IBM SPSS Statistics Base 24

Nota

Antes de utilizar esta información y el producto al que da soporte, lea la información que se incluye en el apartado "Avisos" en la página 211.

Información de producto

Esta edición se aplica a la versión 24, release 0, modificación 0 de IBM SPSS Statistics y a todas las versiones y modificaciones posteriores hasta que se indique lo contrario en nuevas ediciones.

Contenido

Capítulo 1. Libro de códigos

El libro de códigos hace referencia a la información del diccionario, como nombres de variable, etiquetas de variables, etiquetas de valores o valores perdidos, y los estadísticos de resumen de todas o las variables especificadas y conjuntos de respuestas múltiples del conjunto de datos activo. Para variables nominales y ordinales y conjuntos de respuestas múltiples, los estadísticos de resumen incluyen recuentos y porcentajes. Para variables de escala, los estadísticos de resumen incluyen la media, desviación estándar y cuartiles.

Nota: el libro de códigos ignora el estado del archivo segmentado. Esto incluye los grupos de archivos segmentados para imputaciones múltiples de valores perdidos (disponible en la opción adicional Valores perdidos).

Para obtener un libro de códigos

1. Seleccione en los menús:

   Analizar > **Informes** > **Libro de códigos**
2. Pulse en la pestaña Variables.
3. Seleccione una o más variables y/o conjuntos de respuestas múltiples.

Si lo desea, puede:

- Controlar la información de variable que aparece.
- Controlar los estadísticos que aparecen (o excluir todos los estadísticos de resumen).
- Controlar el orden en que aparecen las variables y los conjuntos de respuestas múltiples.
- Cambiar el nivel de medición de cualquier variable en la lista de origen para modificar los estadísticos de resumen que aparecen. Consulte el tema “Pestaña Estadísticos del libro de códigos” en la página 4 para obtener más información.

Cambio del nivel de medición

Puede cambiar temporalmente el nivel de medición de variables. (No puede modificar el nivel de medición de conjuntos de respuestas múltiples. Se tratarán siempre como nominales.)

1. En la lista de origen, pulse con el botón derecho del ratón en una variable.
2. Seleccione un nivel de medición del menú emergente.

Se modificará el nivel de medición temporalmente. En términos prácticos, esto sólo es útil para variables numéricas. El nivel de medición de las variables de cadena está restringido a nominal u ordinal, los cuales reciben el mismo tratamiento por parte del procedimiento del libro de códigos.

Pestaña Resultados de libro de códigos

La pestaña Resultados controla la información de la variable incluida para cada variable y los conjuntos de respuestas múltiples, el orden en que aparecerán las variables y los conjuntos de respuestas múltiples y el contenido de la tabla de información del archivo opcional.

Información de variable

Controla la información del diccionario que se muestra para cada variable.

Posición. Un entero que representa la posición de la variable en el orden de archivo. No está disponible para conjuntos de respuestas múltiples.

Etiqueta. La etiqueta descriptiva asociada con la variable o el conjunto de respuesta múltiple.

Tipo. Tipos de datos fundamentales. Puede ser *Numérico*, *Cadena* o *Conjunto de respuesta múltiple*.

Formato. El formato de visualización de la variable, como *A4*, *F8.2* o *DATE11*. No está disponible para conjuntos de respuestas múltiples.

Nivel de medición. Los valores posibles son *Nominal*, *Ordinal*, *Escala* y *Desconocido*. El valor que aparece es el nivel de medición guardado en el diccionario y no se ve afectado por ninguna sustitución de medición temporal especificada al modificar el nivel de medición en la lista de variables de origen de la pestaña Variables. No está disponible para conjuntos de respuestas múltiples.

Nota: el nivel de medición de las variables numéricas puede ser "desconocido" antes de la primera lectura de datos si el nivel de medición no se ha definido de forma explícita, como lecturas de datos de un origen externo o nuevas variables creadas. Consulte el tema para obtener más información.

Papel. Algunos cuadros de diálogo permiten preseleccionar variables para su análisis en función de papeles definidos.

Etiquetas de valor. Etiquetas descriptivas asociadas con valores de datos específicos.

- Si selecciona Recuento o Porcentaje en la pestaña Estadísticos, las etiquetas de valor definidas se incluyen en la distribución de los resultados incluso si no selecciona las etiquetas de valor aquí.
- Para los conjuntos de dicotomías múltiples, las "etiquetas de valor" son etiquetas de variable de las variables elementales en el conjunto o las etiquetas de valores contados, dependiendo de cómo se define el conjunto. Consulte el tema para obtener más información.

Valores perdidos. Valores perdidos del usuario. Si selecciona Recuento o Porcentaje en la pestaña Estadísticos, las etiquetas de valor definidas se incluyen en la distribución de los resultados incluso si no selecciona los valores perdidos aquí. No está disponible para conjuntos de respuestas múltiples.

Atributos personalizados. Atributos de variable personalizados definidos por el usuario. Los resultados incluyen los nombres y valores de cualquier atributo de variable personalizado asociado con cada variable. Consulte el tema para obtener más información. No está disponible para conjuntos de respuestas múltiples.

Atributos reservados. Atributos de variable de sistema reservados. Puede mostrar atributos del sistema, pero no debe modificarlas. Los nombres de atributos del sistema comienzan por un signo de dólar ($). No se incluyen los atributos que no se muestran, cuyo nombre comienza por "@" o "$@". Los resultados incluyen los nombres y valores de cualquier atributo de sistema asociado con cada variable. No está disponible para conjuntos de respuestas múltiples.

Información de archivo

La tabla de información del archivo opcional puede incluir cualquiera de los atributos de archivos siguientes:

Nombre de archivo. Nombre del archivo de datos de IBM® SPSS Statistics. Si el conjunto de datos no se ha guardado nunca en formato de IBM SPSS Statistics, no existe un nombre de archivo de datos. (Si no aparece un nombre de archivo en la barra de título de la ventana del Editor de datos, el conjunto de datos activo no tiene un nombre de archivo.)

Posición. Ubicación del directorio (carpeta) del archivo de datos de IBM SPSS Statistics. Si el conjunto de datos no se ha guardado nunca en formato de IBM SPSS Statistics, no existe una ubicación.

Número de casos. Número de casos en el conjunto de datos activo. Es el número total de casos, incluyendo los casos que se pueden excluir de los estadísticos de resumen por condiciones de filtro.

Etiqueta. Es la etiqueta del archivo (si tiene alguna) que define el comando `FILE LABEL`.

Documentos. Texto del documento del archivo de datos.

Estado de ponderación. Si se encuentra activada la ponderación, aparece el nombre de la variable de ponderación. Consulte el tema para obtener más información.

Atributos personalizados. Atributos de archivos de datos personalizados definidos por el usuario. Atributos de archivos de datos definidos con el comando `DATAFILE ATTRIBUTE`.

Atributos reservados. Atributos de archivo de datos de sistema reservados. Puede mostrar atributos del sistema, pero no debe modificarlas. Los nombres de atributos del sistema comienzan por un signo de dólar ($). No se incluyen los atributos que no se muestran, cuyo nombre comienza por "@" o "$@". Los resultados incluyen los nombres y valores de los atributos del archivo de datos del sistema.

Orden de visualización de variables

Las siguientes alternativas están disponibles para controlar el orden en que aparecen las variables y los conjuntos de respuestas múltiples.

Alfabético. Orden alfabético por nombre de variable.

Archivo. El orden en que aparecen las variables en el conjunto de datos (el orden en que aparecen en el editor de datos). En orden ascendente, los conjuntos de respuestas múltiples aparecen en último lugar, después de todas las variables seleccionadas.

Nivel de medición. Ordenar por nivel de medición. nominal, ordinal, escala y desconocido. Los conjuntos de respuestas múltiples se consideran nominales.

Nota: el nivel de medición de las variables numéricas puede ser "desconocido" antes de la primera lectura de datos si el nivel de medición no se ha definido de forma explícita, como lecturas de datos de un origen externo o nuevas variables creadas.

Lista de variables. El orden en que aparecen las variables y conjuntos de respuestas múltiples en la lista de variables seleccionadas en la pestaña Variables.

Nombre de atributo personalizado. La lista de opciones de orden de clasificación también incluye los nombres de cualquier atributo de variables personalizadas definidas por el usuario. En orden ascendente, las variables que no tienen la opción de clasificación de atributos al principio, seguidas de las variables que tienen el atributo pero no los valores definidos del atributo, seguidas de las variables con valores definidos para el atributo en orden alfabético de los valores.

Número máximo de categorías

Si el resultado incluye etiquetas de valor, los recuentos o porcentajes de cada valor exclusivo, puede eliminar esta información de la tabla si el número de los valores excede el valor especificado. De forma predeterminada, esta información se elimina si el número de valores exclusivos de la variable es superior a 200.

Pestaña Estadísticos del libro de códigos

La pestaña Estadísticos permite controlar los estadísticos de resumen que se incluyen en los resultados o suprimir la visualización de los estadísticos de resumen completamente.

Recuentos y porcentajes

Para las variables nominales y ordinales, conjuntos de respuestas múltiples y valores de etiquetas de variables de escala, los estadísticos disponibles son:

Recuento. El recuento o número de casos que tienen cada valor (o el rango de valores) de una variable.

Porcentaje. Porcentaje de casos que presenta un valor determinado.

Tendencia y dispersión centrales

Para las variables de escala, los estadísticos disponibles son:

Media. Una medida de tendencia central. El promedio aritmético, la suma dividida por el número de casos.

Desviación estándar. Una medida de dispersión sobre la media. En una distribución normal, el 68% de los casos se encuentra dentro de una desviación estándar de la media y el 95% queda entre dos desviaciones estándar. Por ejemplo, si la edad media es de 45 años, con una desviación estándar de 10, el 95% de los casos estaría entre los 25 y 65 en una distribución normal.

Cuartiles. Muestra los valores correspondientes a los percentiles 25, 50 y 75.

Nota: puede modificar de forma temporal el nivel de medición asociado con una variable (y por lo tanto, modificar los estadísticos de resumen de la variable) en la lista de variables de origen de la pestaña Variables.

Capítulo 2. Frecuencias

El procedimiento Frecuencias proporciona estadísticos y representaciones gráficas que resultan útiles para describir muchos tipos de variables. El procedimiento Frecuencias es un comienzo para empezar a consultar los datos.

Para los informes de frecuencias y los gráficos de barras, puede organizar los valores distintos en orden ascendente o descendente u ordenar las categorías por sus frecuencias. Es posible suprimir el informe de frecuencias cuando una variable posee muchos valores distintos. Puede etiquetar los gráficos con las frecuencias (la opción predeterminada) o con los porcentajes.

Ejemplo. ¿Cuál es la distribución de los clientes de una empresa por tipo de industria? En los resultados podría observar que el 37,5% de sus clientes pertenece a agencias gubernamentales, el 24,9% a corporaciones, el 28,1% a instituciones académicas, y el 9,4% a la industria sanitaria. Con respecto a los datos continuos, cuantitativos, como los ingresos por ventas, podría comprobar que el promedio de ventas de productos es de 3.576 dólares con una desviación estándar de 1.078 dólares.

Estadísticos y gráficos. Frecuencias, porcentajes, porcentajes acumulados, media, mediana, moda, suma, desviación estándar, varianza, amplitud, valores mínimo y máximo, error estándar de la media, asimetría y curtosis (ambos con sus errores estándar), cuartiles, percentiles especificados por el usuario, gráficos de barras, gráficos circulares e histogramas.

Frecuencias: Consideraciones sobre los datos

Datos. Utilice códigos numéricos o cadenas para codificar las variables categóricas (mediciones de nivel nominal u ordinal).

Supuestos. Las tabulaciones y los porcentajes proporcionan una descripción útil para los datos de cualquier distribución, especialmente para las variables con categorías ordenadas o desordenadas. Muchos de los estadísticos de resumen optativos, tales como la media y la desviación estándar, se basan en la teoría normal y son apropiados para las variables cuantitativas con distribuciones simétricas. Los estadísticos robustos, tales como la mediana, los cuartiles y los percentiles son apropiados para las variables cuantitativas que pueden o no cumplir el supuesto de normalidad.

Para obtener tablas de frecuencias

1. Seleccione en los menús:

   Analizar > **Estadísticos descriptivos** > **Frecuencias...**
2. Seleccione una o más variables categóricas o cuantitativas.

Si lo desea, puede:

- Pulsar en **Estadísticos** para obtener estadísticos descriptivos para las variables cuantitativas.
- Pulsar en **Gráficos** para obtener gráficos de barras, gráficos circulares e histogramas.
- Pulsar en **Formato** para determinar el orden en el que se muestran los resultados.

Frecuencias: Estadísticos

Valores percentiles. Los valores de una variable cuantitativa que dividen los datos ordenados en grupos, de forma que un porcentaje de los casos se encuentre por encima y otro porcentaje se encuentre por debajo. Los cuartiles (los percentiles 25, 50 y 75) dividen las observaciones en cuatro grupos de igual tamaño. Si desea un número igual de grupos que no sea cuatro, seleccione **Puntos de corte para n grup**

os iguales. También puede especificar percentiles individuales (por ejemplo, el percentil 95, el valor por debajo del cual se encuentran el 95% de las observaciones).

Tendencia central. Los estadísticos que describen la localización de la distribución, incluyen: Media, Mediana, Moda y Suma de todos los valores.

- *Media*. Una medida de tendencia central. El promedio aritmético, la suma dividida por el número de casos.
- *Mediana*. Es el valor por encima y por debajo del cual se encuentran la mitad de los casos, el percentil 50. Si hay un número par de casos, la mediana es la media de los dos valores centrales, cuando los casos se ordenan en orden ascendente o descendente. La mediana es una medida de tendencia central que no es sensible a los valores atípicos (a diferencia de la media, que puede resultar afectada por unos pocos valores extremadamente altos o bajos).
- *Moda*. El valor que ocurre con mayor frecuencia. Si varios valores comparten la mayor frecuencia de aparición, cada uno de ellos es un modo. El procedimiento de frecuencias devuelve sólo el modo más pequeño de los modos múltiples.
- *Suma*. Suma o total de todos los valores, a lo largo de todos los casos con valores no perdidos.

Dispersión. Los estadísticos que miden la cantidad de variación o de dispersión en los datos, incluyen: Desviación estándar, Varianza, Rango, Mínimo, Máximo y Error estándar de la media.

- *Desv. estándar*. Una medida de dispersión sobre la media. En una distribución normal, el 68% de los casos se encuentra dentro de una desviación estándar de la media y el 95% queda entre dos desviaciones estándar. Por ejemplo, si la edad media es de 45 años, con una desviación estándar de 10, el 95% de los casos estaría entre los 25 y 65 en una distribución normal.
- *Varianza*. Medida de dispersión sobre la media, igual a la suma de las desviaciones al cuadrado de la media divida por el número de casos menos uno. La varianza se mide en unidades que son el cuadrado de las de la variable en cuestión.
- *Rango*. Diferencia entre los valores mayor y menor de una variable numérica; el máximo menos el mínimo.
- *Mínimo*. Se trata del valor menor de una variable numérica.
- *Máximo*. Se trata del valor mayor de una variable numérica.
- *E. T. media*. Medida de cuánto puede variar el valor de la media de una muestra a otra, extraídas éstas de la misma distribución. Puede utilizarse para comparar de forma aproximada la media observada respecto a un valor hipotetizado (es decir, se puede concluir que los dos valores son distintos si la diferencia entre ellos, dividida por el error estándar, es menor que -2 o mayor que +2).

Distribución. Asimetría y curtosis son estadísticos que describen la forma y la simetría de la distribución. Estos estadísticos se muestran con sus errores estándar.

- *Asimetría*. Medida de la asimetría de una distribución. La distribución normal es simétrica y tiene un valor de asimetría igual a 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. 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.

Los valores son puntos medios de grupos. Si los valores de los datos son puntos medios de grupos (por ejemplo, si las edades de todas las personas entre treinta y cuarenta años se codifican como 35), seleccione esta opción para estimar la mediana y los percentiles para los datos originales no agrupados.

Frecuencias: Gráficos

Tipo de gráfico. Los gráficos circulares muestran la contribución de las partes a un todo. Cada porción de un gráfico circular corresponde a un grupo, definido por una única variable de agrupación. Los gráficos de barras muestran el recuento de cada valor o categoría distinta como una barra diferente, permitiendo comparar las categorías de forma visual. Los histogramas también cuentan con barras, pero se representan a lo largo de una escala de intervalos iguales. La altura de cada barra es el recuento de los valores que están dentro del intervalo para una variable cuantitativa. Los histogramas muestran la forma, el centro y la dispersión de la distribución. Una curva normal superpuesta en un histograma ayuda a juzgar si los datos están normalmente distribuidos.

Valores del gráfico. Para los gráficos de barras, puede etiquetar el eje de escala con las frecuencias o los porcentajes.

Frecuencias: Formato

Ordenar por. La tabla de frecuencias se puede organizar respecto a los valores actuales de los datos o respecto al recuento (frecuencia de aparición) de esos valores y la tabla puede organizarse en orden ascendente o descendente. Sin embargo, si solicita un histograma o percentiles, Frecuencias asumirá que la variable es cuantitativa y mostrará sus valores en orden ascendente.

Múltiples variables. Si desea generar tablas de estadísticos para múltiples variables, podrá mostrar todas las variables en una sola tabla (**Comparar variables**), o bien mostrar una tabla de estadísticos independiente para cada variable (**Organizar resultados según variables)**.

Suprimir tablas con varias categorías. Esta opción impide que se muestren tablas que contengan más valores que el número especificado.

Capítulo 3. Descriptivos

El procedimiento Descriptivos muestra estadísticos de resumen univariados para varias variables en una única tabla y calcula valores tipificados (puntuaciones *z*). Las variables se pueden ordenar por el tamaño de sus medias (en orden ascendente o descendente), alfabéticamente o por el orden en el que se seleccionen las variables (el valor predeterminado).

Cuando se guardan las puntuaciones *z*, éstas se añaden a los datos del Editor de datos y quedan disponibles para los gráficos, el listado de los datos y los análisis. Cuando las variables se registran en unidades diferentes (por ejemplo, producto interior bruto per cápita y porcentaje de alfabetización), una transformación de puntuación *z* pondrá las variables en una escala común para poder compararlas visualmente con más facilidad.

Ejemplo. Si cada caso de los datos contiene los totales de ventas diarias de cada vendedor (por ejemplo, una entrada para Bob, una para Kim y una para Brian) recogidas cada día durante varios meses, el procedimiento Descriptivos puede calcular la media diaria de ventas para cada vendedor y ordenar los resultados del promedio de ventas de mayor a menor.

Estadísticos. Tamaño de la muestra, media, mínimo, máximo, desviación estándar, varianza, rango, suma, error estándar de la media, curtosis y asimetría con sus errores estándar.

Descriptivos: Consideraciones sobres los datos

Datos. Utilice variables numéricas después de haberlas inspeccionado gráficamente para registrar errores, valores atípicos y anomalías de distribución. El procedimiento Descriptivos es muy eficaz para archivos grandes (de miles de casos).

Supuestos. La mayoría de los estadísticos disponibles (incluyendo las puntuaciones *z*) se basan en la teoría normal y son adecuados para variables cuantitativas (mediciones a nivel de razón o de intervalo) con distribuciones simétricas. Se deben evitar las variables con categorías no ordenadas o distribuciones asimétricas. La distribución de puntuaciones *z* tiene la misma forma que la de los datos originales; por tanto, el cálculo de puntuaciones *z* no es una solución para los datos con problemas.

Para obtener estadísticos descriptivos

1. Seleccione en los menús:

   Analizar > **Estadísticos descriptivos** > **Descriptivos...**
2. Seleccione una o más variables.

Si lo desea, puede:

- Seleccionar **Guardar valores tipificados como variables** para guardar las puntuaciones *z* como nuevas variables.
- Pulsar en **Opciones** para seleccionar estadísticos opcionales y el orden de presentación.

Descriptivos: Opciones

Media y suma. Se muestra de forma predeterminada la media o promedio aritmético.

Dispersión. Los estadísticos que miden la dispersión o variación en los datos incluyen la desviación estándar, la varianza, el rango, el mínimo, el máximo y el error estándar de la media.

- *Desv. estándar*. Una medida de dispersión sobre la media. En una distribución normal, el 68% de los casos se encuentra dentro de una desviación estándar de la media y el 95% queda entre dos

desviaciones estándar. Por ejemplo, si la edad media es de 45 años, con una desviación estándar de 10, el 95% de los casos estaría entre los 25 y 65 en una distribución normal.

- *Varianza*. Medida de dispersión sobre la media, igual a la suma de las desviaciones al cuadrado de la media divida por el número de casos menos uno. La varianza se mide en unidades que son el cuadrado de las de la variable en cuestión.
- *Rango*. Diferencia entre los valores mayor y menor de una variable numérica; el máximo menos el mínimo.
- *Mínimo*. Se trata del valor menor de una variable numérica.
- *Máximo*. Se trata del valor mayor de una variable numérica.
- *E.T media*. Medida de cuánto puede variar el valor de la media de una muestra a otra, extraídas éstas de la misma distribución. Puede utilizarse para comparar de forma aproximada la media observada respecto a un valor hipotetizado (es decir, se puede concluir que los dos valores son distintos si la diferencia entre ellos, dividida por el error estándar, es menor que -2 o mayor que +2).

Distribución. La curtosis y la asimetría son los estadísticos que caracterizan la forma y simetría de la distribución. Estos estadísticos se muestran con sus errores estándar.

- *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.
- *Asimetría*. Medida de la asimetría de una distribución. La distribución normal es simétrica y tiene un valor de asimetría igual a 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. 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.

Orden de presentación. De forma predeterminada, las variables se muestran en el orden en que se hayan seleccionado. Si lo desea, se pueden mostrar las variables alfabéticamente, por medias ascendentes o por medias descendentes.

Características adicionales del comando DESCRIPTIVES

La sintaxis de comandos también le permite:

- Guardar puntuaciones tipificadas (puntuaciones *z*) para algunas, pero no todas las variables (con el subcomando `VARIABLES`).
- Especificar nombres para las variables nuevas que contienen puntuaciones tipificadas (con el subcomando `VARIABLES`).
- Excluir del análisis los casos con valores perdidos para cualquier variable (con el subcomando `MISSING`).
- Clasificar las variables de la visualización según el valor de cualquier estadística, no sólo la media (con el subcomando `SORT`).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 4. Explorar

El procedimiento Explorar genera estadísticos de resumen y representaciones gráficas, bien para todos los casos o bien de forma separada para grupos de casos. Existen numerosas razones para utilizar este procedimiento: para inspeccionar los datos, identificar valores atípicos, obtener descripciones, comprobar supuestos y caracterizar diferencias entre subpoblaciones (grupos de casos). La inspección de los datos puede mostrar que existen valores inusuales, valores extremos, discontinuidades en los datos u otras peculiaridades. La exploración de los datos puede ayudar a determinar si son adecuadas las técnicas estadísticas que está teniendo en consideración para el análisis de los datos. La exploración puede indicar que necesita transformar los datos si la técnica necesita una distribución normal. O bien, el usuario puede decidir que necesita utilizar pruebas no paramétricas.

Ejemplo. Observe la distribución de los tiempos de aprendizaje de laberintos de una serie de ratas sometidas a cuatro programas de refuerzo diferentes. Para cada uno de los cuatro grupos, se puede observar si la distribución de tiempos es aproximadamente normal y si las cuatro varianzas son iguales. También se pueden identificar los casos con los cinco valores de tiempo mayores y los cinco menores. Los diagramas de caja y los gráficos de tallo y hojas resumen gráficamente la distribución del tiempo de aprendizaje de cada uno de los grupos.

Estadísticos y gráficos. Media, mediana, media recortada al 5%, error estándar, varianza, desviación estándar, mínimo, máximo, rango, rango intercuartil, asimetría y curtosis y sus errores estándar, intervalo de confianza para la media (y el nivel de confianza especificado), percentiles, estimador-M de Huber, estimador en onda de Andrews, estimador-M redescendente de Hampel, estimador biponderado de Tukey, cinco valores mayores y cinco menores, estadístico de Kolmogorov-Smirnov con el nivel de significación de Lilliefors para contrastar la normalidad y estadístico de Shapiro-Wilk. Diagramas de caja, gráficos de tallo y hojas, histogramas, diagramas de normalidad y diagramas de dispersión por nivel con pruebas de Levene y transformaciones.

Explorar: Consideraciones sobre los datos

Datos. El procedimiento Explorar se puede utilizar para las variables cuantitativas (nivel de medición de razón o de intervalo). Una variable de factor (utilizada para dividir los datos en grupos de casos) debe tener un número razonable de valores distintivos (categorías). Estos valores pueden ser de cadena corta o numéricos. La variable de etiquetas de caso, utilizada para etiquetar valores atípicos en los diagramas de caja, puede ser de cadena corta, de cadena larga (los 15 primeros bytes) o numérica.

Supuestos. La distribución de los datos no tiene que ser simétrica ni normal.

Para explorar los datos

1. Seleccione en los menús:

   Analizar > **Estadísticos descriptivos** > **Explorar...**
2. Seleccione una o más variables dependientes.

Si lo desea, puede:

- Seleccionar una o más variables de factor, cuyos valores definirán grupos de casos.
- Seleccionar una variable de identificación para etiquetar los casos.
- Pulse en **Estadísticos** para obtener estimadores robustos, valores atípicos, percentiles y tablas de frecuencias.
- Pulse en **Gráficos** para obtener histogramas, pruebas y gráficos de probabilidad normal y diagramas de dispersión por nivel con estadísticos de Levene.
- Pulse en **Opciones** para manipular los valores perdidos.

Explorar: Estadísticos

Descriptivos. De forma predeterminada se muestran estas medidas de dispersión y de tendencia central. Éstas últimas indican la localización de la distribución, e incluyen la media, la mediana y la media recortada al 5%. Las medidas de dispersión muestran la disimilaridad de los valores, incluyen: los errores estándar, la varianza, la desviación estándar, el mínimo, el máximo, el rango y el rango intercuartil. Los estadísticos descriptivos también incluyen medidas de la forma de la distribución: la asimetría y la curtosis se muestran con sus errores estándar. También se muestra el intervalo de confianza a un nivel del 95%; aunque se puede especificar otro nivel.

Estimadores robustos centrales. Alternativas robustas a la mediana y a la media muestral para estimar la localización. Los estimadores calculados se diferencian por las ponderaciones que aplican a los casos. Se muestran los siguientes: el estimador-M de Huber, el estimador en onda de Andrew, el estimador-M redescendente de Hampel y el estimador biponderado de Tukey.

Valores atípicos. Muestra los cinco valores mayores y los cinco menores con las etiquetas de caso.

Percentiles. Muestra los valores de los percentiles 5, 10, 25, 50, 75, 90 y 95.

Explorar: Gráficos

Diagramas de caja. Estas alternativas controlan la presentación de los diagramas de caja cuando existe más de una variable dependiente. **Niveles de los factores juntos** genera una presentación para cada variable dependiente. En cada una se muestran diagramas de caja para cada uno de los grupos definidos por una variable de factor. **Dependientes juntas** genera una presentación para cada grupo definido por una variable de factor. En cada una se muestran juntos los diagramas de caja de cada variable dependiente. Esta disposición es de gran utilidad cuando las variables representan una misma característica medida en momentos distintos.

Descriptivos. La sección Descriptivos permite seleccionar gráficos de tallo y hojas e histogramas.

Gráficos con pruebas de normalidad. Muestra los diagramas de probabilidad normal y de probabilidad sin tendencia. Se muestra el estadístico de Kolmogorov-Smirnov con un nivel de significación de Lilliefors para contrastar la normalidad. Si se especifican ponderaciones no enteras, se calculará el estadístico de Shapiro-Wilk cuando el tamaño de la muestra ponderada esté entre 3 y 50. Si no hay ponderaciones o éstas son enteras, se calculará el estadístico cuando el tamaño de la muestra esté entre 3 y 5.000.

Dispersión por nivel con prueba de Levene. Controla la transformación de los datos para los diagramas de dispersión por nivel. Para todos los diagramas de dispersión por nivel se muestra la inclinación de la línea de regresión y las pruebas robustas de Levene sobre la homogeneidad de varianza. Si selecciona una transformación, las pruebas de Levene se basarán en los datos transformados. Si no selecciona ninguna variable de factor, no se generará ningún diagrama de dispersión por nivel. **Estimación de potencia** produce un gráfico de los logaritmos naturales de los rangos intercuartiles respecto a los logaritmos naturales de las medianas de todas las casillas, así como una estimación de la transformación de potencia necesaria para conseguir varianzas iguales en las casillas. Un diagrama de dispersión por nivel ayuda a determinar la potencia que precisa una transformación para estabilizar (igualar) las varianzas de los grupos. **Transformados** permite seleccionar una de las alternativas de potencia, quizás siguiendo las recomendaciones de la estimación de potencia, y genera gráficos de los datos transformados. Se trazan el rango intercuartil y la mediana de los datos transformados. **No transformados** genera gráficos de los datos en bruto. Es equivalente a una transformación con una potencia de 1.

Explorar: Transformaciones de potencia

A continuación aparecen las transformaciones de potencia para los diagramas de dispersión por nivel. Para transformar los datos, deberá seleccionar una potencia para la transformación. Puede elegir una de las siguientes alternativas:

- **Log natural.** Transformación de logaritmo natural. Este es el método predeterminado.
- **1/raíz cuadrada.** Para cada valor de los datos se calcula el inverso de la raíz cuadrada.
- **Recíproco.** Se calcula el inverso de cada valor de los datos.
- **Raíz cuadrada.** Se calcula la raíz cuadrada de cada valor de los datos.
- **Cuadrado.** Se calcula el cuadrado de cada valor de los datos.
- **Cubo.** Se calcula el cubo de cada valor de los datos.

Explorar: Opciones

Valores perdidos. Controla el tratamiento de los valores perdidos.

- **Excluir casos según lista.** Los casos con valores perdidos para cualquier variable de factor o variable dependiente se excluyen de todos los análisis. Este es el método predeterminado.
- **Excluir casos según pareja.** Los casos con valores no perdidos para las variables de un grupo (casilla) se incluyen en el análisis de ese grupo. El caso puede tener valores perdidos para las variables utilizadas en otros grupos.
- **Mostrar los valores.** Los valores perdidos para las variables de factor se tratan como una categoría diferente. Todos los resultados se generan para esta categoría adicional. Las tablas de frecuencias incluyen categorías para los valores perdidos. Los valores perdidos para una variable de factor se incluyen pero se etiquetan como perdidos.

Características adicionales del comando EXAMINE

El procedimiento Explorar utiliza la sintaxis de comandos `EXAMINE`. La sintaxis de comandos también le permite:

- Solicitar los gráficos y resultados totales además de los gráficos y los resultados para los grupos definidos por las variables de factor (con el subcomando `TOTAL`).
- Especificar una escala común para un grupo de diagramas de caja (con el subcomando `SCALE`).
- Especificar interacciones de variables de factor (con el subcomando `VARIABLES`).
- Especificar percentiles distintos de los percentiles predeterminados (con el subcomando `PERCENTILES`).
- Calcular percentiles respecto a cualquiera de los cinco métodos (con el subcomando `PERCENTILES`).
- Especificar una transformación de potencia para diagramas de dispersión por nivel (con el subcomando `PLOT`).
- Especificar el número de valores extremos que se van a mostrar (mediante el subcomando `STATISTICS`).
- Especificar parámetros para los estimadores robustos centrales, los estimadores robustos de ubicación (mediante el subcomando `MESTIMATORS`).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 5. Tablas cruzadas

El procedimiento Tablas cruzadas crea tablas bidimensionales y multidimensionales y, además, proporciona una serie de pruebas y medidas de asociación para las tablas bidimensionales. La estructura de la tabla y el hecho de que las categorías estén ordenadas o no determinan las pruebas o medidas que se utilizaban.

Los estadísticos de tablas cruzadas y las medidas de asociación sólo se calculan para las tablas bidimensionales. Si especifica una fila, una columna y un factor de capa (variable de control), el procedimiento Tablas cruzadas crea un panel de medidas y estadísticos asociados para cada valor del factor de capa (o una combinación de valores para dos o más variables de control). Por ejemplo, si *sexo* es un factor de capa para una tabla de *casado* (sí, no) en función de *vida* (vida emocionante, rutinaria o aburrida), los resultados para una tabla bidimensional para las mujeres se calculan de forma independiente de los resultados de los hombres y se imprimen en paneles uno detrás del otro.

Ejemplo. ¿Es más probable que los clientes de las empresas pequeñas sean más rentables en la venta de servicios (por ejemplo, formación y asesoramiento) que los clientes de las empresas grandes? A partir de una tabulación cruzada podría deducir que la prestación de servicios a la mayoría de las empresas pequeñas (con menos de 500 empleados) produce considerables beneficios, mientras que con la mayoría de las empresas de gran tamaño (con más de 2.500 empleados), los beneficios obtenidos son mucho menores.

Estadísticos y medidas de asociación. Chi-cuadrado de Pearson, chi-cuadrado de la razón de verosimilitud, prueba de asociación lineal por lineal, prueba exacta de Fisher, chi-cuadrado corregido de Yates, *r* de Pearson, rho de Spearman, coeficiente de contingencia, phi, *V* de Cramér, lambdas simétricas y asimétricas, tau de Kruskal y Goodman, coeficiente de incertidumbre, gamma, *d* de Somers, tau-*b* de Kendall, tau-*c* de Kendall, coeficiente eta, kappa de Cohen, estimación de riesgo relativo, razón de las ventajas, prueba de McNemar y estadísticos de Cochran y Mantel-Haenszel.

Tablas cruzadas: Consideraciones sobre los datos

Datos. Para definir las categorías de cada variable, utilice valores de una variable numérica o de cadena (ocho bytes o menos). Por ejemplo, para *sexo*, podría codificar los datos como 1 y 2 o como *varón* y *mujer*.

Supuestos. En algunos estadísticos y medidas se asume que hay unas categorías ordenadas (datos ordinales) o unos valores cuantitativos (datos de intervalos o de proporciones), como se explica en la sección sobre los estadísticos. Otros estadísticos son válidos cuando las variables de la tabla tienen categorías no ordenadas (datos nominales). Para los estadísticos basados en chi-cuadrado (phi, *V* de Cramér y coeficiente de contingencia), los datos deben ser una muestra aleatoria de una distribución multinomial.

Nota: las variables ordinales pueden ser códigos numéricos que representen categorías (por ejemplo, 1 = *bajo*, 2 = *medio*, 3 = *alto*) o valores de cadena. Sin embargo, se supone que el orden alfabético de los valores de cadena indica el orden correcto de las categorías. Por ejemplo, en una variable de cadena cuyos valores sean *bajo*, *medio*, *alto*, se interpreta el orden de las categorías como *alto*, *bajo*, *medio* (orden que no es el correcto). Por norma general, se puede indicar que es más fiable utilizar códigos numéricos para representar datos ordinales.

Para obtener tabulaciones cruzadas

1. Seleccione en los menús:

   Analizar > **Estadísticos descriptivos** > **Tablas cruzadas...**
2. Seleccione una o más variables de fila y una o más variables de columna.

Si lo desea, puede:

- Seleccionar una o más variables de control.
- Pulsar en **Estadísticos** para obtener pruebas y medidas de asociación para tablas o subtablas bidimensionales.
- Pulsar en **Casillas** para obtener porcentajes, residuos y valores esperados y observados.
- Pulsar en **Formato** para controlar el orden de las categorías.

Capas de las tablas cruzadas

Si se seleccionan una o más variables de capas, se generará una tabulación cruzada por cada categoría de cada variable de capas (variable de control). Por ejemplo, si emplea una variable de fila, una variable de columna y una variable de capas con dos categorías, obtendrá una tabla bidimensional por cada categoría de la variable de capas. Para crear otra capa de variables de control, pulse en **Siguiente**. Se crean subtablas para cada combinación de categorías para cada variable de la 1ª capa, cada variable de la 2ª capa, y así sucesivamente. Si se solicitan estadísticos y medidas de asociación, se aplicarán sólo a las tablas bidimensionales.

Gráficos de barras agrupadas

Mostrar los gráficos de barras agrupadas. Los gráficos de barras agrupadas ayudan a resumir los datos por grupos de casos. Hay una agrupación de barras por cada valor de la variable especificada en el cuadro Filas. La variable que define las barras dentro de cada agrupación es la variable especificada en el cuadro Columnas. Por cada valor de esta variable hay un conjunto de barras de distinto color o trama. Si especifica más de una variable en Columnas o en Filas, se generará un gráfico de barras agrupadas por cada combinación de dos variables.

Tablas cruzadas mostrando variables de capa en capas de tabla

Mostrar variables de capa en capas de tabla. Puede seleccionar visualizar las variables de capa (variables de control) como capas en la tabla de tabulación cruzada. De esta forma podrá crear vistas que muestren los estadísticos globales de las variables de fila y columna y que permitan la obtención de detalles de las categorías de las variables de capa.

A continuación se muestra un ejemplo que utiliza el archivo de datos *demo.sav* (disponible en el directorio Samples del directorio de instalación) y que se ha obtenido de la siguiente forma:

1. Seleccione *Categoría de ingresos en miles (cating)* como la variable de fila, *Tiene PDA (pda)* como la variable de columna y *Nivel educativo (educ)* como la variable de capa.
2. Seleccione **Mostrar variables de capa en capas de tabla**.
3. Seleccione **Columna** en el cuadro de diálogo subordinado Mostrar en las casillas.
4. Ejecute el procedimiento de Tablas cruzadas, pulse dos veces en la tabla de tabulación cruzada y seleccione **Titulación universitaria** de la lista desplegable Nivel de estudios.

La vista seleccionada de la tabla de tabulación cruzada muestra los estadísticos de encuestados que tienen un título universitario.

Estadísticos de tablas cruzadas

Chi-cuadrado. Para las tablas con dos filas y dos columnas, seleccione **Chi-cuadrado** para calcular el chi-cuadrado de Pearson, el chi-cuadrado de la razón de verosimilitud, la prueba exacta de Fisher y el chi-cuadrado corregido de Yates (corrección por continuidad). Para las tablas 2x2, se calcula la prueba exacta de Fisher cuando una tabla (que no resulte de perder columnas o filas en una tabla mayor) presente una casilla con una frecuencia esperada menor que 5. Para las restantes tablas 2x2 se calcula el chi-cuadrado corregido de Yates. Para las tablas con cualquier número de filas y columnas, seleccione

Chi-cuadrado para calcular el chi-cuadrado de Pearson y el chi-cuadrado de la razón de verosimilitud. Cuando ambas variables de tabla son cuantitativas, **Chi-cuadrado** da como resultado la prueba de asociación lineal por lineal.

Correlaciones. Para las tablas en las que tanto las columnas como las filas contienen valores ordenados, **Correlaciones** da como resultado rho, el coeficiente de correlación de Spearman (sólo datos numéricos). La rho de Spearman es una medida de asociación entre órdenes de rangos. Cuando ambas variables de tabla (factores) son cuantitativas, **Correlaciones** da como resultado *r*, el coeficiente de correlación de Pearson, una medida de asociación lineal entre las variables.

Nominal. Para los datos nominales (sin orden intrínseco, como católico, protestante o judío), puede seleccionar el **Coeficiente de contingencia**, **Phi** (coeficiente) **y V de Cramér**, **Lambda** (lambdas simétricas y asimétricas y tau de Kruskal y Goodman) y el **Coeficiente de incertidumbre**.

- *Coeficiente de contingencia*. Medida de asociación basada en chi-cuadrado. El valor varía 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. El valor máximo posible depende del número de filas y columnas de la tabla.
- *Phi y V de Cramer*. Phi es una medida de asociación basada en chi-cuadrado que conlleva dividir el estadístico de chi-cuadrado por el tamaño de la muestra y extraer la raíz cuadrada del resultado. V de Cramer es una medida de asociación basada en chi-cuadradro.
- *Lambda*. Medida de asociación que refleja la reducción proporcional en el error cuando se utilizan los valores de la variable independiente para pronosticar los valores de la variable dependiente. Un valor igual a 1 significa que la variable independiente pronostica perfectamente la variable dependiente. Un valor igual a 0 significa que la variable independiente no ayuda a pronosticar la variable dependiente.
- *Coeficiente de incertidumbre*. Medida de asociación que refleja la reducción proporcional en el error cuando se utilizan los valores de una variable para pronosticar los valores de la otra variable. Por ejemplo, un valor de 0,83 indica que el conocimiento de una variable reduce en un 83% el error al pronosticar los valores de la otra variable. El programa calcula tanto la versión simétrica como la asimétrica del coeficiente de incertidumbre.

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 sólo 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 sólo 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 sólo 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.

Visualización en casillas de tablas cruzadas

Para ayudarle a descubrir las tramas en los datos que contribuyen a una prueba de chi-cuadrado significativa, el procedimiento Tablas cruzadas muestra las frecuencias esperadas y tres tipos de residuos (desviaciones) que miden la diferencia entre las frecuencias observadas y las esperadas. Cada casilla de la tabla puede contener cualquier combinación de recuentos, porcentajes y residuos seleccionados.

Recuentos. El número de casos realmente observados y el número de casos esperados si las variables de fila y columna son independientes entre sí. Puede optar por ocultar recuentos inferiores un número entero especificado. Los valores ocultos se mostrarán como **<N**, donde **N** es el número entero especificado. El entero especificado debe ser mayor o igual a 2, aunque se permite el valor 0 y especifica que no se hay recuentos ocultos.

Comparar las proporciones de columna. Esta opción calcula comparaciones por pares de proporciones de columnas e indica los pares de columnas (de una fila concreta) que son significativamente diferentes. Las diferencias significativas se indican en la tabla de tabulación cruzada con formato de estilo APA utilizando subíndices de letras y se calculan con un nivel de significación de 0,05. *Nota*: si se especifica esta opción sin seleccionar recuentos observados o porcentajes de columnas, se incluirán los recuentos observados en la tabla de tabulación cruzada, con subíndices de estilo APA indicando los resultados de las pruebas de proporciones de columnas.

- **Corregir valores p (método de Bonferroni).** Las comparaciones por parejas de las proporciones de columnas utilizan la corrección de Bonferroni, que ajusta el nivel de significación observado por el hecho de que se realizan múltiples comparaciones.

Porcentajes. Los porcentajes se pueden sumar a través de las filas o a lo largo de las columnas. También se encuentran disponibles los porcentajes del número total de casos representados en la tabla (una capa). *Nota*: si **Ocultar recuentos pequeños** está seleccionado en el grupo Recuentos, se ocultarán también los porcentajes asociados con recuentos ocultos.

Residuos. Los residuos brutos no tipificados presentan la diferencia entre los valores observados y los esperados. También se encuentran disponibles los residuos tipificados y tipificados corregidos.

- *No tipificados*. La diferencia entre un valor observado y el valor esperado. El valor esperado es el número de casos que se esperaría encontrar en la casilla si no hubiera relación entre las dos variables. Un residuo positivo indica que hay más casos en la casilla de los que habría en ella si las variables de fila y columna fueran independientes.
- *Tipificados*. El residuo dividido por una estimación de su error estándar. Los residuos tipificados, que son conocidos también como los residuos de Pearson o residuos estandarizados, tienen una media de 0 y una desviación estándar de 1.
- *Tipificados corregidos*. El residuo de una casilla (el valor observado menos el valor esperado) dividido por una estimación de su error estándar. El residuo estandarizado resultante viene expresado en unidades de desviación estándar, por encima o por debajo de la media.

Ponderaciones no enteras. Los recuentos de las casillas suelen ser valores enteros, ya que representan el número de casos de cada casilla. Sin embargo, si el archivo de datos está ponderado en un momento determinado por una variable de ponderación con valores fraccionarios (por ejemplo, 1,25), los recuentos de las casillas pueden que también sean valores fraccionarios. Puede truncar o redondear estos valores antes o después de calcular los recuentos de las casillas o bien utilizar recuentos de casillas fraccionarios en la presentación de las tablas y los cálculos de los estadísticos.

- *Redondear recuentos de casillas*. Las ponderaciones de los casos se utilizan tal cual, pero las ponderaciones acumuladas en las casillas se redondean antes de calcular cualquiera de los estadísticos.
- *Truncar recuentos de casillas*. Las ponderaciones de los casos se utilizan tal cual, pero las ponderaciones acumuladas en las casillas se truncan antes de calcular cualquiera de los estadísticos.
- *Redondear ponderaciones de casos*. Se redondean las ponderaciones de los casos antes de utilizarlas.
- *Truncar ponderaciones de casos*. Se truncan las ponderaciones de los casos antes de utilizarlas.
- *No efectuar correcciones*. Las ponderaciones de los casos se utilizan tal cual y se utilizan los recuentos de casillas fraccionales. Sin embargo, cuando se solicitan Estadísticos exactos (disponibles sólo con la opción Pruebas exactas), las ponderaciones acumuladas en las casillas se truncan o redondean antes de calcular los estadísticos de las Pruebas exactas.

Formato de tablas cruzadas

Puede ordenar las filas en orden ascendente o descendente de los valores de la variable de fila.

Capítulo 6. Resumir

El procedimiento Resumir calcula estadísticos de subgrupo para las variables dentro de las categorías de una o más variables de agrupación. Se cruzan todos los niveles de las variables de agrupación. Puede elegir el orden en el que se mostrarán los estadísticos. También se muestran estadísticos de resumen para cada variable a través de todas las categorías. Los valores de los datos en cada categoría pueden mostrarse en una lista o suprimirse. Con grandes conjuntos de datos, tiene la opción de listar sólo los primeros *n* casos.

Ejemplo. ¿Cuál es la media de las ventas por regiones o por tipo de cliente? Podrá descubrir que el importe medio de las ventas es ligeramente superior en la región occidental respecto a las demás regiones, y que la media más alta se da entre los clientes de empresas privadas de la zona occidental .

Estadísticos. Suma, número de casos, media, mediana, mediana agrupada, error estándar de la media, mínimo, máximo, rango, valor de la variable para la primera categoría de la variable de agrupación, valor de la variable para la última categoría de la variable de agrupación, desviación estándar, varianza, curtosis, error estándar de curtosis, asimetría, error estándar de asimetría, porcentaje de la suma total, porcentaje del *N* total, porcentaje de la suma en, porcentaje de *N* en, media geométrica y media armónica.

Resumir: Consideraciones sobre los datos

Datos. Las variables de agrupación son variables categóricas cuyos valores pueden ser numéricos o de cadena. El número de categorías debe ser razonablemente pequeño. Las otras variables deben poder ordenarse mediante rangos.

Supuestos. Algunos de los estadísticos opcionales de subgrupo, como la media y la desviación estándar, se basan en la teoría normal y son adecuados para variables cuantitativas con distribuciones simétricas. Los estadísticos robustos, tales como la mediana y el rango, son adecuados para las variables cuantitativas que pueden o no cumplir el supuesto de normalidad.

Para obtener resúmenes de casos

1. Seleccione en los menús:

   Analizar > **Informes** > **Resúmenes de casos...**
2. Seleccione una o más variables.

Si lo desea, puede:

- Seleccionar una o más variables de agrupación para dividir los datos en subgrupos.
- Pulsar en **Opciones** para cambiar el título de los resultados, añadir un texto al pie debajo de los resultados o excluir los casos con valores perdidos.
- Pulsar en **Estadísticos** para acceder a estadísticos adicionales.
- Seleccionar **Mostrar los casos** para listar los casos en cada subgrupo. De forma predeterminada, el sistema enumera sólo los 100 primeros casos del archivo. Puede aumentar o disminuir el valor de **Limitar los casos a los primeros** *n* o desactivar ese elemento para enumerar todos los casos.

Resumir: Opciones

Resumir permite cambiar el título de los resultados o añadir un texto que aparecerá debajo de la tabla de resultados. Puede controlar el ajuste de las líneas en los títulos y textos escribiendo `\n` en el lugar donde desee insertar una línea de separación.

Además, puede elegir entre mostrar o suprimir los subtítulos para los totales e incluir o excluir los casos con valores perdidos para cualquiera de las variables utilizadas en cualquiera de los análisis. A menudo es aconsejable representar los casos perdidos en los resultados con un punto o un asterisco. Introduzca un carácter, frase o código que desee que aparezca cuando haya un valor perdido; de lo contrario, no se aplicará ningún tratamiento especial a los casos perdidos en los resultados.

Resumir: Estadísticos

Puede elegir uno o más de los siguientes estadísticos de subgrupo para las variables dentro de cada categoría de cada variable de agrupación: suma, número de casos, media, mediana, mediana agrupada, error estándar de la media, mínimo, máximo, rango, valor de la variable para la primera categoría de la variable de agrupación, valor de la variable para la última categoría de la variable de agrupación, desviación estándar, varianza, curtosis, error estándar de curtosis, asimetría, error estándar de asimetría, porcentaje de la suma total, porcentaje del *N* total, porcentaje de la suma en, porcentaje de *N* en, media geométrica y media armónica. El orden en el que aparecen los estadísticos en la lista Estadísticos de casilla es el orden en el que se mostrarán en los resultados. También se muestran estadísticos de resumen para cada variable a través de todas las categorías.

Primero. Muestra el primer valor de los datos encontrado en el archivo de datos.

Media geométrica. La raíz enésima del producto de los valores de los datos, donde n representa el número de casos.

Mediana agrupada. La mediana calculada para los datos que se codifican en grupos. Por ejemplo, con datos de edades, si cada valor de los 30 se ha codificado como 35, cada valor de los 40 como 45 y así sucesivamente, la mediana agrupada es la mediana calculada a partir de los datos codificados.

Media armónica. Se utiliza para estimar el tamaño promedio de un grupo cuando los tamaños de las muestras de los grupos no son iguales. La media armónica es el número total de muestras divido por la suma de los inversos de los tamaños de las muestras.

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.

Último. Muestra el último valor de los datos encontrado en el archivo de datos.

Máximo. Se trata del valor mayor de una variable numérica.

Media. Una medida de tendencia central. El promedio aritmético, la suma dividida por el número de casos.

Mediana. Es el valor por encima y por debajo del cual se encuentran la mitad de los casos, el percentil 50. Si hay un número par de casos, la mediana es la media de los dos valores centrales, cuando los casos se ordenan en orden ascendente o descendente. La mediana es una medida de tendencia central que no es sensible a los valores atípicos (a diferencia de la media, que puede resultar afectada por unos pocos valores extremadamente altos o bajos).

Mínimo. Se trata del valor menor de una variable numérica.

N. Número de casos (observaciones o registros).

Porcentaje del N total. Porcentaje del número total de casos en cada categoría.

Porcentaje de la suma total. Porcentaje de la suma total en cada categoría.

Rango. Diferencia entre los valores mayor y menor de una variable numérica; el máximo menos el mínimo.

Asimetría. Medida de la asimetría de una distribución. La distribución normal es simétrica y tiene un valor de asimetría igual a 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. 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.

Desviación estándar. Una medida de dispersión sobre la media. En una distribución normal, el 68% de los casos se encuentra dentro de una desviación estándar de la media y el 95% queda entre dos desviaciones estándar. Por ejemplo, si la edad media es de 45 años, con una desviación estándar de 10, el 95% de los casos estaría entre los 25 y 65 en una distribución normal.

Error estándar de curtosis. La razón de la curtosis sobre su error estándar puede utilizarse como prueba de normalidad (es decir, se puede rechazar la normalidad si la razón es menor que -2 o mayor que +2). Un valor grande y positivo para la curtosis indica que las colas son más largas que las de una distribución normal; por el contrario, un valor extremo y negativo indica que las colas son más cortas (llegando a tener forma de caja como en la distribución uniforme).

Error estándar de la media. Medida de cuánto puede variar el valor de la media de una muestra a otra, extraídas éstas de la misma distribución. Puede utilizarse para comparar de forma aproximada la media observada respecto a un valor hipotetizado (es decir, se puede concluir que los dos valores son distintos si la diferencia entre ellos, dividida por el error estándar, es menor que -2 o mayor que +2).

Error estándar de asimetría. La razón de la asimetría sobre su error estándar puede utilizarse como una prueba de normalidad (es decir, se puede rechazar la normalidad si la razón es menor que -2 o mayor que +2). Un valor grande y positivo para la asimetría indica una cola larga a la derecha; un valor extremo y negativo indica una cola larga por la izquierda

Suma. Suma o total de todos los valores, a lo largo de todos los casos con valores no perdidos.

Varianza. Medida de dispersión sobre la media, igual a la suma de las desviaciones al cuadrado de la media divida por el número de casos menos uno. La varianza se mide en unidades que son el cuadrado de las de la variable en cuestión.

Capítulo 7. Medias

El procedimiento Medias calcula medias de subgrupo y estadísticos univariados relacionados para variables dependientes dentro de las categorías de una o más variables independientes. Si lo desea, puede obtener el análisis de varianza de un factor, la eta y pruebas de linealidad.

Ejemplo. Mida la cantidad media de grasa absorbida en función de tres tipos distintos de aceite comestible y realice un análisis de varianza de un factor para comprobar si difieren las medias.

Estadísticos. Suma, número de casos, media, mediana, mediana agrupada, error estándar de la media, mínimo, máximo, rango, valor de la variable para la primera categoría de la variable de agrupación, valor de la variable para la última categoría de la variable de agrupación, desviación estándar, varianza, curtosis, error estándar de curtosis, asimetría, error estándar de asimetría, porcentaje de la suma total, porcentaje del *N* total, porcentaje de la suma en, porcentaje de *N* en, media geométrica y media armónica. Las opciones incluyen: análisis de varianza, eta, eta cuadrado y pruebas de linealidad de *R* y R^2.

Medias: Consideraciones sobre los datos

Datos. Las variables dependientes son cuantitativas y las independientes son categóricas. Los valores de las variables categóricas pueden ser numéricos o de cadena.

Supuestos. Algunos de los estadísticos opcionales de subgrupo, como la media y la desviación estándar, se basan en la teoría normal y son adecuados para variables cuantitativas con distribuciones simétricas. Los estadísticos robustos, tales como la mediana son adecuados para las variables cuantitativas que pueden o no cumplir el supuesto de normalidad. El análisis de varianza es robusto a las desviaciones de la normalidad, aunque los datos de cada casilla deberían ser simétricos. El análisis de varianza también supone que los grupos proceden de poblaciones con la misma varianza. Para comprobar este supuesto, utilice la prueba de homogeneidad de las varianzas de Levene, disponible en el procedimiento ANOVA de un factor.

Para obtener medias de subgrupo

1. Seleccione en los menús:

   Analizar > **Comparar medias** > **Medias...**
2. Seleccione una o más variables dependientes.
3. Utilice uno de los siguientes métodos para seleccionar variables independientes categóricas:
   - Seleccionar una o más variables independientes. Se mostrarán resultados individuales para cada variable independiente.
   - Seleccione una o más capas de variables independientes. Cada capa subdivide consecutivamente la muestra. Si tiene una variable independiente en Capa 1 y otra variable independiente en Capa 2, los resultados se mostrarán en una tabla cruzada en contraposición a tablas individuales para cada variable independiente.
4. Si lo desea, pulse en **Opciones** si desea obtener estadísticos opcionales, una tabla de análisis de varianza, eta, eta cuadrado, *R*, y R^2.

Medias: Opciones

You can choose one or more of the following subgroup statistics for the variables within each category of each grouping variable: suma, número de casos, media, mediana, mediana agrupada, error estándar de la media, mínimo, máximo, rango, valor de la variable para la primera categoría de la variable de agrupación, valor de la variable para la última categoría de la variable de agrupación, desviación estándar, varianza, curtosis, error estándar de curtosis, asimetría, error estándar de asimetría, porcentaje

de la suma total, porcentaje del *N* total, porcentaje de la suma en, porcentaje de *N* en, media geométrica, media armónica. Se puede cambiar el orden de aparición de los estadísticos de subgrupo. El orden en el que aparecen en la lista Estadísticos de casilla es el mismo orden que presentarán en los resultados. También se muestran estadísticos de resumen para cada variable a través de todas las categorías.

Primero. Muestra el primer valor de los datos encontrado en el archivo de datos.

Media geométrica. La raíz enésima del producto de los valores de los datos, donde n representa el número de casos.

Mediana agrupada. La mediana calculada para los datos que se codifican en grupos. Por ejemplo, con datos de edades, si cada valor de los 30 se ha codificado como 35, cada valor de los 40 como 45 y así sucesivamente, la mediana agrupada es la mediana calculada a partir de los datos codificados.

Media armónica. Se utiliza para estimar el tamaño promedio de un grupo cuando los tamaños de las muestras de los grupos no son iguales. La media armónica es el número total de muestras divido por la suma de los inversos de los tamaños de las muestras.

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.

Último. Muestra el último valor de los datos encontrado en el archivo de datos.

Máximo. Se trata del valor mayor de una variable numérica.

Media. Una medida de tendencia central. El promedio aritmético, la suma dividida por el número de casos.

Mediana. Es el valor por encima y por debajo del cual se encuentran la mitad de los casos, el percentil 50. Si hay un número par de casos, la mediana es la media de los dos valores centrales, cuando los casos se ordenan en orden ascendente o descendente. La mediana es una medida de tendencia central que no es sensible a los valores atípicos (a diferencia de la media, que puede resultar afectada por unos pocos valores extremadamente altos o bajos).

Mínimo. Se trata del valor menor de una variable numérica.

N. Número de casos (observaciones o registros).

Porcentaje del N total. Porcentaje del número total de casos en cada categoría.

Porcentaje de la suma total. Porcentaje de la suma total en cada categoría.

Rango. Diferencia entre los valores mayor y menor de una variable numérica; el máximo menos el mínimo.

Asimetría. Medida de la asimetría de una distribución. La distribución normal es simétrica y tiene un valor de asimetría igual a 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. 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.

Desviación estándar. Una medida de dispersión sobre la media. En una distribución normal, el 68% de los casos se encuentra dentro de una desviación estándar de la media y el 95% queda entre dos desviaciones estándar. Por ejemplo, si la edad media es de 45 años, con una desviación estándar de 10, el 95% de los casos estaría entre los 25 y 65 en una distribución normal.

Error estándar de curtosis. La razón de la curtosis sobre su error estándar puede utilizarse como prueba de normalidad (es decir, se puede rechazar la normalidad si la razón es menor que -2 o mayor que +2). Un valor grande y positivo para la curtosis indica que las colas son más largas que las de una distribución normal; por el contrario, un valor extremo y negativo indica que las colas son más cortas (llegando a tener forma de caja como en la distribución uniforme).

Error estándar de la media. Medida de cuánto puede variar el valor de la media de una muestra a otra, extraídas éstas de la misma distribución. Puede utilizarse para comparar de forma aproximada la media observada respecto a un valor hipotetizado (es decir, se puede concluir que los dos valores son distintos si la diferencia entre ellos, dividida por el error estándar, es menor que -2 o mayor que +2).

Error estándar de asimetría. La razón de la asimetría sobre su error estándar puede utilizarse como una prueba de normalidad (es decir, se puede rechazar la normalidad si la razón es menor que -2 o mayor que +2). Un valor grande y positivo para la asimetría indica una cola larga a la derecha; un valor extremo y negativo indica una cola larga por la izquierda

Suma. Suma o total de todos los valores, a lo largo de todos los casos con valores no perdidos.

