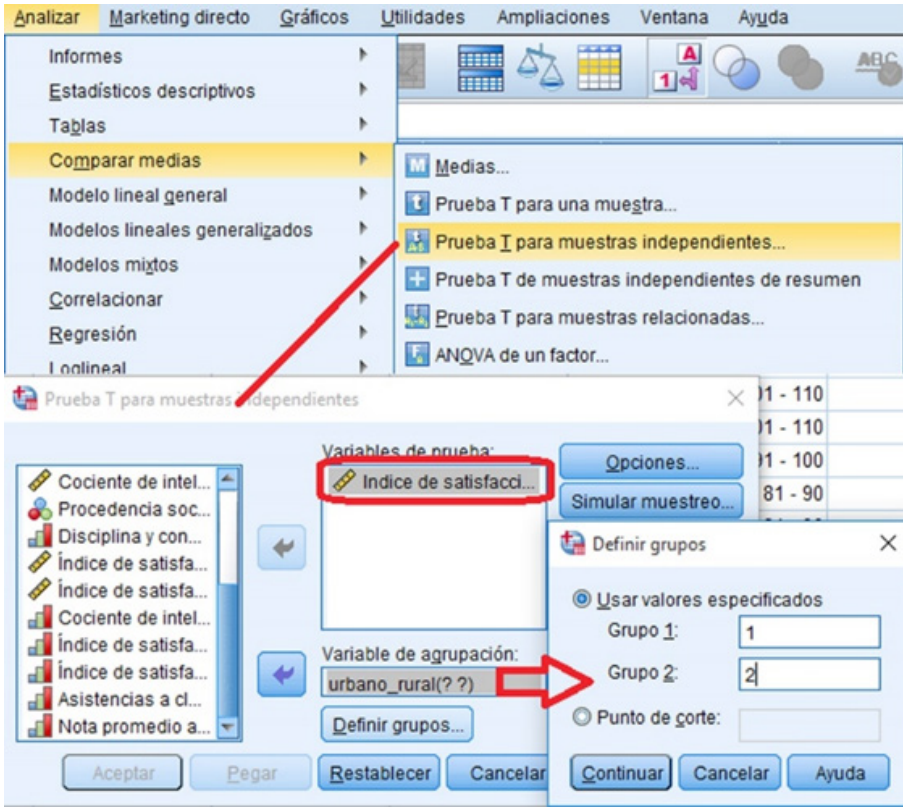
En lo esencial la prueba para dos muestras no relacionadas no difiere dela de dos muestras pareadas, en ambas es propósito del investigador determinar si entre las medias de ambas poblaciones hay diferencias significativas, lo que equivale a determinar si las diferencias de las medias es o no significativa, pero difieren entre otros aspectos en que los eventos que han generado los datos correspondientes a cada muestra no mantienen vínculos de relación o precedencia y que estas relaciones son tan independientes que hasta el tamaño de las muestras pueden ser distintas.

Problema ejemplo: Un sociólogo elabora un test que le permite determinar un índice de satisfacción con la escuela de cada estudiante y lo aplica a una muestra aleatoria de 228 alumnos urbanos y 137 de escuelas rurales. Se desea investigar si hay diferencias significativas entre las satisfacciones por sus escuelas de los alumnos urbanos y los rurales.

Evidentemente que H_0 expresa que no hay diferencia ni para las medias ni para las varianzas de las preferencias, frente a H_1 que expresa lo contrario que hay diferencias significativas en las medias, o en las varianzas, o en ambas.

La diferencia esencial está en la forma en la que hay que organizar los datos, para ello debe seguir el siguiente algoritmo:

1. En la base de datos se dispone de dos variables; dos columnas donde aparecen los índices de satisfacción por la escuela de los alumnos urbanos y los alumnos de áreas rurales.
2. Construya una nueva variable en la que una los datos de ambos grupos de alumnos, es decir, los índices de los alumnos de área urbana y a continuación en la misma columna la de los alumnos de áreas rurales o viceversa.
3. Construya otra nueva variable que asigne un dígito a los alumnos de área urbana y otro a los alumnos de área rural, ejemplo, identifique con 1 a los de área urbana y 2 a los de área rural. ESTE NÚMERO ES CLAVE, con él se puede diferenciar una variable de otra.
4. El procedimiento en SPSS se desarrolla siguiendo los pasos que se muestra en la siguiente imagen y los resultados se dan en la tabla que le sigue:



Estadísticas de grupo					
	Procedencia urbana o rural	N	Media	Desviación estándar	Media de error estándar
Índice de satisfacción por la escuela	De área urbana	228	80,8925	12,12439	,80296
	De área rural	137	80,0691	11,96730	1,02244

Observe que entre las medias y entre las desviaciones hay poca diferencia, esto se corrobora en la siguiente tabla y trae consecuencias para la interpretación de los resultados de la prueba.

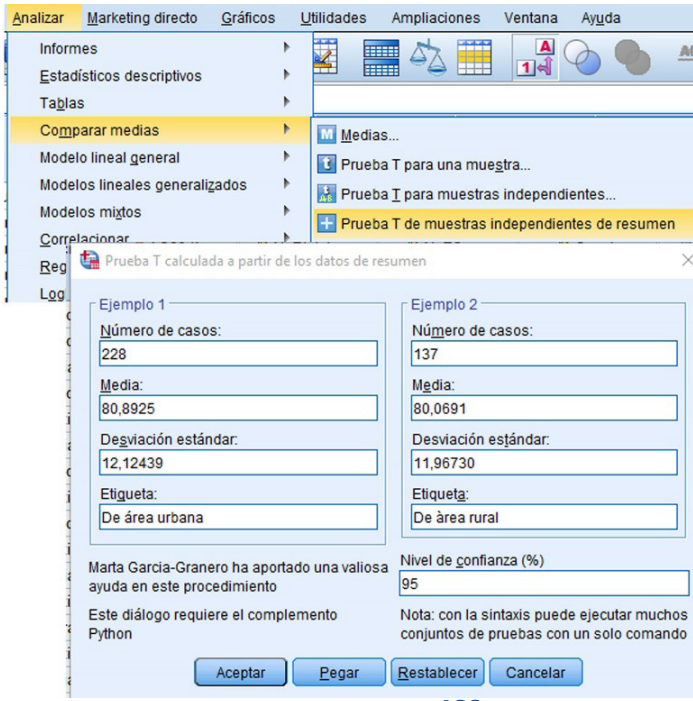
Prueba de muestras independientes										
Prueba de Levene de igualdad de varianzas				prueba t para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	95% de intervalo de confianza de la diferencia	
								Inferior	Superior	
Índice de satisfacción por la escuela	Se asumen varianzas iguales	,093	,760	,631	363	,528	,82341	1,30429	-1,74151	3,38832
	No se asumen varianzas iguales			,633	289,511	,527	,82341	1,30004	-1,73533	3,38214

Esta prueba da el resultado de dos test, uno para ver cómo se comportan las varianzas y otro para ver si existe diferencia entre ambos resultados. En la prueba de Levene para la igualdad de varianzas la probabilidad asociada al estadígrafo fue de p = 0,760 (en rojo), como este valor es superior a 0, 05 que fue el nivel de significación prefijado, se puede decir que *no existen diferencias significativas entre las dos variables estudiadas en cuanto a sus varianzas y por tanto se tiene que asumir que las varianzas son iguales* (marcado en rojo) que es lo que plantea la hipótesis de nulidad, es por ello que cuando se desarrolla el analice de la prueba T para la igualdad de medias de las dos probabilidades asociadas que se muestra en la tabla de resul-

tado se ha de tomar es el que se encuentra arriba,(en rojo) en este caso particular el valor, de las probabilidades, es similar pero eso no implica que siempre sea así, incluso, puede dar un resultado con uno distinto al del otro, este valor, para el caso estudiado, es de p = 0,528, por lo que se puede concluir, con un nivel de confianza del 95% que tampoco existen diferencias significativas entre las medias de las dos variables analizadas. Como conclusión, para estas muestras no hay diferencias significativas en el Índice de satisfacción por la escuela entre estudiantes de zona urbana y zona rural.

SPSS tiene otra opción para pruebas de muestras independientes, pero en ella el usuario debe introducir los parámetros de media, desviación y número de integrantes de la muestra. En la figura se muestra el proceso a seguir con los mismos datos del problema resuelto.

Los resultados del procesamiento son los siguientes:



Estadísticas de grupo					
	Procedencia urbana o rural	N	Media	Desviación estándar	Media de error estándar
Índice de satisfacción por la escuela	De área urbana	228	80,8925	12,12439	,80296
	De área rural	137	80,0691	11,96730	1,02244

Prueba de muestras independientes									
	Prueba de Levene de igualdad de varianzas		prueba t para la igualdad de medias						
							95% de intervalo de confianza de la diferencia		
F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	Inferior	Superior	

Índice de satisfacción por la escuela	Se asumen varianzas iguales	,093	,760	,631	363	,528	,82341	1,30429	-1,74151	3,38832
	No se asumen varianzas iguales			,633	289,511	,527	,82341	1,30004	-1,73533	3,38214

Obsérvese la coincidencia de los resultados con los alcanzados por el procedimiento antes descrito, esta vía es más sencilla y la información puede tomarse de otros procedimientos de estadística descriptiva explicados en epígrafes anteriores.

4.8. Análisis de Varianza de un solo factor o ANOVA

El análisis de la varianza (ANOVA, ANalysis Of VAriance, según terminología inglesa) es una colección de modelos estadísticos y sus procedimientos asociados, en los que la varianza está particionada en ciertos componentes debidos a diferentes variables explicativas.

Las técnicas iniciales del análisis de varianza fueron desarrolladas por el estadístico y genetista R. A. Fisher[xxiii] en los años 1920 y 1930, por lo que también se conoce como “Anova de Fisher” o “análisis de varianza de Fisher”, debido al uso de la distribución F de Fisher como parte del contraste de hipótesis.

El análisis de varianza (ANOVA) de un factor sirve para comparar varios grupos en una variable cuantitativa. Se trata, por tanto, de una generalización de la Prueba T para dos muestras independientes al caso de diseños con más de dos muestras.

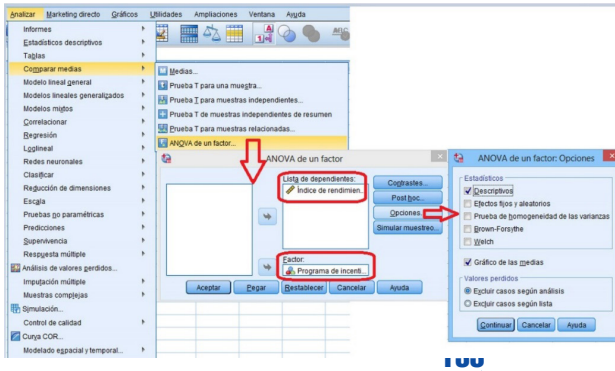
A la variable categórica (nominal u ordinal) que define los grupos que se desean comparar se les llama independiente o factor y se representa por VI. A la variable cuantitativa (de intervalo o razón) en la que se desean comparar los grupos se les llama dependiente y se representa por VD.

Ejemplo, si se quiere averiguar cuál de tres programas distintos de incentivos aumenta de forma más eficaz el rendimiento de un determinado colectivo, se deben seleccionar tres muestras aleatorias de ese colectivo y aplicar a cada una de ellas uno de los tres programas.

Después, se mide el rendimiento de cada grupo y se averigua si existen o no diferencias entre ellos. En este caso se tiene una VI categórica (el tipo de programa de incentivos) cuyos niveles se desean comparar entre sí, y una VD cuantitativa (la medida del rendimiento, en el caso que se ejemplificará será un índice que toma valores entre 0 y 1), en la cual se quieren comparar los tres programas. El ANOVA de un factor permite obtener información sobre el resultado de esa comparación. Es decir, permite concluir si los sujetos sometidos a distintos programas difieren la medida de rendimiento utilizada.

La hipótesis que se pone a prueba en el ANOVA de un factor es que las medias poblacionales (las medias de la VD en cada nivel de la VI) son iguales. Si las medias poblacionales son iguales, eso significa que los grupos no difieren en la VD y que, en consecuencia, la VI o factor es independiente de la VD.

Para el ejemplo planteado su solución mediante SPSS es la siguiente:



En el cuadro de diálogos se ha seleccionado los estadísticos descriptivos y el gráfico de medias por las informaciones que ofrecen:

Descriptivos								
Índice de rendimiento								
	N	Media	Desviación estándar	Error estándar	95% del intervalo de confianza para la media		Mínimo	Máximo
					Límite inferior	Límite superior		
Incentivo 1	47	,58111	,232733	,033948	,51277	,64944	,107	,966
Incentivo 2	61	,62705	,242579	,031059	,56492	,68918	,200	,995
Incentivo 3	49	,69492	,193873	,027696	,63923	,75061	,323	,993
Total	157	,63448	,228466	,018234	,59846	,67049	,107	,995

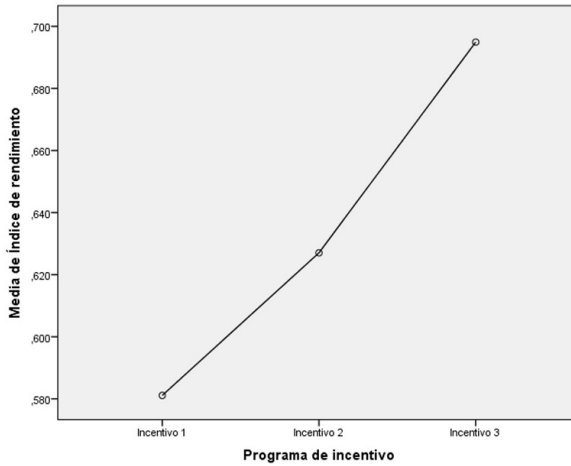
El análisis descriptivo puede dar una primera idea del comportamiento de los datos que debe verificarse con el ANOVA.

ANOVA					
Índice de rendimiento					
	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Entre grupos	,316	2	,158	3,111	,047
Dentro de grupos	7,826	154	,051		

Total	8,143	156			

Como Sig = 0,047< 0,05 se rechaza la hipótesis nula de que las medias poblacionales con iguales, es decir los grupos difieren en la VD y en consecuencia, la VI o factor es independiente de la VD.

El siguiente gráfico de las medias ilustra la referida diferencia.



Nota: Si se desea saber cuáles son los grupos que están afectando la homogeneidad puede activar el botón Post hoc. que aparece en el cuadro de diálogo de la imagen anterior y activar en el nuevo cuadro de diálogo los comandos de las pruebas de HSD de Tukey y/o la de Duncan

4.9. Pruebas no paramétricas

¿Qué es la Estadística no paramétrica?

La Estadística no paramétrica es una rama de la Estadística que estudia las pruebas y modelos estadísticos cuya distribución subyacente no se ajusta a los llamados criterios paramétricos. Su distribución no puede ser definida a priori, pues son los datos observados los que la determinan. La utilización de estos métodos se hace recomendable cuando no se puede asumir que los datos se ajusten a una distribución normal o cuando el nivel de medida empleado no sea, como mínimo, de intervalo.

La estadística paramétrica analiza modelos estadísticos que implican distribuciones continuas con ciertos supuestos básicos para la aplicación de las técnicas a emplear. El principal uso de esos modelos es la estimación de parámetros desconocidos de la población en estudio, para poder hacer pruebas de validación o ensayos de significación y testear así las hipótesis planteadas.

Mientras los supuestos usados en la estadística paramétrica especifican la distribución original (generalmente la normal), hay otros casos en la práctica donde no se puede hacer esto, donde no se puede especificar la forma de distribución original. Se requiere entonces otra metodología de trabajo, una estadística de distribuciones libres, donde no se necesitan hacer supuestos acerca de la distribución poblacional, donde se puede comparar distribuciones entre sí o verificar supuestos a cerca de la forma de la población. Por ejemplo, verificar el supuesto de normalidad necesario para usar el modelo t de Student.

4.10. Ventajas de las pruebas no paramétricas sobre las pruebas paramétricas

1. Por lo general, son fáciles de usar y entender.
2. Eliminan la necesidad de suposiciones restrictivas de las pruebas paramétricas.
3. Se pueden usar con muestras pequeñas.
4. Se pueden usar con datos cualitativos.
5. Se pueden estudiar casos donde no es posible precisar la naturaleza de la distribución.
6. Se pueden estudiar casos donde los supuestos de la forma poblacional son débiles.
7. Es posible aplicar el mismo modelo a casi todas las distribuciones en lugar a una sola.

4.11. Desventajas de las pruebas no paramétricas respecto a las pruebas paramétricas

1. A veces, ignoran, desperdician o pierden información.
2. No son tan eficientes como las paramétricas.

3. Llevan a una mayor probabilidad de no rechazar una hipótesis nula falsa (incurriendo en un error).
4. Las pruebas no paramétricas son pruebas estadísticas que no hacen suposiciones sobre la constitución de los datos de la población.
5. Por lo general, las pruebas paramétricas son más poderosas que las pruebas no paramétricas y deben usarse siempre que sea posible.
6. Es importante observar que, aunque las pruebas no paramétricas no hacen suposiciones sobre la distribución de la población que se muestrea, muchas veces se apoyan en distribuciones muestrales como la normal o la ji cuadrada.
7. Tienen cálculos usualmente más engorrosos.
8. No extraen tanta información como los paramétricos si se aplican al mismo caso.
9. Son menos eficientes si las muestras son grandes.

4.12. Análisis para el caso de una muestra

Prueba binomial

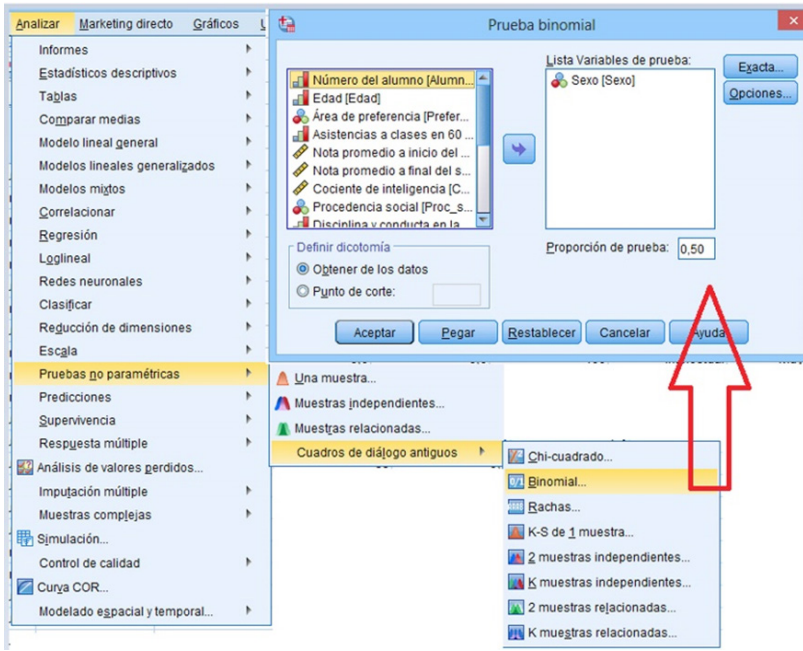
Analiza variables dicotómicas y compara las frecuencias observadas en cada categoría con las que cabría esperar según una distribución binomial de parámetro especificado en la hipótesis nula atendiendo a la siguiente expresión:

$$p(x) = \begin{cases} \binom{n}{x} p^x \; q^{n-x} \; para \; x = 0,1,2,\dots,n \\ 0 \;\; p.o.v \end{cases}$$

El nivel de significación crítico de esta prueba indica la probabilidad de obtener una discrepancia igual o superior a la observada a partir de la muestra si la distribución es la postulada por la hipótesis nula.

En base de datos PROBLEMA_BASE se puede verificar si hay diferencias significativas entre la cantidad de alumnos de sexo femenino y masculino, porque de ser significativamente diferen-

tes esto puede influir a la hora de hacer cualquier estudio donde se deba considerar el sexo. Para ello se debe seguir el camino que a continuación se muestra, donde se ha dejado 0,50 que es la probabilidad que trae por defecto ya que se supone que la población de la que para la población de donde se ha escogido la muestra tiene igual proporción de varones que de hembras y devuelve el resultado que se muestra en la tabla que sigue:



Prueba binomial						
		Categoría	N	Prop. Observada	Prop. de prueba	Significación exacta (bilateral)
Sexo	Grupo 1	Masculino	21	,53	,50	,875
	Grupo 2	Femenino	19	,48		
	Total		40	1,00		

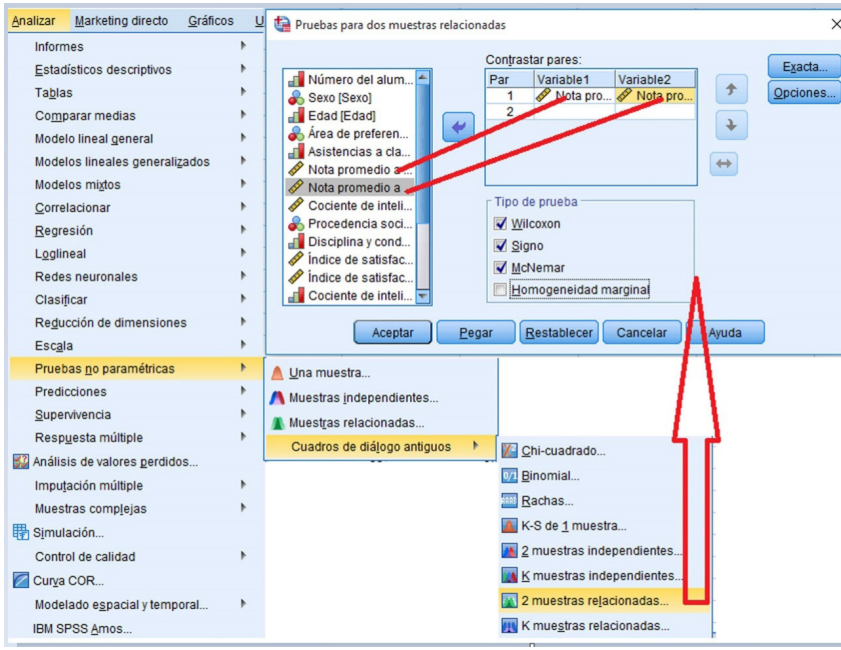
Por ser 0,875 > 0,05 se acepta la hipótesis nula, es decir no es significativa la diferencia existente entre la cantidad de alumnos de sexo masculino y femenino, por lo que cualquier prueba que los involucre pude tomar en consideración este resultado.

Otras pruebas de la estadística no paramétrica para el análisis de una muestra son el test de Kolmogorov-Smirnov y la prueba Chi-cuadrado que han sido tratadas en epígrafes anteriores.

4.13. Análisis para el caso de dos muestras relacionadas

Esta prueba se utiliza principalmente en los diseños de ¨antes y después¨ donde cada persona es usada como su propio control y es aplicada cuando las variables están medidas en una escala nominal u ordinal.

Para la base PROBLEMAS_BASE que se ha utilizado anteriormente, estas pruebas son ideales para comparar los resultados de las notas a inicio y final del semestre, su procesamiento se puede realizar siguiendo el algoritmo que se muestra en la figura adjunta:



El resultado de esta selección se muestra en la siguiente tabla:

Rangos				
		N	Rango promedio	Suma de rangos
Nota promedio a final del semestre - Nota promedio a inicio del semestre	Rangos negativos	18a	15,50	279,00
	Rangos positivos	12b	15,50	186,00
	Empates	10c		
	Total	40		
a. Nota promedio a final del semestre < Nota promedio a inicio del semestre				
b. Nota promedio a final del semestre > Nota promedio a inicio del semestre				
c. Nota promedio a final del semestre = Nota promedio a inicio del semestre				

Estadísticos de prueba[a]	
	Nota promedio a final del semestre - Nota promedio a inicio del semestre
Z	-,962[b]
Sig. asintótica (bilateral)	,336
a. Prueba de rangos con signo de Wilcoxon	
b. Se basa en rangos positivos.	

Resultado: como Sig. asintótica (bilateral)= 0,336 >0,05 Se acepta la hipótesis nula de que no existen cambios significativos entre Nota promedio a final del semestre - Nota promedio a inicio del semestre. Esto es consecuencia de que hay

10 empates (no hay cambios).

18 Retrocesos.

12 Cambios positivos.

La prueba de los signos por su parte indica que un resultado similar al anterior, pero considerando el cambio mediante signos:

Frecuencias		
		N
Nota promedio a final del semestre - Nota promedio a inicio del semestre	Diferencias negativasa	18
	Diferencias positivasb	12
	Empatesc	10
	Total	40
a. Nota promedio a final del semestre < Nota promedio a inicio del semestre		
b. Nota promedio a final del semestre > Nota promedio a inicio del semestre		
c. Nota promedio a final del semestre = Nota promedio a inicio del semestre		

Estadísticos de pruebaa	
	Nota promedio a final del semestre - Nota promedio a inicio del semestre
Z	-,913
Sig. asintótica (bilateral)	,361
a. Prueba de los signos	

Se llega al mismo resultado que por la prueba de rangos con signo de Wilcoxon.

Algunos comentarios sobre el fundamento de estas pruebas:

La prueba de los signos calcula las diferencias entre las dos variables para todos los casos y clasifica las diferencias como positivas, negativas o empatadas. Si se eliminan los empates, el problema se reduce a una binomial conEstadígrafo X igual al número de cambios positivos, en este caso, X = 12 (número de signos +).

La Decisión: aplicando la dócima binomial con n = 30; teniendo en cuenta que hay 10 empates, y p = 0,5 se puede calcular: 0,10024421

$$\{P < 12\} \approx 0{,}1002421 > \frac{0{,}5}{2} = 0{,}025$$

La prueba de Wilcoxon de los rangos con signo tiene en cuenta la información del signo de las diferencias y de la magnitud de las diferencias entre los pares. Dado que la prueba de Wilcoxon de los rangos con signo incorpora más información acerca de los datos, es más potente que la prueba de los signos.

La prueba de McNemar para la significación de los cambios es una dócima chi-cuadrado apropiada para decidir si hay o no diferencia entre dos poblaciones a partir de dos muestras apareadas en escalas nominales dicotómicas que incluyen el caso de los experimentos de antes y después en los que cada individuo o elemento de la muestra está apareado consigo mismo, usándolo como su propio control y se utiliza para verificar si hay o no cambios después, respecto a lo acontecido antes.

Ejemplo: Un especialista ha observado el comportamiento de los niños con trastornos de la conducta antes y después de la realización de un conjunto de actividades que él supone que los hará cambiar. Con la finalidad de comprobar su hipótesis, se escogen 29 niños de estos, se someten a este tratamiento y se clasifican, de acuerdo con su comportamiento en malo y aceptable, tanto antes como después de realizar el conjunto de actividades.

La prueba de homogeneidad marginal. es una extensión de la prueba de McNemar a partir de la respuesta binaria a la respuesta multinomial. Contrasta los cambios de respuesta, utilizando la distribución chi-cuadrado; es útil para detectar cambios de respuesta causados por intervención experimental en diseños antes-después. Esta prueba solo está disponible si se ha instalado “Pruebas exactas.”

4.14. Análisis para el caso de dos muestras independientes

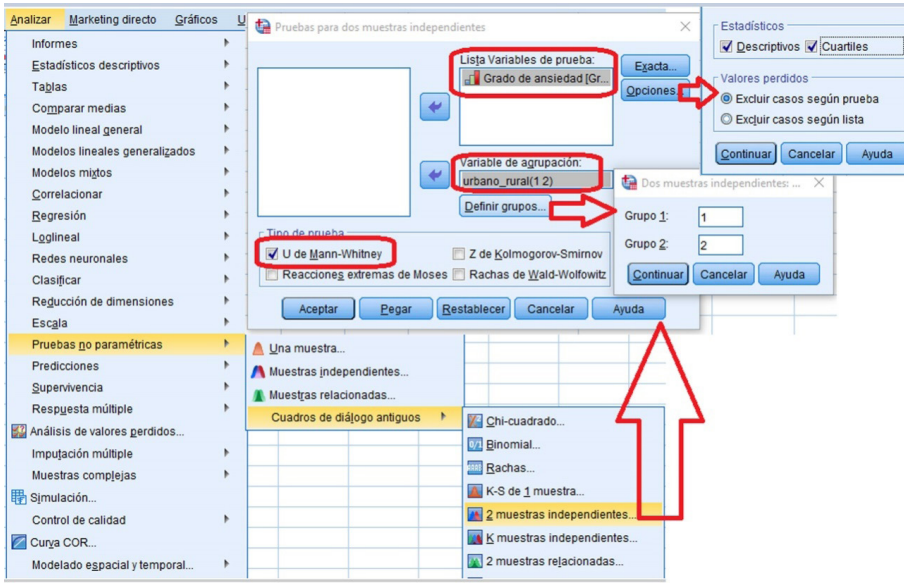
Las muestras son tomadas de formas independientes unas de otra. Ejemplo:

En un estudio sociológico el investigador necesitaba conocer si ante una misma situación los sujetos de las comunidades urbanas alcanzaban un mayor grado de ansiedad que los residentes en zonas rurales. Para ello tomó una muestra de cada zona, 23 sujetos de una localidad urbana y 16 de sujetos de una zona rural, a los cuales se le aplicaron instrumentos estandarizados en una escala de 0 a 20. Los resultados aparecen en la tabla.

SUJETOS URBANO.	G_ANSIEDAD	SUJETOS RURAL	G_ANSIEDAD.
U-1	17	R-1	13
U-2	16	R-2	12
U-3	15	R-3	12
U-4	15	R-4	10
U-5	15	R-5	10
U-6	14	R-6	10
U-7	14	R-7	10
U-8	14	R-8	9
U-9	13	R-9	8
U-10	13	R-10	8
U-11	13	R-11	7
U-12	12	R-12	7
U-13	12	R-13	7
U-14	12	R-14	7
U-15	12	R-15	7
U-16	11	R-16	6
U-17	11		
U-18	10		
U-19	10		
U-20	10		
U-21	8		
U-22	8		
U-23	6		

Para procesarlo con SPSS hay que definir dos variables:

Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida
Grado_ansiedad	Numérico	2	0	Grado de ansiedad	Ninguno	Ninguno	12	Derecha	Ordinal
urbano_rural	Numérico	1	0	Procecedencia urbana o rural	{1, De área urbana}...	Ninguno	25	Derecha	Nominal



En una se almacena el grado de ansiedad de todos los individuos y en la otra la clasificación de las variables; eso facilita el empleo de la siguiente solución del problema.

En la imagen anterior se destaca:

- La prueba utilizada es la U de Mann-Whithney.
- Se toma como variable de prueba el grado de ansiedad.
- La variable de agrupación es la dicotómica urbano/rural, que requiere para su precisión y procesamiento definir los números que se ha asignado a cada grupo.
- Las opciones de estadísticos constituyen una opción necesaria para las decisiones finales.

Con esta información se obtienen los siguientes resultados:

Estadísticos descriptivos								
	N	Media	Desviación estándar	Mínimo	Máximo	Percentiles		
						25	50 (Mediana)	75
Grado de ansiedad	39	10,87	2,966	6	17	8,00	11,00	13,00
Procedencia urbana o rural	39	1,41	,498	1	2	1,00	1,00	2,00

Prueba de Mann-Whitney				
Rangos				
	Procedencia urbana o rural	N	Rango promedio	Suma de rangos
Grado de ansiedad	De área urbana	23	25,22	580,00
	De área rural	16	12,50	200,00
	Total	39		

Estadísticos de prueba[a]	
	Grado de ansiedad
U de Mann-Whitney	64,000
W de Wilcoxon	200,000
Z	-3,451
Sig. asintótica (bilateral)	,001
Significación exacta [2*(sig. unilateral)]	,000[b]
a. Variable de agrupación: Procedencia urbana o rural	
b. No corregido para empates.	

Conclusiones: por ser la significación asintótica menor que 0,05 se rechaza la hipótesis nula de que no hay diferencias significa-

tivas para el grado de ansiedad en la muestra de individuos de áreas urbanas y rurales. Esta diferencia la marca rango promedio de grado de ansiedad que en el área urbana es mayor que en área rural.

¿Qué procesamiento realiza la prueba la U de Mann-Whithney?

Este procesamiento se puede sintetizar en:

1. Un rangueo u ordenación por rango de los datos como se muestra en siguiente tabla:

U-R	Grado de ansiedad	RANGO	
URBANO	17	39	
URBANO	16	38	
URBANO	15	36	
URBANO	15	36	
URBANO	15	36	
URBANO	14	33	
URBANO	14	33	
URBANO	14	33	
URBANO	13	29,5	
URBANO	13	29,5	
URBANO	13	29,5	
URBANO	12	24,5	
URBANO	12	24,5	
URBANO	12	24,5	
URBANO	12	24,5	
URBANO	11	20,5	
URBANO	11	20,5	
URBANO	10	16	
URBANO	10	16	
URBANO	10	16	
URBANO	8	9,5	
URBANO	8	9,5	
URBANO	6	1,5	580
RURAL	13	29,5	
RURAL	12	24,5	
RURAL	12	24,5	
RURAL	10	16	
RURAL	10	16	
RURAL	10	16	
RURAL	10	16	
RURAL	9	12	
RURAL	8	9,5	
RURAL	8	9,5	
RURAL	7	5	
RURAL	7	5	
RURAL	7	5	
RURAL	7	5	
RURAL	7	5	
RURAL	6	1,5	200

2. Cálculo de un estadígrafo de prueba para compararlo con Z= 2,33 punto de corte en la distribución normal correspondiente a α=0,05

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

Donde R1 y R2 equivalen a la suma de los rangos asignados a cada elemento de los grupos respectivos

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\left(\frac{n_1 n_2}{N(N-1)}\right)\left(\frac{N^3 - N}{12} - \sum T\right)}}$$

$$T = \frac{t^3 - t}{12}$$

t es el número de observaciones ligadas para un rango dado.

Donde:
N: Número total de elementos de ambas muestras.
: Número de elementos en la muestra 1. *n*1
: Número de elementos en la muestra 2. *n*2

$$U = 16 \times 23 + \frac{16(16+1)}{2} - 200$$

Cálculo de $\sum T$:

Observamos los siguientes grupos ligados:

• 2 puntajes de 6	• 7 puntajes de 10	• 4 puntajes de 13
• 5 puntajes de 7	• 2 puntajes de 11	• 3 puntajes de 14
• 4 puntajes de 8	• 6 puntajes de 12	• 3 puntajes de 15

$$\sum T = 2\frac{2^3 - 2}{12} + 2\frac{3^3 - 3}{12} + 2\frac{4^3 - 4}{12} + \frac{5^3 - 5}{12} + \frac{6^3 - 6}{12} + \frac{7^3 - 7}{12}$$

$$\sum T = 70{,}5$$

Para el ejemplo analizado los cálculos son:

Sustituyendo en la fórmula se tiene:

$$Z = \frac{304 - \frac{16 \times 23}{2}}{\sqrt{\left(\frac{16 \times 23}{39(39-1)}\right)\left(\frac{39^3 - 39}{12} - 70{,}5\right)}}$$

$$Z = 3{,}45$$

3. Conclusión: Como el valor Z calculado es mayor que la de la región de rechazo, se puede afirmar que existen diferencias significativas en el grado de ansiedad de los residentes en las zonas urbanas y los residentes en las zonas rurales.

En la imagen para el procesamiento del SPSS aparecen otras tres pruebas:

1. *La prueba Z de Kolmogorov-Smirnov*: se basa en la diferencia máxima absoluta entre las funciones de distribución acumulada observadas para ambas muestras. Cuando esta diferencia es significativamente grande, se consideran diferentes las dos distribuciones.
2. *La prueba de rachas de Wald-Wolfowitz*: combina y ordena las observaciones de ambos grupos. Si las dos muestras proceden de una misma población, los dos grupos deben dispersarse aleatoriamente en la clasificación.
3. *La prueba de reacciones extremas de Moses*: presupone que la variable experimental afectará a algunos sujetos en una dirección y a otros sujetos en la dirección opuesta. La prueba contrasta las respuestas extremas comparándolas con un grupo de control. Esta prueba se centra en la amplitud del grupo de control y supone una medida de la influencia de los valores extremos del grupo experimental en la amplitud al combinarse con el grupo de control. El grupo de control se define en el cuadro Grupo 1 del cuadro de diálogo Dos muestras independientes: Definir grupos. Las observaciones de ambos grupos se combinan y ordenan. La amplitud del grupo de control se calcula como la diferencia entre los rangos de los valores mayor y menor del grupo de control más 1. Debido a que los valores atípicos ocasionales pueden distorsionar fácilmente el rango de la amplitud, de manera automática se recorta de cada extremo un 5% de los casos de control.

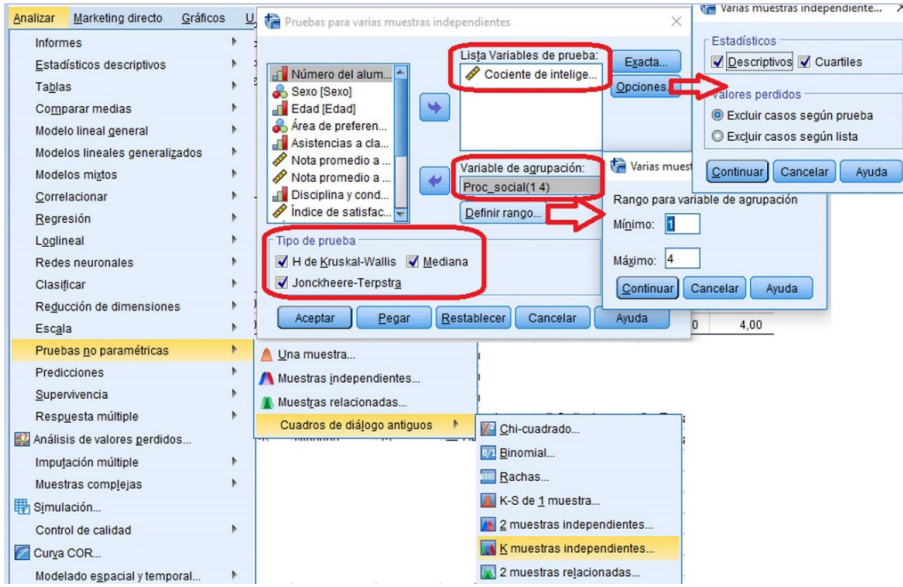
En todo el valor asintótico determina la aceptación o rechazo de la hipótesis nula.