Varianza. Medida de dispersión sobre la media, igual a la suma de las desviaciones al cuadrado de la media divida por el número de casos menos uno. La varianza se mide en unidades que son el cuadrado de las de la variable en cuestión.

Estadísticos para la primera capa

Tabla de Anova y eta. Muestra una tabla de análisis de varianza de un factor y calcula la eta y la eta cuadrado (medidas de asociación) para cada variable independiente de la primera capa.

Contrastes de linealidad. Calcula la suma de cuadrados, los grados de libertad y la media cuadrática asociados a los componentes lineal y no lineal, así como la razón F, la R y la R cuadrado. Si la variable independiente es una cadena corta entonces la linealidad no se calcula.

Capítulo 8. Cubos OLAP

El procedimiento Cubos OLAP (siglas del inglés On-Line Analytic Processing, «Procesamiento analítico interactivo») calcula totales, medias y otros estadísticos univariantes para variables de resumen continuas dentro de las categorías de una o más variables categóricas de agrupación. En la tabla se creará una nueva capa para cada categoría de cada variable de agrupación.

Ejemplo. El total y el promedio de ventas para diversas regiones y líneas de producto, dentro de las regiones.

Estadísticos. Suma, número de casos, media, mediana, mediana agrupada, error estándar de la media, mínimo, máximo, rango, valor de la variable para la primera categoría de la variable de agrupación, valor de la variable para la última categoría de la variable de agrupación, desviación estándar, varianza, curtosis, error estándar de curtosis, asimetría, error estándar de asimetría, porcentaje de casos totales, porcentaje de la suma total, porcentaje de casos totales dentro de las variables agrupadas, porcentaje de la suma total dentro de las variables agrupadas, media geométrica y media armónica.

Cubos OLAP: Consideraciones sobre los datos

Datos. Las variables de resumen son cuantitativas (variables continuas medidas en una escala de intervalo o de razón) y las variables de agrupación son categóricas. Los valores de las variables categóricas pueden ser numéricos o de cadena.

Supuestos. Algunos de los estadísticos opcionales de subgrupo, como la media y la desviación estándar, se basan en la teoría normal y son adecuados para variables cuantitativas con distribuciones simétricas. Los estadísticos robustos, tales como la mediana y el rango, son adecuados para las variables cuantitativas que pueden o no cumplir el supuesto de normalidad.

Para obtener cubos OLAP

1. Seleccione en los menús:

   Analizar > **Informes** > **Cubos OLAP...**
2. Seleccione una o más variables de resumen continuas.
3. Seleccione una o más variables de agrupación categóricas.

Si lo desea:

- Seleccionar diferentes estadísticos de resumen (pulse en **Estadísticos**). Debe seleccionar una o más variables de agrupación para poder seleccionar estadísticos de resumen.
- Calcule las diferencias existentes entre los pares de variables y los pares de grupos definidos por una variable de agrupación (pulse en **Diferencias**).
- Crear títulos de tabla personalizados (pulse en **Título**).
- Oculta recuentos que sean inferiores a un entero especificado. Los valores ocultos se mostrarán como **<N**, donde **N** es el número entero especificado. El número entero especificado debe ser mayor o igual a 2.

Cubos OLAP: Estadísticos

Puede elegir uno o varios de los siguientes estadísticos de subgrupo para las variables de resumen dentro de cada categoría de cada variable de agrupación: Suma, Número de casos, Media, Mediana, Mediana agrupada, Error estándar de la media, Mínimo, Máximo, Rango, Valor de la variable para la primera categoría de la variable de agrupación, Valor de la variable para la última categoría de la variable de agrupación, Desviación estándar, Varianza, Curtosis, Error estándar de curtosis, Asimetría, Error estándar

de asimetría, Porcentaje de casos totales, Porcentaje de la suma total, Porcentaje de casos totales dentro de las variables de agrupación, Porcentaje de la suma total dentro de las variables de agrupación, Media geométrica y Media armónica.

Se puede cambiar el orden de aparición de los estadísticos de subgrupo. El orden en el que aparecen en la lista Estadísticos de casilla es el mismo orden que presentarán en los resultados. También se muestran estadísticos de resumen para cada variable a través de todas las categorías.

Primero. Muestra el primer valor de los datos encontrado en el archivo de datos.

Media geométrica. La raíz enésima del producto de los valores de los datos, donde n representa el número de casos.

Mediana agrupada. La mediana calculada para los datos que se codifican en grupos. Por ejemplo, con datos de edades, si cada valor de los 30 se ha codificado como 35, cada valor de los 40 como 45 y así sucesivamente, la mediana agrupada es la mediana calculada a partir de los datos codificados.

Media armónica. Se utiliza para estimar el tamaño promedio de un grupo cuando los tamaños de las muestras de los grupos no son iguales. La media armónica es el número total de muestras divido por la suma de los inversos de los tamaños de las muestras.

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.

Último. Muestra el último valor de los datos encontrado en el archivo de datos.

Máximo. Se trata del valor mayor de una variable numérica.

Media. Una medida de tendencia central. El promedio aritmético, la suma dividida por el número de casos.

Mediana. Es el valor por encima y por debajo del cual se encuentran la mitad de los casos, el percentil 50. Si hay un número par de casos, la mediana es la media de los dos valores centrales, cuando los casos se ordenan en orden ascendente o descendente. La mediana es una medida de tendencia central que no es sensible a los valores atípicos (a diferencia de la media, que puede resultar afectada por unos pocos valores extremadamente altos o bajos).

Mínimo. Se trata del valor menor de una variable numérica.

N. Número de casos (observaciones o registros).

Porcentaje del N en. Porcentaje del número de casos para la variable de agrupación especificada dentro de las categorías de otras variables de agrupación. Si sólo tiene una variable de agrupación, este valor es idéntico al porcentaje del número de casos total.

Porcentaje de la suma en. Porcentaje de la suma para la variable de agrupación especificada dentro de las categorías de otras variables de agrupación. Si sólo tiene una variable de agrupación, este valor es idéntico al porcentaje de la suma total.

Porcentaje del N total. Porcentaje del número total de casos en cada categoría.

Porcentaje de la suma total. Porcentaje de la suma total en cada categoría.

Rango. Diferencia entre los valores mayor y menor de una variable numérica; el máximo menos el mínimo.

Asimetría. Medida de la asimetría de una distribución. La distribución normal es simétrica y tiene un valor de asimetría igual a 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. 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.

Desviación estándar. Una medida de dispersión sobre la media. En una distribución normal, el 68% de los casos se encuentra dentro de una desviación estándar de la media y el 95% queda entre dos desviaciones estándar. Por ejemplo, si la edad media es de 45 años, con una desviación estándar de 10, el 95% de los casos estaría entre los 25 y 65 en una distribución normal.

Error estándar de curtosis. La razón de la curtosis sobre su error estándar puede utilizarse como prueba de normalidad (es decir, se puede rechazar la normalidad si la razón es menor que -2 o mayor que +2). Un valor grande y positivo para la curtosis indica que las colas son más largas que las de una distribución normal; por el contrario, un valor extremo y negativo indica que las colas son más cortas (llegando a tener forma de caja como en la distribución uniforme).

Error estándar de la media. Medida de cuánto puede variar el valor de la media de una muestra a otra, extraídas éstas de la misma distribución. Puede utilizarse para comparar de forma aproximada la media observada respecto a un valor hipotetizado (es decir, se puede concluir que los dos valores son distintos si la diferencia entre ellos, dividida por el error estándar, es menor que -2 o mayor que +2).

Error estándar de asimetría. La razón de la asimetría sobre su error estándar puede utilizarse como una prueba de normalidad (es decir, se puede rechazar la normalidad si la razón es menor que -2 o mayor que +2). Un valor grande y positivo para la asimetría indica una cola larga a la derecha; un valor extremo y negativo indica una cola larga por la izquierda

Suma. Suma o total de todos los valores, a lo largo de todos los casos con valores no perdidos.

Varianza. Medida de dispersión sobre la media, igual a la suma de las desviaciones al cuadrado de la media divida por el número de casos menos uno. La varianza se mide en unidades que son el cuadrado de las de la variable en cuestión.

Cubos OLAP: Diferencias

Este cuadro de diálogo le permite calcular el porcentaje y las diferencias aritméticas entre las variables de resumen o entre los grupos definidos por una variable de agrupación. Las diferencias se calculan para todas las medidas seleccionadas en el cuadro de diálogo Cubos OLAP: Estadísticos.

Diferencias entre variables. Calcula las diferencias entre pares de variables. Los valores de los estadísticos de resumen para la segunda variable de cada par (la variable Menos) se restan de los valores de los estadísticos de resumen correspondientes a la primera variable del par. En cuanto a las diferencias porcentuales, el valor de la variable de resumen para la variable Menos es el que se usa como denominador. Debe seleccionar al menos dos variables de resumen en el cuadro de diálogo principal para poder especificar las diferencias entre las variables.

Diferencias entre grupos de casos. Calcula las diferencias entre pares de grupos definidos por una variable de agrupación. Los valores de los estadísticos de resumen para la segunda categoría de cada par (la variable Menos) se restan de los valores de los estadísticos de resumen correspondientes a la primera categoría del par. Las diferencias porcentuales utilizan el valor del estadístico de resumen de la categoría

Menos como denominador. Debe seleccionar una o más variables de agrupación en el cuadro de diálogo principal para poder especificar las diferencias entre los grupos.

Cubos OLAP: Título

Puede cambiar el título de los resultados o añadir un texto al pie que aparecerá debajo de la tabla de resultados. También puede controlar el ajuste de las líneas de los títulos y de los textos al pie escribiendo `\n` en el lugar del texto donde desee insertar una línea de separación.

Capítulo 9. Pruebas T

Pruebas T

Hay tres tipos de pruebas *t*:

Prueba T para muestras independientes (prueba T para dos muestras). Compara las medias de una variable para dos grupos de casos. Se ofrecen estadísticos descriptivos para cada grupo y la prueba de Levene sobre la igualdad de las varianzas, así como valores *t* de igualdad de varianzas y varianzas desiguales y un intervalo de confianza al 95% para la diferencia entre las medias.

Prueba T para muestras relacionadas (prueba T dependiente). Compara las medias de dos variables en un solo grupo. Esta prueba también se utiliza para pares relacionados o diseños de estudio de control de casos. El resultado incluye estadísticos descriptivos de las variables que se van a contrastar, la correlación entre ellas, estadísticos descriptivos de las diferencias emparejadas, la prueba *t* y un intervalo de confianza al 95%.

Prueba t para una muestra. Compara la media de una variable con un valor conocido o hipotetizado. Se muestran estadísticos descriptivos para las variables de contraste junto con la prueba *t*. De forma predeterminada, en los resultados se incluye un intervalo de confianza al 95% para la diferencia entre la media de la variable de contraste y el valor hipotetizado de la prueba.

Prueba T para muestras independientes

El procedimiento Prueba T para muestras independientes compara las medias de dos grupos de casos. Lo ideal es que para esta prueba los sujetos se asignen aleatoriamente a dos grupos, de forma que cualquier diferencia en la respuesta sea debida al tratamiento (o falta de tratamiento) y no a otros factores. Este caso no ocurre si se comparan los ingresos medios para hombres y mujeres. El sexo de una persona no se asigna aleatoriamente. En estas situaciones, debe asegurarse de que las diferencias en otros factores no enmascaren o resalten una diferencia significativa entre las medias. Las diferencias de ingresos medios pueden estar sometidas a la influencia de factores como los estudios (y no solamente el sexo).

Ejemplo. Se asigna aleatoriamente un grupo de pacientes con hipertensión arterial a un grupo con placebo y otro con tratamiento. Los sujetos con placebo reciben una pastilla inactiva y los sujetos con tratamiento reciben un nuevo medicamento del cual se espera que reduzca la tensión arterial. Después de tratar a los sujetos durante dos meses, se utiliza la prueba *t* para dos muestras para comparar la tensión arterial media del grupo con placebo y del grupo con tratamiento. Cada paciente se mide una sola vez y pertenece a un solo grupo.

Estadísticos. Para cada variable: tamaño de la muestra, media, desviación estándar y error estándar de la media. Para la diferencia entre las medias: media, error estándar e intervalo de confianza (puede especificar el nivel de confianza). Pruebas: prueba de Levene sobre la igualdad de varianzas y pruebas *t* de varianzas combinadas y separadas sobre la igualdad de las medias.

Prueba T para muestras independientes: Consideraciones sobre los datos

Datos. Los valores de la variable cuantitativa de interés se hallan en una única columna del archivo de datos. El procedimiento utiliza una variable de agrupación con dos valores para separar los casos en dos grupos. La variable de agrupación puede ser numérica (valores como 1 y 2, o 6,25 y 12,5) o de cadena corta (como *sí* y *no*). También puede usar una variable cuantitativa, como la *edad*, para dividir los casos en dos grupos especificando un punto de corte (el punto de corte 21 divide la *edad* en un grupo de menos de 21 años y otro de más de 21).

Supuestos. Para la prueba *t* de igualdad de varianzas, las observaciones deben ser muestras aleatorias independientes de distribuciones normales con la misma varianza de población. Para la prueba *t* de varianzas desiguales, las observaciones deben ser muestras aleatorias independientes de distribuciones normales. La prueba *t* para dos muestras es bastante robusta a las desviaciones de la normalidad. Al contrastar las distribuciones gráficamente, compruebe que son simétricas y que no contienen valores atípicos.

Para obtener una prueba T para muestras independientes

1. Seleccione en los menús:

   Analizar > **Comparar medias** > **Prueba T para muestras independientes...**
2. Seleccione una o más variables de contraste cuantitativas. Se calcula una prueba *t* para cada variable.
3. Seleccione una sola variable de agrupación y pulse en **Definir grupos** para especificar dos códigos para los grupos que desee comparar.
4. Si lo desea, puede pulsar en **Opciones** para controlar el tratamiento de los datos perdidos y el nivel del intervalo de confianza.

Definición de grupos en la prueba T para muestras independientes

Para las variables de agrupación numéricas, defina los dos grupos de la prueba *t* especificando dos valores o un punto de corte:

- **Usar valores especificados.** Escriba un valor para el Grupo 1 y otro para el Grupo 2. Los casos con otros valores quedarán excluidos del análisis. Los números no tienen que ser enteros (por ejemplo, 6,25 y 12,5 son válidos).
- **Punto de corte.** Escriba un número que divida los valores de la variable de agrupación en dos conjuntos. Todos los casos con valores menores que el punto de corte forman un grupo y los casos con valores mayores o iguales que el punto de corte forman el otro grupo.

Para las variables de agrupación de cadena, escriba una cadena para el Grupo 1 y otra para el Grupo 2; por ejemplo *sí* y *no*. Los casos con otras cadenas se excluyen del análisis.

Prueba T para muestras independientes: Opciones

Intervalo de confianza. De forma predeterminada, se muestra un intervalo de confianza al 95% para la diferencia entre las medias. Introduzca un valor entre 1 y 99 para solicitar otro nivel de confianza.

Valores perdidos. Si ha probado varias variables y se han perdido los datos de una o más de ellas, puede indicar al procedimiento qué casos desea incluir (o excluir).

- **Excluir casos según análisis.** Cada prueba *t* utiliza todos los casos que tienen datos válidos para las variables contrastadas. Los tamaños muestrales pueden variar de una prueba a otra.
- **Excluir casos según lista.** Cada prueba *t* utiliza sólo aquellos casos que contienen datos válidos para todas las variables utilizadas en las pruebas *t* solicitadas. El tamaño de la muestra es constante en todas las pruebas.

Prueba T para muestras relacionadas

El procedimiento Prueba T para muestras relacionadas compara las medias de dos variables de un solo grupo. El procedimiento calcula las diferencias entre los valores de las dos variables de cada caso y contrasta si la media difiere de 0.

Ejemplo. En un estudio sobre la hipertensión sanguínea, se toma la tensión a todos los pacientes al comienzo del estudio, se les aplica un tratamiento y se les toma la tensión otra vez. De esta manera, a cada sujeto le corresponden dos medidas, normalmente denominadas medidas *pre* y *post*. Un diseño alternativo para el que se utiliza esta prueba consiste en un estudio de pares relacionados o un estudio de control de casos en el que cada registro en el archivo de datos contiene la respuesta del paciente y de su

sujeto de control correspondiente. En un estudio sobre la tensión sanguínea, pueden emparejarse pacientes y controles por edad (un paciente de 75 años con un miembro del grupo de control de 75 años).

Estadísticos. Para cada variable: media, tamaño de la muestra, desviación estándar y error estándar de la media. Para cada par de variables: correlación, diferencia promedio entre las medias, prueba *t* de intervalo de confianza para la diferencia entre las medias (puede especificarse el nivel de confianza). Desviación estándar y error estándar de la diferencia entre las medias.

Prueba T para muestras relacionadas: Consideraciones sobre los datos

Datos. Especifique dos variables cuantitativas (nivel de medición de intervalo o de razón) para cada prueba de pares. En un estudio de pares relacionados o de control de casos, la respuesta de cada sujeto de la prueba y su sujeto de control correspondiente deberán hallarse en el mismo caso en el archivo de datos.

Supuestos. Las observaciones de cada par deben hacerse en las mismas condiciones. Las diferencias entre las medias deben estar normalmente distribuidas. Las varianzas de cada variable pueden ser iguales o desiguales.

Para obtener una prueba T para muestras relacionadas

1. Seleccione en los menús:

   Analizar > **Comparar medias** > **Prueba T para muestras relacionadas...**
2. Seleccione uno o más pares de variables
3. Si lo desea, puede pulsar en **Opciones** para controlar el tratamiento de los datos perdidos y el nivel del intervalo de confianza.

Prueba T para muestras relacionadas: Opciones

Intervalo de confianza. De forma predeterminada, se muestra un intervalo de confianza al 95% para la diferencia entre las medias. Introduzca un valor entre 1 y 99 para solicitar otro nivel de confianza.

Valores perdidos. Si ha probado varias variables y se han perdido los datos de una o más de ellas, puede indicar al procedimiento qué casos desea incluir (o excluir):

- **Excluir casos según análisis.** Cada prueba *t* utilizará todos los casos que contienen datos válidos para la pareja de variables contrastadas. Los tamaños muestrales pueden variar de una prueba a otra.
- **Excluir casos según lista.** Cada prueba *t* utilizará únicamente los casos que contengan datos válidos para todas las parejas de variables contrastadas. El tamaño de la muestra es constante en todas las pruebas.

Características adicionales del comando T-TEST

La sintaxis de comandos también le permite:

- Producir pruebas t tanto de una sola muestra como de muestras independientes ejecutando un solo comando.
- Contrastar una variable con todas las variables de una lista, en una prueba relacionada (mediante el subcomando `PAIRS`).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Prueba T para una muestra

El procedimiento Prueba T para una muestra contrasta si la media de una sola variable difiere de una constante especificada.

Ejemplos. Un investigador desea comprobar si la puntuación media del coeficiente intelectual de un grupo de alumnos difiere de 100. O bien, un fabricante de copos de cereales puede tomar una muestra de envases de la línea de producción y comprobar si el peso medio de las muestras difiere de 1 kg con un nivel de confianza al 95%.

Estadísticos. Para cada variable de prueba: media, desviación estándar y error estándar de la media. La diferencia promedio entre cada valor de los datos y el valor de contraste hipotetizado, una prueba *t* que contrasta que esta diferencia es 0 y un intervalo de confianza para la diferencia promedio (para el que puede especificarse el nivel de confianza).

Prueba T para una muestra: Consideraciones sobre los datos

Datos. Para contrastar los valores de una variable cuantitativa con un valor de contraste hipotetizado, elija una variable cuantitativa e introduzca un valor de contraste hipotetizado.

Supuestos. Esta prueba asume que los datos están normalmente distribuidos; sin embargo, esta prueba es bastante robusto frente a las desviaciones de la normalidad.

Para obtener una prueba T para una muestra

1. Seleccione en los menús:

   Analizar > **Comparar medias** > **Prueba T para una muestra...**
2. Seleccione una o más variables para contrastarlas con el mismo valor hipotetizado.
3. Introduzca un valor de contraste numérico para compararlo con cada media muestral.
4. Si lo desea, puede pulsar en **Opciones** para controlar el tratamiento de los datos perdidos y el nivel del intervalo de confianza.

Prueba T para una muestra: Opciones

Intervalo de confianza. De forma predeterminada, se muestra un intervalo de confianza al 95% para la diferencia entre la media y el valor de contraste hipotetizado. Introduzca un valor entre 1 y 99 para solicitar otro nivel de confianza.

Valores perdidos. Si ha probado varias variables y se han perdido los datos de una o más de ellas, puede indicar al procedimiento qué casos desea incluir (o excluir).

- **Excluir casos según análisis.** Cada prueba *t* utiliza todos los casos que tienen datos válidos para la variable contrastada. Los tamaños muestrales pueden variar de una prueba a otra.
- **Excluir casos según lista.** Cada prueba *t* utiliza sólo aquellos casos que contienen datos válidos para todas las variables utilizadas en las pruebas *t* solicitadas. El tamaño de la muestra es constante en todas las pruebas.

Características adicionales del comando T-TEST

La sintaxis de comandos también le permite:

- Producir pruebas t tanto de una sola muestra como de muestras independientes ejecutando un solo comando.
- Contrastar una variable con todas las variables de una lista, en una prueba relacionada (mediante el subcomando `PAIRS`).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Características adicionales del comando T-TEST

La sintaxis de comandos también le permite:

- Producir pruebas t tanto de una sola muestra como de muestras independientes ejecutando un solo comando.
- Contrastar una variable con todas las variables de una lista, en una prueba relacionada (mediante el subcomando PAIRS).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 10. ANOVA de un factor

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

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

Ejemplo. Las rosquillas absorben diferentes cantidades de grasa cuando se fríen. Se plantea un experimento utilizando tres tipos de grasas: aceite de cacahuete, aceite de maíz y manteca de cerdo. El aceite de cacahuete y el aceite de maíz son grasas no saturadas y la manteca es una grasa saturada. Además de determinar si la cantidad de grasa absorbida depende del tipo de grasa utilizada, también se podría preparar un contraste a priori para determinar si la cantidad de absorción de la grasa difiere para las grasas saturadas y las no saturadas.

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

ANOVA de un factor: Consideraciones sobre los datos

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

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

Para obtener un análisis de varianza de un factor

1. Seleccione en los menús:

   Analizar > **Comparar medias** > **ANOVA de un factor...**
2. Seleccione una o más variables dependientes.
3. Seleccione una sola variable de factor independiente.

ANOVA de un factor: Contrastes

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

Polinómico. Divide las sumas de cuadrados inter-grupos en componentes de tendencia. Puede contrastar la existencia de tendencia en la variable dependiente a través de los niveles ordenados de la variable de

factor. Por ejemplo, podría contrastar si existe una tendencia lineal (creciente o decreciente) en el salario, a través de los niveles ordenados de la titulación mayor obtenida.

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

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

El orden de los coeficientes es importante porque se corresponde con el orden ascendente de los valores de las categorías de la variable de factor. El primer coeficiente en la lista se corresponde con el menor de los valores de grupo en la variable de factor y el último coeficiente se corresponde con el valor más alto. Por ejemplo, si existen seis categorías en la variable factor, los coeficientes -1, 0, 0, 0, 0,5 y 0,5 contrastan el primer grupo con los grupos quinto y sexto. Para la mayoría de las aplicaciones, la suma de los coeficientes debería ser 0. Los conjuntos que no sumen 0 también se pueden utilizar, pero aparecerá un mensaje de advertencia.

ANOVA de un factor: Contrastes post hoc

Una vez que se ha determinado que existen diferencias entre las medias, las pruebas de rango post hoc y las comparaciones múltiples por parejas permiten determinar qué medias difieren. Las pruebas de rango identifican subconjuntos homogéneos de medias que no se diferencian entre sí. Las comparaciones múltiples por parejas contrastan la diferencia entre cada pareja de medias y generan una matriz donde los asteriscos indican las medias de grupo significativamente diferentes a un nivel alfa de 0,05.

Asumiendo varianzas iguales

La prueba de la diferencia honestamente significativa de Tukey, la GT2 de Hochberg, la prueba de Gabriel y la prueba de Scheffé son pruebas de comparaciones múltiples y pruebas de rango. Otras pruebas de rango disponibles son la *b* de Tukey, S-N-K (Student-Newman-Keuls), Duncan, R-E-G-W *F* (prueba *F* de Ryan-Einot-Gabriel-Welsch), R-E-G-W *Q* (prueba de rango de Ryan-Einot-Gabriel-Welsch) y Waller-Duncan. Las pruebas de comparaciones múltiples disponibles son Bonferroni, Diferencia honestamente significativa de Tukey, Sidak, Gabriel, Hochberg, Dunnett, Scheffé y DMS (diferencia menos significativa).

- *DMS*. Utiliza pruebas t para realizar todas las comparaciones por pares entre las medias de los grupos. La tasa de error no se corrige para realizar múltiples comparaciones.
- *Bonferroni*. Utiliza las pruebas de t para realizar comparaciones por pares entre las medias de los grupos, pero controla la tasa de error global estableciendo que la tasa de error de cada prueba sea igual a la tasa de error por experimento dividida entre el número total de contrastes. Así, se corrige el nivel de significación observado por el hecho de que se están realizando múltiples comparaciones.
- *Sidak*. Prueba de comparaciones múltiples por parejas basada en un estadístico t. La prueba de Sidak corrige el nivel de significación para las comparaciones múltiples y da lugar a límites más estrechos que los de Bonferroni.
- *Scheffe*. Realiza comparaciones múltiples conjuntas por parejas para todas las parejas de combinaciones de las medias posibles. Utiliza la distribución muestral F. Puede utilizarse para examinar todas las combinaciones lineales de grupos de medias posibles, no sólo las comparaciones por parejas.
- *R-E-G-W F*. Procedimiento múltiple por pasos (por tamaño de las distancias) de Ryan-Einot-Gabriel-Welsch que se basa en una prueba F.
- *R-E-G-W Q*. Procedimiento múltiple por pasos (por tamaño de las distancias) de Ryan-Einot-Gabriel-Welsch que se basa en el rango estudentizado.
- *S-N-K*. Realiza todas las comparaciones por parejas entre las medias utilizando la distribución del rango de Student. Con tamaños de muestras iguales, también compara pares de medias dentro de

subconjuntos homogéneos utilizando un procedimiento por pasos Las medias se ordenan de mayor a menor y se comparan primero las diferencias más extremas.

- *Tukey*. Utiliza el estadístico del rango estudentizado para realizar todas las comparaciones por pares entre los grupos. Establece la tasa de error por experimento como la tasa de error para el conjunto de todas las comparaciones por pares.
- *Tukey-b*. Prueba que emplea la distribución del rango estudentizado para realizar comparaciones por pares entre los grupos. El valor crítico es el promedio de los valores correspondientes a la diferencia honestamente significativa de Tukey y al método de Student-Newman-Keuls.
- *Duncan*. Realiza comparaciones por pares utilizando un orden por pasos idéntico al orden usado por la prueba de Student-Newman-Keuls, pero establece un nivel de protección en la tasa de error para la colección de contrastes, en lugar de usar una tasa de error para los contrastes individuales. Utiliza el estadístico del rango estudentizado.
- *GT2 de Hochberg*. Prueba de comparaciones múltiples y de rango que utiliza el módulo máximo estudentizado. Es similar a la prueba de la diferencia honestamente significativa de Tukey.
- *Gabriel*. Prueba de comparación por parejas que utiliza el módulo máximo estudentizado y que es generalmente más potente que la GT2 de Hochberg, si los tamaños de las casillas son desiguales. La prueba de Gabriel se puede convertir en liberal cuando los tamaños de las casillas varían mucho.
- *Waller-Duncan*. Prueba de comparaciones múltiples basada en un estadístico t. Utiliza la aproximación Bayesiana.
- *Dunnett*. Prueba de comparaciones múltiples por parejas que compara un conjunto de tratamientos respecto a una única media de control. La última categoría es la categoría de control predeterminada. Si lo desea, puede seleccionar la primera categoría. Para comprobar que la media de cualquier nivel del factor (excepto la categoría de control) no es igual a la de la categoría de control, utilice una prueba **bilateral**. Para contrastar si la media en cualquier nivel del factor es menor que la de la categoría de control, seleccione **<Control**. Para contrastar si la media en cualquier nivel del factor es mayor que la de la categoría de control, seleccione **>Control**.

No asumiendo varianzas iguales

Las pruebas de comparaciones múltiples que no suponen varianzas iguales son T2 de Tamhane, T3 de Dunnett, Games-Howell y *C* de Dunnett.

- *T2 de Tamhane*. Prueba conservadora de comparación por parejas basada en la prueba t. Esta prueba es adecuada cuando las varianzas son desiguales.
- *T3 de Dunnett*. Prueba de comparación por parejas basada en el módulo máximo estudentizado. Esta prueba es adecuada cuando las varianzas son desiguales.
- *Games-Howell*. Prueba de comparación por parejas que es en ocasiones liberal. Esta prueba es adecuada cuando las varianzas son desiguales.
- *C de Dunnett*. Prueba de comparación por parejas basada en el rango estudentizado. Esta prueba es adecuada cuando las varianzas son desiguales.

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

ANOVA de un factor: Opciones

Estadísticos. Elija uno o más entre los siguientes:

- **Descriptivos.** Calcula los siguientes estadísticos: Número de casos, Media, Desviación estándar, Error estándar de la media, Mínimo, Máximo y los Intervalos de confianza al 95% de cada variable dependiente para cada grupo.
- **Efectos fijos y aleatorios.** Muestra la desviación estándar, el error estándar y un intervalo de confianza del 95% para el modelo de efectos fijos, y el error estándar, un intervalo de confianza del 95% y una estimación de la varianza entre componentes para el modelo de efectos aleatorios.

- **Prueba de homogeneidad de las varianzas.** Calcula el estadístico de Levene para contrastar la igualdad de las varianzas de grupo. Esta prueba no depende del supuesto de normalidad.
- **Brown-Forsythe.** Calcula el estadístico de Brown-Forsythe para contrastar la igualdad de las medias de grupo. Este estadístico es preferible al estadístico *F* si no se supone la igualdad de las varianzas.
- **Welch.** Calcula el estadístico de Welch para contrastar la igualdad de las medias de grupo. Este estadístico es preferible al estadístico *F* si no se supone la igualdad de las varianzas.

Gráfico de las medias. Muestra un gráfico que representa las medias de los subgrupos (las medias para cada grupo definido por los valores de la variable factor).

Valores perdidos. Controla el tratamiento de los valores perdidos.

- **Excluir casos según análisis.** Un caso que tenga un valor perdido para la variable dependiente o la variable de factor en un análisis determinado, no se utiliza en ese análisis. Además, los casos fuera del rango especificado para la variable de factor no se utilizan.
- **Excluir casos según lista.** Se excluyen de todos los análisis los casos con valores perdidos para la variable de factor o para cualquier variable dependiente incluida en la lista de variables dependientes en el cuadro de diálogo principal. Si no se han especificado varias variables dependientes, esta opción no surte efecto.

Características adicionales del comando ONEWAY

La sintaxis de comandos también le permite:

- Obtener estadísticos de efectos fijos y aleatorios. Desviación estándar, error estándar de la media e intervalos de confianza al 95% para el modelo de efectos fijos. Error estándar, intervalos de confianza al 95% y estimación de la varianza entre componentes para el modelo de efectos aleatorios (mediante `STATISTICS=EFFECTS`).
- Especificar niveles alfa para la menor diferencia de significación, Bonferroni, pruebas de comparación múltiple de Duncan y Scheffé (con el subcomando `RANGES`).
- Escribir una matriz de medias, desviaciones estándar y frecuencias, o leer una matriz de medias, frecuencias, varianzas combinadas y grados de libertad para las varianzas combinadas. Estas matrices pueden utilizarse en lugar de los datos en bruto para obtener un análisis de varianza de un factor (con el subcomando `MATRIX`).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 11. MLG Análisis univariante

El procedimiento MLG Univariante proporciona un análisis de regresión y un análisis de varianza para una variable dependiente mediante uno o más factores o variables. Las variables de factor dividen la población en grupos. Con el procedimiento Modelo lineal general se pueden contrastar hipótesis nulas sobre los efectos de otras variables en las medias de varias agrupaciones de una única variable dependiente. Se pueden investigar las interacciones entre los factores así como los efectos de los factores individuales, algunos de los cuales pueden ser aleatorios. Además, se pueden incluir los efectos de las covariables y las interacciones de covariables con los factores. Para el análisis de regresión, las variables (predictoras) independientes se especifican como covariables.

Se pueden contrastar tanto los modelos equilibrados como los no equilibrados. Se considera que un diseño está equilibrado si cada casilla del modelo contiene el mismo número de casos. Además de contrastar hipótesis, MLG Univariante genera estimaciones de los parámetros.

También se encuentran disponibles los contrastes a priori de uso más habitual para contrastar las hipótesis. Además, si una prueba *F* global ha mostrado cierta significación, pueden emplearse las pruebas post hoc para evaluar las diferencias entre las medias específicas. Las medias marginales estimadas ofrecen estimaciones de valores de las medias pronosticados para las casillas del modelo; los gráficos de perfil (gráficos de interacciones) de estas medias permiten observar fácilmente algunas de estas relaciones.

En su archivo de datos puede guardar residuos, valores pronosticados, distancia de Cook y valores de influencia como variables nuevas para comprobar los supuestos.

Ponderación MCP permite especificar una variable usada para aplicar a las observaciones una ponderación diferente en un análisis de mínimos cuadrados ponderados (MCP), por ejemplo para compensar la distinta precisión de las mediciones.

Ejemplo. Se recogen datos de los corredores individuales en el maratón de Chicago durante varios años. El tiempo final de cada corredor es la variable dependiente. Influyen otros factores como el clima (frío, calor o temperatura agradable), los meses de entrenamiento, el número de maratones anteriores y el sexo. La edad se considera una covariable. Observará que el sexo es un efecto significativo y que la interacción del sexo con el clima es significativa.

Métodos. Las sumas de cuadrados de Tipo I, Tipo II, Tipo III y Tipo IV pueden emplearse para evaluar las diferentes hipótesis. Tipo III es el valor predeterminado.

Estadísticos. Las pruebas de rango post hoc y las comparaciones múltiples: Diferencia menos significativa (DMS), Bonferroni, Sidak, Scheffé, Múltiples *F* de Ryan-Einot-Gabriel-Welsch (R-E-G-W-F), Rango múltiple de Ryan-Einot-Gabriel-Welsch, Student-Newman-Keuls (S-N-K), Diferencia honestamente significativa de Tukey, *b* de Tukey, Duncan, GT2 de Hochberg, Gabriel, Pruebas *t* de Waller Duncan, Dunnett (unilateral y bilateral), T2 de Tamhane, T3 de Dunnett, Games-Howell y *C* de Dunnett. Estadísticos descriptivos: medias observadas, desviaciones estándar y frecuencias de todas las variables dependientes en todas las casillas. Prueba de Levene para la homogeneidad de varianzas.

Diagramas. Diagramas de dispersión por nivel, gráficos de residuos, gráficos de perfil (interacción).

MLG Univariante: Consideraciones sobre los datos

Datos. La variable dependiente es cuantitativa. Los factores son categóricos; pueden tener valores numéricos o valores de cadena de hasta ocho caracteres. Pueden tener valores numéricos o valores de cadena de hasta ocho caracteres. Las covariables son variables cuantitativas que están relacionadas con la variable dependiente.

Supuestos. Los datos son una muestra aleatoria de una población normal; en la población, todas las varianzas de las casillas son iguales. El análisis de varianza es robusto a las desviaciones de la normalidad, aunque los datos deberán ser simétricos. Para comprobar los supuestos, puede utilizar la prueba de homogeneidad de varianzas y los gráficos de dispersión por nivel. También puede examinar los residuos y los gráficos de residuos.

Para obtener un análisis MLG Univariante

1. Seleccione en los menús:

   Analizar > **Modelo lineal general** > **Univariante...**
2. Seleccione una variable dependiente.
3. Seleccione variables para Factores fijos, Factores aleatorios y Covariables, en función de los datos.
4. Si lo desea, puede utilizar la Ponderación MCP para especificar una variable de ponderación para el análisis de mínimos cuadrados ponderados. Si el valor de la variable de ponderación es cero, negativo o perdido, el caso queda excluido del análisis. Una variable que ya se haya utilizado en el modelo no puede usarse como variable de ponderación.

MLG: Modelo

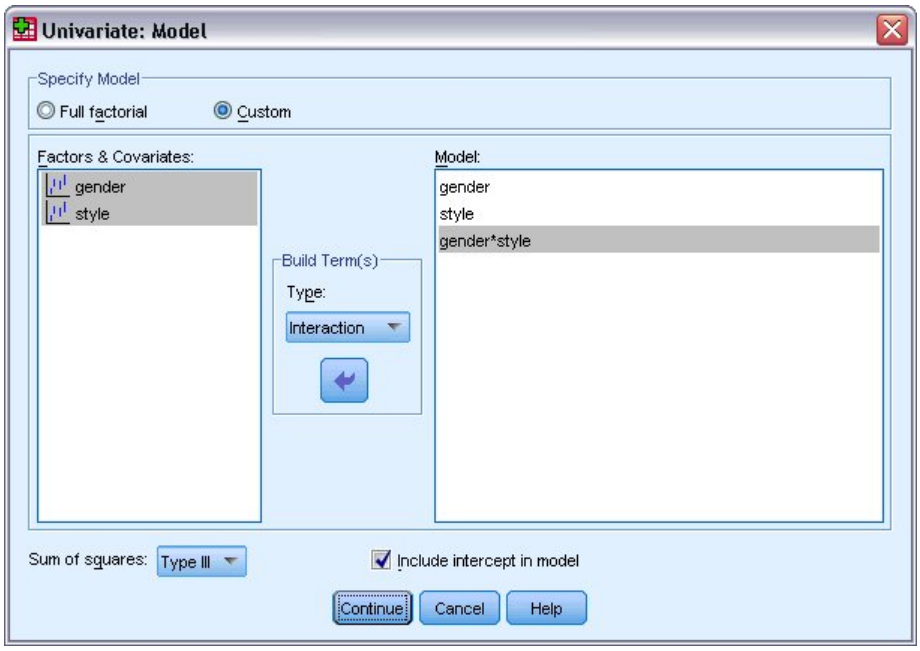


Figura 1. Cuadro de diálogo Univariante: Modelo

Especificar modelo. Un modelo factorial completo contiene todos los efectos principales del factor, todos los efectos principales de las covariables y todas las interacciones factor por factor. No contiene interacciones de covariable. Seleccione **Personalizado** para especificar sólo un subconjunto de interacciones o para especificar interacciones factor por covariable. Indique todos los términos que desee incluir en el modelo.

Factores y Covariables. Muestra una lista de los factores y las covariables.

Modelo. El modelo depende de la naturaleza de los datos. Después de seleccionar **Personalizado**, puede elegir los efectos principales y las interacciones que sean de interés para el análisis.

Suma de cuadrados Determina el método para calcular las sumas de cuadrados. Para los modelos equilibrados y no equilibrados sin casillas perdidas, el método de suma de cuadrados más utilizado es el de Tipo III.

Incluir la intersección en el modelo. La intersección se incluye normalmente en el modelo. Si supone que los datos pasan por el origen, puede excluir la intersección.

Generar términos

Para las covariables y los factores seleccionados:

Interacción. Crea el término de interacción de mayor nivel con todas las variables seleccionadas. Este es el método predeterminado.

Efectos principales. Crea un término de efectos principales para cada variable seleccionada.

Todas de 2. Crea todas las interacciones bidimensionales posibles de las variables seleccionadas.

Todas de 3. Crea todas las interacciones tridimensionales posibles de las variables seleccionadas.

Todas de 4. Crea todas las interacciones tetradimensionales posibles de las variables seleccionadas.

Todas de 5. Crea todas las interacciones quíntuples posibles de las variables seleccionadas.

Suma de cuadrados

Para el modelo, puede elegir un tipo de suma de cuadrados. El Tipo III es el más utilizado y es el tipo predeterminado.

Tipo I. Este método también se conoce como el método de descomposición jerárquica de la suma de cuadrados. Cada término se corrige sólo respecto al término que le precede en el modelo. El método Tipo I para la obtención de sumas de cuadrados se utiliza normalmente para:

- Un modelo ANOVA equilibrado en el que se especifica cualquier efecto principal antes de cualquier efecto de interacción de primer orden, cualquier efecto de interacción de primer orden se especifica antes de cualquier efecto de interacción de segundo orden, y así sucesivamente.
- Un modelo de regresión polinómica en el que se especifica cualquier término de orden inferior antes que cualquier término de orden superior.
- Un modelo puramente anidado en el que el primer efecto especificado está anidado dentro del segundo efecto especificado, el segundo efecto especificado está anidado dentro del tercero, y así sucesivamente. Esta forma de anidamiento solamente puede especificarse utilizando la sintaxis.

Tipo II. Este método calcula cada suma de cuadrados del modelo considerando sólo los efectos pertinentes. Un efecto pertinente es el que corresponde a todos los efectos que no contienen el que se está examinando. El método de suma de cuadrados de Tipo II se utiliza normalmente para:

- Un modelo ANOVA equilibrado.
- Cualquier modelo que sólo tenga efectos de factor principal.
- Cualquier modelo de regresión.
- Un diseño puramente anidado (esta forma de anidamiento solamente puede especificarse utilizando la sintaxis).

Tipo III. Es el método predeterminado. Este método calcula las sumas de cuadrados de un efecto de diseño como las sumas de cuadrados corregidas respecto a cualquier otro efecto que no lo contenga y

ortogonales a cualquier efecto (si existe alguno) que lo contenga. Las sumas de cuadrados de Tipo III tienen una gran ventaja por ser invariables respecto a las frecuencias de casilla, siempre que la forma general de estimabilidad permanezca constante. Así, este tipo de sumas de cuadrados se suele considerar de gran utilidad para un modelo no equilibrado sin casillas perdidas. En un diseño factorial sin casillas perdidas, este método equivale a la técnica de cuadrados ponderados de las medias de Yates. El método de suma de cuadrados de Tipo III se utiliza normalmente para:

- Cualquiera de los modelos que aparecen en los tipos I y II.
- Cualquier modelo equilibrado o desequilibrado sin casillas vacías.

Tipo IV. Este método está diseñado para una situación en la que hay casillas perdidas. Para cualquier efecto *F* en el diseño, si *F* no está contenida en cualquier otro efecto, entonces Tipo IV = Tipo III = Tipo II. Cuando *F* está contenida en otros efectos, el Tipo IV distribuye equitativamente los contrastes que se realizan entre los parámetros en *F* a todos los efectos de nivel superior. El método de suma de cuadrados de Tipo I se utiliza normalmente para:

- Cualquiera de los modelos que aparecen en los tipos I y II.
- Cualquier modelo equilibrado o no equilibrado con casillas vacías.

MLG: Contrastes

Los contrastes se utilizan para contrastar las diferencias entre los niveles de un factor. Puede especificar un contraste para cada factor en el modelo (en un modelo de medidas repetidas, para cada factor inter-sujetos). Los contrastes representan las combinaciones lineales de los parámetros.

MLG Univariante. El contraste de hipótesis se basa en la hipótesis nula **LB** = 0, donde **L** es la matriz de coeficientes de contraste y **B** es el vector de parámetros. Cuando se especifica un contraste, se crea una matriz **L**. Las columnas de la matriz **L** correspondientes al factor coinciden con el contraste. El resto de las columnas se corrigen para que la matriz **L** sea estimable.

Los resultados incluyen un estadístico *F* para cada conjunto de contrastes. Para el contraste de diferencias también se muestran los intervalos de confianza simultáneos de tipo Bonferroni basados en la distribución *t* de Student.

Contrastes disponibles

Los contrastes disponibles son de desviación, simples, de diferencias, de Helmert, repetidos y polinómicos. En los contrastes de desviación y los contrastes simples, es posible determinar que la categoría de referencia sea la primera o la última categoría.

Tipos de contrastes

Desviación. Compara la media de cada nivel (excepto una categoría de referencia) con la media de todos los niveles (media global). Los niveles del factor pueden colocarse en cualquier orden.

Simple. Compara la media de cada nivel con la media de un nivel especificado. Este tipo de contraste resulta útil cuando existe un grupo de control. Puede seleccionar la primera o la última categoría como referencia.

Diferencia. Compara la media de cada nivel (excepto el primero) con la media de los niveles anteriores (a veces también se denominan contrastes de Helmert inversos). (A veces también se denominan contrastes de Helmert inversos).

Helmert. Compara la media de cada nivel del factor (excepto el último) con la media de los niveles siguientes.

Repetidas. Compara la media de cada nivel (excepto el último) con la media del nivel siguiente.

Polinómico. Compara el efecto lineal, cuadrático, cúbico, etc. El primer grado de libertad contiene el efecto lineal a través de todas las categorías; el segundo grado de libertad, el efecto cuadrático, y así sucesivamente. Estos contrastes se utilizan a menudo para estimar las tendencias polinómicas.

MLG: Gráficos de perfil

Los gráficos de perfil (gráficos de interacción) sirven para comparar las medias marginales en el modelo. Un gráfico de perfil es un gráfico de líneas en el que cada punto indica la media marginal estimada de una variable dependiente (corregida respecto a las covariables) en un nivel de un factor. Los niveles de un segundo factor se pueden utilizar para generar líneas diferentes. Cada nivel en un tercer factor se puede utilizar para crear un gráfico diferente. Todos los factores fijos y aleatorios, si existen, están disponibles para los gráficos. Para los análisis multivariantes, los gráficos de perfil se crean para cada variable dependiente. En un análisis de medidas repetidas, es posible utilizar tanto los factores inter-sujetos como los intra-sujetos en los gráficos de perfil. Las opciones MLG - Multivariante y MLG - Medidas repetidas sólo estarán disponibles si tiene instalada la opción Estadísticas avanzadas.

Un gráfico de perfil de un factor muestra si las medias marginales estimadas aumentan o disminuyen a través de los niveles. Para dos o más factores, las líneas paralelas indican que no existe interacción entre los factores, lo que significa que puede investigar los niveles de un único factor. Las líneas no paralelas indican una interacción.

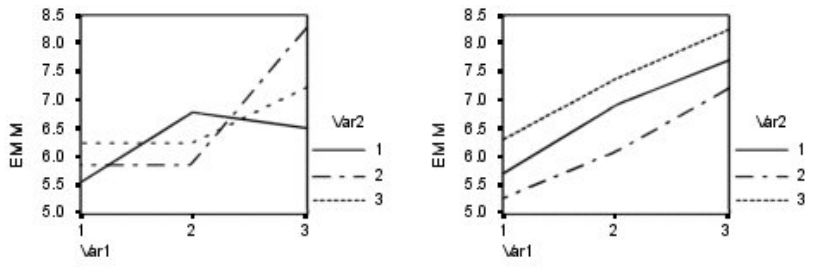


Figura 2. Gráfico no paralelo (izquierda) y gráfico paralelo (derecha)

Después de especificar un gráfico mediante la selección de los factores del eje horizontal y, de manera opcional, los factores para distintas líneas y gráficos, el gráfico deberá añadirse a la lista de gráficos.

Opciones MLG

Este cuadro de diálogo contiene estadísticos opcionales. Los estadísticos se calculan utilizando un modelo de efectos fijos.

Medias marginales estimadas. Seleccione los factores e interacciones para los que desee obtener estimaciones de las medias marginales de la población en las casillas. Estas medias se corrigen respecto a las covariables, si las hay.

- **Comparar los efectos principales.** Proporciona comparaciones por parejas no corregidas entre las medias marginales estimadas para cualquier efecto principal del modelo, tanto para los factores inter-sujetos como para los intra-sujetos. Este elemento sólo se encuentra disponible si los efectos principales están seleccionados en la lista Mostrar las medias para.
- **Ajuste del intervalo de confianza.** Seleccione un ajuste de diferencia menor significativa (DMS), Bonferroni o Sidak para los intervalos de confianza y la significación. Este elemento sólo estará disponible si se selecciona **Comparar los efectos principales**.

Representación. Seleccione **Estadísticos descriptivos** para generar medias observadas, desviaciones estándar y frecuencias para cada variable dependiente en todas las casillas. La opción **Estimaciones del tamaño del efecto** ofrece un valor parcial de eta-cuadrado para cada efecto y cada estimación de parámetros. El estadístico eta cuadrado describe la proporción de variabilidad total atribuible a un factor. Seleccione **Potencia observada** para obtener la potencia de la prueba cuando la hipótesis alternativa se ha

establecido basándose en el valor observado. Seleccione **Estimaciones de los parámetros** para generar las estimaciones de los parámetros, los errores estándar, las pruebas *t*, los intervalos de confianza y la potencia observada para cada prueba. Seleccione **Matriz de coeficientes de contraste** para obtener la matriz **L**.

Las **pruebas de homogeneidad** producen la prueba de homogeneidad de varianzas de Levene para cada variable dependiente en todas las combinaciones de nivel de los factores inter-sujetos sólo para factores inter-sujetos. Las opciones de diagramas de dispersión por nivel y gráfico de los residuos son útiles para comprobar los supuestos sobre los datos. Estos elementos no estarán activado si no hay factores. Seleccione **Gráficos de los residuos** para generar un gráfico de los residuos observados respecto a los pronosticados respecto a los tipificados para cada variable dependiente. Estos gráficos son útiles para investigar el supuesto de varianzas iguales. Seleccione **Falta de ajuste** para comprobar si el modelo puede describir de forma adecuada la relación entre la variable dependiente y las variables independientes. La **función estimable general** permite construir pruebas de hipótesis personales basadas en la función estimable general. Las filas en las matrices de coeficientes de contraste son combinaciones lineales de la función estimable general.

Nivel de significación. Puede que le interese corregir el nivel de significación usado en las pruebas post hoc y el nivel de confianza empleado para construir intervalos de confianza. El valor especificado también se utiliza para calcular la potencia observada para la prueba. Si especifica un nivel de significación, el cuadro de diálogo mostrará el nivel asociado de los intervalos de confianza.

Características adicionales del comando UNIANOVA

La sintaxis de comandos también le permite:

- Especificar efectos anidados en el diseño (utilizando el subcomando `DESIGN`).
- Especificar contrastes de los efectos respecto a una combinación lineal de efectos o un valor (utilizando el subcomando `TEST`).
- Especificar varios contrastes (utilizando el subcomando `CONTRAST`).
- Incluir los valores perdidos del usuario (utilizando el subcomando `MISSING`).
- Especificar criterios EPS (utilizando el subcomando `CRITERIA`).
- Construir una matriz **L**, **M** o **K** personalizada (utilizando los subcomandos `LMATRIX`, `MMATRIX` y `KMATRIX`).
- Para contrastes de desviación o simples, especifique una categoría de referencia intermedia (utilizando el subcomando `CONTRAST`).
- Especificar métricas para contrastes polinómicos (utilizando el subcomando `CONTRAST`).
- Especificar términos de error para comparaciones post hoc (utilizando el subcomando `POSTHOC`).
- Calcular medias marginales estimadas para cualquier factor o interacción de factores entre los factores de la lista de factores (utilizando el subcomando `EMMEANS`).
- Especificar nombres para variables temporales (utilizando el subcomando `SAVE`).
- Construir un archivo de datos de matriz de correlaciones (utilizando el subcomando `OUTFILE`).
- Construir un archivo de datos de matriz que contenga estadísticos de la tabla de ANOVA inter-sujetos (utilizando el subcomando `OUTFILE`).
- Guardar la matriz de diseño en un archivo de datos nuevo (utilizando el subcomando `OUTFILE`).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

MLG: Comparaciones post hoc

Pruebas de comparaciones múltiples post hoc Una vez que se ha determinado que existen diferencias entre las medias, las pruebas de rango post hoc y las comparaciones múltiples por parejas permiten determinar qué medias difieren. Las comparaciones se realizan sobre valores sin corregir. Estas pruebas se utilizan únicamente para factores inter-sujetos fijos. En MLG Medidas repetidas, estas pruebas no están

disponibles si no existen factores inter-sujetos y las pruebas de comparación múltiple post hoc se realizan para la media a través de los niveles de los factores intra-sujetos. Para MLG - Multivariante, las pruebas post hoc se realizan por separado para cada variable dependiente. Las opciones MLG - Multivariante y MLG - Medidas repetidas sólo estarán disponibles si tiene instalada la opción Estadísticas avanzadas.

Las pruebas de diferencia honestamente significativa de Tukey y de Bonferroni son pruebas de comparación múltiple muy utilizadas. La **prueba de Bonferroni**, basada en el estadístico *t* de Student, corrige el nivel de significación observado por el hecho de que se realizan comparaciones múltiples. La **prueba t de Sidak** también corrige el nivel de significación y da lugar a límites más estrechos que los de Bonferroni. La **prueba de diferencia honestamente significativa de Tukey** utiliza el estadístico del rango estudentizado para realizar todas las comparaciones por pares entre los grupos y establece la tasa de error por experimento como la tasa de error para el conjunto de todas las comparaciones por pares. Cuando se contrasta un gran número de pares de medias, la prueba de la diferencia honestamente significativa de Tukey es más potente que la prueba de Bonferroni. Para un número reducido de pares, Bonferroni es más potente.

GT2 de Hochberg es similar a la prueba de la diferencia honestamente significativa de Tukey, pero se utiliza el módulo máximo estudentizado. La prueba de Tukey suele ser más potente. La **prueba de comparación por parejas de Gabriel** también utiliza el módulo máximo estudentizado y es generalmente más potente que la GT2 de Hochberg cuando los tamaños de las casillas son desiguales. La prueba de Gabriel se puede convertir en liberal cuando los tamaños de las casillas varían mucho.

La **prueba t de comparación múltiple por parejas de Dunnett** compara un conjunto de tratamientos con una media de control simple. La última categoría es la categoría de control predeterminada. Si lo desea, puede seleccionar la primera categoría. Asimismo, puede elegir una prueba unilateral o bilateral. Para comprobar que la media de cualquier nivel del factor (excepto la categoría de control) no es igual a la de la categoría de control, utilice una prueba bilateral. Para contrastar si la media en cualquier nivel del factor es menor que la de la categoría de control, seleccione **< Control**. Asimismo, para contrastar si la media en cualquier nivel del factor es mayor que la de la categoría de control, seleccione **> Control**.

Ryan, Einot, Gabriel y Welsch (R-E-G-W) desarrollaron dos pruebas de rangos múltiples por pasos. Los procedimientos múltiples por pasos (por tamaño de las distancias) contrastan en primer lugar si todas las medias son iguales. Si no son iguales, se contrasta la igualdad en los subconjuntos de medias. **R-E-G-W F** se basa en una prueba *F* y **R-E-G-W Q** se basa en un rango estudentizado. Estas pruebas son más potentes que la prueba de rangos múltiples de Duncan y Student-Newman-Keuls (que también son procedimientos múltiples por pasos), pero no se recomiendan para tamaños de casillas desiguales.

Cuando las varianzas son desiguales, utilice **T2 de Tamhane** (prueba conservadora de comparación por parejas basada en una prueba *t*), **T3 de Dunnett** (prueba de comparación por parejas basada en el módulo máximo estudentizado), **prueba de comparación por parejas Games-Howell** (a veces liberal), o **C de Dunnett** (prueba de comparación por parejas basada en el rango estudentizado). Tenga en cuenta que estas pruebas no son válidas y no se realizarán si el modelo tiene múltiples factores.

La **prueba de rango múltiple de Duncan**, Student-Newman-Keuls (**S-N-K**) y **b de Tukey** son pruebas de rango que asignan rangos a medias de grupo y calculan un valor de rango. Estas pruebas no se utilizan con la misma frecuencia que las pruebas anteriormente mencionadas.

La **prueba t de Waller-Duncan** utiliza la aproximación bayesiana. Esta prueba de rango emplea la media armónica del tamaño de la muestra cuando los tamaños muestrales no son iguales.

El nivel de significación de la prueba de **Scheffé** está diseñado para permitir todas las combinaciones lineales posibles de las medias de grupo que se van a contrastar, no sólo las comparaciones por parejas disponibles en esta característica. El resultado es que la prueba de Scheffé es normalmente más conservadora que otras pruebas, lo que significa que se precisa una mayor diferencia entre las medias para la significación.

La prueba de comparación múltiple por parejas de la diferencia menos significativa (**DMS**) es equivalente a varias pruebas *t* individuales entre todos los pares de grupos. La desventaja de esta prueba es que no se realiza ningún intento de corregir el nivel de significación observado para realizar las comparaciones múltiples.

Pruebas mostradas. Se proporcionan comparaciones por parejas para DMS, Sidak, Bonferroni, Games-Howell, T2 y T3 de Tamhane, *C* de Dunnett y T3 de Dunnett. También se facilitan subconjuntos homogéneos para S-N-K, *b* de Tukey, Duncan, R-E-G-W *F*, R-E-G-W *Q* y Waller. La prueba de la diferencia honestamente significativa de Tukey, la GT2 de Hochberg, la prueba de Gabriel y la prueba de Scheffé son pruebas de comparaciones múltiples y pruebas de rango.

Opciones MLG

Este cuadro de diálogo contiene estadísticos opcionales. Los estadísticos se calculan utilizando un modelo de efectos fijos.

Medias marginales estimadas. Seleccione los factores e interacciones para los que desee obtener estimaciones de las medias marginales de la población en las casillas. Estas medias se corrigen respecto a las covariables, si las hay.

- **Comparar los efectos principales.** Proporciona comparaciones por parejas no corregidas entre las medias marginales estimadas para cualquier efecto principal del modelo, tanto para los factores inter-sujetos como para los intra-sujetos. Este elemento sólo se encuentra disponible si los efectos principales están seleccionados en la lista Mostrar las medias para.
- **Ajuste del intervalo de confianza.** Seleccione un ajuste de diferencia menor significativa (DMS), Bonferroni o Sidak para los intervalos de confianza y la significación. Este elemento sólo estará disponible si se selecciona **Comparar los efectos principales**.

Representación. Seleccione **Estadísticos descriptivos** para generar medias observadas, desviaciones estándar y frecuencias para cada variable dependiente en todas las casillas. La opción **Estimaciones del tamaño del efecto** ofrece un valor parcial de eta-cuadrado para cada efecto y cada estimación de parámetros. El estadístico eta cuadrado describe la proporción de variabilidad total atribuible a un factor. Seleccione **Potencia observada** para obtener la potencia de la prueba cuando la hipótesis alternativa se ha establecido basándose en el valor observado. Seleccione **Estimaciones de los parámetros** para generar las estimaciones de los parámetros, los errores estándar, las pruebas *t*, los intervalos de confianza y la potencia observada para cada prueba. Seleccione **Matriz de coeficientes de contraste** para obtener la matriz **L**.