4.15. Análisis para el caso de varias muestras independientes

Este procedimiento contiene varias pruebas no paramétricas, todas ellas diseñadas para analizar datos provenientes de diseños con una variable independiente categórica (con más de dos niveles que definen más de dos grupos o muestras) y una variable dependiente cuantitativa al menos ordinal en la cual interesa comparar las muestras. El procedimiento incluye tres pruebas:

1. La prueba H de Kruskal-Wallis.
2. La prueba de la mediana
3. La prueba de Jonckheere-Terpstra (ésta última solo se incluye en el módulo Pruebas exactas).

Para obtener cualquiera de ellas se siguen los pasos que se muestran en: la siguiente lámina.



Obsérvese la gran similitud con la prueba para dos muestras independientes, por esta vez al definir el rango se pide el menor y el mayor valor. Lo resultados para el ejemplo son los siguientes:

Estadísticos descriptivos								
	N	Media	Desviación estándar	Mínimo	Máximo	Percentiles		
						25	50 (Mediana)	75
Cociente de inteligencia	100	99,65	9,282	83	120	89,00	100,00	106,00

Prueba de Kruskal-Wallis			
Rangos			
	Procedencia social	N	Rango promedio
Cociente de inteligencia	Obrera	18	54,53
	Campesina	12	64,50
	Intelectual	12	27,50
	Clase media-alta	58	51,11
	Total	100	

Estadísticos de prueba[a,b]	
	Cociente de inteligencia
Chi-cuadrado	11,047
gl	3
Sig. asintótica	,011
a. Prueba de Kruskal Wallis	
b. Variable de agrupación: Procedencia social	

El valor asintótico de 0,011 indica que se debe negar la hipóte-

sis nula. Como dato curioso es que esta prueba se sustenta en la distribución Chi-cuadrado.

La prueba Prueba de Kruskal Wallis es una extensión realizada por los autores en 1952 de la prueba de Mann-Whitney para dos muestras independientes. La situación experimental que permite resolver esta prueba es similar a la estudiada a propósito del ANOVA de un factor completamente aleatorizado muestras son aleatoria e independientemente extraídas de J poblaciones para averiguar si las J poblaciones son idénticas o alguna de ellas presenta promedios mayores que otra.

El estadístico de prueba es:

$$K = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{i=1}^{n_i} n_i\left(r_{ij} - \bar{r}\right)^2}$$

Donde:

- n_i es el número de observaciones en el grupo i.
- r_{ij} es el rango (entre todas las observaciones) de la observación j en el grupo i.
- N es el número total de observaciones entre todos los grupos.

$$\bar{r}_i = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$$

$$\bar{r} = \frac{(N+1)}{2}$$

Es el promedio de rij

Dado que

$$\sum_{i=1}^{g}\sum_{i=1}^{n_i} n_i\left(r_{ij} - \bar{r}\right)^2 = \frac{(N-1)N(N+1)}{12}$$

K se puede expresar por

$$K = \frac{12}{N(N+1)} \sum_{i=1}^{g} n_i (\overline{r_i} - \bar{r})^2$$

Este estadístico se contrasta con la prueba

Las ventajas fundamentales de esta prueba frente al estadístico F del ANOVA de un factor completamente aleatorizado son dos:

1. No necesita establecer supuestos sobre las poblaciones originales tan exigentes como los del estadístico F (normalidad, homocedasticidad).
2. Permite trabajar con datos ordinales. Por contra, si se cumplen los supuestos en los que se basa el estadístico F, la potencia de éste es mayor que la que es posible alcanzar con el estadístico H de Kruskal-Wallis.

Ahora bien, teniendo en cuenta que en muchas situaciones reales resulta demasiado arriesgado, suponer normalidad y homocedasticidad (especialmente si las muestras son pequeñas y/o los tamaños muestrales desiguales), y considerando además que en otras situaciones el nivel de medida de los datos puede no ir más allá del ordinal, la prueba de Kruskal-Wallis representa una excelente alternativa al ANOVA de un factor completamente aleatorizado.

Prueba de la mediana					
Frecuencias					
Obrera		Procedencia social			
		Campesina	Intelectual	Clase media-alta	
Cociente de inteligencia	> Mediana	8	8	0	22
	<= Mediana	10	4	12	36

Estadísticos de prueba[a]	
	Cociente de inteligencia
N	100
Mediana	100,00
Chi-cuadrado	11,858[b]
gl	3
Sig. asintótica	,008
a. Variable de agrupación: Procedencia social	
b. 2 casillas (25,0%) han esperado frecuencias menores que 5. La frecuencia mínima de casilla esperada es 4,6.	

De nuevo el valor asintótico permite decidir en negar la hipótesis nula.

La prueba de la mediana es similar a la prueba chi-cuadrado, ella también utiliza una tabla de contingencia dicotómica, dada por los datos que cumplen la condición de ser <= Mediana y los que son > Mediana.

El objetivo de esta prueba es el de contrastarla hipótesis de que las J muestras proceden de poblaciones con la misma mediana y para eso ordena todas las observaciones y calculando la mediana total (la mediana delas n observaciones) y la de cada muestra.

Bajo estas condicione determina en cada muestra, el número de casos con puntuación igual o menor que la mediana (grupo1) y el número de casos con valor mayor que la mediana (grupo = 2), con lo que se construye una tabla de contingencia bidimensional de tamaño 2×J, con las 2 filas correspondientes a los dos grupos dicotomizados por la mediana y las J columnas correspondientes a las J muestras independientes.

Sobre esta tabla se aplica el estadístico chi-cuadrado ya estudiado, bajo la hipótesis nula de los 2 grupos y las J muestras son independientes.

Prueba de Jonckheere-Terpstra[a]	
	Cociente de inteligencia
Número de niveles en Procedencia social	4
N	100
Estadístico J-T observado	1452,000
Estadístico J-T de media	1506,000
Desviación estándar del estadístico J-T	147,164
Estadístico J-T estándar	-,367
Sig. asintótica (bilateral)	,714
a. Variable de agrupación: Procedencia social	

Estos resultados numéricos y en particular el correspondiente a "Sig. Asintótico" resultan contradictorio en relación con lo analizado y es que en realidad la prueba de *Jonckheere-Terpstra* para k muestras no es adecuada para el problema que se desarrolla porque su aplicación supondría a priori que al pasar de un nivel de procedencia social a otro el cociente de inteligencia aumenta o disminuye y todos saben que este supuesto es falso, pero la prueba es particularmente importante para probar alternativas ordenadas y en ese caso es más potente que Kruskal-Wallis.

Por ejemplo, las k poblaciones pueden representar k temperaturas ascendentes. Se contrasta la hipótesis de que diferentes temperaturas producen la misma distribución de respuesta, con la hipótesis alternativa de que cuando la temperatura aumenta, la magnitud de la respuesta aumenta. La hipótesis alternativa se encuentra aquí ordenada; por tanto, la prueba de *Jonckheere-Terpstra* es la prueba más apropiada. Del menor al mayor especifica la hipótesis alternativa de que el parámetro de ubicación del primer grupo es menor o igual que el segundo, que es menor o igual que el tercero, etc. Del mayor al menor especifica la hipótesis alternativa de que el parámetro de ubicación del primer grupo es mayor o igual que el segundo, que es mayor o igual que el tercero, etc. Para ambas opciones, la hipótesis alternativa también asume que las ubicaciones no son todas iguales. Opcionalmente puede solicitar múltiples comparaciones de las muestras k, en comparaciones múltiples todo por parejas o comparaciones por pasos en sentido descendente.

Capítulo V. Análisis de Datos Multivariados (Los inicios)

5.1. ¿Qué es el Análisis de Datos Multivariados?

El análisis multivariante o análisis de datos multivariados no es fácil de definir, pero como primer acercamiento al tema se puede asumir que es un método estadístico utilizado para determinar la contribución de varios factores en un simple evento o resultado, llamando a los factores de estudio (factores de riesgo en bioestadística), variables independientes o variables explicativas y el resultado estudiado es el evento, la variable dependiente o la variable respuesta.

En un sentido amplio, se refiere a todos los métodos estadísticos que analizan simultáneamente medidas múltiples de cada individuo u objeto sometido a investigación.

Cualquier análisis simultáneo de más de dos variables puede ser considerado aproximadamente como un análisis multivariante y en sentido estricto, muchas técnicas multivariantes son extensiones del análisis univariante.

El término multivariante no se usa de la misma forma en la literatura, así, para algunos investigadores, multivariante significa simplemente examinar relaciones entre más de dos variables y otros usan el término solo para problemas en los que se supone que todas las variables múltiples tienen una distribución normal multivariante.

Para ser considerado verdaderamente multivariante, todas las variables deben ser aleatorias y estar interrelacionadas de tal forma que sus diferentes efectos no puedan ser interpretados separadamente con algún sentido.

Algunos autores afirman que el propósito del análisis multivariante es medir, explicar y predecir el grado de relación de los valores teóricos (combinaciones ponderadas de variables), por tanto, el carácter multivariante reside en los múltiples valores teóricos (combinaciones múltiples de variables) y no solo en el número de variables u observaciones.

Algunos autores definen: *"el Análisis Multivariante es un conjunto de métodos estadísticos y matemáticos, destinados a describir e interpretar los datos que provienen de la observación de varias variables estadísticas, estudiadas conjuntamente".*

Una definición más precisa es la siguiente: *"el Análisis Multivariante es la rama de la Estadística y del análisis de datos, que estudia, interpreta y elabora el material estadístico sobre un conjunto de n>1 de variables, que pueden ser cuantitativas, cualitativas o una mezcla".* (Cuadras, 1981)

Según Hair, Anderson, Tatham & Black, (1999), en *"Análisis Multivariante."*

Durante la década de los ochenta se fueron desarrollando los programas estadísticos que facilitaron el análisis de gran cantidad de datos cuyo origen estaba en encuestas o en bases de datos que provenían de fuentes secundarias de información. Los fundamentos teóricos o estadísticos de las técnicas multivariantes que permitían el análisis de estos datos habían sido desarrollados con anterioridad, pero solo cuando los ordenadores tuvieron la capacidad de cálculo y memoria necesarios para llevar a cabo el análisis multivariante, se empezó a generalizar el uso de estas técnicas. ´

Es poco menos que imposible discutir la aplicación de las técnicas multivariantes sin una menciónale impacto de la informática. [...] el amplio desarrollo de la aplicación de los computadores (primero el computador y más recientemente los computadores personales o los microcomputadores) para procesar grandes y complejas bases de datos, ha estimulado de manera impresionante el uso de los métodos de estadística multivariante. Toda la estadística teórica de las técnicas multivariantes actuales fue desarrollada mucho antes de la aparición de los computadores, pero solo cuando estuvo disponible el poder de la informática para realizar cálculos cada vez más complejos llegó a conocerse la existencia de esas técnicas fuera del círculo de los estadísticos teóricos.

Los continuos avances tecnológicos en informática, particularmente en los computadores personales, han puesto a disposición de cualquier investigador interesado el acceso a todos los recursos necesarios para resolver un problema multivariante de casi cualquier dimensión. De hecho, muchos investigadores se llaman a sí mismos analistas de datos en lugar de estadísticos o (en lenguaje llano) «aficionados a lo cuantitativo». Estos analistas de datos han contribuido sustancialmente al aumento del uso y aceptación de la estadística multivariante en los negocios y en la administración. En la comunidad académica, disciplinas de todos los campos del saber han adoptado las técnicas multivariantes, y los académicos deben estar cada vez más versados en las técnicas multivariantes apropiadas para sus investigaciones empíricas. Incluso para personas consólida preparación cuantitativa, la disponibilidad de programas preparados para el análisis multivariante ha facilitado la compleja manipulación de matrices de datos que durante mucho tiempo ha retrasado el crecimiento de técnicas multivariantes.

5.2. ¿Para qué sirve el Análisis ultivariante o multivariados?

La respuesta obliga a precisar los objetivos del Análisis multivariante de datos:

1. Resumir los datos mediante un pequeño conjunto de nuevas variables con la mínima pérdida de información.
2. Encontrar grupos en los datos, si existen.
3. Clasificar nuevas observaciones en grupos definidos.
4. Relacionar dos conjuntos de variables

En las ciencias particulares los métodos multivariantes resuelven diversos problemas, algunos ejemplos son los relacionados con:

- Administración de empresas: para construir tipología de clientes.
- Agricultura: para clasificar terrenos de cultivo por fotografía aérea.

- Arqueología: clasificar restos arqueológicos.
- Biometría: identificar los factores que determinan la forma de un organismo vivo.
- Computación: diseñar algoritmos de clasificación automática.
- Educación: para investigar la efectividad del aprendizaje a distancia.
- Medio Ambiente: dimensiones de la contaminación ambiental.
- Documentación: para clasificar revistas por su contenido.
- Economía: dimensiones del desarrollo económico.
- Geología: clasificar sedimentos.
- Lingüística: encontrar patrones de asociación de palabras.
- Medicina: para identificar tumores.
- Psicología: para identificar factores que componen la inteligencia humana.

Las principales técnicas multivariantes se denominan:

- Análisis de Componentes principales.
- Análisis factorial.
- Análisis discriminante.
- Análisis de Correlación Canónica.
- Análisis de Clúster.
- Análisis de Escalamiento Dimensional.
- Análisis de correspondencias.
- Análisis factorial confirmatorio.
- Modelo de Ecuaciones Estructurales (SEM), análisis causal.
- Análisis conjunto.
- Regresión Lineal Múltiple.
- Regresión Logit y Probit.
- Análisis Manova.

Las relaciones entre algunos de estas técnicas y su empleo se muestran en la siguiente tabla:

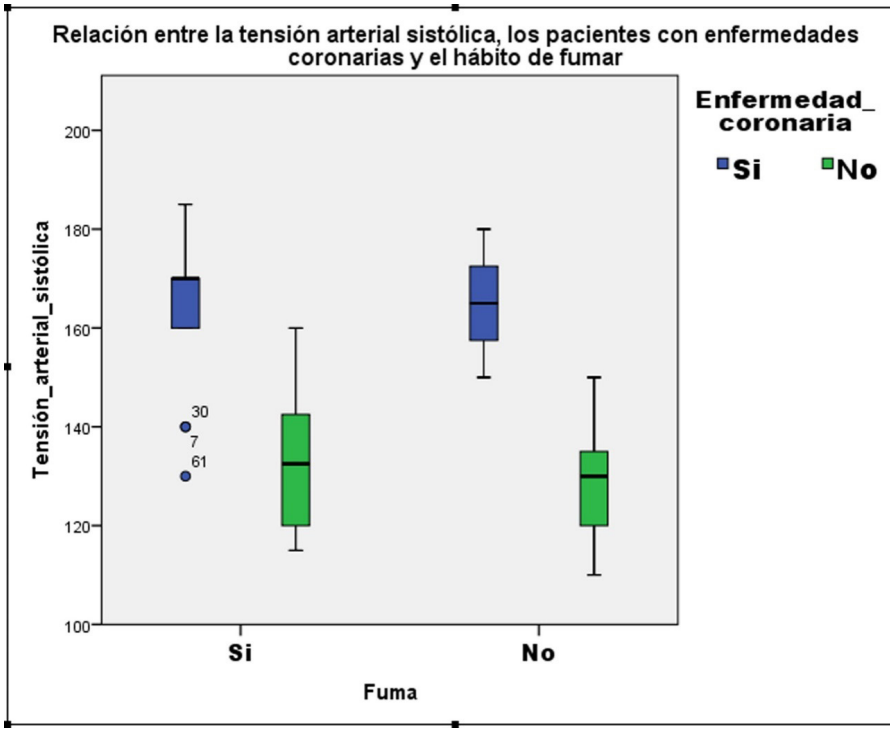
	Análisis de componentes principales	Análisis por factores	Análisis discriminante	Análisis discriminante canónico	Análisis por agrupación	Análisis multivariado de la varianza (MANOVA)	Análisis de variables canónicas	Análisis de correlación canónica
Exploración de las Relaciones entre las variables	A veces	Indudablemente	Nunca	Nunca	Nunca	Nunca	Rara vez	A veces
Cribado de los datos	Indudablemente	A veces	Nunca	Nunca	A veces	Nunca	Nunca	Nunca
Creación de nuevas variables	Lo hace	Lo hace	No lo hace	Lo hace	No lo hace	No lo hace	Lo hace	Lo hace
Predicción de ser miembro de un grupo	No lo hace	No lo hace	Lo hace	Lo hace	Lo hace	No lo hace	No lo hace	No lo hace
Comparación de medias grupales	Posiblemente	Posiblemente	Rara vez	Rara vez	No lo hace	Lo hace	Lo hace	No lo hace
Comparación de grupos de variables	Posiblemente	Posiblemente	Nunca	Nunca	Nunca	Nunca	Nunca	Indudablemente
Verificación de Agrupamientos	Indudablemente	Posiblemente	Nunca	Nunca	Indudablemente	Nunca	Nunca	Nunca
Reducción de la Dimensionalidad	Indudablemente	Indudablemente	Nunca	Indudablemente	Nunca	Nunca	Indudablemente	Indudablemente
Creación de variables significativas	No es probable	Por lo común	Nunca	Posiblemente	Nunca	Nunca	Posiblemente	No es probable

5.3. El análisis de los datos individuales como primer paso del análisis multivariante de datos

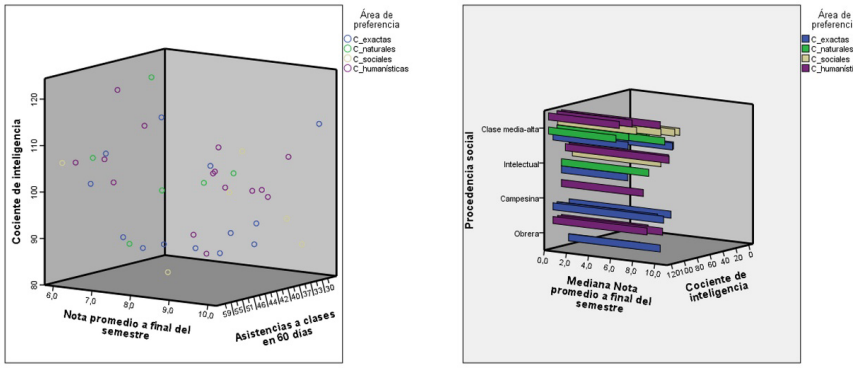
Desde los inicios de este libro se ha insistido que antes de aplicar cualquier técnica estadística es preciso realizar un análisis previo de los datos de que se dispone y este principio se mantiene n el análisis multivariante de datos. Es necesario examinar las variables individuales y las relaciones entre ellas, así como evaluar y solucionar problemas en el diseño de la investigación y en la recogida de datos tales como el tratamiento de la información faltante (o datos ausentes) y la presencia de datos anómalos (o atípicos).

Ahora los gráficos se extienden a representar más de dos variables; un ejemplo son el gráfico múltiple de caja y bigotes que permite analizar, resumir y comparar simultáneamente varios conjuntos de datos univariados que corresponden a los diferentes grupos en que se pueden subdividir los valores de una variable. Este tipo de grafico permite estudiar la simetría de los datos, detectar valores atípicos y representar medias, medianas, rangos y valores extremos para todos los grupos. Por realizar las representaciones de las variables simultáneamente se pueden comparar medias, medianas, rangos, valores extremos, simetrías y valores atípicos de todos los grupos. El gráfico múltiple representará horizontalmente un gráfico de caja y bigotes para cada grupo de valores de la variable en estudio. La siguiente figura muestra el proceso de construcción de uno de ellos:

Se obtiene el siguiente gráfico:



En el gráfico se destacan tres casos atípicos perfectamente identificados por sus números de orden que son enfermos que fuman; por otro lado, la gráfica muestra el comportamiento de la tensión arterial en paciente sano y enfermo que fuman y no fuman. A partir del mismo cuadro de diálogo inicial se pueden construir otros gráficos multivariantes:



5.4. Análisis de componentes principales

Un problema central en el análisis de datos multivariantes es la reducción de la dimensionalidad, esto es, determinar si es posible describir con precisión los valores de p variables de una muestra por un pequeño subconjunto r < p de ellas, de modo que se reduzca la dimensión del problema a costa de una pequeña pérdida de información.

El análisis de componentes principales tiene este objetivo: dada n observaciones de p variables, se analiza si es posible representar adecuadamente esta información con un número menor de variables construidas como combinaciones lineales de las originales.

Obsérvese que en estos dos párrafos se ha precisado el objetivo de esta técnica de análisis multivariado y su método de proceder, la construcción de nuevas variables como combinaciones lineales de las originales.

Ejemplo: Un análisis según A.E.D. de una muestra tomada en una empresa de las 7 primeras variables de la base HATCO, arroja los siguientes resultados:

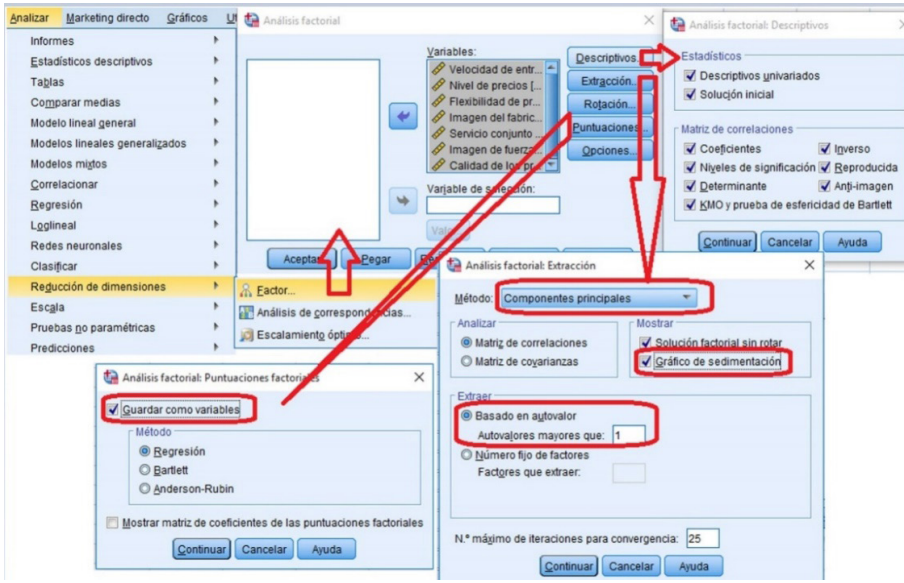
Correlaciones							
	Velocidad de entrega	Nivel de precios	Flexibilidad de precios	Imagen del fabricante	Servicio conjunto	Imagen de fuerza de ventas	Calidad de los productos
Velocidad de entrega	1	-,349**	,509**	,050	,612**	,102	-,483**
Nivel de precios	-,349**	1	-,487**	,272**	,513**	,194	,470**
Flexibilidad de precios	,509**	-,487**	1	-,116	,067	-,062	-,448**
Imagen del fabricante	,050	,272**	-,116	1	,299**	,754**	,200*
Servicio conjunto	,612**	,513**	,067	,299**	1	,263**	-,055
Imagen de fuerza de ventas	,102	,194	-,062	,754**	,263**	1	,191
Calidad de los productos	-,483**	,470**	-,448**	,200*	-,055	,191	1
**. La correlación es significativa en el nivel 0,01 (bilateral).							
*. La correlación es significativa en el nivel 0,05 (bilateral).							

De la tabla se infiere que entre estas variables hay una alta correlación con significaciones en niveles entre 0,05 y 0,01 lo que indica que es posible reducir la dimensionalidad, es decir se

pueden encontrar menos variables que sean combinaciones lineales de las siete variables analizadas y que entre ellas no existan correlaciones con niveles de significación tan altos.

La siguiente imagen muestra las opciones a seleccionar en el menú del SPSS.

En la imagen anterior se destaca:



1. Los estadísticos descriptivos, en particular el índice KMO.
2. De la opción extracción:

a. El método, en este caso componentes principales.

b. En la opción mostrar se activó la pestaña correspondiente a Gráfico de sedimentación".

c. La extracción de los componentes se hará en este caso basado en los autovalores[xxiv] mayores que 1.

3. De la opción "Puntuaciones" se seleccionó "Guardar como variable".

El procesamiento de la información devuelve los siguientes resultados:

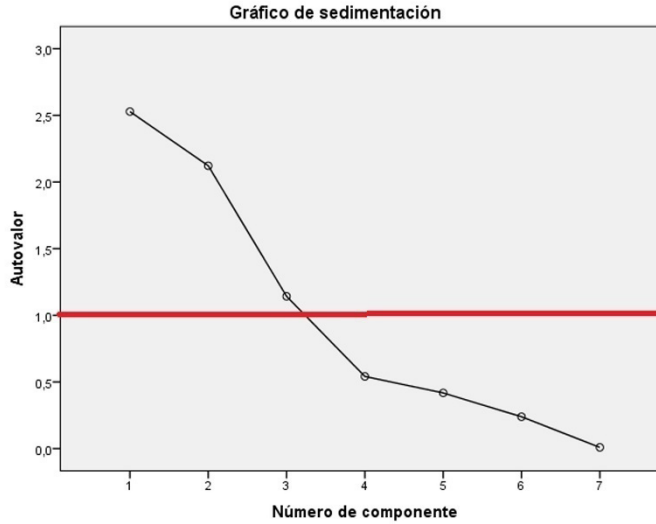
Comunalidades		
	Inicial	Extracción
Velocidad de entrega	1,000	,885
Nivel de precios	1,000	,900
Flexibilidad de precios	1,000	,646
Imagen del fabricante	1,000	,865
Servicio conjunto	1,000	,995
Imagen de fuerza de ventas	1,000	,883
Calidad de los productos	1,000	,620
Método de extracción: análisis de componentes principales.		

Descriptores univariados incluyen la media, la desviación estándar y el número de casos válidos para cada variable; sobre estos resultados ya se conoce su significado y forma de obtenerlos, pero asociado a ellos se obtiene también:

- Las comunalidades iniciales, estimaciones de la varianza compartida o común entre las variables, expresada en la proporción de la variabilidad de cada variable explicada por los factores, la cual en el caso de los componentes principales da 1 como comunalidad inicial de todas las variables.
- Los autovalores y el porcentaje de varianza explicada.

Varianza total explicada						
Componente	Autovalores iniciales			Sumas de extracción de cargas al cuadrado		
	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado
1	2,529	36,123	36,123	2,529	36,123	36,123
2	2,122	30,312	66,436	2,122	30,312	66,436

3	1,142	16,321	82,757	1,142	16,321	82,757
4	,541	7,733	90,490			
5	,418	5,967	96,456			
6	,239	3,414	99,870			
7	,009	,130	100,000			
Método de extracción: análisis de componentes principales.						



En esta tabla se resume el procedimiento de análisis de componentes principales. El propósito del análisis es obtener un número reducido de combinaciones lineales de las 7 variables que expliquen la mayor variabilidad en los datos. En este caso, 3 componentes se han extraído puesto que 3 componentes tuvieron autovalores mayores o iguales que 1,0 (recuérdese que en las cajas de diálogos iniciales se seleccionó "La extracción de los componentes se hará en este caso basado en los autovalores mayores que 1"). En conjunto estos tres componentes explican el 83,3182% de la variabilidad en los datos originales.

Lo explicado se muestra en el gráfico de sedimentación que también se devuelve como resultado.

Otros resultados son los asociados a la matriz de correlación anteriormente referenciadas y asociados a ellas se dan las siguientes tablas:

Inversión de matriz[xxv] de correlaciones							
	Velocidad de entrega	Nivel de precios	Flexibilidad de precios	Imagen del fabricante	Servicio conjunto	Imagen de fuerza de ventas	Calidad de los productos
Velocidad de entrega	36,045	32,354	,117	1,793	-38,934	-1,011	-,063
Nivel de precios	32,354	31,726	1,101	1,454	-36,455	-,678	-,967
Flexibilidad de precios	,117	1,101	1,633	,093	-,751	-,038	,218
Imagen del fabricante	1,793	1,454	,093	2,518	-2,135	-1,783	-,057
Servicio conjunto	-38,934	-36,455	-,751	-2,135	44,024	,858	,690
Imagen de fuerza de ventas	-1,011	-,678	-,038	-1,783	,858	2,396	-,239
Calidad de los productos	-,063	-,967	,218	-,057	,690	-,239	1,616

Prueba de KMO y Bartlett		
Medida Kaiser-Meyer-Olkin de adecuación de muestreo		,445
Prueba de esfericidad de Bartlett	Aprox. Chi-cuadrado	556,191
	gl	21
	Sig.	,000

KMO y prueba de esfericidad de Bartlett. La medida de la adecuación muestral de Kaiser-Meyer-Olkin contrasta si las correlaciones parciales entre las variables son pequeñas. La prueba de esfericidad de Bartlett contrasta si la matriz de correlaciones es una matriz de identidad, que indicaría que el modelo factorial es inadecuado.

Prueba de esfericidad de Barlett

Esta prueba contrasta las siguientes hipótesis: H0: R=1; H1: R <> 1

La hipótesis nula postula que la matriz de correlaciones es una matriz identidad; esto significa que las correlaciones entre las variables son todas igual a cero, puesto que en una matriz identidad los elementos de la diagonal principal son todos unos y, por lo tanto, el valor del determinante es igual a 1. La hipótesis alternativa asume que la matriz de correlaciones es distinta de una matriz identidad o, lo que es lo mismo, que el determinante de la matriz de correlaciones es significativamente distinto de uno.

El determinante de una matriz de correlaciones es un índice de la varianza generalizada de dicha matriz; un determinante próximo a cero indica que una o más variables pueden ser expresadas como una combinación lineal de las otras variables.

Tiene sentido un análisis factorial si podemos rechazar la hipótesis nula, lo cual sería indicativo de que existen correlaciones

entre las variables. En caso de no poder rechazar la hipótesis nula, no tendría sentido un análisis factorial, puesto que esto indicaría que existe poca información redundante y, por tanto, el número de factores necesario para explicar un alto porcentaje de información sería próximo al de variables originales.

Índice KMO de Kaiser-Meyer-Olkin

$$KMO = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}^{\,2}}{\sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}^{\,2} + \sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}^{\,2}}$$

r_{ij} es el coeficiente de correlación entre las variables i-ésima y j-ésima; se excluyen de los sumatorios los campos de aplicación de los sumatorios no es aplicable en los casos i = j, S_{ij} es el coeficiente de correlación parcial entre las variables i-ésima y j-ésima. También se excluyen los casos i = j.

Un índice KMO bajo indica que la intercorrelación entre las variables no es grande y, por lo tanto, el análisis factorial no sería práctico, ya que necesitaríamos casitantos factores como variables para incluir un porcentaje de la información aceptable.

KAISER indica que:

- Un KMO mayor que 0.7 es indicativo de muy alta intercorrelación y, por tanto, indicativo de que el Análisis Factorial / Componentes principales es una técnica muy útil.
- Entre 0,6 y 0,7 el grado de intercorrelación es alto y el Análisis Factorial se considera útil.
- Entre 0.5 y 0.6 el grado de intercorrelación es medio y el Análisis Factorial sería menos útil que en el caso anterior, pero aplicable;
- Un KMO < 0.5 indicaría que el Análisis Factorial no resultaría una técnica útil.

Correlaciones reproducidas									
			Velocidad de entrega	Nivel de precios	Flexibilidad de precios	Imagen del fabricante	Servicio conjunto	Imagen de fuerza de ventas	Calidad de los productos
Correlación reproducida		Velocidad de entrega	,885a	-,248	,613	,074	,606	,105	-,592
		Nivel de precios	-,248	,900a	-,600	,263	,524	,182	,554
		Flexibilidad de precios	,613	-,600	,646a	-,101	,062	-,042	-,612
		Imagen del fabricante	,074	,263	-,101	,865a	,303	,870	,263
		Servicio conjunto	,606	,524	,062	,303	,995a	,264	-,078
		Imagen de fuerza de ventas	,105	,182	-,042	,870	,264	,883a	,215
		Calidad de los productos	-,592	,554	-,612	,263	-,078	,215	,620a

Residuob	Velocidad de entrega		-,101	-,104	-,024	,006	-,003	,110
	Nivel de precios	-,101		,113	,009	-,011	,012	-,084
	Flexibilidad de precios	-,104	,113		-,016	,005	-,020	,164
	Imagen del fabricante	-,024	,009	-,016		-,005	-,116	-,063
	Servicio conjunto	,006	-,011	,005	-,005		-,001	,023
	Imagen de fuerza de ventas	-,003	,012	-,020	-,116	-,001		-,025
	Calidad de los productos	,110	-,084	,164	-,063	,023	-,025	
Método de extracción: análisis de componentes principales.								
a. Comunalidades reproducidas								
b. Los residuos se calculan entre las correlaciones observadas y reproducidas. Existen 8 (38,0%) residuos no redundantes con valores absolutos mayores que 0,05.								

Reproducida. La matriz de correlaciones estimada a partir de la solución del factor. También se muestran las correlaciones de residuos (la diferencia entre la correlación observada y la estimada).

Matrices anti-imagen									
		Velocidad de entrega	Nivel de precios	Flexibilidad de precios	Imagen del fabricante	Servicio conjunto	Imagen de fuerza de ventas	Calidad de los productos
Covarianza anti-imagen	Velocidad de entrega	,028	,028	,002	,020	-,025	-,012	-,001
	Nivel de precios	,028	,032	,021	,018	-,026	-,009	-,019
	Flexibilidad de precios	,002	,021	,612	,023	-,010	-,010	,083
	Imagen del fabricante	,020	,018	,023	,397	-,019	-,296	-,014
	Servicio conjunto	-,025	-,026	-,010	-,019	,023	,008	,010
	Imagen de fuerza de ventas	-,012	-,009	-,010	-,296	,008	,417	-,062
	Calidad de los productos	-,001	-,019	,083	-,014	,010	-,062	,619

Correlación anti-imagen	Velocidad de entrega	,343a	,957	,015	,188	-,977	-,109	-,008
	Nivel de precios	,957	,330a	,153	,163	-,975	-,078	-,135
	Flexibilidad de precios	,015	,153	,933a	,046	-,089	-,019	,134
	Imagen del fabricante	,188	,163	,046	,554a	-,203	-,726	-,028
	Servicio conjunto	-,977	-,975	-,089	-,203	,290a	,084	,082
	Imagen de fuerza de ventas	-,109	-,078	-,019	-,726	,084	,561a	-,121
	Calidad de los productos	-,008	-,135	,134	-,028	,082	-,121	,926a
a. Medidas de adecuación de muestreo (MSA)								

Anti-imagen. El coeficiente de correlación parcial es un indicador de la fuerza de la asociación entre dos variables que elimina la influencia de las otras variables. Si existen factores comunes, esperamos que los coeficientes de correlación parcial sean pequeños. El coeficiente de correlación anti-imagen es el negativo del coeficiente de correlación parcial entre dos variables. Es aplicable el análisis factorial si en la matriz de correlaciones anti-imagen hay muchos coeficientes con valores pequeños.

Este índice se calcula para cada variable, de forma similar al índice KMO.

5.5. Medida de Adecuación de la Muestra (MSA)

$$MSA = \frac{\sum_{j=1}^{n} r_{ij}^2}{\sum_{j=1}^{n} r_{ij}^2 + \sum_{j=1}^{n} s_{ij}^2}$$

Finalmente, SPSS devuelve la siguiente matriz:

Matriz de componente [a]			
	Componente		
	1	2	3
Velocidad de entrega	-,517	,765	,179
Nivel de precios	,796	,095	,507
Flexibilidad de precios	-,697	,368	-,159
Imagen del fabricante	,557	,582	-,464
Servicio conjunto	,197	,791	,575
Imagen de fuerza de ventas	,496	,596	-,530
Calidad de los productos	,740	-,270	-,018
Método de extracción: análisis de componentes principales.			
a. 3 componentes extraídos.			

Esta tabla muestra las ecuaciones de los componentes principales. Por ejemplo, el primer componente principal tiene la ecuación:

0,517*Velocidad de entrega + 0,796*Nivel de precios - 0,697

* Flexibilidad de precios + 0,557 *Imagen del fabricante + 0,197

*Servicio conjunto + 0,496 * Imagen de fuerza de ventas + 0,740

* Calidad de los productos

Sabiendo que los valores de las variables en la ecuación se han estandarizado restándoles su media y dividiéndolos entre sus desviaciones estándar.

Finalmente se generan tres nuevas variables que se incorporan a la base que como se muestra a continuación no están correlacionadas entre sí como se puede ver en la matriz de covarianza que se adjunta.

FAC1_1	Numérico	11	5	REGR factor score	1 for analysis 1	Ninguno	Ninguno	13	Derecha	Escala
FAC2_1	Numérico	11	5	REGR factor score	2 for analysis 1	Ninguno	Ninguno	13	Derecha	Escala
FAC3_1	Numérico	11	5	REGR factor score	3 for analysis 1	Ninguno	Ninguno	13	Derecha	Escala

FAC1_1	FAC2_1	FAC3_1
-,92516	-,39254	-,42639
1,56211	-,16618	-1,36388
1,73292	,59284	1,86752
,00840	-,95015	-1,32314
-,54463	2,57188	-2,10630
,70179	-1,14314	,49571
,41136	1,70835	-1,18173
1,00416	-,98552	,81261
-,64991	,93364	,17482
1,18387	,62777	,53528
-,56726	-,63170	-1,04213
-,24722	-,01907	,21644

Matriz de covarianzas de puntuación de componente

Componente	1	2	3
1	1,000	,000	,000
2	,000	1,000	,000
3	,000	,000	1,000

Método de extracción: análisis de componentes principales.
Puntuaciones de componente.

De esta forma se muestra que de 7 variables altamente correlacionadas se han obtenido 3 nuevas variables que explican a las anteriores en un 82,757% y que no están correlacionadas entre sí.