Las **pruebas de homogeneidad** producen la prueba de homogeneidad de varianzas de Levene para cada variable dependiente en todas las combinaciones de nivel de los factores inter-sujetos sólo para factores inter-sujetos. Las opciones de diagramas de dispersión por nivel y gráfico de los residuos son útiles para comprobar los supuestos sobre los datos. Estos elementos no estarán activado si no hay factores. Seleccione **Gráficos de los residuos** para generar un gráfico de los residuos observados respecto a los pronosticados respecto a los tipificados para cada variable dependiente. Estos gráficos son útiles para investigar el supuesto de varianzas iguales. Seleccione **Falta de ajuste** para comprobar si el modelo puede describir de forma adecuada la relación entre la variable dependiente y las variables independientes. La **función estimable general** permite construir pruebas de hipótesis personales basadas en la función estimable general. Las filas en las matrices de coeficientes de contraste son combinaciones lineales de la función estimable general.

Nivel de significación. Puede que le interese corregir el nivel de significación usado en las pruebas post hoc y el nivel de confianza empleado para construir intervalos de confianza. El valor especificado también se utiliza para calcular la potencia observada para la prueba. Si especifica un nivel de significación, el cuadro de diálogo mostrará el nivel asociado de los intervalos de confianza.

Características adicionales del comando UNIANOVA

La sintaxis de comandos también le permite:

- Especificar efectos anidados en el diseño (utilizando el subcomando DESIGN).
- Especificar contrastes de los efectos respecto a una combinación lineal de efectos o un valor (utilizando el subcomando TEST).
- Especificar varios contrastes (utilizando el subcomando CONTRAST).
- Incluir los valores perdidos del usuario (utilizando el subcomando MISSING).
- Especificar criterios EPS (utilizando el subcomando CRITERIA).
- Construir una matriz **L**, **M** o **K** personalizada (utilizando los subcomandos LMATRIX, MMATRIX y KMATRIX).
- Para contrastes de desviación o simples, especifique una categoría de referencia intermedia (utilizando el subcomando CONTRAST).
- Especificar métricas para contrastes polinómicos (utilizando el subcomando CONTRAST).
- Especificar términos de error para comparaciones post hoc (utilizando el subcomando POSTHOC).
- Calcular medias marginales estimadas para cualquier factor o interacción de factores entre los factores de la lista de factores (utilizando el subcomando EMMEANS).
- Especificar nombres para variables temporales (utilizando el subcomando SAVE).
- Construir un archivo de datos de matriz de correlaciones (utilizando el subcomando OUTFILE).
- Construir un archivo de datos de matriz que contenga estadísticos de la tabla de ANOVA inter-sujetos (utilizando el subcomando OUTFILE).
- Guardar la matriz de diseño en un archivo de datos nuevo (utilizando el subcomando OUTFILE).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

MLG: Guardar

Es posible guardar los valores pronosticados por el modelo, los residuos y las medidas relacionadas como variables nuevas en el Editor de datos. Muchas de estas variables se pueden utilizar para examinar supuestos sobre los datos. Si desea almacenar los valores para utilizarlos en otra sesión de IBM SPSS Statistics, guárdelos en el archivo de datos actual.

Valores pronosticados. Son los valores que predice el modelo para cada caso.

- *No tipificados*. Valor predicho por el modelo para la variable dependiente.
- *Ponderados*. Los valores pronosticados no tipificados ponderados. Sólo están disponibles si se seleccionó previamente una variable de ponderación MCP.
- *Error estándar*. Estimación de la desviación estándar del valor promedio de la variable dependiente para los casos que tengan los mismos valores en las variables independientes.

Diagnósticos. Son medidas para identificar casos con combinaciones poco usuales de valores para los casos y las variables independientes que puedan tener un gran impacto en el modelo.

- *Distancia de Cook*. Una medida de cuánto cambiarían los residuos de todos los casos si un caso particular se excluyera del cálculo de los coeficientes de regresión. Una Distancia de Cook grande indica que la exclusión de ese caso del cálculo de los estadísticos de regresión hará variar substancialmente los coeficientes.
- *Valores de influencia*. Los valores de influencia no centrados. La influencia relativa de una observación en el ajuste del modelo.

Residuos. Un residuo no tipificado es el valor real de la variable dependiente menos el valor predicho por el modelo. También se encuentran disponibles residuos eliminados, estudentizados y tipificados. Si ha seleccionado una variable MCP, contará además con residuos no tipificados ponderados.

- *No tipificados*. Diferencia entre un valor observado y el valor predicho por el modelo.
- *Ponderados*. Los residuos no tipificados ponderados. Sólo están disponibles si se seleccionó previamente una variable de ponderación MCP.
- *Tipificados*. El residuo dividido por una estimación de su error estándar. Los residuos tipificados, que son conocidos también como los residuos de Pearson o residuos estandarizados, tienen una media de 0 y una desviación estándar de 1.
- *Estudentizados*. Residuo dividido por una estimación de su desviación estándar que varía de caso en caso, dependiendo de la distancia de los valores de cada caso en las variables independientes respecto a las medias en las variables independientes.
- *Eliminados*. Residuo para un caso cuando éste se excluye del cálculo de los coeficientes de la regresión. Es igual a la diferencia entre el valor de la variable dependiente y el valor predicho corregido.

Estadísticos de los coeficientes. Escribe una matriz varianza-covarianza de las estimaciones de los parámetros del modelo en un nuevo conjunto de datos de la sesión actual o un archivo de datos externo de IBM SPSS Statistics. Asimismo, para cada variable dependiente habrá una fila de estimaciones de los parámetros, una fila de valores de significación para los estadísticos *t* correspondientes a las estimaciones de los parámetros y una fila de grados de libertad de los residuos. En un modelo multivariante, existen filas similares para cada variable dependiente. Si lo desea, puede usar este archivo matricial en otros procedimientos que lean archivos matriciales.

Opciones MLG

Este cuadro de diálogo contiene estadísticos opcionales. Los estadísticos se calculan utilizando un modelo de efectos fijos.

Medias marginales estimadas. Seleccione los factores e interacciones para los que desee obtener estimaciones de las medias marginales de la población en las casillas. Estas medias se corrigen respecto a las covariables, si las hay.

- **Comparar los efectos principales.** Proporciona comparaciones por parejas no corregidas entre las medias marginales estimadas para cualquier efecto principal del modelo, tanto para los factores inter-sujetos como para los intra-sujetos. Este elemento sólo se encuentra disponible si los efectos principales están seleccionados en la lista Mostrar las medias para.
- **Ajuste del intervalo de confianza.** Seleccione un ajuste de diferencia menor significativa (DMS), Bonferroni o Sidak para los intervalos de confianza y la significación. Este elemento sólo estará disponible si se selecciona **Comparar los efectos principales**.

Representación. Seleccione **Estadísticos descriptivos** para generar medias observadas, desviaciones estándar y frecuencias para cada variable dependiente en todas las casillas. La opción **Estimaciones del tamaño del efecto** ofrece un valor parcial de eta-cuadrado para cada efecto y cada estimación de parámetros. El estadístico eta cuadrado describe la proporción de variabilidad total atribuible a un factor. Seleccione **Potencia observada** para obtener la potencia de la prueba cuando la hipótesis alternativa se ha establecido basándose en el valor observado. Seleccione **Estimaciones de los parámetros** para generar las estimaciones de los parámetros, los errores estándar, las pruebas *t*, los intervalos de confianza y la potencia observada para cada prueba. Seleccione **Matriz de coeficientes de contraste** para obtener la matriz **L**.

Las **pruebas de homogeneidad** producen la prueba de homogeneidad de varianzas de Levene para cada variable dependiente en todas las combinaciones de nivel de los factores inter-sujetos sólo para factores inter-sujetos. Las opciones de diagramas de dispersión por nivel y gráfico de los residuos son útiles para comprobar los supuestos sobre los datos. Estos elementos no estarán activado si no hay factores. Seleccione **Gráficos de los residuos** para generar un gráfico de los residuos observados respecto a los pronosticados respecto a los tipificados para cada variable dependiente. Estos gráficos son útiles para investigar el supuesto de varianzas iguales. Seleccione **Falta de ajuste** para comprobar si el modelo puede describir de forma adecuada la relación entre la variable dependiente y las variables independientes. La

función estimable general permite construir pruebas de hipótesis personales basadas en la función estimable general. Las filas en las matrices de coeficientes de contraste son combinaciones lineales de la función estimable general.

Nivel de significación. Puede que le interese corregir el nivel de significación usado en las pruebas post hoc y el nivel de confianza empleado para construir intervalos de confianza. El valor especificado también se utiliza para calcular la potencia observada para la prueba. Si especifica un nivel de significación, el cuadro de diálogo mostrará el nivel asociado de los intervalos de confianza.

Características adicionales del comando UNIANOVA

La sintaxis de comandos también le permite:

- Especificar efectos anidados en el diseño (utilizando el subcomando DESIGN).
- Especificar contrastes de los efectos respecto a una combinación lineal de efectos o un valor (utilizando el subcomando TEST).
- Especificar varios contrastes (utilizando el subcomando CONTRAST).
- Incluir los valores perdidos del usuario (utilizando el subcomando MISSING).
- Especificar criterios EPS (utilizando el subcomando CRITERIA).
- Construir una matriz **L**, **M** o **K** personalizada (utilizando los subcomandos LMATRIX, MMATRIX y KMATRIX).
- Para contrastes de desviación o simples, especifique una categoría de referencia intermedia (utilizando el subcomando CONTRAST).
- Especificar métricas para contrastes polinómicos (utilizando el subcomando CONTRAST).
- Especificar términos de error para comparaciones post hoc (utilizando el subcomando POSTHOC).
- Calcular medias marginales estimadas para cualquier factor o interacción de factores entre los factores de la lista de factores (utilizando el subcomando EMMEANS).
- Especificar nombres para variables temporales (utilizando el subcomando SAVE).
- Construir un archivo de datos de matriz de correlaciones (utilizando el subcomando OUTFILE).
- Construir un archivo de datos de matriz que contenga estadísticos de la tabla de ANOVA inter-sujetos (utilizando el subcomando OUTFILE).
- Guardar la matriz de diseño en un archivo de datos nuevo (utilizando el subcomando OUTFILE).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 12. Correlaciones bivariadas

El procedimiento Correlaciones bivariadas calcula el coeficiente de correlación de Pearson, la rho de Spearman y la tau-*b* de Kendall con sus niveles de significación. Las correlaciones miden cómo están relacionadas las variables o los órdenes de los rangos. Antes de calcular un coeficiente de correlación, inspeccione los datos para detectar valores atípicos (que pueden generar resultados equívocos) y evidencias de una relación lineal. El coeficiente de correlación de Pearson es una medida de asociación lineal. Dos variables pueden estar perfectamente relacionadas, pero si la relación no es lineal, el coeficiente de correlación de Pearson no será un estadístico adecuado para medir su asociación.

Ejemplo. ¿Está el número de partidos ganados por un equipo de baloncesto correlacionado con el número medio de puntos anotados por partido? Un diagrama de dispersión indica que existe una relación lineal. Al analizar los datos de la temporada 1994–1995 de la NBA, se descubre que el coeficiente de correlación de Pearson (0,581) es significativo al nivel 0,01. Se puede sospechar que cuantos más partidos se ganen por temporada, menos puntos habrán anotado los adversarios. Estas variables están correlacionadas negativamente (-0,401) y la correlación es significativa al nivel 0,05.

Estadísticos. Para cada variable: número de casos sin valores perdidos, media y desviación estándar. Para cada par de variables: coeficiente de correlación de Pearson, rho de Spearman, tau-*b* de Kendall, productos vectoriales de las desviaciones y covarianzas.

Correlaciones bivariadas: Consideraciones sobre los datos

Datos. Utilice variables cuantitativas simétricas para el coeficiente de correlación de Pearson y variables cuantitativas o variables con categorías ordenadas para la rho de Spearman y la tau-*b* de Kendall.

Supuestos. El coeficiente de correlación de Pearson asume que cada pareja de variables es normal bivariada.

Para obtener correlaciones bivariadas

Seleccione en los menús:

Analizar > **Correlaciones** > **Bivariadas...**

1. Seleccione dos o más variables numéricas.

También se encuentran disponibles las siguientes opciones:

- **Coeficientes de correlación.** Para las variables cuantitativas, normalmente distribuidas, seleccione el coeficiente de correlación de **Pearson**. Si los datos no están normalmente distribuidos o tienen categorías ordenadas, seleccione los correspondientes a la **Tau-b de Kendall** o **Spearman**, que miden la asociación entre órdenes de rangos. Los coeficientes de correlación pueden estar entre -1 (una relación negativa perfecta) y +1 (una relación positiva perfecta). Un valor 0 indica que no existe una relación lineal. Al interpretar los resultados, se debe evitar extraer conclusiones de causa-efecto a partir de una correlación significativa.
- **Prueba de significación.** Se pueden seleccionar las probabilidades bilaterales o las unilaterales. Si conoce de antemano la dirección de la asociación, seleccione **Unilateral**. Si no es así, seleccione **Bilateral**.
- **Señalar las correlaciones significativas.** Los coeficientes de correlación significativos al nivel 0,05 se identifican por medio de un solo asterisco y los significativos al nivel 0,01 se identifican con dos asteriscos.

Correlaciones bivariadas: Opciones

Estadísticos. Para las correlaciones de Pearson, se puede elegir una o ambas de estas opciones:

- **Medias y desviaciones estándar.** Se muestran para cada variable. También se muestra el número de casos con valores no perdidos. Los valores perdidos se consideran según cada variable individual, sin tener en cuenta la opción elegida para la manipulación de los valores perdidos.
- **Desviaciones de productos vectoriales y covarianzas.** Se muestran para cada pareja de variables. El producto vectorial de las desviaciones es igual a la suma de los productos de las variables corregidas respecto a la media. Éste es el numerador del coeficiente de correlación de Pearson. La covarianza es una medida no tipificada de la relación entre dos variables, igual a la desviación del producto vectorial dividido por *N*-1.

Valores perdidos. Puede elegir uno de los siguientes:

- **Excluir casos según pareja.** Se excluyen del análisis los casos con valores perdidos para una o ambas variables de la pareja que forma un coeficiente de correlación. Debido a que cada coeficiente está basado en todos los casos que tienen códigos válidos para esa pareja concreta de variables, en cada cálculo se utiliza la mayor cantidad de información disponible. Esto puede dar como resultado un grupo de coeficientes basados en un número de casos variable.
- **Excluir casos según lista.** Se excluyen de todas las correlaciones los casos con valores perdidos para cualquier variable.

Características adicionales de los comandos CORRELATIONS y NONPAR CORR

La sintaxis de comandos también le permite:

- Escribir una matriz de correlaciones para correlaciones de Pearson que pueda ser utilizada en lugar de los datos en bruto, con el fin de obtener otros análisis como el análisis factorial (con el subcomando `MATRIX`).
- Obtener correlaciones de todas las variables de una lista con todas las variables de una segunda lista (utilizando la palabra clave `WITH` en el subcomando `VARIABLES`).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 13. Correlaciones parciales

El procedimiento Correlaciones parciales calcula los coeficientes de correlación parcial, los cuales describen la relación lineal existente entre dos variables mientras se controlan los efectos de una o más variables adicionales. Las correlaciones son medidas de asociación lineal. Dos variables pueden estar perfectamente relacionadas, pero si la relación no es lineal, el coeficiente de correlación no es un estadístico adecuado para medir su asociación.

Ejemplo. ¿Existe alguna relación entre la financiación sanitaria y las tasas de enfermedad? Aunque cabe esperar que dicha relación sea negativa, un estudio describe una correlación *positiva* significativa: si la financiación sanitaria aumenta, las tasas de enfermedad parecen disminuir. Sin embargo, si se controla la tasa de visitas de visitadores médicos, se elimina prácticamente la correlación positiva observada. La financiación sanitaria y las tasas de enfermedad sólo parecen estar relacionadas positivamente debido a que más personas tienen acceso a la sanidad si la financiación aumenta, lo que tiene como resultado que los médicos y hospitales informen de más enfermedades.

Estadísticos. Para cada variable: número de casos sin valores perdidos, media y desviación estándar. Matrices de correlación de orden cero y parcial, con grados de libertad y niveles de significación.

Correlaciones parciales: Consideraciones sobre los datos

Datos. Utilice variables cuantitativas y simétricas.

Supuestos. El procedimiento Correlaciones parciales supone que cada par de variables es normal bivariante.

Para obtener correlaciones parciales

1. Seleccione en los menús:

   Analizar > **Correlaciones** > **Parcial...**
2. Seleccione dos o más variables numéricas para las que se van a calcular las correlaciones parciales.
3. Elija una o más variables numéricas de control.

También se encuentran disponibles las siguientes opciones:

- **Prueba de significación.** Se pueden seleccionar las probabilidades bilaterales o las unilaterales. Si conoce de antemano la dirección de la asociación, seleccione **Unilateral**. Si no es así, seleccione **Bilateral**.
- **Mostrar el nivel de significación real.** De forma predeterminada, se muestran la probabilidad y los grados de libertad para cada coeficiente de correlación. Si anula la selección de este elemento, los coeficientes significativos al nivel 0,05 se identifican con un asterisco, los coeficientes significativos al nivel 0,01 se identifican con un asterisco doble y se eliminan los grados de libertad. Este ajuste afecta a las matrices de correlación parcial y de orden cero.

Correlaciones parciales: Opciones

Estadísticos. Puede elegir una o ambas de las siguientes opciones:

- **Medias y desviaciones estándar.** Se muestran para cada variable. También se muestra el número de casos con valores no perdidos.
- **Correlaciones de orden cero.** Se muestra una matriz de las correlaciones simples entre todas las variables, incluyendo las variables de control.

Valores perdidos. Puede elegir una de las siguientes alternativas:

- **Excluir casos según lista.** Se excluyen de todos los cálculos los casos que presenten valores perdidos para cualquier variable, incluso si es para las variables de control.
- **Excluir casos según pareja.** Para el cálculo de las correlaciones de orden cero, en las que se basan las correlaciones parciales, no se utilizará un caso si tiene valores perdidos en una o ambas variables de un par. La eliminación según pareja aprovecha el máximo de los datos que sean posibles. Sin embargo, el número de casos puede variar de unos coeficientes a otros. Cuando se activa esta opción, los grados de libertad para un coeficiente parcial determinado se basan en el número menor de casos utilizado en el cálculo de cualquiera de las correlaciones de orden cero necesarias para el cálculo de dicho coeficiente parcial.

Características adicionales del comando PARTIAL CORR

La sintaxis de comandos también le permite:

- Leer una matriz de correlaciones de orden cero o escribir una matriz de correlaciones parciales (mediante el subcomando `MATRIX`).
- Obtener correlaciones parciales entre dos listas de variables (mediante la palabra clave `WITH` en el subcomando `VARIABLES`).
- Obtener análisis múltiples (mediante varios subcomandos `VARIABLES`).
- Especificar otros valores para solicitar (por ejemplo, las correlaciones parciales tanto de primer como de segundo orden) cuando tiene dos variables de control (mediante el subcomando `VARIABLES`).
- Suprimir coeficientes redundantes (mediante el subcomando `FORMAT`).
- Mostrar una matriz de correlaciones simples cuando algunos coeficientes no se pueden calcular (mediante el subcomando `STATISTICS`).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 14. Distancias

Este procedimiento calcula una variedad de estadísticos que miden las similitudes o diferencias (distancias), entre pares de variables o entre pares de casos. Estas medidas de similaridad o de distancia se pueden utilizar después con otros procedimientos, como análisis factorial, análisis de clústeres o escalamiento multidimensional, para ayudar en el análisis de conjuntos de datos complejos.

Ejemplo. ¿Es posible medir similaridades entre pares de automóviles en función de ciertas características, como tipo de motor, consumo y potencia? Al calcular las similitudes entre los coches, se puede obtener una noción de qué coches son similares entre sí y cuáles son diferentes. Para un análisis más formal, puede considerar la aplicación de un análisis jerárquico de clústeres o escalamiento multidimensional a las similitudes para explorar la estructura subyacente.

Estadísticos. Las medidas de diferencia (distancia) para datos de un intervalo son Distancia euclídea, Distancia euclídea al cuadrado, Chebychev, bloque, Minkowski o personalizada; para datos de recuento, medida de chi-cuadrado o phi-cuadrado; para datos binarios, Distancia euclídea, Distancia euclídea al cuadrado, diferencia de tamaño, diferencia de configuración, varianza, forma o Lance y Williams. Las medidas de similitud para datos de intervalos son correlación de Pearson o coseno; para datos binarios, Russel y Rao, concordancia simple, Jaccard, Dice, Rogers y Tanimoto, Sokal y Sneath 1, Sokal y Sneath 2, Sokal y Sneath 3, Kulczynski 1, Kulczynski 2, Sokal y Sneath 4, Hamann, Lambda, *D* de Anderberg, *Y* de Yule, *Q* de Yule, Ochiai, Sokal y Sneath 5, correlación Phi de 4 puntos o dispersión.

Para obtener matrices de distancias

1. Seleccione en los menús:

   Analizar > **Correlaciones** > **Distancias...**
2. Seleccione al menos una variable numérica para calcular distancias entre casos o seleccione al menos dos variables numéricas para calcular distancias entre variables.
3. Seleccione una alternativa en el grupo Calcular distancias para calcular proximidades entre casos o entre variables.

Distancias: Medidas de disimilaridad

En el grupo Medida, seleccione la alternativa que corresponda al tipo de datos (intervalo, recuento o binario); a continuación, de la lista desplegable, seleccione una las medidas que corresponda a dicho tipo de datos. Las medidas disponibles, por tipo de dato, son:

- **Datos de intervalo.** Distancia euclídea, Distancia euclídea al cuadrado, Chebychev, Bloque, Minkowski o Personalizada.
- **Datos de recuento.** Medida de chi-cuadrado o Medida de phi-cuadrado.
- **Datos binarios.** Distancia euclídea, Distancia euclídea al cuadrado, Diferencia de tamaño, Diferencia de configuración, Varianza, Forma o Lance y Williams. (Introduzca valores para Presente y Ausente para especificar cuáles son los dos valores representativos; las Distancias ignorarán todos los demás valores.)

El grupo Transformar valores permite estandarizar los valores de los datos para casos o variables *antes* de calcular proximidades. Estas transformaciones no se pueden aplicar a los datos binarios. Los métodos disponibles de estandarización son: Puntuaciones *z*, Rango -1 a 1, Rango 0 a 1, Magnitud máxima de 1, Media de 1 o Desviación estándar 1.

El grupo Transformar medidas permite transformar los valores generados por la medida de distancia. Se aplican después de calcular la medida de distancia. Las opciones disponibles son: Valores absolutos, Cambiar el signo y Cambiar la escala al rango 0–1.

Distancias: Medidas de similaridad

En el grupo Medida, seleccione la alternativa que corresponda al tipo de datos (intervalo o binario); a continuación, de la lista desplegable, seleccione una las medidas que corresponda a dicho tipo de datos. Las medidas disponibles, por tipo de dato, son:

- **Datos de intervalo.** Correlación de Pearson o Coseno.
- **Datos binarios.** Russel y Rao, Concordancia simple, Jaccard, Dice, Rogers y Tanimoto, Sokal y Sneath 1, Sokal y Sneath 2, Sokal y Sneath 3, Kulczynski 1, Kulczynski 2, Sokal y Sneath 4, Hamann, Lambda, *D* de Anderberg, *Y* de Yule, *Q* de Yule, Ochiai, Sokal y Sneath 5, Correlación Phi de 4 puntos o Dispersión. (Introduzca valores para Presente y Ausente para especificar cuáles son los dos valores representativos; las Distancias ignorarán todos los demás valores.)

El grupo Transformar valores permite estandarizar los valores de los datos para casos o variables antes de calcular proximidades. Estas transformaciones no se pueden aplicar a los datos binarios. Los métodos disponibles de estandarización son: Puntuaciones *z*, Rango -1 a 1, Rango 0 a 1, Magnitud máxima de 1, Media de 1 y Desviación estándar 1.

El grupo Transformar medidas permite transformar los valores generados por la medida de distancia. Se aplican después de calcular la medida de distancia. Las opciones disponibles son: Valores absolutos, Cambiar el signo y Cambiar la escala al rango 0–1.

Características adicionales del comando PROXIMITIES

El procedimiento Distancias utiliza la sintaxis de comandos `PROXIMITIES`. La sintaxis de comandos también le permite:

- Especificar cualquier número entero como la potencia para la medida de distancia de Minkowski.
- Especificar cualquier número entero como la potencia y la raíz para una medida de distancia personalizada.

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 15. Modelos lineales

Los modelos lineales predicen un objetivo continuo basándose en relaciones lineales entre el objetivo y uno o más predictores.

Los modelos lineales son relativamente simples y proporcionan una fórmula matemática fácil de interpretar para la puntuación. Las propiedades de estos modelos se comprenden bien y se pueden crear rápidamente en comparación con el resto de tipos de modelos (como redes neuronales o árboles de decisión) en el mismo conjunto de datos.

Ejemplo. Una correduría de seguros con recursos limitados para investigar las reclamaciones de seguros de los asegurados desea crear un modelo para estimar los costes de las reclamaciones. Al desplegar este modelo a los centros de servicio, los representantes pueden introducir información de la reclamación mientras están al teléfono con un cliente y obtener inmediatamente el coste "esperado" de la reclamación en función de datos de archivo.

Requisitos de campo. Debe haber un objetivo y al menos una entrada. De forma predeterminada, los campos con las funciones predefinidas de Ambos o Ninguno no se utilizan. El objetivo debe ser continuo (escala). No hay ninguna restricción de nivel de medición en los predictores (entradas); los campos categóricos (nominal y ordinal) se utilizan como factores en el modelo y los campos continuos se utilizan como covariables.

Nota: Si un campo categórico tiene más de 1000 categorías, el procedimiento no se ejecuta y no se genera ningún modelo.

Para obtener un modelo lineal

Esta característica requiere la opción Statistics Base.

Seleccione en los menús:

Analizar > **Regresión** > **Modelos lineales automáticos...**

1. Asegúrese de que hay al menos un destino y una entrada.
2. Pulse en **Opciones de generación** para especificar cualquier configuración de generación y modelado.
3. Pulse en **Opciones de modelo** para guardar puntuaciones en el conjunto de datos activo y exportar el modelo en un archivo externo.
4. Pulse en **Ejecutar** para ejecutar el procedimiento y crear los objetos Modelo.

Objetivos

¿Cuál es su objetivo principal? Seleccione el objetivo adecuado.

- **Crear un modelo estándar.** El método crea un único modelo para pronosticar el objetivo utilizando los predictores. Por lo general, los modelos estándar son más fáciles de interpretar y pueden ser más rápidos de puntuar que conjuntos de datos potenciados, empaquetados o grandes.
- **Mejorar la precisión de modelos (boosting).** El método crea un modelo de conjunto utilizando potenciación, que genera una secuencia de modelos para obtener predicciones más precisas. Los conjuntos pueden tardar más en generarse y puntuarse que un modelo estándar.

  La potenciación produce una sucesión de "modelos de componente", cada uno creado con el conjunto de datos al completo. Antes de crear cada modelo de componente, los archivos se ponderan basándose en los residuos de los anteriores modelos de componente. Los casos con muchos residuos reciben ponderaciones de análisis relativamente mayores para que el próximo modelo de componente se centre

en predecir bien estos archivos. Juntos, estos modelos de componente forman un modelo de conjunto. El modelo de conjunto puntúa nuevos resultados usando una regla de combinación; las reglas disponibles dependen del nivel de medición del objetivo.

- **Mejorar la estabilidad de modelos (bagging).** El método crea un modelo de conjunto utilizando bagging (agregación de simulación de muestreo), que genera modelos múltiples para obtener predicciones más fiables. Los conjuntos pueden tardar más en generarse y puntuarse que un modelo estándar.

  El la agregación de simulación de muestreo (bagging) produce réplicas del conjunto de datos de entrenamiento haciendo muestras con sustitución del conjunto de datos original. Así crea muestras de simulación de muestreo del mismo tamaño que el conjunto de datos original. Después se crea un "modelo de componente" en cada réplica. Juntos, estos modelos de componente forman un modelo de conjunto. El modelo de conjunto puntúa nuevos resultados usando una regla de combinación; las reglas disponibles dependen del nivel de medición del objetivo.
- **Crear un modelo para conjuntos de datos muy grandes (requiere IBM SPSS Statistics Server).** El método crea un modelo de conjunto dividiendo el conjunto de datos en bloques de datos separados. Seleccione esta opción si su conjunto de datos es demasiado grande para generar cualquiera de los modelos anteriores o para la generación incremental de modelos. Esta opción emplea menos tiempo en su generación, pero puede tardar más en puntuarse que un modelo estándar. Esta opción requiere conectividad con IBM SPSS Statistics Server.

Consulte “Conjuntos” en la página 64 para ver la información de configuración relacionada con boosting, bagging y conjuntos de datos de gran tamaño.

Conceptos básicos

Preparar automáticamente datos. Esta opción permite el procedimiento de transformar de forma interna el destino y predictores para maximizar el poder predictivo del modelo; las transformaciones se guardan con el modelo y se aplican a los nuevos datos para su puntuación. Las versiones originales de los campos transformados se excluyen del modelo. De forma predeterminada, se realiza la siguiente preparación automática de datos.

- **Fecha y hora.** Cada predictor de fecha se transforma en un nuevo predictor continuo que contiene el tiempo transcurrido desde una fecha de referencia (01-01-1970). Cada predictor de hora se transforma en un nuevo predictor continuo que contiene el tiempo transcurrido desde una hora de referencia (00:00:00).
- **Ajustar nivel de medición.** Los predictores continuos con menos de 5 valores distintos se reestructuran como predictores ordinales. Los predictores ordinales con más de 10 valores distintos se reestructuran como predictores continuos.
- **Tratamiento de valores atípicos.** Los valores de los predictores continuos que recaen más allá de un valor de corte (3 desviaciones estándar de la media) se establecen con el valor de corte.
- **Manejo de valores perdidos.** Los valores perdidos de los predictores nominales se sustituyen por el modo de la partición de entrenamiento. Los valores perdidos de los predictores ordinales se sustituyen por la mediana de la partición de entrenamiento. Los valores perdidos de los predictores continuos se sustituyen por la media de la partición de entrenamiento.
- **Fusión supervisada.** Hace un modelo más parsimonioso reduciendo el número de campos que deben procesarse junto con el destino. Las categorías similares se identifican en función de la relación entre la entrada y destino. Las categorías que no son significativamente diferentes (es decir, que tienen un valor p superior al valor 0,1) se fusionan. Tenga en cuenta que si todas las categorías se combinan en una, las versiones original y derivada del campo se excluyen del modelo porque no tienen ningún valor como predictor.

Nivel de confianza. Éste es el nivel de confianza que se utiliza para calcular las estimaciones de intervalos de los coeficientes de modelos en la vista Coeficientes. Especifique un valor mayor que 0 y menor que 100. El valor predeterminado es 95.

Selección de modelos

Método de selección de modelos. Seleccione uno de los métodos de selección de modelos (a continuación se encuentran los detalles) o **Incluya todos los predictores**, que simplemente introduce todos los predictores disponibles como términos del modelo de efectos principales. De forma predeterminada, se utiliza **Pasos sucesivos hacia adelante**.

Selección de Pasos sucesivos hacia adelante. Comienza sin efectos en el modelo y añade y elimina efectos paso por paso hasta que ya no se puedan añadir o eliminar según los criterios de los pasos sucesivos.

- **Criterios para entrada/eliminación.** Éste es el estadístico utilizado para determinar si debe añadirse o eliminarse un efecto del modelo. **Criterio de información (AICC)** se basa en la similitud del conjunto de entrenamiento que se le da al modelo, y se ajusta para penalizar modelos excesivamente complejos. **Estadísticos de F** se utiliza en una prueba estadística de la mejora en el error de modelo. **R cuadrado corregida** se basa en el ajuste del conjunto de entrenamiento, y se ajusta para penalizar modelos excesivamente complejos. **Criterio de prevención sobreajustado (ASE)** se basa en el ajuste (error cuadrado medio o ASE) del conjunto de prevención sobreajustado. El conjunto de prevención sobreajustado es una submuestra aleatoria de aproximadamente el 30% del conjunto de datos original que no se utiliza para enseñar el modelo.

  Si se selecciona otro criterio que no sea **Estadísticos de F**, se añadirá al modelo cada paso del efecto que se corresponda con el aumento positivo mayor en el criterio. Se eliminará cualquier efecto en el modelo que se corresponda con una disminución en el criterio.

  Si se selecciona **Estadísticos de F** como criterio, cada paso en el efecto que tenga el valor *p* más pequeño inferior al umbral especificado, se añadirá **Incluir efectos con valores p inferiores a** al modelo. El valor predeterminado es 0.05. Cualquier efecto en el modelo con un valor *p* superior al umbral especificado, **Eliminar efectos con valores p mayores que**, será eliminado. El valor predeterminado es 0.10.
- **Personalizar número máximo de efectos en el modelo final.** De forma predeterminada, pueden introducirse todos los efectos disponibles en el modelo. Del mismo modo, si el algoritmo por pasos sucesivos termina con un paso con el número máximo de efectos especificado, el algoritmo se detiene con el conjunto actual de efectos.
- **Personalizar número máximo de pasos.** El algoritmo por pasos sucesivos termina tras un cierto número de pasos. De forma predeterminada, es 3 veces el número de efectos disponibles. Del mismo modo, especifique un entero positivo para el número máximo de pasos.

Selección de mejores subconjuntos. Comprueba "todos los modelos posibles", o al menos el subconjunto más grande de los modelos posibles que los pasos sucesivos hacia adelante, para seleccionar el mejor según el criterio de mejores subconjuntos. **Criterio de información (AICC)** se basa en la similitud del conjunto de entrenamiento que se le da al modelo, y se ajusta para penalizar modelos excesivamente complejos. **R cuadrado corregida** se basa en el ajuste del conjunto de entrenamiento, y se ajusta para penalizar modelos excesivamente complejos. **Criterio de prevención sobreajustado (ASE)** se basa en el ajuste (error cuadrado medio o ASE) del conjunto de prevención sobreajustado. El conjunto de prevención sobreajustado es una submuestra aleatoria de aproximadamente el 30% del conjunto de datos original que no se utiliza para enseñar el modelo.

Se selecciona el modelo con el valor mayor del criterio como el mejor modelo.

Nota: La selección de mejores subconjuntos requiere más trabajo computacional que la selección por pasos sucesivos hacia adelante. Cuando los mejores subconjuntos se procesan junto con boosting, bagging y conjuntos de datos de gran tamaño, la generación de un modelo estándar generado mediante una selección por pasos sucesivos hacia delante puede tardar considerablemente más tiempo.

Conjuntos

Estos ajustes determinan el comportamiento de la agrupación que se produce cuando los conjuntos de datos de gran tamaño o de boosting o bagging son obligatorios en Objetivos. Las opciones no aplicables al objetivo seleccionado se ignorarán.

Bagging y conjuntos de datos muy grandes. Al puntuar un conjunto, ésta es la regla utilizada para combinar los valores pronosticados a partir de los modelos básicos para calcular el valor de puntuación del conjunto.

- **Regla de combinación predeterminada para objetivos continuos.** Los valores pronosticados de conjunto para objetivos continuos pueden combinarse mediante la media o mediana de los valores pronosticados a partir de los modelos básicos.

Tenga en cuenta que cuando el objetivo es mejorar la precisión del modelo, se ignoran las selecciones de reglas de combinación. El boosting siempre utiliza un voto de mayoría ponderada para puntuar objetivos categóricos y una mediana ponderada para puntuar objetivos continuos.

Boosting y bagging. Especifique el número de modelos básicos que debe generarse cuando el objetivo es mejorar la precisión o estabilidad del modelo; en el caso del bagging, se trata del número de muestras de simulación de muestreo. Debe ser un número entero positivo.

Avanzado

Replicar resultados. Al establecer una semilla aleatoria podrá replicar análisis. El generador de números aleatorios se utiliza para seleccionar qué registros están en el conjunto de prevención sobreajustado. Especifique un entero o pulse en **Generar**, lo que creará un entero pseudo-aleatorio entre 1 y 2147483647, ambos inclusive. El valor predeterminado es 54752075.

Opciones de modelos

Guardar valores predichos en el conjunto de datos. El nombre de variable predeterminado es *PredictedValue*.

Exportar modelo. Escribe el modelo en un archivo `.zip` externo. Puede utilizar este archivo de modelo para aplicar la información del modelo a otros archivos de datos para puntuarlo. Especifique un nombre de archivo exclusivo válido. Si la especificación de archivo hace referencia a un archivo existente, se sobrescribirá el archivo.

Resumen del modelo

La vista Resumen de modelos es una instantánea, un resumen de un vistazo del modelo y su ajuste.

Tabla. La tabla identifica algunos ajustes de modelo de alto nivel, incluyendo:

- Nombre del destino especificado en la pestaña Campos,
- Si la preparación automática de los datos se realizó tal como se especificaba en los ajustes de Procedimientos básicos,
- El criterio de selección y el método de selección de modelo especificado en los ajustes de Selección de modelo. También se muestra el valor del criterio de selección del modelo final, y viene presentado en un formato cuanto más pequeño mejor.

Gráfico. El gráfico muestra la precisión del modelo final, que se presenta en el formato mayor es mejor. El valor es $100 \times R^2$ ajustado para el modelo final.

Preparación automática de datos

Esta vista muestra información a cerca de qué campos se excluyen y cómo los campos transformados se derivaron en el paso de preparación automática de datos (ADP). Para cada campo que fue transformado o excluido, la tabla enumera el nombre del campo, su papel en el análisis y la acción tomada por el paso ADP. Los campos se clasifican por orden alfabético ascendente de nombres de campo. Las acciones que puede realizarse para cada campo incluyen:

- **Derivar duración: meses** calcula el tiempo transcurrido en meses a partir de los valores de un campo que contiene las fechas hasta la fecha actual del sistema.
- **Derivar duración: horas** calcula el tiempo transcurrido en horas a partir de los valores de un campo que contiene las horas hasta la hora actual del sistema.
- **Cambiar el nivel de medición de continuo a ordinal** reestructura los campos continuos que tienen menos de 5 valores exclusivos como campos ordinales.
- **Cambiar el nivel de medición de ordinal a continuo** reestructura los campos ordinales que tienen menos de 10 valores exclusivos como campos continuos.
- **Recortar valores atípicos** define los valores de los predictores continuos que recaen más allá de un valor de corte (3 desviaciones estándar de la media) con el valor de corte.
- **Reemplazar los valores perdidos** sustituye los valores perdidos en los campos nominales con el modo, en los campos ordinales con la mediana y en los campos continuos con la media.
- **Fundir categorías para maximizar la asociación con el destino** identifica las categorías de predictor "similares" en función de la relación entre la entrada y el destino. Las categorías que no son significativamente diferentes (es decir, que tienen un valor *p* superior al valor 0,05) se fusionan.
- **Excluir predictor constante / tras tratamiento de valores atípicos / tras fundir categorías** elimina los predictores que tienen un único valor, posiblemente después de que se hayan realizado otras acciones de ADP.

Importancia de predictor

Normalmente, desea centrar sus esfuerzos de modelado en los campos del predictor que importan más y considera eliminar o ignorar las que importan menos. El predictor de importancia de la variable le ayuda a hacerlo indicando la importancia relativa de cada predictor en la estimación del modelo. Como los valores son relativos, la suma de los valores de todos los predictor de la visualización es 1.0. La importancia del predictor no está relacionada con la precisión del modelo. Sólo está relacionada con la importancia de cada predictor para realizar una predicción, independientemente de si ésta es precisa o no.

Predicho por observado

Muestra un diagrama de dispersión agrupado de los valores predichos en el eje vertical por los valores observados en el eje horizontal. En teoría, los puntos deberían encontrarse en una línea de 45 grados; esta vista puede indicar si hay registros para los que el modelo realiza una predicción particularmente malo.

Residuos

Muestra un gráfico de diagnosis de los residuos del modelo.

Estilos de gráfico. Existen varios estilos de visualización diferentes, que son accesibles desde la lista desplegable **Estilo**.

- **Histograma.** Se trata de un histograma en intervalos de los residuos estudentizados de una superposición de la distribución normal. Los modelos lineales asumen que los residuos tienen una distribución normal, de forma que el histograma debería estar muy cercano a la línea continua.
- **Gráfico p-p.** Se trata de un gráfico probabilidad-probabilidad en intervalos que compara los residuos estudentizados con una distribución normal. Si la curva de los puntos representados es menos

pronunciada que la línea normal, los residuos muestran una variabilidad mayor que una distribución normal; si la curva es más pronunciada, los residuos muestran una variabilidad inferior que una distribución normal. Si los puntos representados tienen una curva con forma en S, la distribución de los residuos es asimétrica.

Valores atípicos

Esta tabla enumera los registros que ejercen una influencia excesiva sobre el modelo, y muestra el ID de registro (si se especifica en la pestaña Campos), el valor objetivo y la distancia de Cook. La distancia de Cook es una medida de cuánto cambiarían los residuos de todos los registros si un registro en particular se excluyera del cálculo de los coeficientes del modelo. Una distancia de Cook grande indica que la exclusión de un registro cambia sustancialmente los coeficientes, y por lo tanto debe considerarse relevante.

Los registros relevantes deben examinarse cuidadosamente para determinar si puede darles menos importancia en la estimación del modelo, truncar los valores atípicos a algún umbral aceptable o eliminar los registros relevantes completamente.

Efectos

Esta vista muestra el tamaño de cada efecto en el modelo.

Estilos. Existen varios estilos de visualización diferentes, que son accesibles desde la lista desplegable **Estilo**.

- **Diagrama.** Es un gráfico en el que los efectos se clasifican desde arriba hacia abajo con una importancia de predictores descendente. Las líneas de conexión del diagrama se ponderan tomando como base la significación del efecto, con un grosor de línea mayor co rrespondiente a efectos con mayor significación (valores *p* inferiores). Al pasar el ratón sobre una línea de conexión se muestra una ayuda contextual que muestra el valor *p* y la importancia del efecto. Este es el valor predeterminado.
- **Tabla.** Se trata de una tabla ANOVA para el modelo completo y los efectos de modelo individuales. Los efectos individuales se clasifican desde arriba hacia abajo con una importancia de predictores descendente. Tenga en cuenta que, de forma predeterminada, la tabla se contrae para mostrar únicamente los resultados del modelo general. Para ver los resultados de los efectos de modelo individuales, pulse en la casilla **Modelo corregido** de la tabla.

Importancia del predictor. Existe un control deslizante Importancia del predictor que controla qué predictores se muestran en la vista. Esto no cambia el modelo, simplemente le permite centrarse en los predictores más importantes. De forma predeterminada, se muestran los 10 efectos más importantes.

Significación. Existe un control deslizante Significación que controla aún más qué efectos se muestran en la vista, a parte de los que se muestran tomando como base la importancia de predictor. Se ocultan los efectos con valores de significación superiores al valor del control deslizante. Esto no cambia el modelo, simplemente le permite centrarse en los efectos más importantes. El valor predeterminado es 1.00, de modo que no se filtran efectos tomando como base la significación.

Coeficientes

Esta vista muestra el valor de cada coeficiente en el modelo. Tenga en cuenta que los factores (predictores categóricos) tienen codificación de indicador dentro del modelo, de modo que los **efectos** que contienen los factores generalmente tendrán múltiples **coeficientes** asociados: uno por cada categoría exceptuando la categoría que corresponde al parámetro (referencia) redundante.

Estilos. Existen varios estilos de visualización diferentes, que son accesibles desde la lista desplegable **Estilo**.

- **Diagrama.** Es un gráfico que muestra la intersección primero, y luego clasifica los efectos desde arriba hacia abajo con una importancia de predictores descendente. Dentro de los efectos que contienen factores, los coeficientes se clasifican en orden ascendente de valores de datos. Las líneas de conexión del diagrama se colorean tomando como base el signo del coeficiente (consulte la descripción del diagrama) y se ponderan en función de la significación del coeficiente, con un grosor de línea mayor correspondiente a coeficientes con mayor significación (valores *p* inferiores). Al pasar el ratón sobre una línea de conexión se muestra una ayuda contextual que muestra el valor del coeficiente, su valor *p* y la importancia del efecto con el que está asociado el parámetro. Este es el estilo predeterminado.
- **Tabla.** Muestra los valores, las pruebas de significación y los intervalos de confianza para los coeficientes de modelos individuales. Tras la intersección, los efectos se clasifican desde arriba hacia abajo con una importancia de predictores descendente. Dentro de los efectos que contienen factores, los coeficientes se clasifican en orden ascendente de valores de datos. Tenga en cuenta que, de forma predeterminada, la tabla se contrae para mostrar únicamente los resultados de coeficiente, significación e importancia de cada parámetro de modelo. Para ver el error estándar, el estadístico *t* y el intervalo de confianza, pulse en la casilla **Coeficiente** de la tabla. Si pasa el ratón sobre el nombre de un parámetro de modelo en la tabla aparece una ayuda contextual que muestra el nombre del parámetro, el efecto con el que está asociado y (para los predictores categóricos) las etiquetas de valor asociadas con el parámetro de modelo. Esto puede ser especialmente útil para ver las nuevas categorías creadas cuando la preparación de datos automática fusiona categorías similares de un predictor categórico.

Importancia del predictor. Existe un control deslizante Importancia del predictor que controla qué predictores se muestran en la vista. Esto no cambia el modelo, simplemente le permite centrarse en los predictores más importantes. De forma predeterminada, se muestran los 10 efectos más importantes.

Significación. Existe un control deslizante Significación que controla aún más qué coeficientes se muestran en la vista, a parte de los que se muestran tomando como base la importancia de predictor. Se ocultan los coeficientes con valores de significación superiores al valor del control deslizante. Esto no cambia el modelo, simplemente le permite centrarse en los coeficientes más importantes. El valor predeterminado es 1.00, de modo que no se filtran coeficientes tomando como base la significación.

Medias estimadas

Son gráficos representados para predictores significativos. El gráfico muestra el valor estimado de modelo del objetivo en el eje vertical de cada valor del predictor en el eje horizontal, que alberga el resto de los predictores constantes. Proporciona una visualización útil de los efectos de los coeficientes de cada predictor en el objetivo.

Nota: si no hay predictores significativos, no se generan medias estimadas.

Resumen de generación de modelos

Cuando se selecciona un algoritmo de selección de modelos que no sea **Ninguno**, proporciona algunos detalles del proceso de generación de modelos.

Pasos sucesivos hacia adelante. Cuando la selección por pasos hacia adelante es el algoritmo de selección, la tabla muestra los últimos 10 pasos en el algoritmo de selección por pasos hacia adelante. Para cada paso, se muestran el valor del criterio de selección y los efectos en el modelo en ese paso. Esto ofrece el sentido del grado de contribución de cada paso al modelo. Cada columna le permite clasificar las filas, de modo que es posible ver con mayor facilidad qué efectos hay en un paso en particular.

Mejores subconjuntos. Cuando Mejores subconjuntos es el algoritmo de selección, la tabla muestra los 10 modelos principales. Para cada modelo, se muestran el valor del criterio de selección y los efectos en el modelo. Esto ofrece un sentido de la estabilidad de los modelos principales; si tienden a tener muchos efectos similares con pocas diferencias, puede tenerse una confianza casi completa en el modelo "principal"; si tienden a tener muchos efectos diferentes, algunos efectos pueden ser demasiado parecidos

y deberían combinarse (o eliminar uno). Cada columna le permite clasificar las filas, de modo que es posible ver con mayor facilidad qué efectos hay en un paso en particular.

Capítulo 16. Regresión lineal

La regresión lineal estima los coeficientes de la ecuación lineal, con una o más variables independientes, que mejor prediga el valor de la variable dependiente. Por ejemplo, puede intentar predecir el total de ventas anuales de un vendedor (la variable dependiente) a partir de variables independientes tales como la edad, la formación y los años de experiencia.

Ejemplo. ¿Están relacionados el número de partidos ganados por un equipo de baloncesto en una temporada con la media de puntos que el equipo marca por partido? Un diagrama de dispersión indica que estas variables están relacionadas linealmente. El número de partidos ganados y la media de puntos marcados por el equipo adversario también están relacionados linealmente. Estas variables tienen una relación negativa. A medida que el número de partidos ganados aumenta, la media de puntos marcados por el equipo adversario disminuye. Con la regresión lineal es posible modelar la relación entre estas variables. Puede utilizarse un buen modelo para predecir cuántos partidos ganarán los equipos.

Estadísticos. Para cada variable: número de casos válidos, media y desviación estándar. Para cada modelo: coeficientes de regresión, matriz de correlaciones, correlaciones parciales y semiparciales, *R* multiple, *R* cuadrado, *R* cuadrado corregida, cambio en *R* cuadrado, error estándar de la estimación, tabla de análisis de varianza, valores pronosticados y residuos. Además, intervalos de confianza al 95% para cada coeficiente de regresión, matriz de varianzas-covarianzas, factor de inflación de la varianza, tolerancia, prueba de Durbin-Watson, medidas de distancia (Mahalanobis, Cook y valores de influencia), DfBeta, DfAjuste, intervalos de predicción e información de diagnóstico por caso. Gráficos: diagramas de dispersión, gráficos parciales, histogramas y gráficos de probabilidad normal.

Regresión lineal: Consideraciones sobre los datos

Datos. Las variables dependiente e independientes deben ser cuantitativas. Las variables categóricas, como la religión, estudios principales o el lugar de residencia, han de recodificarse como variables binarias (dummy) o como otros tipos de variables de contraste.

Supuestos. Para cada valor de la variable independiente, la distribución de la variable dependiente debe ser normal. La varianza de distribución de la variable dependiente debe ser constante para todos los valores de la variable independiente. La relación entre la variable dependiente y cada variable independiente debe ser lineal y todas las observaciones deben ser independientes.

Para obtener un análisis de regresión lineal

1. Seleccione en los menús:

   Analizar > **Regresión** > **Lineal...**
2. En el cuadro de diálogo Regresión lineal, seleccione una variable numérica dependiente.
3. Seleccione una más variables numéricas independientes.

Si lo desea, puede:

- Agrupar variables independientes en bloques y especificar distintos métodos de entrada para diferentes subconjuntos de variables.
- Elegir una variable de selección para limitar el análisis a un subconjunto de casos que tengan valores particulares para esta variable.
- Seleccionar una variable de identificación de casos para identificar los puntos en los diagramas.
- Seleccione una variable numérica de Ponderación MCP para el análisis de mínimos cuadrados ponderados.

MCP. Permite obtener un modelo de mínimos cuadrados ponderados. Los puntos de los datos se ponderan por los inversos de sus varianzas. Esto significa que las observaciones con varianzas grandes tienen menor impacto en el análisis que las observaciones asociadas a varianzas pequeñas. Si el valor de la variable de ponderación es cero, negativo o perdido, el caso queda excluido del análisis.

Métodos de selección de variables en el análisis de regresión lineal

La selección del método permite 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.

- *Introducir (Regresión)*. Procedimiento para la selección de variables en el que todas las variables de un bloque se introducen en un solo paso.
- *Por pasos*. En cada paso se introduce la variable independiente que no se encuentre ya en la ecuación y que tenga la probabilidad para F más pequeña, si esa probabilidad es suficientemente pequeña. Las variables ya introducidas en la ecuación de regresión se eliminan de ella si su probabilidad para F llega a ser suficientemente grande. El método termina cuando ya no hay más variables candidatas a ser incluidas o eliminadas.
- *Eliminar*. Procedimiento para la selección de variables en el que las variables de un bloque se eliminan en un solo paso.
- *Eliminación hacia atrás*. Procedimiento de selección de variables en el que se introducen todas las variables en la ecuación y después se van excluyendo una tras otra. Aquella variable que tenga la menor correlación parcial con la variable dependiente será la primera en ser considerada para su eliminación. Si satisface el criterio de eliminación, se eliminará. Tras haber excluido la primera variable, se pondrá a prueba aquella variable, de las que queden en la ecuación, que presente una correlación parcial más pequeña. El procedimiento termina cuando ya no quedan en la ecuación variables que satisfagan el criterio de eliminación.
- *Selección hacia adelante*. Procedimiento de selección de variables por pasos en el que las variables se introducen secuencialmente en el modelo. La primera variable que se considerará introducir en la ecuación será la que tenga mayor correlación, positiva o negativa, con la variable dependiente. Dicha variable se introducirá en la ecuación sólo si cumple el criterio de entrada. Si se introduce la primera variable, a continuación se considerará la variable independiente cuya correlación parcial sea la mayor y que no esté en la ecuación. El procedimiento termina cuando ya no quedan variables que cumplan el criterio de entrada.

Los valores de significación de los resultados se basan en el ajuste de un único modelo. Por ello, estos valores de significación no suelen ser válidos cuando se emplea un método por pasos (pasos sucesivos, hacia adelante o hacia atrás).

Todas las variables deben superar el criterio de tolerancia para que puedan ser introducidas en la ecuación, independientemente del método de entrada especificado. El nivel de tolerancia predeterminado es 0,0001. Tampoco se introduce una variable si esto provoca que la tolerancia de otra ya presente en el modelo se sitúe por debajo del criterio de tolerancia.

Todas las variables independientes seleccionadas se añaden a un mismo modelo de regresión. Sin embargo, puede especificar distintos métodos de introducción para diferentes subconjuntos de variables. Por ejemplo, puede introducir en el modelo de regresión un bloque de variables que utilice la selección por pasos sucesivos, y un segundo bloque que emplee la selección hacia adelante. Para añadir un segundo bloque de variables al modelo de regresión, pulse en **Siguiente**.

Regresión lineal: Establecer regla

Los casos definidos por la regla de selección se incluyen en el análisis. Por ejemplo, si selecciona una variable, elija **igual que** y escriba 5 para el valor; de este modo, solamente se incluirán en el análisis los casos para los cuales la variable seleccionada tenga un valor igual a 5. También se permite un valor de cadena.

Regresión lineal: Gráficos

Los gráficos pueden ayudar a validar los supuestos de normalidad, linealidad e igualdad de las varianzas. También son útiles para detectar valores atípicos, observaciones poco usuales y casos de influencia. Tras guardarlos como nuevas variables, dispondrá en el Editor de datos de los valores pronosticados, los residuos y otra información de diagnóstico, con los cuales podrá poder crear gráficos respecto a las variables independientes. Se encuentran disponibles los siguientes gráficos:

Diagramas de dispersión. Puede representar cualquier combinación por parejas de la lista siguiente: la variable dependiente, los valores pronosticados tipificados, los residuos tipificados, los residuos eliminados, los valores pronosticados corregidos, los residuos estudentizados o los residuos eliminados estudentizados. Represente los residuos tipificados frente a los valores pronosticados tipificados para contrastar la linealidad y la igualdad de las varianzas.

Lista de variables de origen. Lista la variable dependiente (DEPENDNT) y las siguientes variables pronosticadas y de residuos: Valores pronosticados tipificados (*ZPRED), Residuos tipificados (*ZRESID), Residuos eliminados (*DRESID), Valores pronosticados corregidos (*ADJPRED), Residuos estudentizados (*SRESID) y Residuos estudentizados eliminados (*SDRESID).

Generar todos los gráficos parciales. Muestra los diagramas de dispersión de los residuos de cada variable independiente y los residuos de la variable dependiente cuando se regresan ambas variables por separado sobre las restantes variables independientes. En la ecuación debe haber al menos dos variables independientes para que se generen los gráficos parciales.

Gráficos de residuos tipificados. Puede obtener histogramas de los residuos tipificados y gráficos de probabilidad normal que comparen la distribución de los residuos tipificados con una distribución normal.

Si se solicita cualquier gráfico, se muestran los estadísticos de resumen para los valores pronosticados tipificados y los residuos tipificados (***ZPRED** y ***ZRESID**).

Regresión lineal: almacenamiento de variables nuevas

Puede guardar los valores pronosticados, los residuos y otros estadísticos útiles para la información de diagnóstico. Cada selección añade una o más variables nuevas a su archivo de datos activo.

Valores pronosticados. Son los valores que el modelo de regresión pronostica para cada caso.

- *No tipificados*. Valor predicho por el modelo para la variable dependiente.
- *Tipificados*. Transformación de cada valor predicho a su forma tipificada. Es decir, se sustrae el valor predicho medio al valor predicho y el resultado se divide por la desviación estándar de los valores pronosticados. Los valores pronosticados tipificados tienen una media de 0 y una desviación estándar de 1.
- *Corregidos*. Valor predicho para un caso cuando dicho caso no se incluye en los cálculos de los coeficientes de regresión.
- *E.T. del predicción promedio*. Error estándar de los valores pronosticados. Estimación de la desviación estándar del valor promedio de la variable dependiente para los casos que tengan los mismos valores en las variables independientes.

Distancias. Son medidas para identificar casos con combinaciones poco usuales de valores para las variables independientes y casos que puedan tener un gran impacto en el modelo.

- *Mahalanobis*. Medida de cuánto difieren del promedio para todos los casos los valores en las variables independientes de un caso dado. Una distancia de Mahalanobis grande identifica un caso que tenga valores extremos en una o más de las variables independientes.
- *De Cook*. Una medida de cuánto cambiarían los residuos de todos los casos si un caso particular se excluyera del cálculo de los coeficientes de regresión. Una Distancia de Cook grande indica que la exclusión de ese caso del cálculo de los estadísticos de regresión hará variar substancialmente los coeficientes.
- *Valores de influencia*. Mide la influencia de un punto en el ajuste de la regresión. Influencia centrada varía entre 0 (no influye en el ajuste) a (N-1)/N.

Intervalos de predicción. Los límites superior e inferior para los intervalos de predicción individual y promedio.

- *Media*. Límites inferior y superior (dos variables) para el intervalo de predicción de la respuesta pronosticada promedio.
- *Individual*. Límites superior e inferior (dos variables) del intervalo de predicción para la variable dependiente para un caso individual.
- *Intervalo de confianza*. Introduzca un valor entre 1 y 99,99 para especificar el nivel de confianza para los dos intervalos de predicción. Debe seleccionar Media o Individuos antes de introducir este valor. Los valores habituales de los intervalos de confianza son 90, 95 y 99.

Residuos. El valor actual de la variable dependiente menos el valor predicho por la ecuación de regresión.

- *No tipificados*. Diferencia entre un valor observado y el valor predicho por el modelo.
- *Tipificados*. El residuo dividido por una estimación de su error estándar. Los residuos tipificados, que son conocidos también como los residuos de Pearson o residuos estandarizados, tienen una media de 0 y una desviación estándar de 1.
- *Estudentizado*. Residuo dividido por una estimación de su desviación estándar que varía de caso en caso, dependiendo de la distancia de los valores de cada caso en las variables independientes respecto a las medias en las variables independientes.
- *Eliminado*. Residuo para un caso cuando éste se excluye del cálculo de los coeficientes de la regresión. Es igual a la diferencia entre el valor de la variable dependiente y el valor predicho corregido.
- *Eliminados estudentizados*. Residuo eliminado para un caso dividido por su error estándar. La diferencia entre un residuo eliminado estudentizado y su residuo estudentizado asociado indica la diferencia que implica el eliminar un caso sobre su propia predicción.

Estadísticos de influencia. El cambio en los coeficientes de regresión (DfBeta) y en los valores pronosticados (DfAjuste) que resulta de la exclusión de un caso particular. También están disponibles los valores tipificados para las DfBeta y para las DfAjuste, junto con la razón entre covarianzas.

- *DfBetas*. La diferencia en el valor de beta es el cambio en el valor de un coeficiente de regresión que resulta de la exclusión de un caso particular. Se calcula un valor para cada término del modelo, incluyendo la constante.
- *DfBeta tipificada*. Valor de la diferencia en beta tipificada. El cambio tipificado en un coeficiente de regresión cuando se elimina del análisis un caso particular. Puede interesarle examinar aquellos casos cuyos valores absolutos sean mayores que 2 dividido por la raíz cuadrada de N, donde N es el número de casos. Se calcula un valor para cada término del modelo, incluyendo la constante.
- *DfAjuste*. La diferencia en el valor ajustado es el cambio en el valor predicho que resulta de la exclusión de un caso particular.
- *DfAjuste tipificada*. Diferencia tipificada en el valor ajustado. El cambio, tipificado, en el valor predicho que resulta de la exclusión de un caso particular. Puede interesarle examinar aquellos valores

tipificados cuyo valor absoluto sea mayor que 2 dividido por la raíz cuadrada de p/N, donde p es el número de variables independientes en la ecuación y N es el número de casos.

- *Razón entre covarianzas*. Razón del determinante de la matriz de covarianza con un caso particular excluido del cálculo de los coeficientes de regresión, respecto al determinante de la matriz de covarianza con todos los casos incluidos. Si la razón se aproxima a 1, el caso no altera significativamente la matriz de covarianza.

Estadísticos de los coeficientes. Almacena los coeficientes de regresión en un conjunto de datos o en un archivo de datos. Los conjuntos de datos están disponibles para su uso posterior durante la misma sesión, pero no se guardarán como archivos a menos que se hayan guardado explícitamente antes de que finalice la sesión. El nombre de un conjunto de datos debe cumplir las normas de denominación de variables.

Exportar información del modelo a un archivo XML. Las estimaciones de los parámetros y (si lo desea) sus covarianzas se exportan al archivo especificado en formato XML (PMML). Puede utilizar este archivo de modelo para aplicar la información del modelo a otros archivos de datos para puntuarlo.

Regresión lineal: Estadísticos

Se encuentran disponibles los siguientes estadísticos:

Coeficientes de regresión. Estimaciones muestra el coeficiente de regresión *B*, el error estándar de *B*, el coeficiente beta tipificado, el valor de *t* para *B* y el nivel de significación bilateral de *t*. **Intervalos de confianza** muestra intervalos de confianza con el nivel de confianza especificado para cada coeficiente de regresión o una matriz de covarianzas. **Matriz de covarianzas** muestra una matriz de varianzas-covarianzas de los coeficientes de regresión, con las covarianzas fuera de la diagonal y las varianzas en la diagonal. También se muestra una matriz de correlaciones.

Ajuste del modelo. Presenta una lista de las variables introducidas y eliminadas del modelo y muestra los siguientes estadísticos de bondad de ajuste: *R* múltiple, R^{cuadrado} y R^{cuadrado} corregida, error estándar de la estimación y tabla de análisis de la varianza.

Cambio en el estadístico R cuadrado. Cambio en el estadístico R^{cuadrado} que se produce al añadir o eliminar una variable independiente. Si es grande el cambio en R^{cuadrado} asociado a una variable, esto significa que esa variable es un buen predictor de la variable dependiente.

Descriptivos. Proporciona el número de casos válidos, la media y la desviación estándar para cada variable en el análisis. También muestra una matriz de correlaciones con el nivel de significación unilateral y el número de casos para cada correlación.

Correlación parcial. La correlación remanente entre dos variables después de haber eliminado la correlación debida a su asociación mutua con otras variables. La correlación entre una variable dependiente y una variable independiente cuando se han eliminado de ambas los efectos lineales de las otras variables independientes del modelo.

Correlación de componente. La correlación entre la variable dependiente y una variable independiente cuando se han eliminado de la variable independiente los efectos lineales de las otras variables independientes del modelo. Está relacionada con el cambio en R cuadrado cuando una variable se añade a una ecuación. En ocasiones e denomina correlación semiparcial.

Diagnósticos de colinealidad. La colinealidad (o multicolinealidad) es una situación no deseable en la que una de las variables independientes es una función lineal de otras variables independientes. Muestra los autovalores de la matriz de productos vectoriales no centrada y escalada, los índices de condición y las proporciones de la descomposición de la varianza junto con los factores de inflación de la varianza (FIV) y las tolerancias para las variables individuales.

Residuos. Presenta la prueba de Durbin-Watson sobre la correlación serial de los residuos y la información de diagnóstico por casos para los casos que cumplan el criterio de selección (los valores atípicos por encima de *n* desviaciones estándar).

Regresión lineal: Opciones

Se encuentran disponibles las siguientes opciones:

Criterios del método por pasos. Estas opciones son aplicables si se ha especificado el método de selección de variables hacia adelante, hacia atrás o por pasos. Las variables se pueden introducir o eliminar del modelo dependiendo de la significación (probabilidad) del valor de *F* o del propio valor de *F*.

- *Usar probabilidad de F*. Una variable se introduce en el modelo si el nivel de significación de su valor de F es menor que el valor de entrada, y se elimina si el nivel de significación de su valor de F es mayor que el valor de Eliminación. La entrada debe ser menor que la eliminación y ambos valores deben ser positivos. Para introducir más variables en el modelo, aumente el valor de entrada. Para eliminar más variables del modelo, disminuya el valor de eliminación.
- *Usar valor de F*. Una variable se introduce en el modelo si su valor de F es mayor que el valor de entrada, y se elimina si su valor de F es menor que el valor de Eliminación. La entrada debe ser mayor que la eliminación y ambos valores deben ser positivos. Para introducir más variables en el modelo, disminuya el valor de entrada. Para eliminar más variables del modelo, eleve el valor de eliminación.

Incluir la constante en la ecuación. De forma predeterminada, el modelo de regresión incluye un término constante. Si se anula la selección de esta opción se obtiene la regresión que pasan por el origen, lo cual se hace raramente. Algunos resultados de la regresión que pasan por el origen no son comparables con los resultados de la regresión que sí incluyen una constante. Por ejemplo, R^2 no puede interpretarse de la manera habitual.

Valores perdidos. Puede elegir uno de los siguientes:

- **Excluir casos según lista.** Sólo se incluirán en el análisis los casos con valores válidos para todas las variables.
- **Excluir casos según pareja.** Los casos con datos completos para la pareja de variables correlacionadas se utilizan para calcular el coeficiente de correlación en el cual se basa el análisis de regresión. Los grados de libertad se basan en el *N* mínimo de las parejas.
- **Reemplazar por la media.** Se emplean todos los casos en los cálculos, sustituyendo las observaciones perdidas por la media de la variable.

Características adicionales del comando REGRESSION

La sintaxis de comandos también le permite:

- Escribir una matriz de correlaciones o leer una matriz (en lugar de los datos en bruto) con el fin de obtener el análisis de regresión (mediante el subcomando `MATRIX`).
- Especificar los niveles de tolerancia (mediante el subcomando `CRITERIA`).
- Obtener múltiples modelos para las mismas variables dependientes u otras diferentes (mediante los subcomandos `METHOD` y `DEPENDENT`).
- Obtener estadísticos múltiples (mediante los subcomandos `DESCRIPTIVE` y `VARIABLES`).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 17. Regresión ordinal

La regresión ordinal permite dar forma a la dependencia de una respuesta ordinal politómica sobre un conjunto de predictores, que pueden ser factores o covariables. El diseño de la regresión ordinal se basa en la metodología de McCullagh (1980, 1998) y en la sintaxis se hace referencia al procedimiento como `PLUM`.

El análisis de regresión lineal ordinario implica minimizar las diferencias de la suma de los cuadrados entre una variable de respuesta (la dependiente) y una combinación ponderada de las variables predictoras (las independientes). Los coeficientes estimados reflejan cómo los cambios en los predictores afectan a la respuesta. Se considera que la respuesta es numérica, en el sentido en que los cambios en el nivel de la respuesta son equivalentes en todo el rango de la respuesta. Por ejemplo, la diferencia de altura entre una persona que mide 150 cm y una que mide 140 cm es de 10 cm, que tiene el mismo significado que la diferencia de altura entre una persona que mide 210 cm y una que mide 200 cm. Estas relaciones no se mantienen necesariamente con las variables ordinales, en las que la elección y el número de categorías de respuesta pueden ser bastante arbitrarios.

Ejemplo. La regresión ordinal podría utilizarse para estudiar la reacción de los pacientes con respecto a una dosis de un fármaco. Las reacciones posibles podrían clasificarse como *ninguna*, *ligera*, *moderada* o *grave*. La diferencia entre una reacción ligera y una moderada es difícil o imposible de cuantificar y se basa en la apreciación. Además, la diferencia entre una respuesta ligera y una moderada podría ser superior o inferior a la diferencia entre una respuesta moderada y una grave.

Estadísticos y gráficos. Frecuencias observadas y esperadas y frecuencias acumuladas, residuos de Pearson para las frecuencias y las frecuencias acumuladas, probabilidades observadas y esperadas, probabilidades acumuladas observadas y esperadas para cada categoría de respuesta por patrón en las covariables, matrices de correlaciones asintóticas y de covarianzas entre las estimaciones de los parámetros, chi-cuadrado de Pearson y chi-cuadrado de la razón de verosimilitud, estadísticos de bondad de ajuste, historial de iteraciones, contraste del supuesto de líneas paralelas, estimaciones de los parámetros, errores estándar, intervalos de confianza y estadísticos R^2 de Cox y Snell, de Nagelkerke y de McFadden.

Regresión ordinal: Consideraciones sobre los datos

Datos. Se asume que la variable dependiente es ordinal y puede ser numérica o de cadena. El orden se determina al clasificar los valores de la variable dependiente en orden ascendente. El valor inferior define la primera categoría. Se asume que las variables de factor son categóricas. Las covariables deben ser numéricas. Observe que al usar más de una covariable continua, se puede llegar a crear una tabla de probabilidades de casilla muy grande.

Supuestos. Sólo se permite una variable de respuesta y debe especificarse. Además, para cada patrón distinto de valores en las variables independientes, se supone que las respuestas son variables multinomiales independientes.

Procedimientos relacionados. La regresión logística nominal utiliza modelos similares para las variables dependientes nominales.

Obtener una regresión ordinal

1. Seleccione en los menús:

   Analizar > **Regresión** > **Ordinal...**
2. Seleccione una variable dependiente.
3. Pulse en **Aceptar**.

Regresión ordinal: Opciones

El cuadro de diálogo Opciones le permite ajustar los parámetros utilizados en el algoritmo de estimación iterativo, seleccionar un nivel de confianza para las estimaciones de los parámetros y seleccionar una función de enlace.

Iteraciones. El algoritmo iterativo puede personalizarse.

- **Número máximo de iteraciones.** Especifique un número entero no negativo. Si se especifica el 0, el procedimiento devolverá las estimaciones iniciales.
- **Máxima subdivisión por pasos.** Especifique un número entero positivo.
- **Convergencia del logaritmo de la verosimilitud.** El algoritmo se detiene si el cambio absoluto o relativo en el logaritmo de la verosimilitud es inferior a este valor. Si se especifica 0, no se utiliza el criterio.
- **Convergencia de los parámetros.** El algoritmo se detiene si el cambio absoluto o relativo en cada una de las estimaciones de los parámetros es inferior a este valor. Si se especifica 0, no se utiliza el criterio.

Intervalo de confianza. Especifique un valor mayor o igual a 0 e inferior a 100.

Delta. El valor añadido a las frecuencias de casilla de cero. Especifique un valor no negativo inferior a 1.

Tolerancia para la singularidad. Utilizada para comprobar los predictores con alta dependencia. Seleccione un valor en la lista de opciones.

Función de enlace. La función de enlace es una transformación de las probabilidades acumuladas que permiten la estimación del modelo. Se encuentran disponibles las cinco funciones de enlace siguientes.

- **Logit.** $f(x)=\log(x/(1-x))$. Se utiliza típicamente para categorías uniformemente distribuidas.
- **Log-log complementario.** $f(x)=\log(-\log(1-x))$. Se utiliza normalmente cuando las categorías más altas son más probables.
- **Log-log negativo.** $f(x)=-\log(-\log(x))$. Se utiliza normalmente cuando las categorías más bajas son más probables.
- **Probit.** $f(x)=\Phi^{-1}(x)$. Se utiliza normalmente cuando la variable latente sigue una distribución normal.
- **Cauchit (Cauchy inversa).** $f(x)=\tan(\pi(x-0.5))$. Se utiliza normalmente cuando la variable latente tiene muchos valores extremos.

Resultados de la regresión ordinal

El cuadro de diálogo Resultados le permite generar tablas que se pueden visualizar en el Visor y guardar variables en el archivo de trabajo.

Representación. Genera tablas correspondientes a:

- **Imprimir el historial de iteraciones.** El logaritmo de la verosimilitud y las estimaciones de los parámetros se imprimen con la frecuencia de iteraciones a imprimir especificada. Siempre se imprimen la primera y la última iteración.
- **Estadísticos de bondad de ajuste.** Estadísticos chi-cuadrado de Pearson y chi-cuadrado de la razón de verosimilitud. Estos estadísticos se calculan según la clasificación especificada en la lista de variables.
- **Estadísticos de resumen.** R^2 de Cox y Snell, de Nagelkerke y de McFadden.
- **Estimaciones de los parámetros.** Estimaciones de los parámetros, errores estándar e intervalos de confianza.
- **Correlación asintótica de las estimaciones de los parámetros.** Matriz de las correlaciones entre las estimaciones de los parámetros.
- **Covarianza asintótica de las estimaciones de los parámetros.** Matriz de las covarianzas entre las estimaciones de los parámetros.

- **Información de casilla.** Frecuencias esperadas y observadas y frecuencias acumuladas, residuos de Pearson para las frecuencias y las frecuencias acumuladas, probabilidades esperadas y observadas y probabilidades esperadas y observadas de cada categoría de respuesta según el patrón en las covariables. Tenga en cuenta que en el caso de modelos con muchos patrones de covariables (por ejemplo, modelos con covariables continuas), esta opción puede generar una tabla grande y poco manejable.
- **Prueba de líneas paralelas.** Prueba correspondiente a la hipótesis de que los parámetros de ubicación son equivalentes en todos los niveles de la variable dependiente. Esta prueba está disponible únicamente para el modelo de sólo ubicación.

Variables guardadas. Guarda las siguientes variables en el archivo de trabajo:

- **Probabilidades de respuesta estimadas.** Probabilidades estimadas por el modelo para la clasificación de un patrón de factores/covariables en las categorías de respuesta. El número de probabilidades es igual al número de categorías de respuesta.
- **Categoría pronosticada.** La categoría de respuesta con la mayor probabilidad estimada para un patrón de factores/covariables.
- **Probabilidad de la categoría pronosticada.** Probabilidades estimada de la clasificación de un patrón factores/covariables en la categoría pronosticada. Esta probabilidad también es el máximo de las probabilidades estimadas para el patrón de factores/covariables.
- **Probabilidad de la categoría real.** Probabilidad estimada de la clasificación de un patrón de factores/covariables en la categoría real.

Imprimir log-verosimilitud. Controla la representación del logaritmo de la verosimilitud. **Incluir constante multinomial** ofrece el valor completo de la verosimilitud. Para comparar los resultados con los productos que no incluyan la constante, puede seleccionar la opción de excluirla.

Modelo de ubicación de la regresión ordinal

El cuadro de diálogo Ubicación le permite especificar el modelo de ubicación para el análisis.

Especificar modelo. Un modelo de efectos principales contiene los efectos principales de las covariables y los factores, pero no contiene efectos de interacción. Puede crear un modelo personalizado para especificar subconjuntos de interacciones entre los factores o bien interacciones entre las covariables.

Factores y covariables. Muestra una lista de los factores y las covariables.

Modelo de ubicación. El modelo depende de los efectos principales y de los de interacción que seleccione.

Generar términos

Para las covariables y los factores seleccionados:

Interacción. Crea el término de interacción de mayor nivel con todas las variables seleccionadas. Este es el método predeterminado.

Efectos principales. Crea un término de efectos principales para cada variable seleccionada.

Todas de 2. Crea todas las interacciones bidimensionales posibles de las variables seleccionadas.

Todas de 3. Crea todas las interacciones tridimensionales posibles de las variables seleccionadas.

Todas de 4. Crea todas las interacciones tetradimensionales posibles de las variables seleccionadas.

Todas de 5. Crea todas las interacciones quíntuples posibles de las variables seleccionadas.

Modelo de escala de la regresión ordinal

El cuadro de diálogo Escala le permite especificar el modelo de escala para el análisis.

Factores y covariables. Muestra una lista de los factores y las covariables.

Modelo de escala. El modelo depende de los efectos principales y de los de interacción que seleccione.

Generar términos

Para las covariables y los factores seleccionados:

Interacción. Crea el término de interacción de mayor nivel con todas las variables seleccionadas. Este es el método predeterminado.

Efectos principales. Crea un término de efectos principales para cada variable seleccionada.

Todas de 2. Crea todas las interacciones bidimensionales posibles de las variables seleccionadas.

Todas de 3. Crea todas las interacciones tridimensionales posibles de las variables seleccionadas.

Todas de 4. Crea todas las interacciones tetradimensionales posibles de las variables seleccionadas.

Todas de 5. Crea todas las interacciones quíntuples posibles de las variables seleccionadas.

Características adicionales del comando PLUM

Se puede personalizar la regresión ordinal si se pegan las selecciones en una ventana de sintaxis y se edita la sintaxis del comando `PLUM` resultante. La sintaxis de comandos también le permite:

- Crear contrastes de hipótesis personalizados especificando las hipótesis nulas como combinaciones lineales de los parámetros.

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 18. Estimación curvilínea

El procedimiento Estimación Curvilínea genera estadísticos de estimación curvilínea por regresión y gráficos relacionados para 11 modelos diferentes de estimación curvilínea por regresión. Se produce un modelo diferente para cada variable dependiente. También se pueden guardar valores predichos, residuos e intervalos de predicción como nuevas variables.

Ejemplo. Un proveedor de servicios de Internet realiza un seguimiento del porcentaje de tráfico de correo electrónico infectado de virus en la red a lo largo del tiempo. Un diagrama de dispersión revela que la relación es no lineal. Se puede ajustar un modelo lineal a los datos y comprobar la validez de los supuestos y la bondad de ajuste del modelo.

Estadísticos. Para cada modelo: coeficientes de regresión, *R* múltiple, *R* $^{\text{cuadrado}}$, *R* $^{\text{cuadrado}}$ corregida, error estándar de la estimación, tabla de análisis de varianza, valores pronosticados, residuos e intervalos de predicción. Modelos: lineal, logarítmico, inverso, cuadrático, cúbico, de potencia, compuesto, curva-S, logístico, de crecimiento y exponencial.

Estimación curvilínea: Consideraciones sobre los datos

Datos. Las variables dependiente e independientes deben ser cuantitativas. Si selecciona **Tiempo** del conjunto de datos activo como variable independiente (en lugar de una variable), el procedimiento Estimación curvilínea generará una variable de tiempo en la que la distancia temporal entre los casos es uniforme. Si se selecciona **Tiempo**, la variable dependiente debe ser una medida de serie temporal. El análisis de series temporales requiere una estructura particular para los archivos de datos, de manera que cada caso (cada fila) represente un conjunto de observaciones en un momento determinado del tiempo y que la distancia temporal entre los casos sea uniforme.

Supuestos. Represente los datos gráficamente para determinar cómo se relacionan las variables dependientes e independiente (linealmente, exponencialmente, etc.). Los residuos de un buen modelo deben distribuirse de forma aleatoria y normal. Si se utiliza un modelo lineal, se deben cumplir los siguientes supuestos: para cada valor de la variable independiente, la distribución de la variable dependiente debe ser normal. La varianza de distribución de la variable dependiente debe ser constante para todos los valores de la variable independiente. La relación entre la variable dependiente y la variable independiente debe ser lineal y todas las observaciones deben ser independientes.

Para obtener una estimación curvilínea

1. Seleccione en los menús:

   Analizar > **Regresión** > **Estimación curvilínea...**
2. Seleccione una o más variables dependientes. Se produce un modelo diferente para cada variable dependiente.
3. Seleccione una variable independiente (seleccione una variable del conjunto de datos activo o **Tiempo**).
4. Si lo desea:

- Seleccionar una variable para etiquetar los casos en los diagramas de dispersión. Para cada punto en el diagrama de dispersión, se puede utilizar la herramienta de Identificación de puntos para mostrar el valor de la variable utilizada en Etiquetas de caso.
- Pulsar en **Guardar** para guardar los valores pronosticados, los residuos y los intervalos de predicción como nuevas variables.

También se encuentran disponibles las siguientes opciones:

- **Incluir la constante en la ecuación.** Estima un término constante en la ecuación de regresión. La constante se incluye de forma predeterminada.
- **Representar los modelos.** Representa los valores de la variable dependiente y cada modelo seleccionado frente a la variable independiente. Se genera un gráfico distinto para cada variable dependiente.
- **Ver tabla de ANOVA.** Muestra una tabla de análisis de varianza de resumen para cada modelo seleccionado.

Modelos del procedimiento Estimación curvilínea

Se puede seleccionar uno o más modelos de estimación curvilínea por regresión. Para determinar qué modelo utilizar, represente los datos. Si las variables parecen estar relacionadas linealmente, utilice un modelo de regresión lineal simple. Cuando las variables no estén relacionadas linealmente, intente transformar los datos. Cuando la transformación no resulte útil, puede necesitar un modelo más complicado. Inspeccione un diagrama de dispersión de los datos; si el diagrama se parece a una función matemática reconocible, ajuste los datos a ese tipo de modelo. Por ejemplo, si los datos se parecen a una función exponencial, utilice un modelo exponencial.

Lineal. Modelo cuya ecuación es Y = b0 + (b1 * t). Los valores de la serie se modelan como una función lineal del tiempo.

Logarítmico. Modelo cuya ecuación es Y = b0 + (b1 * ln(t)).

Inverso. Modelo cuya ecuación es Y = b0 + (b1 / t).

Cuadrático. Modelo cuya ecuación es Y = b0 + (b1 * t) + (b2 * t**2). El modelo cuadrático puede utilizarse para modelar una serie que "despega" o una serie que se amortigua.

Cúbico. Modelo definido por la ecuación Y = b0 + (b1 * t) + (b2 * t**2) + (b3 * t**3).

Potencia. Modelo cuya ecuación es Y = b0 * (t**b1) ó ln(Y) = ln(b0) + (b1 * ln(t)).

Compuesto. Modelo cuya ecuación es Y = b0 * (b1**t) o ln(Y) = ln(b0) + (ln(b1) * t).

Curva-S. Modelo cuya ecuación es Y = e**(b0 + (b1/t)) o ln(Y) = b0 + (b1/t).

Logística. Modelo cuya ecuación es Y = 1 / (1/u + (b0 * (b1**t))) o ln(1/y-1/u) = ln (b0) + (ln(b1) * t) donde u es el valor del límite superior. Después de seleccionar Logístico, especifique un valor para el límite superior que se utilizará en la ecuación de regresión. El valor debe ser un número positivo mayor que el valor máximo de la variable dependiente.

Crecimiento. Modelo cuya ecuación es Y = e**(b0 + (b1 * t)) ó ln(Y) = b0 + (b1 * t).

Exponencial. Modelo cuya ecuación es Y = b0 * (e**(b1 * t)) ó ln(Y) = ln(b0) + (b1 * t).

Estimación curvilínea: Guardar

Guardar variables. Para cada modelo seleccionado se pueden guardar los valores pronosticados, los residuos (el valor observado de la variable dependiente menos el valor predicho por el modelo) y los intervalos de predicción (sus límites superior e inferior). En la ventana de resultados, se muestran en una tabla los nombres de las nuevas variables y las etiquetas descriptivas.

Pronosticar casos. En el conjunto de datos activo, si se selecciona **Tiempo** como variable independiente en lugar de una variable, se puede especificar un período de predicción que vaya más allá del final de la serie temporal. Puede elegir una de las siguientes alternativas:

- **Desde el período de estimación hasta el último caso.** Pronostica los valores para todos los casos del archivo, basándose en los casos del período de estimación. El período de estimación, que se muestra en la parte inferior del cuadro de diálogo, se define con el subcuadro de diálogo Rango de la opción Seleccionar casos en el menú Datos. Si no se ha definido un período de estimación, se utilizan todos los casos para pronosticar los valores.
- **Predecir hasta.** Predice los valores hasta la fecha especificada, hora o número de observación, basándose en los casos del período de estimación. Esta característica se puede utilizar para prever valores más allá del último caso de la serie temporal. Las variables definidas actualmente determinan los cuadros de texto disponibles para especificar el final del período de predicción. Si no existen variables de fecha definidas, se puede especificar el número de la observación (caso) final.

Utilice la opción de Definir fechas en el menú Datos para crear las variables de fecha.

Capítulo 19. Regresión por mínimos cuadrados parciales

El procedimiento Regresión de mínimos cuadrados parciales estima los modelos de regresión de los mínimos cuadrados parciales (PLS, también denominados "proyección a la estructura latente"). La PLS es una técnica de predicción alternativa a la regresión de mínimos cuadrados ordinarios (OLS), a la correlación canónica o al modelado de ecuaciones estructurales, y resulta particularmente útil cuando las variables predictoras están muy correlacionadas o cuando el número de predictores es superior al número de casos.

La PLS combina las características del análisis de componentes principales y la regresión múltiple. En primer lugar, extrae un conjunto de factores latentes que explica en la mayor medida posible la covarianza entre las variables dependientes e independientes. A continuación, un paso de regresión pronostica los valores de las variables dependientes mediante la descomposición de las variables independientes.

Tablas. La proporción de la varianza explicada (por factor latente), las ponderaciones y las cargas de los factores latentes, la importancia de la variable independiente en proyección (VIP) y las estimaciones de los parámetros de la regresión (por variable dependiente) se generan de forma predeterminada.

Gráficos. La variable independiente en proyección (VIP), las puntuaciones factoriales, las ponderaciones factoriales de los tres primeros factores latentes y la distancia al modelo se generan desde la pestaña Options.

Consideraciones sobre los datos de la regresión de mínimos cuadrados parciales

Nivel de medición. Las variables (predictoras) dependientes e independientes pueden ser de escala, nominales u ordinales. El procedimiento supone que se ha asignado el nivel de medición adecuado a todas las variables, aunque puede cambiar temporalmente el nivel de medición de una variable pulsando el botón derecho la variable en la lista de variables de origen y seleccionando un nivel de medición en el menú emergente. El procedimiento trata por igual las variables categóricas (nominales u ordinales).

Codificación de la variable categórica. El procedimiento recodifica temporalmente las variables dependientes categóricas utilizando la codificación "una de *c*" durante el procedimiento. Si hay *c* categorías de una variable, la variable se almacena como *c* vectores, con la primera categoría denotada como (1,0,...,0), la siguiente categoría (0,1,0,...,0), ... y la última categoría (0,0,...,0,1). Las variables dependientes categóricas se representan mediante una variable auxiliar o dummy; es decir, omitiendo simplemente el indicador correspondiente a la categoría de referencia.

Ponderaciones de frecuencia. Los valores de ponderación se redondean al número entero más cercano antes de utilizarlos. Los casos con ponderaciones perdidas o ponderaciones inferiores a 0,5 no se emplearán en los análisis.

Valores perdidos. Los valores perdidos del usuario y del sistema se consideran no válidos.

Cambio de escala. Todas las variables del modelo se centran y tipifican, incluidas las variables indicador que representan variables categóricas.

Para obtener la regresión de mínimos cuadrados parciales

Seleccione en los menús:

Analizar > **Regresión** > **Mínimos cuadrados parciales...**

1. Seleccione al menos una variable dependiente.

2. Seleccione al menos una variable independiente.

Si lo desea, puede:

- Especificar una categoría de referencia para las variables dependientes categóricas (nominales u ordinales).
- Especificar la variable que se utilizará como identificador exclusivo para los resultados por casos y los conjuntos de datos guardados.
- Especificar un límite máximo para el número de factores latentes que se extraerán.

Requisitos

El procedimiento Regresión de mínimos cuadrados parciales es un comando de extensión Python y necesita IBM SPSS Statistics - Essentials for Python, que se instala de forma predeterminada con el producto IBM SPSS Statistics. También necesita las bibliotecas NumPy y SciPy de Python, que están disponibles libremente.

Nota: Para los usuarios que trabajen en modo de análisis distribuido (requiere IBM SPSS Statistics Server), NumPy y SciPy deben estar instalados en el servidor. Póngase en contacto con el administrador del sistema para obtener ayuda.

Usuarios de Windows y Mac

Para Windows y Mac, NumPy y SciPy deben instalarse en una versión diferente de Python 2.7 disntinta de la versión que se instala con IBM SPSS Statistics. Si no tiene una versión diferente de Python 2.7, puede descargarla desde http://www.python.org. A continuación, instale NumPy y SciPy para Python versión 2.7. Los instaladores están disponibles desde http://www.scipy.org/Download.

Para habilitar el uso de NumPy y SciPy, debe establecer la ubicación de Python en la versión de Python 2.7 donde ha instalado NumPy y SciPy. La ubicación de Python se establece desde la pestaña Ubicaciones de archivos en el diálogo Opciones (Editar > Opciones).

Usuarios de Linux

Se recomienda que el usuario descargue el origen y genere NumPy y SciPy. El origen está disponible desde http://www.scipy.org/Download. Puede instalar NumPy y SciPy en la versión de Python 2.7 que se ha instalado con IBM SPSS Statistics. Se encuentra en el directorio `Python` bajo la ubicación donde se ha instalado IBM SPSS Statistics.

Si elige instalar NumPy y SciPy en una versión de Python 2.7 distinta de la versión que se ha instalado con IBM SPSS Statistics, debe establecer la ubicación de Python para que indique dicha versión. La ubicación de Python se establece desde la pestaña Ubicaciones de archivos en el diálogo Opciones (Editar > Opciones).

Servidor de Windows y Unix

NumPy y SciPy deben estar instalada en el servidor en una versión diferente de Python 2.7 de la versión que se instala con IBM SPSS Statistics. Si no hay una versión diferente de Python 2.7 en el servidor, entonces se puede descargar desde http://www.python.org. NumPy y SciPy para Python 2.7 están disponibles en http://www.scipy.org/Download. Para habilitar el uso de NumPy y SciPy, la ubicación de Python para el servidor debe establecerse en la versión de Python 2.7 en la que se han instalado NumPy y SciPy. La ubicación de Python se establece desde IBM SPSS Statistics Administration Console.

Modelo

Especificar efectos del modelo. El modelo de efectos principales contiene todos los efectos principales de los factores y de las covariables. Seleccione **Personalizado** para especificar las interacciones. Indique todos los términos que desee incluir en el modelo.

Factores y Covariables. Muestra una lista de los factores y las covariables.

Modelo. El modelo depende de la naturaleza de los datos. Después de seleccionar **Personalizado**, puede elegir los efectos principales y las interacciones que sean de interés para el análisis.

Generar términos

Para las covariables y los factores seleccionados:

Interacción. Crea el término de interacción de mayor nivel con todas las variables seleccionadas. Este es el método predeterminado.

Efectos principales. Crea un término de efectos principales para cada variable seleccionada.

Todas de 2. Crea todas las interacciones bidimensionales posibles de las variables seleccionadas.

Todas de 3. Crea todas las interacciones tridimensionales posibles de las variables seleccionadas.

Todas de 4. Crea todas las interacciones tetradimensionales posibles de las variables seleccionadas.

Todas de 5. Crea todas las interacciones quíntuples posibles de las variables seleccionadas.

Opciones

La pestaña Opciones permite al usuario guardar y representar las estimaciones de los modelos para los determinados casos, factores latentes y predictores.

Para cada tipo de datos, especifique el nombre del conjunto de datos. Los nombres de los conjuntos de datos deben ser exclusivos. Si introduce el nombre de un conjunto de datos ya existente, se reemplazarán los contenidos. En otro caso, se creará un nuevo conjunto de datos.

- **Guardar estimaciones para casos individuales.** Guarda las siguientes estimaciones de modelos por casos: valores pronosticados, residuos, distancia respecto al modelo del factor latente y puntuaciones de los factores latentes. También representa las puntuaciones de los factores latentes.
- **Guardar estimaciones para factores latentes.** Guarda las cargas y las ponderaciones de los factores latentes. También representa las ponderaciones de factores latentes.
- **Guardar estimaciones para variables independientes.** Guarda las estimaciones de los parámetros de regresión y la importancia de la variable en la proyección (VIP). También representa la VIP por factor latente.

Capítulo 20. Análisis vecino más cercano

Análisis de vecinos más próximos es un método para clasificar casos basándose en su parecido a otros casos. En el aprendizaje automático, se desarrolló como una forma de reconocer patrones de datos sin la necesidad de una coincidencia exacta con patrones o casos almacenados. Los casos parecidos están próximos y los que no lo son están alejados entre sí. Por lo tanto, la distancia entre dos casos es una medida de disimilaridad.

Los casos próximos entre sí se denominan "vecinos". Cuando se presenta un nuevo caso (reserva), se calcula su distancia con respecto a los casos del modelo. Las clasificaciones de los casos más parecidos (los vecinos más próximos) se cuadran y el nuevo caso se incluye en la categoría que contiene el mayor número de vecinos más próximos.

Puede especificar el número de vecinos más próximos que deben examinarse; este valor se denomina *k*.

El método Análisis de vecinos más próximos también puede utilizarse para calcular valores para un destino continuo. En esta situación, la media o el valor objetivo medio de los vecinos más próximos se utiliza para obtener el valor predicho del nuevo caso.

Análisis de vecino más próximo: Consideraciones sobre los datos

Objetivo y características. El objetivo y las características 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.

  El análisis de vecinos más próximos trata por igual las variables nominales u ordinales. El procedimiento supone que se ha asignado el nivel de medición adecuado a cada variable. No obstante, puede cambiar temporalmente el nivel de medición para una variable pulsando con el botón derecho en la variable en la lista de variables de origen y seleccionar un nivel de medición en el menú emergente.

Un icono situado junto a cada variable de la lista de variables identifica el nivel de medición y el tipo de datos.

Tabla 1. Iconos de nivel de medición

	Numérico	**Cadena**	**Fecha**	**Hora**
Escala (Continuo)		n/a		
Ordinal				

Tabla 1. Iconos de nivel de medición (continuación)

	Numérico	Cadena	Fecha	Hora
Nominal				

Codificación de la variable categórica. El procedimiento recodifica temporalmente predictores categóricos y variables dependientes utilizando la codificación "una de *c*" para todo el procedimiento. Si hay *c* categorías de una variable, la variable se almacena como vectores c, con la primera categoría denotada (1,0,...,0), la siguiente categoría (0,1,0,...,0), ..., y la última categoría (0,0,...,0,1).

Este esquema de codificación aumenta la dimensionalidad del espacio de características. En concreto, el número total de dimensiones es el número de predictores de escala más el número de categorías en todos los predictores categóricos. Como resultado, este esquema de codificación puede conllevar un entrenamiento más lento. Si el entrenamiento de vecinos más próximos avanza muy lentamente, pruebe a reducir el número de categorías en los predictores categóricos combinando categorías similares o eliminando los casos que tengan categorías extremadamente raras antes de ejecutar el procedimiento.

Toda codificación "una de *c*" se basa en los datos de entrenamiento, incluso si se define una muestra reservada (consulte “Particiones” en la página 90). De este modo, si las muestras reservadas contienen casos con categorías de predictores que no están presentes en los datos de entrenamiento, esos casos no se puntúan. Si las muestras reservadas contienen casos con categorías de variables dependientes que no están presentes en los datos de entrenamiento, esos casos se puntúan.

Cambio de escala. Fe forma predeterminada, las características de escala se normalizan. Todo cambio de escala se realiza basándose en los datos de entrenamiento, incluso si se define una muestra reservada (consulte “Particiones” en la página 90). Si especifica una variable para definir particiones, es importante que las funcaracterísticasciones tengan distribuciones similares en todas las muestras reservadas, de entrenamiento o comprobación. Utilice, por ejemplo, el procedimiento Explorar para examinar las distribuciones en las particiones.

Ponderaciones de frecuencia. Este procedimiento ignora las ponderaciones de frecuencia.

Replicación de los resultados. El procedimiento utiliza la generación de números aleatorios durante la asignación aleatoria de particiones y pliegues de validación cruzada. Si desea duplicar los resultados de forma exacta, además de utilizar los mismos ajustes de procedimiento, defina una semilla para el Tornado de Mersenne (consulte “Particiones” en la página 90), o utilice variables para definir particiones y pliegues de validación cruzada.

Para obtener un análisis de vecino más próximo

Seleccione en los menús:

Analizar > **Clasificar** > **Vecino más próximo...**

1. Especifique una o más características que puedan constituir variables independientes o predictores en caso de haber un destino.

   Destino (opcional). Si no hay ningún destino (variable dependiente o respuesta) especificado, el procedimiento encontrará únicamente los *k* vecinos más próximos, sin realizar ninguna clasificación ni predicción.

   Características de escala de normalización. Las características normalizadas tienen el mismo rango de valores, lo que puede mejorar el rendimiento del algoritmo de estimación. Se utilizará la normalización ajustada [2*(x−min)/(max−min)]1. Los valores normalizados ajustados quedan comprendidos entre -1 y 1.

Identificador de caso focal (opcional). Esto le permite marcar casos de especial interés. Por ejemplo, un investigador desea determinar si las puntuaciones de las pruebas de un distrito escolar (el caso focal) son comparables con las de distritos escolares similares. Utiliza un análisis de vecinos más próximos para encontrar los distritos escolares más parecidos con respecto a un conjunto dado de características. Después compara las puntuaciones de las pruebas del distrito escolar focal con las de los vecinos más próximos.

Los casos focales también deben emplearse en estudios clínicos para seleccionar casos de control similares a los casos clínicos. Los casos focales se muestran en la tabla de *k* vecinos más próximos y distancias, el gráfico de espacio de características, el gráfico de homólogos y el mapa de cuadrantes. La información sobre casos focales se guarda en los archivos especificados en la pestaña Resultados.

Los casos con un valor positivo en la variable especificada se tratan como casos focales. No es posible especificar una variable sin valores positivos.

Etiqueta de caso (opcional). Los casos se etiquetan utilizando estos valores en el gráfico de espacio de características, el gráfico de homólogos y el mapa de cuadrantes.

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.

Vecinos

Número de vecinos más próximos (k) Especifique el número de vecinos más próximos. Tenga en cuenta que el uso de un número mayor de vecinos no implica que el modelo resultante sea más preciso.

Si se especifica un destino en la pestaña Variables, puede especificar un rango de valores y permitir que el procedimiento seleccione el "mejor" número de vecinos de ese rango. El método para determinar el número de vecinos más próximos depende de si se solicita la selección de características en la pestaña Características.

- Si la selección de características está activada, ésta se realizará para cada valor de *k* en el rango solicitado, y se seleccionará la *k* y el conjunto de funciones compañero con la menor tasa de error (o el menor error cuadrático si el destino es escala).
- Si la selección de características no está activada, se utilizará la validación cruzada de pliegue en *V* para seleccionar el “mejor” número de vecinos. Consulte la pestaña Partición para tener control sobre la asignación de pliegues.

Cálculo de distancias. Es la métrica utilizada para especificar la métrica de distancia empleada para medir la similitud de los casos.

- **Métrica euclídea.** La distancia entre dos casos, x e y, es la raíz cuadrada de la suma, sobre todas las dimensiones, de las diferencias cuadradas entre los valores de esos casos.

- **Métrica de bloques de ciudad.** La distancia entre dos casos es la suma, en todas las dimensiones, de las diferencias absolutas entre los valores de esos casos. También se conoce como la distancia de Manhattan.

Además, si se especifica un destino en la pestaña Variables, puede optar por ponderar características según su importancia normalizada a la hora de calcular distancias. La importancia que una característica tiene para un predictor se calcula en función de la relación entre la tasa de error o errores cuadráticos del modelo sin el predictor y la tasa de error o errores cuadráticos del modelo completo. La importancia normalizada se calcula volviendo a ponderar los valores de importancia de la característica para que sumen 1.

Predicciones del destino de escala. Si se especifica un destino de escala en la pestaña Variables, especificará si el valor predicho se calcula en función de la media o del valor de mediana de los vecinos más próximos.

Características

La pestaña Características le permite seleccionar y especificar opciones para la selección de características cuando se especifica un destino en la pestaña Variables. De forma predeterminada, todas las características se tienen en cuenta para la selección de características, pero es posible seleccionar un subconjunto de características para forzarlas en el modelo.

Criterio de parada. En cada paso, la característica cuya suma al modelo dé lugar al menor error (calculado como la tasa de error de un destino categórico y el error cuadrático de un destino de escala) se tiene en cuenta para su inclusión en el conjunto de modelos. La selección continúa hasta que se cumple la condición especificada.

- **Número de características especificadas.** El algoritmo añade un número fijo de características además de las forzadas en el modelo. Especifique un número entero positivo. Si se disminuyen los valores de número que se puede seleccionar se obtiene un modelo más reducido, lo que supone el riesgo de perder importantes características. Si se aumentan los valores de número que se puede seleccionar se incluirán todas las características importantes, pero se corre el riesgo de añadir características que aumenten el error del modelo.
- **Cambio mínimo de la tasa de errores absolutos.** El algoritmo se detiene cuando el cambio de la tasa de errores absolutos indica que el modelo no puede mejorarse más añadiendo nuevas características. Especifique un número positivo. Si se reducen los valores del cambio mínimo se incluirán más características, pero puede que se incluyan características que no añadan gran valor al modelo. Si se aumentan los valores del cambio mínimo se excluirán más características, pero puede que se pierdan características importantes para el modelo. El valor “óptimo” de cambio mínimo dependerá de sus datos y de la aplicación. Consulte el Registro de errores de selección de características en los resultados para poder evaluar qué características son más importantes. Consulte el tema “Registro de errores de selección de características” en la página 95 para obtener más información.

Particiones

La pestaña Particiones le permite dividir el conjunto de datos en conjuntos de entrenamiento y reserva y, siempre que proceda, asignar casos a pliegues de validación cruzada.

Particiones de entrenamiento y reserva. Este grupo especifica el método de crear particiones en el conjunto de datos activo correspondientes a las muestras de entrenamiento y reserva. La **muestra de entrenamiento** comprende los registros de datos utilizados para entrenar el modelo de vecino más próximo; cierto porcentaje de casos del conjunto de datos debe asignarse a la muestra de entrenamiento para poder obtener un modelo. La **muestra reservada** es otro conjunto independiente de registros de datos que se utiliza para evaluar el modelo final; el error de la muestra reservada ofrece una estimación "sincera" de la capacidad predictora del modelo, ya que los casos reservados no se utilizan para crear el modelo.

- **Asignar casos a particiones aleatoriamente.** Especifique el porcentaje de casos que se asignarán a la muestra de entrenamiento. El resto se asignan a la muestra reservada.
- **Utilizar variable para asignar los casos.** Especifique una variable numérica que asigne cada caso del conjunto de datos activo a la muestra de entrenamiento o reserva. Los casos con un valor positivo de la variable se asignarán a la muestra de entrenamiento, los casos con un valor 0 o negativo se asignarán a la muestra reservada. Los casos con un valor perdido del sistema se excluirán del análisis. Todos los valores perdidos de usuario de la variable de partición se tratarán siempre como válidos.

Pliegues de validación cruzada. La validación cruzada de pliegue en *V* se utiliza para determinar el "mejor" número de vecinos. Por razones de rendimiento, no está disponible con la selección de características.

La validación cruzada divide la muestra en un número de submuestras o pliegues. A continuación, se generan los modelos de vecino más próximo, que no incluyen los datos de cada submuestra. El primer modelo se basa en todos los casos excepto los correspondientes al primer pliegue de la muestra; el segundo modelo se basa en todos los casos excepto los del segundo pliegue de la muestra y así sucesivamente. Para cada modelo se calcula el error aplicando el modelo a la submuestra que se excluyó al generarlo. El "mejor" número de vecinos más próximos será el que produzca el menor error entre los pliegues.

- **Asignar casos a pliegues aleatoriamente.** Especifique el número de pliegues que se utilizarán para la validación cruzada. El procedimiento asigna aleatoriamente casos a los pliegues, numerados de 1 a *V*, que es el número de pliegues.
- **Utilizar variable para asignar los casos.** Especifique una variable numérica que asigne cada caso del conjunto de datos activo a un pliegue. La variable debe ser numérica y adoptar valores de 1 a *V*. Si falta algún valor de este rango, y en cualquier segmento si los archivos de segmentación están en vigor, se producirá un error.

Definir semilla para tornado de Mersenne. Si se establece una semilla es posible replicar análisis. El uso de este control es parecido a establecer el tornado de Mersenne como generador activo y especificar un punto de inicio fijo en el cuadro de diálogo Generadores de números aleatorios, con la importante diferencia de que la definición de la semilla de este cuadro de diálogo mantendrá el estado actual del generador de números aleatorios y restaurará dicho estado cuando haya terminado el análisis.

Guardado

Nombres de las variables guardadas. La generación automática de nombres garantiza que conserva todo su trabajo. Los nombres personalizados le permiten descartar/reemplazar los resultados de las ejecuciones anteriores sin eliminar antes las variables guardadas en el Editor de datos.

Variables a guardar

- **Valor o categoría pronosticados.** Esta opción guarda el valor predicho para el destino de escala o la categoría predicha para un destino categórico.
- **Probabilidad predicha.** Esta opción guarda las probabilidades pronosticadas para un destino categórico. Para cada una de las primeras *n* categorías se guarda una variable diferente, donde *n* se especifica en el control **Máximo de categorías para guardar para un destino categórico**.
- **Variables de particiones de entrenamiento y reserva.** Si los casos se asignan aleatoriamente a las muestras de entrenamiento y reserva de la pestaña Particiones, esta opción guarda el valor de la partición (entrenamiento y reserva) a la que se ha asignado el caso.
- **Variable de pliegues de validación cruzada.** Si los casos se asignan aleatoriamente a los pliegues de validación cruzada de la pestaña Particiones, esta opción guarda el valor del pliegue al que se ha asignado el caso.

Resultados

Salida del visor

- **Resumen de procesamiento de casos.** Muestra la tabla de resumen de procesamiento de casos, que resume el número de casos incluidos y excluidos en el análisis, en total y por muestras de entrenamiento y reservadas.
- **Gráficos y tablas.** Muestra los resultados relacionados con los modelos, incluyendo tablas y gráficos. Las tablas de la vista de modelo incluyen los *k* vecinos más próximos y las distancias de casos focales, la clasificación de variables de respuesta categórica y un resumen de errores. El resultado gráfico de la vista de modelo incluye un registro de errores de selección, un gráfico de importancia de características, un gráfico de espacio de características, un gráfico de homólogos y un mapa de cuadrante. Consulte el tema "Vista de modelo" para obtener más información.

Archivos

- **Exportar modelo a XML.** Puede utilizar este archivo de modelo para aplicar la información del modelo a otros archivos de datos para puntuarlo. Esta opción no se encuentra disponible si se han definido archivos segmentados.
- **Exportar distancias entre casos focales y k vecinos más próximos.** En cada caso focal, se crea una variable distinta para cada uno de los *k* vecinos más próximos del caso focal (de la muestra de entrenamiento) y las *k* distancias más próximas correspondientes.

Opciones

Valores perdidos del usuario. Para que un caso se incluya en el análisis, las variables categóricas deben tener valores válidos para dicho caso. Estos controles permiten decidir si los valores perdidos del usuario se deben tratar como válidos entre las variables categóricas.

Los valores perdidos del sistema y perdidos para las variables de escala siempre se tratan como no válidos.

Vista de modelo

Cuando seleccione **Gráficos y tablas** en la pestaña Resultados, el procedimiento creará un objeto de modelo de vecino más próximo en el visor. Al activar (pulsando dos veces) este objeto se obtiene una vista interactiva del modelo. La vista de modelos tiene una ventana con dos paneles:

- El primer panel muestra una descripción general del modelo denominado vista principal.
- El segundo panel muestra uno de los dos tipos de vistas:

  Una vista de modelos auxiliar muestra más información sobre el modelo, pero no se centra en el propio modelo.

  Una vista enlazada es una vista que muestra detalles sobre una característica del modelo cuando el usuario desglosa parte de la vista principal.

De forma predeterminada, el primer panel muestra el espacio de características y el segundo muestra el gráfico de importancia de variables. Si el gráfico de importancia de variables no está disponible, es decir, si **Ponderar características por importancia** no se ha seleccionado en la pestaña Funciones, se mostrará la primera vista disponible en la lista desplegable Ver.

Cuando una vista no tiene ninguna información disponible, se desactiva este texto de elemento en la lista desplegable Ver.

Espacio de características

El gráfico de espacio de características es un gráfico interactivo del espacio de características (o un subespacio, si hay más de 3 características). Cada eje representa una característica del modelo, y la ubicación de los puntos del gráfico muestran los valores de dichas características para casos de las particiones de entrenamiento y reserva.

Claves. Además los valores de las características, los puntos del gráfico indican otra información.

- La forma indica la partición a la que pertenece un punto, ya sea Entrenamiento o Reserva.
- El color y el sombreado de un punto indican el valor del destino de ese caso: cada valor de color diferente representa las categorías de un destino categórico y las sombras indican el rango de valores de un destino continuo. El valor indicado para la partición de entrenamiento es el valor observado, mientras que en el caso de la partición de reserva, representa el valor predicho. Si no se especifica ningún destino, esta clave no aparece.
- Los titulares más gruesos indican que un caso es focal. Los casos focales se muestran en relación con sus *k* vecinos más próximos.

Controles e interactividad. Una serie de controles del gráfico le permite explorar el espacio de características.

- Puede seleccionar qué subconjunto de características mostrar en el gráfico y modificar qué funciones se representan en las dimensiones.
- Los "casos focales" son simplemente puntos seleccionados en el gráfico del espacio de características. Si ha especificado una variable de caso focal, los puntos que representan los casos focales se seleccionarán inicialmente. Sin embargo, cualquier punto puede convertirse en un caso focal si lo selecciona. A la selección de puntos se aplican los controles "normales", es decir, si pulsa en un punto éste se selecciona y se cancela la selección de todos los demás y si pulsa Control y el ratón sobre un punto éste se añadirá al conjunto de puntos seleccionados. Las vistas enlazadas, como el gráfico de homólogos, se actualizarán automáticamente en función de los casos seleccionados en el espacio de características.
- Puede modificar el número de vecinos más próximos (*k*) para mostrar casos focales.
- Al pasar el ratón sobre un punto del gráfico se mostrará una ayuda contextual con el valor de la etiqueta de caso o un número de caso si las etiquetas de caso no se definen, así como los valores de destino observados y pronosticados.
- Un botón "Restablecer" le permite devolver el espacio de características a su estado original.

Adición y eliminación de campos/variables

Puede añadir nuevos campos/variables al espacio de características o eliminar los que se visualizan en este momento.

Paleta de variables

Debe visualizar la paleta de variables antes de que pueda añadir y eliminar variables. Para visualizar la paleta de variables, el visor de modelos deberá estar en modo de edición y deberá seleccionarse un caso en el espacio de características.

1. Para poner el visor de modelos en modo de edición, elija en los menús:

   Ver > **Modo de edición**
2. Una vez en Modo Edición, pulse sobre cualquier caso del espacio de características.
3. Para visualizar la paleta de variables, elija en los menús:

   Ver > **Paletas** > **Variables**

   La paleta de variables enumera todas las variables del espacio de características. El icono junto al nombre de variable indica el nivel de medición de la variable.

4. Para cambiar temporalmente el nivel de medición de una variable, pulse con el botón derecho en la variable de la paleta de variables y seleccione una opción.

Zonas de variables

Las variables se añaden a "zonas" del espacio de características. Para visualizar las zonas, empiece arrastrando una variable desde la paleta de variables o seleccionando **Mostrar zonas**.

El espacio de características tiene zonas para los ejes *x*, *y* y *z*.

Desplazamiento de variables a zonas

Éstas son algunas reglas generales y sugerencias para desplazar variables a zonas:

- Para desplazar una variable a una zona, pulse y arrastre la variable desde la paleta de variables y suéltela en la zona. Si selecciona **Mostrar zonas**, también puede pulsar con el botón derecho en una zona y seleccionar una variable que desee añadir a la zona.
- Si arrastra una variable de la paleta de variables a una zona que ya esté ocupada por otra variable, la nueva variable sustituirá a la anterior.
- Si arrastra una variable de una zona a una zona que ya esté ocupada por otra variable, las variables intercambiarán posiciones.
- Si pulsa en la X de una zona, eliminará la variable de dicha zona.
- Si hay varios elementos gráficos en la visualización, cada elemento gráfico puede tener sus propias zonas de variables asociadas. Primero, seleccione el elemento gráfico.

Importancia de la variable

Normalmente, desea centrar sus esfuerzos de modelado en las variables que importan más y considera eliminar o ignorar las que importan menos. El gráfico de importancia de la variable le ayuda a hacerlo indicando la importancia relativa de cada variable en la estimación del modelo. Como las variables son relativas, la suma de los valores de todas las variables de la visualización es 1,0. La importancia de variable no está relacionada con la precisión del modelo. Sólo está relacionada con la importancia de cada variable para realizar una predicción, independientemente de si ésta es precisa o no.

Homólogos

Este gráfico muestra los casos focales y sus *k* vecinos más próximos en cada característica y en el destino. Está disponible si se selecciona un caso focal en el espacio de características.

Forma de enlace. El gráfico Homólogos se enlaza con el espacio de características de dos formas.

- Los casos seleccionados (focal) en el espacio de características se muestran en el gráfico Homólogos, juntos con sus *k* vecinos más próximos.
- El valor de *k* seleccionado en el espacio de características se utiliza en el gráfico Homólogos.

Distancias de vecinos más próximos

Esta tabla muestra los *k* vecinos más próximos y las distancias de casos focales únicamente. Está disponible si se especifica un identificador de caso focal en la pestaña Variables, y sólo muestra los casos focales identificados por esta variable.

Cada fila de:

- La columna **Caso focal** contiene el valor de la variable de etiqueta de caso del caso focal; si las etiquetas de caso no se definen, esta columna contendrá el número de caso del caso focal.
- La i^a columna del grupo Vecinos más próximos contiene el valor de la variable de etiqueta de caso del i^o vecino más próximo al caso focal; si las etiquetas de caso no se definen, esta columna contendrá el número de caso del i^o vecino más próximo al caso focal.

- La *i*ª columna del grupo Distancias más próximas contiene la distancia del *i*º vecino más próximo al caso focal.

Mapa de cuadrantes

Este gráfico muestra los casos focales y sus *k* vecinos más próximos en un diagrama de dispersión (o gráfico de puntos, dependiendo del nivel de medición del destino) con el destino en el eje *y* y una característica de escala en el eje *x*, panelado por características. Está disponible si hay un destino y se selecciona un caso focal en el Espacio de características.

- Se dibujan líneas de referencia para las variables continuas en las medias variables en la partición de entrenamiento.

Registro de errores de selección de características

Señala en la vista de gráfico el error (la tasa de error o de error cuadrático, dependiendo del nivel de medición del destino) en el eje *y* para el modelo con la característica enumerada en el eje *x* (además de todas las características a la izquierda del eje *x*). Este gráfico está disponible si hay un destino y la selección de características está activada.

Registro de errores de selección de k

Señala en la vista de gráfico el error (la tasa de error o de error cuadrático, dependiendo del nivel de medición del destino) en el eje *y* para el modelo con el número de vecinos más próximos (*k*) en el eje *x*. Este gráfico está disponible si hay un destino y la selección de *k* está activada.

Registro de errores de selección de características y k

Estos son los gráficos de selección de características (consulte "Registro de errores de selección de características"), panelados por *k*. Este gráfico está disponible si hay un destino y la selección de características y *k* están activadas.

Tabla de clasificación

Esta tabla muestra la clasificación cruzada de los valores observados en comparación con los valores pronosticados del destino, en función de la partición. Está disponible si hay un destino y es categórico.

- La fila **(Perdidos)** de la partición de reserva contiene casos de reserva con los valores perdidos en el destino. Estos casos contribuyen a los valores de Muestra reservada: Valores de Porcentaje global, pero no a los valores de Porcentaje correcto.

Resumen de error

Esta tabla está disponible si hay una variable objetivo. Muestra el error asociado con el modelo, la suma de cuadrados de un destino continuo y la tasa de error (100%, porcentaje global correcto) de un destino categórico.

Capítulo 21. Análisis discriminante

El análisis discriminante crea un modelo predictivo para la pertenencia al grupo. El modelo está compuesto por una función discriminante (o, para más de dos grupos, un conjunto de funciones discriminantes) basada en combinaciones lineales de las variables predictoras que proporcionan la mejor discriminación posible entre los grupos. Las funciones se generan a partir de una muestra de casos para los que se conoce el grupo de pertenencia; posteriormente, las funciones pueden ser aplicadas a nuevos casos que dispongan de mediciones para las variables predictoras pero de los que se desconozca el grupo de pertenencia.

Nota: la variable de agrupación puede tener más de dos valores. Los códigos de la variable de agrupación han de ser números enteros y es necesario especificar sus valores máximo y mínimo. Los casos con valores fuera de estos límites se excluyen del análisis.

Ejemplo. Por término medio, las personas de los países de zonas templadas consumen más calorías por día que las de los trópicos, y una proporción mayor de la población de las zonas templadas vive en núcleos urbanos. Un investigador desea combinar esta información en una función para determinar cómo de bien un individuo es capaz de discriminar entre los dos grupos de países. El investigador considera además que el tamaño de la población y la información económica también pueden ser importantes. El análisis discriminante permite estimar los coeficientes de la función discriminante lineal, que tiene el aspecto de la parte derecha de una ecuación de regresión lineal múltiple. Es decir, utilizando los coeficientes *a*, *b*, *c* y *d*, la función es:

```
D = a * clima + b * urbanos + c * población + d * producto interior bruto per cápita
```

Si estas variables resultan útiles para discriminar entre las dos zonas climáticas, los valores de *D* serán diferentes para los países templados y para los tropicales. Si se utiliza un método de selección de variables por pasos, quizás no se necesite incluir las cuatro variables en la función.

Estadísticos. Para cada variable: medias, desviaciones estándar, ANOVA univariado. Para cada análisis: *M* de Box, matriz de correlaciones intra-grupos, matriz de covarianzas intra-grupos, matriz de covarianzas de los grupos separados, matriz de covarianzas total. Para cada función discriminante canónica: autovalores, porcentaje de varianza, correlación canónica, lambda de Wilks, chi-cuadrado. Para cada paso: probabilidades previas, coeficientes de la función de Fisher, coeficientes de función no tipificados, lambda de Wilks para cada función canónica.