5.6. El análisis factorial

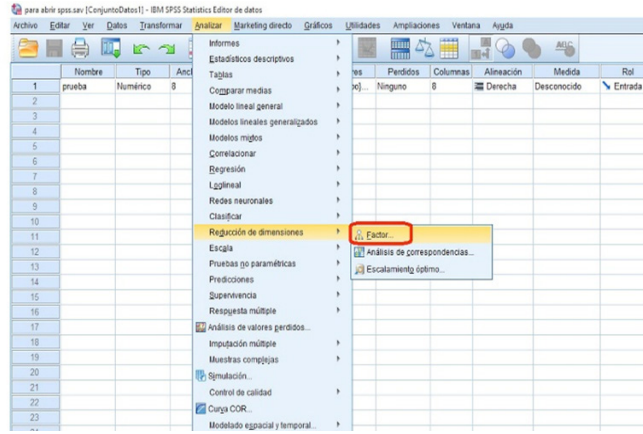
En numerosas áreas de Psicología y de Ciencias del Comportamiento no es posible medir directamente las variables que interesan; por ejemplo, los conceptos de inteligencia y de clase social. En estos casos es necesario recoger medidas indirectas que estén relacionadas con los conceptos que interesan. Las variables que interesan reciben el nombre de variables latentes y la metodología que las relaciona con variables observadas recibe el nombre de Análisis Factorial.

De lo anterior se infiere que el análisis factorial es una técnica de reducción de datos que sirve para encontrar grupos homogéneos de variables a partir de un conjunto numeroso de variables. Esos grupos homogéneos se forman con las variables que correlacionan mucho entre sí y procurando, inicialmente, que unos grupos sean independientes de otros.

El Análisis Factorial puede ser exploratorio o confirmatorio. El análisis exploratorio se caracteriza porque no se conocen a prio-

ri el número de factores y es en la aplicación empírica donde se determina este número. Por el contrario, en el análisis de tipo confirmatorio los factores están fijados a priori, utilizándose contrastes de hipótesis para su corroboración.

El análisis de componentes principales y el análisis factorial son dos técnicas conceptualmente distintas, aunque el procedimiento matemático es similar en ambas, por eso los grandes paquetes estadísticos, como el SPSS, los incluyen en el mismo procedimiento (FACTOR) las técnicas necesarias para realizar ambos análisis. Para comprender mejor el análisis factorial se comparará con el análisis del componente principal:

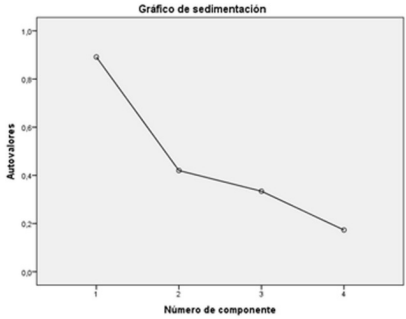


5.7. Comparación de análisis factorial con el análisis del componente principal

Componente principal	Análisis factorial
No se tiene hipótesis previa, pero se sabe que el 100% de la variabilidad de las K variables se explica por K factores, cada uno de los cuales es combinación lineal de las variables originales.	Se supone que hay una parte común, COMUNALIDAD, de la variabilidad de las variables, explicada por factores comunes no observables. Cada variable tiene una parte de su variabilidad no común propia de cada variable; a esta variabilidad no común se llama factor único. Se asume que los factores únicos correspondientes a las variables son independientes entre sí.

No se pretende sustituir las K variables por K factores, a veces de difícil interpretación, pero los factores recogen la variabilidad de las variables originales de forma desigual. En muchas ocasiones, pocos factores recogen un porcentaje de variabilidad alto; por lo tanto, se podría explicar la mayor parte de la variabilidad original a partir de ellos.	Se distinguen dos tipos de análisis factorial, el exploratorio AFE y el confirmativo AFC. En el AFE el investigador no tiene a priori una hipótesis acerca del número de factores comunes; éstos se seleccionan durante el análisis. En el AFC, el investigador parte de la hipótesis de que existe un número determinado de factores, los cuales tienen un significado determinado. Ejemplo, en el problema de las asistencias y las notas en las asignaturas se puede asumir que la asistencia tiene incidencia sobre las notas en las asignaturas.
Modelo matemático: no se tiene a priori ninguna hipótesis acerca de la cualidad de los factores. El modelo parte de la base de que se tiene invariables inicialmente y, a partir de ellas, se han calculado K factores linealmente independientes y ortogonales. Conceptualmente, el modelo anterior indica que el 100% de la información de la variable se explica por los K factores. Se Llama COMUNALIDAD a la proporción de la variabilidad de cada variable explicada por los factores.	*Modelo matemático:* Xij es el valor de la j-ésima variable correspondiente al i-ésimo caso, Fij son los coeficientes factoriales correspondientes al i-ésimo caso y aij las puntuaciones factoriales, Uj es el factor único correspondiente a la j-ésima variable. La diferencia del modelo del análisis factorial respecto al decompo-nentes principales es que el análisis factorial supone que la variabilidad de cada variable tiene una parte explicable por factores comunes y otra independiente de las demás variables.

A partir de la matriz de varianzas covarianzas o de la matriz de correlaciones, se calculan los autovalores de la matriz. A partir de estos autovalores se realiza el cálculo de los correspondientes autovectores. Si se tienen K variables iniciales, la matriz de varianzas covarianzas y la matriz de correlaciones tienen dimensión K X K, y a partir de ellas se extraen K autovalores, los cuales darán origen a K autovectores. Cada autovector define un eje correspondiente a un factor. Los K ejes definidos corresponden a K factores ortogonales. La variabilidad total de la información original, está recogida en estos K factores.



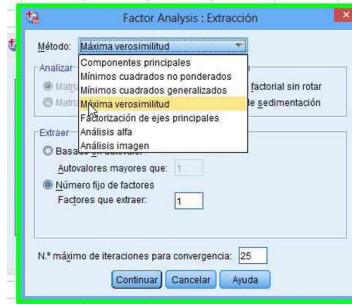
El porcentaje de variabilidad que recoge cada factor suele ser muy distinto, de tal forma que unos pocos factores (COMPONENTES PRINCIPALES) explican gran parte de la variabilidad total.

En un modelo factorial, se parte de la base de que solo una parte de la variabilidad de cada variable depende de factores comunes y, por lo tanto, se busca una comunalidad inicial para cada variable. Para ello se construye un modelo de regresión múltiple para cada variable. En cada uno de los modelos figura una variable distinta como variable dependiente y el resto como variables independientes. El coeficiente de determinación del modelo en que cada variable figura como variable dependiente se considera como comunalidad inicial.

Ejemplo, supongamos que se dispone de las variables PESO, TALLA y EDAD; se construye un modelo de regresión múltiple en el que la variable dependiente sea el PESO, y la EDAD y la TALLA las variables independientes de dicho modelo. A continuación, se construye otro modelo de regresión múltiple, en el que la variable dependiente sea la EDAD y el PESO y la TALLA las variables independientes.

Por último, se construye un tercer modelo en el que la variable dependiente sea la TALLA y el PESO y la EDAD las variables independientes.

Las características de los factores vienen condicionadas por la matriz de correlaciones; muchas correlaciones altas entre las variables, es indicativo de información redundante y pocos factores explicarán gran parte de la variabilidad total. Por el contrario, correlaciones pequeñas entre las variables son indicativas de poca información redundante y, por lo tanto, se necesitan muchos factores para explicar una parte sustancial de la variabilidad.	Si los coeficientes de determinación de los tres modelos han sido 0.7 para el PESO, 0.57 para la EDAD y 0.64 para la TALLA. Dichos coeficientes de determinación se considerarán como COMUNALIDADES iniciales en los modelos factoriales.
Fases de análisis de componentes principales: Elección de los componentes principales.	Fases análisis de un análisis factorial: Examen de la matriz de correlaciones de todas las variables que constituyen los datos originales. Extracción de los factores comunes.
Rotación de los ejes. Representaciones gráficas. Cálculo de las puntuaciones factoriales.	
Elección de los componentes principales: La elección de los ejes factoriales se realiza de tal manera que el primer factor recoja la máxima proporción posible de la variabilidad de la nube de puntos original. La variabilidad de la proyección de la nube de puntos sobre el eje definido por el factor debe ser la máxima posible. El segundo factor debe recoger la máxima variabilidad posible no recogida por el primer factor y así	**Examen de la matriz de correlaciones:** Un análisis factorial tiene sentido si existen altas correlaciones entre las variables; esto es indicativo de información redundante o, lo que es lo mismo, que algunas variables aportan información que en gran parte llevan también otras variables, y ello es indicativo de la existencia de factores comunes.

sucesivamente, hasta la selección de los K factores. De los K factores posibles, se eligen aquellos que recojan el porcentaje de variabilidad que se estime suficiente. A los factores elegidos se les llama COMPONENTES PRINCIPALES	Note que, en el análisis de componentes principales, no tiene sentido el examen de la matriz de correlaciones, ya que no se tienen hipótesis de la existencia de factores comunes. Los métodos para la comprobación analítica del grado de intercorrelación entre las variables fueron tratados en epígrafes anteriores: Prueba de esfericidad de Barlett Índice KMO de Kaiser-Meyer-Olkin Correlación antiimagen Medida de adecuación de la muestra (MSA) Correlación múltiple.
	Extracción de los factores comunes: Los métodos más utilizados y que incluyen los principales paquetes estadísticos como SPSS son: Máxima verosimilitud. Factorización de ejes principales. Factorización alfa. Factorización de imagen. Mínimos cuadrados no ponderados. Mínimos cuadrados generalizados. 

Rotación de los ejes:

Las características que deben tener los factores, para que sean fácilmente interpretables, son las siguientes:

Las cargas factoriales de un factor con las variables deben ser próximas a 1 o próximas a cero.

Una variable debe tener cargas factoriales elevadas con un solo factor. Ha de intentarse que la mayor parte de la variabilidad de una variable sea explicada por un solo factor.

No deben existir factores con cargas factoriales similares. Si dos o más factores tienen cargas factoriales altas o bajas con las mismas variables, en realidad explican lo mismo y serían redundantes, lo cual sería un contrasentido puesto que el análisis factorial intenta eliminar la redundancia.

Las tres características anteriores son difíciles de cumplir por los factores originales, pero es posible conseguirlo rotando los factores.

Las rotaciones pueden ser ortogonales u oblicuas. Estas rotaciones permiten que comunalidades de cada variable se conservan, aunque cambian las cargas factoriales, puesto que los ejes son distintos al ser rotados, pero la variabilidad explicada de cada variable permanece inalterada. Las rotaciones ortogonales más importantes son la rotación VARIMAX y la rotación CUARTIMAX.

ROTACIÓN VARIMAX. Este método maximiza la varianza de los factores. Cada columna de la matriz factorial rotada tendrá cargas factoriales altas con algunas variables y bajas con otras, lo cual facilitará la interpretación.

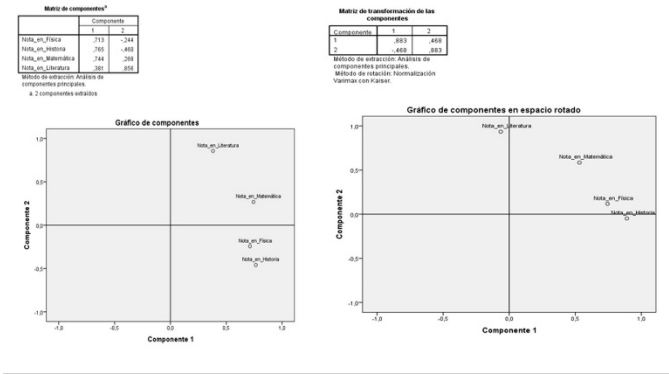
La rotación VARIMAX es la que realiza SPSS por defecto, aunque puede realizar otras rotaciones si se le indica.

ROTACIÓN CUARTIMAX. Trata de simplificar las filas de la matriz factorial, de esta manera, cada variable tendrá una correlación alta con pocos factores y baja con los demás, lo cual facilitará la interpretación.

Rotaciones oblicuasLas rotaciones oblicuas pretenden los mismos objetivos que las ortogonales. En general, solo se realizan cuando las rotaciones ortogonales no logran su objetivo.

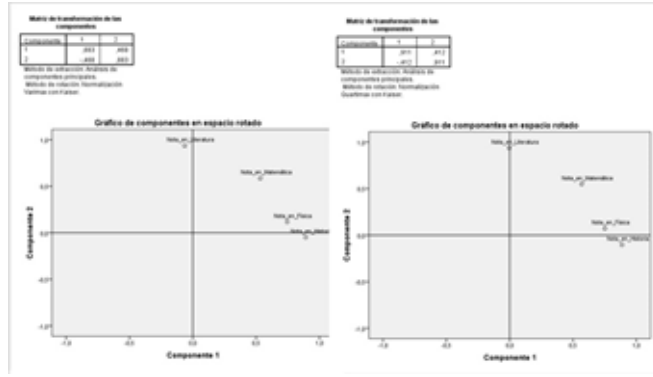
En una rotación oblicua, las comunalidades no se mantienen y la interpretación es bastante más compleja que en las rotaciones ortogonales.

Representación gráfica:



El fin de un análisis de componentes principales es conseguir reducir las variables explicativas, obtener un número de componentes menor que el de variables y dar una interpretación práctica de los mismos.

A fin de conseguir una buena interpretació n de los factores, una de las fases fundamentales del análisis factorial es la representación gráfica. La representación se hace tomando factores dos a dos y proyectando las variables sobre los planos determinados por cada par de ejes factoriales.



Las coordenadas de las variables, en el espacio definido por los componentes principales, son los coeficientes factoriales de la matriz rotada, en caso de que los ejes hayan sido rotados.

Puntuaciones factoriales individuales:

En ocasiones, puede ser interesante conocer las puntuaciones que tienen los CP para cada caso, lo cual nos permitirá entre otras cosas representar los casos en el espacio de los CP. Las puntuaciones factoriales para cada caso de la muestra pueden calcularse según la expresión:

$$F_{ij} = a_{i1}Z_{1j} + a_{i2}Z_{2j} + \cdots + a_{ik}z_{kj}$$

$$= \sum_{s=1}^{k} F_{is}z_{sj}$$

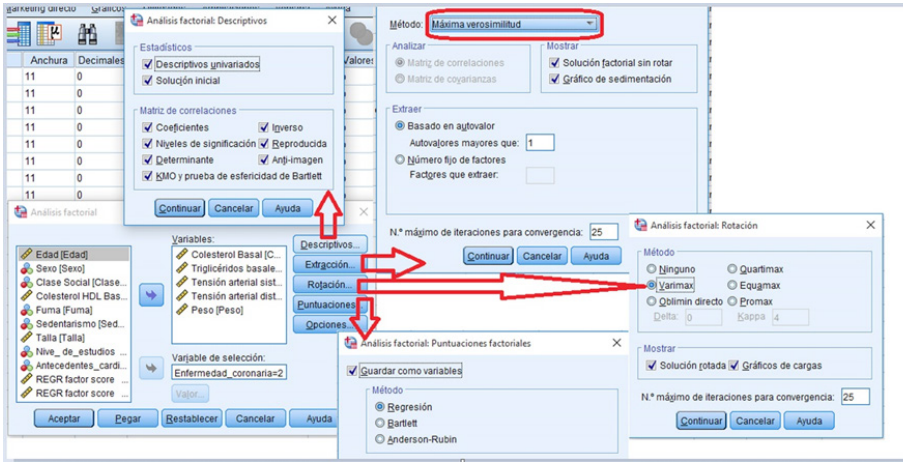
Fij representa la puntuación del i-ésimo componente, correspondiente al j-ésimo caso de la muestra, K indica el número de variables, así representa la puntuación factorial correspondiente a la e-sésima variable y al i-ésimo componente y Zsj representa el valor estandarizado de la e-sésima variable correspondiente al j-ésimo caso.

5.8. Ejemplo de análisis factorial exploratorio

Sea la base de datos ENFERMEDADES CORONARIAS (Anexo 3).

En principio, se tiene como hipótesis que existen factores comunes que pueden resumir la variabilidad de las variables asociadas a las enfermedades coronarias como son: colesterol basal, triglicéridos basales, tensión arterial sistólica, tensión arterial diastólica y peso. El fundamento de esta técnica es que el inves-

tigador cree que existen factores comunes asociados a las variables originales. En este estadio el investigador no sabe cuántos son los factores comunes; el número de factores se determinará explorando los auto valores de la matriz de correlaciones y los factores posibles, de ahí el nombre de Análisis Factorial Exploratoria. En la siguiente imagen se muestran los distintos métodos seleccionados para el procesamiento de la información.



A partir de esta selección se obtienen los siguientes resultados:

Análisis factorial exploratorio

Estadísticos descriptivos [a]			
	Media	Desviación estándar	N de análisis
Colesterol Basal	218,83	18,688	48
Triglicéridos basales	140,88	39,883	48
Tensión arterial sistólica	131,00	11,666	48
Tensión arterial diastólica	76,56	7,795	48
Peso	65,83	9,228	48

a. Solo se utilizan los casos para los cuales Enfermedad coronaria = Padece la enfermedad en la fase de análisis.

Matriz de correlaciones [a,b]

		Colesterol Basal	Triglicéridos basales	Tensión arterial sistólica	Tensión arterial diastólica	Peso
Correlación	Colesterol Basal	1,000	-,036	-,117	,140	,152
	Triglicéridos basales	-,036	1,000	,178	,313	,155
	Tensión arterial sistólica	-,117	,178	1,000	,426	,361
	Tensión arterial diastólica	,140	,313	,426	1,000	,400
	Peso	,152	,155	,361	,400	1,000

Sig. (unilateral)	Colesterol Basal		,403	,214	,171	,151
	Triglicéridos basales	,403		,114	,015	,146
	Tensión arterial sistólica	,214	,114		,001	,006
	Tensión arterial diastólica	,171	,015	,001		,002
	Peso	,151	,146	,006	,002	
a. Solo se utilizan los casos para los cuales Enfermedad coronaria = Padece la enfermedad en la fase de análisis.						
b. Determinante = ,534						

Inversión de matriz de correlaciones a					
	Colesterol Basal	Triglicéridos basales	Tensión arterial sistólica	Tensión arterial diastólica	Peso
Colesterol Basal	1,096	,090	,277	-,224	-,191
Triglicéridos basales	,090	1,120	-,031	-,333	-,044
Tensión arterial sistólica	,277	-,031	1,362	-,474	-,340
Tensión arterial diastólica	-,224	-,333	-,474	1,470	-,331
Peso	-,191	-,044	-,340	-,331	1,291

a. Solo se utilizan los casos para los cuales Enfermedad coronaria = Padece la enfermedad en la fase de análisis.

Prueba de KMO y Bartlett[a]		
Medida Kaiser-Meyer-Olkin de adecuación de muestreo		,619
Prueba de esfericidad de Bartlett	Aprox. Chi-cuadrado	27,908
	gl	10
	Sig.	,002
a. Solo se utilizan los casos para los cuales Enfermedad coronaria = Padece la enfermedad en la fase de análisis.		

Observe que por ser el KMO = 0,619, "el grado de intercorrelación es alto y el Análisis Factorial se considera útil."

Matrices anti-imagen[a]			
		Colesterol Basal	Triglicéridos basales
Covarianza anti-imagen	Colesterol _Basal	,912	,074
	Triglicéridos basales	,074	,893
	Tensión arterial sistólica	,185	-,020
	Tensión arterial diastólica	-,139	-,202
	Peso	-,135	-,030
Correlación anti-imagen	Colesterol Basal	,335b	,082
	Triglicéridos basales	,082	,671b
	Tensión arterial sistólica	,227	-,025
	Tensión arterial diastólica	-,176	-,260
	Peso	-,161	-,036

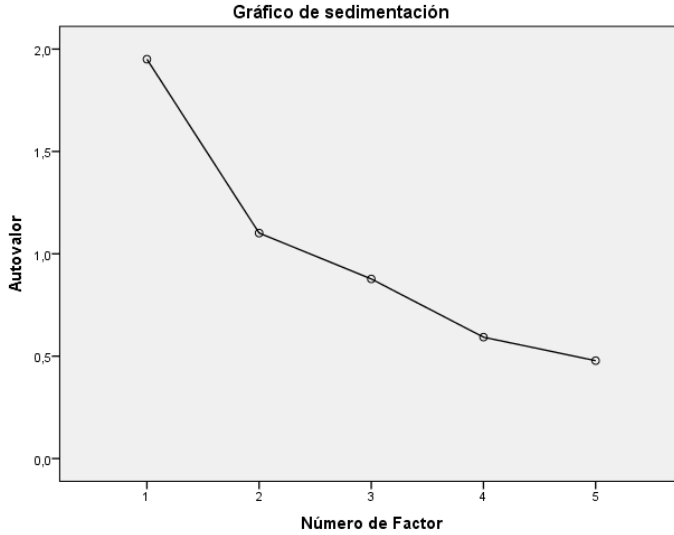
Matrices anti-imagen[a]				
		Tensión arterial sistólica	Tensión arterial diastólica	Peso
Covarianza anti-imagen	Colesterol Basal	,185	-,139	-,135
	Triglicéridos basales	-,020	-,202	-,030
	Tensión arterial sistólica	,734	-,236	-,193
	Tensión arterial diastólica	-,236	,680	-,174
	Peso	-,193	-,174	,775
Correlación anti-imagen	Colesterol Basal	,227	-,176	-,161
	Triglicéridos basales	-,025	-,260	-,036
	Tensión arterial sistólica	,608b	-,335	-,256
	Tensión arterial diastólica	-,335	,631b	-,240
	Peso	-,256	-,240	,691b
Solo se utilizan los casos para los cuales Enfermedad coronaria = Padece la enfermedad en la fase de análisis.				
b. Medidas de adecuación de muestreo (MSA)				

Comunalidades [a,b]		
	Inicial	Extracción
Colesterol Basal	,088	,999
Triglicéridos basales	,107	,135
Tensión arterial sistólica	,266	,395
Tensión arterial diastólica	,320	,551
Peso	,225	,318

Método de extracción: máxima probabilidad [a,b]
a. Solo se utilizan los casos para los cuales Enfermedad coronaria = Padece la enfermedad en la fase de análisis.
b. Se han encontrado una o más estimaciones de comunalidad mayores que 1 durante las iteraciones. La solución resultante se debe interpretar con precaución.

Varianza total explicada[a]

Factor	Autovalores iniciales			Sumas de extracción de cargas al cuadrado		
	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado
1	1,950	39,008	39,008	1,057	21,141	21,141
2	1,101	22,028	61,035	1,341	26,829	47,970
3	,877	17,546	78,582			
4	,593	11,858	90,440			
5	,478	9,560	100,000			



Varianza total explicada [a]			
Factor	Sumas de rotación de cargas al cuadrado		
	Total	% de varianza	% acumulado
1	1,345	26,893	26,893
2	1,054	21,077	47,970
3			
4			
5			
Método de extracción: máxima probabilidad. a			
Solo se utilizan los casos para los cuales Enfermedad coronaria = Padece la enfermedad en la fase de análisis.			

Estos dos factores explican en un 47,97 las otras 5 variables.

Matriz factorial a,b				
	Factor			
	1	2		
Colesterol Basal	,999	,000		
Triglicéridos basales	-,036	,366		
Tensión arterial sistólica	-,117	,617		
Tensión arterial diastólica	,141	,729		
Peso	,152	,543		
Método de extracción: máxima verosimilitud. a,b				
a. 2 factores extraídos. 5 iteraciones necesarias.				
b. Solo se utilizan los casos para los cuales Enfermedad coronaria = Padece la enfermedad en la fase de análisis.				
Prueba de bondad de ajustea				
Chi-cuadrado			gl	Sig.
,849			1	,357

a. Solo se utilizan los casos para los cuales Enfermedad coronaria = Padece la enfermedad en la fase de análisis.

Correlaciones reproducidas [a]			
		Colesterol Basal	Triglicéridos basales
Correlación reproducida	Colesterol Basal	,999[b]	-,036
	Triglicéridos basales	-,036	,135[b]
	Tensión arterial sistólica	-,117	,230
	Tensión arterial diastólica	,140	,262
	Peso	,152	,193
Residuo[c]	Colesterol Basal		-1,793E-5
	Triglicéridos basales	-1,793E-5	
	Tensión arterial sistólica	-9,559E-6	-,053
	Tensión arterial diastólica	4,601E-6	,052
	Peso	1,239E-5	-,038

Correlaciones reproducidas[a]				
		Tensión arterial sistólica	Tensión arterial diastólica	Peso
Correlación reproducida	Colesterol Basal	-,117	,140	,152
	Triglicéridos basales	,230	,262	,193
	Tensión arterial sistólica	,395[b]	,434	,317
	Tensión arterial diastólica	,434	,551[b]	,417
	Peso	,317	,417	,318[b]

Residuoc	Colesterol Basal	-9,559E-6	4,601E-6	1,239E-5
	Triglicéridos basales	-,053	,052	-,038
	Tensión arterial sistólica		-,008	,044
	Tensión arterial diastólica	-,008		-,018
	Peso	,044	-,018	

Método de extracción: máxima probabilidad. [a]

Solo se utilizan los casos para los cuales Enfermedad coronaria = Padece la enfermedad en la fase de análisis.

b. Comunalidades reproducidas

c. Los residuos se calculan entre las correlaciones observadas y reproducidas. Existen 2 (20,0%) residuos no redundantes con valores absolutos mayores que 0,05.

Matriz de factor rotado[a,b]

	Factor	
	1	2
Colesterol Basal	,016	,999
Triglicéridos basales	,365	-,042
Tensión arterial sistólica	,615	-,127
Tensión arterial diastólica	,731	,129
Peso	,546	,143

Método de extracción: máxima verosimilitud.

Método de rotación: Varimax con normalización Kaiser.[a,b]

a. La rotación ha convergido en 3 iteraciones.

Solo se utilizan los casos para los cuales Enfermedad coronaria = Padece la enfermedad en la fase de análisis.

Matriz de transformación factorial[a]

Factor	1	2

1	,017	1,000
2	1,000	-,017
Método de extracción: máxima verosimilitud. Método de rotación: Varimax con normalización Kaiser.[a]		
a. Solo se utilizan los casos para los cuales Enfermedad coronaria = Padece la enfermedad en la fase de análisis.		

Matriz de coeficiente de puntuación factorial [a]		
	Factor	
	1	2
Colesterol Basal	-,047	1,000
Triglicéridos basales	,124	-,002
Tensión arterial sistólica	,300	-,005
Tensión arterial diastólica	,477	-,008

Peso	,234	-,004
Método de extracción: máxima verosimilitud. Método de rotación: Varimax con normalización Kaiser. Método de puntuaciones factoriales: Regresión. [a]		
a. Solo se utilizan los casos para los cuales Enfermedad coronaria = Padece la enfermedad en la fase de análisis.		

Matriz de covarianzas de puntuación factorial[a]		
Factor	1	2
1	,706	,005
2	,005	,999
Método de extracción: máxima verosimilitud. Método de rotación: Varimax con normalización Kaiser. Método de puntuaciones factoriales: Regresión.[a]		
a. Solo se utilizan los casos para los cuales Enfermedad coronaria = Padece la enfermedad en la fase de análisis.		

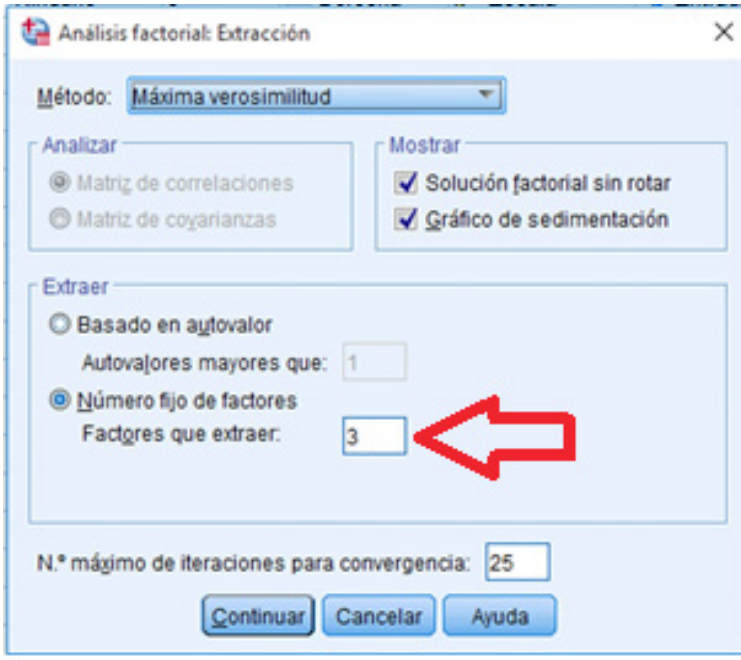
5.9. Ejemplo de análisis factorial confirmatorio

En el análisis factorial confirmatorio, el investigador además de una hipótesis previa sobre la existencia de factores comunes, tiene otra sobre el número de factores.

Ejemplo:

A partir del ejemplo antes analizado el investigador tiene evidencia que, si a las variables asociadas a las enfermedades coronarias se añade la edad y la talla, entonces con tres factores es suficiente para explicar todas las variables consideradas.

Para ello, además de incrementar las variables, debe modificar el cuadro de diálogo adjunto y escribir el número de factores a extraer.



Con esta modificación se obtiene el siguiente resultado:

Varianza total explicada[a]

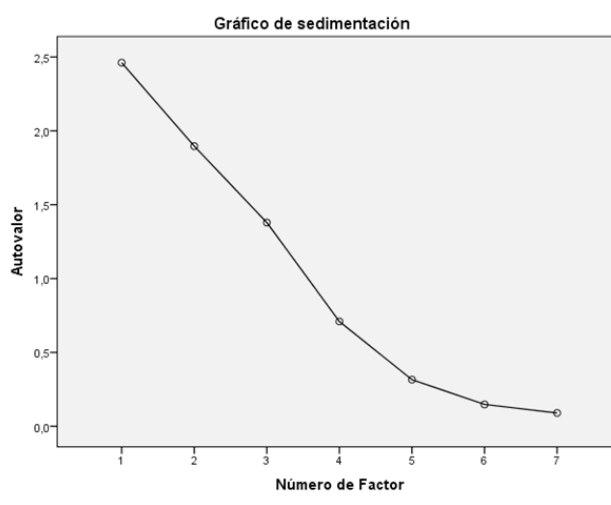
Factor	Autovalores iniciales			Sumas de extracción de cargas al cuadrado		
	Total	% de varian-za	% acu-mulado	Total	% de varian-za	% acu-mulado
1	2,461	35,163	35,163	2,077	29,676	29,676
2	1,896	27,090	62,253	1,502	21,460	51,137

3	1,379	19,703	81,956	1,605	22,928	74,065
4	,709	10,135	92,091			
5	,316	4,509	96,600			
6	,148	2,113	98,713			
7	,090	1,287	100,000			

Varianza total explicada [a]			
	Sumas de rotación de cargas al cuadrado		
Factor	Total	% de varianza	% acumulado
1	1,944	27,770	27,770
2	1,746	24,949	52,719
3	1,494	21,346	74,065
4			
5			
6			
7			
Método de extracción: máxima probabilidad. [a]			
a. Solo se utilizan los casos para los cuales Enfermedad coronaria = Si en la fase de análisis.			

Observe que ahora estos tres factores explican el 74,065% de la varianza, superior al 47,970% del ejemplo anterior del análisis factorial exploratorio.

El gráfico de sedimentación muestra tres factores con auto valores con valores superiores a 1.

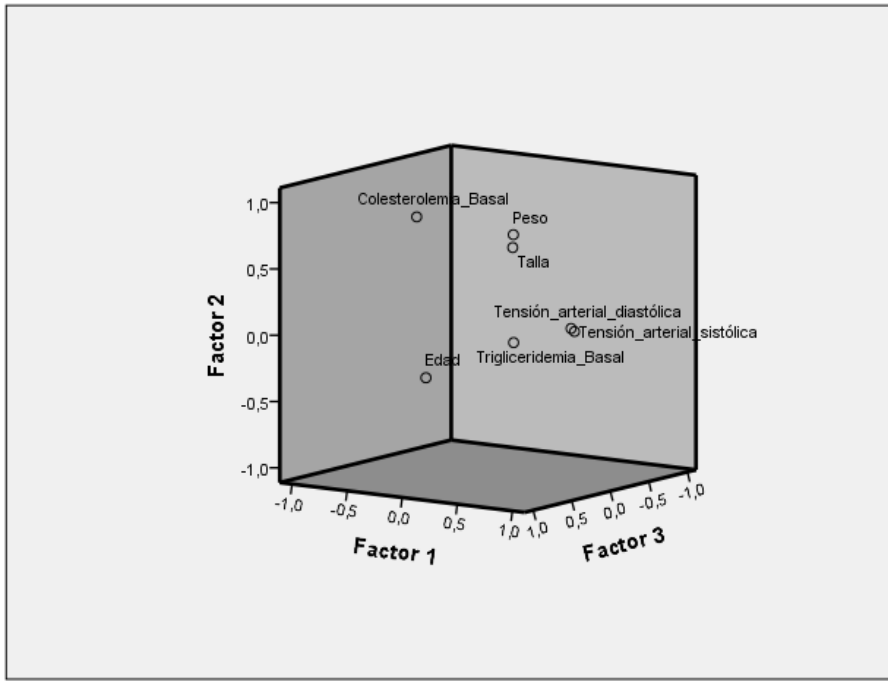


Matriz factorial[a,b]			
	Factor		
	1	2	3
Colesterol Basal	-,308	,900	,307
Tensión arterial sistólica	,689	-,220	,479
Tensión arterial diastólica	,856	-,100	,507
Peso	-,111	,292	,664
Edad	,739	,491	-,461
Triglicérido Basal	-,464	-,465	-,038
Talla	,044	,303	,608
Método de extracción: máxima verosimilitud. [a,b]			
a. 3 factores extraídos. 8 iteraciones necesarias.			
b. Solo se utilizan los casos para los cuales Enfermedad coronaria = Si en la fase de análisis.			

Matriz de factor rotado[a,b]			
	Factor		
	1	2	3
Clesterol Basal	-,425	,847	,316
Tensión arterial sistólica	,858	,081	,102
Tensión arterial diastólica	,950	,139	,279
Peso	,153	,710	-,111
Edad	,116	-,217	,969
Triglicérido Basal	-,184	-,208	-,596
Talla	,235	,639	,013
Método de extracción: máxima verosimilitud. Método de rotación: Varimax con normalización Kaiser.[a,b]			
a. La rotación ha convergido en 5 iteraciones.			
b. Solo se utilizan los casos para los cuales Enfermedad coronaria = Si en la fase de análisis.			

Matriz de transformación factorial [a]			
Factor	1	2	3
1	,750	-,220	,623
2	-,396	,605	,690
3	,529	,765	-,367
Método de extracción: máxima verosimilitud. Método de rotación: Varimax con normalización Kaiser.[a]			
a. Solo se utilizan los casos para los cuales Enfermedad coronaria = Si en la fase de análisis.			

Gráfico de factor en espacio de factores rotados



16	FAC1_1	Numérico	11	5	REGR factor score 1 for analysis 1	Ninguno	Ninguno	13	Derecha	Escala
17	FAC2_1	Numérico	11	5	REGR factor score 2 for analysis 1	Ninguno	Ninguno	13	Derecha	Escala
18	FAC3_1	Numérico	11	5	REGR factor score 3 for analysis 1	Ninguno	Ninguno	13	Derecha	Escala
19	FAC1_2	Numérico	11	5	REGR factor score 1 for analysis 2	Ninguno	Ninguno	13	Derecha	Escala

FAC1_1	FAC2_1	FAC3_1
,71160	1,37328	-2,83379
-,75471	-1,25797	,56400
-1,16531	-,48195	-2,60989
-1,62187	-1,30423	,16240
,67675	,32536	-,31715
-1,94777	-,64048	-1,32141
-1,20117	-1,42310	-,60902
-2,16056	-,89999	-,00638
,96902	,36423	,55882

El procesamiento de la información incorpora al fichero tres nuevas variables como se muestra a continuación:

Capítulo VI. El análisis discriminante y la regresión logística

6.1. El Análisis Factorial Discriminante

Un problema que conduce al análisis discriminante se da al intentar elegir una técnica analítica apropiada para resolver problemas en los que aparece una variable dependiente categórica y varias variables independientes métricas.

Por ejemplo, si se desea distinguir entre riesgo de crédito alto y bajo. Si tuviéramos una medida métrica del riesgo de crédito, se podría utilizar la regresión multivariante. Pero puede ocurrir que solo se pueda conocer si alguien se encuentra en una categoría de riesgo bueno o malo. Esta no es la medida de tipo métrico requerida para el análisis de regresión múltiple. A problemas de este tipo la estadística multivariada tiene dos respuestas:

a. El análisis discriminante.

b. La regresión logística.

Ambas son las técnicas estadísticas apropiadas cuando la variable dependiente es categórica (nominal o no métrica) y las variables independientes son métricas, pero en muchos casos, la variable dependiente consta de dos grupos o clasificaciones, por ejemplo, masculino frente a femenino o alto frente a bajo; en otras situaciones, se incluyen más de dos casos, como en una clasificación de tres grupos que comprenda clasificaciones bajas, medias y altas.

Lo esencial del análisis discriminante es que cuando se dispone de dos o más grupos de elementos, de los cuales se conocen los datos correspondientes a varias variables numéricas, se plantean los dos problemas siguientes:

A) Explicar la pertenencia de un elemento a un grupo determinado, en función de los valores de las variables disponibles ¿Qué

variables explican la clasificación en grupos distintos? ¿Cuáles de estas variables son más importantes en la discriminación?

B) Predecir a qué grupo pertenece o pertenecerá un elemento del que se conocen los valores de una serie de variables.

Según cuál sea el interés, se utilizarán uno de los dos métodos siguientes:

- Análisis factorial discriminante (AFD).
- Funciones discriminantes.

Tiene como objetivo primordial explicar la pertenencia de un individuo a un determinado grupo. Este método también permite realizar predicciones, asignando a cada individuo al grupo más cercano a su puntuación factorial, pero el método de las funciones discriminantes es más potente en cuanto a predicciones.

6.2. Funciones discriminantes

Este método pretende predecir la pertenencia de un individuo a un determinado grupo, en base a la probabilidad calculada, conocidos una serie de datos. El método de las funciones discriminantes calcula las probabilidades de pertenecer a un determinado grupo según técnicas de decisión bayesianas[xxvi].

$$P(G_i/D) = \frac{P(D/G_i)P(G_i)}{\sum_{i=1}^{S} P(D/G_i)P(G_i)}$$

En muchos casos prácticos se utilizan los dos métodos. Primero el Análisis factorial discriminante para determinar las variables explicativas, y después el método de las funciones discriminantes, a fin de calcular las probabilidades de pertenecer a un grupo, según los valores de una serie de variables.