Análisis discriminante: Consideraciones sobre los datos

Datos. La variable de agrupación debe tener un número limitado de categorías distintas, codificadas como números enteros. Las variables independientes que sean nominales deben ser recodificadas a variables auxiliares o de contraste.

Supuestos. Los casos deben ser independientes. Las variables predictoras deben tener una distribución normal multivariada y las matrices de varianzas-covarianzas intra-grupos deben ser iguales en todos los grupos. Se asume que la pertenencia al grupo es mutuamente exclusiva (es decir, ningún caso pertenece a más de un grupo) y exhaustiva de modo colectivo (es decir, todos los casos son miembros de un grupo). El procedimiento es más efectivo cuando la pertenencia al grupo es una variable verdaderamente categórica; si la pertenencia al grupo se basa en los valores de una variable continua (por ejemplo, un cociente de inteligencia alto respecto a uno bajo), considere el uso de la regresión lineal para aprovechar la información más rica ofrecida por la propia variable continua.

Para obtener un análisis discriminante

1. Seleccione en los menús:

Analizar > **Clasificar** > **Discriminante...**

2. Seleccione una variable de agrupación con valores enteros y pulse en **Definir rango** para especificar las categorías de interés.
3. Seleccione las variables independientes o predictoras. (Si la variable de agrupación no tiene valores enteros, la opción Recodificación automática en el menú Transformar creará una variable que los tenga).
4. Seleccione el método de introducción de las variables independientes.
   - **Introducir independientes juntas.** Introducir simultáneamente todas las variables independientes que satisfacen el criterio de tolerancia.
   - **Usar método de inclusión por pasos.** Utiliza el análisis por pasos para controlar la entrada y la eliminación de variables.
5. Si lo desea, seleccione casos mediante una variable de selección.

Análisis discriminante: Definir rango

Especifique los valores mínimo y máximo de la variable de agrupación para el análisis. Los casos con valores fuera de este rango no se utilizan en el análisis discriminante, pero sí se clasifican en uno de los grupos existentes a partir de los resultados que obtengan en el análisis. Los valores mínimo y máximo deben ser números enteros.

Análisis discriminante: Seleccionar casos

Para seleccionar casos para el análisis:

1. En el cuadro de diálogo Análisis discriminante, seleccione una variable de selección.
2. Pulse en **Valor** para introducir un número entero como valor de selección.

Sólo se utilizan los casos con el valor especificado en la variable de selección para derivar las funciones discriminantes. Tanto para los casos seleccionados como para los no seleccionados se generan resultados de clasificaciones y estadísticos. Este proceso ofrece un mecanismo para clasificar casos nuevos basados en datos previos o para dividir los datos en subconjuntos de contraste y comprobación para realizar procedimientos de validación en el modelo generado.

Análisis discriminante: Estadísticos

Descriptivos. Las opciones disponibles son: Medias (que incluye las desviaciones estándar), ANOVAs univariados y prueba *M* de Box.

- *Medias*. Muestra la media y desviación estándar totales y las medias y desviaciones estándar de grupo, para las variables independientes.
- *ANOVAs univariados*. Realiza un análisis de varianza de un factor sobre la igualdad de las medias de grupo para cada variable independiente.
- *M de Box*. Contraste sobre la igualdad de las matrices de covarianza de los grupos. Para tamaños de muestras suficientemente grandes, un valor de p no significativo quiere decir que no hay suficiente evidencia de que las varianzas sean diferentes. Esta prueba es sensible a las desviaciones de la normalidad multivariada.

Coeficientes de la función. Las opciones disponibles son: Coeficientes de clasificación de Fisher y Coeficientes no tipificados.

- *De Fisher*. Muestra los coeficientes de la función de clasificación de Fisher que pueden utilizarse directamente para la clasificación. Se obtiene un conjunto de coeficientes para cada grupo, y se asigna un caso al grupo para el que tiene una mayor puntuación discriminante (valor de función de clasificación).
- *No tipificados*. Muestra los coeficientes de la función discriminante sin estandarizar.

Matrices. Las matrices de coeficientes disponibles para las variables independientes son las de: Correlación intra-grupos, Covarianza intra-grupos, Covarianza de grupos separados y Covarianza total.

- *Correlación intra-grupos*. Muestra la matriz de correlaciones intra-grupos combinada, que se obtiene de promediar las matrices de covarianza individuales para todos los grupos antes de calcular las correlaciones.
- *Covarianza intra-grupos*. Muestra la matriz de covarianza intra-grupos combinada, la cual puede diferir de la matriz de covarianza total. La matriz se obtiene de promediar, para todos los grupos, las matrices de covarianza individuales.
- *Covarianza de grupos separados*. Muestra las matrices de covarianza de cada grupo por separado.
- *Covarianza total*. Muestra la matriz de covarianza para todos los casos, como si fueran una única muestra.

Análisis discriminante: Método de inclusión por pasos

Método. Seleccione el estadístico que se va a utilizar para introducir o eliminar nuevas variables. Las alternativas disponibles son la lambda de Wilks, la varianza no explicada, la distancia de Mahalanobis, la menor razón *F* y la *V* de Rao. Con la *V* de Rao se puede especificar el incremento mínimo de *V* para introducir una variable.

- *Lambda de Wilks*. Método para la selección de variables por pasos del análisis discriminante que selecciona las variables para su introducción en la ecuación basándose en cuánto contribuyen a disminuir la lambda de Wilks. En cada paso se introduce la variable que minimiza la lambda de Wilks global.
- *Varianza no explicada*. En cada paso se introduce la variable que minimiza la suma de la variación no explicada entre los grupos.
- *Distancia de Mahalanobis*. Medida de cuánto difieren del promedio para todos los casos los valores en las variables independientes de un caso dado. Una distancia de Mahalanobis grande identifica un caso que tenga valores extremos en una o más de las variables independientes.
- *Menor razón F*. Método para la selección de variables en los análisis por pasos que se basa en maximizar la razón F, calculada a partir de la distancia de Mahalanobis entre los grupos.
- *V de Rao*. Medida de las diferencias entre las medias de los grupos. También se denomina la traza de Lawley-Hotelling. En cada paso, se incluye la variable que maximiza el incremento de la V de Rao. Después de seleccionar esta opción, introduzca el valor mínimo que debe tener una variable para poder incluirse en el análisis.

Criterios. Las alternativas disponibles son **Usar valor de F** y **Usar probabilidad de F**. Introduzca valores para introducir y eliminar variables.

- *Usar valor de F*. Una variable se introduce en el modelo si su valor de F es mayor que el valor de entrada, y se elimina si su valor de F es menor que el valor de Eliminación. La entrada debe ser mayor que la eliminación y ambos valores deben ser positivos. Para introducir más variables en el modelo, disminuya el valor de entrada. Para eliminar más variables del modelo, eleve el valor de eliminación.
- *Usar probabilidad de F*. Una variable se introduce en el modelo si el nivel de significación de su valor de F es menor que el valor de entrada, y se elimina si el nivel de significación de su valor de F es mayor que el valor de Eliminación. La entrada debe ser menor que la eliminación y ambos valores deben ser positivos. Para introducir más variables en el modelo, aumente el valor de entrada. Para eliminar más variables del modelo, disminuya el valor de eliminación.

Representación. Resumen de los pasos muestra los estadísticos para todas las variables después de cada paso; **F para distancias por parejas** muestra una matriz de razones *F* por parejas para cada pareja de grupos.

Análisis discriminante: Clasificar

Probabilidades previas. Esta opción determina si se corrigen los coeficientes de clasificación teniendo en cuenta la información previa sobre la pertenencia a los grupos.

- **Todos los grupos iguales.** Se suponen probabilidades previas iguales para todos los grupos. Esta opción no tiene ningún efecto sobre los coeficientes.
- **Calcular según tamaños de grupos.** Los tamaños de los grupos observados de la muestra determinan las probabilidades previas de la pertenencia a los grupos. Por ejemplo, si el 50% de las observaciones incluidas en el análisis corresponden al primer grupo, el 25% al segundo y el 25% al tercero, se corregirán los coeficientes de clasificación para aumentar la probabilidad de la pertenencia al primer grupo respecto a los otros dos.

Representación. Las opciones de presentación disponibles son: Resultados por casos, Tabla de resumen y Clasificación dejando uno fuera.

- *Resultados para cada caso*. Se muestran, para cada caso, los códigos del grupo real de pertenencia, el grupo pronosticado, las probabilidades posteriores y las puntuaciones discriminantes.
- *Tabla de resumen*. Número de casos correcta e incorrectamente asignados a cada uno de los grupos, basándose en el análisis discriminante. En ocasiones se denomina "Matriz de Confusión".
- *Clasificación dejando uno fuera*. Se clasifica cada caso del análisis mediante la función derivada de todos los casos, excepto el propio caso. También se conoce como método U.

Reemplazar los valores perdidos con la media. Seleccione esta opción para sustituir la media de una variable independiente para un valor perdido sólo durante la fase de clasificación.

Usar matriz de covarianzas. Existe la opción de clasificar los casos utilizando una matriz de covarianzas intra-grupos o una matriz de covarianzas de los grupos separados.

- *Intra-grupos*. Se utiliza la matriz de covarianza intra-grupos combinada para clasificar los casos.
- *Grupos separados*. Para la clasificación se utilizan las matrices de covarianza de los grupos separados. Dado que la clasificación se basa en las funciones discriminantes y no en las variables originales, esta opción no siempre es equivalente a la discriminación cuadrática.

Diagramas. Las opciones de gráficos disponibles son: Grupos combinados, Grupos separados y Mapa territorial.

- *Grupos combinados*. Crea un diagrama de dispersión, con todos los grupos, de los valores en las dos primeras funciones discriminantes. Si sólo hay una función, en su lugar se muestra un histograma.
- *Grupos separados*. Crea diagramas de dispersión, de los grupos por separado, para los valores en las dos primeras funciones discriminantes. Si sólo hay una función, en su lugar se muestra un histograma.
- *Mapa territorial*. Gráfico de las fronteras utilizadas para clasificar los casos en grupos a partir de los valores en las funciones. Los números corresponden a los grupos en los que se clasifican los casos. La media de cada grupo se indica mediante un asterisco situado dentro de sus fronteras. No se mostrará el mapa si sólo hay una función discriminante.

Análisis discriminante: Guardar

Es posible añadir variables nuevas al archivo de datos activo. Las opciones disponibles son las de grupo de pertenencia pronosticado (una única variable), puntuaciones discriminantes (una variable para cada función discriminante en la solución) y probabilidades de pertenencia al grupo según las puntuaciones discriminantes (una variable para cada grupo).

También se puede exportar información del modelo al archivo especificado en formato XML. Puede utilizar este archivo de modelo para aplicar la información del modelo a otros archivos de datos para puntuarlo.

Características adicionales del comando DISCRIMINANT

La sintaxis de comandos también le permite:

- Realizar varios análisis discriminantes (con un comando) y controlar el orden en el que se introducen las variables (mediante el subcomando ANALYSIS).
- Especificar probabilidades previas para la clasificación (mediante el subcomando PRIORS).
- Mostrar matrices de estructura y de configuración rotadas (mediante el subcomando ROTATE).
- Limitar el número de funciones discriminantes extraídas (mediante el subcomando FUNCTIONS).
- Restringir la clasificación a los casos que están seleccionados (o no seleccionados) para el análisis (mediante el subcomando SELECT).
- Leer y analizar una matriz de correlaciones (mediante el subcomando MATRIX).
- Escribir una matriz de correlaciones para su análisis posterior (mediante el subcomando MATRIX).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 22. Análisis factorial

El análisis factorial intenta identificar variables subyacentes, o **factores**, que expliquen la configuración de las correlaciones dentro de un conjunto de variables observadas. El análisis factorial se suele utilizar en la reducción de los datos para identificar un pequeño número de factores que explique la mayoría de la varianza observada en un número mayor de variables manifiestas. También puede utilizarse para generar hipótesis relacionadas con los mecanismos causales o para inspeccionar las variables para análisis subsiguientes (por ejemplo, para identificar la colinealidad antes de realizar un análisis de regresión lineal).

El procedimiento de análisis factorial ofrece un alto grado de flexibilidad:

- Existen siete métodos de extracción factorial disponibles.
- Existen cinco métodos de rotación disponibles, entre ellos el oblimin directo y el promax para rotaciones no ortogonales.
- Existen tres métodos disponibles para calcular las puntuaciones factoriales; y las puntuaciones pueden guardarse como variables para análisis adicionales.

Ejemplo. ¿Qué actitudes subyacentes hacen que las personas respondan a las preguntas de una encuesta política de la manera en que lo hacen? Examinando las correlaciones entre los elementos de la encuesta se deduce que hay una superposición significativa entre los diversos subgrupos de elementos (las preguntas sobre los impuestos tienden a estar correlacionadas entre sí, las preguntas sobre temas militares también están correlacionadas entre sí, y así sucesivamente). Con el análisis factorial, se puede investigar el número de factores subyacentes y, en muchos casos, identificar lo que los factores representan conceptualmente. Adicionalmente, se pueden calcular las puntuaciones factoriales para cada encuestado, que pueden utilizarse en análisis subsiguientes. Por ejemplo, es posible generar un modelo de regresión logística para predecir el comportamiento de voto basándose en las puntuaciones factoriales.

Estadísticos. Para cada variable: número de casos válidos, media y desviación estándar. Para cada análisis factorial: matriz de correlaciones de variables, incluidos niveles de significación, determinante, inversa; matriz de correlaciones reproducida, que incluye anti-imagen; solución inicial (comunalidades, autovalores y porcentaje de varianza explicada); KMO (medida de la adecuación muestral de Kaiser-Meyer-Olkin) y prueba de esfericidad de Bartlett; solución sin rotar, que incluye cargas factoriales, comunalidades y autovalores; y solución rotada, que incluye la matriz de configuración rotada y la matriz de transformación. Para rotaciones oblicuas: las matrices de estructura y de configuración rotadas; matriz de coeficientes para el cálculo de las puntuaciones factoriales y matriz de covarianzas entre los factores. Gráficos: gráfico de sedimentación y gráfico de las cargas de los dos o tres primeros factores.

Análisis factorial: Consideraciones sobre los datos

Datos. Las variables deben ser cuantitativas a nivel de *intervalo* o de *razón*. Los datos categóricos (como la religión o el país de origen) no son adecuados para el análisis factorial. Los datos para los cuales razonablemente se pueden calcular los coeficientes de correlación de Pearson, deberían ser adecuados para el análisis factorial.

Supuestos. Los datos deben tener una distribución normal bivariada para cada pareja de variables y las observaciones deben ser independientes. El modelo de análisis factorial especifica que las variables vienen determinadas por los factores comunes (los factores estimados por el modelo) y por factores exclusivos (los cuales no se superponen entre las distintas variables observadas); las estimaciones calculadas se basan en el supuesto de que ningún factor único está correlacionado con los demás, ni con los factores comunes.

Para obtener un análisis factorial

1. Seleccione en los menús:
   Analizar > **Reducción de dimensiones** > **Factor...**
2. Seleccione las variables para el análisis factorial.

Selección de casos en el análisis factorial

Para seleccionar casos para el análisis:

1. Seleccione una variable de selección.
2. Pulse en **Valor** para introducir un número entero como valor de selección.

En el análisis factorial, sólo se usarán los casos con ese valor para la variable de selección.

Análisis factorial: Descriptivos

Estadísticos. Los descriptores univariados incluyen la media, la desviación estándar y el número de casos válidos para cada variable. La **solución inicial** muestra las comunalidades iniciales, los autovalores y el porcentaje de varianza explicada.

Matriz de correlaciones. Las opciones disponibles son: coeficientes, niveles de significación, determinante, inversa, reproducida, anti-imagen y KMO y prueba de esfericidad de Bartlett.

- *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.
- *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).
- *Anti-imagen*. La matriz de correlaciones anti-imagen contiene los negativos de los coeficientes de correlación parcial y la matriz de covarianza anti-imagen contiene los negativos de las covarianzas parciales. En un buen modelo factorial la mayoría de los elementos no diagonales deben ser pequeños. En la diagonal de la matriz de correlaciones anti-imagen se muestra la medida de adecuación muestral para esa variable.

Análisis factorial: Extracción

Método. Permite especificar el método de extracción factorial. Los métodos disponibles son: Componentes principales, Mínimos cuadrados no ponderados, Mínimos cuadrados generalizados, Máxima verosimilitud, factorización de Ejes principales, factorización Alfa y factorización Imagen.

- *Análisis de componentes principales*. Método para la extracción de factores utilizada para formar combinaciones lineales no correlacionadas de las variables observadas. El primer componente tiene la varianza máxima. Las componentes sucesivas explican progresivamente proporciones menores de la varianza y no están correlacionadas unas con otras. El análisis principal de las componentes se utiliza para obtener la solución factorial inicial. No se puede utilizar cuando una matriz de correlaciones es singular.
- *Método de mínimos cuadrados no ponderados*. Método de extracción de factores que minimiza la suma de los cuadrados de las diferencias entre las matrices de correlación observada y reproducida, ignorando las diagonales.
- *Método de Mínimos cuadrados generalizados*. Método de extracción de factores que minimiza la suma de los cuadrados de las diferencias entre las matrices de correlación observada y reproducida. Las correlaciones se ponderan por el inverso de su exclusividad, de manera que las variables que tengan un valor alto de exclusividad reciban una ponderación menor que aquéllas que tengan un valor bajo de exclusividad.
- *Método de máxima verosimilitud*. Método de extracción factorial que proporciona las estimaciones de los parámetros que con mayor probabilidad ha producido la matriz de correlaciones observada, si la

muestra procede de una distribución normal multivariada. Las correlaciones se ponderan por el inverso de la exclusividad de las variables, y se emplea un algoritmo iterativo.

- *Factorización de ejes principales*. Método para la extracción de factores que parte de la matriz de correlaciones original con los cuadrados de los coeficientes de correlación múltiple insertados en la diagonal principal como estimaciones iniciales de las comunalidades. Las cargas factoriales resultantes se utilizan para estimar de nuevo las comunalidades que reemplazan a las estimaciones previas de comunalidad en la diagonal. Las iteraciones continúan hasta que el cambio en las comunalidades, de una iteración a la siguiente, satisfaga el criterio de convergencia para la extracción.
- *Alfa*. Método de extracción factorial que considera a las variables incluidas en el análisis como una muestra del universo de las variables posibles. Este método maximiza el Alfa de Cronbach para los factores.
- *Factorización imagen*. Método para la extracción de factores, desarrollado por Guttman y basado en la teoría de las imágenes. La parte común de una variable, llamada la imagen parcial, se define como su regresión lineal sobre las restantes variables, en lugar de ser una función de los factores hipotéticos.

Analizar. Permite especificar o una matriz de correlaciones o una matriz de covarianzas.

- **Matriz de correlaciones.** Es útil si las variables de su análisis se miden sobre escalas distintas.
- **Matriz de covarianzas.** Es útil si se desea aplicar el análisis factorial a varios grupos con distintas varianzas para cada variable.

Extraer. Se pueden retener todos los factores cuyos autovalores excedan un valor especificado o retener un número específico de factores.

Representación. Permite solicitar la solución factorial sin rotar y el gráfico de sedimentación de los autovalores.

- *Solución factorial sin rotar*. Muestra las cargas factoriales sin rotar (la matriz de configuración factorial), las comunalidades y los autovalores de la solución factorial.
- *Gráfico de sedimentación*. Gráfico de la varianza que se asocia a cada factor. Este gráfico se utiliza para determinar cuántos factores se deben retenerse. Típicamente el gráfico muestra una clara ruptura entre la pronunciada inclinación de los factores más importantes y el descenso gradual de los restantes (los sedimentos).

Nº máximo de iteraciones para convergencia. Permite especificar el número máximo de pasos que el algoritmo puede seguir para estimar la solución.

Análisis factorial: Rotación

Método. Permite seleccionar el método de rotación factorial. Los métodos disponibles son: varimax, equamax, quartimax, oblimin directo y promax.

- *Método Varimax*. Método de rotación ortogonal que minimiza el número de variables que tienen cargas altas en cada factor. Simplifica la interpretación de los factores.
- *Criterio Oblimin directo*. Método para la rotación oblicua (no ortogonal). Si delta es igual a cero (el valor predeterminado) las soluciones son las más oblicuas. A medida que delta se va haciendo más negativo, los factores son menos oblicuos. Para anular el valor predeterminado 0 para delta, introduzca un número menor o igual que 0,8.
- *Método quartimax*. Método de rotación que minimiza el número de factores necesarios para explicar cada variable. Simplifica la interpretación de las variables observadas.
- *Método Equamax* . Método de rotación que es combinación del método varimax, que simplifica los factores, y el método quartimax, que simplifica las variables. Se minimiza tanto el número de variables que saturan alto en un factor como el número de factores necesarios para explicar una variable.
- *Rotación Promax*. Rotación oblicua que permite que los factores estén correlacionados. Esta rotación se puede calcular más rápidamente que una rotación oblimin directa, por lo que es útil para conjuntos de datos grandes.

Representación. Permite incluir los resultados de la solución rotada, así como los gráficos de las cargas para los dos o tres primeros factores.

- *Solución rotada.* Debe seleccionarse un método de rotación para obtener la solución rotada. Para las rotaciones ortogonales, se muestran la matriz de configuración rotada y la matriz de transformación de factor. Para las rotaciones oblicuas, se muestran las matrices de correlaciones de factor, estructura y patrón.
- *Diagrama de las cargas factoriales*. Representación tridimensional de las cargas factoriales para los tres primeros factores. En una solución de dos factores, se representa un diagrama bidimensional. Si sólo se extrae un factor no se muestra el gráfico. Si se solicita la rotación, los diagramas representan las soluciones rotadas.

Nº máximo de iteraciones para convergencia. Permite especificar el número máximo de pasos que el algoritmo puede seguir para llevar a cabo la rotación.

Análisis factorial: Puntuaciones factoriales

Guardar como variables. Crea una nueva variable para cada factor en la solución final.

Método. Los métodos alternativos para calcular las puntuaciones factoriales son: regresión, Bartlett, y Anderson-Rubin.

- *Método de regresión*. Método para estimar los coeficientes de las puntuaciones factoriales. Las puntuaciones que se producen tienen una media de 0 y una varianza igual al cuadrado de la correlación múltiple entre las puntuaciones factoriales estimadas y los valores factoriales verdaderos. Las puntuaciones puede correlacionarse incluso si los factores son ortogonales.
- *Puntuaciones de Bartlett*. Método para estimar los coeficientes de las puntuaciones factoriales. Las puntuaciones resultantes tienen una media de 0. Se minimiza la suma de cuadrados de los factores exclusivos sobre el rango de las variables.
- *Método de Anderson-Rubin*. Método para calcular los coeficientes para las puntuaciones factoriales; es una modificación del método de Bartlett, que asegura la ortogonalidad de los factores estimados. Las puntuaciones resultantes tienen una media 0, una desviación estándar de 1 y no correlacionan entre sí.

Mostrar matriz de coeficientes de las puntuaciones factoriales. Muestra los coeficientes por los cuales se multiplican las variables para obtener puntuaciones factoriales. También muestra las correlaciones entre las puntuaciones factoriales.

Análisis factorial: Opciones

Valores perdidos. Permite especificar el tratamiento que reciben los valores perdidos. Las selecciones disponibles son: Excluir casos *según lista*, Excluir casos *según pareja* y Reemplazar por la media.

Formato de presentación de los coeficientes. Permite controlar aspectos de las matrices de resultados. Los coeficientes se ordenan por tamaño y se suprimen aquellos cuyos valores absolutos sean menores que el valor especificado.

Características adicionales del comando FACTOR

La sintaxis de comandos también le permite:

- Especificar los criterios de convergencia para la iteración durante la extracción y la rotación.
- Especificar gráficos factoriales rotados individuales.
- Especificar el número de puntuaciones factoriales que se van a guardar.
- Especificar valores diagonales para el método de factorización del eje principal.
- Escribir matrices de correlación o matrices de carga factorial en el disco para su análisis posterior.
- Leer y analizar matrices de correlación o matrices de carga factorial.

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 23. Selección de procedimientos para la agrupación en clústeres

Los análisis de clústeres se pueden realizar mediante los procedimientos de análisis de clústeres en dos fases, jerárquico o de K-medias. Cada uno de estos procedimientos emplea un algoritmo distinto en la creación de clústeres y contiene opciones que no están disponibles en los otros.

Análisis de clústeres en dos fases. En algunas aplicaciones, se puede seleccionar como método el procedimiento Análisis de clústeres en dos fases. Ofrece una serie de características exclusivas que se detallan a continuación:

- Selección automática del número más apropiado de clústeres y medidas para la selección de los distintos modelos de clúster.
- Posibilidad de crear modelos de clúster basados al mismo tiempo en variables categóricas y continuas.
- Posibilidad de guardar el modelo de clúster en un archivo XML externo y, a continuación, leer el archivo y actualizar el modelo de clúster con datos más recientes.

Asimismo, el procedimiento Análisis de clústeres en dos fases puede analizar archivos de datos grandes.

Análisis de clústeres jerárquico. El uso del procedimiento Análisis de clústeres jerárquico se limita a archivos de datos más pequeños (cientos de objetos por agrupar en clústeres) y ofrece una serie de características exclusivas que se detallan a continuación:

- Posibilidad de agrupar en clústeres casos o variables.
- Posibilidad de calcular un rango de soluciones posibles y guardar los clústeres de pertenencia para cada una de dichas soluciones.
- Distintos métodos de formación de clústeres, transformación de variables y medida de disimilaridad entre clústeres.

Siempre que todas las variables sean del mismo tipo, el procedimiento Análisis de clústeres jerárquico podrá analizar variables de intervalo (continuas), de recuento o binarias.

Análisis de clústeres de K-medias. El uso del procedimiento Análisis de clústeres de K-medias se limita a datos continuos y requiere que el usuario especifique previamente el número de clústeres y ofrece una serie de características exclusivas que se detallan a continuación:

- Posibilidad de guardar las distancias desde los centros de los clústeres hasta los distintos objetos.
- Posibilidad de leer los centros de los clústeres iniciales y guardar los centros de los clústeres finales desde un archivo IBM SPSS Statistics externo.

Asimismo, el procedimiento Análisis de clústeres de K-medias puede analizar archivos de datos grandes.

Capítulo 24. Análisis de clústeres en dos fases

El procedimiento Análisis de clústeres en dos fases es una herramienta de exploración diseñada para descubrir las agrupaciones naturales (o clústeres) de un conjunto de datos que, de otra manera, no sería posible detectar. El algoritmo que emplea este procedimiento incluye varias atractivas características que lo hacen diferente de las técnicas de agrupación en clústeres tradicionales:

- **Tratamiento de variables categóricas y continuas.** Al suponer que las variables son independientes, es posible aplicar una distribución normal multinomial conjunta en las variables continuas y categóricas.
- **Selección automática del número de clústeres.** Mediante la comparación de los valores de un criterio de selección del modelo para diferentes soluciones de agrupación en clústeres, el procedimiento puede determinar automáticamente el número óptimo de clústeres.
- **Escalabilidad.** Mediante la generación de un árbol de características de clústeres (CF) que resume los registros, el algoritmo en dos fases puede analizar archivos de datos de gran tamaño.

Ejemplo. Las empresas minoristas y de venta de productos para el consumidor suelen aplicar técnicas de agrupación en clústeres a los datos que describen los hábitos de consumo, sexo, edad, nivel de ingresos, etc. de los clientes. Estas empresas adaptan sus estrategias de desarrollo de productos y de marketing en función de cada grupo de consumidores para aumentar las ventas y el nivel de fidelidad a la marca.

Medida de distancia. Esta opción determina cómo se calcula la similaridad entre dos clústeres.

- **Log-verosimilitud.** La medida de la verosimilitud realiza una distribución de probabilidad entre las variables. Las variables continuas se supone que tienen una distribución normal, mientras que las variables categóricas se supone que son multinomiales. Se supone que todas las variables son independientes.
- **Euclídea.** La medida euclídea es la distancia según una "línea recta" entre dos clústeres. Sólo se puede utilizar cuando todas las variables son continuas.

Número de clústeres. Esta opción permite especificar cómo se va a determinar el número de clústeres.

- **Determinar automáticamente.** El procedimiento determinará automáticamente el número "óptimo" de clústeres, utilizando el criterio especificado en el grupo Criterio de agrupación en clústeres. Si lo desea, introduzca un entero positivo para especificar el número máximo de clústeres que el procedimiento debe tener en cuenta.
- **Especificar número fijo.** Permite fijar el número de clústeres de la solución. Introduzca un número entero positivo.

Recuento de variables continuas. Este grupo proporciona un resumen de las especificaciones acerca de la tipificación de variables continuas realizadas en el cuadro de diálogo Opciones. Consulte el tema “Opciones del análisis de clústeres en dos fases” en la página 112 para obtener más información.

Criterio de agrupación en clústeres. Esta opción determina cómo el algoritmo de agrupación en clústeres determina el número de clústeres. Se puede especificar tanto el criterio de información bayesiano (BIC) como el criterio de información de Akaike (AIC).

Consideraciones sobre los datos para el análisis de clústeres en dos fases

Datos. Este procedimiento trabaja tanto con variables continuas como categóricas. Los casos representan los objetos a agrupar en clústeres y las variables representan los atributos en los que se va a basar la agrupación en clústeres.

Orden de casos. Observe que el árbol de características de clústeres y la solución final pueden depender del orden de los casos. Para minimizar los efectos del orden, ordene los casos aleatoriamente. Puede que

desee obtener varias soluciones distintas con los casos ordenados en distintos órdenes aleatorios para comprobar la estabilidad de una solución determinada. En situaciones en que esto resulta difícil debido a unos tamaños de archivo demasiado grandes, se pueden sustituir varias ejecuciones por una muestra de casos ordenados con distintos órdenes aleatorios.

Supuestos. La medida de la distancia de la verosimilitud supone que las variables del modelo de clúster son independientes. Además, se supone que cada variable continua tiene una distribución normal (de Gauss) y que cada variable categórica tiene una distribución multinomial. Las comprobaciones empíricas internas indican que este procedimiento es bastante robusto frente a las violaciones tanto del supuesto de independencia como de las distribuciones, pero aún así es preciso tener en cuenta hasta qué punto se cumplen estos supuestos.

Utilice el procedimiento de correlaciones bivariadas para comprobar la independencia de dos variables continuas. Utilice el procedimiento de tablas cruzadas para comprobar la independencia de dos variables categóricas. Utilice el procedimiento de medias para comprobar la independencia entre una variable continua y una categórica. Utilice el procedimiento de exploración para comprobar la normalidad de una variable continua. Utilice el procedimiento de prueba de chi-cuadrado para comprobar si una variable categórica tiene especificada una distribución multinomial.

Para obtener un análisis de clústeres en dos fases

1. Seleccione en los menús:

   Analizar > **Clasificar** > **Clúster de bietápico...**
2. Seleccione una o varias variables categóricas o continuas.

Si lo desea, puede:

- Ajustar los criterios utilizados para construir los clústeres.
- Seleccionar los ajustes para el tratamiento del ruido, la asignación de memoria, la tipificación de las variables y la entrada del modelo de clúster.
- Solicitar resultados del visor de modelos.
- Guardar los resultados del modelo en el archivo de trabajo o en un archivo XML externo.

Opciones del análisis de clústeres en dos fases

Tratamiento del valor atípico. Este grupo permite tratar los valores atípicos de manera especial durante la agrupación en clústeres si se llena el árbol de características de los clústeres (CF). El árbol CF se considera lleno si no puede aceptar ningún caso más en un nodo hoja y no hay ningún nodo hoja que se pueda dividir.

- Si selecciona el tratamiento del ruido y el árbol CF se llena, se hará volver a crecer después de colocar los casos existentes en hojas poco densas en una hoja de "ruido". Se considera que una hoja es poco densa si contiene un número de casos inferior a un determinado porcentaje de casos del máximo tamaño de hoja. Tras volver a hacer crecer el árbol, los valores atípicos se colocarán en el árbol CF en caso de que sea posible. Si no es así, se descartarán los valores atípicos.
- Si no selecciona el tratamiento del ruido y el árbol CF se llena, se hará volver a crecer utilizando un umbral del cambio en distancia mayor. Tras la agrupación en clústeres final, los valores que no se puedan asignar a un clúster se considerarán valores atípicos. Al clúster de valores atípicos se le asigna un número de identificación de -1 y no se incluirá en el recuento del número de clústeres.

Asignación de memoria. Este grupo permite especificar la cantidad máxima de memoria en megabytes (MB) que puede utilizar el algoritmo de agrupación en clústeres. Si el procedimiento supera este máximo, utilizará el disco para almacenar la información que no se pueda colocar en la memoria. Especifique un número mayor o igual que 4.

- Consulte con el administrador del sistema si desea conocer el valor máximo que puede especificar en su sistema.

- Si este valor es demasiado bajo, es posible que el algoritmo no consiga obtener el número correcto de clústeres.

Tipificación de variables. El algoritmo de agrupación en clústeres trabaja con variables continuas tipificadas. Todas las variables continuas que no estén tipificadas deben dejarse como variables en la lista Para tipificar. Para ahorrar algún tiempo y trabajo para el ordenador, puede seleccionar todas las variables continuas que ya haya tipificado como variables en la lista Asumidas como tipificadas.

Opciones avanzadas

Criterios de ajuste del árbol CF. Los siguientes ajustes del algoritmo de agrupación en clústeres se aplican específicamente al árbol de características de clústeres (CF) y deberán cambiarse con cuidado:

- **Umbral del cambio en distancia inicial.** Éste es el umbral inicial que se utiliza para hacer crecer el árbol CF. Si se ha insertado una determinada hoja en el árbol CF que produciría una densidad inferior al umbral, la hoja no se dividirá. Si la densidad supera el umbral, se dividirá la hoja.
- **Nº máximo de ramas (por nodo hoja).** Número máximo de nodos hijo que puede tener un nodo hoja.
- **Máxima profundidad de árbol.** Número máximo de niveles que puede tener un árbol CF.
- **Máximo número posible de nodos.** Indica el número máximo de nodos del árbol CF que puede generar potencialmente el procedimiento, de acuerdo con la función $(b^{d+1} - 1) / (b - 1)$, donde b es el número máximo de ramas y d es la profundidad máxima del árbol. Tenga en cuenta que un árbol CF excesivamente grande puede agotar los recursos del sistema y afectar negativamente al rendimiento del procedimiento. Como mínimo, cada nodo requiere 16 bytes.

Actualización del modelo de clúster. Este grupo permite importar y actualizar un modelo de clúster generado en un análisis anterior. El archivo de entrada contiene el árbol CF en formato XML. A continuación, se actualizará el modelo con los datos existentes en el archivo activo. Debe seleccionar los nombres de variable en el cuadro de diálogo principal en el mismo orden en que se especificaron en el análisis anterior. El archivo XML permanecerá inalterado, a no ser que escriba específicamente la nueva información del modelo en el mismo nombre de archivo. Consulte el tema “Resultados de análisis de clústeres en dos fases” para obtener más información.

Si se ha especificado una actualización del modelo de clúster, se utilizarán las opciones pertenecientes a la generación del árbol CF que se especificaron para el modelo original. Concretamente, se utilizarán los ajustes del modelo guardado acerca de la medida de distancia, el tratamiento del ruido, la asignación de memoria y los criterios de ajuste del árbol CF, por lo que se ignorarán todos los ajustes de estas opciones que se hayan especificado en los cuadros de diálogo.

Nota: al realizar una actualización del modelo de clúster, el procedimiento supone que ninguno de los casos seleccionados en el conjunto de datos activo se utilizó para crear el modelo de clúster original. El procedimiento también supone que los casos utilizados en la actualización del modelo proceden de la misma población que los casos utilizados para crear el modelo; es decir, se supone que las medias y las varianzas de las variables continuas y los niveles de las variables categóricas son los mismos en ambos conjuntos de casos. Si los conjuntos de casos "nuevo" y "antiguo" proceden de poblaciones heterogéneas, deberá ejecutar el procedimiento Análisis de clústeres en dos fases para los conjuntos combinados de casos para obtener los resultados óptimos.

Resultados de análisis de clústeres en dos fases

Resultado. Este grupo proporciona opciones para la presentación los resultados de la agrupación en clústeres.

- **Tablas dinámicas.** Los resultados se muestran en tablas dinámicas.
- **Gráficos y tablas en el Visor de modelos.** Los resultados se muestran en el Visor de modelos.

- **Campos de evaluación.** Calcula los datos del clúster de las variables que no se han utilizado en su creación. Los campos de evaluación se pueden mostrar junto con las características de entrada del visor de modelos seleccionándolas en el cuadro de diálogo subordinado Visualización. Los campos con valores perdidos se ignoran.

Archivo de datos de trabajo. Este grupo permite guardar las variables en el conjunto de datos activo.

- **Crear variable del clúster de pertenencia.** Esta variable contiene un número de identificación de clúster para cada caso. El nombre de esta variable es *tsc_n*, donde *n* es un número entero positivo que indica el ordinal de la operación de almacenamiento del conjunto de datos activo realizada por este procedimiento en una determinada sesión.

Archivos XML. El modelo de clúster final y el árbol CF son dos tipos de archivos de resultados que se pueden exportar en formato XML.

- **Exportar modelo final.** También se puede exportar el modelo de clúster final al archivo especificado en formato XML (PMML). Puede utilizar este archivo de modelo para aplicar la información del modelo a otros archivos de datos para puntuarlo.
- **Exportar árbol CF.** Esta opción le permite guardar el estado actual del árbol del clúster y actualizarlo más tarde utilizando datos más nuevos.

El visor de clústeres

Los modelos de clúster se suelen utilizar para buscar grupos (o clústeres) de registros similares basados en las variables examinadas, donde la similitud entre los miembros del mismo grupo es alta y es baja entre miembros de grupos diferentes. Los resultados pueden utilizarse para identificar las asociaciones que, de otra manera, no serían aparentes. Por ejemplo, mediante el análisis de clústeres de preferencias del cliente, de nivel de ingresos y de hábitos de consumo, se podría identificar los tipos de clientes con más probabilidad de responder a una campaña de marketing particular.

Existen dos métodos para interpretar los resultados de una presentación de clústeres:

- Examinar los clústeres para determinar las características exclusivas de cada clúster. *¿Contiene uno de los clústeres todos los socios con un alto nivel de ingresos? ¿Contiene este clúster más registros que otros?*
- Examinar los campos de todos los clústeres para determinar la forma en que los valores se distribuyen en ellos. *¿Determina el nivel de educación la pertenencia a un clúster? ¿Distingue la puntuación de crédito alto entre la pertenencia a un clúster o a otro?*

Puede utilizar las vistas principales y las diferentes vistas vinculadas en el visor de clústeres para obtener una mayor perspectiva que le ayuda a responder a estas preguntas.

Si desea ver información sobre el modelo de clúster, active el objeto Visor de modelos pulsando dos veces sobre él en el visor.

Visor de clústeres

El Visor de clústeres se compone de dos paneles, la vista principal en la parte izquierda y la vista relacionada o auxiliar de la derecha. Hay dos vistas principales:

- Resumen del modelo (predeterminado). Consulte el tema “Vista Resumen del modelo” en la página 115 para obtener más información.
- Clústeres. Consulte el tema “Vista de clústeres” en la página 115 para obtener más información.

Hay cuatro vistas relacionadas/auxiliares:

- Importancia del predictor. Consulte el tema “Vista Importancia del predictor de clústeres” en la página 117 para obtener más información.
- Tamaños de clústeres (predeterminado). Consulte el tema “Vista de tamaños de clústeres” en la página 117 para obtener más información.

- Distribución de casillas. Consulte el tema “Vista Distribución de casillas” en la página 117 para obtener más información.
- Comparación de clústeres. Consulte el tema “Vista Comparación de clústeres” en la página 117 para obtener más información.

Vista Resumen del modelo

La vista Resumen del modelo muestra una instantánea o resumen del modelo de clúster, incluyendo una medida de silueta de la cohesión y separación de clústeres sombreada para indicar resultados pobres, correctos o buenos. Esta instantánea le permite comprobar rápidamente si la calidad es insuficiente, en cuyo caso puede optar por volver al nodo de modelado para cambiar los ajustes del modelo de clúster para producir mejores resultados.

Los resultados serán pobres, correctos o buenos de acuerdo con el trabajo de Kaufman y Rousseeuw (1990) sobre la interpretación de estructuras de clústeres. En la vista Resumen del modelo, un resultado "bueno" indica que los datos reflejan una evidencia razonable o sólida de que existe una estructura de clústeres, de acuerdo con la valoración Kaufman y Rousseeuw; una resultado "correcto" indica que ese evidencia es débil, y un resultado "pobre" significa que, según esa valoración, no hay evidencias obvias.

Las medias de medida de silueta, en todos los registros, (B−A) / max(A,B), donde A es la distancia del registro al centro de su clúster y B es la distancia del registro al centro del clúster más cercano al que no pertenece. Un coeficiente de silueta de 1 podría implicar que todos los casos están ubicados directamente en los centros de sus clústeres. Un valor de -1 significaría que todos los casos se encuentran en los centros de clúster de otro clúster. Un valor de 0 implica, de media, que los casos están equidistantes entre el centro de su propio clúster y el siguiente clúster más cercano.

El resumen incluye una tabla que contiene la siguiente información:

- **Algoritmo.** El algoritmo de agrupación en clústeres utilizado, por ejemplo, "Dos fases".
- **Características de entrada.** El número de campos, también conocidos como **entradas** o **predictores**.
- **Clústeres.** Número de clústeres de la solución.

Vista de clústeres

La vista Clústeres contiene una cuadrícula de clústeres por características que incluye nombres de clústeres, tamaños y perfiles para cada clúster.

Las columnas de la cuadrícula contienen la siguiente información:

- **Clúster.** Números de clústeres creados por el algoritmo.
- **Etiqueta.** Etiquetas aplicadas a cada clúster (está en blanco de forma predeterminada). Pulse dos veces la casilla para introducir una etiqueta que describa el contenido del clúster, por ejemplo "Compradores de automóviles de lujo".
- **Descripción.** Cualquier descripción de los contenidos de los clústeres (está en blanco de forma predeterminada). Pulse dos veces la casilla para introducir una descripción del clúster, por ejemplo "Más de 55 años de edad, profesionales, con ingresos superiores a 100.000 €".
- **Tamaño.** El tamaño de cada clúster como porcentaje de la muestra general del clúster. Cada casilla de tamaño de la cuadrícula muestra una barra vertical que muestra el porcentaje de tamaño del clúster, un porcentaje de tamaño en formato numérico y los recuentos de casos de clúster.
- **Características.** Los predictores o entradas individuales, ordenados por importancia general de forma predeterminada. Si hay columnas con tamaños iguales, se muestran en orden ascendente en función de los miembros del clúster.

  La importancia general de la característica se indica por el color del sombreado del fondo de la casilla, siendo más oscuro cuanto más importante sea la característica. Una guía sobre la tabla indica la importancia vinculada a cada color de casilla de característica.

Cuando pasa el ratón por una casilla, se muestra el nombre completo/etiqueta de la característica y el valor de importancia de la casilla. Es posible que aparezca más información, en función de la vista y tipo

de característica. En la vista Centros de clústeres, esto incluye la estadística de casilla y el valor de la casilla, por ejemplo: "Media: 4,32". En las características categóricas, la casilla muestra el nombre de la categoría (modal) más frecuente y su porcentaje.

En la vista Clústeres, puede seleccionar varias formas de mostrar la información de clústeres:

- Transponer clústeres y características. Consulte el tema "Transponer clústeres y características" para obtener más información.
- Clasificar características. Consulte el tema "Clasificar características" para obtener más información.
- Clasificar clústeres. Consulte el tema "Clasificar clústeres" para obtener más información.
- Seleccionar contenido de casilla. Consulte el tema "Contenido de casilla" para obtener más información.

Transponer clústeres y características: De forma predeterminada, los clústeres se muestran como columnas y las características se muestran como filas. Para invertir esta visualización, pulse el botón **Transponer clústeres y características** a la izquierda de los botones **Clasificar características**. Por ejemplo, puede que desea hacer esto cuando se muestren muchos clústeres para reducir la cantidad de desplazamiento horizontal necesario para visualizar los datos.

Clasificar características: Los botones **Clasificar características por** le permiten seleccionar la cantidad de casillas de características:

- **Importancia global** Este es el orden de clasificación predeterminado. Las características se clasifican en orden descendente de importancia general y el orden de clasificación es el mismo entre los distintos clústeres. Si hay características que empatan en valores de importancia, éstas se muestran en orden de clasificación ascendente según el nombre.
- **Importancia dentro del clúster** Las características se clasifican con respecto de su importancia para cada clúster. Si hay características que empatan en valores de importancia, éstas se muestran en orden de clasificación ascendente según el nombre. Si esta opción está seleccionada, el orden de clasificación suele variar en los diferentes clústeres.
- **Nombre.** Las características se clasifican por nombre en orden alfabético.
- **Orden de los datos** Las características se clasifican por orden en el conjunto de datos.

Clasificar clústeres: De forma predeterminada, los clústeres se clasifican en orden de tamaño descendente. Los botones **Clasificar clústeres por** le permiten ordenarlos por nombre en orden alfabético o, si ha creado etiquetas de exclusivas, por orden de etiqueta alfanumérico.

Las características con la misma etiqueta se clasifican por nombre de clúster. Si los clústeres se clasifican por etiqueta y modifica la etiqueta de un clúster, el orden de clasificación se actualiza automáticamente.

Contenido de casilla: Los botones **Casillas** le permiten cambiar la visualización del contenido de casillas de características y campos de evaluación.

- **Centros de los clústeres.** De forma predeterminada, las casillas muestran nombres/etiquetas de características y la tendencia central para cada combinación de clúster/característica. La media se muestra para los campos continuos y el modo (categoría más frecuente) con porcentaje de categoría para los campos categóricos.
- **Distribuciones absolutas.** Muestra nombres/etiquetas de características y distribuciones absolutas de las características de cada clúster. En el caso de las características categóricas, la visualización muestra gráficos de barras superpuestas con las categorías ordenadas en orden ascendente de valores de datos. En las características continuas, la visualización muestra un gráfico de densidad suave que utiliza los mismos puntos finales e intervalos para cada clúster.

  La visualización en color rojo oscuro muestra la distribución de clústeres, mientras que la más clara representa los datos generales.

- **Distribuciones relativas** Muestra los nombres/etiquetas de características y las distribuciones relativas en las casillas. En general, las visualizaciones son similares a las mostradas para las distribuciones absolutas, sólo que en su lugar se muestran distribuciones relativas.

  La visualización en color rojo oscuro muestra la distribución de clústeres, mientras que la más clara representa los datos generales.
- **Vista básica.** Si hay muchos clústeres, puede resultar difícil ver todos los detalles sin desplazarse. Para reducir la cantidad de desplazamiento, seleccione esta vista para cambiar la visualización a una versión más compacta de la tabla.

Vista Importancia del predictor de clústeres

La vista Importancia del predictor muestra la importancia relativa de cada campo en la estimación del modelo.

Vista de tamaños de clústeres

La vista Tamaños de clústeres muestra el gráfico circular que contiene cada clúster. El tamaño de porcentaje de cada clúster se muestra en cada sector, pase el ratón sobre cada porción para mostrar el recuento de esa porción.

Bajo el gráfico, una tabla enumera la siguiente información de tamaño:

- El tamaño del clúster más pequeño (un recuento y porcentaje del conjunto).
- El tamaño del clúster mayor (un recuento y porcentaje del conjunto).
- La proporción entre el tamaño del mayor clúster y el del menor.

Vista Distribución de casillas

La vista Distribución de casillas muestra un gráfico expandido y más detallado de la distribución de los datos para cualquier casilla que seleccione en el panel principal Clústeres.

Vista Comparación de clústeres

La vista Comparación de clústeres se compone de un diseño en estilo de cuadrícula, con características en las filas y clústeres seleccionados en las columnas. Esta vista le ayuda a entender mejor los factores de los que se componen los clústeres, y le permite ver las diferencias entre los clústeres no sólo con respecto a los datos generales, sino entre sí.

Para seleccionar clústeres para su visualización, pulse en la parte superior de la columna del clúster en el panel principal Clústeres. Pulse las teclas Ctrl o Mayús y pulse para seleccionar o cancelar la selección de más de un clúster para su comparación.

Nota: puede seleccionar que se muestren hasta cinco clústeres.

Los clústeres se muestran en el orden en que se seleccionaron, mientras que el orden de los campos viene determinado por la opción **Clasificar características por**. Si selecciona **Importancia dentro del clúster**, los campos siempre se clasifican por importancia general.

Los gráficos de fondo muestran las distribuciones generales de cada característica:

- Las características categóricas aparecen como gráficos de puntos, donde el tamaño del punto indica la categoría más frecuente/modal para cada clúster (por característica).
- Las características continuas se muestran como diagramas de caja, que muestran las medianas globales y los rangos intercuartiles.

En estas vistas de fondo aparecen superpuestos diagramas de caja para los clústeres seleccionados:

- En las características continuas hay marcadores de puntos cuadrados y líneas horizontales que indican el rango de mediana e intercuartil de cada clúster.
- Cada clúster viene representado por un color distinto, que se muestra en la parte superior de la vista.

Navegación en el Visor de clústeres

El visor de clústeres es una pantalla interactiva. Puede:

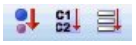
- Seleccionar un campo o clúster para ver más detalles.
- Comparar clústeres para seleccionar elementos de interés.
- Alterar la visualización.
- Transponer ejes.

Uso de las barras de herramientas

Puede controlar la información que aparece en los paneles izquierdo y derecho mediante las opciones de la barra de herramientas. Puede cambiar la orientación de la pantalla (de arriba a abajo, de izquierda a derecha, o de derecha a izquierda) mediante los controles de la barra de herramientas. Además, también puede restablecer el visor a los ajustes predeterminados, y abrir un cuadro de diálogo para especificar el contenido de la vista Clústeres en el panel principal.

Las opciones **Clasificar características por**, **Clasificar clústeres por**, **Casillas** y **Mostrar** sólo están disponibles cuando selecciona la vista **Clústeres** en el panel principal. Consulte el tema "Vista de clústeres" en la página 115 para obtener más información.

Tabla 2. Iconos de barra de herramientas.

Icono	Tema
	Consulte Transponer clústeres y características
	Consulte Clasificar características por
	Consulte Clasificar clústeres por
	Consulte Casillas

Control de la vista Clústeres

Para controlar qué se muestra en la vista Clústeres del panel principal, pulse el botón **Mostrar** y se abrirá el cuadro de diálogo Mostrar.

Características. Está seleccionado de forma predeterminada. Para ocultar todas las características de entrada, cancele la selección de la casilla de verificación.

Campos de evaluación Seleccione los campos de evaluación (campos que no se usan para crear el modelo de clúster, sino que se envían al visor de modelos para evaluar los clústeres) que desea mostrar, ya que ninguno se muestra de forma predeterminada. *Nota* El campo de evaluación debe ser una cadena con más de un valor. Esta casilla de verificación no está disponible si no hay ningún campo de evaluación disponible.

Descripciones de clústeres Está seleccionado de forma predeterminada. Para ocultar todas las casillas de descripción de clúster, cancele la selección de la casilla de verificación.

Tamaños de clústeres Está seleccionado de forma predeterminada. Para ocultar todas las casillas de tamaño de clúster, cancele la selección de la casilla de verificación.

Número máximo de categorías Especifique el número máximo de categorías que se mostrarán en gráficos de características categóricas. El valor predeterminado es 20.

Filtrado de registros

Si desea obtener más información sobre los casos de un determinado clúster o grupo de clústeres, puede seleccionar un subconjunto de registros para realizar un análisis más detallado en los clústeres seleccionados.

1. Seleccione los clústeres en la vista Clúster del Visor de clústeres. Mantenga pulsada la tecla Ctrl al mismo tiempo que pulsa el botón del ratón para seleccionar varios clústeres.
2. Seleccione en los menús:

   Generar > **Filtrar registros...**
3. Introduzca un nombre de variable de filtro. Los registros de los clústeres seleccionados recibirán un valor igual a 1 para este campo. Todos los demás registros recibirán un valor igual a 0 y se excluirán de los análisis subsiguientes hasta que se modifique el estado del filtro.
4. Pulse en **Aceptar**.

Capítulo 25. Análisis de clústeres jerárquico

Este procedimiento intenta identificar grupos relativamente homogéneos de casos (o de variables) basándose en las características seleccionadas, mediante un algoritmo que comienza con cada caso (o cada variable) en un clúster diferente y combina los clústeres hasta que sólo queda uno. Es posible analizar las variables brutas o elegir de entre una variedad de transformaciones de estandarización. Las medidas de distancia o similaridad se generan mediante el procedimiento Proximidades. Los estadísticos se muestran en cada etapa para ayudar a seleccionar la mejor solución.

Ejemplo. ¿Existen grupos identificables de programas televisivos que atraigan a audiencias similares dentro de cada grupo? Con el análisis de clústeres jerárquico, podría agrupar los programas de TV (los casos) en grupos homogéneos basados en las características del espectador. Esto se puede utilizar para identificar segmentos de mercado. También puede agrupar ciudades (los casos) en grupos homogéneos, de manera que se puedan seleccionar ciudades comparables para probar diversas estrategias de marketing.

Estadísticos. Historial de conglomeración, matriz de distancias (o similaridades) y pertenencia a los clústeres para una solución única o una serie de soluciones. Gráficos: dendrogramas y diagramas de témpanos.

Análisis de clústeres jerárquico: Consideraciones sobre los datos

Datos. Las variables pueden ser cuantitativas, binarias o datos de recuento. El escalamiento de las variables es un aspecto importante, ya que las diferencias en el escalamiento pueden afectar a las soluciones en clústeres. Si las variables muestran grandes diferencias en el escalamiento (por ejemplo, una variable se mide en dólares y la otra se mide en años), debería considerar la posibilidad de estandarizarlas (esto puede llevarse a cabo automáticamente mediante el propio procedimiento Análisis de clústeres jerárquico).

Orden de casos. Si hay distancias empatadas o similitudes en los datos de entrada o si éstas se producen entre los clústeres actualizados durante la unión, la solución de clúster resultante puede depender del orden de los casos del archivo. Puede que desee obtener varias soluciones distintas con los casos ordenados en distintos órdenes aleatorios para comprobar la estabilidad de una solución determinada.

Supuestos. Las medidas de distancia o similaridad empleadas deben ser adecuadas para los datos analizados (véase el procedimiento Proximidades para obtener más información sobre la elección de las medidas de distancia y similaridad). Asimismo, debe incluir todas las variables relevantes en el análisis. Si se omiten variables de interés la solución obtenida puede ser equívoca. Debido a que el análisis de clústeres jerárquico es un método exploratorio, los resultados deben considerarse provisionales hasta que sean confirmados mediante otra muestra independiente.

Para obtener un análisis de clústeres jerárquico

1. Seleccione en los menús:

   Analizar > **Clasificar** > **Clústeres jerárquicos...**
2. Si está aglomerando casos, seleccione al menos una variable numérica. Si está aglomerando variables, seleccione al menos tres variables numéricas.

Si lo desea, puede seleccionar una variable de identificación para etiquetar los casos.

Análisis de clústeres jerárquico: Método

Método de agrupación en clústeres. Las opciones disponibles son: Vinculación inter-grupos, Vinculación intra-grupos, Vecino más próximo, Vecino más lejano, Agrupación de centroides, Agrupación de medianas y Método de Ward.

Medida. Permite especificar la medida de distancia o similaridad que será empleada en la aglomeración. Seleccione el tipo de datos y la medida de distancia o similaridad adecuada:

- **Intervalo.** Distancia euclídea, Distancia euclídea al cuadrado, Coseno, Correlación de Pearson, Chebychev, Bloque, Minkowski y Personalizada.
- **Recuentos.** Las opciones disponibles son: Medida de chi-cuadrado y Medida de phi-cuadrado.
- **Binaria.** Las opciones disponibles son: Distancia euclídea, Distancia euclídea al cuadrado, Diferencia de tamaño, Diferencia de configuración, Varianza, Dispersión, Forma, Concordancia simple, Correlación phi de 4 puntos, Lambda, *D* de Anderberg, Dice, Hamann, Jaccard, Kulczynski 1, Kulczynski 2, Lance y Williams, Ochiai, Rogers y Tanimoto, Russel y Rao, Sokal y Sneath 1, Sokal y Sneath 2, Sokal y Sneath 3, Sokal y Sneath 4, Sokal y Sneath 5, *Y* de Yule y *Q* de Yule.

Transformar valores. Permite estandarizar los valores de los datos, para los casos o las variables, antes de calcular las proximidades (no está disponible para datos binarios). Los métodos disponibles de estandarización son: Puntuaciones *z*, Rango -1 a 1, Rango 0 a 1, Magnitud máxima de 1, Media de 1 y Desviación estándar 1.

Transformar medidas. Permite transformar los valores generados por la medida de distancia. Se aplican después de calcular la medida de distancia. Las opciones disponibles son: Valores absolutos, Cambiar el signo y Cambiar la escala al rango 0–1.

Análisis de clústeres jerárquico: Estadísticos

Historial de conglomeración. Muestra los casos o clústeres combinados en cada etapa, las distancias entre los casos o los clústeres que se combinan, así como el último nivel del proceso de aglomeración en el que cada caso (o variable) se unió a su clúster correspondiente.

Matriz de proximidades. Proporciona las distancias o similaridades entre los elementos.

Clúster de pertenencia. Muestra el clúster al cual se asigna cada caso en una o varias etapas de la combinación de los clústeres. Las opciones disponibles son: Solución única y Rango de soluciones.

Análisis de clústeres jerárquico: Gráficos

Dendrograma. Muestra un *dendrograma*. Los dendrogramas pueden emplearse para evaluar la cohesión de los clústeres que se han formado y proporcionar información sobre el número adecuado de clústeres que deben conservarse.

Témpanos. Muestra un *diagrama de témpanos*, que incluye 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.

Análisis de clústeres jerárquico: Guardar variables nuevas

Clúster de pertenencia. Permite guardar los clústeres de pertenencia para una solución única o un rango de soluciones. Las variables guardadas pueden emplearse en análisis posteriores para explorar otras diferencias entre los grupos.

Características adicionales de la sintaxis de comandos CLUSTER

El procedimiento Clúster jerárquico utiliza la sintaxis de comandos `CLUSTER`. La sintaxis de comandos también le permite:

- Utilizar varios métodos de agrupación en un único análisis.
- Leer y analizar una matriz de proximidades.
- Escribir una matriz de distancias para su análisis posterior.
- Especificar cualquier valor para la potencia y la raíz en la medida de distancia personalizada (potencia).
- Especificar nombres para variables guardadas.

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 26. Análisis de clústeres de K-medias

Este procedimiento intenta identificar grupos de casos relativamente homogéneos basándose en las características seleccionadas y utilizando un algoritmo que puede gestionar un gran número de casos. Sin embargo, el algoritmo requiere que el usuario especifique el número de clústeres. Puede especificar los centros iniciales de los clústeres si conoce de antemano dicha información. Puede elegir uno de los dos métodos disponibles para clasificar los casos: la actualización de los centros de los clústeres de forma iterativa o sólo la clasificación. Asimismo, puede guardar la pertenencia a los clústeres, información de la distancia y los centros de los clústeres finales. Si lo desea, puede especificar una variable cuyos valores sean utilizados para etiquetar los resultados por casos. También puede solicitar los estadísticos *F* de los análisis de varianza. Aunque estos estadísticos son oportunistas (ya que el procedimiento trata de formar grupos que de hecho difieran), el tamaño relativo de los estadísticos proporciona información acerca de la contribución de cada variable a la separación de los grupos.

Ejemplo. ¿Cuáles son los grupos identificables de programas de televisión que atraen audiencias parecidas dentro de cada grupo? Con el análisis de clústeres de *k*-medias, podría agrupar los programas de televisión (los casos) en *k* grupos homogéneos, basados en las características del televidente. Este proceso se puede utilizar para identificar segmentos de mercado. También puede agrupar ciudades (los casos) en grupos homogéneos, de manera que se puedan seleccionar ciudades comparables para probar diversas estrategias de marketing.

Estadísticos. Solución completa: centros iniciales de los clústeres, tabla de ANOVA. Cada caso: información del clúster, distancia desde el centro del clúster.

Análisis de clústeres de K-medias: Consideraciones sobre los datos

Datos. Las variables deben ser cuantitativas en el nivel de intervalo o de razón. Si las variables son binarias o recuentos, utilice el procedimiento Análisis de clústeres jerárquicos.

Orden de casos y centro de clústeres iniciales. El algoritmo predeterminado para elegir centros de clústeres iniciales no es invariable con respecto a la ordenación de casos. La opción **Usar medias actualizadas** del cuadro de diálogo Iterar hace que la solución resultante dependa potencialmente del orden de casos con independencia de cómo se eligen los centros de clústeres iniciales. Si va a utilizar alguno de estos métodos, puede que desee obtener varias soluciones distintas con los casos ordenados en distintos órdenes aleatorios para comprobar la estabilidad de una solución determinada. La especificación de los centros de clústeres iniciales y la no utilización de la opción **Usar medias actualizadas** evita los problemas relacionados con el orden de casos. No obstante, la ordenación de los centros de clústeres iniciales puede afectar a la solución en caso de haber distancias empatadas desde los casos a los centros de clústeres. Para evaluar la estabilidad de una solución determinada, puede comparar los resultados de los análisis con las distintas permutaciones de los valores de centros iniciales.

Supuestos. Las distancias se calculan utilizando la distancia euclídea simple. Si desea utilizar otra medida de distancia o de similaridad, utilice el procedimiento Análisis de clústeres jerárquicos. El escalamiento de variables es una consideración importante. Si sus variables utilizan diferentes escalas (por ejemplo, una variable se expresa en dólares y otra, en años), los resultados podrían ser equívocos. En estos casos, debería considerar la estandarización de las variables antes de realizar el análisis de clústeres de *k*-medias (esta tarea se puede hacer en el procedimiento Descriptivos). Este procedimiento supone que ha seleccionado el número apropiado de clústeres y que ha incluido todas las variables relevantes. Si ha seleccionado un número inapropiado de clústeres o ha omitido variables relevantes, los resultados podrían ser equívocos.

Para obtener un análisis de clústeres de K-medias

1. Seleccione en los menús:

Analizar > **Clasificar** > **Clúster de K medias...**

2. Seleccione las variables que se van a utilizar en el análisis de clústeres.
3. Especifique el número de clústeres. (Este número no debe ser inferior a 2 ni superior al número de casos del archivo de datos.)
4. Seleccione **Iterar y clasificar** o **Sólo clasificar**.
5. Si lo desea, seleccione una variable de identificación para etiquetar los casos.

Eficacia del análisis de clústeres de K-medias

El comando de análisis de clústeres de *k*-medias es eficaz principalmente porque no calcula las distancias entre todos los pares de casos, como hacen muchos algoritmos de agrupación en clústeres, como el utilizado por el comando de agrupación en clústeres jerárquica.

Para conseguir la máxima eficacia, tome una muestra de los casos y seleccione el método **Iterar y clasificar** para determinar los centros de los clústeres. Seleccione **Escribir finales en**. A continuación, restaure el archivo de datos completo, seleccione el método **Sólo clasificar** y seleccione **Leer iniciales de** para clasificar el archivo completo utilizando los centros estimados a partir de la muestra. Se puede escribir y leer desde un archivo o conjunto de datos. Los conjuntos de datos están disponibles para su uso posterior durante la misma sesión, pero no se guardarán como archivos a menos que se hayan guardado explícitamente antes de que finalice la sesión. El nombre de un conjunto de datos debe cumplir las normas de denominación de variables. Consulte el tema para obtener más información.

Análisis de clústeres de K-medias: Iterar

Nota: estas opciones sólo están disponibles si se selecciona el método **Iterar y clasificar** en el cuadro de diálogo Análisis de clústeres de K-medias.

Nº máximo de iteraciones. Limita el número de iteraciones en el algoritmo *k*-medias. La iteración se detiene después de este número de iteraciones, incluso si no se ha satisfecho el criterio de convergencia. Este número debe estar entre el 1 y el 999.

Para reproducir el algoritmo utilizado por el comando Quick Cluster en las versiones previas a la 5.0, establezca **Máximo de iteraciones** en 1.

Criterio de convergencia. Determina cuándo cesa la iteración. Representa una proporción de la distancia mínima entre los centros iniciales de los clústeres, por lo que debe ser mayor que 0 pero no mayor que 1. Por ejemplo, si el criterio es igual a 0,02, la iteración cesará si una iteración completa no mueve ninguno de los centros de los clústeres en una distancia superior al dos por ciento de la distancia menor entre cualquiera de los centros iniciales.

Usar medias actualizadas. Permite solicitar la actualización de los centros de los clústeres tras la asignación de cada caso. Si no selecciona esta opción, los nuevos centros de los clústeres se calcularán después de la asignación de todos los casos.

Análisis de clústeres de K-medias: Guardar

Puede guardar información sobre la solución como nuevas variables para que puedan ser utilizadas en análisis subsiguientes:

Clúster de pertenencia. Crea una nueva variable que indica el clúster final al que pertenece cada caso. Los valores de la nueva variable van desde el 1 hasta el número de clústeres.

Distancia desde centro del clúster. Crea una nueva variable que indica la distancia euclídea entre cada caso y su centro de clasificación.

Análisis de clústeres de K-medias: Opciones

Estadísticos. Puede seleccionar los siguientes estadísticos: Centros de clústeres iniciales, Tabla de ANOVA e Información del clúster para cada caso.

- *Centros de clústeres iniciales*. Primera estimación de las medias de las variables para cada uno de los clústeres. De forma predeterminada, se selecciona entre los datos un número de casos debidamente espaciados igual al número de clústeres. Los centros iniciales de los clústeres se utilizan como criterio para una primera clasificación y, a partir de ahí, se van actualizando.
- *tabla de ANOVA*. Muestra una tabla de análisis de varianza que incluye las pruebas F univariadas para cada variable de aglomeración. Las pruebas F son sólo descriptivas y las probabilidades resultantes no se deben interpretar. La tabla de ANOVA no se mostrará si se asignan todos los casos a un único clúster.
- *Información del clúster para cada caso*. Muestra, para cada caso, el clúster final asignado y la distancia euclídea entre el caso y el centro del clúster utilizado para clasificar el caso. También muestra la distancia euclídea entre los centros de los clústeres finales.

Valores perdidos. Las opciones disponibles son: **Excluir casos según lista** o **Excluir casos según pareja**.

- **Excluir casos según lista.** Excluye los casos con valores perdidos para cualquier variable de agrupación del análisis.
- **Excluir casos según pareja.** Asigna casos a los clústeres en función de las distancias que se calculan desde todas las variables con valores no perdidos.

Características adicionales del comando QUICK CLUSTER

El procedimiento de clústeres de K-medias utiliza la sintaxis de comandos `QUICK CLUSTER`. La sintaxis de comandos también le permite:

- Aceptar los primeros casos *k* como primeros centros de clústeres iniciales y, por lo tanto, evitar la lectura de datos que normalmente se utiliza para calcularlos.
- Especificar los centros de clústeres iniciales directamente como parte de la sintaxis de comandos.
- Especificar nombres para variables guardadas.

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 27. Pruebas no paramétricas

Las pruebas no paramétricas hacen supuestos mínimos acerca de la distribución subyacente de los datos. Las pruebas que están disponibles en estos cuadros de diálogo, se pueden agrupar en tres categorías amplias en función de cómo se organizan los datos:

- Una prueba para una muestra analiza un campo.
- Una prueba para muestras relacionadas compara dos o más campos para el mismo conjunto de casos.
- Una prueba para muestras independientes analiza un campo que se agrupa por categorías de otro campo.

Pruebas no paramétricas para una muestra

Una prueba no paramétricas para una muestra identifica diferencias en campos únicos mediante una o más pruebas no paramétricas. Las pruebas no paramétricas no dan por hecho que sus datos sigan la distribución normal.

¿Cuál es su objetivo? Los objetivos le permiten especificar rápidamente ajustes de prueba diferentes y comunes.

- **Comparar automáticamente datos observados con el valor hipotetizado.** Este objetivo aplica la prueba binomial a campos categóricos con sólo dos categorías, la prueba de chi-cuadrado al resto de campos categóricos y la prueba de Kolmogorov-Smirnov a campos continuos.
- **Probar la aleatoriedad de la secuencia.** Este objetivo utiliza la prueba de rachas para comprobar la aleatoriedad de la secuencia observada de valores de datos.
- **Análisis personalizado.** Seleccione esta opción si desea modificar manualmente la configuración de la prueba de la pestaña Configuración. Tenga en cuenta que esta configuración se selecciona automáticamente si realiza cambios posteriores a muchas opciones de la pestaña Configuración que sean incompatibles con los del objetivo seleccionado actualmente.

Para obtener Pruebas no paramétricas para una muestra

Seleccione en los menús:

Analizar > **Pruebas no paramétricas** > **Una muestra...**

1. Pulse en **Ejecutar**.

Si lo desea, puede:

- Especifique un objetivo en la pestaña Objetivos.
- Especifique asignaciones de campo en la pestaña Campos.
- Especifique la configuración de experto en la pestaña Configuración.

Pestaña Campos

La pestaña Campos especifica los campos que se deben comprobar.

Utilizar papeles predefinidos. Esta opción utiliza información de campos existentes. Todos los campos con un papel predefinido como Entrada o Ambos se utilizarán como campos de prueba. Al menos un campo de prueba es necesario.

Utilizar asignaciones de campos personalizadas. Esta opción le permite sobrescribir papeles de campos. Después de seleccionar esta opción, especifique los campos siguientes.

- **Campos de prueba.** Seleccione uno o más campos de prueba.

Pestaña Configuración

La pestaña Configuración contiene diferentes grupos de ajustes que puede modificar para ajustar con precisión la forma en que el algoritmo procesa sus datos. Si realiza algún cambio en la configuración predeterminada que sea incompatible con el objetivo seleccionado actualmente, la pestaña Objetivo se actualiza automáticamente para seleccionar la opción **Personalizar análisis**.

Seleccionar pruebas

Estos ajustes especifican las pruebas que realizarán en los campos especificados en la pestaña Campos.

Seleccione automáticamente las pruebas en función de los datos. Esta configuración aplica la prueba binomial a campos categóricos con sólo dos categorías válidas (sin valores perdidos), la prueba de chi-cuadrado al resto de campos categóricos y la prueba de Kolmogorov-Smirnov a campos continuos.

Personalizar pruebas. Esta configuración permite especificar las pruebas que se ejecutarán.

- **Comparar la probabilidad binaria observada con el valor hipotetizado (prueba binomial).** La prueba binomial se puede aplicar a todos los campos. Produce una prueba de una muestra que comprueba si la distribución observada de un campo de distintivo (un campo categórico con sólo dos categorías) es el mismo que lo que se espera de una distribución binomial especificada. Además, puede solicitar intervalos de confianza. Consulte "Opciones de prueba binomiales" para obtener más información sobre la configuración de prueba.
- **Comparar las probabilidades observadas con el valor hipotetizado (prueba de chi-cuadrado).** La prueba de chi-cuadrado se aplica a campos nominales y ordinales. Produce una prueba de una muestra que calcula un estadístico chi-cuadrado basado en las diferencias entre las frecuencias observadas y esperadas de las categorías de un campo. Consulte "Opciones de prueba de chi-cuadrado" en la página 131 para obtener más información sobre la configuración de prueba.
- **Probar la distribución observada con el valor hipotetizado (prueba de Kolmogorov-Smirnov).** La prueba de Kolmogorov-Smirnov se aplica a campos continuos y ordinales. Produce una prueba de una muestra de si la función de distribución acumulada de muestra de un campo es homogénea con una distribución uniforme, normal, Poisson o exponencial. Consulte "Opciones de Kolmogorov-Smirnov" en la página 131 para obtener más información sobre la configuración de prueba.
- **Comparar mediana con el valor hipotetizado (prueba de Wilcoxon de los rangos con signo).** La prueba de Wilcoxon de los rangos con signo se aplica a los campos continuos y ordinales. Produce una prueba para una muestra del valor de mediana de un campo. Especifique un número como la mediana hipotetizada.
- **Probar la aleatoriedad de la secuencia (prueba de rachas).** La prueba de rachas se aplica a todos los campos. Produce una prueba de una muestra de si la secuencia de valores de un campo de dicotomías es aleatoria. Consulte "Prueba de rachas: Opciones" en la página 131 para obtener más información sobre la configuración de prueba.

Opciones de prueba binomiales: La prueba binomial está diseñada para campos de distintivo (campos categóricos con sólo dos categorías), pero se aplica a todos los campos mediante reglas para definir "éxito".

Proporción hipotetizada. Especifica la proporción esperada de registros definidos como "éxitos", o *p*. Especifique un valor mayor que 0 y menor que 1. El valor predeterminado es 0,5.

Intervalo de confianza. Los siguientes métodos permiten calcular intervalos de confianza de datos binarios:

- **Clopper-Pearson (exacto).** Un intervalo exacto basado en la distribución binomial acumulada.
- **Jeffreys.** Un intervalo Bayesian basado en la distribución posterior de *p* que utiliza la opción Jeffreys anterior.
- **Razón de verosimilitud.** Un intervalo basado en la función de verosimilitud para *p*.

Definir éxito para campos categóricos. Especifica cómo se define "éxito", el valor(s) de datos se comprueba en la proporción hipotetizada, en los campos categóricos.

- **Utilizar primera categoría encontrada en los datos** realiza la prueba binomial que utiliza el primer valor encontrado en la muestra para definir "éxito". Esta opción sólo es aplicable a los campos nominal u ordinal con sólo dos valores; el resto de campos categóricos especificados en la pestaña Campos en los que se utiliza esta opción no se comprobarán. Este es el valor predeterminado.
- **Especificar valores de éxito** realiza la prueba binomial que utiliza la lista especificada de valores para definir "éxito". Especifique una lista de valores de cadena o numérico. No es necesario que los valores de la lista estén en la muestra.

Definir éxito para campos continuos. Especifica cómo se define "éxito", el valor(s) de datos se comprueba en el valor de prueba, en los campos categóricos. Éxito se define como valores iguales o menores que un punto de corte.

- **Punto medio de muestra** define el punto de corte en la media de los valores mínimo o máximo.
- **Punto de corte personalizado** permite especificar un valor para el punto de corte.

Opciones de prueba de chi-cuadrado: Todas las categorías tienen la misma probabilidad. Produce la misma frecuencia entre todas las categorías en la muestra. Este es el método predeterminado.

Personalizar probabilidad esperada. Permite especificar frecuencias desiguales para una lista de categorías específica. Especifique una lista de valores de cadena o numérico. No es necesario que los valores de la lista estén en la muestra. En la columna **Categoría**, especifique los valores de categorías. En la columna **Frecuencia relativa**, especifique un valor superior a 0 para cada categoría. Las frecuencias personalizadas se consideran porcentajes de forma que, por ejemplo, especificar frecuencias de 1, 2 y 3 es equivalente a especificar frecuencias de 10, 20 y 30, y especificar que 1/6 de los registros se esperan en la primera categoría, 1/3 en la segunda y 1/2 en la tercera. Si se especifican probabilidades esperadas personalizadas, los valores de categorías personalizadas deben incluir todos los valores de campo de los datos; de lo contrario la prueba no se realiza en ese campo.

Opciones de Kolmogorov-Smirnov: Este cuadro de diálogo especifica las distribuciones que se deben comprobar y los parámetros de las distribución hipotetizada.

Normal. Utilizar datos muestrales utiliza la media observada y la desviación estándar, **Personalizado** le permite especificar valores.

Uniforme. Utilizar datos muestrales utiliza los valores observados mínimos y máximos, **Personalizado** le permite especificar valores.

Exponential. Media muestral utiliza la media observada, **Personalizado** le permite especificar valores.

Poisson. Media muestral utiliza la media observada, **Personalizado** le permite especificar valores.

Prueba de rachas: Opciones: La prueba de rachas está diseñada para campos de distintivo (campos categóricos con sólo dos categorías), pero se puede aplica a todos los campos mediante reglas para definir grupos.

Definir grupos para campos categóricos. Se encuentran disponibles las siguientes opciones:

- **Sólo hay 2 categorías en la muestra** realiza la prueba de rachas utilizando los valores encontrados en la muestra para definir los grupos. Esta opción sólo es aplicable a los campos nominal u ordinal con sólo dos valores; el resto de campos categóricos especificados en la pestaña Campos en los que se utiliza esta opción no se comprobarán.
- **Recodificar datos en 2 categorías** realiza la prueba de rachas utilizando la lista de valores especificada para definir uno de los grupos. El resto de valores de muestra definen el otro grupo. No es necesario que todos los valores de la lista estén presentes en la muestra, pero debe haber al menos un registro en cada grupo.

Definir punto de corte para campos continuos. Especifica cómo se definen los grupos para campos continuos. El primer grupo se define como valores iguales o menores que un punto de corte.

- **Mediana muestral** define el punto de corte en la mediana muestral.
- **Media muestral** define el punto de corte en la media muestral.
- **Personalizado** permite especificar un valor para el punto de corte.

Opciones de prueba

Nivel de significación. Especifica el nivel de significación (alfa) de todas las pruebas. Especifica un valor numérico entre 0 y 1. 0,05 es el valor predeterminado.

Intervalo de confianza (%). Esto especifica el nivel de confianza de todos los intervalos de confianza producidos. Especifique un valor numérico entre 0 y 100. El valor predeterminado es 95.

Casos excluidos. Especifica cómo se determinan las pruebas caso por caso.

- **Excluir casos según lista** significa que los registros con valores perdidos de cualquier campo que se nombran en la pestaña Campos se excluyen de todos los análisis.
- **Excluir casos según prueba** significa que los registros con valores perdidos para un campo que se utiliza para una prueba específica se omiten de esa prueba. Si se realizan varias pruebas en el análisis, cada prueba se evalúa por separado.

Valores perdidos del usuario

Valores perdidos del usuario para campos categóricos. Para que un registro se incluya en el análisis, los campos categóricos deben tener valores válidos para dicho caso. Estos controles permiten decidir si los valores perdidos del usuario se deben tratar como válidos entre los campos categóricos. Los valores perdidos del sistema y los valores perdidos de campos continuos siempre se tratan como no válidos.

Características adicionales del comando NPTESTS

La sintaxis de comandos también le permite:

- Especifique pruebas de una muestra, muestras independientes y muestras relacionadas en una única ejecución del procedimiento.

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Pruebas no paramétricas para muestras independientes

Las pruebas no paramétricas para muestras independientes identifican diferencias entre dos o más grupos utilizando una o más pruebas no paramétricas. Las pruebas no paramétricas no dan por hecho que sus datos sigan la distribución normal.

¿Cuál es su objetivo? Los objetivos le permiten especificar rápidamente ajustes de prueba diferentes y comunes.

- **Comparar automáticamente distribuciones entre grupos.** Este objetivo aplica la prueba U de Mann-Whitney para datos con 2 grupos o la prueba ANOVA de 1 factor de Kruskal-Wallis para datos con grupos *k* .
- **Comparar medianas entre grupos.** Este objetivo utiliza la prueba de la mediana para comparar las medianas observadas entre grupos.
- **Análisis personalizado.** Seleccione esta opción si desea modificar manualmente la configuración de la prueba de la pestaña Configuración. Tenga en cuenta que esta configuración se selecciona automáticamente si realiza cambios posteriores a muchas opciones de la pestaña Configuración que sean incompatibles con los del objetivo seleccionado actualmente.

Para obtener pruebas no paramétricas para muestras independientes

Seleccione en los menús:

Analizar > **Pruebas no paramétricas** > **Muestras independientes...**

1. Pulse en **Ejecutar**.

Si lo desea, puede:

- Especifique un objetivo en la pestaña Objetivos.
- Especifique asignaciones de campo en la pestaña Campos.
- Especifique la configuración de experto en la pestaña Configuración.

Pestaña Campos

La pestaña Campos especifica los campos que se deben comprobar y el campo que se utilizará para definir grupos.

Utilizar papeles predefinidos. Esta opción utiliza información de campos existentes. Todos los campos continuos y ordinales con un rol predefinido como Destino o Ambos se utilizarán como campos de prueba. Si hay un único campo categórico con un papel predefinido como Entrada se utilizará como un campo de agrupación. De lo contrario, no se utilizará de forma predeterminada ningún campo de agrupación y deberá utilizar asignaciones de campos personalizadas. Se requiere al menos un campo de prueba y un campo de agrupación.

Utilizar asignaciones de campos personalizadas. Esta opción le permite sobrescribir papeles de campos. Después de seleccionar esta opción, especifique los campos siguientes.

- **Campos de prueba.** Seleccione uno o más campos continuos u ordinales.
- **Grupos.** Seleccione un campo categórico.

Pestaña Configuración

La pestaña Configuración contiene diferentes grupos de ajustes que puede modificar para ajustar con precisión con la que el algoritmo procesa sus datos. Si realiza algún cambio en la configuración predeterminada que sea incompatible con el objetivo seleccionado actualmente, la pestaña Objetivo se actualiza automáticamente para seleccionar la opción **Personalizar análisis**.

Seleccionar pruebas

Estos ajustes especifican las pruebas que realizarán en los campos especificados en la pestaña Campos.

Seleccione automáticamente las pruebas en función de los datos. Esta configuración aplica la prueba U de Mann-Whitney para datos con 2 grupos o la prueba ANOVA de 1 factor de Kruskal-Wallis para datos con grupos *k* .

Personalizar pruebas. Esta configuración permite especificar las pruebas que se ejecutarán.

- **Comparar distribuciones entre grupos.** Producen pruebas para muestras independientes si las muestras son de la misma población.

  U de Mann-Whitney (2 muestras) utiliza el nivel de cada caso para comprobar si los grupos se extraen de la misma población. El primer valor del campo de agrupación define el grupo de control y el segundo define el grupo de comparación. Si el campo de agrupación tiene más de dos valores, esta prueba no se ejecuta.

  Kolmogorov-Smirnov (2 muestras) es sensible a cualquier diferencia en la mediana, dispersión, asimetría, etcétera entre las dos distribuciones. Si el campo de agrupación tiene más de dos valores, esta prueba no se ejecuta.

Probar la aleatoriedad de la secuencia (Wald-Wolfowitz para 2 muestras) produce una prueba de rachas con la pertenencia al grupo como criterio. Si el campo de agrupación tiene más de dos valores, esta prueba no se ejecuta.

ANOVA de 1 factor de Kruskal-Wallis (k muestras) es una extensión de la prueba U de Mann-Whitney y el análogo no paramétrico de análisis de varianza de un factor. Opcionalmente puede solicitar múltiples comparaciones de las muestras *k*, en comparaciones múltiples **todo por parejas** o comparaciones **por pasos en sentido descendente**.

Probar alternativas ordenadas (prueba de Jonckheere-Terpstra para k muestras) es una alternativa más potente que Kruskal-Wallis si las muestras *k* tienen un orden natural. 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. **De menor a mayor** especifica la hipótesis alternativa de que el parámetro de ubicación del primer grupo es menor o igual que el segundo, el cual es menor o igual que el tercero, y así sucesivamente. **De mayor a menor** especifica la hipótesis alternativa de que el parámetro de ubicación del primer grupo es mayor o igual que el segundo, el cual es mayor o igual que el tercero, y así sucesivamente. Para ambas opciones, la hipótesis alternativa también supone 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**.

- **Comparar rangos entre grupos.** Produce una prueba de muestras independientes de si las muestras tienen el mismo rango. **Reacciones extremas de Moses (2 muestras)** comprueba un grupo de control con un grupo de comparación. El primer valor en orden ascendente del campo de agrupación define el grupo de control y el segundo define el grupo de comparación. Si el campo de agrupación tiene más de dos valores, esta prueba no se ejecuta.
- **Comparar medianas entre grupos.** Produce una prueba de muestras independientes de si las muestras tienen la misma mediana. **Prueba de la mediana (k muestras)** puede utilizar la mediana muestral combinada (calculada con todos los registros del conjunto de datos) o un valor personalizado como la mediana hipotetizada. Opcionalmente puede solicitar múltiples comparaciones de las muestras *k*, en comparaciones múltiples **todo por parejas** o comparaciones **por pasos en sentido descendente**.
- **Estimar intervalos de confianza entre grupos.** **Hodges-Lehman (2 muestras)** produce una estimación muestras independientes y el intervalo de confianza para la diferencia en las medianas de los dos grupos. Si el campo de agrupación tiene más de dos valores, esta prueba no se ejecuta.

Opciones de prueba

Nivel de significación. Especifica el nivel de significación (alfa) de todas las pruebas. Especifica un valor numérico entre 0 y 1. 0,05 es el valor predeterminado.

Intervalo de confianza (%). Esto especifica el nivel de confianza de todos los intervalos de confianza producidos. Especifique un valor numérico entre 0 y 100. El valor predeterminado es 95.

Casos excluidos. Especifica cómo se determinan las pruebas caso por caso. **Excluir casos según lista** significa que los registros con valores perdidos de cualquier campo que se nombran en cualquier subcomando se excluyen de todos los análisis. **Excluir casos según prueba** significa que los registros con valores perdidos para un campo que se utiliza para una prueba específica se omiten de esa prueba. Si se realizan varias pruebas en el análisis, cada prueba se evalúa por separado.

Valores perdidos del usuario

Valores perdidos del usuario para campos categóricos. Para que un registro se incluya en el análisis, los campos categóricos deben tener valores válidos para dicho caso. Estos controles permiten decidir si los valores perdidos del usuario se deben tratar como válidos entre los campos categóricos. Los valores perdidos del sistema y los valores perdidos de campos continuos siempre se tratan como no válidos.

Características adicionales del comando NPTESTS

La sintaxis de comandos también le permite:

- Especifique pruebas de una muestra, muestras independientes y muestras relacionadas en una única ejecución del procedimiento.

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Pruebas no paramétricas de muestras relacionadas

Identifica diferencias entre dos o más campos relacionados mediante una o más pruebas no paramétricas. Las pruebas no paramétricas no dan por hecho que sus datos sigan la distribución normal.

Consideraciones sobre los datos. Cada registro corresponde a un sujeto concreto para el que se almacenan dos o más mediciones relacionadas en campos separados del conjunto de datos. Por ejemplo, es posible analizar un estudio sobre la efectividad de un plan de dietas mediante pruebas no paramétricas de muestras relacionadas si el peso de cada sujeto se mide a intervalos regulares y se almacena como campos como *Peso previo a la dieta*, *Peso intermedio* y *Peso tras la dieta*. Estos campos están "relacionados".

¿Cuál es su objetivo? Los objetivos le permiten especificar rápidamente ajustes de prueba diferentes y comunes.

- **Comparar automáticamente datos observados con datos hipotetizados.** Este objetivo aplica la prueba de McNemar a datos categóricos cuando se especifican 2 campos, la prueba Q de Cochran datos categóricos cuando se especifican más de 2 campos, la prueba de Wilcoxon de los rangos con signo a datos continuos cuando se especifican 2 campos y ANOVA de 2 vías de Friedman por rangos a datos continuos cuando se especifican más de 2 campos.
- **Análisis personalizado.** Seleccione esta opción si desea modificar manualmente la configuración de la prueba de la pestaña Configuración. Tenga en cuenta que esta configuración se selecciona automáticamente si realiza cambios posteriores a muchas opciones de la pestaña Configuración que sean incompatibles con los del objetivo seleccionado actualmente.

Cuando se especifican campos de diferentes niveles de medición, primero se separan por nivel de medición y después se aplica la prueba adecuada a cada grupo. Por ejemplo, si selecciona **Comparar automáticamente datos observados con el valor hipotetizado** como objetivo y especifica 3 campos continuos y 2 campos nominales, se aplicará la prueba de Friedman a los campos continuos y la prueba de McNemar a los campos nominales.

Para obtener pruebas no paramétricas para muestras relacionadas

Seleccione en los menús:

Analizar > **Pruebas no paramétricas** > **Muestras relacionadas...**

1. Pulse en **Ejecutar**.

Si lo desea, puede:

- Especifique un objetivo en la pestaña Objetivos.
- Especifique asignaciones de campo en la pestaña Campos.
- Especifique la configuración de experto en la pestaña Configuración.

Pestaña Campos

La pestaña Campos especifica los campos que se deben comprobar.

Utilizar papeles predefinidos. Esta opción utiliza información de campos existentes. Todos los campos con un papel predefinido como Destino o Ambos se utilizarán como campos de prueba. Se requieren al menos dos campos de prueba.

Utilizar asignaciones de campos personalizadas. Esta opción le permite sobrescribir papeles de campos. Después de seleccionar esta opción, especifique los campos siguientes.

- **Campos de prueba.** Seleccione dos o más campos. Cada campo corresponde a una muestra relacionada diferente.

Pestaña Configuración

La pestaña Configuración contiene diferentes grupos de ajustes que puede modificar para ajustar con precisión con la que el procedimiento procesa sus datos. Si realiza algún cambio en la configuración predeterminada que sea incompatible con el resto de objetivos, la pestaña Objetivo se actualiza automáticamente para seleccionar la opción**Personalizar análisis**.

Seleccionar pruebas

Estos ajustes especifican las pruebas que realizarán en los campos especificados en la pestaña Campos.

Seleccione automáticamente las pruebas en función de los datos. Esta configuración aplica la prueba de McNemar a datos categóricos cuando se especifican 2 campos, la prueba Q de Cochran datos categóricos cuando se especifican más de 2 campos, la prueba de Wilcoxon de los rangos con signo a datos continuos cuando se especifican 2 campos y ANOVA de 2 vías de Friedman por rangos a datos continuos cuando se especifican más de 2 campos.

Personalizar pruebas. Esta configuración permite especificar las pruebas que se ejecutarán.

- **Probar si hay cambios en datos binario. La prueba de McNemar (2 muestras)** se puede aplicar a campos categóricos. Produce una prueba de muestras relacionadas de si las combinaciones de valores entre dos campos de distintivo (campos categóricos con dos valores únicamente) son igualmente probables. Si hay más de dos campos especificados en la pestaña Campos, esta prueba no se realiza. Consulte “Prueba de McNemar: definir éxito” en la página 137 para obtener más información sobre la configuración de prueba. **Q de Cochran (k muestras)** se puede aplicar a campos categóricos. Produce una prueba de muestras relacionadas de si las combinaciones de valores entre *k* campos de distintivo (campos categóricos con dos valores únicamente) son igualmente probables. Opcionalmente puede solicitar múltiples comparaciones de las muestras *k*, en comparaciones múltiples **todo por parejas** o comparaciones **por pasos en sentido descendente**. Consulte “Prueba de Cochran: definir éxito” en la página 137 para obtener más información sobre la configuración de prueba.
- **Probar si hay cambios en datos mutinomiales. Prueba de homogeneidad marginal (2 muestras)** produce una prueba de muestras relacionadas de si combinaciones de valores entre dos campos ordinales emparejados son igualmente probables. La prueba de homogeneidad marginal se suele utilizar en situaciones de medidas repetidas. Se trata de una extensión de la prueba de McNemar a partir de la respuesta binaria a la respuesta multinomial. Si hay más de dos campos especificados en la pestaña Campos, esta prueba no se realiza.
- **Comparar diferencia de la mediana con el valor hipotetizado.** Cada una de estas pruebas produce una prueba de muestras relacionadas de si la diferencia de la mediana entre dos campos es diferente de 0. La prueba se aplica a campos continuos y ordinales. Si hay más de dos campos especificados en la pestaña Campos, estas pruebas no se realizan.
- **Estimar intervalo de confianza.** Produce un cálculo de muestras relacionadas y un intervalo de confianza para la diferencia de la mediana entre dos campos emparejados. Esta prueba se aplica a campos continuos y ordinales. Si hay más de dos campos especificados en la pestaña Campos, esta prueba no se realiza.
- **Cuantificar asociaciones. Coeficiente de concordancia de Kendall (k muestras)** produce un coeficiente de concordancia entre evaluadores, donde cada registro es un valor de evaluador de varios elementos (campos). Opcionalmente puede solicitar múltiples comparaciones de las muestras *k*, en comparaciones múltiples **todo por parejas** o comparaciones **por pasos en sentido descendente**.

- **Comparar distribuciones. Friedman's 2-way ANOVA by ranks (k samples)** produces a related samples test of whether *k* related samples have been drawn from the same population. Opcionalmente puede solicitar múltiples comparaciones de las muestras *k*, en comparaciones múltiples **todo por parejas** o comparaciones **por pasos en sentido descendente**.

Prueba de McNemar: definir éxito: La prueba de McNemar está diseñada para campos de distintivo (campos categóricos con sólo dos categorías), pero se aplica a todos los campos categóricos mediante reglas para definir "éxito".

Definir éxito para campos categóricos. Especifica cómo se define "éxito" en los campos categóricos.

- **Utilizar primera categoría encontrada en los datos** realiza la prueba que utiliza el primer valor encontrado en la muestra para definir "éxito". Esta opción sólo es aplicable a los campos nominal u ordinal con sólo dos valores; el resto de campos categóricos especificados en la pestaña Campos en los que se utiliza esta opción no se comprobarán. Este es el método predeterminado.
- **Especificar valores de éxito** realiza la prueba que utiliza la lista especificada de valores para definir "éxito". Especifique una lista de valores de cadena o numérico. No es necesario que los valores de la lista estén en la muestra.

Prueba de Cochran: definir éxito: La prueba de Q de Cochran está diseñada para campos de distintivo (campos categóricos con sólo dos categorías), pero se aplica a todos los campos categóricos mediante reglas para definir "éxito".

Definir éxito para campos categóricos. Especifica cómo se define "éxito" en los campos categóricos.

- **Utilizar primera categoría encontrada en los datos** realiza la prueba que utiliza el primer valor encontrado en la muestra para definir "éxito". Esta opción sólo es aplicable a los campos nominal u ordinal con sólo dos valores; el resto de campos categóricos especificados en la pestaña Campos en los que se utiliza esta opción no se comprobarán. Este es el método predeterminado.
- **Especificar valores de éxito** realiza la prueba que utiliza la lista especificada de valores para definir "éxito". Especifique una lista de valores de cadena o numérico. No es necesario que los valores de la lista estén en la muestra.

Opciones de prueba

Nivel de significación. Especifica el nivel de significación (alfa) de todas las pruebas. Especifica un valor numérico entre 0 y 1. 0,05 es el valor predeterminado.

Intervalo de confianza (%). Esto especifica el nivel de confianza de todos los intervalos de confianza producidos. Especifique un valor numérico entre 0 y 100. El valor predeterminado es 95.

Casos excluidos. Especifica cómo se determinan las pruebas caso por caso.

- **Excluir casos según lista** significa que los registros con valores perdidos de cualquier campo que se nombran en cualquier subcomando se excluyen de todos los análisis.
- **Excluir casos según prueba** significa que los registros con valores perdidos para un campo que se utiliza para una prueba específica se omiten de esa prueba. Si se realizan varias pruebas en el análisis, cada prueba se evalúa por separado.

Valores perdidos del usuario

Valores perdidos del usuario para campos categóricos. Para que un registro se incluya en el análisis, los campos categóricos deben tener valores válidos para dicho caso. Estos controles permiten decidir si los valores perdidos del usuario se deben tratar como válidos entre los campos categóricos. Los valores perdidos del sistema y los valores perdidos de campos continuos siempre se tratan como no válidos.

Características adicionales del comando NPTESTS

La sintaxis de comandos también le permite:

- Especifique pruebas de una muestra, muestras independientes y muestras relacionadas en una única ejecución del procedimiento.

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Vista de modelo

Vista de modelos

Este procedimiento crea un objeto Visor de modelos en el visor. Al activar (pulsando dos veces) este objeto se obtiene una vista interactiva del modelo. La vista de modelos se compone de una ventana con dos paneles, la vista principal en la parte izquierda y la vista relacionada o auxiliar de la derecha.

Hay dos vistas principales:

- Resumen de hipótesis. Esta es la vista predeterminada. Consulte el tema "Resumen de hipótesis" para obtener más información.
- Resumen de intervalo de confianza. Consulte el tema "Resumen de intervalo de confianza" para obtener más información.

Hay siete vistas relacionadas/auxiliares:

- Prueba de una muestra. Es la vista predeterminada si se han solicitado pruebas de una muestra. Consulte el tema "Pruebas de una muestra" en la página 139 para obtener más información.
- Prueba de muestras relacionadas. Es la vista predeterminada si se han solicitado pruebas de muestras relacionadas y no pruebas de una muestra. Consulte el tema "Prueba de muestras relacionadas" en la página 140 para obtener más información.
- Prueba de muestras independientes. Es la vista predeterminada si no se han solicitado pruebas de muestras relacionadas ni pruebas de una muestra. Consulte el tema "Prueba de muestras independientes" en la página 141 para obtener más información.
- Información de campos categóricos. Consulte el tema "Información de campos categóricos" en la página 142 para obtener más información.
- Información de campos continuos. Consulte el tema "Información de campos continuos" en la página 142 para obtener más información.
- Comparaciones por parejas. Consulte el tema "Comparaciones por parejas" en la página 142 para obtener más información.
- Subconjuntos homogéneos. Consulte el tema "Subconjuntos homogéneos" en la página 142 para obtener más información.

Resumen de hipótesis

La vista Resumen de modelos es una instantánea, un resumen de un vistazo de las pruebas no paramétricas. Enfatiza las hipótesis y decisiones nulas, centrando la atención en los valores *p* más significativos.

- Cada fila corresponde a una prueba distinta. Al pulsar en una fila se muestra información adicional acerca de la prueba en la vista enlazada.
- Al pulsar en el encabezado de cualquier columna las filas se ordenan por los valores de esa columna.
- El botón **Restablecer** le permite devolver el Visor de modelos a su estado original.
- La lista desplegable **Filtro de campos** le permite mostrar únicamente las pruebas que incluyeron el campo seleccionado.

Resumen de intervalo de confianza

Resumen de intervalo de confianza le muestra los intervalos de confianza producidos por las pruebas no paramétricas.

- Cada fila corresponde a un intervalo de confianza distinto.

- Al pulsar en el encabezado de cualquier columna las filas se ordenan por los valores de esa columna.

Pruebas de una muestra

La vista de prueba de una muestra incluye detalles relacionados con cualquier prueba no paramétrica de una muestra solicitada. La información que se muestra depende de la prueba seleccionada.

- La lista desplegable **Prueba** le permite seleccionar un tipo concreto de prueba de una muestra.
- La lista desplegable **Campo(s)** le permite seleccionar un campo que se haya comprobado mediante la prueba seleccionada en la lista desplegable **Prueba**.

Prueba binomial

La prueba binomial muestra un gráfico de barras apiladas y una tabla de pruebas.

- El gráfico de barras apiladas muestra las frecuencias observadas e hipotetizadas de las categorías "éxito" y "fallo" del campo de prueba, con los "fallos" apilados sobre los "éxitos". Al pasar el ratón sobre una barra se muestran los porcentajes de categoría en una ayuda contextual. Las diferencias visibles en las barras indican que el campo de prueba puede no tener la distribución binomial hipotetizada.
- La tabla muestra detalles de la prueba.

Prueba de chi-cuadrado

La vista Prueba de chi-cuadrado muestra un gráfico de barras apiladas y una tabla de pruebas.

- El gráfico de barras agrupadas muestra las frecuencias observadas e hipotetizadas para cada categoría del campo de pruebas. Al pasar el ratón sobre una barra se muestran las frecuencias observadas e hipotetizadas y sus diferencias (residuales) en una ayuda contextual. Las diferencias visibles entre las barras observadas y las hipotetizadas indican que el campo de prueba puede no tener la distribución hipotetizada.
- La tabla muestra detalles de la prueba.

Prueba de Wilcoxon de los rangos con signo

La vista Prueba de Wilcoxon de los rangos con signo muestra un histograma y una tabla de pruebas.

- El histograma incluye líneas verticales que muestran las medianas observadas e hipotéticas.
- La tabla muestra detalles de la prueba.

Prueba de rachas

La vista Prueba de rachas muestra un gráfico y una tabla de pruebas.

- El gráfico muestra la distribución normal con el número observado de rachas marcado con una línea vertical. Tenga en cuenta que cuando se realiza la prueba exacta, ésta no se basa en la distribución normal.
- La tabla muestra detalles de la prueba.

Prueba de Kolmogorov-Smirnov

La vista Prueba de Kolmogorov-Smirnov muestra un histograma y una tabla de pruebas.

- El histograma incluye una superposición de la función de densidad de probabilidad para la distribución exponencial, Poisson, normal o uniforme hipotetizada. Tenga en cuenta que la prueba se basa en distribuciones acumuladas, y las Diferencias más extremas indicadas en la tabla deben interpretarse con respecto a las distribuciones acumuladas.
- La tabla muestra detalles de la prueba.

Prueba de muestras relacionadas

La vista de prueba de una muestra incluye detalles relacionados con cualquier prueba no paramétrica de una muestra solicitada. La información que se muestra depende de la prueba seleccionada.

- La lista desplegable **Prueba** le permite seleccionar un tipo concreto de prueba de una muestra.
- La lista desplegable **Campo(s)** le permite seleccionar un campo que se haya comprobado mediante la prueba seleccionada en la lista desplegable **Prueba**.

Prueba de McNemar

La vista Prueba de McNemar muestra un gráfico de barras apiladas y una tabla de pruebas.

- El gráfico de barras agrupadas muestra las frecuencias observadas e hipotetizadas para las casillas no diagonales de la tabla 2x2 definida por los campos de prueba.
- La tabla muestra detalles de la prueba.

Prueba de los signos

La vista Prueba de los signos muestra un histograma apilado y una tabla de pruebas.

- El histograma apilado muestra las diferencias entre los campos, usando el signo de la diferencia como el campo de apilado.
- La tabla muestra detalles de la prueba.

Prueba de Wilcoxon de los rangos con signo

La vista Prueba de Wilcoxon de los rangos con signo muestra un histograma apilado y una tabla de pruebas.

- El histograma apilado muestra las diferencias entre los campos, usando el signo de la diferencia como el campo de apilado.
- La tabla muestra detalles de la prueba.

Prueba de homogeneidad marginal

La vista Prueba de homogeneidad marginal muestra un gráfico de barras apiladas y una tabla de pruebas.

- El gráfico de barras agrupadas muestra las frecuencias observadas para las casillas no diagonales de la tabla definida por los campos de prueba.
- La tabla muestra detalles de la prueba.

Prueba Q de Cochran

La vista Prueba Q de Cochran muestra un gráfico de barras apiladas y una tabla de pruebas.

- El gráfico de barras apiladas muestra las frecuencias observadas de las categorías "éxito" y "fallo" de los campos de prueba, con los "fallos" apilados sobre los "éxitos". Al pasar el ratón sobre una barra se muestran los porcentajes de categoría en una ayuda contextual.
- La tabla muestra detalles de la prueba.

Análisis de dos factores de Friedman de varianza por rangos

La vista Análisis de dos factores de Friedman de varianza por rangos muestra histogramas panelados y una tabla de pruebas.

- Los histogramas muestran la distribución observada de rangos, panelados por los campos de pruebas.
- La tabla muestra detalles de la prueba.

Coeficiente de concordancia de Kendall

La vista Coeficiente de concordancia de Kendall muestra histogramas panelados y una tabla de pruebas.

- Los histogramas muestran la distribución observada de rangos, panelados por los campos de pruebas.
- La tabla muestra detalles de la prueba.

Prueba de muestras independientes

La vista Prueba de muestras independientes incluye detalles relacionados con cualquier prueba no paramétrica de muestras independientes solicitada. La información que se muestra depende de la prueba seleccionada.

- La lista desplegable **Prueba** le permite seleccionar un tipo concreto de prueba de muestra independiente.
- La lista desplegable **Campo(s)** le permite seleccionar una combinación de prueba y campo de agrupación que se haya comprobado mediante la prueba seleccionada en la lista desplegable **Prueba**.

Prueba de Mann-Whitney

La vista Prueba de Mann-Whitney muestra un gráfico de pirámide de población y una tabla de pruebas.

- El gráfico de pirámide de población muestra histogramas seguidos en función de las categorías del campo de agrupación, anotando el número de registros de cada grupo y el rango promedio del grupo.
- La tabla muestra detalles de la prueba.

Prueba de Kolmogorov-Smirnov

La vista Prueba de Kolmogorov-Smirnov muestra un gráfico de pirámide de población y una tabla de pruebas.

- El gráfico de pirámide de población muestra histogramas seguidos en función de las categorías del campo de agrupación, anotando el número de registros de cada grupo. Las líneas de distribución acumulada observadas pueden mostrarse u ocultarse pulsando el botón **Acumulado**.
- La tabla muestra detalles de la prueba.

Prueba de rachas de Wald-Wolfowitz

La vista Prueba de rachas de Wald-Wolfowitz muestra un gráfico de barras apiladas y una tabla de pruebas.

- El gráfico de pirámide de población muestra histogramas seguidos en función de las categorías del campo de agrupación, anotando el número de registros de cada grupo.
- La tabla muestra detalles de la prueba.

Prueba de Kruskal-Wallis

La vista Prueba de Kruskal-Wallis muestra diagramas de caja y una tabla de pruebas.

- Se muestran diagramas de caja distintos para cada categoría del campo de agrupación. Al pasar el ratón sobre un cuadro se muestra el rango promedio en una ayuda contextual.
- La tabla muestra detalles de la prueba.

Prueba de Jonckheere-Terpstra

La vista Prueba de Jonckheere-Terpstra muestra diagramas de caja y una tabla de pruebas.

- Se muestran diagramas de caja distintos para cada categoría del campo de agrupación.
- La tabla muestra detalles de la prueba.

Prueba de Moses de reacción extrema

La vista Prueba de Moses de reacción extrema muestra diagramas de caja y una tabla de pruebas.

- Se muestran diagramas de caja distintos para cada categoría del campo de agrupación. Las etiquetas de punto pueden mostrarse u ocultarse pulsando el botón **ID de registro**.
- La tabla muestra detalles de la prueba.

Prueba de la mediana

La vista Prueba de la mediana muestra diagramas de caja y una tabla de pruebas.

- Se muestran diagramas de caja distintos para cada categoría del campo de agrupación.
- La tabla muestra detalles de la prueba.

Información de campos categóricos

La vista Información de campos categóricos muestra un gráfico de barras para el campo categórico seleccionado en la lista desplegable **Campo(s)**. La lista de campos disponibles está restringida a los campos categóricos utilizados en la prueba seleccionada actualmente en la vista Resumen de hipótesis.

- Al pasar el ratón sobre una barra se muestran los porcentajes de categoría en una ayuda contextual.

Información de campos continuos

La vista Información de campos continuos muestra un histograma del campo continuo seleccionado en la lista desplegable **Campo(s)**. La lista de campos disponibles está restringida a los campos continuos utilizados en la prueba seleccionada actualmente en la vista Resumen de hipótesis.

Comparaciones por parejas

La vista Comparaciones por parejas muestra un gráfico de distancias de red y una tabla de comparaciones producidas por pruebas no paramétricas de muestras *k* cuando se solicitan múltiples comparaciones de pares.

- El gráfico de distancias de red es una representación gráfica de la tabla de comparaciones en la que las distancias entre nodos de la red corresponden a las diferencias entre las muestras. Las líneas amarillas corresponden a diferencias estadísticamente importantes, mientras que las líneas negras corresponden a diferencias no significativas. Al pasar el ratón por una línea de la red se muestra una ayuda contextual con la significación corregida de la diferencia entre los nodos conectados por la línea.
- La tabla de comparaciones muestra los resultados numéricos de todas las comparaciones de parejas. Cada fila corresponde a una comparación de parejas distinta. Al pulsar en el encabezado de cualquier columna las filas se ordenan por los valores de esa columna.

Subconjuntos homogéneos

La vista Subconjuntos homogéneos muestra una tabla de comparaciones generadas por pruebas no paramétricas de muestras *k* cuando se solicitan múltiples comparaciones por pasos.

- Cada fila del grupo Muestra corresponde a una muestra relacionada distinta (representada en los datos mediante campos distintos). Las muestras que no son muy diferentes estadísticamente se agrupan en los mismos subconjuntos de color, y hay una columna separada por cada subconjunto identificado. Cuando todas las muestras son muy diferentes estadísticamente, hay un subconjunto separado para cada muestra. Si ninguna de las muestras es muy diferente estadísticamente, hay un único subconjunto.
- Se calcula una estadística de prueba, un valor de significación y un valor de significación corregida para cada subconjunto que contenga más de una muestra.

Características adicionales del comando NPTESTS

La sintaxis de comandos también le permite:

- Especifique pruebas de una muestra, muestras independientes y muestras relacionadas en una única ejecución del procedimiento.

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Cuadros de diálogo antiguos

Hay diferentes cuadros de diálogo "antiguos" que también realizan pruebas no paramétricas. Estos cuadros de diálogo admiten la funcionalidad que ofrece la opción Pruebas exactas.

Prueba de chi-cuadrado. Tabula una variable en categorías y calcula un estadístico de chi-cuadrado basándose en las diferencias entre las frecuencias observadas y las esperadas.

Prueba binomial. Compara la frecuencia observada en cada categoría de una variable dicotómica con las frecuencias esperadas en la distribución binomial.

Prueba de rachas. Comprueba si el orden de aparición de dos valores de una variable es aleatorio.

Prueba de Kolmogorov-Smirnov para una muestra. Compara la función de distribución acumulada observada de una variable con una distribución teórica especificada, que puede ser normal, uniforme, exponencial o de Poisson.

Pruebas para dos muestras independientes. Compara dos grupos de casos en una variable. Se encuentran disponibles la prueba *U* de Mann-Whitney, la prueba de Kolmogorov-Smirnov para dos muestras, la prueba de Moses de reacciones extremas y la prueba de rachas de Wald-Wolfowitz.

Pruebas para dos muestras relacionadas. Compara las distribuciones de dos variables. La prueba de Wilcoxon de los rangos con signo, la prueba de signos y la prueba de McNemar.

Pruebas para varias muestras independientes. Compara dos o más grupos de casos en una variable. Se encuentran disponibles la prueba de Kruskal-Wallis, la prueba de la mediana y la prueba de Jonckheere-Terpstra.

Pruebas para varias muestras relacionadas. Compara las distribuciones de dos o más variables. Se encuentran disponibles la prueba de Friedman, la *W* de Kendall y la *Q* de Cochran.

Además, para todas las pruebas citadas anteriormente están disponibles los cuartiles y la media, la desviación estándar, el mínimo, el máximo y, por último, el número de casos no perdidos.

Prueba de chi-cuadrado

El procedimiento Prueba de chi-cuadrado tabula una variable en categorías y calcula un estadístico de chi-cuadrado. Esta prueba de bondad de ajuste compara las frecuencias observadas y esperadas en cada categoría para contrastar que todas las categorías contengan la misma proporción de valores o que cada categoría contenga una proporción de valores especificada por el usuario.

Ejemplos. La prueba de chi-cuadrado podría utilizarse para determinar si una bolsa de caramelos contiene en igualdad de proporción caramelos de color azul, marrón, verde, naranja, rojo y amarillo. También podría utilizarse para ver si una bolsa de caramelos contiene un 5% de color azul, un 30% de color marrón, un 10% de color verde, un 20% de color naranja, un 15% de color rojo y un 15% de color amarillo.

Estadísticos. Media, desviación estándar, mínimo, máximo y cuartiles. Número y porcentaje de casos perdidos y no perdidos; número de casos observados y esperados de cada categoría; residuos y estadístico de chi-cuadrado.

Prueba de chi-cuadrado: Consideraciones sobre los datos

Datos. Use variables categóricas numéricas ordenadas o no ordenadas (niveles de medición ordinal o nominal). Para convertir las variables de cadena en variables numéricas, utilice el procedimiento Recodificación automática, disponible en el menú Transformar.

Supuestos. Las pruebas no paramétricas no requieren supuestos sobre la forma de la distribución subyacente. Se asume que los datos son una muestra aleatoria. Las frecuencias esperadas para cada categoría deberán ser 1 como mínimo. No más de un 20% de las categorías deberán tener frecuencias esperadas menores que 5.

Para obtener una prueba de chi-cuadrado

1. Seleccione en los menús:

   Analizar > **Pruebas no paramétricas** > **Cuadros de diálogo antiguos** > **Chi-cuadrado...**
2. Seleccione una o más variables de contraste. Cada variable genera una prueba independiente.
3. Si lo desea, puede pulsar en **Opciones** para obtener estadísticos descriptivos, cuartiles y controlar el tratamiento de los datos perdidos.

Prueba de chi-cuadrado: Rango y valores esperados

Rango esperado. De forma predeterminada, cada valor distinto de la variable se define como una categoría. Para establecer categorías dentro de un rango específico, seleccione **Usar rango especificado** e introduzca valores enteros para los límites inferior y superior. Se establecerán categorías para cada valor entero dentro del rango inclusivo y los casos con valores fuera de los límites se excluirán. Por ejemplo, si se especifica 1 como límite inferior y 4 como límite superior, únicamente se utilizarán los valores enteros entre 1 y 4 para la prueba de chi-cuadrado.

Valores esperados. De forma predeterminada, todas las categorías tienen valores esperados iguales. Las categorías pueden tener proporciones esperadas especificadas por el usuario. Seleccione **Valores**, introduzca un valor mayor que 0 para cada categoría de la variable de contraste y, a continuación, pulse en **Añadir**. Cada vez que se añade un valor, éste aparece al final de la lista de valores. El orden de los valores es importante; corresponde al orden ascendente de los valores de categoría de la variable de contraste. El primer valor de la lista corresponde al valor de grupo mínimo de la variable de contraste y el último valor corresponde al valor máximo. Los elementos de la lista de valores se suman y, a continuación, cada valor se divide por esta suma para calcular la proporción de casos esperados en la categoría correspondiente. Por ejemplo, una lista de valores de 3, 4, 5, 4 especifica unas proporciones esperadas de 3/16, 4/16, 5/16 y 4/16.

Prueba de chi-cuadrado: Opciones

Estadísticos. Puede elegir uno o los dos estadísticos de resumen.

- **Descriptivos.** Muestra la media, la desviación estándar, el mínimo, el máximo y el número de casos no perdidos.
- **Cuartiles.** Muestra los valores correspondientes a los percentiles 25, 50 y 75.

Valores perdidos. Controla el tratamiento de los valores perdidos.

- **Excluir casos según prueba.** Cuando se especifican varias pruebas, cada una se evalúa separadamente respecto a los valores perdidos.
- **Excluir casos según lista.** Los casos con valores perdidos para cualquier variable se excluyen de todos los análisis.

Características adicionales del comando NPAR TESTS (Prueba de chi-cuadrado)

La sintaxis de comandos también le permite:

- Especificar valores mínimos y máximos o frecuencias esperadas diferentes para diferentes variables (con el subcomando `CHISQUARE`).
- Contrastar la misma variable respecto a diferentes frecuencias esperadas o utilizar diferentes rangos (con el subcomando `EXPECTED`).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Prueba binomial

El procedimiento Prueba binomial compara las frecuencias observadas de las dos categorías de una variable dicotómica con las frecuencias esperadas en una distribución binomial con un parámetro de probabilidad especificado. De forma predeterminada, el parámetro de probabilidad para ambos grupos es 0,5. Para cambiar las probabilidades, puede introducirse una proporción de prueba para el primer grupo. La probabilidad del segundo grupo será 1 menos la probabilidad especificada para el primer grupo.

Ejemplo. Si se lanza una moneda al aire, la probabilidad de que salga cara es 1/2. Basándose en esta hipótesis, se lanza una moneda al aire 40 veces y se anotan los resultados (cara o cruz). De la prueba binomial, podría deducir que en 3/4 de los lanzamientos salió cara y que el nivel de significación observado es pequeño (0,0027). Estos resultados indican que no es verosímil que la probabilidad de que salga cara sea 1/2; probablemente la moneda presenta una tendencia a caer por un sentido determinado.

Estadísticos. Media, desviación estándar, mínimo, máximo, número de casos no perdidos y cuartiles.

Prueba binomial: Consideraciones sobre los datos

Datos. Las variables de contraste deben ser numéricas y dicotómicas. Para convertir las variables de cadena en variables numéricas, utilice el procedimiento Recodificación automática, disponible en el menú Transformar. Una **variable dicotómica** es una variable que sólo puede tomar dos valores posibles: *sí* o *no*, *verdadero* o *falso*, 0 o 1, etc. El primer valor encontrado en los datos define el primer grupo y el otro valor define el segundo grupo. Si las variables no son dicotómicas, debe especificar un punto de corte. El punto de corte asigna los casos con valores menores o iguales que el punto de corte del primer grupo y asigna el resto de los casos a un segundo grupo.

Supuestos. Las pruebas no paramétricas no requieren supuestos sobre la forma de la distribución subyacente. Se asume que los datos son una muestra aleatoria.

Para obtener una prueba binomial

1. Seleccione en los menús:

   Analizar > **Pruebas no paramétricas** > **Cuadros de diálogo antiguos** > **Binomial...**
2. Seleccione una o más variables de contraste numéricas.
3. Si lo desea, puede pulsar en **Opciones** para obtener estadísticos descriptivos, cuartiles y controlar el tratamiento de los datos perdidos.

Prueba binomial: Opciones

Estadísticos. Puede elegir uno o los dos estadísticos de resumen.

- **Descriptivos.** Muestra la media, la desviación estándar, el mínimo, el máximo y el número de casos no perdidos.
- **Cuartiles.** Muestra los valores correspondientes a los percentiles 25, 50 y 75.

Valores perdidos. Controla el tratamiento de los valores perdidos.

- **Excluir casos según prueba.** Cuando se especifican varias pruebas, cada una se evalúa separadamente respecto a los valores perdidos.
- **Excluir casos según lista.** Se excluirán de todos los análisis los casos con valores perdidos de cualquier variable.