El método de Análisis Discriminante permite obtener un valor teórico, es decir, una combinación lineal de dos (o más) varia-

bles independientes que discrimine mejor entre los grupos definidos a priori. La discriminación se lleva a cabo estableciendo las ponderaciones del valor teórico para cada variable de tal forma que maximicen la varianza entre grupos frente a la varianza intragrupos. La combinación lineal para el análisis discriminante, también conocida como función discriminante, se deriva de una ecuación que adopta la siguiente forma:

$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + W_3 X_{3k} + \cdots + W_n X_{nk}$$

z_{jk} =puntuación z discriminante de la función discriminante j para el objeto k

a = constante

w_i = ponderación discriminante para la variable independiente i

X_k = variable independiente i para el objeto k

Para ejemplificar lo explicado se puede tomar la base datos ya utilizada en el análisis factorial y en ella se puede observar que ninguna variable predictora de la enfermedad (Colesterol Basal, Colesterol HDL Basal, Triglicérido Basal, Tensión arterial sistólica, o Tensión arterial diastólica) por sí sola permite a un cardiólogo diagnosticarle a un paciente que padece una enfermedad coronaria, pero una combinación lineal de todos estos indicadores permite hacerlo.

La combinación lineal antes referida se convierte en una función discriminante, de modo que el valor de la función discriminante para un individuo determinado se calcula sustituyendo los valores correspondientes a las variables de cada individuo en la función discriminante. Al valor obtenido se le denomina puntuación discriminante.

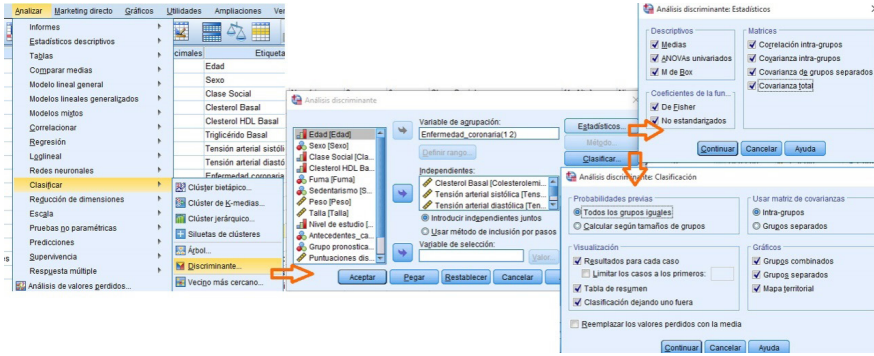
Con las puntuaciones discriminantes de todos los individuos enfermos es posible calculas media y también la de los individuos no enfermos y con esta información determinar si la puntuación de un individuo en particular está *más próxima* al valor medio del grupo

de sanos o del valor medio del grupo de enfermos y con ellos hacer una valoración de cuan acertado han sido los diagnósticos.

En el párrafo anterior se escribió con toda intencionalidad la expresión *más próxima* porque la proximidad de la combinación lineal de los parámetros de un individuo a la combinación lineal del promedio de las personas sanas o enfermas depende de la distancia que se tome; en este caso se toma la distancia de Mahalanobis, la cual se define por la expresión:

$$H^2{}_{ab} = (n-g)\sum_{i=1}^{p}\sum_{j=1}^{p} w_{ij}{}^{*}\left(\bar{X}_i^{(a)} - \bar{X}_i^{(b)}\right)\left(\bar{X}_j^{(a)} - \bar{X}_j^{(b)}\right)$$

Donde n es el número de casos válidos, *g* es el número de grupos, $\bar{X}_i^{(a)}$ es la media del grupo a en la i-ésima variable independiente, es la media del grupo b en la i-ésima variable independiente, y $w_{ij}{}^{*}$ es un elemento de la inversa de la matriz de varianza-covarianza intra-grupos. Esta distancia fue introducida por Mahalanobis[xxvii] en 1936. Su utilidad radica en determinar la similitud entre dos variables aleatorias multidimensionales; difiere de la distancia euclídea en que tiene en cuenta la correlación entre las variables aleatorias. El procesamiento del ejemplo con SPSS se desarrolla según la siguiente secuencia de imágenes:



La ejecución arroja los siguientes resultados

Estadísticas de grupo					
Enfermedad coronaria		Media	Desviación estándar No ponderados	N válido (por lista)	
				Ponderados	
Si	Colesterol Basal	291,95	79,214	19	19,000
	Triglicérido Basal	139,42	39,928	19	19,000
	Tensión arterial sistólica	163,95	14,774	19	19,000
	Tensión arterial diastólica	94,84	7,182	19	19,000
No	Colesterol Basal	219,73	17,537	48	48,000
	Triglicérido Basal	138,79	37,452	48	48,000
	Tensión arterial sistólica	130,58	11,694	48	48,000
	Tensión arterial diastólica	76,67	7,875	48	48,000
Total	Colesterol Basal	240,21	54,827	67	67,000
	Triglicérido Basal	138,97	37,865	67	67,000
	Tensión arterial sistólica	140,04	19,660	67	67,000
	Tensión arterial diastólica	81,82	11,241	67	67,000

Prueba de igualdad de medias de grupos					
	Lambda de Wilks	F	gl1	gl2	Sig.
Colesterol Basal	,642	36,220	1	65	,000
Triglicérido Basal	1,000	,004	1	65	,952
Tensión arterial sistólica	,406	95,098	1	65	,000
Tensión arterial diastólica	,461	76,052	1	65	,000

Matrices dentro de grupos combinados [a]				
		Colesterol Basal	Triglicérido Basal	Tensión arterial sistólica
Covarianza	Colesterol Basal	1960,037	-206,589	-97,407
	Triglicérido Basal	-206,589	1455,732	-13,965
	Tensión arterial sistólica	-97,407	-13,965	159,333
	Tensión arterial diastólica	-14,223	35,460	51,233
Correlación	Colesterol Basal	1,000	-,122	-,174
	Triglicérido Basal	-,122	1,000	-,029
	Tensión arterial sistólica	-,174	-,029	1,000
	Tensión arterial diastólica	-,042	,121	,528

Matrices dentro de grupos combinados [a]	
	Tensión arterial diastólica

Covarianza	Clesterol Basal	-14,223	
	Triglicérido Basal	35,460	
	Tensión arterial sistólica	51,233	
	Tensión arterial diastólica	59,126	
Correlación	Clesterol Basal	-,042	
	Triglicérido Basal	,121	
	Tensión arterial sistólica	,528	
	Tensión arterial diastólica	1,000	

a. La matriz de covarianzas tiene 65 grados de libertad.

Matrices de covarianzas [a]			
Enfermedad coronaria	Colesterol Basal	Triglicéri-do Basal	Tensión arterial sistólica

Si	Colesterol Basal	6274,830	-907,588	-308,392
	Triglicérido Basal	-907,588	1594,257	-221,477
	Tensión arterial sistólica	-308,392	-221,477	218,275
	Tensión arterial diastólica	-112,509	-105,930	90,658
No	Colesterol Basal	307,563	61,879	-16,605
	Triglicérido Basal	61,879	1402,679	65,507
	Tensión arterial sistólica	-16,605	65,507	136,759
	Tensión arterial diastólica	23,418	89,610	36,135
Total	Colesterol Basal	3005,986	-194,085	401,006
	Triglicérido Basal	-194,085	1433,757	-9,423
	Tensión arterial sistólica	401,006	-9,423	386,498
	Tensión arterial diastólica	256,705	37,282	175,523

Matrices de covarianzas [a]		
Enfermedad coronaria		Tensión arterial diastólica
Si	Colesterol Basal	-112,509
	Triglicérido Basal	-105,930
	Tensión arterial sistólica	90,658
	Tensión arterial diastólica	51,585

No	Colesterol Basal	23,418
	Triglicérido Basal	89,610
	Tensión arterial sistólica	36,135
	Tensión arterial diastólica	62,014
Total	Colesterol Basal	256,705
	Triglicérido Basal	37,282
	Tensión arterial sistólica	175,523
	Tensión arterial diastólica	126,361

a. La matriz de covarianzas total tiene 66 grados de libertad.

Análisis 1

Prueba de Box de la igualdad de matrices de covarianzas

Logaritmo de los determinantes		
Enfermedad coronaria	Rango	Logaritmo del determinante
Si	4	23,698
No	4	21,696
Dentro de grupos combinados	4	23,607

Los logaritmos naturales y los rangos de determinantes impresos son los de las matrices de covarianzas de grupo.

Resultados de prueba		
M de Box		88,131
F	Aprox.	8,021
	gl1	10
	gl2	5581,110
	Sig.	,000
Prueba la hipótesis nula de las matrices de covarianzas de población iguales.		

Resumen de funciones discriminantes canónicas

Autovalores				
Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	2,632a	100,0	100,0	,851
a. Se utilizaron las primeras 1 funciones discriminantes canónicas en el análisis.				

Lambda de Wilks				
Prueba de funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.

1	,275	81,261	4	,000

Coeficientes de función discriminante canónica estandarizados		
1	Función	
Clesterol Basal	,599	
Triglicérido Basal	,058	
Tensión arterial sistólica	,680	
Tensión arterial diastólica	,326	
Matriz de estructuras		
	Función	
	1	
Tensión arterial sistólica	,746	
Tensión arterial diastólica	,667	
Colesterol Basal	,460	
Triglicérido Basal	,005	

Correlaciones dentro de grupos combinados entre las variables discriminantes y las funciones discriminantes canónicas estandarizadas

Variables ordenadas por el tamaño absoluto de la correlación dentro de la función.

Coeficientes de la función discriminante canónica	
	Función
	1
Clesterol Basal	,014
Triglicérido Basal	,002
Tensión arterial sistólica	,054

Tensión arterial diastólica	,042
(Constante)	-14,472
Coeficientes no estandarizados	

Funciones en centroides de grupo	
	Función
Enfermedad coronaria	1
Si	2,540
No	-1,005
Las funciones discriminantes canónicas sin estandarizar se han evaluado en medias de grupos	

Estadísticas de clasificación

Resumen de proceso de clasificación		
Procesado		70
Excluido	Códigos de grupo perdidos o fuera de rango	0
	Como mínimo, falta una variable discriminatoria	0
Utilizado en resultado		70

Probabilidades previas para grupos			
Enfermedad coronaria	Previa	Casos utilizados en análisis	
		No ponderados	Ponderados
Si	,500	19	19,000
No	,500	48	48,000

Total	1,000	67	67,000

Coeficientes de función de clasificación		
	Enfermedad coronaria	
	Si	No
Colesterol Basal	,212	,164
Triglicérido Basal	,115	,110
Tensión arterial sistólica	,913	,722
Tensión arterial diastólica	,795	,644
(Constante)	-152,264	-98,233

Esta ventana muestra las funciones usadas para clasificar observaciones. Hay una función para cada uno de los 2 niveles de Enfermedad coronaria. Por ejemplo, la función usada para el primer nivel de Enfermedad coronaria es

-152,264+ 0,2120* Colesterol Basal + 0,795*Tensión arterial diastólica + 0,913*Tensión arterial sistólica + 0,115* Triglicérido Basal

Se utilizan estas funciones para predecir a que nivel de Enfermedad coronaria pertenecen las nuevas observaciones.

Estadísticas por casos								
	Número del caso	Número de predictores con valores perdidos	Grupo real	Grupo superi				
				Grupo pronosticado	P(D>d | G=g)		P(G=g | D=d)	
					p	gl		
Original	1		1	1	,578	1	1,000	
	2		2	1**	,083	1	,533	
	3		2	2	,587	1	,987	
	4		2	2	,705	1	,993	
	5		1	1	,362	1	1,000	
	6		2	2	,745	1	,994	
	7		1	2**	,387	1	,962	
	8		2	2	,429	1	,970	
	9		1	1	,204	1	1,000	
	10	1	2	2	,936	1	,999	
	11		2	2	,380	1	1,000	
	12	1	2	2	,938	1	,998	
	13		2	2	,804	1	,996	
	14		1	1	,396	1	1,000	
	15	1	1	1	,759	1	,999	
	16		1	1	,196	1	,846	
	17		2	2	,218	1	,872	
	18		2	2	,910	1	,999	
	19		2	2	,837	1	,999	
	20		2	2	,805	1	,999	
	21		2	2	,337	1	1,000	

ior		Segundo grupo superior			Puntuaciones discriminantes
	Distancia de Mahalanobis al cuadrado para centroide	Grupo	P(G=g | D=d)	Distancia de Mahalanobis al cuadrado para centroide	Función 1
	,310	2	,000	16,824	3,096
	3,011	2	,467	3,276	,805
	,295	1	,013	9,012	-,462
	,144	1	,007	10,024	-,626
	,831	2	,000	19,863	3,451
	,106	1	,006	10,367	-,680
	,748	1	,038	7,184	-,140
	,625	1	,030	7,589	-,215
	1,615	2	,000	23,197	3,811
	,006	1	,001	13,143	-1,085
	,770	1	,000	19,561	-1,883
	,006	1	,002	12,024	-,928
	,062	1	,004	10,868	-,757
	,721	2	,000	19,311	3,389
	,094	2	,001	14,841	2,847
	1,668	2	,154	5,079	1,248
	1,518	1	,128	5,351	,227
	,013	1	,001	13,385	-1,119
	,042	1	,001	14,073	-1,211
	,061	1	,001	14,384	-1,253
	,922	1	,000	20,302	-1,966

	22		1	1	,426	1	1,000	
	23		2	2	,532	1	,983	
	24		2	2	,636	1	,990	
	25		2	1**	,116	1	,670	
	26		2	2	,430	1	,970	
	27		2	2	,901	1	,999	
	28		2	2	,880	1	,999	
	29		2	2	,994	1	,998	
	30		1	2**	,851	1	,996	
	31		2	2	,195	1	1,000	
	32		2	2	,614	1	,989	
	33		2	1**	,210	1	,862	
	34		1	1	,692	1	1,000	
	35		2	2	,960	1	,998	
	36		2	2	,680	1	,992	
	37		2	2	,231	1	1,000	
	38		2	2	,655	1	,991	
	39		1	1	,709	1	,993	
	40		2	2	,511	1	1,000	
	41		2	2	,259	1	,907	
	42		2	2	,407	1	,966	
	43		2	2	,878	1	,999	
	44		2	2	,449	1	1,000	
	45		1	1	,833	1	,996	
	46		2	2	,092	1	1,000	
	47		1	1	,928	1	,997	
	48		2	2	,846	1	,999	
	49		2	2	,191	1	1,000	
	50		2	2	,744	1	,999	

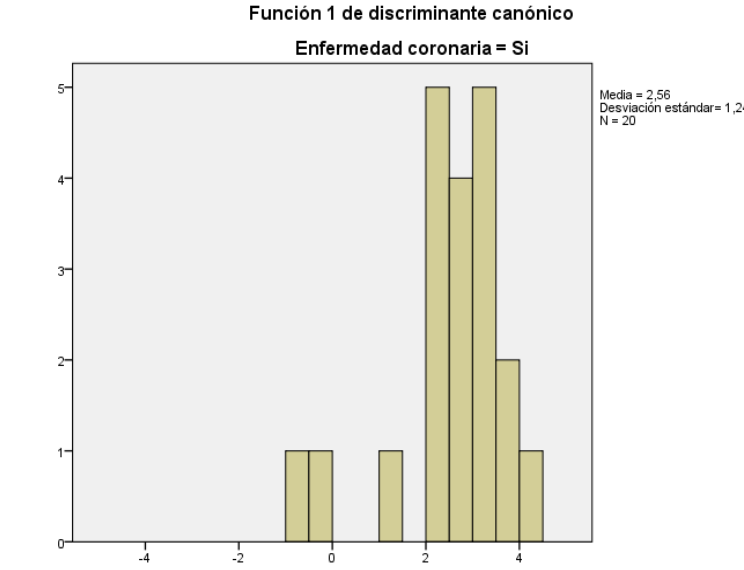
	,633	2	,000	18,846	3,336
	,390	1	,017	8,532	-,381
	,224	1	,010	9,440	-,532
	2,472	2	,330	3,893	,968
	,624	1	,030	7,594	-,216
	,015	1	,001	13,467	-1,130
	,023	1	,001	13,664	-1,156
	,000	1	,002	12,620	-1,012
	,035	1	,004	11,271	-,817
	1,679	1	,000	23,437	-2,301
	,254	1	,011	9,248	-,501
	1,575	2	,138	5,247	1,285
	,157	2	,000	15,538	2,936
	,003	1	,002	12,215	-,955
	,170	1	,008	9,817	-,593
	1,433	1	,000	22,489	-2,202
	,200	1	,009	9,597	-,558
	,139	2	,007	10,062	2,167
	,432	1	,000	17,664	-1,663
	1,275	1	,093	5,838	,124
	,686	1	,034	7,382	-,177
	,024	1	,001	13,681	-1,159
	,574	1	,000	18,514	-1,763
	,045	2	,004	11,117	2,329
	2,830	1	,000	27,330	-2,688
	,008	2	,003	11,934	2,449
	,038	1	,001	13,981	-1,199
	1,711	1	,000	23,555	-2,313
	,107	1	,001	14,993	-1,332

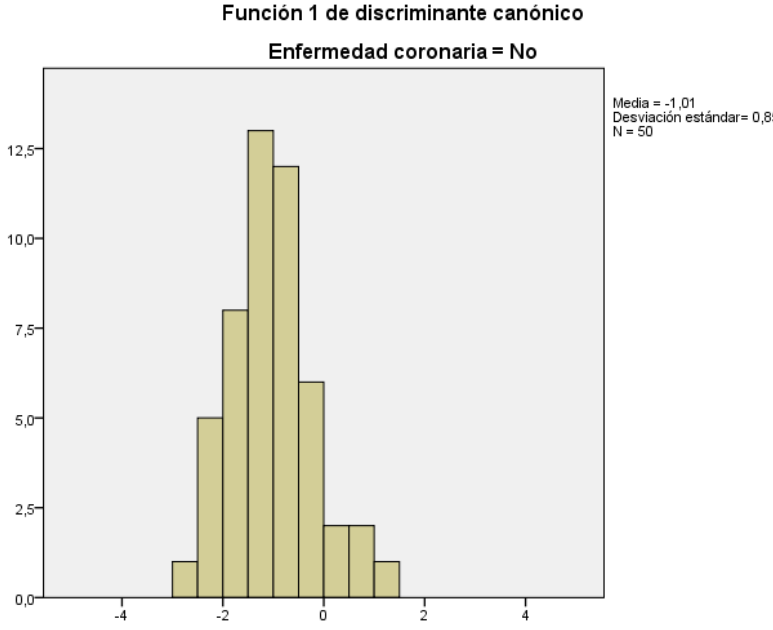
	51		1	1	,793	1	,995	
	52		1	1	,400	1	1,000	
	53		2	2	,902	1	,997	
	54		2	2	,459	1	1,000	
	55		2	2	,685	1	1,000	
	56		2	2	,673	1	1,000	
	57		2	2	,283	1	1,000	
	58		1	1	,813	1	,999	
	59		1	1	,936	1	,999	
	60		2	2	,995	1	,998	
	61		1	1	,118	1	1,000	
	62		1	1	,715	1	,993	
	63		2	2	,701	1	,993	
	64		2	2	,838	1	,996	
	65		2	2	,439	1	1,000	
	66		1	1	,258	1	1,000	
	67		2	2	,448	1	,973	
	68		2	2	,517	1	1,000	
	69		2	2	,509	1	1,000	
	70		2	2	,183	1	1,000	

**. Caso clasificado incorrectamente

	,069	2	,005	10,782	2,278
	,709	2	,000	19,251	3,382
	,015	1	,003	11,712	-,882
	,550	1	,000	18,376	-1,747
	,164	1	,000	15,607	-1,411
	,178	1	,000	15,737	-1,427
	1,154	1	,000	21,340	-2,080
	,056	2	,001	14,298	2,776
	,006	2	,001	13,142	2,620
	,000	1	,002	12,618	-1,012
	2,444	2	,000	26,099	4,103
	,133	2	,007	10,114	2,175
	,148	1	,007	9,993	-,621
	,042	1	,004	11,164	-,801
	,598	1	,000	18,653	-1,779
	1,277	2	,000	21,861	3,670
	,576	1	,027	7,763	-,246
	,420	1	,000	17,587	-1,654
	,436	1	,000	17,691	-1,666
	1,777	1	,000	23,799	-2,338

Gráficos de grupos separados





Resultados de clasificación a					
		Enfermedad coronaria	Pertenencia a grupos pronosticada		Total
			Si	No	
Original	Recuento	Si	18	2	20
		No	3	47	50
	%	Si	90,0	10,0	100,0
		No	6,0	94,0	100,0
a. 92,9% de casos agrupados originales clasificados correctamente.					

6.3. Aplicaciones del análisis discriminante

El problema de discriminación aparece en muchas situaciones en que se necesita clasificar elementos con información incompleta. Por ejemplo:

- Los sistemas automáticos de concesión de créditos (credit scoring) implantados en muchas instituciones financieras tienen que utilizar variables medibles hoy (ingresos, antigüedad en el trabajo, patrimonio, etc.) para prever el comportamiento futuro.
- En ingeniería este problema se ha estudiado con el nombre de reconocimiento de patrones (pattern recognition), para diseñar máquinas capaces de clasificar de manera automática. Por ejemplo, reconocer voces y sonidos, clasificar billetes o monedas, reconocer caracteres escritos en una pantalla de ordenador o clasificar cartas según el distrito postal.
- Asignar un texto escrito de procedencia desconocida a uno de varios autores por las frecuencias de utilización de palabras.
- Asignar una partitura musical o un cuadro a un artista, una declaración de impuestos como potencialmente defraudadora o no.
- Una empresa como en riesgo de quiebra o no.

- Las enseñanzas de un centro como teóricas y aplicadas.
- Un paciente como enfermo de cáncer o no.
- Un nuevo método de fabricación como eficaz o no.

6.4. La regresión logística

Al estudiar el análisis discriminante en el apartado anterior se dijo que es una técnica estadística apropiada cuando la variable dependiente es categórica (nominal o no métrica) y las variables independientes son métricas; bajo esta condición, en muchos casos, la variable dependiente consta de varios grupos de clasificación, pero en otros casos tan numerosos como los anteriores, solo existen dos grupos o clasificaciones, por ejemplo, masculino frente a femenino o alto frente a bajo.

El análisis discriminante tiene la capacidad de tratar tanto dos grupos como grupos múltiples (tres o más). Cuando se incluyen dos clasificaciones, la técnica es conocida como análisis discriminante de dos grupos. Cuando se identifican tres o más clasificaciones, la técnica es conocida como análisis discriminante múltiple (MDA); pero la regresión logística, que se estudiará en este epígrafe, también conocida como análisis logit, está restringida en su forma básica a dos grupos, aunque en formulaciones alternativas muy específicas puede considerar más de dos grupos.

De lo expresado se puede concluir que la regresión logística es un tipo especial de regresión que se utiliza para predecir y explicar una variable categórica binaria (dos grupos) en lugar de una medida dependiente métrica y su valor teórico es similar a la del valor teórico en la regresión múltiple, de ahí que cuando se conocen los supuestos básicos de ambas, estas técnicas proporcionan resultados predictivos y clasificatorios comparables y emplean medidas de validación similares.

Pese a las similitudes expuestas, la regresión logística tiene la ventaja de verse menos afectada que el análisis discriminante cuando no se cumplen los supuestos básicos, concretamente la

normalidad de las variables. Además, posibilita el empleo de variables no métricas por medio de su codificación con variables ficticias, tal como puede hacerse en la regresión. La regresión logística está limitada, sin embargo, a la predicción de tan solo la medida dependiente de dos grupos. Por tanto, como se ha dicho, cuando la medida de la variable dependiente está formada por dos o más grupos, lo adecuado es aplicar un análisis discriminante.

Se reitera que la regresión logística supone una alternativa respecto al análisis discriminante que puede resultar más «cómoda» a muchos investigadores debido a sus parecidos con la regresión múltiple. Su robustez frente a condiciones en los datos que pueden afectar negativamente al análisis discriminante como puede ser la regresión logística resulta la técnica de estimación preferida por diferentes grupos de investigadores.

Aunque se ha dicho que la regresión logística es semejante a la regresión múltiple; su principal diferencia radica en que, en la logística, la variable dependiente suele ser binaria (es decir, toma solo dos valores posibles), en tanto que, en la múltiple, esa variable dependiente es continua.

La regresión logística tiene la ventaja de verse menos afectada que el análisis discriminante cuando no se cumplen supuestos básicos como la normalidad de las variables, pudiendo además empelar variables no métricas por medio de su codificación con variables ficticias, tal como puede hacerse en la regresión. La regresión logística está limitada, sin embargo, a la predicción de tan solo la medida dependiente de dos grupos.

6.5. El modelo de regresión logística

Un modelo de regresión con variable dependiente binomial (modelo logístico o modelo de regresión logística) será un modelo que permita estudiar si dicha variable discreta depende o no, de otra u otras variables. Si una variable binomial de parámetro p es independiente de otra variable X, se cumple $(p|X=x) = p$, para cualquier valor x de la variable X.

Este modelo se materializa en una función en la que p aparece dependiendo de X y de unos coeficientes cuya investigación permite abordar la relación de dependencia. Para una única variable independiente X, el modelo de regresión logística toma la forma:

$$\ln\left({}^{p}/_{q}\right) = a_0 + a_1 X$$

$$\ln\left({}^{p}/_{q}\right) = a_0 + a_1 X_1 + a_2 X_2 + a_3 X_3 \ldots + a_n X_n$$

$$\ln\left({}^{p}/_{q}\right) = a_0 + a_1 X \Leftrightarrow ln\left(\frac{p}{1-p}\right) = a_0 + a_1 X \Leftrightarrow \frac{p}{1-p} = e^{a_0 + a_1 X} \Leftrightarrow p$$

$$= \frac{e^{a_0 + a_1 X}}{1 + e^{-(a_0 + a_1 X)}} \Leftrightarrow p = \frac{1}{1 + e^{-(a_0 + a_1 X)}}$$

$$p = \frac{1}{1 + e^{-(a_0 + a_1 X_1 + a_2 X_2 + a_3 X_3 \ldots + a_n X_n)}}$$

6.6. Ejemplo de aplicación de la regresión logística

Continuando con la base de datos sobre enfermedades coronarias es posible determinar los factores (variables independientes) que inciden sobre la enfermedad (variable dependiente), tanto los dados en forma numérica como categóricas, las seleccionadas para el ejemplo a desarrollar son: Sexo, Triglicérido Basal, Tensión arterial sistólica, Tensión arterial diastólica, Fuma, Sedentarismo y Peso; a partir de ellas se plantean don hipótesis:

H0. Las variables independientes no influyen significativamente sobre la variable dependiente.

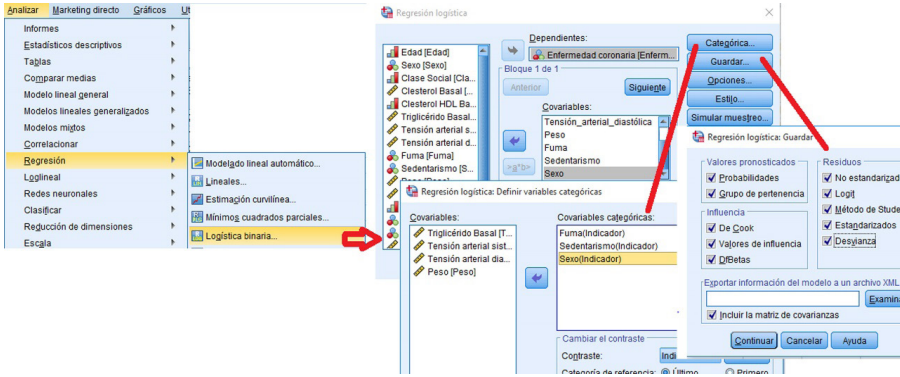
H1. Las variables independientes influyen significativamente sobre la variable dependiente.

Lo expresado en las hipótesis se trascribe a términos estadísticos que es necesario encontrar parámetros que puedan ser evaluados para que, según los valores obtenidos, sea posible rechazar o no la hipótesis nula, para ello el modelo de regresión logística simple es válido si a1 es significativamente distinto de

cero; si se remite a la fórmula del modelo se verá que a1 es el coeficiente de regresión logística muestral y es un estimador de A1 que es el coeficiente de regresión logística poblacional. El que a1 sea significativamente distinto de cero indica que es muy poco probable que A1 sea cero.

Las hipótesis operativas son las siguientes: H0. $A_1 = 0$; H1. $A_1 \neq 0$

La siguiente imagen ilustra el proceso se selección de las opciones en SPSS.



6.7. Métodos de selección de variables en el análisis de regresión logística

Otro aspecto a tener en cuenta es la selección del método para procesar la información y especificar cómo se introducen las variables independientes en el análisis. Utilizando distintos métodos se pueden construir diversos modelos de regresión a partir del mismo conjunto de variables. Según la ayuda del SPSS los principales métodos utilizados son:

- *Intro*. Procedimiento para la selección de variables en el que todas las variables de un bloque se introducen en un solo paso.
- *Selección hacia adelante (Condicional)*. Método de selección por pasos que contrasta la entrada basándose en la

significación del estadístico de puntuación y contrasta la eliminación basándose en la probabilidad de un estadístico de la razón de verosimilitud que se basa en estimaciones condicionales de los parámetros.

- *Selección hacia adelante (razón de verosimilitud)*. Método de selección por pasos hacia adelante que contrasta la entrada basándose en la significación del estadístico de puntuación y contrasta la eliminación basándose en la probabilidad del estadístico de la razón de verosimilitud, que se basa en estimaciones de la máxima verosimilitud parcial.
- *Selección hacia adelante (Wald)*. Método de selección por pasos hacia adelante que contrasta la entrada basándose en la significación del estadístico de puntuación y contrasta la eliminación basándose en la probabilidad del estadístico de Wald.
- *Eliminación hacia atrás (Condicional)*. Selección por pasos sucesivos hacia atrás El contraste para la eliminación se basa en la probabilidad del estadístico de la razón de verosimilitud, el cual se basa a su vez en las estimaciones condicionales de los parámetros.
- *Eliminación hacia atrás (razón de verosimilitud)*. Selección por pasos sucesivos hacia atrás El contraste para la eliminación se fundamenta en la probabilidad del estadístico de la razón de verosimilitud, el cual se fundamenta en estimaciones de máxima verosimilitud parcial.
- *Eliminación hacia atrás (Wald)*. Selección por pasos sucesivos hacia atrás El contraste para la eliminación se basa en la probabilidad del estadístico de Wald.

6.8. Resultados de la aplicación del método

Codificaciones de variables categóricas			
		Frecuencia (1)	Codificación de parámetro
Sexo	Masculino	33	1,000
	Femenino	33	,000

Sedentarismo	Si	30	1,000
	No	36	,000
Fuma	Si	33	1,000
	No	33	,000

Bloque 0: Bloque de inicio

En este bloque de inicio se calcula la verosimilitud de un modelo que solo tiene el término constante (a ó b0). Puesto que la verosimilitud L es un número muy pequeño (comprendido entre 0 y 1), generalmente se da el logaritmo neperiano de la verosimilitud (LL), que es un número negativo, o dos veces el logaritmo neperiano de la verosimilitud (-2LL), que es un número positivo.

Historial de iteraciones [a,b,c]			
		Logaritmo de la verosimilitud -2	Coeficientes
Iteración		Constante	
Paso 0	1	79,277	,848
	2	79,232	,905
	3	79,232	,906
a. La constante se incluye en el modelo.			
b. Logaritmo de la verosimilitud -2 inicial: 79,232			
c. La estimación ha terminado en el número de iteración 3 porque las estimaciones de parámetro han cambiado en menos de ,001.			

El estadístico -2LL mide hasta qué punto un modelo se ajusta bien a los datos. El resultado de esta medición recibe también el nombre de *desviación*. Cuanto más pequeño sea el valor, mejor será el ajuste. En este primer paso solo se ha introducido el término constante en el modelo; en la tabla se muestra un resumen del proceso iterativo de estimación del primer parámetro (b0). El proceso ha necesitado tres ciclos para estimar correctamente el término constante, porque la variación de -2LL entre el segundo y tercer bucle ha cambiado en menos del criterio fijado por el programa (0,001). También nos muestra el valor del parámetro calculado (b0 = 0,906).

Tabla de clasificación [a,b]					
	Observado		Pronosticado		
		Enfermedad coronaria		Porcentaje correcto	
		Si	No		
Paso 0	Enfermedad coronaria	Si	0	19	,0
		No	0	47	100,0
	Porcentaje global				71,2
a. La constante se incluye en el modelo.					
b. El valor de corte es ,500					

En la tabla anterior se presenta la clasificación de los casos según su ocurrencia y según la predicción realizada en función del modelo nulo. Como puede observarse, habría un 100% de acierto del pronóstico de no enfermos y ningún acierto en el pronóstico de no enfermos, por lo cual en este primer modelo han sido correctamente clasificados el 71,2% de los casos.

Variables en la ecuación							
		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	,906	,272	11,099	1	,001	2,474
Las variables no están en la ecuación							
				Puntuación	gl	Sig.	
Paso 0	Variables	Triglicérido Basal		,003	1	,959	
		Tensión arterial diastólica		36,110	1	,000	
		Peso		20,847	1	,000	
		Sedentarismo(1)		12,072	1	,001	
		Fuma(1)		16,629	1	,000	
	Estadísticos globales			40,550	5	,000	

En la tabla se presentan los parámetros del modelo nulo: B o constante, el error estándar correspondiente, el estadístico Wald28, los grados de libertad del estadístico, el nivel de significación y el Exponencial de B. El estadístico Wald es significativo, es decir que B difiere significativamente de 0 y por lo tanto produce cambio sobre la variable dependiente.

En la ecuación de regresión solo aparece, en este primer bloque, la constante, habiendo quedado fuera todas las variables. Sin embargo, como se verá en la subtabla inferior, por tener una significación estadística asociada al índice de Wald de 0,000, el proceso automático por pasos continuará, incorporándola a la ecuación.

Bloque 1: Método = Avanzar por pasos (razón de verosimilitud)

En este bloque se emplea el criterio de la razón de la verosimilitud (RV) para contrastar las nuevas variables a introducir o sacar del modelo.

Historial de iteraciones a,b,c,d				
Iteración		Logaritmo de la verosimilitud -2 Constante	Coeficientes	
			Tensión arterial diastólica	
Paso 1	1	44,027	10,600	-,119
	2	36,225	16,836	-,189
	3	34,458	21,346	-,238
	4	34,283	23,298	-,260
	5	34,280	23,575	-,263
	6	34,280	23,580	-,263
	7	34,280	23,580	-,263
a. Método: Avanzar por pasos (razón de verosimilitud)				
b. La constante se incluye en el modelo.				
c. Logaritmo de la verosimilitud -2 inicial: 79,232				
d. La estimación ha terminado en el número de iteración 7 porque las estimaciones de parámetro han cambiado en menos de ,001.				

En la primera tabla se muestra el proceso de iteración, que ahora se realiza para varios coeficientes, la constante (ya incluida en el anterior paso) y las variables numéricas y las categóricas. Obsérvese la disminución del -2LL respecto al paso anterior (el modelo solo con la constante tenía un valor de este estadístico de 79,232, mientras que ahora se reduce a 34,280), y el proceso termina con 7 iteraciones. Los coeficientes calculados son para la constante b_0= 23,580 y para la variable Tensión arterial diastólica b_1= -0,263

La ecuación del modelo ajustado es

$$\text{Enfermedad coronaria} = \frac{e^{23,580-0,263*\text{Tensión arterial diastólica}}}{(1+e^{23,580-0,263*\text{Tensión arterial diastólica}})}$$

Las tablas que siguen aportan información sobre el ajuste del modelo con estas estimaciones. La probabilidad de los resultados observados en el estudio, dadas las estimaciones de los parámetros, es lo que se conoce por verosimilitud; como se ha dicho, por ser este un número pequeño (habitualmente menor de uno) se emplea el -2LL ("menos dos veces el logaritmo neperiano de la verosimilitud").

En la siguiente tabla (prueba ómnibus sobre los coeficientes del modelo) se muestra una prueba Chi Cuadrado que evalúa la hipótesis nula de que los coeficientes (P) de todos los términos (excepto la constante) incluidos en el modelo son cero. El estadístico Chi Cuadrado para este contraste es la diferencia entre el valor de -2LL para el modelo solo con la constante y el valor de -2LL para el modelo actual:

$$Chi\ cuadrado = (-2LLMODELO\ 0) - (-2LLMODELO\ 1)$$
$$= 79,232 - 34,280 = 44,952$$

Pruebas ómnibus de coeficientes de modelo				
		Chi-cuadrado	gl	Sig.
Paso 1	Paso	44,951	1	,000
	Bloque	44,951	1	,000
	Mode-lo	44,951	1	,000

Como puede verse en la tabla de la Prueba Ómnibus, el programa ofrece tres entradas: Paso, Bloque y Modelo.

- La fila primera (PASO) es la correspondiente al cambio de verosimilitud (de -2LL) entre pasos sucesivos en la construcción del modelo, contrastando la H0 de que los coeficientes de las variables añadidas en el último paso son cero.
- La segunda fila (BLOQUE) es el cambio en -2LL entre bloques de entrada sucesivos durante la construcción del modelo. Si como es habitual en la práctica se introducen las variables en un solo bloque, el Chi Cuadrado del Bloque es el mismo que el Chi Cuadrado del Modelo.
- La tercera fila (MODELO) es la diferencia entre el valor de -2LL para el modelo solo con la constante y el valor de -2LL para el modelo actual.

En este ejemplo coinciden los tres valores. La significación estadística (0,000) indica que el modelo mejora el ajuste de forma significativa con las nuevas variables introducida respecto a lo que se tenía inicialmente.

Resumen del modelo			
Paso	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	34,280a	,494	,707
a. La estimación ha terminado en el número de iteración 7 porque las estimaciones de parámetro han cambiado en menos de ,001.			