Características adicionales del comando NPAR TESTS (Prueba binomial)

La sintaxis de comandos también le permite:

- Seleccionar grupos específicos (y excluir otros grupos) si una variable tiene más de dos categorías (mediante el subcomando `BINOMIAL`).
- Especificar diferentes probabilidades o puntos de corte para diferentes variables (mediante el subcomando `BINOMIAL`).

- Contrastar la misma variable respecto a diferentes probabilidades o puntos de corte (mediante el subcomando EXPECTED).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Prueba de rachas

El procedimiento Prueba de rachas contrasta si es aleatorio el orden de aparición de dos valores de una variable. Una racha es una secuencia de observaciones similares. Una muestra con un número excesivamente grande o excesivamente pequeño de rachas sugiere que la muestra no es aleatoria.

Ejemplos. Suponga que se realiza una encuesta a 20 personas para saber si comprarían un producto. Si todas estas personas fueran del mismo sexo, se pondría seriamente en duda la supuesta aleatoriedad de la muestra. La prueba de rachas se puede utilizar para determinar si la muestra fue extraída de manera aleatoria.

Estadísticos. Media, desviación estándar, mínimo, máximo, número de casos no perdidos y cuartiles.

Prueba de rachas: Consideraciones sobre los datos

Datos. Las variables deben ser numéricas. Para convertir las variables de cadena en variables numéricas, utilice el procedimiento Recodificación automática, disponible en el menú Transformar.

Supuestos. Las pruebas no paramétricas no requieren supuestos sobre la forma de la distribución subyacente. Utilice muestras de distribuciones de probabilidad continua.

Para obtener una prueba de rachas

1. Seleccione en los menús:

   Analizar > **Pruebas no paramétricas** > **Cuadros de diálogo antiguos** > **Rachas...**
2. Seleccione una o más variables de contraste numéricas.
3. Si lo desea, puede pulsar en **Opciones** para obtener estadísticos descriptivos, cuartiles y controlar el tratamiento de los datos perdidos.

Prueba de rachas: punto de corte

Punto de corte. Especifica un punto de corte para dicotomizar las variables seleccionadas. Puede utilizar como punto de corte los valores observados para la media, la mediana o la moda, o bien un valor especificado. Los casos con valores menores que el punto de corte se asignarán a un grupo y los casos con valores mayores o iguales que el punto de corte se asignarán a otro grupo. Se lleva a cabo una prueba para cada punto de corte seleccionado.

Prueba de rachas: Opciones

Estadísticos. Puede elegir uno o los dos estadísticos de resumen.

- **Descriptivos.** Muestra la media, la desviación estándar, el mínimo, el máximo y el número de casos no perdidos.
- **Cuartiles.** Muestra los valores correspondientes a los percentiles 25, 50 y 75.

Valores perdidos. Controla el tratamiento de los valores perdidos.

- **Excluir casos según prueba.** Cuando se especifican varias pruebas, cada una se evalúa separadamente respecto a los valores perdidos.
- **Excluir casos según lista.** Los casos con valores perdidos para cualquier variable se excluyen de todos los análisis.

Características adicionales del comando NPAR TESTS (Prueba de rachas)

La sintaxis de comandos también le permite:

- Especificar puntos de corte diferentes para las distintas variables (con el subcomando RUNS).
- Contrastar la misma variable con distintos puntos de corte personalizados (con el subcomando RUNS).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Prueba Kolmogorov-Smirnov de una muestra

El procedimiento Prueba de Kolmogorov-Smirnov para una muestra compara la función de distribución acumulada observada de una variable con una distribución teórica determinada, que puede ser la normal, la uniforme, la de Poisson o la exponencial. La *Z* de Kolmogorov-Smirnov se calcula a partir de la diferencia mayor (en valor absoluto) entre las funciones de distribución acumuladas teórica y observada. Esta prueba de bondad de ajuste contrasta si las observaciones podrían razonablemente proceder de la distribución especificada.

Ejemplo. Muchas pruebas paramétricas requieren que las variables se distribuyan de forma normal. La prueba de Kolmogorov-Smirnov para una muestra se puede utilizar para comprobar que una variable (por ejemplo *ingresos*) se distribuye normalmente.

Estadísticos. Media, desviación estándar, mínimo, máximo, número de casos no perdidos y cuartiles.

Prueba de Kolmogorov-Smirnov para una muestra: Consideraciones sobre los datos

Datos. Utilice variables cuantitativas (a nivel de medición de razón o de intervalo).

Supuestos. La prueba de Kolmogorov-Smirnov asume que los parámetros de la distribución de prueba se han especificado previamente. Este procedimiento estima los parámetros a partir de la muestra. La media y la desviación estándar de la muestra son los parámetros de una distribución normal, los valores mínimo y máximo de la muestra definen el rango de la distribución uniforme, la media muestral es el parámetro de la distribución de Poisson y la media muestral es el parámetro de la distribución exponencial. La capacidad de la prueba para detectar desviaciones a partir de la distribución hipotetizada puede disminuir gravemente. Para contrastarla con una distribución normal con parámetros estimados, considere la posibilidad de utilizar la prueba de K-S Lilliefors (disponible en el procedimiento Explorar).

Para obtener una prueba de Kolmogorov-Smirnov para una muestra

1. Seleccione en los menús:

   Analizar > **Pruebas no paramétricas** > **Cuadros de diálogo antiguos** > **K-S de 1 muestra...**
2. Seleccione una o más variables de contraste numéricas. Cada variable genera una prueba independiente.
3. Si lo desea, puede pulsar en **Opciones** para obtener estadísticos descriptivos, cuartiles y controlar el tratamiento de los datos perdidos.

Prueba de Kolmogorov-Smirnov para una muestra: Opciones

Estadísticos. Puede elegir uno o los dos estadísticos de resumen.

- **Descriptivos.** Muestra la media, la desviación estándar, el mínimo, el máximo y el número de casos no perdidos.
- **Cuartiles.** Muestra los valores correspondientes a los percentiles 25, 50 y 75.

Valores perdidos. Controla el tratamiento de los valores perdidos.

- **Excluir casos según prueba.** Cuando se especifican varias pruebas, cada una se evalúa separadamente respecto a los valores perdidos.
- **Excluir casos según lista.** Los casos con valores perdidos para cualquier variable se excluyen de todos los análisis.

Características adicionales del comando NPAR TESTS (Prueba de Kolmogorov-Smirnov para una muestra)

El lenguaje de sintaxis de comandos también permite especificar los parámetros de la distribución de prueba (con el subcomando K-S).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Pruebas para dos muestras independientes

El procedimiento Pruebas para dos muestras independientes compara dos grupos de casos existentes en una variable.

Ejemplo. Se han desarrollado nuevos correctores dentales diseñados para que sean más cómodos y estéticos, así como para facilitar un progreso más rápido en la realineación de la dentadura. Para averiguar si el nuevo corrector debe llevarse tanto tiempo como el modelo antiguo, se eligen 10 niños al azar para que lleven este último y otros 10 niños para que usen el nuevo. Mediante la prueba *U* de Mann-Whitney podría descubrir que, de media, los niños que llevaban el nuevo corrector tenían que llevarlo puesto menos tiempo que los que llevaban el antiguo.

Estadísticos. Media, desviación estándar, mínimo, máximo, número de casos no perdidos y cuartiles. Pruebas: *U* de Mann-Whitney, reacciones extremas de Moses, *Z* de Kolmogorov-Smirnov, rachas de Wald-Wolfowitz.

Pruebas para dos muestras independientes: Consideraciones sobre los datos

Datos. Utilice variables numéricas que puedan ser ordenables.

Supuestos. Utilice muestras independientes y aleatorias. La prueba Mann-Whitney *U* comprueba la igualdad de las dos distribuciones. Para utilizarla para comprobar sus diferencias en la ubicación entre dos distribuciones, se debe asumir que las distribuciones tienen la misma forma.

Para obtener pruebas para dos muestras independientes

1. Seleccione en los menús:

   Analizar > **Pruebas no paramétricas** > **Cuadros de diálogo antiguos** > **2 muestras independientes...**
2. Seleccione una o más variables numéricas.
3. Seleccione una variable de agrupación y pulse en **Definir grupos** para segmentar el archivo en dos grupos o muestras.

Tipos de pruebas para dos muestras independientes.

Tipo de prueba. Hay cuatro pruebas disponibles para contrastar si dos muestras (grupos) independientes proceden de una misma población.

La **prueba U de Mann-Whitney** es la más conocida de las pruebas para dos muestras independientes. Es equivalente a la prueba de la suma de rangos de Wilcoxon y a la prueba de Kruskal-Wallis para dos grupos. La prueba de Mann-Whitney contrasta si dos poblaciones muestreadas son equivalentes en su posición. Las observaciones de ambos grupos se combinan y clasifican, asignándose el rango de promedio en caso de producirse empates. El número de empates debe ser pequeño en relación con el número total de observaciones. Si la posición de las poblaciones es idéntica, los rangos deberían mezclarse aleatoriamente entre las dos muestras. La prueba calcula el número de veces que una puntuación del grupo 1 precede a una puntuación del grupo 2 y el número de veces que una puntuación del grupo 2 precede a una puntuación del grupo 1. El estadístico *U* de Mann-Whitney es el menor de estos dos números. También se muestra el estadístico *W* de la suma de rangos de Wilcoxon. *W* es la suma de los rangos del grupo en el rango menor, salvo que los grupos tengan el mismo rango medio, en cuyo caso es la suma de rangos del grupo que se nombra en último lugar en el cuadro de diálogo Dos muestras independientes: Definir grupos.

La **prueba Z de Kolmogorov-Smirnov** y la **prueba de rachas de Wald-Wolfowitz** son pruebas más generales que detectan las diferencias entre las posiciones y las formas de las distribuciones. La prueba 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. 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.

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.

Pruebas para dos muestras independientes: Definir grupos

Para segmentar el archivo en dos grupos o muestras, introduzca un valor entero para el Grupo 1 y otro valor para el Grupo 2. Los casos con otros valores se excluyen del análisis.

Pruebas para dos muestras independientes: Opciones

Estadísticos. Puede elegir uno o los dos estadísticos de resumen.

- **Descriptivos.** Muestra la media, la desviación estándar, el mínimo, el máximo y el número de casos no perdidos.
- **Cuartiles.** Muestra los valores correspondientes a los percentiles 25, 50 y 75.

Valores perdidos. Controla el tratamiento de los valores perdidos.

- **Excluir casos según prueba.** Cuando se especifican varias pruebas, cada una se evalúa separadamente respecto a los valores perdidos.
- **Excluir casos según lista.** Los casos con valores perdidos para cualquier variable se excluyen de todos los análisis.

Características adicionales del comando NPAR TESTS (Dos muestras independientes)

El lenguaje de sintaxis de comandos permite especificar el número de casos que se recortarán en la prueba de Moses (con el subcomando `MOSES`).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Pruebas para dos muestras relacionadas

El procedimiento Pruebas para dos muestras relacionadas compara las distribuciones de dos variables.

Ejemplo. En general, cuando una familia vende su casa ¿logra obtener la cantidad que pide inicialmente? Si aplica la prueba de Wilcoxon de los rangos con signo a 10 casas, podría descubrir que siete familias reciben menos cantidad de la solicitada, una recibe más y dos familias reciben el precio solicitado.

Estadísticos. Media, desviación estándar, mínimo, máximo, número de casos no perdidos y cuartiles. Pruebas: Wilcoxon de los rangos con signo, signo, McNemar. Si se ha instalado la opción Pruebas exactas (disponible sólo en los sistemas operativos Windows), la prueba de homogeneidad marginal también estará disponible.

Pruebas para dos muestras relacionadas: Consideraciones sobre los datos

Datos. Utilice variables numéricas que puedan ser ordenables.

Supuestos. Aunque no se suponen distribuciones en particular para las dos variables, se supone que la distribución de población de las diferencias emparejadas es simétrica.

Para obtener pruebas para dos muestras relacionadas

1. Seleccione en los menús:

   Analizar > **Pruebas no paramétricas** > **Cuadros de diálogo antiguos** > **2 muestras relacionadas...**
2. Seleccione uno o más pares de variables.

Dos muestras relacionadas: Tipos de pruebas

Las pruebas de esta sección comparan las distribuciones de dos variables relacionadas. La prueba apropiada depende del tipo de datos.

Si los datos son continuos, use la prueba de los signos o la prueba de Wilcoxon de los rangos con signo. 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 las dos variables tienen una distribución similar, el número de diferencias positivas y negativas no difiere de forma significativa. 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.

Si los datos son binarios, use la **prueba de McNemar**. Esta prueba se utiliza normalmente en una situación de medidas repetidas, en la que la respuesta de cada sujeto se obtiene dos veces, una antes y otra después de que ocurra un evento especificado. La prueba de McNemar determina si el índice de respuesta inicial (antes del evento) es igual al índice de respuesta final (después del evento). Esta prueba es útil para detectar cambios en las respuestas causadas por la intervención experimental en los diseños del tipo antes-después.

Si los datos son categóricos, use la **prueba de homogeneidad marginal**. Se trata de 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, y es útil para detectar cambios de respuesta causados por intervención experimental en diseños antes-después. La prueba de homogeneidad marginal sólo está disponible si se ha instalado Pruebas exactas.

Pruebas para dos muestras relacionadas: Opciones

Estadísticos. Puede elegir uno o los dos estadísticos de resumen.

- **Descriptivos.** Muestra la media, la desviación estándar, el mínimo, el máximo y el número de casos no perdidos.
- **Cuartiles.** Muestra los valores correspondientes a los percentiles 25, 50 y 75.

Valores perdidos. Controla el tratamiento de los valores perdidos.

- **Excluir casos según prueba.** Cuando se especifican varias pruebas, cada una se evalúa separadamente respecto a los valores perdidos.
- **Excluir casos según lista.** Los casos con valores perdidos para cualquier variable se excluyen de todos los análisis.

Características adicionales del comando NPAR TESTS (Dos muestras relacionadas)

El lenguaje de sintaxis de comandos también permite contrastar una variable con cada variable de la lista.

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Pruebas para varias muestras independientes

El procedimiento Pruebas para varias muestras independientes compara dos o más grupos de casos respecto a una variable.

Ejemplo. ¿Es diferente el tiempo medio en que se fundirán las bombillas de 100 vatios de tres marcas distintas? A partir del análisis de varianza de un factor de Kruskal-Wallis, puede comprobar que las tres marcas sí se diferencian en su vida media.

Estadísticos. Media, desviación estándar, mínimo, máximo, número de casos no perdidos y cuartiles. Pruebas: *H* de Kruskal-Wallis, de la mediana.

Pruebas para varias muestras independientes: Consideraciones sobre los datos

Datos. Utilice variables numéricas que puedan ser ordenables.

Supuestos. Utilice muestras independientes y aleatorias. La prueba *H* de Kruskal-Wallis requiere que las muestras comparadas tengan formas similares.

Para obtener pruebas para varias muestras independientes

1. Seleccione en los menús:

   Analizar > **Pruebas no paramétricas** > **Cuadros de diálogo antiguos** > **K muestras independientes...**
2. Seleccione una o más variables numéricas.
3. Seleccione una variable de agrupación y pulse en **Definir rango** para especificar los valores enteros máximo y mínimo para la variable de agrupación.

Tipos de pruebas para varias muestras independientes

Se hallan disponibles tres pruebas para determinar si varias muestras independientes proceden de la misma población. La prueba *H* de Kruskal-Wallis, la prueba de la mediana y la prueba de Jonckheere-Terpstra contrastan si varias muestras independientes proceden de la misma población.

La prueba **H de Kruskal-Wallis**, una extensión de la prueba *U* de Mann-Whitney, es el análogo no paramétrico del análisis de varianza de un factor y detecta las diferencias en la localización de las distribuciones. La **prueba de la mediana**, que es una prueba más general pero no tan potente, detecta diferencias distribucionales en la localización y en la forma. La prueba *H* de Kruskal-Wallis y la prueba de la mediana suponen que no existe una ordenación *a priori* de las poblaciones *k* de las cuales se extraen las muestras.

Cuando *existe* una ordenación natural *a priori* (ascendente o descendente) de las poblaciones *k*, la prueba **Jonckheere-Terpstra** es más potente. 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. La prueba de Jonckheere-Terpstra estará disponible sólo si ha instalado el módulo adicional Pruebas exactas.

Varias muestras independientes: Definir rango

Para definir el rango, introduzca valores enteros para el **mínimo** y el **máximo** que se correspondan con las categorías mayor y menor de la variable de agrupación. Se excluyen los casos con valores fuera de los límites. Por ejemplo, si indica un valor mínimo de 1 y un valor máximo de 3, únicamente se utilizarán los valores enteros entre 1 y 3. Debe indicar ambos valores y el valor mínimo ha ser menor que el máximo.

Varias muestras independientes: Opciones

Estadísticos. Puede elegir uno o los dos estadísticos de resumen.

- **Descriptivos.** Muestra la media, la desviación estándar, el mínimo, el máximo y el número de casos no perdidos.
- **Cuartiles.** Muestra los valores correspondientes a los percentiles 25, 50 y 75.

Valores perdidos. Controla el tratamiento de los valores perdidos.

- **Excluir casos según prueba.** Cuando se especifican varias pruebas, cada una se evalúa separadamente respecto a los valores perdidos.
- **Excluir casos según lista.** Los casos con valores perdidos para cualquier variable se excluyen de todos los análisis.

Características adicionales del comando NPAR TESTS (K muestras independientes)

El lenguaje de sintaxis de comandos permite especificar un valor distinto al de la mediana observada para la prueba de la mediana (mediante el subcomando `MEDIAN`).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Pruebas para varias muestras relacionadas

El procedimiento Pruebas para varias muestras relacionadas compara las distribuciones de dos o más variables.

Ejemplo. ¿Asocia la gente diferentes niveles de prestigio a doctores, abogados, policías y profesores? Se pide a diez personas que ordenen estas cuatro profesiones por orden de prestigio. La prueba de Friedman indica que la gente asocia diferentes niveles de prestigio con estas cuatro profesiones.

Estadísticos. Media, desviación estándar, mínimo, máximo, número de casos no perdidos y cuartiles. Pruebas: Friedman, *W* de Kendall y *Q* de Cochran.

Pruebas para varias muestras relacionadas: Consideraciones sobre los datos

Datos. Utilice variables numéricas que puedan ser ordenables.

Supuestos. Las pruebas no paramétricas no requieren supuestos sobre la forma de la distribución subyacente. Utilice muestras aleatorias y dependientes.

Para obtener pruebas para varias muestras relacionadas

1. Seleccione en los menús:

   Analizar > **Pruebas no paramétricas** > **Cuadros de diálogo antiguos** > **K muestras relacionadas...**
2. Seleccione dos o más variables de contraste numéricas.

Tipos de prueba en el procedimiento Pruebas para varias muestras relacionadas

Hay tres pruebas disponibles para comparar las distribuciones de diversas variables relacionadas.

La **prueba de Friedman** es el equivalente no paramétrico de un diseño de medidas repetidas para una muestra o un análisis de varianza bidimensional con una observación por casilla. Friedman contrasta la hipótesis nula de que las *k* variables relacionadas procedan de la misma población. Para cada caso, a las *k* variables se les asignan los rangos 1 a *k*. El estadístico de contraste se basa en estos rangos.

La **W de Kendall** es una normalización del estadístico de Friedman. La prueba *W* de Kendall se puede interpretar como el coeficiente de concordancia, que es una medida de acuerdo entre evaluadores. Cada caso es un juez o evaluador y cada variable es un elemento o persona que está siendo evaluada. Para cada variable, se calcula la suma de rangos. La *W* de Kendall varía entre 0 (no hay acuerdo) y 1 (acuerdo completo).

La prueba **Q de Cochran** es idéntica a la prueba de Friedman pero se puede aplicar cuando todas las respuestas son binarias. Esta prueba es una extensión de la prueba de McNemar para la situación de *k* muestras. La *Q* de Cochran contrasta la hipótesis de que diversas variables dicotómicas relacionadas tienen la misma media. Las variables se miden al mismo individuo o a individuos emparejados.

Pruebas para varias muestras relacionadas: Estadísticos

Puede elegir estadísticos.

- **Descriptivos.** Muestra la media, la desviación estándar, el mínimo, el máximo y el número de casos no perdidos.
- **Cuartiles.** Muestra los valores correspondientes a los percentiles 25, 50 y 75.

Características adicionales del comando NPAR TESTS (K pruebas relacionadas)

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 28. Análisis de respuestas múltiples

Análisis de respuestas múltiples

Se ofrecen dos procedimientos para analizar los conjuntos de categorías múltiples y de dicotomías múltiples. El procedimiento Frecuencias de respuestas múltiples muestra tablas de frecuencias. El procedimiento Tablas cruzadas de respuestas múltiples muestra tabulaciones cruzadas de dos y tres dimensiones. Antes de utilizar cualquiera de estos procedimientos, deberá definir conjuntos de respuestas múltiples.

Ejemplo. Este ejemplo ilustra el uso de elementos de respuestas múltiples en un estudio de investigación de mercado. Los datos son ficticios y no deben interpretarse como reales. Una línea aérea podría hacer una encuesta a los pasajeros que realicen una determinada ruta para evaluar las líneas aéreas de la competencia. En este ejemplo, American Airlines desea conocer el uso que hacen sus pasajeros de otras líneas aéreas en la ruta Chicago-Nueva York y la importancia relativa del horario y el servicio a la hora de seleccionar una línea aérea. El encargado del vuelo proporciona a cada pasajero un breve cuestionario durante el embarque. La primera pregunta dice: rodee con un círculo todas las líneas aéreas con la que haya volado al menos una vez en los últimos seis meses en este mismo trayecto: American, United, TWA, USAir, Otras. Se trata de una pregunta de respuestas múltiples, ya que el pasajero puede marcar más de una respuesta. Sin embargo, la pregunta no se puede codificar directamente, ya que una variable sólo puede tener un valor para cada caso. Deberá utilizar distintas variables para correlacionar las respuestas con cada pregunta. Existen dos formas de hacerlo. Una consiste en definir una variable para cada una de las opciones (por ejemplo, American, United, TWA, USAir y Otras). Si el pasajero marca United, a la variable *united* se le asignará el código 1; en caso contrario se le asignará 0. Éste es un **método de dicotomías múltiples** de correlación de variables. La otra forma de correlacionar respuestas es el **método de categorías múltiples**, en el que se estima el número máximo de posibles respuestas a la pregunta y se configura el mismo número de variables, con códigos para especificar la línea aérea utilizada. Examinando una muestra de cuestionarios, podría observarse que ningún usuario ha volado en más de tres líneas aéreas diferentes en esta ruta durante los últimos seis meses. Aún más, se observará que debido a la liberalización de las líneas aéreas, aparecen otras 10 en la categoría Otras. Con el método de respuestas múltiples, definiría tres variables, cada una codificada como 1 = *american*, 2 = *united*, 3 = *twa*, 4 = *usair*, 5 = *delta* y así sucesivamente. Si un pasajero determinado marca American y TWA, la primera variable tendrá el código 1, la segunda el 3 y la tercera un código de valor perdido. Otro pasajero podría haber marcado American e introducido Delta. Así, la primera variable tendrá el código 1, la segunda el 5 y la tercera un código de valor perdido. Por el contrario, si utiliza el método de dicotomías múltiples, terminará con 14 variables independientes. Aunque cualquiera de los métodos de correlación anteriores es viable para este estudio, el método seleccionado dependerá de la distribución de respuestas.

Definir conjuntos de respuestas múltiples

El procedimiento Definir conjuntos de respuestas múltiples agrupa variables elementales en conjuntos de categorías múltiples y de dicotomías múltiples, para los que se pueden obtener tablas de frecuencias y tabulaciones cruzadas. Se pueden definir hasta 20 conjuntos de respuestas múltiples. Cada conjunto debe tener un nombre exclusivo. Para eliminar un conjunto, resáltelo en la lista de conjuntos de respuestas múltiples y pulse en **Borrar**. Para cambiar un conjunto, resáltelo en la lista, modifique cualquier característica de la definición del conjunto y pulse en **Cambiar**.

Las variables elementales se pueden codificar como dicotomías o categorías. Para utilizar variables dicotómicas, seleccione **Dicotomías** para crear un conjunto de dicotomías múltiples. Introduzca un valor entero en Valor contado. Cada variable que tenga al menos una aparición del valor contado se convierte en una categoría del conjunto de dicotomías múltiples. Seleccione **Categorías** para crear un conjunto de categorías múltiples con el mismo rango de valores que las variables que lo componen. Introduzca valores enteros para los valores máximo y mínimo del rango para las categorías del conjunto de

categorías múltiples. El procedimiento suma cada valor entero distinto en el rango inclusivo para todas las variables que lo componen. Las categorías vacías no se tabulan.

A cada conjunto de respuestas múltiples se le debe asignar un nombre exclusivo de hasta siete caracteres. El procedimiento coloca delante del nombre asignado un signo dólar ($). No se pueden utilizar los siguientes nombres reservados: *casenum*, *sysmis*, *jdate*, *date*, *time*, *length* y *width*. El nombre del conjunto de respuestas múltiples sólo se encuentra disponible para su uso en los procedimientos de respuestas múltiples. No se puede hacer referencia a nombres de conjuntos de respuestas múltiples en otros procedimientos. Si lo desea, puede introducir una etiqueta de variable descriptiva para el conjunto de respuestas múltiples. La etiqueta puede tener hasta 40 caracteres.

Para definir conjuntos de respuestas múltiples

1. Seleccione en los menús:

   Analizar > **Respuesta múltiple** > **Definir conjuntos de variables...**
2. Seleccione dos o más variables.
3. Si las variables están codificadas como dicotomías, indique qué valor desea contar. Si las variables están codificadas como categorías, defina el rango de las categorías.
4. Escriba un nombre exclusivo para cada conjunto de respuestas múltiples.
5. Pulse **Añadir** para añadir el conjunto de respuestas múltiples a la lista de conjuntos definidos.

Frecuencias de respuestas múltiples

El procedimiento Frecuencias de respuestas múltiples produce tablas de frecuencias para conjuntos de respuestas múltiples. En primer lugar es necesario definir uno o más conjuntos de respuestas múltiples (véase "Definir conjuntos de respuestas múltiples").

Para los conjuntos de dicotomías múltiples, los nombres de categorías que se muestran en los resultados proceden de etiquetas de variable definidas para variables elementales del grupo. Si las etiquetas de variable no están definidas, los nombres de las variables se utilizarán como etiquetas. Para los conjuntos de categorías múltiples, las etiquetas de categoría proceden de las etiquetas de valor de la primera variable del grupo. Si las categorías perdidas para la primera variable están presentes para otras variables del grupo, defina una etiqueta de valor para las categorías perdidas.

Valores perdidos. Los casos con valores perdidos se excluyen en base a tabla por tabla. Si lo desea, puede seleccionar una de las opciones siguientes o ambas:

- **Excluir los casos según lista dentro de las dicotomías.** Excluye los casos con valores perdidos en cualquier variable de la tabulación del conjunto de dicotomías múltiples. Esto sólo se aplica a conjuntos de respuestas múltiples definidos como conjuntos de dicotomías. De forma predeterminada, un caso se considera perdido para un conjunto de dicotomías múltiples si ninguna de sus variables que lo componen contiene el valor contado. Los casos con valores perdidos en algunas variables, pero no en todas, se incluyen en las tabulaciones del grupo si al menos una variable contiene el valor contado.
- **Excluir los casos según lista dentro de las categorías.** Excluye los casos con valores perdidos en cualquier variable de la tabulación del conjunto de categorías múltiples. Esto sólo se aplica a conjuntos de respuestas múltiples definidos como conjuntos de categorías. De forma predeterminada, un caso se considera perdido para un conjunto de categorías múltiples sólo si ninguno de sus componentes tiene valores válidos dentro del rango definido.

Ejemplo. Cada variable creada a partir de una pregunta de una encuesta es una variable elemental. Para analizar un elemento de respuestas múltiples, deberá combinar las variables en uno o dos tipos de conjuntos de respuestas múltiples: un conjunto de dicotomías múltiples o un conjunto de categorías múltiples. Por ejemplo, si una encuesta sobre líneas aéreas preguntara al encuestado cuál de las tres líneas (American, United, TWA) ha utilizado durante los seis últimos meses y usted utilizara variables dicotómicas y definiera un **conjunto de dicotomías múltiples**, cada una de las tres variables del conjunto se convertiría en una categoría de la variable de grupo. Las frecuencias y los porcentajes de las tres líneas

aéreas se muestran en una tabla de frecuencias. Si observa que ningún encuestado ha mencionado más de dos líneas aéreas, podría crear dos variables, cada una con tres códigos, uno para cada línea aérea. Si define un **conjunto de categorías múltiples**, los valores se tabulan añadiendo los mismos códigos en las variables elementales juntas. El conjunto de valores resultantes es igual a los de cada una de las variables elementales. Por ejemplo, 30 respuestas para United son la suma de las cinco respuestas de United para la línea aérea 1 y las 25 respuestas de United para la línea aérea 2. Las frecuencias y los porcentajes de las tres líneas aéreas se muestran en una tabla de frecuencias.

Estadísticos. Tablas de frecuencias que muestran recuentos, porcentajes de respuestas, porcentajes de casos, número de casos válidos y número de casos perdidos.

Frecuencias de respuestas múltiples: Consideraciones sobre los datos

Datos. Utilice conjuntos de respuestas múltiples.

Supuestos. Las frecuencias y los porcentajes proporcionan una descripción útil de los datos de cualquier distribución.

Procedimientos relacionados. El procedimiento Definir conjuntos de respuestas múltiples permite definir este tipo de conjuntos.

Para obtener frecuencias de respuestas múltiples

1. Seleccione en los menús:

   Analizar > **Respuesta múltiple** > **Frecuencias...**
2. Seleccione uno o más conjuntos de respuestas múltiples.

Tablas cruzadas de respuestas múltiples

El procedimiento Tablas cruzadas de respuestas múltiples presenta en forma de tabulación cruzada conjuntos de respuestas múltiples, variables elementales o una combinación. También puede obtener porcentajes de casillas basados en casos o respuestas, modificar la gestión de los valores perdidos u obtener tabulaciones cruzadas emparejadas. Antes debe definir uno o varios conjuntos de respuestas múltiples (véase “Para definir conjuntos de respuestas múltiples”).

Para los conjuntos de dicotomías múltiples, los nombres de categorías que se muestran en los resultados proceden de etiquetas de variable definidas para variables elementales del grupo. Si las etiquetas de variable no están definidas, los nombres de las variables se utilizarán como etiquetas. Para los conjuntos de categorías múltiples, las etiquetas de categoría proceden de las etiquetas de valor de la primera variable del grupo. Si las categorías perdidas para la primera variable están presentes para otras variables del grupo, defina una etiqueta de valor para las categorías perdidas. El procedimiento muestra las etiquetas de categoría por columnas en tres líneas, con un máximo de ocho caracteres por línea. Para evitar la división de palabras, puede invertir los elementos de las filas y las columnas o volver a definir las etiquetas.

Ejemplo. Tanto los conjuntos de categorías múltiples como los conjuntos de dicotomías múltiples se pueden presentar en forma de tabulación cruzada con otras variables de este procedimiento. Un estudio sobre pasajeros de líneas aéreas solicita a éstos la siguiente información: marque las líneas aéreas con las que ha volado al menos una vez en los seis últimos meses (American, United, TWA). ¿Qué considera más importante a la hora de seleccionar un vuelo, el horario o el servicio? Seleccione sólo uno. Después de introducir los datos como dicotomías o categorías múltiples y combinarlos en un conjunto, puede presentar en forma de tabulación cruzada las selecciones de línea aérea con la pregunta relativa al servicio o al horario.

Estadísticos. Tabulación cruzada con recuentos de casilla, fila, columna y totales, así como porcentajes de casillas, filas, columnas y totales. Los porcentajes de casillas se basan en casos o respuestas.

Tablas cruzadas de respuestas múltiples: Consideraciones sobre los datos

Datos. Utilice conjuntos de respuestas múltiples o variables categóricas numéricas.

Supuestos. Las frecuencias y los porcentajes proporcionan una útil descripción de los datos de cualquier distribución.

Procedimientos relacionados. El procedimiento Definir conjuntos de respuestas múltiples permite definir este tipo de conjuntos.

Para obtener tablas cruzadas de respuestas múltiples

1. Seleccione en los menús:

   Analizar > **Respuesta múltiple** > **Tablas cruzadas...**
2. Seleccione una o más variables numéricas o conjuntos de respuestas múltiples para cada dimensión de la tabulación cruzada.
3. Defina el rango de cada variable elemental.

Si lo desea, puede obtener una tabulación cruzada bidimensional para cada categoría de una variable de control o conjunto de respuestas múltiples. Seleccione uno o varios elementos para la lista Capas.

Tablas de respuestas múltiples: Definir rangos de las variables

Los rangos de valores deben definirse para cualquier variable elemental de la tabulación cruzada. Introduzca los valores enteros de categoría máximos y mínimos que desee tabular. Las categorías que estén fuera del rango se excluyen del análisis. Se entiende que los valores que estén dentro del rango inclusivo son enteros (los no enteros quedan truncados).

Tablas cruzadas de respuestas múltiples: Opciones

Porcentajes de casillas. Los recuentos de casillas siempre se muestran. Puede elegir entre mostrar los porcentajes de fila, los de columna o los de tabla bidimensional (totales).

Porcentajes basados en. Los porcentajes de casillas pueden basarse en casos (o encuestados). Esta opción no estará disponible si selecciona la concordancia de variables en conjuntos de categorías múltiples. También se pueden basar en las respuestas. Para los conjuntos de dicotomías múltiples, el número de respuestas es igual al número de valores contados por los casos. Para los conjuntos de categorías múltiples, el número de respuestas es el número de valores del rango definido.

Valores perdidos. Puede elegir una o ambas de las siguientes opciones:

- **Excluir los casos según lista dentro de las dicotomías.** Excluye los casos con valores perdidos en cualquier variable de la tabulación del conjunto de dicotomías múltiples. Esto sólo se aplica a conjuntos de respuestas múltiples definidos como conjuntos de dicotomías. De forma predeterminada, un caso se considera perdido para un conjunto de dicotomías múltiples si ninguna de sus variables que lo componen contiene el valor contado. Los casos con valores perdidos para algunas variables, pero no todas, se incluyen en las tabulaciones del grupo si al menos una variable contiene el valor contado.
- **Excluir los casos según lista dentro de las categorías.** Excluye los casos con valores perdidos en cualquier variable de la tabulación del conjunto de categorías múltiples. Esto sólo se aplica a conjuntos de respuestas múltiples definidos como conjuntos de categorías. De forma predeterminada, un caso se considera perdido para un conjunto de categorías múltiples sólo si ninguno de sus componentes tiene valores válidos dentro del rango definido.

De forma predeterminada, cuando se presentan dos conjuntos de categorías múltiples en forma de tabulación cruzada, el procedimiento tabula cada variable del primer grupo con cada variable del segundo y suma las frecuencias de cada casilla; de esta forma, algunas respuestas pueden aparecer más de una vez en una tabla. Puede seleccionar la opción siguiente:

Emparejar las variables entre los conjuntos de respuesta. Empareja la primera variable del primer grupo con la primera variable del segundo, y así sucesivamente. Si selecciona esta opción, el procedimiento basará los porcentajes de casillas en las respuestas en lugar de hacerlo en los encuestados. El emparejamiento no está disponible para conjuntos de dicotomías múltiples o variables elementales.

Características adicionales del comando MULT RESPONSE

La sintaxis de comandos también le permite:

- Obtener tablas de tabulación cruzada con un máximo de cinco dimensiones (con el subcomando `BY`).
- Cambiar las opciones de formato de salida, incluyendo la supresión de etiquetas de valor (con el subcomando `FORMAT`).

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 29. Informes de los resultados

Informes de los resultados

Los listados de casos y los estadísticos descriptivos son herramientas básicas para estudiar y presentar los datos. Puede obtener listados de casos con el Editor de datos o el procedimiento Resumir, frecuencias y estadísticos descriptivos con el procedimiento Frecuencias, y estadísticos de subpoblación con el procedimiento Medias. Cada uno utiliza un formato diseñado para que la información sea clara. Si desea ver la información con otro formato, las opciones Informe de estadísticos en filas e Informe de estadísticos en columnas le ofrecen el control que precisa para presentar los datos.

Informe de estadísticos en filas

Informe de estadísticos en filas genera informes en los cuales se presentan distintos estadísticos de resumen en filas. También se encuentran disponibles listados de los casos, con o sin estadísticos de resumen.

Ejemplo. Una empresa con una cadena de tiendas registra los datos de sus empleados, incluyendo el salario, el cargo, la tienda y la sección en la que trabaja cada uno. Se podría generar un informe que proporcione los datos individuales de cada empleado (listado) desglosados por tienda y sección (variables de segmentación), con estadísticos de resumen (por ejemplo, el salario medio) por tienda, sección y sección dentro de cada tienda.

Columnas de datos. Muestra una lista de las variables del informe para las que desea obtener el listado de los casos o los estadísticos de resumen y controla el formato de presentación de las columnas de datos.

Salto de columna. Muestra una lista de las variables de segmentación opcionales que dividen el informe en grupos y controla los estadísticos de resumen y los formatos de presentación de Salto de columna. Si hay varias variables de segmentación, se creará un grupo distinto para cada una de las categorías de las variables de segmentación dentro de las categorías de la variable de segmentación anterior en la lista. Las variables de segmentación deben ser variables categóricas discretas que dividan los casos en un número limitado de categorías con sentido. Los valores individuales de cada variable de segmentación aparecen ordenados en una columna distinta situada a la izquierda de todas las columnas de datos.

Informe. Controla las características globales del informe, incluyendo los estadísticos de resumen globales, la presentación de los valores perdidos, la numeración de las páginas y los títulos.

Mostrar casos. Muestra los valores reales (o etiquetas de valor) de las variables de la columna de datos para cada caso. Esto genera un informe a modo de listado, que puede ser mucho más largo que un informe de resumen.

Vista previa. Muestra sólo la primera página del informe. Esta opción es útil para ver una vista previa del formato del informe sin tener que procesar el informe completo.

Los datos están ordenados. Para los informes con variables de segmentación, el archivo de datos se debe ordenar por los valores de estas variables antes de generar el informe. Si el archivo de datos ya está ordenado por estos valores, se puede ahorrar tiempo de procesamiento seleccionando esta opción. Esta opción es especialmente útil después de generar la vista previa de un informe.

Para obtener un informe de resumen: Estadísticos en filas

1. Seleccione en los menús:

   Analizar > **Informes** > **Informe de estadísticos en filas...**

2. Seleccione una o más variables para las columnas de datos. En el informe se genera una columna para cada variable seleccionada.
3. Para los informes ordenados y mostrados por subgrupos, seleccione una o más variables para Salto de columnas.
4. Para los informes con estadísticos de resumen para los subgrupos definidos por las variables de segmentación, seleccione la variable de segmentación de la lista de variables de Salto de columna y pulse en **Resumen**, en la sección Salto de columna, para especificar las medidas del resumen.
5. Para los informes con estadísticos de resumen globales, pulse en **Resumen** para especificar las medidas de resumen.

Formato de los saltos y de las columnas de datos del informe

Los cuadros de diálogo de formato controlan los títulos y el ancho de las columnas, la alineación del texto y la presentación de los valores de los datos o de las etiquetas de valor. El formato de las columnas de datos controla el formato de las columnas de datos situadas en la parte derecha de la página del informe. La opción Formato de salto controla el formato de los saltos de columna situadas en la parte izquierda.

Título de la columna. Para la variable seleccionada, controla el título de la columna. Los títulos largos se ajustan de forma automática dentro de la columna. Utilice la tecla Intro para insertar manualmente líneas de separación donde desee ajustar los títulos.

Posición de valor en la columna. Para la variable seleccionada, controla la alineación de los valores de los datos o de las etiquetas de valor dentro de la columna. La alineación de los valores o de las etiquetas no afecta a la alineación de los encabezados de las columnas. Puede sangrar el contenido de la columna por un número específico de caracteres o centrar el contenido.

Contenido de la columna. Para la variable seleccionada, controla la presentación de los valores de los datos o de las etiquetas de valor definidas. Los valores de los datos siempre se muestran para cualquier valor que no tenga etiquetas de valor definidas. No se encuentra disponible para las columnas de datos en los informes estadísticos en columnas.

Líneas de resumen finales y Líneas de resumen del informe

Los dos cuadros de diálogo Líneas de resumen controlan la presentación de los estadísticos de resumen para los grupos de ruptura y para el informe entero. Líneas de resumen controla los estadísticos de subgrupo para cada categoría definida por las variables de segmentación. Líneas de resumen finales controla los estadísticos globales que se muestran al final del informe.

Los estadísticos de resumen disponibles son: suma, media, valor mínimo, valor máximo, número de casos, porcentaje de casos por encima y por debajo de un valor especificado, porcentaje de casos dentro de un rango de valores especificado, desviación estándar, curtosis, varianza y asimetría.

Opciones de saltos del informe

Opciones de saltos controla el espaciado y la paginación de la información de la categoría de salto.

Control de página. Controla el espaciado y la paginación para las categorías de la variable de segmentación seleccionada. Puede especificar un número de líneas en blanco entre las categorías de segmentación o empezar cada categoría de segmentación en una página nueva.

Líneas en blanco antes de los estadísticos. Controla el número de líneas en blanco entre las etiquetas o los datos de la categoría de ruptura y los estadísticos de resumen. Esta opción es especialmente útil para

los informes combinados que incluyan tanto el listado de los casos individuales como los estadísticos de resumen para las categorías de segmentación; en estos informes puede insertar un espacio entre el listado de los casos y los estadísticos de resumen.

Opciones del informe

Informe: Opciones controla el tratamiento y la presentación de los valores perdidos y la numeración de las páginas del informe.

Excluir casos con valores perdidos según lista. Elimina (del informe) cualquier caso con valores perdidos para cualquier variable del informe.

Los valores perdidos aparecen como. Permite especificar el símbolo que representa los valores perdidos en el archivo de datos. Este símbolo sólo puede tener un carácter y se utiliza para representar tanto los valores *perdidos del sistema* como los valores *perdidos del usuario*.

Numerar las páginas desde la. Permite especificar un número de página para la primera página del informe.

Diseño del informe

Informe: Diseño controla el ancho y alto de cada página del informe, la ubicación del informe dentro de la página y la inserción de etiquetas y líneas en blanco.

Diseño de página. Controla los márgenes de las páginas expresados en líneas (extremos superior e inferior) y caracteres (a la izquierda y a la derecha) y la alineación del informe entre los márgenes.

Títulos y pies de página. Controla el número de líneas que separan los títulos y los pies de página del cuerpo del informe.

Salto de columna. Controla la presentación de los saltos de columna. Si se especifican diversas variables de segmentación, pueden situarse en columnas diferentes o en la primera columna. Si se colocan todas en la primera columna, se generará un informe más estrecho.

Títulos de columna. Controla la presentación de los títulos de columna, incluyendo el subrayado de títulos, el espacio entre los títulos y el cuerpo del informe y la alineación vertical de los títulos de columna.

Filas de col. datos y etiquetas de salto. Controla la ubicación de la información de las columnas de datos (valores de datos o estadísticos de resumen) en relación con las etiquetas de salto al principio de cada categoría de ruptura. La primera fila de información puede empezar en la misma línea que la etiqueta de categoría de ruptura o en un número de líneas posterior especificado. Esta sección no se encuentra disponible para los informes de estadísticos en columnas.

Títulos del informe

Informe: Títulos controla el contenido y la ubicación de los títulos y los pies de página del informe. Puede especificar un máximo de diez líneas de títulos de página y otras tantas de pies de página, con componentes justificados a la izquierda, en el centro y a la derecha en cada línea.

Si inserta variables en los títulos o en los pies de página, la etiqueta de valor o el valor de la variable actual aparecerá en el título o en el pie de página. Para los títulos se mostrará la etiqueta de valor correspondiente al valor de la variable al principio de la página. Para los pies de página, esta etiqueta se mostrará al final de la página. Si no hay etiqueta de valor, se mostrará el valor real.

Variables especiales. Las variables especiales *DATE* y *PAGE* permiten insertar la fecha actual o el número de página en cualquier línea de un encabezado o pie del informe. Si el archivo de datos contiene variables llamadas *DATE* o *PAGE*, no podrá utilizar estas variables en los títulos ni en los pies del informe.

Informe de estadísticos en columnas

Informe de estadísticos en columnas genera informes de resumen en los que diversos estadísticos de resumen aparecen en columnas distintas.

Ejemplo. Una empresa con una cadena de tiendas registra la información de los empleados, incluyendo el salario, el cargo y la sección en la que trabaja cada uno. Se podría generar un informe que proporcione los estadísticos de salario resumidos (por ejemplo, media, mínimo y máximo) para cada sección.

Columnas de datos. Muestra una lista de las variables del informe para las que se desea obtener estadísticos de resumen y controla el formato de presentación y los estadísticos de resumen mostrados para cada variable.

Salto de columna. Muestra una lista de las variables de segmentación opcionales que dividen el informe en grupos y controla los formatos de presentación de los saltos de columna. Si hay varias variables de segmentación, se creará un grupo distinto para cada una de las categorías de las variables de segmentación dentro de las categorías de la variable de segmentación anterior en la lista. Las variables de segmentación deben ser variables categóricas discretas que dividan los casos en un número limitado de categorías con sentido.

Informe. Controla las características globales del informe, incluyendo la presentación de los valores perdidos, la numeración de las páginas y los títulos.

Vista previa. Muestra sólo la primera página del informe. Esta opción es útil para ver una vista previa del formato del informe sin tener que procesar el informe completo.

Los datos están ordenados. Para los informes con variables de segmentación, el archivo de datos se debe ordenar por los valores de estas variables antes de generar el informe. Si el archivo de datos ya está ordenado por estos valores, se puede ahorrar tiempo de procesamiento seleccionando esta opción. Esta opción es especialmente útil después de generar la vista previa de un informe.

Para obtener un informe de resumen: Estadísticos en columnas

1. Seleccione en los menús:

   Analizar > **Informes** > **Informe de estadísticos en columnas...**
2. Seleccione una o más variables para las columnas de datos. En el informe se genera una columna para cada variable seleccionada.
3. Para cambiar la medida de resumen para una variable, seleccione la variable de la lista de columnas de datos y pulse en **Resumen**.
4. Para obtener más de una medida de resumen para una variable, seleccione la variable en la lista de origen y desplácela hasta la lista Columnas de datos varias veces, una para cada medida que desee obtener.
5. Para mostrar una columna con la suma, la media, la razón o cualquier otra función de las columnas existentes, pulse en **Insertar total**. Al hacerlo se situará una variable llamada *total* en la lista Columnas de datos.
6. Para los informes ordenados y mostrados por subgrupos, seleccione una o más variables para Salto de columnas.

Función Columna de resumen total

Líneas de resumen controla el estadístico de resumen mostrado para la variable de las columnas de datos seleccionada.

Los estadísticos de resumen disponibles son: suma, media, valor mínimo, valor máximo, número de casos, porcentaje de casos por encima y por debajo de un valor especificado, porcentaje de casos dentro de un rango de valores especificado, desviación estándar, varianza, curtosis y asimetría.

Columna de resumen total

Columna de resumen controla los estadísticos de resumen del total que resumen dos o más columnas de datos.

Los estadísticos de resumen del total son la suma de columnas, la media de columnas, el mínimo, el máximo, la diferencia entre los valores de dos columnas, el cociente de los valores de una columna dividido por los valores de otra y el producto de los valores de las columnas multiplicados entre sí.

Suma de columnas. La columna *total* es la suma de las columnas de la lista Columna de resumen.

Media de columnas. La columna *total* es la media de las columnas de la lista Columna de resumen.

Mínimo de columnas. La columna *total* es el mínimo de las columnas de la lista Columna de resumen.

Máximo de columnas. La columna *total* es el máximo de las columnas de la lista Columna de resumen.

1ª columna - 2ª columna. La columna *total* es la resta de las columnas de la lista Columna de resumen. Esta lista debe contener, exactamente, dos columnas.

1ª columna / 2ª columna. La columna *total* es el cociente de las columnas de la lista Columna de resumen. Esta lista debe contener, exactamente, dos columnas.

% 1ª columna / 2ª columna. La columna *total* es el porcentaje de la primera columna dividido por la segunda columna de la lista Columna de resumen. Esta lista debe contener, exactamente, dos columnas.

Producto de columnas. La columna *total* es el producto de las columnas de la lista Columna de resumen.

Formato de columna de informe

Las opciones de formato de datos y salto de columna para Informe de estadísticos en columnas son iguales que las que se describen para Informes de estadísticos en filas.

Opciones de Salto de columna para los estadísticos en el informe

Opciones de saltos controla la presentación del subtotal, el espaciado y la paginación para las categorías de segmentación.

Subtotal. Controla los subtotales mostrados para cada categoría de segmentación.

Control de página. Controla el espaciado y la paginación para las categorías de la variable de segmentación seleccionada. Puede especificar un número de líneas en blanco entre las categorías de segmentación o empezar cada categoría de segmentación en una página nueva.

Líneas en blanco antes del subtotal. Controla el número de líneas en blanco entre los datos de las categorías de ruptura y los subtotales.

Opciones de columnas para los estadísticos en el informe

Opciones controla la presentación de los totales finales y de los valores perdidos y la paginación de los informes de estadísticos en columnas.

Total final. Muestra y etiqueta un total global para cada columna que aparece al final de la columna.

Valores perdidos. Permite excluir los valores perdidos del informe o seleccionar un único carácter para indicar estos valores.

Diseño de informe para Informe de estadísticos en columnas

Las opciones de diseño de informe para Informe de estadísticos en columnas son iguales que las que se describen para Informe de estadísticos en filas.

Características adicionales del comando REPORT

La sintaxis de comandos también le permite:

- Mostrar funciones de resumen diferentes en las columnas de una única línea de resumen.
- Insertar líneas de resumen en las columnas de datos para variables que no sean la variable de la columna de datos o para diversas combinaciones (funciones compuestas) de las funciones de resumen.
- Utilizar Mediana, Moda, Frecuencia y Porcentaje como funciones de resumen.
- Controlar de forma más precisa el formato de presentación de los estadísticos de resumen.
- Insertar líneas en blanco en diversos puntos de los informes.
- Insertar líneas en blanco después de cada *n*-ésimo caso de los informes en formato de listado.

Debido a la complejidad de la sintaxis de `REPORT`, a la hora de generar un nuevo informe con sintaxis puede resultarle útil, para aproximar el informe generado a partir de los cuadros de diálogo, copiar y pegar la sintaxis correspondiente y depurar esa sintaxis para generar exactamente el informe que le interese.

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 30. Análisis de fiabilidad

El análisis de fiabilidad permite estudiar las propiedades de las escalas de medición y de los elementos que componen las escalas. El procedimiento Análisis de fiabilidad calcula un número de medidas de fiabilidad de escala que se utilizan normalmente y también proporciona información sobre las relaciones entre elementos individuales de la escala. Se pueden utilizar los coeficientes de correlación intraclase para calcular estimaciones de la fiabilidad inter-evaluadores.

Ejemplo. ¿El cuestionario mide la satisfacción del cliente de manera útil? El análisis de fiabilidad le permitirá determinar el grado en que los elementos del cuestionario se relacionan entre sí, obtener un índice global de la replicabilidad o de la consistencia interna de la escala en su conjunto e identificar elementos problemáticos que deberían ser excluidos de la escala.

Estadísticos. Descriptivos para cada variable y para la escala, estadísticos de resumen comparando los elementos, correlaciones y covarianzas entre elementos, estimaciones de la fiabilidad, tabla de ANOVA, coeficientes de correlación intraclase, T^{cuadrado} de Hotelling y prueba de aditividad de Tukey.

Modelos. Están disponibles los siguientes modelos de fiabilidad:

- **Alfa (Cronbach).** Este modelo es un modelo de consistencia interna, que se basa en la correlación entre elementos promedio.
- **Dos mitades.** Este modelo divide la escala en dos partes y examina la correlación entre dichas partes.
- **Guttman.** Este modelo calcula los límites inferiores de Guttman para la fiabilidad verdadera.
- **Paralelo.** Este modelo asume que todos los elementos tienen varianzas iguales y varianzas error iguales a través de las réplicas.
- **Paralelo estricto.** Este modelo asume los supuestos del modelo paralelo y también asume que las medias son iguales a través de los elementos.

Análisis de fiabilidad: Consideraciones sobre los datos

Datos. Los datos pueden ser dicotómicos, ordinales o de intervalo, pero deben estar codificados numéricamente.

Supuestos. Las observaciones deben ser independientes y los errores no deben estar correlacionados entre los elementos. Cada par de elementos debe tener una distribución normal bivariada. Las escalas deben ser aditivas, de manera que cada elemento esté linealmente relacionado con la puntuación total.

Procedimientos relacionados. Si desea explorar la dimensionalidad de los elementos de la escala (para comprobar si es necesario más de un constructo para explicar el patrón de puntuaciones en los elementos), utilice el Análisis factorial o el Escalamiento multidimensional. Para identificar grupos homogéneos de variables, use el análisis de clústeres jerárquico para agrupar las variables en clústeres.

Para obtener un análisis de fiabilidad

1. Seleccione en los menús:

   Analizar > **Escala** > **Análisis de fiabilidad...**
2. Seleccione dos o más variables como componentes potenciales de una escala aditiva.
3. Elija un modelo de la lista desplegable Modelo.

Análisis de fiabilidad: Estadísticos

Puede seleccionar diversos estadísticos que describen la escala y sus elementos. Los estadísticos de los que se informa de forma predeterminada incluyen el número de casos, el número de elementos y las estimaciones de la fiabilidad, según se explica a continuación:

- **Modelos Alfa.** Coeficiente Alfa; para datos dicotómicos, éste es equivalente al coeficiente 20 de Kuder-Richardson (KR20).
- **Modelos Dos mitades.** Correlación entre formas, fiabilidad de dos mitades de Guttman, fiabilidad de Spearman-Brown (longitud igual y desigual) y coeficiente alfa para cada mitad.
- **Modelos de Guttman.** Coeficientes de fiabilidad lambda 1 a lambda 6.
- **Modelos Paralelo y Estrictamente paralelo.** Prueba de bondad de ajuste del modelo; estimaciones de la varianza error, varianza común y varianza verdadera; correlación común entre elementos estimada; fiabilidad estimada y estimación de la fiabilidad no sesgada.

Descriptivos para. Genera estadísticos descriptivos para las escalas o los elementos a través de los casos.

- **Elemento.** Genera estadísticos descriptivos para los elementos a través de los casos.
- **Escalas.** Genera estadísticos descriptivos para las escalas.
- **Escala si se elimina el elemento.** Muestra estadísticos de resumen para comparar cada elemento con la escala compuesta por otros elementos. Los estadísticos incluyen la media de escala y la varianza si el elemento fuera a eliminarse de la escala, la correlación entre el elemento y la escala compuesta por otros elementos, y alfa de Cronbach si el elemento fuera a eliminarse de la escala.

Resúmenes. Proporciona estadísticos descriptivos sobre las distribuciones de los elementos a través de todos los elementos de la escala.

- *Medias*. Estadísticos de resumen de las medias de los elementos. Se muestran el máximo, el mínimo y el promedio de las medias de los elementos, el rango y la varianza de las medias de los elementos, y la razón de la mayor media sobre la menor media de los elementos.
- *Varianzas*. Estadísticos de resumen de las varianzas de los elementos. Se muestran el máximo, el mínimo y el promedio de las varianzas de los elementos, el rango y la varianza de las varianzas de los elementos y la razón de la mayor varianza sobre la menor varianza de los elementos.
- *Covarianzas*. Estadísticos de resumen de las covarianzas entre elementos. Se muestran el máximo, el mínimo y el promedio de las covarianzas entre elementos, el rango y la varianza de las covarianzas entre elementos, y la razón de la mayor sobre la menor covarianza entre elementos.
- *Correlaciones*. Estadísticos de resumen para las correlaciones entre elementos. Se muestran el máximo, el mínimo y el promedio de las correlaciones entre elementos, el rango y la varianza de las correlaciones entre elementos, y la razón de la mayor correlación sobre la menor correlación entre elementos.

entre elementos. Genera las matrices de correlaciones o covarianzas entre los elementos.

Tabla de ANOVA. Produce pruebas de medias iguales.

- *Prueba F*. Muestra la tabla de un análisis de varianza de medidas repetidas.
- *Chi-cuadrado de Friedman*. Muestra el chi-cuadrado de Friedman y el coeficiente de concordancia de Kendall. Esta opción es adecuada para datos que se encuentren en el formato de rangos. La prueba de chi-cuadrado sustituye a la prueba F habitual en la tabla de ANOVA.
- *Chi-cuadrado de Cochran*. Muestra la Q de Cochran. Esta opción es adecuada para datos dicotómicos El estadístico Q sustituye a la F habitual en la tabla de ANOVA.

T-cuadrado de Hotelling. Genera una prueba multivariante sobre la hipótesis nula de que todos los elementos de la escala tienen la misma media.

Prueba de aditividad de Tukey. Genera un contraste sobre el supuesto de que no existe una interacción multiplicativa entre los elementos.

Coeficiente de correlación intraclase. Genera medidas sobre la consistencia o sobre el acuerdo de los valores entre los propios casos.

- **Modelo.** Seleccione el modelo para calcular el coeficiente de correlación intraclase. Los modelos disponibles son: Mixto de dos factores, aleatorio de dos factores y aleatorio de un factor. Seleccione **Mixto de dos factores**, si los efectos de personas son aleatorios y los efectos de elementos son fijos, **Aleatorio de dos factores** si los efectos de personas y los efectos de elementos son aleatorios; o **Aleatorio de un factor** si los efectos de personas son aleatorios.
- **Tipo.** Seleccione el tipo de índice. Los tipos disponibles son: Consistencia y Acuerdo absoluto.
- **Intervalo de confianza.** Especifica el nivel para el intervalo de confianza. El valor predeterminado es 95%.
- **Valor de prueba.** Especifica el valor hipotetizado para el coeficiente, para el contraste de hipótesis. Este valor es el valor con el que se compara el valor observado. El valor predeterminado es 0.

Características adicionales del comando RELIABILITY

La sintaxis de comandos también le permite:

- Leer y analizar una matriz de correlaciones.
- Escribir una matriz de correlaciones para su análisis posterior.
- Especificar divisiones que no sean mitades iguales para el método de dos mitades.

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 31. Escalamiento multidimensional

El escalamiento multidimensional trata de encontrar la estructura de un conjunto de medidas de distancia entre objetos o casos. Esta tarea se logra asignando las observaciones a posiciones específicas en un espacio conceptual (normalmente de dos o tres dimensiones) de modo que las distancias entre los puntos en el espacio concuerden al máximo con las disimilaridades dadas. En muchos casos, las dimensiones de este espacio conceptual son interpretables y se pueden utilizar para comprender mejor los datos.