En resumen, del modelo se expresa:

- -2 log de la verosimilitud (-2LL) mide hasta qué punto un modelo se ajusta bien a los datos. El resultado de esta medición recibe también el nombre de “deviance (desviación)”. Cuanto más pequeño sea el valor, mejor será el ajuste.
- La R cuadrado de Cox y Snell es un coeficiente de determinación generalizado que se utiliza para estimar la proporción de varianza de la variable dependiente explicada por las variables predictoras (independientes). La R cuadrado de Cox y Snell se basa en la comparación del log de la verosimilitud (LL) para el modelo respecto al log de la verosimilitud (LL) para un modelo de línea base. Sus valores oscilan entre 0 y 1. En este caso es un valor (0,494) un poco inferior a 0,5 pero indica que el 49,4 % de la variación de la variable dependiente es explicada por la variable incluida en el modelo.
- La R cuadrado de Nagelkerke es una versión corregida de la R cuadrado de Cox y Snell. La R cuadrado de Cox y Snell tiene un valor máximo inferior a 1, incluso para un modelo “perfecto”. La R cuadrado de Nagelkerke corrige la escala del estadístico para cubrir el rango completo de 0 a 1. En el caso que se estudia este coeficiente es 0,707, superior a 0,5

Prueba de Hosmer y Lemeshow			
Paso	Chi-cuadrado	gl	Sig.
1	4,918	6	,554

Esta es otra prueba para evaluar la bondad del ajuste de un modelo de regresión logística. Parte de la idea de que, si el ajuste es bueno, un valor alto de la probabilidad predicha (p) se asociará con el resultado 1 de la variable binomial dependiente, mientras que un valor bajo de p (próximo a cero) corresponderá -en la mayoría de las ocasiones- con el resultado Y=0. Se trata de calcular, para cada observación del conjunto de datos, las probabilidades de la variable dependiente que predice el

modelo, ordenarlas, agruparlas y calcular, a partir de ellas, las frecuencias esperadas, y compararlas con las observadas mediante una prueba Chi-cuadrado.

Como se observa en la tabla para el primer modelo Chi cuadrado no es significativo lo cual indica un ben ajuste del modelo, en el sentido que la hipótesis nula que se contrasta es que no existen diferencias entre las frecuencias de los casos observados y las frecuencias de los casos pronosticados

Sobre este razonamiento, una forma de evaluar la ecuación de regresión y el modelo obtenido es construir una tabla 2×2 clasificando a todos los individuos de la muestra según la concordancia de los valores observados con los predichos o estimados por el modelo, de forma similar a como se evalúan las pruebas diagnósticas. Un modelo puede considerarse aceptable si tanto la especificidad como la sensibilidad tienen un nivel alto, de al menos el 75%.

Tabla de contingencia para la prueba de Hosmer y Lemeshow						
Observado		Enfermedad coronaria = Si		Enfermedad coronaria = No		Total
		Esperado	Observado	Esperado		
Paso 1	1	7	6,654	0	,346	7
	2	7	6,486	1	1,514	8
	3	2	3,646	5	3,354	7
	4	1	1,130	4	3,870	5
	5	2	,727	8	9,273	10
	6	0	,292	13	12,708	13
	7	0	,056	10	9,944	10
	8	0	,010	6	5,990	6

Tabla de clasificación [a]					
	Observado		Pronosticado		
		Enfermedad coronaria		Porcentaje correcto	
		Si	No		
Paso 1	Enfermedad coronaria	Si	16	3	84,2
		No	6	41	87,2
	Porcentaje global				86,4
a. El valor de corte es ,500					

En la tabla de clasificación se constata que el modelo tiene un pronóstico alto (84,2%) de coincidencia con el diagnóstico de si en el padecimiento de la enfermedad de un (87,0%) de coincidencia con el no, empleando solo una constante y una única variable predictora (Tensión arterial sistólica)

Variables en la ecuación									
		B	Error estándar	Wald	gl	Sig.	Exp(B) Inferior	95% C.I. para EXP(B)	
								Superior	
Paso 1[a]	Tensión arterial diastólica	-,263	,065	16,557	1	,000	,769	,677	,873
	Constante	23,580	5,747	16,834	1	,000	17401398380,000		
a. Variables especificadas en el paso 1: Tensión arterial diastólica.									

Con estos datos se constatar se puede construir la ecuación de regresión logística del ejemplo estudiado mediante la ecuación:

$$P(\text{Enfermedad coronaria} = 2(\text{sano})) = \frac{e^{23{,}580-0{,}263*\text{Tensión arterial diastólica}}}{(1+e^{23{,}580-0{,}263*\text{Tensión arterial diastólica}})}$$

De modo que con persona con una tensión arterial diastólica de 97 resulta

$$P(\text{Enfermedad coronaria} = 2(\text{sano})) = \frac{e^{23{,}580-0{,}263*97}}{(1+e^{23{,}580-0{,}263*97})} = \frac{0{,}145003123}{1{,}145003123}$$

$$= 0{,}126639937$$

Lo cual indica que la probabilidad de que este individuo esté sano es solo de un 13%

Matriz de correlaciones			
		Constante	Tensión arterial diastólica
Paso 1	Constante	1,000	-,997
	Tensión arterial diastólica	-,997	1,000

Modelo si el término se ha eliminado					
Variable		Logaritmo de la verosimilitud de modelo	Cambio en el logaritmo de la verosimilitud -2	gl	Sig. del cambio
Paso 1	Tensión arterial diastólica	-39,616	44,951	1	,000

La tabla anterior muestra una evaluación de cuánto perdería el modelo obtenido si se eliminara la variable incluida (Tensión arterial diastólica) en este paso, ya que en los métodos automáticos de construcción del modelo por pasos el proceso evalúa la inclusión y la exclusión de variables. Si dicha variable se elimina; la significación estadística asociada (Sig. del cambio) fuese mayor que el criterio de exclusión establecido, la variable se elimi-

naría del modelo en el paso siguiente. En nuestro ejemplo, como el cambio de verosimilitud es estadísticamente significativo (p = 0,00), la variable en cuestión queda en el modelo.

Las variables no están en la ecuación					
			Puntuación	gl	Sig.
Paso 1	Variables	Triglicérido Basal	,041	1	,839
		Peso	1,549	1	,213
		Sedentarismo(1)	,873	1	,350
		Fuma(1)	3,732	1	,053
	Estadísticos globales		5,010	4	,286

```
             Step number: 1

             Observed Groups and Predicted Probabilities

      16 +                                                                                                   N+
         I                                                                                                   NI
         I                                                                                                   NI
F        I                                                                                                   NI
R     12 +                                                                                               N N+
E        I                                                                                               N NI
Q        I                                                                                          N    N NI
U        I                                                                                          N    N NI
E      8 +                                                                                          N    N N+
N        I                   N                           N                                            N    N NI
C        I                   S                           N                                            N    N NI
Y        I      S            S                           N                             N              N    N NI
       4 +      S            S                           N                             N              N    N N+
         I      S            S                           N                             N              N    N NI
         I S    S            S                           S                             N              S    N NI
         I S    S     S      S                           S                             S              S  N N NI
Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------
  Prob:   0       ,1        ,2        ,3        ,4        ,5        ,6        ,7        ,8        ,9         1
  Group:  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

          Predicted Probability is of Membership for No
          The Cut Value is ,50
          Symbols: S - Si
                   N - No
          Each Symbol Represents 1 Case.
```

Lista por casos[b]						
Caso	Estado seleccionado a	Observado	Pronosticado	Grupo pronosticado	Variable temporal	
		Enfermedad coronaria			Resid	ZResid
30	S	S**	,927	N	-,927	-3,571
61	S	S**	,927	N	-,927	-3,571
a. S = Seleccionado, U = casos sin seleccionar, y ** = casos clasificados incorrectamente.						
b. Se listan los casos con residuos estudentizados mayores que 2,000.						

6.9. Correlación canónica

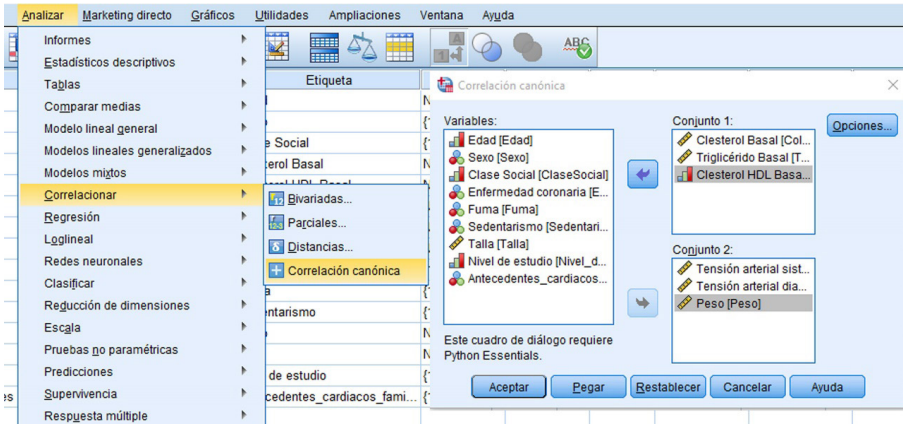
El análisis de correlación canónica (CCA: Canonical Correlation Analysis) es un método de análisis multivariante desarrollado por Harold Hotelling[xxix]. Es una generalización de la correlación múltiple que se aplica en los problemas de regresión múltiple. Recuerde que , el coeficiente de determinación, de los problemas de regresión es la proporción de la variabilidad existente en una variable dependiente que se explica por un conjunto de variables predictoras y se llama coeficiente de correlación múltiple. El coeficiente de correlación múltiple también se puede interpretar como una medida de la correlación máxima que se puede alcanzar entre la variable dependiente y cualquier combinación lineal de las variables predictoras.

El análisis de la correlación canónica es una técnica estadística utilizada para analizar la relación entre múltiples variables dependientes (o endógenas) métricas y varias variables independientes (o exógenas) también métricas. El objetivo esencial del análisis de la correlación canónica es utilizar las variables independientes, cuyos valores son conocidos, para predecir las variables criterio (dependientes) seleccionadas por el investigador.

El procedimiento Correlaciones Canónicas está diseñado para ayudar a identificar asociaciones entre dos conjuntos de variables. Esto lo hace encontrando combinaciones lineales de las variables en los dos conjuntos que exhiban correlaciones fuertes. El par de combinaciones lineales con la correlación más fuerte forman el primer conjunto de *variables canónicas*. El segundo conjunto de variables canónicas es el par de combinaciones lineales que muestran la siguiente correlación más fuerte entre todas las combinaciones que no están correlacionadas con el primer conjunto. Frecuentemente, un número pequeño de pares puede ser usado para cuantificar la relación que existe entre los dos conjuntos.

El objetivo del análisis no lineal de la correlación canónica es analizar las relaciones entre dos o más grupos de variables. En el análisis de correlación canónica hay dos grupos de variables numéricas: por ejemplo, un grupo de variables, formado por los ítems demográficos en un grupo de encuestados, y un grupo de variables, con respuestas a un grupo de ítems de actitud. El análisis de correlación canónica estándar es una técnica estadística que busca una combinación lineal de un grupo de variables y una combinación lineal de un segundo grupo de variables correlacionadas al máximo.

6.10. Ejemplo de aplicación de la correlación canónica



En la base de datos sobre enfermedades coronarias es posible determinar la correlación que existe entre dos grupos de variables predictoras: por un lado, Colesterol Basal, Colesterol HDL Basal, Triglicérido Basal y por el otro, Tensión arterial sistólica, Tensión arterial diastólica y el peso: La siguiente imagen ilustra el proceso se selección de las opciones en SPSS.

6.11. Resultados de la aplicación del método

Correlaciones canónicas							
	Correlación	Autovalor	Estadístico de Wilks	F	Número D.F	Denominador D.F.	Sig.
1	,609	,588	,607	3,701	9,000	146,175	,000
2	,155	,025	,964	,567	4,000	122,000	,687
3	,111	,013	,988	,780	1,000	62,000	,381
H0 de prueba de Wilks significa que las correlaciones de la fila actual y las siguientes son cero							

Coeficiente de correlación canónica estandarizada del conjunto 1				
Variable	1	2	3	
Colesterolemia_Basal	-,931	,245	-,295	
Trigliceridemia_Basal	-,300	-,975	-,134	
Colesterolemia_HDL_Basal	,295	,272	-,946	

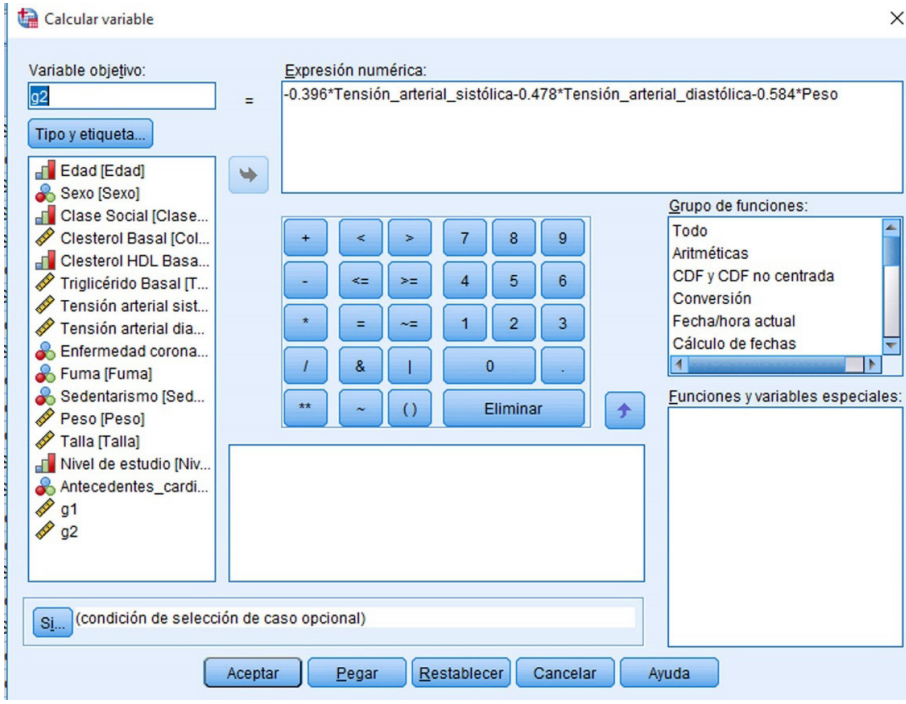
Coeficiente de correlación canónica estandarizada del conjunto 2			
Variable	1	2	3
Tensión_arterial_sistólica	,196	1,442	,944
Tensión_arterial_diastólica	-,462	-1,645	,371
Peso	-,797	,369	-,945

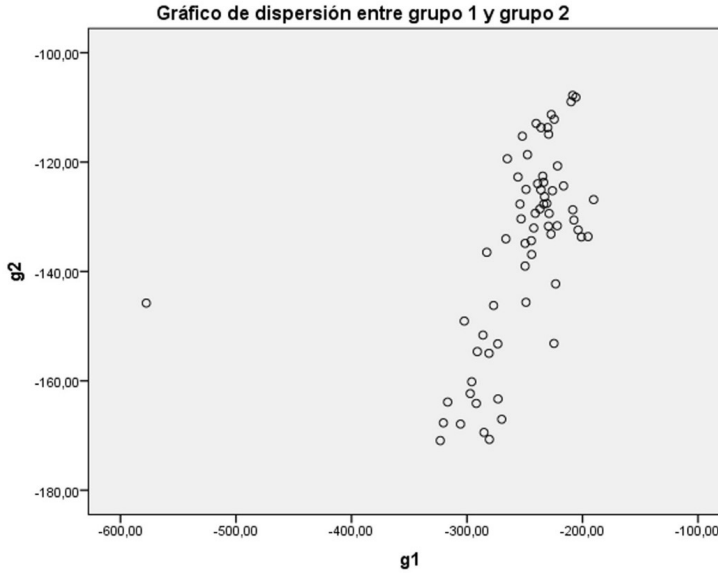
En las tres tablas anteriores se resume el método explicado anteriormente. En las dos tablas comprendidas bajo el título “Co-

eficiente de correlación canónica estandarizada del conjunto (1 / 2)" se muestran las combinaciones lineales de dos conjuntos de variables que tienen la mayor correlación entre ellas. En este caso, se formaron 3 conjuntos de combinaciones lineales. El primer conjunto de combinaciones lineales es:

- 0,931*Colesterolemia_Basal - 0,300* Trigliceridemia_Basal + 0,295* Colesterolemia_HDL_Basal
- 0,462*Tensión_arterial_diastólica - 0,196*Tensión_arterial_sistólica + 0,797*Peso

Para estas tablas las variables fueron primero estandarizadas restándoles primero sus medias y dividiéndolas entre sus desviaciones estándar. Si el lector desea puede construir estas variables con el SPSS y desarrollar el gráfico de dispersión como se muestra a continuación:





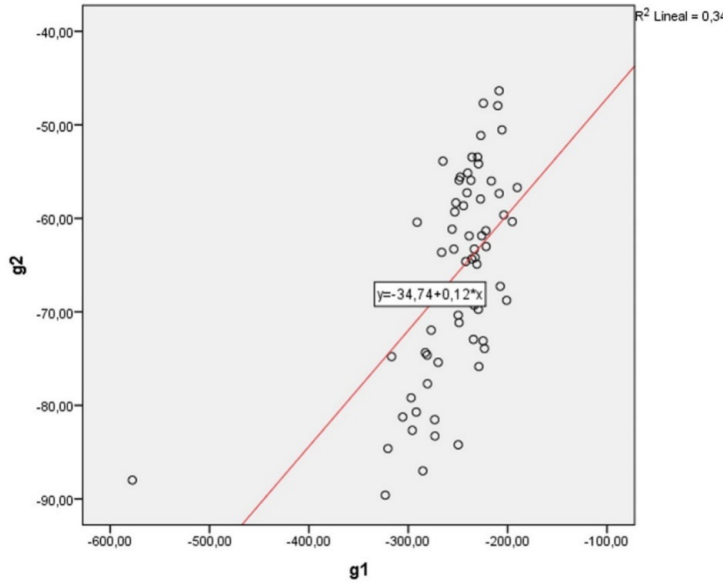
Entre estas dos nuevas variables se puede establecer las siguientes relaciones:

Estadísticos descriptivos			
	Media	Desviación estándar	N
g1	-252,8299	51,39233	67
g2	-66,0061	10,90892	68

Correlaciones			
		g1	g2
g1	Correlación de Pearson	1	,588**
	Sig. (bilateral)		,000
	Suma de cuadrados y productos vectoriales	174317,322	21573,302
	Covarianza	2641,172	331,897
	N	67	66

g2	Correlación de Pearson	,588**	1
	Sig. (bilateral)	,000	
	Suma de cuadrados y productos vectoriales	21573,302	7973,308
	Covarianza	331,897	119,005
	N	66	68
**. La correlación es significativa en el nivel 0,01 (bilateral).			

También es posible determinar la recta de regresión:



La primera tabla (Correlaciones canónicas) muestra las correlaciones estimadas entre cada conjunto de variables canónicas. Dado que uno de los valores-P es menor que 0,05, ese conjunto (conjunto 1) tiene una correlación estadísticamente significativa con un nivel de confianza del 95,0%.

En esta primera tabla también se da el estadístico lambda de Wilks, que expresa la proporción de variabilidad total no debida a las diferencias entre los grupos; permite contrastar la hipótesis

nula de que las medias multivariantes de los grupos (los centroides) son iguales. Wilks (1932), basándose en el principio de razón de verosimilitud generalizada (según el cual la varianza generalizada de un espacio multivariante puede ser calculada mediante el determinante de la matriz de dispersión), planteó el estadístico A, definido como:

$$A = \frac{Suma\ de\ los\ cuadrados\ intragrupos}{Suma\ de\ cuadrado\ total} = \frac{|S|}{|T|}$$

Donde S es la matriz de varianzas-covarianzas *combinada*, calculada a partir de las matrices de varianzas-covarianzas de cada grupo, y T es la matriz de varianzas-covarianzas total, calculada sobre todos los casos como si pertenecieran a un único grupo. Cuando los grupos se encuentren superpuestos en el espacio multidimensional, los valores del numerador y del denominador serán aproximadamente iguales y su cociente valdrá 1; a medida que los grupos se vayan separando más y más, la variabilidad *inter-grupos* irá aumentando y la variabilidad *intra-grupos* se irá haciendo comparativamente menor respecto a la variabilidad *total* y disminuye así el valor del cociente. Por tanto, valores próximos a 1 indicarán un gran parecido entre los grupos, mientras que valores próximos a 0 indicarán una gran diferencia entre ellos.

En el caso que nos ocupa para el conjunto 1 que tiene una correlación estadísticamente significativa con un nivel de confianza del 95,0%. El estadístico de Wilks es de 0,607, lo cual indica que el parecido entre los grupos es adecuado por ser mayor que 0,5.

Coeficiente de correlación canónica no estandarizada del conjunto 1			
Variable	1	2	3
Colesterolemia_Basal	-,017	,004	-,005
Trigliceridemia_Basal	-,008	-,026	-,004
Colesterolemia_HDL_Basal	,037	,034	-,119

Coeficiente de correlación canónica no estandarizada del conjunto 2			
Variable	1	2	3
Tensión_arterial_sistólica	,010	,073	,048
Tensión_arterial_diastólica	-,041	-,145	,033
Peso	-,066	,031	-,078

Cargas canónicas del conjunto 1			
Variable	1	2	3
Colesterolemia_Basal	-,929	,313	-,199
Trigliceridemia_Basal	-,146	-,938	-,315
Colesterolemia_HDL_ Basal	,310	,035	-,950

Cargas canónicas del conjunto 2			
Variable	1	2	3
Tensión_arterial_sistólica	-,651	,336	,680
Tensión_arterial_diastólica	-,786	-,260	,561
Peso	-,959	,234	-,158

Las cargas canónicas, también denominadas correlaciones de estructura canónica, miden la correlación lineal simple entre una variable original observada del conjunto dependiente o independiente y el valor teórico canónico del conjunto. Las cargas canónicas reflejan la varianza que la variable observada compare con el valor teórico canónico, y puede ser interpretada como una carga factorial para valorar la contribución relativa de cada variable a cada función canónica. Se considera cada función canónica independiente de forma separada, y se calcula la correlación dentro del conjunto entre variables y valores teóricos. Cuanto mayor sea el coeficiente, mayor es la importancia que tiene para calcular el valor teórico canónico. Los criterios para determinar la significación de las correlaciones de estructura canónica también son los mismos que con las cargas factoriales.

Cargas cruzadas del conjunto 1			
Variable	1	2	3
Colesterolemia_Basal	-,565	,049	-,022
Trigliceridemia_Basal	-,089	-,145	-,035
Colesterolemia_HDL_Basal	,189	,005	-,106
Cargas cruzadas del conjunto 2			
Variable	1	2	3
Tensión_arterial_sistólica	-,396	,052	,076
Tensión_arterial_diastólica	-,478	-,040	,063
Peso	-,584	,036	-,018

La carga cruzada consiste en correlacionar cada una de las variables dependientes originales observadas directamente con el valor teórico canónico independiente, y viceversa.

Proporción de la varianza explicada				
Variable canónica	Conjunto 1 por sí mismo	Conjunto 1 por conjunto 2	Conjunto 2 por sí mismo	Conjunto 2 por conjunto 1
1	,327	,121	,654	,242
2	,326	,008	,079	,002
3	,347	,004	,268	,003

Capítulo VII. Conglomerados y correspondencias

7.1. Análisis de conglomerados (clúster)

Cluster (a veces castellanizado como *clúster*) es un término inglés encontrado en varios tecnicismos. La traducción literal al castellano es *racimo*, conjunto, *grupo o cúmulo*, Conglomerado, pero ¿qué es realmente el análisis de conglomerados o clúster?

El análisis clúster es la denominación de un grupo de técnicas multivariantes cuyo principal propósito es agrupar objetos basándose en las características que poseen. El análisis clúster clasifica objetos (es decir, encuestados, productos u otras entidades) de tal forma que cada objeto es muy parecido a los que hay en el conglomerado con respecto a algún criterio de selección predeterminado.

Los conglomerados de objetos resultantes deberían mostrar un alto grado de homogeneidad interna (dentro del conglomerado) y un alto grado de heterogeneidad externa (entre conglomerados). Por tanto, si la clasificación es acertada, los objetos dentro de los conglomerados estarán muy próximos cuando se representen gráficamente, y los diferentes grupos estarán muy alejados.

Al parecer lo dicho es claro y hasta elemental pero la idea de un clúster o grupo resulta compleja para concretar su definición exacta, por esto existen múltiples algoritmos de agrupamiento, aunque existe como elemento común que se trata de un grupo de datos, aunque los investigadores utilicen diferentes modelos de agrupación con algoritmos que difieren entre sí, lo que hace variar las propiedades de cada subgrupo producto de la clasificación.

Aunque resulta difícil clasificar los algoritmos de agrupación existe un consenso de que estos modelos pueden ser:

- De conectividad, como los agrupamientos jerárquicos basados en la distancia de las conexiones.
- De centroide: organizan los grupos en base a un solo vector medio.

- De distribución: los grupos a partir de distribuciones estadísticas.
- De densidad: definen los grupos como regiones densas conectadas en el espacio de los datos.
- De sub-espacios: conocido como Co-clustering o two-mode-clustering, siguiendo este modelo los grupos se forman con las dos características, que aportan, por un lado, el ser miembros del grupo y por otro atendiendo a los atributos relevantes.
- De grupo: se trata del empleo de algoritmos que no proporcionan un modelo refinado para sus resultados y solo ofrecen la información de la agrupación.
- Basados en grafo: cada dos nodos en el subconjunto están conectados por una arista.

También los agrupamientos pueden clasificarse en:

- Agrupamiento Duro: cada objeto pertenece o no pertenece a un solo grupo.
- Agrupamiento Suave o difuso: cada objeto pertenece o no a un grupo según un grado de pertenencia.
- Agrupamiento con partición estricta con ruido: E posible que existan objetos que no pertenezcan a grupo alguno.
- Agrupamiento con solapamiento: contrario a agrupamiento duro, los objetos pueden pertenecer a más de un grupo.
- Agrupamiento jerárquico: objetos que pertenecen a un grupo hijo también pertenecen al grupo padre
- Agrupamiento de subespacios: contrario a agrupamiento con solapamiento, dentro de un único sub-espacio definido, los grupos deben solaparse.

7.2. Utilidad de análisis por conglomerados o clúster

El análisis clúster es muy útil cuando un investigador desea desarrollar las hipótesis concernientes a la naturaleza de los datos o para examinar las hipótesis previamente establecidas. Por ejemplo:

- Un investigador puede creer que las actitudes hacia el consumo de refrescos normales frente a «lights» podrían utilizarse para separar a los consumidores de refrescos en segmentos lógicos o grupos. El análisis clúster puede clasificar consumidores de refrescos por sus actitudes hacia los refrescos normales frente a los «light», y los conglomerados resultantes, si los hay, pueden perfilarse mediante diferencias y similitudes demográficas.
- Desde la derivación de taxonomías en biología para la agrupación de todos los organismos vivientes a clasificaciones psicológicas basadas en la personalidad y otros rasgos personales, pasando por los análisis de segmentación de los mercados, el análisis clúster ha tenido siempre una fuerte tradición en la agrupación de individuos. Esta tradición se ha extendido a la clasificación de objetos e incluye la estructura de mercado, análisis de similitudes y diferencias entre productos nuevos y evaluación del rendimiento de empresas para identificar agrupaciones basadas en las estrategias de las empresas u orientaciones estratégicas. El resultado ha sido una profusión de aplicaciones en casi todas las áreas de investigación, creando no solo una riqueza de conocimiento en el uso del análisis de conglomerados sino también la necesidad de una mejor comprensión de la técnica para minimizar su mala utilización.

7.3. Inconvenientes del análisis de clúster

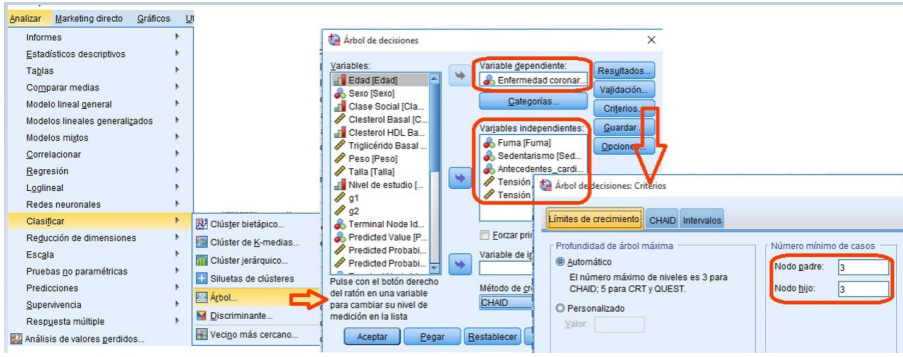
- El análisis clúster puede caracterizarse como descriptivo, ateórico y no inferencial.

- El análisis clúster no tiene bases estadísticas sobre las cuales deducir inferencias estadísticas para una población a partir de una muestra, y se utiliza fundamentalmente como una técnica de exploratoria.
- Las soluciones no son únicas, en la medida en que la pertenencia al conglomerado para cualquier número de soluciones depende de muchos elementos del procedimiento y se pueden obtener muchas soluciones diferentes variando uno o más de estos elementos.
- El análisis clúster siempre creará conglomerados, a pesar de la existencia de una «auténtica» estructura en los datos.
- La solución clúster es totalmente dependiente de las variables utilizadas como base para la medida de similitud. La adición o destrucción de variables relevantes puede tener un impacto substancial sobre la solución resultante. Por tanto, el investigador debe tener particular cuidado en evaluar el impacto de cada decisión implicada en el desarrollo de un análisis clúster.

7.4. Conglomerados jerárquicos

Un primer acercamiento al tema de los conglomerados lleva a los conglomerados jerárquicos que consisten en la construcción de una estructura en forma de árbol. Una característica importante de los procedimientos jerárquicos es que los resultados obtenidos en un paso previo siempre necesitan encajarse dentro de los resultados del siguiente paso, creando algo parecido a un árbol.

Continuando con el fichero de las enfermedades coronaria se propone hacer varias clasificaciones atendiendo a las variables que se escojan y los métodos de clasificación. La siguiente imagen muestra el inicio del análisis por conglomerados jerárquicos con los cuadros de diálogos más significativos:



Aunque los árboles de decisión son identificados en muchos textos que tratan el tema dentro de las técnicas de minería de datos, en este, los autores han decidido mantenerlos dentro del análisis de clúster como conglomerados jerárquicos; por otro lado, por la importancia y claridad de las orientaciones que sobre este tema da la ayuda del SPSS se ha hecho de la misma la siguiente síntesis:

7.5. Árboles de decisión (tomado de la ayuda del SPSS)

El procedimiento Árbol de decisión crea un modelo de clasificación basado en árboles y clasifica casos en grupos o pronostica valores de una variable (criterio) dependiente basada en valores de variables independientes (predictores). El procedimiento proporciona herramientas de validación para análisis de clasificación exploratorios y confirmatorios.

El procedimiento se puede utilizar para:

Segmentación. Identifica las personas que pueden ser miembros de un grupo específico.

Estratificación. Asigna los casos a una categoría de entre varias, por ejemplo, grupos de alto riesgo, bajo riesgo y riesgo intermedio.

Predicción. Crea reglas y las utiliza para predecir eventos futuros, como la verosimilitud de que una persona cause mora en un

crédito o el valor de reventa potencial de un vehículo o una casa.

Reducción de datos y clasificación de variables. Selecciona un subconjunto útil de predictores a partir de un gran conjunto de variables para utilizarlo en la creación de un modelo paramétrico formal.

Identificación de interacción. Identifica las relaciones que pertenecen solo a subgrupos específicos y las especifica en un modelo paramétrico formal.

Fusión de categorías y discretización de variables continuas. Vuelve a codificar las variables continuas y las categorías de los predictores del grupo, con una pérdida mínima de información.

Ejemplo. Un banco desea categorizar a los solicitantes de créditos en función de si representan o no un riesgo crediticio razonable. Se basa en varios factores, e incluye las valoraciones del crédito conocidas de clientes anteriores, se puede generar un modelo para pronosticar si es probable que los clientes futuros causen mora en sus créditos.

Un análisis basado en árboles ofrece algunas características atractivas:

- Permite identificar grupos homogéneos con alto o bajo riesgo.
- Facilita la creación de reglas para realizar predicciones sobre casos individuales.

Consideraciones de los datos.

Datos. Las variables dependientes e independientes pueden ser:

- *Nominal*. Una variable puede ser tratada como nominal cuando sus valores representan categorías que no obedecen a una clasificación intrínseca. Por ejemplo, el departamento de la compañía en el que trabaja un empleado. Algunos ejemplos de variables nominales son: región, código postal o confesión religiosa.

- *Ordinal.* Una variable puede ser tratada como ordinal cuando sus valores representan categorías con alguna clasificación intrínseca. Por ejemplo, los niveles de satisfacción con un servicio, que abarquen desde muy insatisfecho hasta muy satisfecho. Entre los ejemplos de variables ordinales se incluyen escalas de actitud que representan el grado de satisfacción o confianza y las puntuaciones de evaluación de las preferencias.

- *Escalas*. Una variable puede tratarse como escala (continua) cuando sus valores representan categorías ordenadas con una métrica con significado, por lo que son adecuadas las comparaciones de distancia entre valores. Son ejemplos de variables de escala: la edad en años y los ingresos en dólares.

Ponderaciones de frecuencia. Si se encuentra activada la ponderación, las ponderaciones fraccionarias se redondearán al número entero más cercano; de esta manera, a los casos con un valor de ponderación menor que 0,5 se les asignará una ponderación de 0 y, por consiguiente, se verán excluidos del análisis.

Supuestos. Este procedimiento supone que se ha asignado el nivel de medición adecuado a todas las variables del análisis; además, algunas características suponen que todos los valores de la variable dependiente incluidos en el análisis tienen etiquetas de valor definidas.

Nivel de medición. El nivel de medición afecta a los tres cálculos; por lo tanto, todas las variables deben tener asignado el nivel de medición adecuado. De forma predeterminada, se supone que las variables numéricas son de escala y que las variables de cadena son nominales, lo cual podría no reflejar con exactitud el verdadero nivel de medición. Un icono junto a cada variable en la lista de variables identifica el tipo de variable.

Iconos de nivel de medición	
Icono	Nivel de medición
	Escala
	Nominal
	Ordinal

Puede cambiar de forma temporal el nivel de medición de una variable; para ello, pulse con el botón derecho del ratón en la variable en la lista de variables de origen y seleccione un nivel de medición del menú emergente.

Etiquetas de valor. La interfaz del cuadro de diálogo para este procedimiento supone que, o todos los valores no perdidos de una variable dependiente categórica (nominal, ordinal) tienen etiquetas de valores definidas, o que ninguno de ellos las tiene. Algunas características no estarán disponibles a menos que como mínimo dos valores no perdidos de la variable dependiente categórica tengan etiquetas de valor. Si al menos dos valores no perdidos tienen etiquetas de valor definidas, todos los demás casos con otros valores que no tengan etiquetas de valor se excluirán del análisis.

Puede utilizar Definir propiedades de variable como ayuda en el proceso de definición del nivel de medición y de las etiquetas de valor.

Para obtener árboles de decisión.

Esta característica requiere la opción Árboles de decisión.

1. Seleccione en los menús:

   Analizar > Clasificar > Árbol...

2. Seleccione una variable dependiente.
3. Seleccionar una o más variables independientes.
4. Seleccione un método de crecimiento.

Si lo desea, puede:

- Cambiar el nivel de medición para cualquier variable de la lista de origen.
- Forzar que la primera variable en la lista de variables independientes en el modelo sea la primera variable de segmentación.
- Seleccionar una variable de influencia que defina cuánta influencia tiene un caso en el proceso de crecimiento de un árbol. Los casos con valores de influencia inferiores tendrán menos influencia, mientras que los casos con valores superiores tendrán más. Los valores de la variable de influencia deben ser valores positivos.
- Validar el árbol.
- Personalizar los criterios de crecimiento del árbol.
- Guardar los números de nodos terminales, valores pronosticados y probabilidades pronosticadas como variables.
- Guardar el modelo en formato XML (PMML).

Campos con un nivel de medición desconocido

La alerta de nivel de medición se muestra si el nivel de medición de una o más variables (campos) del conjunto de datos es desconocido. Como el nivel de medición afecta al cálculo de los resultados de este procedimiento, todas las variables deben tener un nivel de medición definido.

Explorar datos. Lee los datos del conjunto de datos activo y asigna el nivel de medición predefinido en cualquier campo con un nivel de medición desconocido. Si el conjunto de datos es grande, puede llevar algún tiempo.

Asignar manualmente. Abre un cuadro de diálogo que contiene todos los campos con un nivel de medición desconocido. Puede utilizar este cuadro de diálogo para asignar el nivel de medición a esos campos. También puede asignar un nivel de medición en la Vista de variables del Editor de datos.

Como el nivel de medición es importante para este procedimiento, no puede acceder al cuadro de diálogo para ejecutar este procedimiento hasta que se hayan definido todos los campos en el nivel de medición.

Cambio del nivel de medición

1. En la lista de origen, pulse con el botón derecho del ratón en la variable.
2. Seleccione un nivel de medición del menú emergente.

Esto modifica de forma temporal el nivel de medición para su uso en el procedimiento Árbol de decisión.

Para modificar permanentemente el nivel de medición de una variable, consulte Nivel de medición de variable.

Métodos de crecimiento

Los métodos de crecimiento disponibles son:

CHAID. Detección automática de interacciones mediante chi-cuadrado (CHi-square Automatic Interaction Detection). En cada paso, CHAID elige la variable independiente (predictora) que presenta la interacción más fuerte con la variable dependiente. Las categorías de cada predictor se funden si no son significativamente distintas respecto a la variable dependiente.

CHAID exhaustivo. Una modificación del CHAID que examina todas las divisiones posibles de cada predictor.

CRT. Árboles de clasificación y regresión (Classification and Regression Trees). CRT divide los datos en segmentos para que sean lo más homogéneos que sea posible respecto a la variable

dependiente. Un nodo terminal en el que todos los casos toman el mismo valor en la variable dependiente es un nodo homogéneo y puro.

QUEST. Árbol estadístico rápido, insesgado y eficiente (Quick, Unbiased, Efficient Statistical Tree). Método rápido y que evita el sesgo que presentan otros métodos al favorecer los predictores con muchas categorías. Solo puede especificarse QUEST si la variable dependiente es nominal.

Cada método presenta ventajas y limitaciones, entre las que se incluyen:

Características del método de crecimiento

Feature	CHAID*	CRT	QUEST
Basado en chi-cuadrado**	X		
Variables (predictoras) independientes sustitutas		X	X
Poda de árboles		X	X
División de nodos multinivel	X		
División de nodos binarios		X	X
Variables de influencia	X	X	
Probabilidades previas		X	X
Costes de clasificación errónea	X	X	X
Cálculo rápido	X		X

*Incluye CHAID exhaustivo.

**QUEST también utiliza una medida de chi-cuadrado para variables independientes nominales.

Es significativo comentar la pestaña límites de crecimiento destacada en la lámina del menú inicial; al respecto la ayuda del SPSS plantea:

Límites de crecimiento

La pestaña Límites de crecimiento permite limitar el número de niveles del árbol y controlar el número de casos mínimo para nodos padre y para nodos hijo.

Máxima profundidad de árbol. Controla el número máximo de niveles de crecimiento por debajo del nodo raíz. El ajuste Automática limita el árbol a tres niveles por debajo del nodo raíz para los métodos CHAID y CHAID exhaustivo y a cinco niveles para los métodos CRT y QUEST.

Número de casos mínimo. Controla el número de casos mínimo para los nodos. Los nodos que no cumplen estos criterios no se dividen.

El aumento de los valores mínimos tiende a generar árboles con menos nodos.

La disminución de dichos valores mínimos generará árboles con más nodos.

Para archivos de datos con un número pequeño de casos, es posible que, en ocasiones, los valores predeterminados de 100 casos para nodos padre y de 50 casos para nodos hijo den como resultado árboles sin ningún nodo por debajo del nodo raíz; en este caso, la disminución de los valores mínimos podría generar resultados más útiles.

Dado que en la muestra del ejemplo se tienen 70 casos se cambiaron los valores predeterminados de 100 nodos padres y 50 nodos hijos por 3 padres y tres hijos.

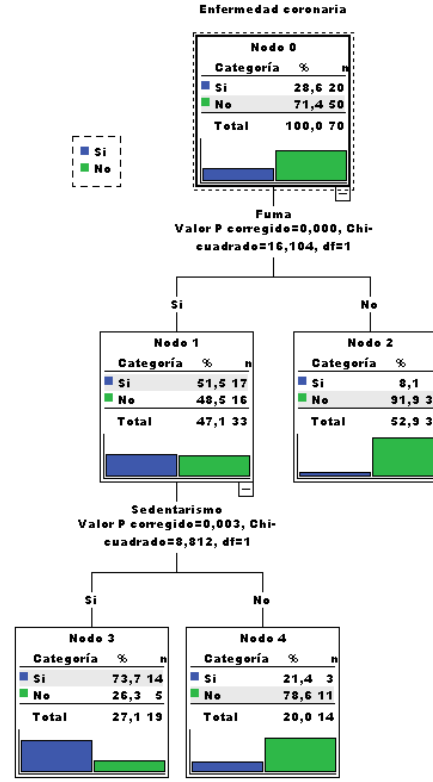
7.6. Resultados de un análisis mediante un árbol de decisiones

Resumen del modelo		
Especificaciones	Método de crecimiento	CHAID
	Variable dependiente	Enfermedad coronaria
	Variables independientes	Fuma, Sedentarismo, Antecedentes_cardiacos_familiares
	Validación	Ninguna
	Máxima profundidad del árbol	3
	Casos mínimos en nodo padre	3
	Casos mínimos en nodo hijo	3
Resultados	Variables independientes incluidas	Fuma, Sedentarismo
	Número de nodos	5
	Número de nodos terminales	3
	Profundidad	2

Obsérvese que:

1. Solamente se tomaron como variables independientes tres variables categóricas.
2. Aunque se dieron tres variables independientes el método solo considera dos: Fuma y Sedentarismo, porque el método CHAID ha descartado la variable Antecedentes_cardiacos_familiares como significativa para un análisis de incidencia dependencia.
3. En este árbol se ha incluido un gráfico de barras además de la tabla de frecuencias que implícitamente se da, esto es posible obtenerlo a partir de un menú de edición que aparece al pulsar el clic secundario sobre el gráfico.

4. En cada nodo se da el chi-cuadrado correspondiente, dato que el método CHAID se basa en él para mostrar solo las tablas donde hay diferencias significativas en la tabla de contingencia que se presenta.
5. Del árbol se infiere que hay un 28% de individuos que padecen la enfermedad, de ellos el 51,5% fuma y de los que fuman el 73,7% son sedentarios. En resumen, hay 14 personas (20%) que son sedentarios, fuman y padecen la enfermedad.



Las siguientes tablas complementan la información del árbol.

Categoría de objetivo: Si

Ganancias para nodos						
Nodo	Nodo		Ganancia		Respuesta	Índice
	N	Porcentaje	N	Porcentaje		
3	19	27,1%	14	70,0%	73,7%	257,9%
4	14	20,0%	3	15,0%	21,4%	75,0%
2	37	52,9%	3	15,0%	8,1%	28,4%
Método de crecimiento: CHAID Variable dependiente: Enfermedad coronaria						

Categoría de objetivo: No

Ganancias para nodos						
Nodo	Nodo		Ganancia		Res-puesta	Índice
	N	Porcen-taje	N	Porcen-taje		
2	37	52,9%	34	68,0%	91,9%	128,6%
4	14	20,0%	11	22,0%	78,6%	110,0%
3	19	27,1%	5	10,0%	26,3%	36,8%
Método de crecimiento: CHAID Variable dependiente: Enfermedad coronaria						

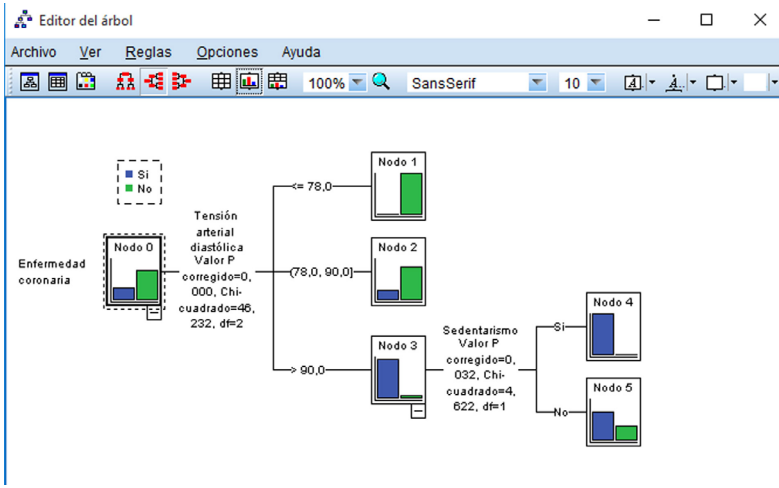
Riesgo	
Estimación	Desv. Error
,157	,043
Método de crecimiento: CHAID Variable dependiente: Enfermedad coronaria	

Clasificación			
Observado	Pronosticado		
	Si	No	Porcentaje correcto
Si	14	6	70,0%
No	5	45	90,0%
Porcentaje global	27,1%	72,9%	84,3%
Método de crecimiento: CHAID Variable dependiente: Enfermedad coronaria			

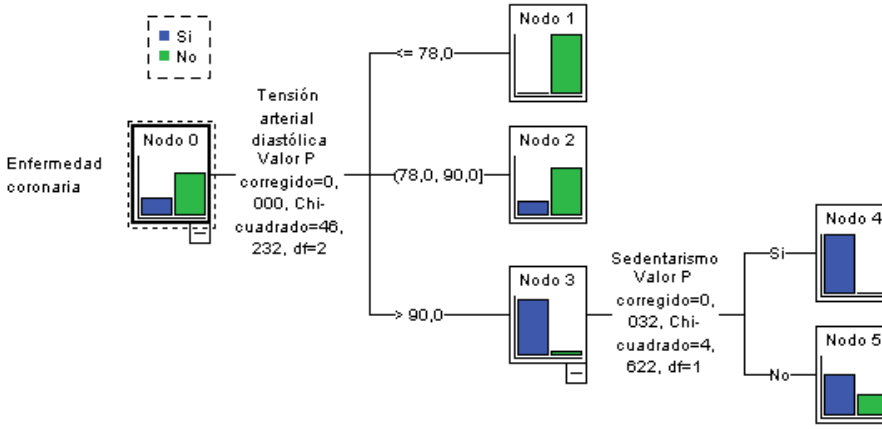
Resumen del modelo		
Especificaciones	Método de crecimiento	CHAID
	Variable dependiente	Enfermedad coronaria
	Variables independientes	Fuma, Sedentarismo, Antecedentes_cardiacos_familiares, Tensión arterial sistólica, Tensión arterial diastólica
	Validación	Ninguna
	Máxima profundidad del árbol	3
	Casos mínimos en nodo padre	3
	Casos mínimos en nodo hijo	3
Resultados	Variables independientes incluidas	Tensión arterial diastólica, Sedentarismo
	Número de nodos	6
	Número de nodos terminales	4
	Profundidad	2

Obsérvese que:

1. Se añadieron al análisis dos variables de escalas relacionas con la tensión arterial.
2. El método CHAID priorizó ahora como variables independientes la Tensión arterial diastólica y el Sedentarismo
3. Se dio al gráfico una orientación horizontal y solo se presenta el gráfico de barra, en la siguiente imagen se muestra el procedimiento de edición utilizado.



El gráfico evidencia la relación entre la presión arterial diastólica superior a 90 y el sedentarismo.



Otras informaciones complementarias se dan en las siguientes tablas.

Categoría de objetivo: Si

Ganancias para nodos						
	Nodo		Ganancia			
Nodo	N	Porcentaje	N	Porcentaje	Respuesta	Índice
4	13	18,6%	13	65,0%	100,0%	350,0%
5	3	4,3%	2	10,0%	66,7%	233,3%
2	23	32,9%	5	25,0%	21,7%	76,1%
1	31	44,3%	0	0,0%	0,0%	0,0%
Método de crecimiento: CHAID Variable dependiente: Enfermedad coronaria						

Categoría de objetivo: No

Ganancias para nodos						
	Nodo		Ganancia			
Nodo	N	Porcentaje	N	Porcentaje	Respuesta	Índice
1	31	44,3%	31	62,0%	100,0%	140,0%
2	23	32,9%	18	36,0%	78,3%	109,6%
5	3	4,3%	1	2,0%	33,3%	46,7%
4	13	18,6%	0	0,0%	0,0%	0,0%
Método de crecimiento: CHAID Variable dependiente: Enfermedad coronaria						

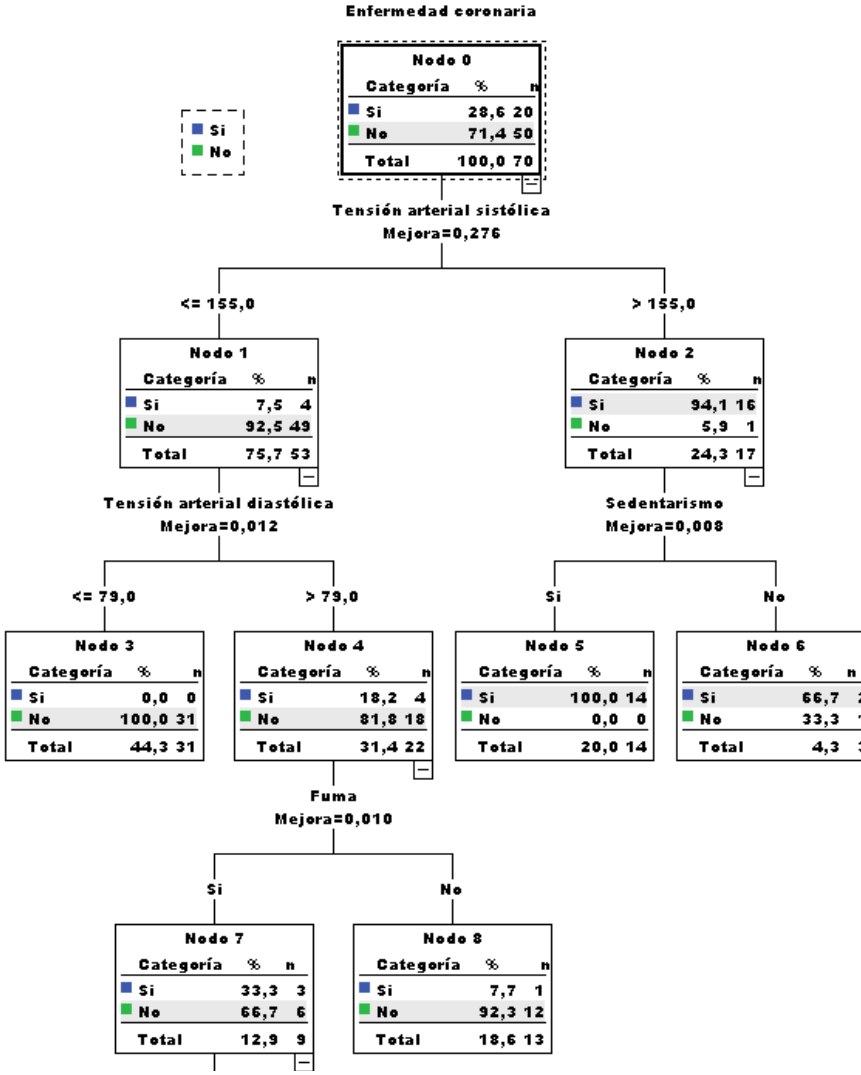
Riesgo	
Estimación	Desv. Error
,086	,033
Método de crecimiento: CHAID Variable dependiente: Enfermedad coronaria	

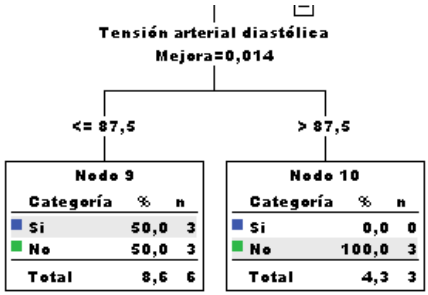
Clasificación			
Observado	Pronosticado		
	Si	No	Porcentaje correcto
Si	15	5	75,0%
No	1	49	98,0%
Porcentaje global	22,9%	77,1%	91,4%
Método de crecimiento: CHAID Variable dependiente: Enfermedad coronaria			

Resumen del modelo		
Especificaciones	Método de crecimiento	CRT
	Variable dependiente	Enfermedad coronaria
	Variables independientes	Fuma, Sedentarismo, Antecedentes_cardiacos_familiares, Tensión arterial sistólica, Tensión arterial diastólica
	Validación	Ninguna
	Máxima profundidad del árbol	5
	Casos mínimos en nodo padre	3
	Casos mínimos en nodo hijo	3
Resultados	Variables independientes incluidas	Tensión arterial sistólica, Tensión arterial diastólica, Antecedentes_cardiacos_familiares, Fuma, Sedentarismo
	Número de nodos	11
	Número de nodos terminales	6
	Profundidad	4

Obsérvese que:

1. Se han mantenido las variables, pero ha cambiado el método, ahora el árbol es mayor porque toma en consideración todas las variables, pero las ordena según su relación con que un individuo esté o no enfermo.





Por lo antes analizado no se requiere en este caso mayores explicaciones y como en los casos anteriores las tablas finales complementan la explicación

Categoría de objetivo: Si

Ganancias para nodos						
Nodo	Nodo		Ganancia		Respuesta	Índice
	N	Porcentaje	N	Porcentaje		
5	14	20,0%	14	70,0%	100,0%	350,0%
6	3	4,3%	2	10,0%	66,7%	233,3%
9	6	8,6%	3	15,0%	50,0%	175,0%
8	13	18,6%	1	5,0%	7,7%	26,9%
3	31	44,3%	0	0,0%	0,0%	0,0%
10	3	4,3%	0	0,0%	0,0%	0,0%
Método de crecimiento: CRT Variable dependiente: Enfermedad coronaria						

Categoría de objetivo: No

Ganancias para nodos						
Nodo	Nodo		Ganancia		Respuesta	Índice
	N	Porcentaje	N	Porcentaje		
3	31	44,3%	31	62,0%	100,0%	140,0%

10	3	4,3%	3	6,0%	100,0%	140,0%
8	13	18,6%	12	24,0%	92,3%	129,2%
9	6	8,6%	3	6,0%	50,0%	70,0%
6	3	4,3%	1	2,0%	33,3%	46,7%
5	14	20,0%	0	0,0%	0,0%	0,0%

Método de crecimiento: CRT

Variable dependiente: Enfermedad coronaria

Riesgo

Estimación	Desv. Error
,071	,031

Método de crecimiento: CRT

Variable dependiente: Enfermedad coronaria

Clasificación

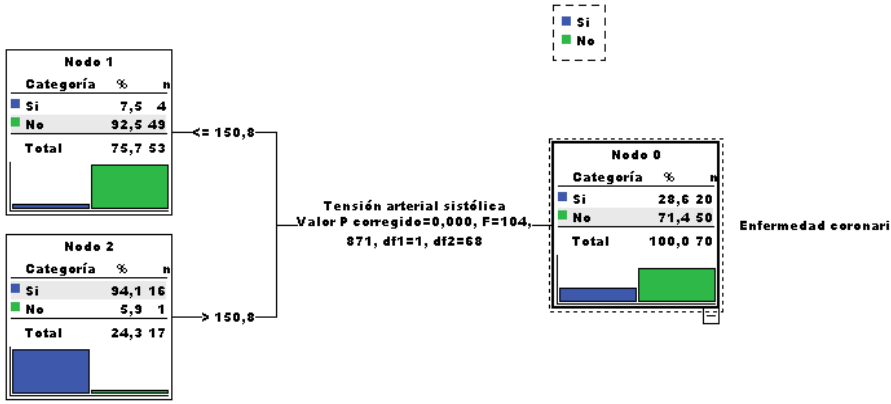
Observado	Pronosticado		
	Si	No	Porcentaje correcto
Si	19	1	95,0%
No	4	46	92,0%
Porcentaje global	32,9%	67,1%	92,9%

Método de crecimiento: CRT

Variable dependiente: Enfermedad coronaria

Resumen del modelo		
Especificaciones	Método de crecimiento	QUEST
	Variable dependiente	Enfermedad coronaria
	Variables independientes	Fuma, Sedentarismo, Antecedentes_cardiacos_familiares, Tensión arterial sistólica, Tensión arterial diastólica
	Validación	Ninguna
	Máxima profundidad del árbol	5
	Casos mínimos en nodo padre	3
	Casos mínimos en nodo hijo	3
Resultados	Variables independientes incluidas	Tensión arterial sistólica, Tensión arterial diastólica
	Número de nodos	3
	Número de nodos terminales	2
	Profundidad	1

En este caso el método QUEST como indica su concepción es muy sintético y prioriza en su clasificación una sola variable; con el propósito de mostrar otra opción se cambió la orientación a horizontal y de derecha a izquierda con tablas y gráficos.



Categoría de objetivo: Si

Ganancias para nodos						
	Nodo		Ganancia			
Nodo	N	Porcentaje	N	Porcentaje	Respuesta	Índice
2	17	24,3%	16	80,0%	94,1%	329,4%
1	53	75,7%	4	20,0%	7,5%	26,4%
Método de crecimiento: QUEST Variable dependiente: Enfermedad coronaria						

Categoría de objetivo: No

Ganancias para nodos						
	Nodo		Ganancia			
Nodo	N	Porcentaje	N	Porcentaje	Respuesta	Índice
1	53	75,7%	49	98,0%	92,5%	129,4%
2	17	24,3%	1	2,0%	5,9%	8,2%
Método de crecimiento: QUEST Variable dependiente: Enfermedad coronaria						

Riesgo						
Estimación	Desv. Error					
,071	,031					
Método de crecimiento: QUEST Variable dependiente: Enfermedad coronaria						

Clasificación			
	Pronosticado		
Observado	Si	No	Porcentaje correcto
Si	16	4	80,0%
No	1	49	98,0%
Porcentaje global	24,3%	75,7%	92,9%
Método de crecimiento: QUEST Variable dependiente: Enfermedad coronaria			

7.7. Dendrograma

Un *dendograma*, es un gráfico que ilustra cómo se van haciendo las subdivisiones o los agrupamientos, etapa a etapa. Partiendo de tantos grupos iniciales como individuos se estudian, se trata de conseguir agrupaciones sucesivas entre ellos de forma que progresivamente se vayan integrando en clústeres los cuales, a su vez, se unirán entre sí en un nivel superior formando grupos mayores que más tarde se juntarán hasta llegar al clúster final que contiene todos los casos analizados, por esta razón, los procedimientos de aglomeración son denominados a veces como métodos de construcción.

Cuando el proceso de obtención de conglomerados procede en dirección opuesta al método de aglomeración, se denomina método divisivo. En los métodos divisivos, empezamos con un gran conglomerado que contiene todas las observaciones

(objetos). En los pasos sucesivos, las observaciones que son más diferentes se dividen y se construyen conglomerados más pequeños. Este proceso continúa hasta que cada observación es un conglomerado en sí mismo.

Todas estas agrupaciones se hacen bajo la concepción de que los elementos que se agrupan están a la misma distancia y este es un concepto clave que complejiza el problema porque hay distintos criterios e distancia como son:

$$\textit{Distancia euclidiana}: \|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$$

$$\textit{Distancia euclidiana al cuadrado}: \|a-b\|_2^{\,2} = \sum_i (a_i - b_i)^2$$

$$\textit{Distancia Manhattan}: \|a-b\|_1 = \sum_i |a_i - b_i|$$

$$\textit{Distancia máxima}: \|a-b\|_\infty = \max_i |a_i - b_i|$$

$$\textit{Distancia Mahalanobis}: \sqrt{(a-b)^T S^{-1}(a-b)} \textit{ donde } S \textit{ es la matriz de covarianza}$$

$$\textit{Similitud coseno}: \frac{a \cdot b}{\|a\|\|b\|}$$

Según estas distancias se construyen los conglomerados según los siguientes algoritmos más habituales:

- *Encadenamiento simple:*

Se basa en la distancia mínima. Encuentra los dos objetos separados por la distancia más corta y las coloca en el primer conglomerado. A continuación, se encuentra la distancia más corta, y o bien un tercer objeto se une a los dos primeros para formar un conglomerado o se forma un nuevo conglomerado de dos miembros. El proceso continúa hasta que todos los objetos se encuentran en un conglomerado. Este procedimiento también se ha denominado como el enfoque del vecino más cercano.

La distancia entre dos conglomerados cualquiera es la distancia más corta desde cualquier punto en un conglomerado a cual-

quier punto en el otro. Dos conglomerados se fusionan en cualquier nivel por el vínculo más corto o más fuerte entre ellos. Fue una regla aplicada en el ejemplo del principio de este capítulo. Los problemas se producen, sin embargo, cuando los conglomerados están mal definidos. En tales casos, los procedimientos de encadenamientos simples pueden formar largas y sinuosas cadenas, y eventualmente todos los individuos pueden situarse en una cadena. Los individuos que se encuentran en los límites opuestos de una cadena pueden ser muy diferentes.

- Encadenamiento completo:

Es parecido al del encadenamiento simple excepto en que el criterio de aglomeración se basa en la distancia máxima. Por esta razón, a veces se le denomina como aproximación del vecino más lejano o método del diámetro.

La distancia máxima entre individuos de cada conglomerado representa la esfera más reducida (diámetro mínimo) que puede incluir todos los objetos en ambos conglomerados. A este método se le denomina encadenamiento completo porque todos los objetos de un conglomerado se vinculan con el resto a alguna distancia máxima o por la mínima similitud. Podemos decir que la similitud dentro del grupo es igual al diámetro del grupo. Esta técnica elimina el problema identificado para el encadenamiento simple.

- Encadenamiento medio:

El método comienza igual que los métodos de encadenamiento simple o completo, pero el criterio de aglomeración es la distancia media de todos los individuos de un conglomerado con todos los individuos de otro.

Tales técnicas no dependen de los valores extremos, como se hace en el encadenamiento simple o completo y la partición se basa en todos los miembros de los conglomerados en lugar de un par único de miembros extremos.

El enfoque del encadenamiento medio tiende a combinar los conglomerados con variaciones reducidas dentro del conglo-

merado. También tiende a estar sesgado hacia la producción de conglomerados con aproximadamente la misma varianza.

- Método de Ward:

En el método de Ward, la distancia entre dos conglomerados es la suma de los cuadrados entre dos conglomerados sumados para todas las variables. Es decir, se calcula la media de todas las variables de cada clúster, luego se calcula la distancia euclídea al cuadrado entre cada individuo y la media de su grupo y después se suman las distancias de todos los casos. En cada paso, los clústeres que se forman son aquéllos que resultan con el menor incremento en la suma total de las distancias al cuadrado intraclúter. Como en los métodos anteriores la métrica utilizada es la euclídea o la euclídea al cuadrado.

- Método del centroide:

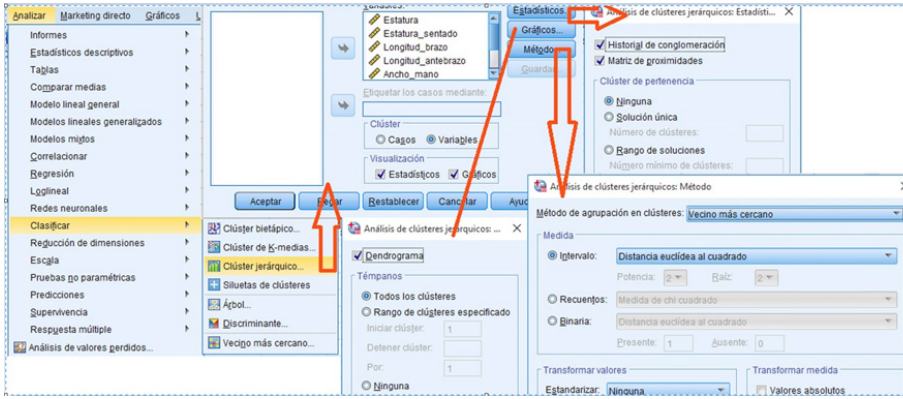
En el método del centroide la distancia entre los dos conglomerados es la distancia (normalmente Euclídea simple o cuadrada) entre sus centroides. Los centroides de los grupos son los valores medios de las observaciones de las variables en el valor teórico del conglomerado, de modo que cada vez que se agrupan los individuos, se calcula un nuevo centroide.

Los centroides de los grupos cambian a medida que se fusionan conglomerados. En otras palabras, existe un cambio en un centroide de un grupo cada vez que un nuevo individuo o grupo de individuos se añade al conglomerado existente.

Estos métodos son más populares entre los biólogos, pero pueden producir resultados desordenados y a menudo confusos. La confusión se produce a causa de los cambios, esto es, casos donde la distancia entre los centroides de un par puede ser menor que la distancia entre los centroides de otro par fusionado en una combinación anterior, pero tiene por ventaja que se ve menos afectada por los valores atípicos que otros métodos jerárquicos.

7.8. Resultados de un análisis mediante dendrograma

El estudio se realizará a partir de la base DIMENSIONES CORPORALES del anexo 4. El inicio del análisis se debe hacer según se muestra en la siguiente lámina:



Para la distancia euclídea y el método del vecino más cercano se tienen los siguientes resultados:

Matriz de proximidades

Caso	Entrada de archivo matricial							
	Estatura	Estatura sentado	Longitud brazo	Longitud antebrazo	Ancho mano	Longitud muslo	Longitud pierna	Longitud pie
Estatura	,000	454,302	759,792	784,700	839,008	702,626	712,375	911,095
Estatura sentado	454,302	,000	306,055	330,882	385,130	249,091	258,843	457,133

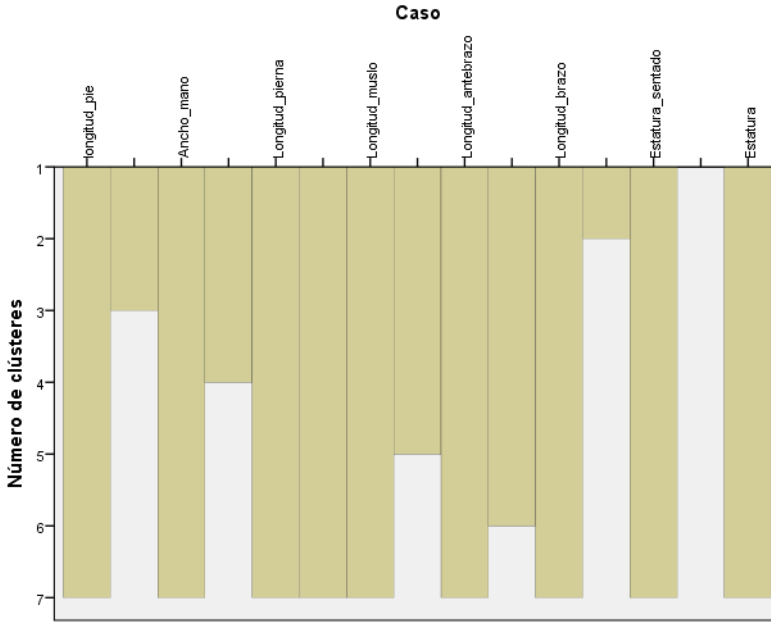
	Longitud pie	Longitud pierna	Longitud muslo	Ancho mano	Longitud antebrazo	Longitud brazo
Longitud pie	911,095	712,375	702,626	839,008	784,700	759,792
	457,133	258,843	249,091	385,130	330,882	306,055
	151,616	48,170	57,837	79,560	26,484	,000
	126,543	72,724	82,992	54,683	,000	26,484
	72,281	126,933	136,774	,000	54,683	79,560
	208,883	13,544	,000	136,774	82,992	57,837
	198,948	,000	13,544	126,933	72,724	48,170
	,000	198,948	208,883	72,281	126,543	151,616

La matriz proporciona las distancias o similaridades entre los elementos.

Historial de conglomeración						
Etapa	Clúster combinado		Coeficientes	Primera aparición del clúster de etapa		Etapa siguiente
	Clúster 1	Clúster 2		Clúster 1	Clúster 2	
1	6	7	13,544	0	0	3
2	3	4	26,484	0	0	3

3	3	6	48,170	2	1	4
4	3	5	54,683	3	0	5
5	3	8	72,281	4	0	6
6	2	3	249,091	0	5	7
7	1	2	454,302	0	6	0

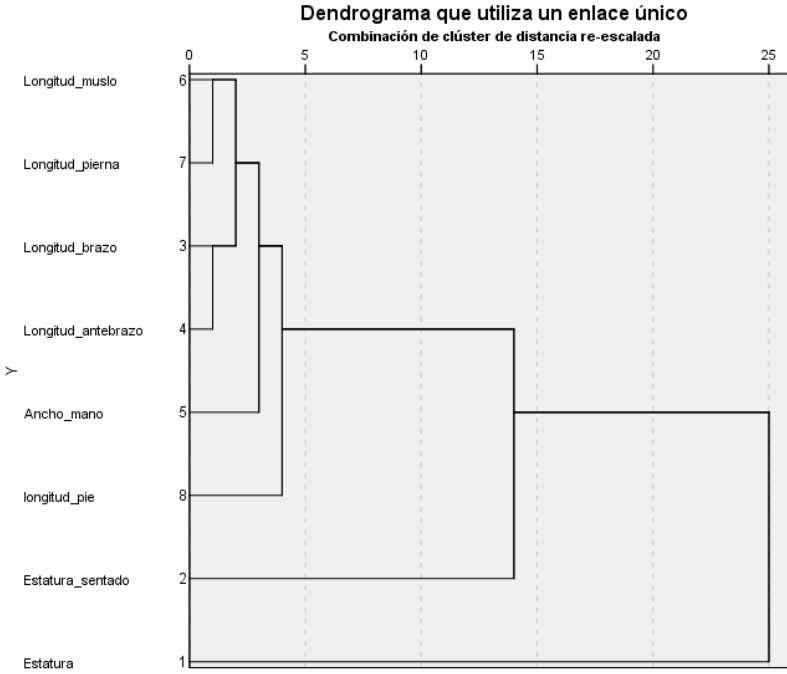
Esta matriz muestra el orden en que se va generando el dendograma.



Este gráfico es conocido como témpanos en el mismo se incluyen todos los clústeres o un rango especificado de clústeres. Los diagramas de témpanos muestran información sobre cómo se combinan los casos en los clústeres, en cada iteración del análisis. La orientación permite seleccionar un diagrama vertical u horizontal.

Como resultado de la matriz de Historial de conglomeración y del gráfico de témpanos se tiene el dendograma que expresa

las relaciones de distancias entre los conjuntos de datos que forman las variables.



En el dendograma se puede observar las relaciones de distancias entre las dimensiones corporales, así, las más *cercanas* son las dimensiones del muslo y la pierna, casi al mismo nivel se dan las dimensiones del brazo y el antebrazo; esos dos pares de "vecinos cercanos" se agrupan en un nuevo conglomerado y así sigue la construcción del dendograma, tal como se indica en la matriz Historial de conglomeración; es recomendable seguir este historial comparándolo con el gráfico.

Con respecto a la distancia euclídea cuadrada y el método de Ward se tienen los siguientes resultados:

Matriz de proximidades

Caso	Entrada de archivo matricial									
	Estatura (1)	Estatura Sentado (2)	Longitud brazo (3)	Longitud antebrazo (4)	Ancho mano (5)	Longitud muslo (6)	Longitud pierna (7)	Longitud pie (8)		
(1)	,000	206390,290	577284,170	615753,840	703934,160	493683,880	507478,470	830094,990		
(2)	206390,290	,000	93669,720	109483,190	148324,890	62046,390	66999,860	208970,820		
(3)	577284,170	93669,720	,000	701,410	6329,850	3345,110	2320,360	22987,380		
(4)	615753,840	109483,190	701,410	,000	2990,240	6887,600	5288,770	16013,170		
(5)	703934,160	148324,890	6329,850	2990,240	,000	18707,220	16111,990	5224,570		

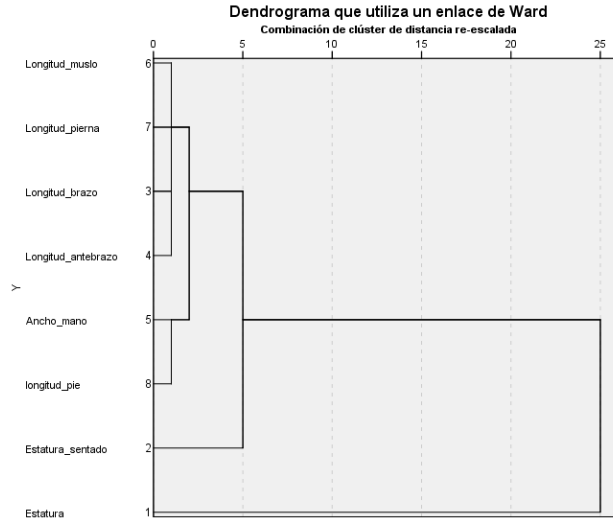
(6)	493683,880	62046,390	3345,110	6887,600	18707,220	,000	183,430	43632,250
(7)	507478,470	66999,860	2320,360	5288,770	16111,990	183,430	,000	39580,280
(8)	830094,990	208970,820	22987,380	16013,170	5224,570	43632,250	39580,280	,000

Historial de conglomeración

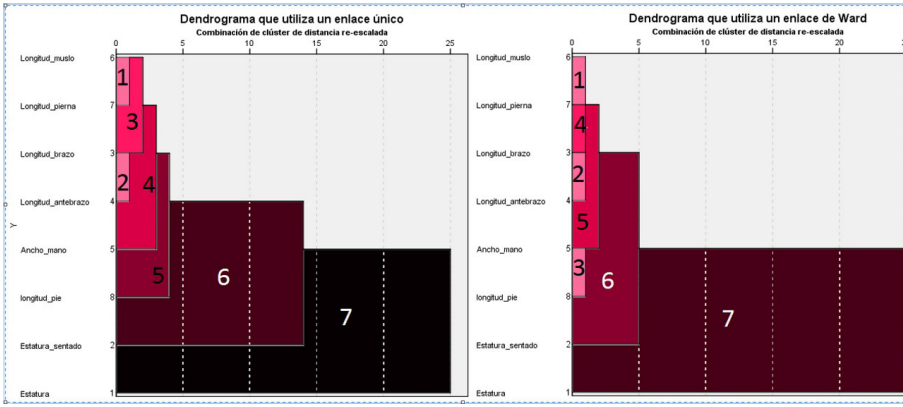
Etapa	Clúster combinado		Coeficientes	Primera aparición del clúster de etapa		Etapa siguiente
	Clúster 1	Clúster 2		Clúster 1	Clúster 2	
1	6	7	91,715	0	0	4
2	3	4	442,420	0	0	4
3	5	8	3054,705	0	0	5

4	3	6	7293,955	2	1	5
5	3	5	31717,272	4	3	6
6	2	3	125685,500	0	5	7
7	1	2	601802,288	0	6	0

Al igual que en el caso anterior se tiene el dendograma según el método de Ward.



La siguiente gráfica ilustra comparativamente el proceso de construcción de los dos dendrograma estudiados según distancias y métodos diferentes.



Los dendrograma también pueden construirse respecto a los casos o individuos estudiados en lugar de las variables como se ha hecho, en este caso los resultados son:

Un fragmento de matriz de proximidades entre los casos se adjunta a continuación:

Matriz de proximidades

Caso	Distancia euclídea														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0,0	6,5	7,1	6,6	11,5	2,6	7,9	2,8	9,0	9,1	5,4	8,0	10,1	8,0	3,1
2	6,5	0,0	4,1	4,1	16,4	5,0	13,7	6,6	14,7	12,5	4,3	7,6	14,6	3,6	4,9
3	7,1	4,1	0,0	3,7	16,3	5,9	14,0	7,9	14,7	11,7	5,0	6,0	14,7	3,5	6,3
4	6,6	4,1	3,7	0,0	15,8	5,5	13,3	6,1	14,0	11,3	5,7	6,7	13,8	5,5	5,3
5	11,5	16,4	16,3	15,8	0,0	11,8	6,2	12,0	3,8	6,4	13,7	12,3	3,1	16,2	12,2

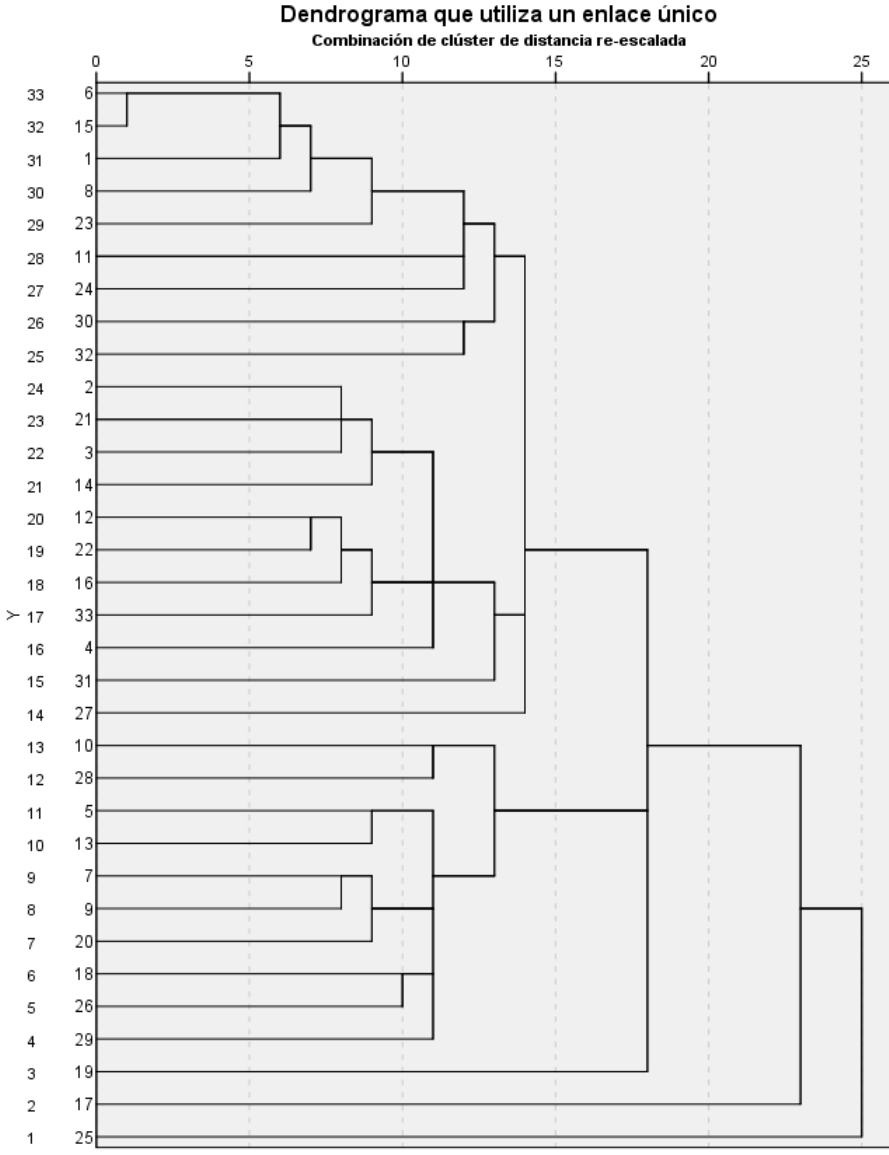
6	2,6	5,0	5,9	5,5	11,8	0,0	8,7	3,4	9,8	8,4	4,2	6,3	10,1	6,1	1,5
7	7,9	13,7	14,0	13,3	6,2	8,7	0,0	8,3	2,8	7,3	11,5	12,0	5,5	14,5	9,1
8	2,8	6,6	7,9	6,1	12,0	3,4	8,3	0,0	9,5	9,3	6,0	8,7	10,2	8,8	2,9
9	9,0	14,7	14,7	14,0	3,8	9,8	2,8	9,5	0,0	6,2	12,2	11,8	3,5	15,1	10,1
10	9,1	12,5	11,7	11,3	6,4	8,4	7,3	9,3	6,2	0,0	10,0	7,3	4,6	12,0	8,9
11	5,4	4,3	5,0	5,7	13,7	4,2	11,5	6,0	12,2	10,0	0,0	6,2	12,4	5,2	5,0
12	8,0	7,6	6,0	6,7	12,3	6,3	12,0	8,7	11,8	7,3	6,2	0,0	10,7	5,7	6,9
13	10,1	14,6	14,7	13,8	3,1	10,1	5,5	10,2	3,5	4,6	12,4	10,7	0,0	14,6	10,3
14	8,0	3,6	3,5	5,5	16,2	6,1	14,5	8,8	15,1	12,0	5,2	5,7	14,6	0,0	6,4
15	3,1	4,9	6,3	5,3	12,2	1,5	9,1	2,9	10,1	8,9	5,0	6,9	10,3	6,4	0,0

Por su parte el historial de conglomeración se muestra en la siguiente tabla.

Historial de conglomeración						
Etapa	Clúster combinado		Coeficientes	Primera aparición del clúster de etapa		Etapa siguiente
	Clúster 1	Clúster 2		Clúster 1	Clúster 2	
1	6	15	1,517	0	0	2
2	1	6	2,627	0	1	4
3	12	22	2,793	0	0	6

4	1	8	2,796	2	0	13
5	7	9	2,830	0	0	9
6	12	16	2,864	3	0	12
7	2	21	2,902	0	0	8
8	2	3	2,966	7	0	10
9	7	20	3,098	5	0	17
10	2	14	3,103	8	0	15
11	5	13	3,150	0	0	19
12	12	33	3,156	6	0	15
13	1	23	3,186	4	0	22
14	18	26	3,233	0	0	16
15	2	12	3,403	10	12	20
16	18	29	3,423	14	0	17
17	7	18	3,507	9	16	19
18	10	28	3,511	0	0	24
19	5	7	3,527	11	17	24
20	2	4	3,533	15	0	25
21	11	24	3,604	0	0	22
22	1	11	3,650	13	21	26
23	30	32	3,734	0	0	26
24	5	10	3,841	19	18	29
25	2	31	3,852	20	0	27
26	1	30	3,948	22	23	27
27	1	2	3,991	26	25	28
28	1	27	4,099	27	0	30
29	5	19	4,716	24	0	30
30	1	5	4,828	28	29	31
31	1	17	5,649	30	0	32
32	1	25	6,198	31	0	0

En correspondencia con estas matrices se obtiene el siguiente dendograma:



7.9. Análisis de correspondencias

El análisis de correspondencias es una técnica descriptiva desarrollada por Jean-Paul Benzécri[xxx]. Se aplica al estudio de tablas de contingencia y es conceptualmente similar al análisis de componentes principales con la diferencia de que en éste los datos se escalan de modo que filas y columnas se tratan de modo equivalente.

El análisis de correspondencias descompone el estadístico del test de la chi-cuadrado asociado a una tabla de contingencia en componentes ortogonales. Dado que se trata de una técnica descriptiva, puede aplicarse hasta en circunstancias en las que una tabla de contingencia no resulta apropiada.

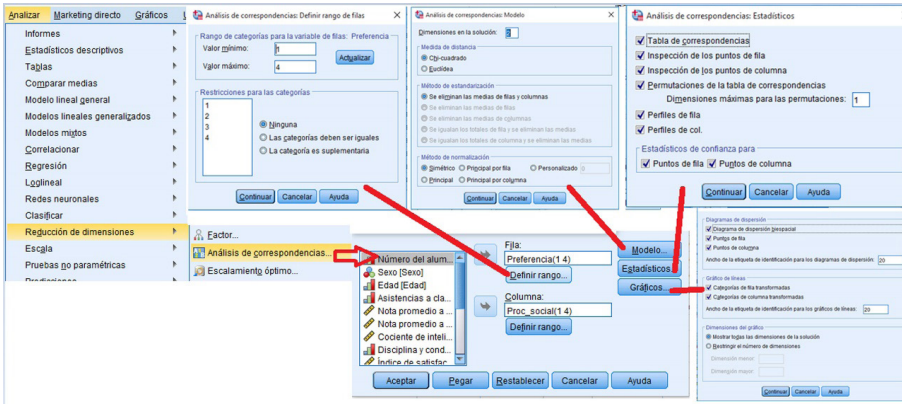
Uno de los objetivos del análisis de correspondencias es describir las relaciones existentes entre dos variables nominales, recogidas en una tabla de correspondencias, sobre un espacio de pocas dimensiones, mientras que al mismo tiempo se describen las relaciones entre las categorías de cada variable. Para cada variable, las distancias sobre un gráfico entre los puntos de categorías reflejan las relaciones entre las categorías, con las categorías similares representadas próximas unas a otras. La proyección de los puntos de una variable sobre el vector desde el origen hasta un punto de categoría de la otra variable describe la relación entre ambas variables.

El análisis de las tablas de contingencia a menudo incluye examinar los perfiles de fila y de columna, así como contrastar la independencia a través del estadístico de chi-cuadrado. Sin embargo, el número de perfiles puede ser bastante grande y la prueba de chi-cuadrado no revelará la estructura de la dependencia. El procedimiento Tablas cruzadas ofrece varias medidas y pruebas de asociación, pero no puede representar gráficamente ninguna relación entre las variables.

Por su parte, con el análisis factorial se describen las relaciones existentes entre variables en un espacio de pocas dimensiones, pero requiere datos de intervalo y el número de observaciones

debe ser cinco veces el número de variables. Por su parte, el análisis de correspondencias asume que las variables son nominales y permite describir las relaciones entre las categorías de cada variable, así como la relación entre las variables. Además, el análisis de correspondencias se puede utilizar para analizar cualquier tabla de medidas de correspondencia que sean positivas.

Para el tratamiento en SPSS el análisis de correspondencia se sigue la secuencia que se muestra en la siguiente figura.



Sobre algunas de las opciones del Menú la ayuda del SSPSS da indicaciones como las siguientes:

Consideraciones sobre los datos: Las variables categóricas que se van a analizar se encuentran escaladas a nivel nominal. Para los datos agregados o para una medida de correspondencia distinta de las frecuencias, utilice una variable de ponderación con valores de similaridad positivos.

Supuestos. El máximo número de dimensiones utilizado en el procedimiento depende del número de categorías activas de fila y de columna y del número de restricciones de igualdad. Si no se utilizan criterios de igualdad y todas las categorías son activas, la dimensionalidad máxima es igual al número de categorías de la variable con menos categorías menos uno.

Por ejemplo, si una variable dispone de cinco categorías y la otra de cuatro, el número máximo de dimensiones es tres. Las categorías suplementarias no son activas. Por ejemplo, si una variable dispone de cinco categorías, dos de las cuales son suplementarias, y la otra variable dispone de cuatro categorías, el número máximo de dimensiones es dos. Considere todos los conjuntos de categorías con restricción de igualdad como una única categoría. Por ejemplo, si una variable dispone de cinco categorías, tres de las cuales tienen restricción de igualdad, dicha variable se debe tratar como si tuviera tres categorías en el momento de calcular la dimensionalidad máxima. Dos de las categorías no tienen restricción y la tercera corresponde a las tres categorías restringidas. Si se especifica un número de dimensiones superior al máximo, se utilizará el valor máximo.

Modelo: Con este botón se especifica el número de dimensiones, la medida de distancia, el método de estandarización y el método de normalización.

Medida de distancia: por defecto se usa la distancia chi-cuadrado.

Método de normalización: esta es una de las decisiones más importantes, ya que, dependiendo del método, se producirán soluciones que, aunque equivalentes, pueden ser diferentes. Se usarán:

1. Simétrico: en este caso la inercia se reparte igualmente entre filas y columnas. Se usa este método para examinar las diferencias entre las categorías de las dos variables.
2. Principal: se utilizará este método si se desea examinar las diferencias entre las categorías de una o de ambas variables en lugar de las diferencias entre las dos variables.
3. Principal por fila: este método se usa para examinar las diferencias entre las categorías de la variable de filas.
4. Principal por columna: para examinar las diferencias entre las categorías de la variable de columnas.
5. Personalizado: Otro método que defina el usuario.

Estadísticos: El cuadro de diálogo Estadísticos permite especificar los resultados numéricos producidos.

- Tabla de correspondencias. Es la tabulación cruzada de las variables de entrada con los totales marginales de fila y columna.
- Inspección de los puntos de fila. Para cada categoría de fila, las puntuaciones, la masa, la inercia, la contribución a la inercia de la dimensión y la contribución de la dimensión a la inercia del punto.
- Inspección de los puntos de columna. Para cada categoría de columna, las puntuaciones, la masa, la inercia, la contribución a la inercia de la dimensión y la contribución de la dimensión a la inercia del punto.
- Perfiles de fila. Para cada categoría de fila, la distribución a través de las categorías de la variable de columna.
- Perfiles de col. Para cada categoría de columna, la distribución a través de las categorías de la variable de fila.
- Permutaciones de la tabla de correspondencias. La tabla de correspondencias reorganizada de tal manera que las filas y las columnas estén en orden ascendente de acuerdo con las puntuaciones en la primera dimensión. Si lo desea, puede especificar el número de la dimensión máxima para el que se generarán las tablas permutadas. Se generará una tabla permutada para cada dimensión desde 1 hasta el número especificado.
- Estadísticos de confianza para puntos de fila. Incluye la desviación estándar y las correlaciones para todos los puntos de fila no suplementarios.
- Estadísticos de confianza para puntos de columna. Incluye la desviación estándar y las correlaciones para todos los puntos de columna no suplementarios.

El cuadro de diálogo Gráficos permite especificar qué gráficos se van a generar. Diagramas de dispersión. Produce una matriz de todos los gráficos por parejas de las dimensiones. Los diagramas de dispersión disponibles incluyen:

- Diagrama de dispersión biespacial. Produce una matriz de diagramas conjuntos de los puntos de fila y de columna. Si está seleccionada la normalización principal, el diagrama de dispersión biespacial no estará disponible.
- Puntos de fila. Produce una matriz de diagramas de los puntos de fila.
- Puntos de columna. Produce una matriz de diagramas de los puntos de columna.

Si lo desea, puede especificar el número de caracteres de etiqueta de valor que se va a utilizar al etiquetar los puntos. Este valor debe ser un entero no negativo menor o igual que 20.

- Gráficos de línea. Produce un gráfico para cada dimensión de la variable seleccionada. Los gráficos de líneas disponibles incluyen:
    - Categorías de fila transformadas. Produce un gráfico de los valores originales para las categorías de fila frente a las puntuaciones de fila correspondientes.
    - Categorías de columna transformadas. Produce un gráfico de los valores originales para las categorías de columna frente a las puntuaciones de columna correspondientes.

Si lo desea, puede especificar el número de caracteres de etiqueta de valor que se va a utilizar al etiquetar los ejes de categorías. Este valor debe ser un entero no negativo menor o igual que 20.

Dimensiones del gráfico. Permite controlar las dimensiones que se muestran en los resultados.

- Muestra todas las dimensiones de la solución. Todas las dimensiones de la solución se muestran en un diagrama de dispersión matricial.
- Restringe el número de dimensiones Las dimensiones mostradas se restringen a los pares representados. Si restringe las dimensiones, deberá seleccionar las dimensiones menor y mayor que se van a representar. La dimensión menor puede variar desde 1 hasta el número de dimensiones de la solución menos 1 y se representa respecto a las dimensiones mayores. El valor de la dimensión mayor puede oscilar variar desde 2 hasta el número de dimensiones de la solución e indica la dimensión mayor que se utilizará al representar los pares de dimensiones. Esta especificación se aplica a todos los gráficos multidimensionales solicitados.

7.10. Resultados de un análisis mediante análisis de correspondencia

De la base de datos "PROBLEMAA BASE" (Anexo 2) se analizará la correspondencia entre las variables "Área de preferencia" y "Procedencia social".

Tabla de correspondencias

Área de preferencia	Procedencia social				
	Obrera	Campesina	Intelectual	Clase media-alta	Margen activo
C_exactas	2	4	1	6	13
C_naturales	0	0	1	5	6
C_sociales	0	0	1	5	6
C_humanísticas	5	1	2	7	15
Margen activo	7	5	5	23	40

En esta tabla se muestra la frecuencia absoluta observada.

Perfiles de fila					
Área de preferencia	Procedencia social				
	Obrera	Campesina	Intelectual	Clase media-alta	Margen activo
C_exactas	,154	,308	,077	,462	1,000
C_naturales	,000	,000	,167	,833	1,000
C_sociales	,000	,000	,167	,833	1,000
C_humanísticas	,333	,067	,133	,467	1,000
Masa	,175	,125	,125	,575	

En esta tabla se muestra la frecuencia relativa observada por filas.

Perfiles de columna					
Área de preferencia	Procedencia social				
	Obrera	Campesina	Intelectual	Clase media-alta	Masa
C_exactas	,286	,800	,200	,261	,325
C_naturales	,000	,000	,200	,217	,150
C_sociales	,000	,000	,200	,217	,150
C_humanísticas	,714	,200	,400	,304	,375
Margen activo	1,000	1,000	1,000	1,000	

En esta tabla se muestra la frecuencia relativa observada por columnas.

Resumen						
Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Proporción de inercia	
					Contabilizado para	Acumulado
1	,437	,191			,634	,634
2	,332	,110			,366	1,000
Total		,302	12,067	,210a	1,000	1,000

Resumen											
Dimensión	Valor singular de confianza										
	Desviación estándar			Correlación							
		2									
1	,090			,139							
2	,160										
Total											
a. 9 grados de libertad											
Puntos de fila generalesa											
Área de preferencia	Masa		Puntuación en dimensión			Inercia		Contribución			
		1	2			Del punto en la inercia de dimensión					
				1							
C_exactas	,325		,662	-,597		,101		,326			
C_naturales	,150		-,966	-,258		,064		,320			
C_sociales	,150		-,966	-,258		,064		,320			
C_humanísticas	,375		,199	,724		,072		,034			
Total activo	1,000					,302		1,000			
Puntos de fila generalesa											
Área de preferencia		Contribución									
	Del punto en la inercia de dimensión		De la dimensión en la inercia del punto								

	2	1	2	Total
C_exactas	,349	,618	,382	1,000
C_naturales	,030	,949	,051	1,000
C_sociales	,030	,949	,051	1,000
C_humanísticas	,591	,090	,910	1,000
Total activo	1,000			

a. Normalización simétrica

Puntos de columna generales[a]

Procedencia social	Masa	Puntuación en dimensión		Inercia	Contribución
		1	2		Del punto en la inercia de dimensión
					1
Obrera	,175	,757	1,042	,107	,230
Campesina	,125	1,303	-1,002	,134	,485
Intelectual	,125	-,399	,202	,010	,045
Clase media-alta	,575	-,427	-,143	,050	,240
Total activo	1,000			,302	1,000

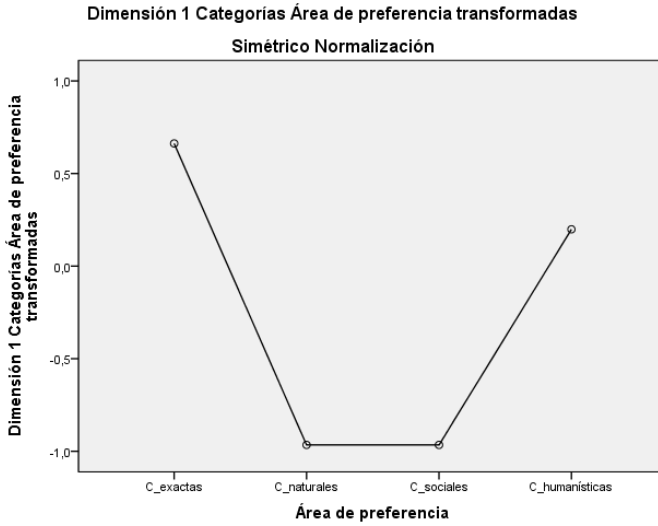
Puntos de columna generalesa

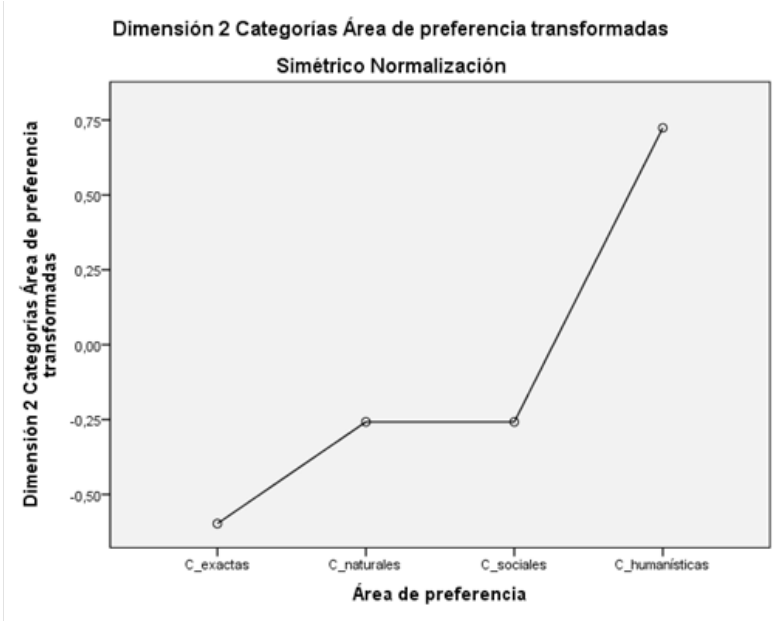
Procedencia social	Contribución			
	Del punto en la inercia de dimensión	De la dimensión en la inercia del punto		
	2	1	2	Total
Obrera	,572	,410	,590	1,000
Campesina	,378	,690	,310	1,000
Intelectual	,015	,837	,163	1,000
Clase media-alta	,035	,921	,079	1,000
Total activo	1,000			

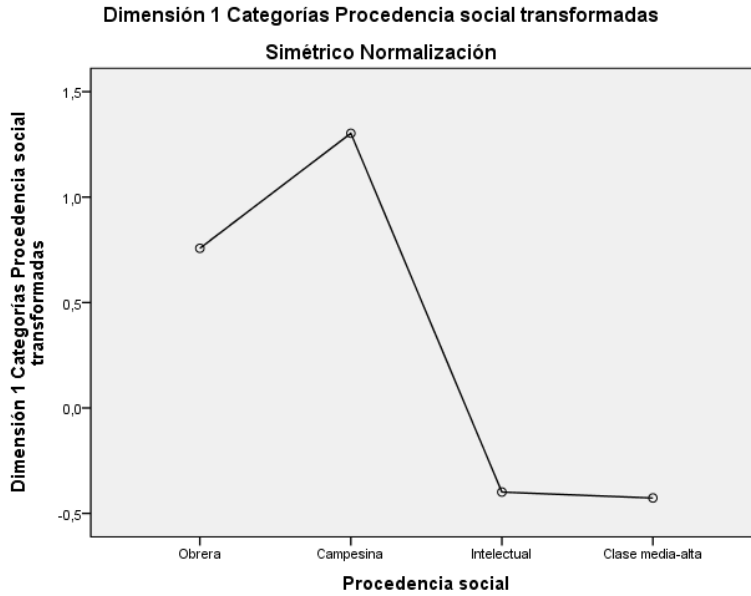
a. Normalización simétrica

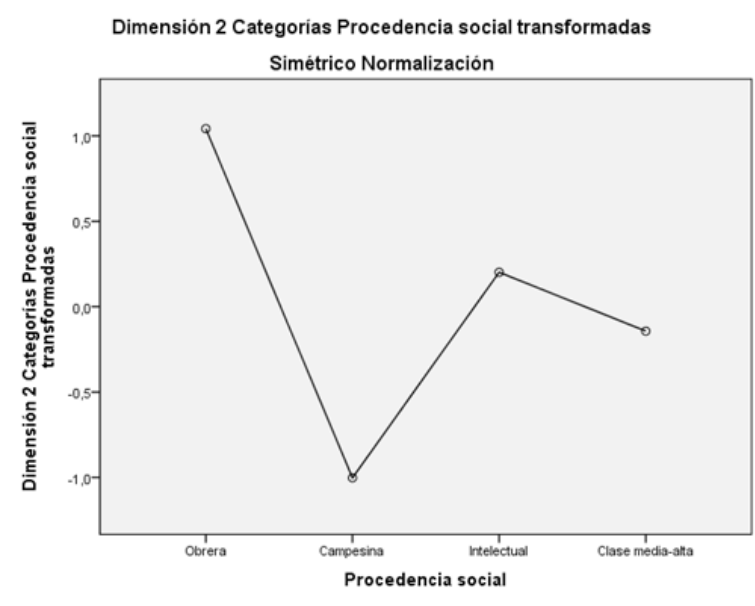
Puntos de fila de confianza			
Área de preferencia	Desviación estándar en la dimensión		Correlación
	1	2	1-2
C_exactas	,645	,524	,945
C_naturales	,243	,772	-,842
C_sociales	,243	,772	-,842
C_humanísticas	,769	,266	-,625

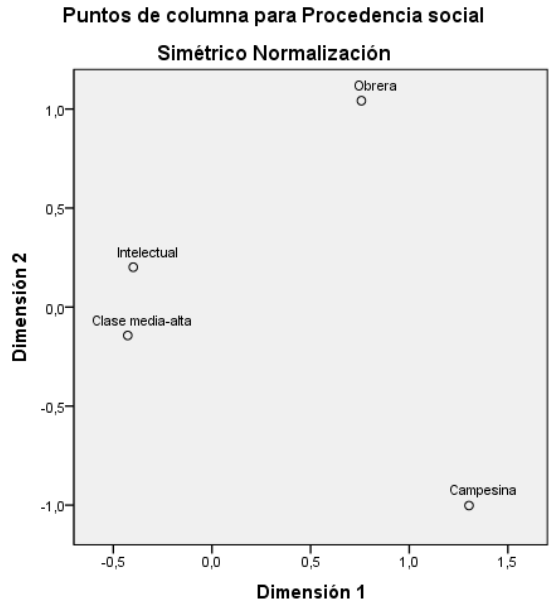
Puntos de columna de confianza			
Procedencia social	Desviación estándar en la dimensión		Correlación
	1	2	1-2
Obrera	1,021	,634	-,867
Campesina	,978	,826	,943
Intelectual	,231	,228	,758
Clase media-alta	,157	,264	-,637

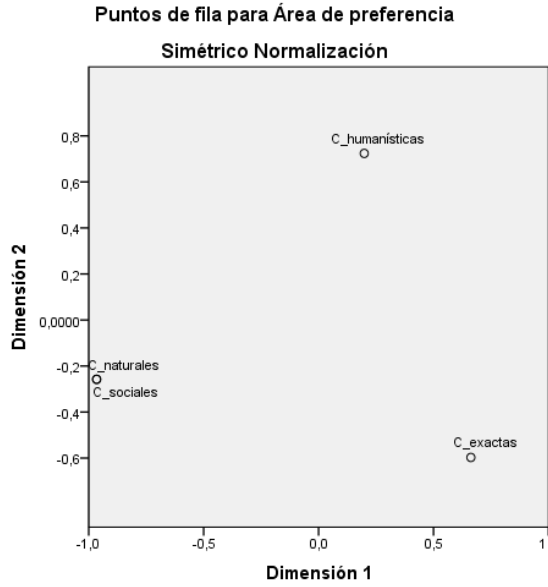


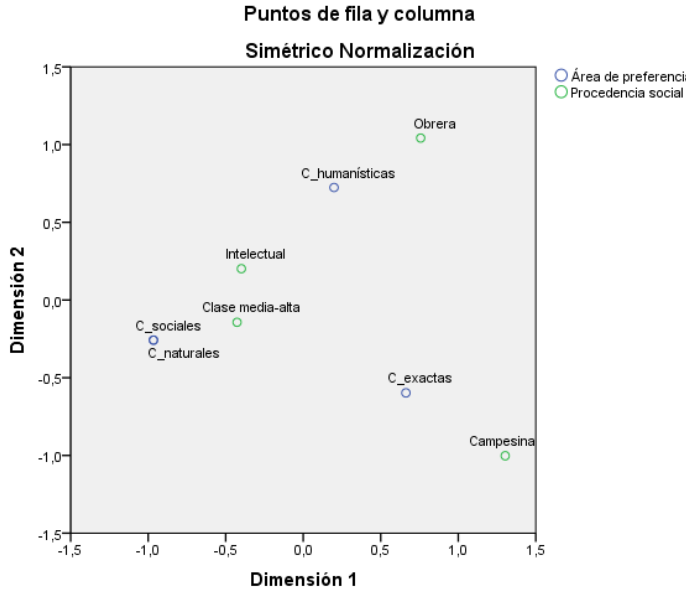

Dimensión 2 Categorías Área de preferencia transformadas
Simétrico Normalización
Dimensión 2 Categorías Área de preferencia transformadas
0,75
0,50
0,25
0,00
-0,25
-0,50
C_exactas
C_naturales
C_sociales
C_humanísticas
Área de preferencia


Dimensión 1 Categorías Procedencia social transformadas
Simétrico Normalización
Dimensión 1 Categorías Procedencia social transformadas
1,5
1,0
0,5
0,0
-0,5
Obrera
Campesina
Intelectual
Clase media-alta
Procedencia social


Dimensión 2 Categorías Procedencia social transformadas
Simétrico Normalización
Dimensión 2 Categorías Procedencia social transformadas
1,0
0,5
0,0
-0,5
-1,0
Obrera
Campesina
Intelectual
Clase media-alta
Procedencia social


Puntos de columna para Procedencia social
Simétrico Normalización
Obrera
Intelectual
Clase media-alta
Campesina
Dimensión 2
1,0
0,5
0,0
-0,5
-1,0
-0,5
0,0
0,5
1,0
1,5
Dimensión 1


Puntos de fila para Área de preferencia
Simétrico Normalización
C_humanísticas
C_naturales
C_sociales
C_exactas
Dimensión 2
Dimensión 1
0,8
0,6
0,4
0,2
0,0000
-0,2
-0,4
-0,6
-1,0
-0,5
0,0
0,5
1,0


Puntos de fila y columna
Simétrico Normalización
Área de preferencia
Procedencia social
Obrera
C_humanísticas
Intelectual
Clase media-alta
C_sociales
C_naturales
C_exactas
Campesina
Dimensión 2
Dimensión 1
1,5
1,0
0,5
0,0
-0,5
-1,0
-1,5

Referencias bibliográficas

Álvarez Cáceres, R. (1995). *Estadística multivariante y no paramétrica con SPSS Aplicación a las ciencias de la salud.* Madrid: Ediciones Díaz de Santos, S.A.

Batanero, C., & Díaz, C. (2011). ***Estadística con proyectos***. Granada: Universidad de Granada.

Batanero, C., & Godino, J. D. (2001). *Análisis de datos y su didáctica*. Granada: Universidad de Granada.

Camacho Rosales, J. (2001). *Estadistica con SPSS (versión 9) para Windows.* México: Alfaomega Grupo Editor.

Castañeda, M. B., Cabrera, A. F., & Navarro, Y. &. (2010). *Procesamiento de datos y análisis estadísticos utilizando SPSS*. Porto Alegre: EDIPUCRS.

Freixa, M., Salafranca, L., Guàrdia, J., Ferrer, R., & i Turbany, J. (1992). *Análisis exploratorio de datos: nuevas técnicas estadísticas*. Barcelona: PPU.

Hair, J., Anderson, R., Tatham, R., & Black, W. (1999). *Análisis multivariante*. Madrid: Prentice Hall Iberia. S.R.L.

International Business Machines. (2012). *Manual del usuario del sistema básico de IBM SPSS Statistics 21.* Nueva York: IBM.

Monterde i Bort, H., & Perea Lara, M. (1991). *El enfoque del análisis exploratorio de datos.* Valencia: Benetusser.

Pérez López, C. (2004). *Técnicas de Análisis Multivariante de Datos*. Madrid: Pearson Educación, S.A..

Pérez-Medinilla, Y. T., Crespo Borges, T., & Ríos-Rodríguez, L. R. (Noviembre-Diciembre de 2015). Análisis exploratorio de datos a través de mapas conceptuales. *Revista IPLAC*, 96-105.

Silva Rodríguez, M. (2002). *Pedagogía, tablas de contingencia y validación de hipótesis científico-pedagógicas*. La Habana: Pueblo y Educación.

Tukey, J. V. (1997). *Fundamentals of exploratory analysis of variance*. New York: A Wiley-Interscience Publication.

Tukey, J. W. (1977). *Exploratory Data Analysis*. New York: Addison-Wesley Publishing Company.

Anexos

Anexo 1. Base HATCO

Número	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14
1	4,1	0,6	6,9	4,7	2,4	2,3	5,2	0	32	4,2	1	0	1	1
2	1,8	3,0	6,3	6,6	2,5	4,0	8,4	1	43	4,3	0	1	0	1
3	3,4	5,2	5,7	6,0	4,3	2,7	8,2	1	48	5,2	0	1	1	2
4	2,7	1,0	7,1	5,9	1,8	2,3	7,8	1	32	3,9	0	1	1	1
5	6,0	0,9	9,6	7,8	3,4	4,6	4,5	0	58	6,8	1	0	1	3
6	1,9	3,3	7,9	4,8	2,6	1,9	9,7	1	45	4,4	0	1	1	2
7	4,6	2,4	9,5	6,6	3,5	4,5	7,6	0	46	5,8	1	0	1	1
8	1,3	4,2	6,2	5,1	2,8	2,2	6,9	1	44	4,3	0	1	0	2
9	5,5	1,6	9,4	4,7	3,5	3,0	7,6	0	63	5,4	1	0	1	3
10	4,0	3,5	6,5	6,0	3,7	3,2	8,7	1	54	5,4	0	1	0	2
11	2,4	1,6	8,8	4,8	2,0	2,8	5,8	0	32	4,3	1	0	0	1
12	3,9	2,2	9,1	4,6	3,0	2,5	8,3	0	47	5,0	1	0	1	2
13	2,8	1,4	8,1	3,8	2,1	1,4	6,6	1	39	4,4	0	1	0	1
14	3,7	1,5	8,6	5,7	2,7	3,7	6,7	0	38	5,0	1	0	1	1
15	4,7	1,3	9,9	6,7	3,0	2,6	6,8	0	54	5,9	1	0	0	3
16	3,4	2,0	9,7	4,7	2,7	1,7	4,8	0	49	4,7	1	0	0	3
17	3,2	4,1	5,7	5,1	3,6	2,9	6,2	0	38	4,4	1	1	1	2
18	4,9	1,8	7,7	4,3	3,4	1,5	5,9	0	40	5,6	1	0	0	2
19	5,3	1,4	9,7	6,1	3,3	3,9	6,8	0	54	5,9	1	0	1	3
20	4,7	1,3	9,9	6,7	3,0	2,6	6,8	0	55	6,0	1	0	0	3
21	3,3	0,9	8,6	4,0	2,1	1,8	6,3	0	41	4,5	1	0	0	2
22	3,4	0,4	8,3	2,5	1,2	1,7	5,2	0	35	3,3	1	0	0	1
23	3,0	4,0	9,1	7,1	3,5	3,4	8,4	0	55	5,2	1	1	0	3
24	2,4	1,5	6,7	4,8	1,9	2,5	7,2	1	36	3,7	0	1	0	1

25	5,1	1,4	8,7	4,8	3,3	2,6	3,8	0	49	4,9	1	0	0	2
26	4,6	2,1	7,9	5,8	3,4	2,8	4,7	0	49	5,9	1	0	1	3
27	2,4	1,5	6,6	4,8	1,9	2,5	7,2	1	36	3,7	0	1	0	1
28	5,2	1,3	9,7	6,1	3,2	3,9	6,7	0	54	5,8	1	0	1	3
29	3,5	2,8	9,9	3,5	3,1	1,7	5,4	0	49	5,4	1	0	1	3
10	4,1	3,7	5,9	5,5	3,9	3,0	8,4	1	46	5,1	0	1	0	2
31	3,0	3,2	6,0	5,3	3,1	3,0	8,0	1	43	3,3	0	1	0	1
32	2,8	3,8	8,9	6,9	3,3	3,2	8,2	0	53	5,0	1	1	0	3
33	5,2	2,0	9,3	5,9	3,7	2,4	4,6	0	60	6,1	1	0	0	3
34	3,4	3,7	6,4	5,7	3,5	3,4	8,4	1	47	3,8	0	1	0	1
35	2,4	1,0	7,7	3,4	1,7	1,1	6,2	1	35	4,1	0	1	0	1
36	1,8	3,3	7,5	4,5	2,5	2,4	7,6	1	39	3,6	0	1	1	1
37	3,6	4,0	5,8	5,8	3,7	2,5	9,3	1	44	4,8	0	1	1	2
38	4,0	0,9	9,1	5,4	2,4	2,6	7,3	0	46	5,1	1	0	1	3
39	0,0	2,1	6,9	5,4	1,1	2,6	8,9	1	29	3,9	0	1	1	1
40	2,4	2,0	6,4	4,5	2,1	2,2	8,8	1	28	3,3	0	1	1	1
41	1,9	3,4	7,6	4,6	2,6	2,5	7,7	1	40	3,7	0	1	1	1
42	5,9	0,9	9,6	7,8	3,4	4,6	4,5	0	58	6,7	1	0	1	3
43	4,9	2,3	9,3	4,5	3,6	1,3	6,2	0	53	5,9	1	0	0	3
44	5,0	1,3	8,6	4,7	3,1	2,5	3,7	0	48	4,8	1	0	0	2
45	2,0	2,6	6,5	3,7	2,4	1,7	8,5	1	38	3,2	0	1	1	1
46	5,0	2,5	9,4	4,6	3,7	1,4	6,3	0	54	6,0	1	0	0	3
47	3,1	1,9	10,0	4,5	2,6	3,2	3,8	0	55	4,9	1	0	1	3
48	3,4	3,9	5,6	5,6	3,6	2,3	9,1	1	43	4,7	0	1	1	2
49	5,8	0,2	8,8	4,5	3,0	2,4	6,7	0	57	4,9	1	0	1	3
50	5,4	2,1	8,0	3,0	3,8	1,4	5,2	0	53	3,8	1	0	1	3
51	3,7	0,7	8,2	6,0	2,1	2,5	5,2	0	41	5,0	1	0	0	2
52	2,6	4,8	8,2	5,0	3,6	2,5	9,0	1	53	5,2	0	1	1	2
53	4,5	4,1	6,3	5,9	4,3	3,4	8,8	1	50	5,5	0	1	0	2
54	2,8	2,4	6,7	4,9	2,5	2,6	9,2	1	32	3,7	0	1	1	1

55	3,8	0,8	8,7	2,9	1,6	2,1	5,6	0	39	3,7	1	0	0	1
56	2,9	2,6	7,7	7,0	2,8	3,6	7,7	0	47	4,2	1	1	1	2
57	4,9	4,4	7,4	6,9	4,6	4,0	9,6	1	62	6,2	0	1	0	2
58	5,4	2,5	9,6	5,5	4,0	3,0	7,7	0	65	6,0	1	0	0	3
59	4,3	1,8	7,6	5,4	3,1	2,5	4,4	0	46	5,6	1	0	1	3
60	2,3	4,5	8,0	4,7	3,3	2,2	8,7	1	50	5,0	0	1	1	2
61	3,1	1,9	9,9	4,5	2,6	3,1	3,8	0	54	4,8	1	0	1	3
62	5,1	1,9	9,2	5,8	3,6	2,3	4,5	0	60	6,1	1	0	0	3
63	4,1	1,1	9,3	5,5	2,5	2,7	7,4	0	47	5,3	1	0	1	3
64	3,0	3,8	5,5	4,9	3,4	2,6	6,0	0	36	4,2	1	1	1	2
65	1,1	2,0	7,2	4,7	1,6	3,2	10,0	1	40	3,4	0	1	1	1
66	3,7	1,4	9,0	4,5	2,6	2,3	6,8	0	45	4,9	1	0	0	2
67	4,2	2,5	9,2	6,2	3,3	3,9	7,3	0	59	6,0	1	0	0	3
68	1,6	4,5	6,4	5,3	3,0	2,5	7,1	1	46	4,5	0	1	0	2
69	5,3	1,7	8,5	3,7	3,5	1,9	4,8	0	58	4,3	1	0	0	3
70	2,3	3,7	8,3	5,2	3,0	2,3	9,1	1	49	4,8	0	1	1	2
71	3,6	5,4	5,9	6,2	4,5	2,9	8,4	1	50	5,4	0	1	1	2
72	5,6	2,2	8,2	3,1	4,0	1,6	5,3	0	55	3,9	1	0	1	3
73	3,6	2,2	9,9	4,8	2,9	1,9	4,9	0	51	4,9	1	0	0	3
74	5,2	1,3	9,1	4,5	3,3	2,7	7,3	0	60	5,1	1	0	1	3
75	3,0	2,0	6,6	6,6	2,4	2,7	8,2	1	41	4,1	0	1	0	1
76	4,2	2,4	9,4	4,9	3,2	2,7	8,5	0	49	5,2	1	0	1	2
77	3,8	0,8	8,3	6,1	2,2	2,6	5,3	0	42	5,1	1	0	0	2
78	3,3	2,6	9,7	3,3	2,9	1,5	5,2	0	47	5,1	1	0	1	3
79	1,0	1,9	7,1	4,5	1,5	3,1	9,9	1	39	3,3	0	1	1	1
80	4,5	1,6	8,7	4,6	3,1	2,1	6,8	0	56	5,1	1	0	0	3
81	5,5	1,8	8,7	3,8	3,6	2,1	4,9	0	59	4,5	1	0	0	3
82	3,4	4,6	5,5	8,2	4,0	4,4	6,3	0	47	5,6	1	1	1	2
83	1,6	2,8	6,1	6,4	2,3	3,8	8,2	1	41	4,1	0	1	0	1
84	2,3	3,7	7,6	5,0	3,0	2,5	7,4	0	37	4,4	1	1	0	1

85	2,6	3,0	8,5	6,0	2,8	2,8	6,8	1	53	5,6	0	1	0	2
86	2,5	3,1	7,0	4,2	2,8	2,2	9,0	1	43	3,7	0	1	1	1
87	2,4	2,9	8,4	5,9	2,7	2,7	6,7	1	51	5,5	0	1	0	2
88	2,1	3,5	7,4	4,8	2,8	2,3	7,2	0	36	4,3	1	1	0	1
89	2,9	1,2	7,3	6,1	2,0	2,5	8,0	1	34	4,0	0	1	1	1
90	4,3	2,5	9,3	6,3	3,4	4,0	7,4	0	60	6,1	1	0	0	3
91	3,0	2,8	7,8	7,1	3,0	3,8	7,9	0	49	4,4	1	1	1	2
92	4,8	1,7	7,6	4,2	3,3	1,4	5,8	0	39	5,5	1	0	0	2
93	3,1	4,2	5,1	7,8	3,6	4,0	5,9	0	43	5,2	1	1	1	2
94	1,9	2,7	5,0	4,9	2,2	2,5	8,2	1	36	3,6	0	1	0	1
95	4,0	0,5	6,7	4,5	2,2	2,1	5,0	0	31	4,0	1	0	1	1
96	0,6	1,6	6,4	5,0	0,7	2,1	8,4	1	25	3,4	0	1	1	1
97	6,1	0,5	9,2	4,8	3,3	2,8	7,1	0	60	5,2	1	0	1	3
98	2,0	2,8	5,2	5,0	2,4	2,7	8,4	1	38	3,7	0	1	0	1
99	3,1	2,2	6,7	6,8	2,6	2,9	8,4	1	42	4,3	0	1	0	1
100	2,5	1,8	9,0	5,0	2,2	3,0	6,0	0	33	4,4	1	0	0	1

x1	Velocidad de entrega	x2	Nivel de precios
x2	Nivel de precios	x3	Flexibilidad de precios
x4	Imagen del fabricante	x5	Servicio conjunto
x6	imagen de fuerza de ventas	x7	Calidad de producto
x8	Tamaño de empresa	1 = grande y O = pequeña	
x9	Nivel de fidelidad	x10	Nivel de satisfacción
x11	Compra al detalle	1 = emplea la aproximación al análisis del valor total, evaluando cada compra por separado y O = uso de la compra detallada.	

x12	Estructura de adquisición	1 = adquisición centralizada y 0 = adquisición descentralizada.
x13	Tipo de industria	1 = industria de la clase A y 0 = otras industrias.
x14	Tipo de situación de compra	1 =nueva tarea, 2 =recompra similar modificada y 3 =recompra simple

Anexo 2. Problema base

Alumno #	Sexo	Edad	Preferencia	AC60	NPIS	NPFS	CI	PS	DC	.ISF	ISE
1	1	15	1	49	8,9	9,1	107	2	2	0,67	0,74
2	2	17	3	44	6,4	6,4	88	1	4	0,74	0,71
3	2	16	3	59	6,3	6	106	4	4	0,83	0,82
4	2	16	1	45	6,6	6,1	106	2	1	0,91	0,65
5	2	16	4	30	7,6	7,3	100	4	4	0,51	0,73
6	1	17	3	35	10	10	88	4	1	0,78	0,67
7	2	18	1	37	8,8	8,3	89	1	4	0,78	0,71
8	1	15	4	41	8,6	8,6	100	4	3	0,9	0,64
9	2	17	1	51	7,3	7,5	88	4	1	0,66	0,81
10	2	18	4	45	6,5	6,3	100	1	3	0,95	0,75
11	2	16	3	33	8,1	8,3	106	4	4	0,69	0,77
12	1	17	4	41	9,1	9,3	100	4	2	0,53	0,96
13	1	18	2	53	7,8	7,3	89	4	1	0,81	0,83
14	1	17	4	33	9,8	9,5	106	1	0	0,7	0,89
15	1	16	3	39	10	9,9	94	4	2	0,51	0,81
16	1	15	2	55	6,4	6,5	107	4	1	0,72	0,91
17	1	17	4	59	9,5	9,4	94	3	4	0,86	0,6
18	2	18	1	42	9,8	9,5	89	4	1	0,52	0,74
19	2	16	2	46	8,2	7,7	100	4	4	0,65	0,79
20	1	17	4	49	8,6	9	88	3	4	0,74	0,97
21	1	17	4	43	10	10	100	4	3	0,9	0,84
22	1	18	1	43	9,3	9,7	94	4	4	0,77	0,69
23	2	16	4	45	7,3	7,1	113	1	4	0,99	0,96
24	2	18	1	49	6	6	100	3	4	0,92	0,72

25	2	16	1	30	10	10	113	2	0	0,74	0,97
26	1	17	2	40	7,9	7,9	100	3	2	0,96	0,7
27	1	15	2	35	6,3	6,1	120	4	2	0,89	0,94
28	1	15	4	45	6,8	6,4	120	4	0	0,56	0,64
29	1	15	4	55	6,5	6,8	107	4	2	0,99	0,68
30	2	18	3	43	9	9	100	4	4	0,87	0,65
31	2	18	3	49	7,7	8	83	3	4	0,9	0,89
32	1	15	1	40	6,3	6,8	113	4	4	0,96	0,75
33	2	17	1	46	9	9,2	88	2	0	0,59	0,6
34	2	18	1	30	6,8	6,8	83	4	3	0,97	0,9
35	1	15	4	58	10	9,9	113	1	2	0,78	0,92
36	1	17	2	49	9,8	9,7	106	4	2	0,72	0,94
37	1	17	4	31	7,5	7,4	100	2	1	0,57	0,6
38	2	17	4	58	6,2	6,2	106	4	3	0,75	0,82
39	1	16	4	40	9,9	9,4	100	1	4	0,53	0,67
40	2	17	1	45	7,2	7,6	88	4	4	0,69	0,96

Sexo	1: MASCULINO; 2: FEMENINO		
Preferencias	1: C_EXACTAS; 2:C_NATURALES; 3:C_SOCIALES 4:C_HUMANÍSTICAS		
PS	1: OBRERA ; 2: CAMPESINA; 3: INTELECTUAL; 4: CLASE MEDIA		
DC	0: MUY MALA; 1: MALA; 2: REGULAR; 3: BUENA; 4: MUY BUENA	CI	Cociente de inteligencia
PS	Procedencia social	DC	Disciplina y conducta
ISF	Índice satisfacción con familia	ISE	Índice satisfacción con escuela

Anexo 3. Enfermedades coronarias

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16
1	42	1	1	292	60	230	170	97	1	1	1	98	176	3	1
2	64	0	2	235	34	134	150	90	0	0	1	74	155	2	1
3	47	1	2	200	53	245	140	80	0	1	0	74	170	1	1
4	56	0	3	200	42	100	145	75	0	1	0	67	160	1	0
5	54	1	2	300	45	132	175	100	1	1	1	87	165	2	0
6	48	1	2	215	32	154	145	67	0	0	0	67	170	2	0
7	57	0	3	216	43	175	140	85	1	1	0	57	164	1	0
8	52	0	1	254	45	100	143	70	0	0	0	69	160	2	1
9	67	1	2	310	47	140	175	105	1	1	1	76	177	2	0
10	46	0	2	237	37		430	70	0	0	1	56	160	3	0
11	58	1	3	220	36	120	120	70	0	0	1	56	160	3	3
12	62	0	2	233	45		130	75	0	0	1	76	165	1	0
13	49	1	1	240	38	125	120	90	0	1	1	83	173	3	0
14	56	0	2	295	44	98	180	95	1	1	0	65	170	2	0
15	63	1	2	310	39		165	95	1	0	0		173	3	0
16	64	0	2	268	41	132	150	90	1	0	0	75	158	1	1
17	67	0	3	243	43	176	140	85	0	1	0	65	165	1	1
18	49	0	2	239	54	137	125	75	0	0	1	57	161	2	0
19	53	1	2	198	32	87	135	75	0	1	1	76	168	2	0
20	59	1	2	218	45	134	120	85	0	0	0		175		1
21	65	1	3	215	39	110	120	70	0	0	1	68	168	1	0
22	67	0	2	254	38	149	180	105	1	0	1	78	165	2	1
23	49	0	1	218	46	176	135	85	0	0	1	65	159	2	0
24	53	0	2	221	46	189	135	80	0	0	0	57	150	3	1
25	57	1	2	237	48	223	150	90	0	1	0	87	185	2	0
26	47	1	1	244	45	230	130	85	0	0	1	76	178	3	0
27	58	0	2	223	43	234	130	70	0	1	1	59	153	1	1
28	48	0	2	198	37	198	125	85	0	0	1	62	155	2	0

29	51	1	2	234	43	112	125	80	0	0	1	62	155	2	0
30	49	0	3	175	38	234	140	80	0	1	0	65	153	1	0
31	68	1	2	230	43	110	110	70	0	0	0	78	159	2	1
32	58	0	2	248	47	109	135	75	0	0	0	78	190	2	1
33	54	0	2	218	36	108	160	95	0	1	0	76	170	2	0
34	59	1	1	285	38	104	170	100	1	1	1	93	172	3	1
35	45	0	2	253	53	120	125	75	0	1	0	64	160	2	0
36	53	0	2	187	28	98	145	80	0	0	0	65	167	2	0
37	43	1	2	208	39	156	120	65	0	1	1	87	179	1	1
38	57	0	2	246	44	127	130	80	0	0	0	65	167	2	0
39	64	1	2	275	26	180	160	95	1	1	1	69	175	2	1
40	43	0	2	218	56	143	120	75	0	1	0	54	165	3	0
41	47	1	3	231	43	140	150	75	0	1	1	67	160	1	1
42	58	1	1	200	31	154	140	90	0	1	0	75	170	1	1
43	58	1	2	214	56	156	130	75	0	0	0	76	170	2	0
44	48	0	2	230	38	110	120	70	0	0	0	56	150	2	1
45	62	1	2	280	36	103	160	100	1	1	1	75	167	1	1
46	54	0	1	198	32	103	115	65	0	1	0	54	160	3	1
47	67	0	2	285	31	100	165	95	1	1	1	70	150	2	1
48	68	1	1	201	39	106	130	80	0	1	0	70	180	2	1
49	55	0	2	206	46	101	120	65	0	1	0	50	189	2	0
50	50	1	2	223	45	139	125	75	0	0	0	68	172		0
51	53	1	1	290	34	120	160	95	1	1	1	88	165	3	1
52	63	1	2	315	40	130	170	100	1	1	1	90	170	2	1
53	60	0	2	220	50	145	130	80	0	0	0	65	150	3	0
54	46	0	2	230	32	158	115	75	0	0	0	58	168	2	0
55	45	1	2	175	32	123	140	70	0	0	0	65	170	2	0
56	53	1	2	213	36	128	130	70	0	0	0	69	175	1	0
57	59	0	2	220	57	130	120	65	0	0	0	56	164	3	0
58	62	1	2	287	38	120	170	95	1	1	1	88	165	2	1
59	60	1	2	290	40	130	170	90	1	1	1	89	162	3	1

60	62	0	2	209	48	120	135	75	0	0	0	60	170		0
61	58	1	2	590	36	130	130	80	1	1	1	96	175	2	1
62	57	1	1	260	39	142	165	95	1	1	1	90	170		1
63	49	0	1	202	56	123	140	80	0	0	0	60	170	3	0
64	61	0	2	214	45	150	125	90	0	0	0	60	175		0
65	52	0	2	231	45	128	115	75	0	0	0	54	164	2	0
66	59	1	2	280	34	100	185	100	1	1	1	85	164	2	1
67	50	1	2	220	60	134	150	70	0	0	0	69	165	2	0
68	46	1	2	233	54	109	115	78	0	0	0	70	175	1	0
69	44	0	1	215	50	130	125	70	0	0	1	50	160	2	0
70	60	0	2	202	48	120	120	65	0	0	1	52	165	2	0
X1	paciente #														
X2	Edad														
X3	Sexo						1: MASCULINO; 2: FEMENINO								
X4	Clase Social						1: ALTA; 2: MEDIA; 3 : BAJA								
X5	Colesterolemia Basal														
X6	Colesterolemia HDL Basal														
X7	Triglicéridemia Basal														
X8	Tensión arterial sistólica														
X9	Tensión arterial diastólica														
X10	Enfermedad coronaria											1: SI; 2: NO			
X11	Fuma											1: SI; 2: NO			
X12	Sedentarismo											1: SI; 2: NO			
X13	Peso														
X14	Talla														
X15	Nivel de estudios						1: PRIMARIO; 2: MEDI0; 3 : SUPERIOR								
X16	Antecedentes cardiacos Familiares											1: SI; 2: NO			

Anexo 4. Dimensiones corporales

Estatura	Estatura sentado	Longitud brazo	Longitud antebrazo	
165,80	88,70	31,80	28,10	
169,80	90,00	32,40	29,10	
170,70	87,70	33,60	29,50	
170,90	87,10	31,00	28,20	
157,50	81,30	32,10	27,30	
165,90	88,20	31,80	29,00	
158,70	86,10	30,60	27,80	
166,00	88,70	30,20	26,90	
158,70	83,70	31,10	27,10	
161,50	81,20	32,30	27,80	
167,30	88,60	34,80	27,30	
167,40	83,20	34,30	30,10	
159,20	81,50	31,00	27,30	
170,00	87,90	34,20	30,90	
166,30	88,30	30,60	28,80	
169,00	85,60	32,60	28,80	
156,20	81,60	31,00	25,60	
159,60	86,60	32,70	25,40	
155,00	82,00	30,30	26,60	
161,10	84,10	29,50	26,60	
170,30	88,10	34,00	29,30	
167,80	83,90	32,50	28,60	
163,10	88,10	31,70	26,90	
165,80	87,00	33,20	26,30	
175,40	89,60	35,20	30,10	
159,80	85,60	31,50	27,10	
166,00	84,90	30,50	28,10	

	Ancho mano	Longitud muslo	Longitud interior pierna	Longitud pie
	18,70	40,30	38,90	6,70
	18,30	43,30	42,70	6,40
	20,70	43,70	41,10	7,20
	18,60	43,70	40,60	6,70
	17,50	38,10	39,60	6,60
	18,60	42,00	40,60	6,50
	18,40	40,00	37,00	5,90
	17,50	41,60	39,00	5,90
	18,30	38,90	37,50	6,10
	19,10	42,80	40,10	6,20
	18,30	43,10	41,80	7,30
	19,20	43,40	42,20	6,80
	17,50	39,80	39,60	4,90
	19,40	43,10	43,70	6,30
	18,30	41,80	41,00	5,90
	19,10	42,70	42,00	6,00
	17,00	44,20	39,00	5,10
	17,70	42,00	37,50	5,00
	17,30	37,90	36,10	5,20
	17,80	38,60	38,20	5,90
	18,20	43,20	41,40	5,90
	20,20	43,30	42,90	7,20
	18,10	40,10	39,00	5,90
	19,50	43,20	40,70	5,90
	19,10	45,10	44,50	6,30
	19,20	42,30	39,00	5,70
	17,80	41,20	43,00	6,10

161,20	84,10	32,80	29,20	
160,40	84,30	30,50	27,80	
164,30	85,00	35,00	27,80	
165,50	82,60	36,20	28,60	
167,20	85,00	33,60	27,10	
167,20	83,40	33,50	29,70	

	18,40	42,60	41,10	5,90
	16,80	41,00	39,80	6,00
	19,00	47,20	42,40	5,00
	20,20	45,00	42,30	5,60
	19,80	46,00	41,60	5,60
	19,40	45,20	44,00	5,20

Notas al final


John W. Tukey

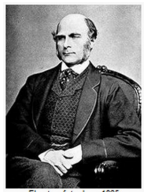
i El 25 de julio del 2000 murió John Wilder Tukey a los 85 años de edad. Fue uno de los grandes talentos estadísticos del siglo XX, con una notable influencia en la Visualización de Información. Su contribución mejor conocida es la de la transformada rápida de Fourier (FFT), pero también su libro Exploratory Data Analysis (1977) es el clásico sobre este tema.


George Box

ii George Box, químico, matemático, estadístico inglés, nacido en. 1919. Acuñó, en 1953, el término Robustez para designar procedimientos estadísticos que dan resultados aceptables cuando no se cumplen totalmente los supuestos en que se basan. Sin embargo, el tema de la Estadística Robusta toma importancia a partir de 1960, con P.Huber y F.R.Hampel.

iii El término fue introducido por el alemán William Stern y adoptado posteriormente por otros investigadores. El CI se calculaba dividiendo la edad mental de la persona por su edad cronológica, y multiplicando este valor por 100. 100 punto significa que el individuo posee una edad mental ajustada a su edad cronológica y una inferior o superior indica que el sujeto se sitúa por debajo o por encima a la media de la población general. Los test de inteligencias actuales han abandonado esta estrategia metodológica, y el cálculo del CI se realiza mediante una comparación estadística respecto a un grupo de muestra. Los CI siguen una distribución normal en campana, con la mayoría de las puntuaciones agrupadas en torno a 100. Aproximadamente dos de cada tres personas arroja una puntuación entre 85 y 115, mientras que el 19 de cada 20 personas tiene una puntuación entre 70 y 130. Una persona con una puntuación de 130 es considerada sobredotada, mientras que una puntuación inferior a 70 apunta a una deficiencia.

iv El índice de satisfacción familiar expresa la mayor o menor satisfacción que siente el alumno con su ambiente familiar. Varía entre 0 y 1. Cuanto más cercano a cero más insatisfacción y a la inversa al acercarse a 1.



v Francis Galton (Sparkbrook, Birmingham, 16 de febrero de 1822 - Haslemere, Surrey, 17 de enero de 1911) fue un polímata, antropólogo, geógrafo, explorador, inventor, meteorólogo, estadístico, psicólogo y eugenista británico con un amplio espectro de intereses. No tuvo cátedras universitarias y realizó la mayoría de sus investigaciones por su cuenta, las que fueron fundamentales para la constitución de la ciencia de la estadística:

- Inventó el uso de la línea de regresión, siendo el primero en explicar el fenómeno de la regresión a la media.
- En las décadas de 1870 y 1880 fue pionero en el uso de la distribución normal.
- Inventó la máquina Quincunx, un instrumento para demostrar la ley del error y la distribución normal.
- Descubrió las propiedades de la distribución normal bivariada y su relación con el análisis de regresión.
- En 1888 introdujo el concepto de correlación, posteriormente desarrollado por Pearson y Sperman.

vi La prueba de Kolmogórov-Smirnov (también prueba K-S) es prueba no paramétrica que determina la bondad de ajuste de dos distribuciones de probabilidad entre sí.

vii Andréi Nikoláyevich Kolmogórov (Андре́й Никола́евич Колмого́ров) (Tambov, 25 de abril de 1903 - Moscú, 20 de octubre de 1987). Matemático soviético que hizo progresos importantes en los campos de la teoría de la probabilidad y de la topología. Estructuró el sistema axiomático de la teoría de la probabilidad a partir de la teoría de conjuntos. Trabajó en lógica constructivista; en las series de Fourier; en turbulencias y mecánica clásica. Fundó la teoría de la complejidad algorítmica.

viii Test de Shapiro–Wilk se usa para contrastar la normalidad de un conjunto de datos. Se plantea como hipótesis nula que una muestra x1, ..., xn proviene de una población normalmente distribuida. Fue publicado en 1965 por Samuel Shapiro y Martin Wilk. Se considera uno de los test más potentes para el contraste de normalidad, sobre todo para muestras pequeñas (n<50).

ix Contingencia: En lógica y filosofía, la contingencia es el modo de ser de lo que no es necesario ni imposible, sino que puede ser o no ser el caso. En general la contingencia se predica de los estados de cosas, los hechos, los eventos o las proposiciones. De la relación entre necesidad, posibilidad y contingencia se tiene que:

1. Todo lo que es contingente es posible.
2. No todo lo que es posible es contingente, pues aquello que es necesario también es posible, pero no es contingente.
3. No todo lo que no es necesario es contingente, pues lo que es imposible no es ni necesario ni contingente.

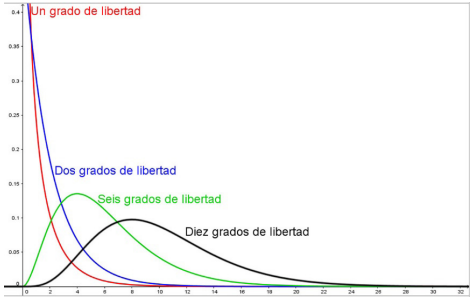
De modo que la relación entre variables que se muestra en una tabla de contingencia es posible (1) pero requiere de una demostración (prueba chi-cuadrado) porque no todo lo posible es contingente (2), ni todo lo necesario es contingente (3).

x Karl Pearson (Londres 27 de marzo de 1857-Londres, 27 de abril de 1936) fue un prominente científico, matemático y pensador británico, que estableció la disciplina de la *estadística matemática*. Desarrolló una intensa investigación sobre la aplicación de los métodos estadísticos en la biología y fue el fundador de la bioestadística.

xi Para su estudio la estadística se clasifica en estadística paramétrica y no paramétrica, la primera comprende los procedimientos estadísticos y de decisión que están basados en las distribuciones de los datos reales, que se determinan usando un número finito de parámetros (medias, desviaciones), bajo el supuesto de que tales datos se ajustan a distribuciones establecidas como la distribución normal (ya mencionada en este libro); ahora bien, cuando los datos no son del tipo escala o cuando se desconoce totalmente qué distribución siguen los datos entonces se deben aplicar los test no paramétricos.

xii En estadística inferencial, un contraste de hipótesis, un test de hipótesis o una prueba de significación es un procedimiento para juzgar si una propiedad que se supone en una población estadística es compatible con lo observado en una muestra de dicha población. Fue iniciada por Ronald Fisher y fundamentada posteriormente por Jerzy Neyman y Karl Pearson.

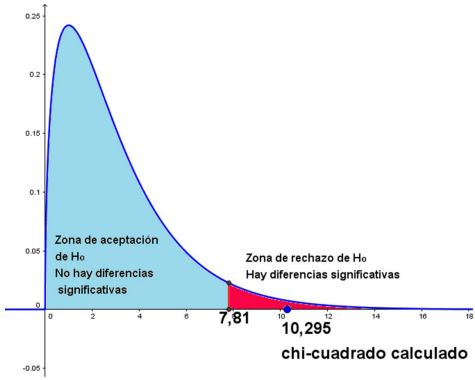
xiii Para la solución de este problema, a cada valor de χ^2 se hace corresponder un valor de probabilidad de modo tal que existe una función que define esta correspondencia (distribución χ^2) según los grado de libertad (r-1)(c-1). Este tratamiento mediante el empleo de distribuciones de probabilidad también contribuye a eliminar la posibilidad de que no haya influido el azar al obtener la frecuencia observada. Para una familiarización con la distribución χ^2 en la siguiente ilustración se presentan gráficas que correspondientes a los grados de libertad 1,2,6 y 10.



xiv Grados de libertad, es una expresión introducida por Ronald Fisher, con la que se indica que, de un conjunto de observaciones, los grados de libertad están dados por el número de valores que pueden ser asignados de forma arbitraria, antes de que el resto de las variables tomen un valor automáticamente, producto de establecerse las que son libres, esto, con el fin de compensar e igualar un resultado el cual se ha conocido previamente. Se encuentran mediante la fórmula , donde n= número de sujetos en la muestra que pueden tomar un valor y r es el número de sujetos cuyo valor dependerá del que tomen los miembros de la muestra que son libres. Generalmente se expresa por gl=(Númerodefilas-1)*(númerodecolumnas-1).

xv Este valor se puede encontrar en tablas de la distribución que aparecen en los libros de texto o se puede calcular mediante asistentes matemáticos u hojas de cálculo.

xvi La correspondencia entre el gráfico de la distribución chi-cuadrada, los datos que plantea el problema se ilustra en el siguiente gráfico y la posibilidad de aceptar o rechazar la hipótesis nula se ilustra en el siguiente gráfico.



xvii Lógica polivalente es un sistema lógico que rechaza el principio del tercero excluido de las lógicas bivalentes y admite más valores de verdad que los tradicionales *verdadero y falso.*

xviii Jan Łukasiewicz (21 de diciembre de 1878 - 13 de febrero de 1956) fue un matemático, lógico y filósofo polaco que nació en Leópolis, Galitzia (actual Ucrania). Su trabajo se centró en la lógica. Trabajó en lógica plurivalente, incluyendo su propio cálculo de tres valores de verdad, la primera lógica de cálculo no clásica. También se dedicó a otras áreas de la filosofía, aproximándose a los aspectos humanos de la creación de la teoría científica con ideas similares a las de Karl Popper.

Es autor, entre otras obras, de Elementos de lógica matemática, La silogística de Aristóteles desde el punto de vista de la lógica formal moderna, Sobre la teoría intuicionista de la deducción, Un sistema de lógica modal, y El principio de individuación.

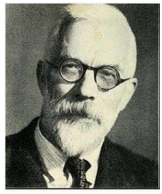
xix Alfred Tarski (1902—1983) lógico, matemático y filósofo polaco. Influyó en toda la investigación lógica posterior a la Segunda Guerra Mundial. Hizo

aportaciones destacadas en teoría de conjuntos y la lógica polivalente entre otras.

xx Charles Spearman. Psicólogo inglés, (1863-1945) Realizó importantes aportes a la psicología y a la estadística, desarrollando el Análisis Factorial. Su método, inscrito en las matemáticas experimentales, estudia las dimensiones del campo empírico. Sus aportes metodológicos no solo se han constituido en herramientas fundamentales para algunos ámbitos de la psicología, sino que son instrumentos para la ciencia estadística. El desarrollo de Spearman es útil en todas las ciencias sociales que requieran de técnicas de estadística correlacional para poder interpretar la información recogida.

xxi Mínimos cuadrados es una técnica de análisis numérico enmarcada dentro de la optimización matemática, en la que, dados un conjunto de pares ordenados —variable independiente, variable dependiente— y una familia de funciones, se intenta encontrar la función continua, dentro de dicha familia, que mejor se aproxime a los datos (un "mejor ajuste"), de acuerdo con el criterio de mínimo error cuadrático.

xxii Siméon Denis Poisson (Pithiviers, Francia, 21 de junio de 1781 - Sceaux (Altos del Sena), Francia, 25 de abril de 1840) fue un físico y matemático francés al que se le conoce por sus diferentes trabajos en el campo de la electricidad y por sus publicaciones acerca de la geometría diferencial y la teoría de probabilidades.



xxiii Sir Ronald Aylmer Fisher (Londres, Reino Unido, 17 de febrero de 1890 – Adelaida, Australia, 29 de julio de 1962) fue un estadístico y biólogo que usó la matemática para combinar las leyes de Mendel con la selección natural. En 1919 Fisher empezó a trabajar en Rothamsted Research, una estación agrícola experimental donde desarrolló el análisis de la varianza para analizar una gran cantidad de datos que generaron los cultivos plantados desde los años 1840.

xxv Matriz invertible: En matemáticas, en particular en álgebra lineal, una matriz cuadrada A de orden n se dice que es invertible, no singular, no degenerada o regular si existe otra matriz cuadrada de orden n, llamada matriz inversa de A y representada como A–1, tal que A x A–1= In, donde In es la matriz identidad de orden n y el producto utilizado es el producto de matrices usual.

xxvi La probabilidad de Bayes se sustenta en el siguiente teorema de la probabilidad total:

La probabilidad del suceso A que pueda ocurrir solo a condición de que aparezca uno de los sucesos mutuamente excluyentes B_1, B_2, B_3, …B_n que forman un grupo completo, es igual a la suma de los productos de las probabilidades de cada uno de estos sucesos por la correspondiente probabilidad condicional del suceso A.

$P(A) = P(B_1).PB_1(A) + P(B_2).PB_2(A) + \ldots. P(B_n).PB_n(A)$

Fórmula de Bayes

$$P_A(B_i) = \frac{P(B_i) \cdot P_{(B_i)}(A)}{P(B_1) \cdot P_{(B_1)}(A) + P(B_2) \cdot P_{(B_2)}(A) + \cdots + P(B_n) \cdot P_{(B_n)}(A)}$$

Thomas Bayes (Londres, Inglaterra, ~1702 - Tunbridge Wells, 1761) fue un matemático británico. Su padre fue ministro presbiteriano. Posiblemente De Movre, autor del afamado libro La doctrina de las probabilidades, fue su maestro particular, pues se sabe que por ese entonces ejercía como profesor en Londres. Bayes fue ordenado, al igual que su padre, como ministro disidente, y en 1731 se convirtió en reverendo de la iglesia presbiteriana en Tunbridge Wells; aparentemente trató de retirarse en 1749, pero continuó ejerciendo hasta 1752, y permaneció en ese lugar hasta su muerte.

Estudió el problema de la determinación de la probabilidad de las causas a través de los efectos observados. El teorema que lleva su nombre se refiere a la probabilidad de un suceso condicionado por la ocurrencia de otro suceso. Más específicamente, con su teorema se resuelve el problema conocido como "de la probabilidad inversa". Esto es, valorar probabilísticamente las posibles condiciones que rigen supuesto que se ha observado cierto suceso. Se trata de probabilidad "inversa" en el sentido de que la "directa" sería la probabilidad de observar algo supuesto que rigen ciertas condiciones.

Los cultores de la inferencia bayesiana (basada en dicho teorema) afirman que la trascendencia de la probabilidad inversa reside en que es ella la que realmente interesa a la ciencia, dado que procura sacar conclusiones generales (enunciar leyes) a partir de lo objetivamente observado, y no viceversa. Los restos de Bayes descansan en el cementerio londinense de Bunhill Fields. La traducción de la inscripción en su tumba es "Reverendo Thomas Bayes. Hijo de los conocidos Joshua y Ann Bayes. 7 de abril de 1761. En reconocimiento al importante trabajo que realizó Thomas Bayes en materia de probabilidades, su tumba fue restaurada en 1969 con donativos realizados por estadísticos de todo el mundo.

Miembro de la Royal Society desde 1742, Bayes fue uno de los primeros en utilizar la probabilidad inductivamente y establecer una base matemática para la inferencia probabilística. Actualmente, con base en su obra, se ha desarrollado una poderosa teoría que ha conseguido notables aplicaciones en las más diversas áreas del conocimiento. Especial connotación han tenido los sistemas para detección de spam en el ambiente de Internet. En el campo sanitario, el enfoque de la inferencia bayesiana experimenta un desarrollo sostenido, especialmente en lo que concierne al análisis de ensayos clínicos, donde dicho enfoque ha venido interesando de manera creciente a las agencias reguladoras de los medicamentos, tales como la norteamericana FDA (Food and Drug Agency).

xxvii Prasanta Chandra Mahalanobis (Bangla: প্রশান্তচন্দ্রমহলানবিস) (29 de junio) de 1893– 28 de junio de 1972) fue un científico indio que destacó en estadística aplicada. Su contribución más conocida es la distancia de Mahalanobis, una medida de distancia estadística. Realizó trabajos pioneros en las variaciones antropométricas en la india. Fundó el Instituto Indio de Estadística, y contribuyó al campo de las encuestas a gran escala.

xxviii La prueba de Wald es una prueba estadística paramétrica nombrada así en honor del estadístico Abraham Wald. Cada vez que hay una relación dentro o entre los datos se puede expresar un modelo estadístico con los parámetros a ser estimados a partir de una muestra, la prueba de Wald se utiliza para poner a prueba el verdadero valor del parámetro basado en la estimación de la muestra.


Abraham Wald

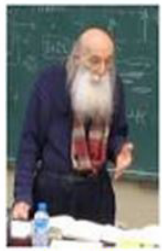
Abraham Wald (31 de octubre de 1902 Cluj-Napoca, Rumania - 13 de diciembre de 1950, Travancore, India) fue un matemático que hizo importantes contribuciones a la teoría de la decisión, la geometría, la economía y que fundó el análisis secuencial.

Hasta que ingresó en la universidad fue educado por sus padres ya que era judío y los sábados no podía ir a la escuela, como era obligatorio en el sistema escolar húngaro. En 1931 se graduó en la Universidad de Viena con el título de doctor en matemáticas.

Pudo emigrar a los Estados Unidos gracias a la invitación de la Comisión Cowles para la Investigación Económica cuando los nazis invadieron Austria en 1938 y fue perseguido junto a su familia debido a su condición de judío. Murió en un accidente aéreo en la India mientras realizaba un viaje para dar una conferencia invitado por el gobierno indio.

xxix Harold Hotelling (29 de septiembre de 1895 - 26 de diciembre de 1973) fue un matemático, estadístico e influyente economista. Fue Profesor Asociado de Matemáticas en la Universidad de Stanford desde 1927 hasta 1931, miembro de la facultad de la Universidad de Columbia desde 1931 hasta 1946, y Profesor de Estadística Matemática en la Universidad de Carolina del Norte en Chapel Hill desde 1946 hasta su muerte. Una calle en Chapel Hill lleva su nombre. Pionero en la combinación de Estadística Matemática y Economía. También trabajo con Ronald Fisher y aplicó algunas de sus técnicas. En particular al periodismo, ciencia política, demografía y alimentación. Es conocido en Estadística por sus trabajos en Análisis Multivariante, en particular por la distribución de probabilidad T-Cuadrada de Hotelling, una generalización de la T de Student y su uso en el contraste de hipóte-

sis estadístico y en las regiones de confianza. También introdujo el análisis de la correlación canónica.



Jean-Paul Benzécri

xxx Jean-Paul Benzécri (1932 -), estadístico francés. Estudió en la Escuela Normal Superior y trabajó como profesor del Instituto de Estadística de la Universidad de París VI. Se lo considera fundador de la escuela francesa de análisis estadístico de datos durante los años 1960-1990. Ha desarrollado técnicas estadísticas entre las que destaca la del análisis de correspondencias.

Según Monterde y Perea el Análisis Exploratorio de Datos (AED) es, "[…] por una parte, una perspectiva o actitud sobre el análisis de datos, en la que se exhorta a que el investigador adopte una actitud activa en y hacia el análisis de los mismos, como un medio para sugerir nuevas hipótesis de trabajo. Por otra parte, se compone de un renovado utillaje conceptual e instrumental respecto a lo que podríamos llamar Estadística Descriptiva "clásica", con el fin de optimizar la cantidad de información que los datos recogidos puedan ofrecer al investigador, […]"

Bajo esta concepción, los autores presentan el paquete estadístico SPSS, de modo que no se trate de un clásico manual de instrucciones y comandos, aunque por supuestos estos están presentes, pero combinados con los elementos teóricos que permitan comprender al lector los fundamentos de AED complementados con ejemplos tomados de cuatro bases de datos probadas en investigaciones y en textos de reconocido nivel científico; con ellos se lustra el proceder de un investigador que siga los métodos de AED, desde la opción "Explorar" del menú inicial "Analizar" y "estadísticos descriptivos" hasta los complejos métodos de la estadística multivariada que analizan simultáneamente varias medidas de cada individuo u objeto sometido a investigación.

EDITORIAL UNIVERSO SUR

ISBN: 978-959-257-493-9

UMET UNIVERSIDAD METROPOLITANA