Si las variables se han medido objetivamente, puede utilizar el escalamiento multidimensional como técnica de reducción de datos (el procedimiento Escalamiento multidimensional permitirá calcular las distancias a partir de los datos multivariados, si es necesario). El escalamiento multidimensional puede también aplicarse a valoraciones subjetivas de disimilaridad entre objetos o conceptos. Además, el procedimiento Escalamiento multidimensional puede tratar datos de disimilaridad procedentes de múltiples fuentes, como podrían ser múltiples evaluadores o múltiples encuestados.

Ejemplo. ¿Cómo percibe el público las diferencias entre distintos coches? Si posee datos de las valoraciones de similaridad emitidas por los encuestados sobre las diferentes marcas y modelos de coches, puede utilizar el escalamiento multidimensional para identificar las dimensiones que describan las preferencias de los consumidores. Puede encontrar, por ejemplo, que el precio y el tamaño de un vehículo definen un espacio de dos dimensiones, capaz de explicar las similaridades de las que informan los encuestados.

Estadísticos. Para cada modelo: Matriz de datos, Matriz de datos escalada óptimamente, S-stress (de Young), estrés (de Kruskal), R^2, Coordenadas de los estímulos, estrés promedio y R^2 para cada estímulo (modelos RMDS). Para modelos de diferencias individuales (INDSCAL): ponderaciones del sujeto e índice de peculiaridad para cada sujeto. Para cada matriz en los modelos de escalamiento multidimensional replicado: estrés y R^2 para cada estímulo. Gráficos: coordenadas de los estímulos (de dos o tres dimensiones), diagrama de dispersión de las disparidades frente a las distancias.

Escalamiento multidimensional: Consideraciones sobre los datos

Datos. Si los datos son de disimilaridad, todas las disimilaridades deben ser cuantitativas y deben estar medidas en la misma métrica. Si los datos son datos multivariantes, las variables pueden ser datos cuantitativos, binarios o de recuento. El escalamiento de las variables es un tema importante, ya que las diferencias en el escalamiento pueden afectar a la solución. Si las variables tienen grandes diferencias en el escalamiento (por ejemplo, una variable se mide en dólares y otra en años), debe considerar la posibilidad de tipificarlas (este proceso puede llevarse a cabo automáticamente con el propio procedimiento Escalamiento multidimensional).

Supuestos. El procedimiento Escalamiento multidimensional está relativamente libre de supuestos distribucionales. Compruebe que selecciona el nivel de medición adecuado (ordinal, de intervalo, o de razón) en el cuadro de diálogo Escalamiento multidimensional: Opciones para asegurar que los resultados se calculan correctamente.

Procedimientos relacionados. Si su objetivo es la reducción de los datos, un método alternativo a tener en cuenta es el análisis factorial, sobre todo si las variables son cuantitativas. Si desea identificar grupos de casos similares, considere complementar el análisis de escalamiento multidimensional con un análisis de clústeres jerárquico o de *k*-medias.

Para obtener un análisis de escalamiento multidimensional

1. Seleccione en los menús:

   Analizar > **Escala** > **Escalamiento multidimensional...**

2. Seleccione al menos cuatro variables para su análisis.
3. En el grupo Distancias, seleccione **Los datos son distancias** o **Crear distancias a partir de datos**.
4. Si selecciona **Crear distancias a partir de datos**, también podrá seleccionar una variable de agrupación para matrices individuales. La variable de agrupación puede ser numérica o de cadena.

Si lo desea, tiene la posibilidad de:

- Especifique la forma de la matriz de distancias cuando los datos sean distancias.
- Especifique la medida de distancia que hay que utilizar al crear distancias a partir de datos.

Escalamiento multidimensional: Forma de los datos

Si el conjunto de datos activo representa distancias entre uno o dos conjuntos de objetos, especifique la forma de la matriz de datos para obtener los resultados correctos.

Nota: no puede seleccionar **Cuadrada simétrica** si el cuadro de diálogo Modelo especifica la condicionalidad de filas.

Escalamiento multidimensional: Crear la medida a partir de los datos

El escalamiento multidimensional utiliza datos de disimilaridad para crear una solución de escalamiento. Si los datos son datos multivariantes (los valores de las variables que se han medido), debe crear los datos de disimilaridad para poder calcular una solución de escalamiento multidimensional. Puede especificar los detalles para la creación de las medidas de disimilaridad a partir de los datos.

Medida. Le permite especificar la medida de disimilaridad para el análisis. Seleccione una opción del grupo Medida que se corresponda con el tipo de datos y, a continuación, elija una de las medidas de la lista desplegable correspondiente a ese tipo de medida. Las opciones disponibles son:

- **Intervalo**. Distancia euclídea, Distancia euclídea al cuadrado, Chebychev, Bloque, Minkowski o Personalizada.
- **Recuentos**. Medida de chi-cuadrado o Medida de phi-cuadrado.
- **Binario**. Distancia euclídea, Distancia euclídea al cuadrado, Diferencia de tamaño, Diferencia de configuración, Varianza o Lance y Williams.

Crear matriz de proximidades. Le permite elegir la unidad de análisis. Las opciones son Entre variables o Entre casos.

Transformar valores. En determinados casos, como cuando las variables se miden en escalas muy distintas, puede que desee tipificar los valores antes de calcular las proximidades (no es aplicable a datos binarios). Seleccione un método de estandarización en la lista desplegable Estandarizar. Si no se requiere ninguna estandarización, seleccione **Ninguno**.

Escalamiento multidimensional: Modelo

La estimación correcta de un modelo de escalamiento multidimensional depende de aspectos que atañen a los datos y al modelo en sí.

Nivel de medición. Permite especificar el nivel de los datos. Las opciones son Ordinal, Intervalo y Razón. Si las variables son ordinales, al seleccionar **Desempatar observaciones empatadas** se solicitará que sean consideradas como variables continuas, de forma que los empates (valores iguales para casos diferentes) se resuelvan óptimamente.

Condicionalidad. Permite especificar qué comparaciones tienen sentido. Las opciones son Matriz, Fila o Incondicional.

Dimensiones. Permite especificar la dimensionalidad de la solución o soluciones del escalamiento. Se calcula una solución para cada número del rango especificado. Especifique números enteros entre 1 y 6; se permite un mínimo de 1 sólo si selecciona **Distancia euclídea** como modelo de escalamiento. Para una solución única, especifique el mismo número para el mínimo y el máximo.

Modelo de escalamiento. Permite especificar los supuestos bajo los que se realiza el escalamiento. Las opciones disponibles son Distancia euclídea o Distancia euclídea de diferencias individuales (también conocida como INDSCAL). Para el modelo de Distancia euclídea de diferencias individuales, puede seleccionar **Permitir ponderaciones negativas de los sujetos**, si es adecuado para los datos.

Escalamiento multidimensional: Opciones

Puede especificar opciones para el análisis de escalamiento multidimensional.

Representación. Permite seleccionar varios tipos de resultados. Las opciones disponibles son Gráficos de grupo, Gráficos para los sujetos individuales, Matriz de datos y Resumen del modelo y de las opciones.

Criterios. Permite determinar cuándo debe detenerse la iteración. Para cambiar los valores predeterminados, introduzca valores para la **Convergencia de s-stress**, el **Valor mínimo de s-stress** y el **Nº máximo de iteraciones**.

Tratar distancias menores que n como perdidas. Las distancias menores que este valor se excluyen del análisis.

Características adicionales del comando ALSCAL

La sintaxis de comandos también le permite:

- Utilizar tres tipos de modelos adicionales, conocidos como ASCAL, AINDS y GEMSCAL en la literatura sobre escalamiento multidimensional.
- Realizar transformaciones polinómicas en los datos de intervalo y proporción.
- Analizar similaridades (en lugar de distancias) con datos ordinales.
- Analizar datos nominales.
- Guardar varias matrices de coordenadas y ponderación en archivos y volverlas a leer para análisis.
- Restringir el desplegamiento multidimensional.

Consulte la *Referencia de sintaxis de comandos* para obtener información completa de la sintaxis.

Capítulo 32. Estadísticos de la razón

El procedimiento Estadísticos de la razón proporciona una amplia lista de estadísticos de resumen para describir la razón entre dos variables de escala.

Se pueden ordenar los resultados por los valores de una variable de agrupación, en orden ascendente o descendente. Se puede eliminar de los resultados el informe de los estadísticos de la razón y almacenar los resultados en un archivo externo.

Ejemplo. ¿Existe una buena uniformidad en la razón entre el precio de tasación y el precio de venta de viviendas en cada una de las cinco regiones? En los resultados, se puede descubrir que la distribución de las razones varía considerablemente entre regiones.

Estadísticos. Mediana, media, media ponderada, intervalos de confianza, coeficiente de dispersión (CDD), coeficiente de variación centrado en la mediana, coeficiente de variación centrado en la media, el diferencial de precio (DRV), desviación estándar, desviación absoluta promedio (DAP), rango, valores mínimos y máximos y el índice de concentración calculado dentro de un rango o porcentaje (especificados por el usuario) respecto a la razón mediana.

Estadísticos de la razón: Consideraciones sobre los datos

Datos. Utilice códigos numéricos o cadenas para codificar las variables de agrupación (mediciones de nivel nominal u ordinal).

Supuestos. Las variables que definen el numerador y el denominador de la razón deben ser variables de escala, que toman valores positivos.

Para obtener estadísticos de la razón

1. Seleccione en los menús:

   Analizar > **Estadísticos descriptivos** > **Razón...**
2. Seleccione una variable de numerador.
3. Seleccione una variable de denominador.

Si lo desea:

- Seleccione una variable de agrupación y especificar el orden de los grupos en los resultados.
- Elija si desea mostrar los resultados en el Visor.
- Elija si desea guardar los resultados en un archivo externo para un uso posterior y especificar el nombre del archivo en el que se van a guardar los resultados.

Estadísticos de la razón

Tendencia central. Las medidas de tendencia central son estadísticos que describen la distribución de las razones.

- **Mediana.** Un valor tal que el número de razones menores que este valor es igual al número de razones mayores que el mismo.
- **Media.** El resultado de sumar las razones y dividir la suma entre el número total de razones.
- **Media ponderada.** El resultado de dividir la media del numerador entre la media del denominador. La media ponderada es también la media de las razones ponderadas por el denominador.

- **Intervalos de confianza.** Muestra los intervalos de confianza para la media, la mediana y la media ponderada (si se solicita). Especifique un valor mayor o igual que 0 y menor que 100 como nivel de confianza.

Dispersión. Estos estadísticos miden la cantidad de variación o de dispersión entre los valores observados.

- **DAP.** La desviación absoluta promedio es el resultado de sumar las desviaciones absolutas de las razones respecto a la mediana y dividir el resultado entre el número total de razones.
- **CDD.** El coeficiente de dispersión es el resultado de expresar la desviación absoluta promedio como un porcentaje de la mediana.
- **DRP.** El diferencial relativo al precio, también conocido como el índice de regresibilidad, es el resultado de dividir la media por la media ponderada.
- **CDV centrado en la mediana.** El coeficiente de variación centrado en la mediana es el resultado de expresar la raíz de la media cuadrática de las desviaciones respecto a la mediana como un porcentaje de la mediana.
- **CDV centrado en la media.** El coeficiente de variación centrado en la media es el resultado de expresar la desviación estándar como un porcentaje de la media.
- **Desviación estándar.** La desviación estándar es el resultado de sumar las desviaciones cuadráticas de las razones respecto a la media, dividir la suma por el número total de razones menos uno y extraer la raíz cuadrada positiva.
- **Rango.** El rango es el resultado de restar la razón mínima de la razón máxima.
- **Mínimo.** El mínimo es la razón menor.
- **Máximo.** El máximo es la razón mayor.

Índice de concentración. El coeficiente de concentración mide el porcentaje de razones que están dentro de un intervalo. Se puede calcular de dos maneras:

- **Razones dentro del.** En este caso, el intervalo se define de forma explícita especificando los valores superior e inferior del intervalo. Introduzca valores para las proporciones superior e inferior y pulse en **Añadir** para obtener un intervalo.
- **Razones en.** En este caso, el intervalo se define de forma implícita al especificar el porcentaje de la mediana. Introduzca un valor entre 0 y 100 y pulse en **Añadir**. El límite inferior del intervalo será igual a $(1 - 0.01 \times \text{valor}) \times \text{mediana}$, y el límite superior será igual a $(1 + 0.01 \times \text{valor}) \times \text{mediana}$.

Capítulo 33. Curvas COR

Este procedimiento es un método útil para evaluar la realización de esquemas de clasificación en los que exista una variable con dos categorías por las que se clasifiquen los sujetos.

Ejemplo. Un banco tiene interés en clasificar a sus clientes dependiendo de si se retrasarán o no en el pago de sus préstamos; por tanto, se desarrollan métodos especiales para tomar estas decisiones. Las curvas COR se pueden utilizar para evaluar el grado de acierto de estos métodos.

Estadísticos. Es un área situada bajo la curva COR con un intervalo de confianza y puntos de coordenadas de la curva COR. Gráficos: Curva COR.

Métodos. Se puede calcular la estimación del área situado bajo la curva COR de forma paramétrica o no paramétrica mediante un modelo exponencial binegativo.

Curvas COR: Consideraciones sobre los datos

Datos. Las variables de contraste son cuantitativas. Las variables de contraste suelen estar constituidas por probabilidades, resultantes de un análisis discriminante o de una regresión logística, o bien compuestas por puntuaciones atribuidas en una escala arbitraria que indican el «grado de convicción» que tiene un evaluador de que el sujeto pueda pertenecer a una u otra categoría. La variable de estado puede ser de cualquier tipo e indicar la categoría real a la que pertenece un sujeto. El valor de la variable de estado indica la categoría que se debe considerar *positiva*.

Supuestos. Se considera que los números ascendentes de la escala del evaluador representan la creciente convicción de que el sujeto pertenece a una categoría. Por el contrario, los números descendentes representan la creciente convicción de que el sujeto pertenece a la otra categoría. El usuario deberá elegir qué dirección es *positiva*. También se considera que se conoce la categoría *real* a la que pertenece el sujeto.

Para obtener una curva COR

1. Seleccione en los menús:
   Analizar > **Curva COR...**
2. Seleccione una o más variables de probabilidad de contraste.
3. Elija una variable de estado.
4. Identifique el valor *positivo* para la variable de estado.

Curvas COR: Opciones

Puede seleccionar las opciones siguientes para su análisis:

Clasificación. Permite especificar si se debe incluir o excluir el valor del punto de corte al realizar una clasificación *positiva*. Este ajuste no afecta a los resultados.

Dirección de la prueba. Permite especificar la dirección de la escala según la categoría *positiva*.

Parámetros para el error estándar del área. Permite especificar el método de estimación del error estándar del área situada bajo la curva. Los métodos disponibles son el no paramétrico y el exponencial binegativo. También se puede establecer el nivel para el intervalo de confianza. El rango disponible es entre el 50,1% y el 99,9%.

Valores perdidos. Permite especificar el tratamiento que reciben los valores perdidos.

Capítulo 34. Simulación

Los modelos predictivos, como una regresión lineal, requieren un conjunto de entradas conocidas para predecir un resultado o valor de destino. En muchas aplicaciones del mundo real, sin embargo, los valores de las entradas son inciertos. La simulación permite explicar la incertidumbre de las entradas en modelos predictivos y evaluar la posibilidad de varios resultados del modelo en presencia de esa incertidumbre. Por ejemplo, tiene un modelo de beneficio que incluye el coste de los materiales como una entrada, pero hay incertidumbre en ese coste por la volatilidad del mercado. Puede utilizar la simulación para modelar esa incertidumbre y determinar el efecto que tiene en los beneficios.

La simulación de IBM SPSS Statistics utiliza el método de Monte Carlo. Las entradas inciertas se modelan con distribuciones de probabilidad (como la distribución triangular), y los valores simulados de esas entradas se generan a partir de esas distribuciones. Las entradas cuyos valores se conocen se mantienen fijas en los valores conocidos. El modelo predictivo se evalúa utilizando un valor simulado para cada entrada incierta y los valores fijos de las entradas conocidas para calcular el destino (u destinos) del modelo. El proceso se repite muchas veces (normalmente decenas de miles o cientos de miles de veces), resultando en una distribución de los valores de destino que es posible utilizar para responder las preguntas de una naturaleza probabilística. En el contexto de IBM SPSS Statistics, cada repetición del proceso genera un caso diferente (registro) de datos que consiste en el conjunto de valores simulados de las entradas inciertas, los valores de las entradas fijas y el destino (o destinos) predichos del modelo.

También puede simular datos en ausencia de un modelo predictivo especificando distribuciones de probabilidad para las variables que se van a simular. Cada caso de datos generado consta del conjunto de valores simulados para las variables especificadas.

Para ejecutar una simulación, necesita especificar datos como el modelo predictivo, las distribuciones de probabilidad de las entradas inciertas, las correlaciones entre esas entradas y los valores de entradas fijas. Una vez haya especificado todos los detalles de una simulación, puede ejecutarla y, opcionalmente, guardar las especificaciones en un archivo de **plan de simulación**. Puede compartir el plan de simulación con otros usuarios, que pueden ejecutar la simulación sin necesidad de comprender los detalles de su creación.

Existen dos interfaces disponibles para trabajar con simulaciones. El Generador de simulaciones es una interfaz avanzada para usuarios que diseñan y ejecutan simulaciones. Proporciona el conjunto completo de funciones para diseñar una simulación, guardar las especificaciones en un archivo de plan de simulación, especificar los resultados y ejecutar la simulación. Puede crear una simulación basada en un archivo de modelo de IBM SPSS, o en un conjunto de ecuaciones personalizadas que defina en el Generador de simulaciones. También puede cargar un plan de simulación existente en el Generador de simulaciones, modificar cualquiera de los ajustes y ejecutar la simulación, con la opción adicional de guardar el plan actualizado. Para los usuarios que tengan un plan de simulación y desean preferentemente ejecutar la simulación, existe una interfaz más simple. Permite modificar ajustes que permiten ejecutar la simulación en condiciones diferentes, pero no proporciona todas las funciones del Generador de simulaciones para el diseño de simulaciones.

Para diseñar una simulación basada en un archivo de modelo

1. Seleccione en los menús:

   Analizar > **Simulación...**
2. Pulse en **Seleccionar Archivo modelo SPSS** y en **Continuar**.
3. Abra el archivo de modelo.

Un modelo de archivo es un archivo XML que contiene el PMML de modelo creado desde IBM SPSS Statistics o IBM SPSS Modeler. Consulte el tema "Pestaña Modelo" en la página 182 para obtener más información.

4. En la pestaña Simulación (del Generador de simulaciones), especifique las distribuciones de probabilidad de las entradas simuladas y los valores de entradas fijas. Si el conjunto de datos activo contiene datos históricos de entradas simuladas, pulse en **Ajustar todas** para determinar automáticamente la distribución que se ajusta mejor para cada entrada, y para determinar las correlaciones entre ellas. Para cada entrada simulada que no se ajuste a los datos históricos, debe especificar explícitamente una distribución seleccionando un tipo de distribución e introduciendo los parámetros necesarios.
5. Pulse en **Ejecutar** para ejecutar la simulación. De forma predeterminada, el plan de simulación, que especifica los detalles de la simulación, se guarda en la ubicación especificada en los ajustes de almacenamiento.

Se encuentran disponibles las siguientes opciones:

- Modificar la ubicación del plan de simulación guardado.
- Especificar correlaciones conocidas entre entradas simuladas.
- Calcular automáticamente una tabla de contingencia de asociaciones entre entradas categóricas y utilizar dichas asociaciones cuando los datos se generan para esas entradas.
- Especificar los análisis de sensibilidad para investigar el efecto de variar el valor de una entrada fija o variar un parámetro de distribución de una entrada simulada.
- Especificar opciones avanzadas como, por ejemplo, establecer el número máximo de casos para generar o solicitar muestreos de cola.
- Personalizar los resultados.
- Guardar los datos simulados en un archivo de datos.

Para diseñar una simulación basada en ecuaciones personalizadas

1. Seleccione en los menús:

   Analizar > **Simulación...**
2. Pulse en **Escribir las ecuaciones** y en **Continuar**.
3. Pulse en **Nueva ecuación** en la pestaña Modelo (del Generador de simulaciones) para definir cada ecuación del modelo predictivo.
4. Pulse la pestaña Simulación y especifique las distribuciones de probabilidad de entradas simuladas y valores de entradas fijas. Si el conjunto de datos activo contiene datos históricos de entradas simuladas, pulse en **Ajustar todas** para determinar automáticamente la distribución que se ajusta mejor para cada entrada, y para determinar las correlaciones entre ellas. Para cada entrada simulada que no se ajuste a los datos históricos, debe especificar explícitamente una distribución seleccionando un tipo de distribución e introduciendo los parámetros necesarios.
5. Pulse **Ejecutar** para ejecutar la simulación. De forma predeterminada, el plan de simulación, que especifica los detalles de la simulación, se guarda en la ubicación especificada en los ajustes de almacenamiento.

Se encuentran disponibles las siguientes opciones:

- Modificar la ubicación del plan de simulación guardado.
- Especificar correlaciones conocidas entre entradas simuladas.
- Calcular automáticamente una tabla de contingencia de asociaciones entre entradas categóricas y utilizar dichas asociaciones cuando los datos se generan para esas entradas.
- Especificar los análisis de sensibilidad para investigar el efecto de variar el valor de una entrada fija o variar un parámetro de distribución de una entrada simulada.

- Especificar opciones avanzadas como, por ejemplo, establecer el número máximo de casos para generar o solicitar muestreos de cola.
- Personalizar los resultados.
- Guardar los datos simulados en un archivo de datos.

Diseñar una simulación sin un modelo predictivo

1. Elija en los menús:

   Analizar > **Simulación...**
2. Pulse **Crear datos simulados** y pulse **Continuar**.
3. En la pestaña Modelo (en el Generador de simulaciones), seleccione los campos que desea simular. Puede seleccionar los campos del conjunto de datos activo o puede definir campos nuevos pulsando **Nuevo**.
4. Pulse en la pestaña Simulación y especifique las distribuciones de probabilidad para los campos que se van a simular. Si el conjunto de datos activo contiene datos históricos para cualquiera de estos campos, pulse **Ajustar todo** para determinar la distribución que más se ajusta a todos los datos y determinar las correlaciones entre los campos. Para los campos que no se ajustan a los datos históricos, debe especificar de forma explícita una distribución seleccionando un tipo de distribución y especificando los parámetros necesarios.
5. Pulse en **Ejecutar** para ejecutar la simulación. De forma predeterminada, los datos simulados se guardan en el conjunto de datos nuevo especificado en la Configuración de guardar. Además, el plan de simulación, que especifica los detalles de la simulación, se guarda en la ubicación especificada en la Configuración de guardar.

Se encuentran disponibles las siguientes opciones:

- Modificar la ubicación para los datos simulados o el plan de simulación guardado.
- Especificar correlaciones conocidas entre campos simulados.
- Calcular automáticamente una tabla de contingencia de asociaciones entre campos categóricos y utilizar dichas asociaciones cuando se generan los datos para dichos campos.
- Especificar los análisis de sensibilidad para investigar el efecto de variar un parámetro de distribución para un campo simulado.
- Especificar opciones avanzadas como, por ejemplo, establecer el número de casos para generar.

Para ejecutar una simulación de un plan de simulación

Existen dos opciones disponibles para ejecutar una simulación a partir de un plan de simulación. Puede utilizar el cuadro de diálogo Ejecutar simulación, que está diseñado para ejecutarse a partir de un plan de simulación, o bien puede utilizar el Generador de simulaciones.

Para utilizar el cuadro de diálogo Ejecutar simulación:

1. Seleccione en los menús:

   Analizar > **Simulación...**
2. Pulse en **Abrir un plan de simulación existente**.
3. Asegúrese de que la casilla de verificación **Abrir en el Generador de simulaciones** no está seleccionada y pulse en **Continuar**.
4. Abra el plan de simulación.
5. Pulse en **Ejecutar** en el cuadro de diálogo Ejecutar simulación.

Para ejecutar la simulación del Generador de simulaciones:

1. Seleccione en los menús:

   Analizar > **Simulación...**

2. Pulse en **Abrir un plan de simulación existente**.
3. Seleccione la casilla de verificación **Abrir en el Generador de simulaciones** y pulse en **Continuar**.
4. Abra el plan de simulación.
5. Modifique los ajustes que desee en la pestaña Simulación.
6. Pulse en **Ejecutar** para ejecutar la simulación.

También puede intentar lo siguiente:

- Configurar o modificar los análisis de sensibilidad para investigar el efecto de variar el valor de una entrada fija o variar un parámetro de distribución de una entrada simulada.
- Reajustar las distribuciones y correlaciones de entradas simuladas en nuevos datos.
- Cambiar la distribución de una entrada simulada.
- Personalizar los resultados.
- Guardar los datos simulados en un archivo de datos.

Generador de simulaciones

El Generador de simulaciones proporciona la gama completa de funciones para diseñar y ejecutar simulaciones. Permite ejecutar las siguientes tareas generales:

- Diseñar y ejecutar una simulación de un modelo de IBM SPSS definido en un archivo de modelo PMML.
- Diseñar y ejecutar una simulación de un modelo predictivo definido por un conjunto de ecuaciones personalizadas que especifique.
- Diseñar y ejecutar una simulación que genera datos en ausencia de un modelo predictivo.
- Ejecutar una simulación basada en un plan de simulación existente, pudiendo modificar los ajustes del plan.

Pestaña Modelo

Para simulaciones basadas en un modelo predictivo, la pestaña Modelo especifica el origen del modelo. Para simulaciones que no incluyen un modelo predictivo, la pestaña Modelo especifica los campos que se van a simular.

Seleccione un archivo de modelo SPSS. Esta opción especifica que el modelo predictivo se define en un archivo de modelo IBM SPSS. Un archivo de modelo de IBM SPSS es un archivo XML o un archivador de archivos comprimidos (archivo .zip) que contiene el PMML de modelo creado desde IBM SPSS Statistics o IBM SPSS Modeler. Los modelos predictivos están creados por procedimientos, como Regresión lineal y Árboles de decisión en IBM SPSS Statistics, y se puede exportar a un archivo de modelo. Puede utilizar un archivo de modelo diferente pulsando en **Examinar** y desplazándose al archivo que desee.

Modelos PMML admitidos por Simulación

- Regresión lineal
- Modelo lineal automático
- Modelo lineal generalizado
- Modelo mixtos lineales generalizados
- Modelo lineal general
- Regresión logística binaria
- Regresión logística multinomial
- Regresión multinomial ordinal
- Regresión de Cox
- Árbol

- Árbol potenciado (C5)
- Discriminante
- Clúster de dos fases
- Clúster de K-medias
- Red neuronal
- Conjunto de reglas (Lista de decisiones)

Nota:
- No se admite en Simulación el uso de modelos PMML que tengan múltiples campos de destino (variables) o divisiones.
- Los valores de entradas de cadena en modelos de regresión logística binaria están limitados a bytes en el modelo. Si está adaptando tales entradas de cadena al conjunto de datos activo, asegúrese de que los valores de los datos no superan 8 bytes de longitud. Los valores de datos que superan 8 bytes se excluyen de la distribución categórica asociada de la entrada, y aparecen sin coincidencias en la tabla de salida Categorías sin coincidencias.

Escriba las ecuaciones para el modelo. Esta opción especifica que el modelo predictivo se compone de una o más ecuaciones personalizadas que puede crear el usuario. Cree las ecuaciones pulsando en **Nueva ecuación**. Se abrirá el Editor de ecuaciones. Puede modificar ecuaciones existentes, copiarlas para utilizarlas como plantillas de nuevas ecuaciones, reordenarlas y eliminarlas.
- El Generador de simulaciones no admite sistemas de ecuaciones simultáneas o ecuaciones que no son lineales en la variable destino.
- Las ecuaciones personalizadas se evalúan en el orden en que se especifican. Si la ecuación de un destino especificado depende de otro destino, el otro destino se debe definir mediante una ecuación anterior.

  Por ejemplo, teniendo en cuenta el conjunto de ecuaciones siguiente, la ecuación de *beneficios* depende de los valores de *ingresos* y *gastos*, por lo que las ecuaciones de *ingresos* y *gastos* deben anteceder a la ecuación de *beneficios*.

```
ingresos= precio*volumen
gastos= fijos + volumen*(unidad_coste_materiales + unidad_costes_laborales)
beneficios = ingresos - gastos
```

Crear datos sumulados sin un modelo. Seleccione esta opción para simular datos sin un modelo predictivo. Especifique los campos que deben simularse seleccionando los campos del conjunto de datos activo o pulsando **Nuevo** para definir campos nuevos.

Editor de ecuaciones

El Editor de ecuaciones permite crear o modificar una ecuación personalizada de su modelo predictivo.
- La expresión de la ecuación puede contener campos del conjunto de datos activo o nuevos campos de entrada que define en el Editor de ecuaciones.
- Puede especificar propiedades del destino como su nivel de medición, etiquetas de valores y si los resultados están generados para el destino.
- Puede utilizar los destinos de ecuaciones definidas anteriormente como entradas de la ecuación actual, permitiéndole crear ecuaciones acopladas.
- Puede adjuntar un comentario descriptivo a la ecuación. Los comentarios se muestran junto con la ecuación de la pestaña Modelo.

1. Introduzca el nombre del destino. Opcionalmente, pulse en **Editar** en el cuadro de texto Destino para abrir el cuadro de diálogo Entradas definidas, permitiendo cambiar las propiedades predeterminadas del destino.
2. Para crear una expresión, puede pegar los componentes en el campo Expresión numérica o escribir directamente en dicho campo.

- Puede crear su expresión utilizando campos del conjunto de datos activo o puede definir nuevas entradas pulsando en el botón **Nuevo**. Se abrirá el cuadro de diálogo Definir entradas.
- Puede pegar las funciones seleccionando un grupo de la lista Grupo de funciones y pulsando dos veces en la función de la lista de funciones (o seleccione la función y pulse en la flecha que se encuentra sobre la lista Grupo de funciones). Especifique los parámetros indicados por los signos de interrogación. El grupo de funciones con la etiqueta **Todo** contiene una lista de todas las funciones variables. En un área reservada del cuadro de diálogo se muestra una breve descripción de la función actualmente seleccionada.
- Las constantes de cadena deben ir entre comillas.
- Si los valores contienen decimales, debe utilizarse una coma(,) como indicador decimal.

Nota: la simulación no admite ecuaciones personalizadas con destinos de cadena.

Entradas definidas: El cuadro de diálogo Entradas definidas permite definir nuevas entradas y las propiedades de los destinos.

- Si una entrada que se va a utilizar en una ecuación no existe en el conjunto de datos activo, debe definirlo para que se pueda utilizar en la ecuación.
- Si simula datos con un modelo predictivo, debe definir todas las entradas simuladas que no existan en el conjunto de datos activo.

Nombre. Especifique el nombre del destino o la entrada.

Destino. Puede especificar el nivel de medición de un destino. El nivel de medición predeterminado es continuo. También puede especificar si se creará el resultado para este destino. Por ejemplo, en el caso de un grupo de ecuaciones acopladas es posible que solo le interese el resultado del destino de la ecuación final, por lo que suprimiría el resultado del resto de destinos.

Entrada que se simulará. Especifica que los valores de la entrada se simularán de acuerdo con una distribución de probabilidad especificada (la distribución de probabilidad se especifica en la pestaña Simulación). El nivel de medición determina el conjunto predeterminado de distribuciones que se consideran al encontrar la mejor distribución de la entrada (pulsando **Ajustar** o **Ajustar todas** en la pestaña de Simulación). Por ejemplo, si el nivel de medición es continuo, la distribución normal (adecuada para datos continuos) se consideraría, pero la distribución binomial no.

Nota: Seleccione el nivel de medición de Cadena para las entradas de cadena. Las entradas de cadena que se simularán están restringidas a la distribución categórica.

Entrada de valor fijo. Especifica que se conoce el valor de la entrada y que se fijará en el valor conocido. Las entradas fijas pueden ser numéricas o de cadena. Especifique un valor para la entrada fija. Los valores de cadena no deben ir entre comillas.

Etiquetas de valor. Puede especificar etiquetas de valores de destino, entradas simuladas y fijas. Las etiquetas de valores se utilizan en gráficos y tablas de resultados.

Pestaña Simulación

La pestaña Simulación especifica todas las propiedades de la simulación diferentes a la del modelo predictivo. Puede ejecutar las siguientes tareas generales en la pestaña Simulación:

- Especificar las distribuciones de probabilidad de las entradas simuladas y valores de entradas fijas.
- Especificar correlaciones entre entradas simuladas. Para entradas categóricas, puede especificar que se utilicen las asociaciones que existen entre las entradas del conjunto de datos activos cuando se generen datos para dichas entradas.
- Especificar opciones avanzadas como muestreos de cola y criterios para ajustar distribuciones a datos históricos.

- Personalizar los resultados.
- Especificar la ubicación en la que guardar el plan de simulación y opcionalmente, guardar los datos simulados.

Campos simulados

Para ejecuar una simulación, cada campo de entrada debe especificarse como fijo o simulado. Las entradas simuladas son aquellas cuyos valores son inciertos y se generarán a partir de una distribución de probabilidad especificada. Cuando los datos históricos están disponibles para que se simulen las entradas, las distribuciones que se ajusten mejor a los datos se podrán determinar de forma automática, junto a las correlaciones entre dichas entradas. También puede especificar manualmente las distribuciones o correlaciones si los datos históricos no están disponibles o si necesita distribuciones o correlaciones específicas.

Las entradas fijas son aquellas cuyos valores se conocen y permanecen constantes en cada caso generado en la simulación. Por ejemplo, tiene un modelo de regresión lineal de ventas como una función de un número de entradas incluyendo el precio y desea mantener el precio fijo en el mercado actual. A continuación especificaría el precio como entrada fija.

Para simulaciones basadas en un modelo predictivo, cada predictor del modelo es un campo de entrada para la simulación. Para simulaciones que no incluyen un modelo predictivo, los campos que se especifican en la pestaña Modelo son los campos para la simulación.

Ajuste automático de distribuciones y cálculo de correlaciones de entradas simuladas. Si la base de datos activa contiene datos históricos de las entradas que desee simular, puede encontrar automáticamente las distribuciones que mejor se ajustan a esas entradas, así como determinar las correlaciones entre ellas. Los pasos son los siguientes:

1. Compruebe que todas las entradas que desea simular coinciden con el campo correcto en el conjunto de datos activo. Las entradas se listan en la columna Entrada y la columna Ajustar muestra el campo correspondiente en el conjunto de datos activos. Puede relacionar una entrada con un campo diferente en el conjunto de datos activo seleccionando un elemento diferente en la lista desplegable Ajustar a.

   Un valor de *-Ninguno-* en la columna Ajustar a indica que la entrada no se ha podido relacionar automáticamente a un campo en el conjunto de datos activo. De forma predeterminada, las entradas se relacionan con los campos del conjunto en el nombre y el nivel de medición y tipo (numéricas o de cadena). Si el conjunto de datos activo no contiene datos históricos de la entrada, especifique manualmente la distribución de la entrada o especifique la entrada como entrada fija, tal y como se describe a continuación.
2. Pulse en **Ajustar todas**.

La distribución con mejor ajuste y sus parámetros asociados se muestran en la columna Distribución junto con una representación de la distribución superpuesta en un histograma (o gráfico de barras) de los datos históricos. Las correlaciones entre las entradas simuladas se muestran en los ajustes de correlaciones. Puede examinar los resultados de ajuste y personalizar el ajuste de distribución automático de una entrada concreta seleccionando la fila de la entrada y pulsando en **Detalles del ajuste**. Consulte el tema “Detalles del ajuste” en la página 187 para obtener más información.

Puede ejecutar el ajuste de distribución automático de una entrada concreta seleccionando la fila de la entrada y pulsando en **Ajustar**. Las correlaciones de todas las entradas simuladas que corresponden con los campos del conjunto de datos activo también se calculan automáticamente.

Nota:

- Los casos con valores que faltan para cualquier entrada simulada se excluyen del ajuste de distribución, el cálculo de correlaciones y el cálculo de la tabla de contingencia opcional (para entradas con una distribución categórica). De forma opcional, puede especificar si los valores que faltan del

usuario de entradas con una distribución categórica se tratan como válidos. De forma predeterminada, se tratan como valores que faltan. Para obtener más información, consulte el tema “Opciones avanzadas” en la página 189.

- Para las entradas continuas y ordinales, si no se puede encontrar ningún ajuste aceptable para cualquiera de las distribuciones probadas, se sugiere la distribución empírica como el ajuste más cercano. En entradas continuas, la distribución empírica es la función de distribución acumulada de los datos históricos. En entradas ordinales, la distribución empírica es la distribución categórica de los datos históricos.

Distribuciones de especificación manuales. Puede especificar manualmente la distribución de probabilidad de cualquier entrada simulada seleccionando la distribución de la lista desplegable **Tipo** e introduciendo los parámetros de distribución de la cuadrícula Parámetros. Una vez haya introducido los parámetros de una distribución, un gráfico de muestra de la distribución, basado en los parámetros específicos, se mostrará junto a la cuadrícula Parámetros. A continuación se incluyen algunas notas sobre distribuciones concretas:

- **Categóricas.** La distribución categórica describe un campo de entrada que tiene un número fijo de valores, referido como categorías. Cada categoría tiene una probabilidad asociada, de forma que la suma de las probabilidades de todas las categorías es igual a uno. Para especificar una categoría, pulse con el botón derecho del ratón en la columna de la izquierda de la cuadrícula Parámetros y especifique el valor de la categoría. Especifique la probabilidad asociada con la categoría en la columna derecha.

  Nota: Las entradas categóricas de un modelo PMML tienen categorías que se determinan a partir del modelo y no se pueden modificar.
- **Binomial negativa - Fallos.** Describe la distribución del número de fallos en una secuencia de intentos antes de detectar un número especificado de éxitos. El parámetro *thresh* es el número especificado de éxitos y el parámetro *prob* es la probabilidad de éxito en cualquier prueba.
- **Binomial negativa - Intentos.** Describe la distribución del número de intentos necesarios para detectar un número especificado de éxitos. El parámetro *thresh* es el número especificado de éxitos y el parámetro *prob* es la probabilidad de éxito en cualquier prueba.
- **Rango.** Esta distribución consiste en un conjunto de intervalos con una probabilidad asignada a cada intervalo de forma que la suma de probabilidades en todos los intervalos es igual a 1. Los valores en un intervalo concreto se generan a partir de una distribución uniforme definida en ese intervalo. Los intervalos se especifican introduciendo un valor mínimo, un valor máximo y una probabilidad asociada.

  Por ejemplo, usted cree que el coste de una materia prima tiene un 40% de posibilidades de caer entre 10 - 15 dólares por unidad y un 60% de posibilidades de caer entre 15 - 20 dólares por unidad. Modelaría el coste con una distribución del rango de los dos intervalos [10 - 15] y [15 - 20], definiendo la probabilidad asociada con el primer intervalo a 0,4 y la probabilidad asociada con el segundo intervalo a 0,6. Los intervalos no tienen que ser contiguos y se pueden superponer. Por ejemplo, podría especificar los intervalos 10 - 15 y 20 - 25 o 10 - 15 y 13 - 16.
- **Weibull.** El parámetro *c* es un parámetro de ubicación opcional, que especifica dónde se encuentra el origen de la distribución.

Los parámetros de las siguientes distribuciones tienen el mismo significado en las funciones variables aleatorias asociadas disponibles en el cuadro de diálogo Calcular variable: Bernoulli, Beta, Binomial, Exponencial, Gamma, Lognormal, Binomial negativa (Intentos y Fallos), Normal, Poisson y Uniforme.

Especificación de entradas fijas. Especifica una entrada fija seleccionando Fija en la lista desplegable **Tipo** de la columna Distribución, e introduciendo el valor fijo. El valor puede ser numérico o cadena, dependiendo de si la entrada es numérica o de cadena. Los valores de cadena no deben ir entre comillas.

Especificación de límites en valores simulados. La mayoría de distribuciones admiten la especificación de los límites superiores e inferiores de los valores simulados. Puede especificar un límite inferior introduciendo un valor en el cuadro de texto **Mín** y especificar un límite superior introduciendo un valor en el cuadro de texto **Máx**.

Bloqueo de entradas. El bloqueo de una entrad, seleccionando la casilla de verificación en la columna con el icono de bloqueo, excluye la entrada del ajuste de distribución automático. Esto es muy útil cuando especifica manualmente una distribución o un valor fijo y desea garantizar que no se verá afectada por el ajuste de distribución automático. El bloqueo también es útil si desea compartir el plan de simulación con los usuarios que lo ejecutarán en el diálogo Ejecutar simulación y desea evitar los cambios en algunas entradas. En este sentido, las especificaciones para entradas bloqueadas no se pueden modificar en el diálogo Ejecutar simulación.

Análisis de sensibilidad. Los análisis de sensibilidad permiten investigar el efecto de los cambios sistemáticos en una entrada fija o en un parámetro de distribución de una entrada estimulada generando un conjunto independiente de casos simulados (una simulación separada) para cada valor especificado. Para especificar los análisis de sensibilidad, seleccione una entrada fija o simulada y pulse en **Análisis de sensibilidad**. El análisis de sensibilidad está limitado a una única entrada fija o un parámetro de distribución único de una entrada simulada. Consulte el tema “Análisis de sensibilidad” en la página 188 para obtener más información.

Iconos de estado de Ajustar

Los iconos de la columna Ajustar a indican el estado de ajuste de cada campo de entrada.

Tabla 3. Iconos de estado.

Icono	Descripción
	No se ha especificado una distribución para la entrada y la entrada no se ha especificado como fija. Para ejecutar la simulación, debe especificar una distribución para esta entrada o definirla como fija y especificar el valor fijo.
	La entrada se ha fijado anteriormente a un campo que no existe en el conjunto de datos activo. No se necesita ninguna acción salvo que desee volver a ajustar la distribución de la entrada al conjunto de datos activo.
	La mejor distribución de ajuste se ha sustituido por una distribución diferente del cuadro de diálogo Detalles del ajuste.
	La entrada se define a la mejor distribución de ajuste.
	La distribución se ha especificado manualmente o se han especificado las iteraciones del análisis de sensibilidad para esta entrada.

Detalles del ajuste: El cuadro de diálogo Detalles del ajuste muestra los resultados del ajuste de distribución automático de una entrada concreta. Las distribuciones se ordenan por idoneidad de ajuste, con la distribución de mejor ajuste en primer lugar. Puede sustituir la distribución de mejor ajuste seleccionando el botón de opción de la distribución que desee en la columna Utilizar. Al seleccionar un botón de opción en la columna Utilizar también muestra un gráfico de distribución superpuesto en un histograma (o gráfico de barras) de los datos históricos de esa entrada.

Estadísticos de ajuste. De forma predeterminada, en los campos continuos, la prueba Anderson-Darling se utiliza para determinar su idoneidad. Del mismo modo, y para los campos continuos únicamente, puede especificar la prueba Kolmogorov-Smirnoff de idoneidad, seleccionando esa opción en los ajustes Opciones avanzadas. En entradas continuas, los resultados de ambas pruebas se muestran en la columna

Estadísticos de ajuste (A para Anderson-Darling y K para Kolmogorov-Smirnoff), usando la prueba elegida para ordenar las distribuciones. En las entradas ordinales y nominales se utiliza la comprobación de chi-cuadrado. También se muestran los valores p asociados con las pruebas.

Parámetros. Los parámetros de distribución asociados con cada distribución ajustada se muestran en la columna Parámetros. Los parámetros de las siguientes distribuciones tienen el mismo significado en las funciones variables aleatorias asociadas disponibles en el cuadro de diálogo Calcular variable: Bernoulli, Beta, Binomial, Exponencial, Gamma, Lognormal, Binomial negativa (Intentos y Fallos), Normal, Poisson y Uniforme. Consulte el tema para obtener más información. En la distribución categórica, los nombres de los parámetros son las categorías y los valores de parámetros son las probabilidades asociadas.

Reajuste con un conjunto de distribución personalizado. De forma predeterminada, el nivel de medición de la entrada se utiliza para determinar el conjunto de distribuciones consideradas para el reajuste de distribución automático. Por ejemplo, las distribuciones continuas como lognormal y gamma se consideran cuando se ajusta una entrada continua; pero las distribuciones discretas como Poisson y binomial, no. Puede elegir un subconjunto de las distribuciones predeterminado seleccionando las distribuciones en la columna Reajustar. También puede sustituir el conjunto predeterminado de distribuciones seleccionando un nivel de medición diferente en la lista desplegable **Tratar como (Medida)** y seleccionando las distribuciones en la columna Reajustar. Pulse en **Ejecutar reajuste** para volver a ajustar con el conjunto de distribución personalizado.

Nota:

- Los casos con valores que faltan para cualquier entrada simulada se excluyen del ajuste de distribución, el cálculo de correlaciones y el cálculo de la tabla de contingencia opcional (para entradas con una distribución categórica). De forma opcional, puede especificar si los valores que faltan del usuario de entradas con una distribución categórica se tratan como válidos. De forma predeterminada, se tratan como valores que faltan. Para obtener más información, consulte el tema “Opciones avanzadas” en la página 189.
- Para las entradas continuas y ordinales, si no se puede encontrar ningún ajuste aceptable para cualquiera de las distribuciones probadas, se sugiere la distribución empírica como el ajuste más cercano. En entradas continuas, la distribución empírica es la función de distribución acumulada de los datos históricos. En entradas ordinales, la distribución empírica es la distribución categórica de los datos históricos.

Análisis de sensibilidad: Los análisis de sensibilidad permiten investigar el efecto de modificar una entrada fija o un parámetro de distribución de una entrada simulada en un conjunto específico de valores. Se genera un conjunto independiente de casos simulados (una simulación separada) para cada valor especificado, lo que le permite investigar el efecto de modificar la entrada. Cada conjunto de casos simulados se denomina **iteración**.

Iterar. Esta opción permite especificar el conjunto de valores en el que se modificará la entrada.

- Si modifica el valor de un parámetro de distribución, seleccione el parámetro de la lista desplegable. Introduzca el conjunto de valores en el valor Parámetro mediante la cuadrícula de iteraciones. Al pulsar en **Continuar** se añadirán los valores especificados en la cuadrícula Parámetros de la entrada asociada, con un índice que especifica el número de iteración del valor.
- En las distribuciones categóricas y de rango, las probabilidades de las categorías o intervalos respectivamente se pueden variar, pero los valores de las categorías y los puntos finales de los intervalos no se pueden modificar. Seleccione una categoría o intervalo en la lista desplegable y especifique el conjunto de probabilidades en el valor del parámetro mediante la cuadrícula de iteración. Las probabilidades de otras categorías o intervalos se ajustarán automáticamente.

Sin iteraciones. Utilice esta opción para cancelar las iteraciones de una entrada. Si pulsa en **Continuar** eliminará las iteraciones.

Correlaciones

Los campos de entrada que se van a simular con frecuencia se sabe que están correlacionados como, por ejemplo, altura y ponderación. Las correlaciones entre las entradas que se simularán se deben tener en cuenta para garantizar que los valores simulados mantengan esas correlaciones.

Recalcular correlaciones al ajustar. Esta opción especifica que las correlaciones entre las entradas simuladas se calculen automáticamente al ajustar las distribuciones en el conjunto de datos activo mediante las acciones **Ajustar todas** o **Ajustar** en los ajustes de Campos simulados.

No recalcular correlaciones al ajustar. Seleccione esta opción si desea especificar manualmente las correlaciones y evitar que se sobrescriban al ajustar automáticamente distribuciones en el conjunto de datos activo. Los valores entrados en la cuadrícula de correlaciones deben estar entre -1 y 1. Un valor de 0 especifica que no existe ninguna correlación entre el par de entradas asociadas.

Restablecer. Restablece todas las correlaciones a 0.

Utilizar tabla de contingencia de varios factores ajustada para entradas con una distribución categórica. Para entradas con una distribución categórica, puede calcular automáticamente una tabla de contingencia de varios factores a partir del conjunto de datos activo que describe las asociaciones entre dichas entradas. La tabla de contingencia se utiliza cuando se generan los datos para dichas entradas. Si opta por guardar el plan de simulación, la tabla de contingencia se guarda en el archivo de plan y se utilizan cuando se ejecuta el plan.

- **Calcular tabla de contingencia a partir del conjunto de datos activo.** Si trabaja con un plan de simulación existente que contiene una tabla de contingencia, puede volver a calcular la tabla de contingencia a partir del conjunto de datos activo. Esta acción sustituye la tabla de contingencia del archivo de plan cargado.
- **Utilizar tabla de contingencia del plan de simulación cargado.** De forma predeterminada, cuando carga un plan de simulación que contiene una tabla de contingencia, se utiliza la tabla del plan. Puede volver a calcular la tabla de contingencia a partir del conjunto de datos activo seleccionando **Calcular tabla de contingencia a partir del conjunto de datos activo**.

Opciones avanzadas

Número máximo de casos. Especifica el número máximo de casos de datos simulados y los valores de destinos asociados que se generarán. Si se especifica el análisis de sensibilidad, es el número máximo de casos de cada iteración.

Destino de criterios de parada. Si su modelo predictivo contiene más de un destino, podrá seleccionar el destino al que se aplican los criterios de parada.

Criterios de parada. Estas opciones especifican los criterios para detener la simulación, potencialmente antes de generar el número máximo de casos permitidos.

- **Continuar hasta alcanzar el máximo.** Especifica que los casos simulados se generarán hasta que se alcance el número máximo de casos.
- **Detener cuando las colas se hayan muestreado.** Utilice esta opción si desea asegurarse de que una de las colas de una distribución de destino especificada se ha muestreado correctamente. Los casos simulados se generarán hasta que se complete el muestreo de la cola específica o se alcance el número máximo de casos. Si su modelo predictivo contiene múltiples destinos, seleccione el destino al que se aplicarán los criterios en la lista desplegable **Destino de criterios de parada**.

  Tipo. Puede definir el límite de la región de cola especificando un valor de destino como 10.000.000 o un percentil como el 99º. Si selecciona Valor en la lista desplegable **Tipo**, introduzca el valor del límite en el cuadro de texto Valor y utilice la lista desplegable **Lado** para especificar si es el límite de la región de cola izquierda o la región de cola derecha. Si selecciona Percentil en la lista desplegable **Tipo**, introduzca un valor en el cuadro de texto Percentil.

Frecuencia. Especifique el número de valores del destino que debe estar en la región para garantizar que la cola se ha muestreado correctamente. La generación de los casos se detendrá cuando se alcance este número.

- **Detener si el intervalo de confianza de la media está en el umbral especificado.** Utilice esta opción si desea asegurarse de que la media de un destino concreto se conoce con un grado especificado de precisión. Los casos simulados se generarán hasta que se alcance el grado específico de precisión o el número máximo de casos. Para utilizar esta opción, especifique un nivel de confianza y un umbral. Los casos simulados se generarán hasta que el intervalo de confianza asociado con el nivel de confianza esté dentro del umbral. Por ejemplo, puede utilizar esta opción para especificar que los casos se generan hasta que el intervalo de confianza de la media al 95% del nivel de confianza esté en el 5% del valor medio. Si su modelo predictivo contiene múltiples destinos, seleccione el destino al que se aplicarán los criterios en la lista desplegable **Destino de criterios de parada**.

  Tipo de umbral. Puede especificar el umbral como un valor numérico o como porcentaje de la media. Si selecciona Valor en la lista desplegable **Tipo de umbral**, introduzca el umbral en el cuadro de texto Umbral como valor. Si selecciona Porcentaje en la lista desplegable **Tipo de umbral**, introduzca un valor en el cuadro de texto Umbral como porcentaje.

Número de casos de muestreo. Especifica el número de casos que se utilizarán al ajustar automáticamente las distribuciones de las entradas simuladas en el conjunto de datos activo. Si su conjunto de datos es muy grande, puede que desee limitar el número de casos utilizados para el ajuste de distribución. Si selecciona **Limita a N casos**, se utilizarán los primeros N casos.

Bondad de criterios de ajuste (Continua). En entradas continuas, puede utilizar la prueba Anderson-Darling o la prueba Kolmogorov-Smirnoff de bondad de ajuste para ordenar distribuciones al ajustar distribuciones de entradas simuladas en el conjunto de datos activo. La prueba Anderson-Darling está seleccionada de forma predeterminada y está especialmente recomendada si desea garantizar el mejor ajuste posible en las regiones de cola.

Distribución empírica. En entradas continuas, la distribución empírica es la función de distribución acumulada de los datos históricos. Puede especificar el número de intervalos utilizados para calcular las distribuciones empíricas de las entradas continuas. El valor predeterminado es 100 y el máximo es 1.000.

Replicar resultados. Al establecer una semilla aleatoria podrá replicar las simulaciones. Especifique un entero o pulse en **Generar**, lo que creará un entero pseudo-aleatorio entre 1 y 2147483647, ambos inclusive. El valor predeterminado es 629111597.

Nota: Para una semilla aleatoria determinada, los resultados se duplican a menos que cambie el número de hebras. En un sistema determinado, el número de hebras es constante a menos que lo cambie ejecutando la sintaxis de comando `SET THREADS`. El número de hebras puede cambiar si se ejecuta la simulación en un sistema diferente porque se utiliza un algoritmo interno para determinar el número de hebras para cada sistema.

Valores perdidos del usuario para entradas con una distribución categórica. Estos controles especifican si los valores perdidos del usuario de entradas con una distribución categórica se tratan como válidos. Los valores perdidos del sistema para todos los tipos de entrada se tratan siempre como no válidos. Todas las entradas deben tener valores válidos para que un caso se incluya en el ajuste de distribución, el cálculo de correlaciones y el cálculo de la tabla de contingencia opcional.

Funciones de densidad

Estos ajustes permiten personalizar los resultados de las funciones de densidad de probabilidad y las funciones de distribución acumuladas de destinos continuos, así como gráficos de barras de los valores pronosticados de destinos categóricos.

Función de densidad de probabilidad (PDF). La función de densidad de probabilidad muestra la distribución de valores de destino. En destinos continuos, permite determinar la probabilidad de que el destino esté dentro de una región concreta. En destinos categóricos (destinos con un nivel de medición

del nominal u ordinal), se genera un gráfico de barras que muestra en porcentaje de casos que entran dentro de cada categoría del destino. Existen otras opciones para los destinos categóricos de modelos PMML disponibles con los valores de Categoría para los ajustes de informes que se describen a continuación.

Para modelos de clúster de dos fases y modelos de clúster de K-medias, se genera un gráfico de barras de la pertenencia a clústeres.

Función de distribución acumulada (CDF). La función de distribución acumulada muestra la probabilidad de que el valor del destino sea menor o igual que un valor especificado. Solo está disponible para destinos continuos.

Posiciones del deslizador. Puede especificar las posiciones iniciales de las líneas de referencia móviles en gráficos PDF y CDF. Los valores que se especifican para las líneas inferior y superior hacen referencia a las posiciones a lo largo del eje horizontal, no a percentiles. Puede eliminar la línea inferior seleccionando **-Infinity** o puede eliminar la línea superior seleccionando **Infinity**. De forma predeterminada, las líneas se sitúan en los percentiles 5 y 95. Cuando se muestran varias funciones de distribución en un único gráfico (debido a varios destinos o resultados de iteraciones del análisis de sensibilidad), el valor predeterminado hace referencia a la distribución de la primera iteración o del primer destino.

Líneas de referencia (Continuas). Puede solicitar varias líneas de referencia verticales para añadirlas a funciones de densidad de probabilidad y funciones de distribución acumulada para destinos continuos.

- **Sigmas.** Puede añadir líneas de referencia por encima o por debajo de un número especificado de desviaciones estándar desde la media de un objetivo.
- **Percentiles.** Ahora puede añadir líneas de referencia a uno o dos valores de los percentiles de la distribución de un objetivo introduciendo los valores en los cuadros de texto inferior y superior. Por ejemplo, un valor de 95 en el cuadro de texto superior representa el percentil 95, que es el valor por debajo del cual cae el 95% de las observaciones. Del mismo modo, un valor de 5 en el cuadro de texto inferior representa el percentil 5, que es el valor por debajo del cual cae el 5% de las observaciones.
- **Líneas de referencia personalizadas.** Puede añadir líneas de referencia a los valores especificados del destino.

Nota: Cuando se muestran varias funciones de distribución en un único gráfico (debido a varios destinos o resultados de iteraciones del análisis de sensibilidad), las líneas de referencia sólo se aplican a la distribución de la primera iteración o del primer destino. Puede añadir líneas de referencia a otras distribuciones desde el diálogo Opciones de gráfico, al que se accede desde el gráfico PDF o CDF.

Superponer resultados de destinos continuos diferentes. En el caso de múltiples destinos continuos, especifica si las funciones de distribución de todos estos destinos se muestran en un único gráfico, con un gráfico de funciones de densidad de probabilidad y otra para funciones de distribución acumuladas. Si esta opción no se selecciona, los resultados de cada destino se mostrarán en un gráfico diferente.

Valores de categoría que se incluirán en el informe. En modelos PMML con destinos categóricos, el resultado del modelo es un conjunto de probabilidades pronosticadas, una de cada categoría, en la que el valor de destino entra dentro de cada categoría. La categoría con la mayor probabilidad se toma como la categoría pronosticada y se utiliza en la generación del gráfico de barras que se describe en el ajuste **Función de densidad de probabilidad** anterior. Si selecciona **Categoría pronosticada** se generará el gráfico de barras. Si selecciona **Probabilidades pronosticadas** se generarán histogramas de la distribución de las probabilidades pronosticadas de cada una de las categorías del destino.

Agrupación de análisis de sensibilidad. Las simulaciones que incluyen análisis de sensibilidad generan un conjunto independiente de valores de destino pronosticados para cada iteración que define el análisis (una iteración para cada valor de la entrada que se está variando). Si existen iteraciones, el gráfico de

barras de la categoría pronosticada de un destino categórico se muestra como un gráfico de barras clúster que incluye los resultados de todas las iteraciones. Puede seleccionar agrupar las categorías o agrupar las iteraciones.

Resultado

Gráficos de tornado. Los gráficos de tornado son gráficos de barra que muestran relaciones entre destinos y entradas simuladas utilizando una variedad de métricas.

- **Correlación del destino con la entrada.** Esta opción crea un gráfico de tornado de los coeficientes de correlación entre un destino especificado y cada una de sus entradas simuladas. Este tipo de gráfico de tornado no admite destinos con un nivel de medición nominal u ordinal ni entradas simuladas con una distribución categórica.
- **Contribución a la varianza.** Esta opción crea un gráfico de tornado que muestra la contribución a la varianza de un destino de cada una de sus entradas simuladas, lo que permite evaluar el grado en el que cada entrada contribuye a la incertidumbre general del destino. Este tipo de gráfico de tornado no admite destinos con niveles de medición nominales u ordinales, o entradas simuladas con cualquiera de las distribuciones siguientes: categórica, Bernoulli, binomial, Poisson o binomial negativa.
- **Sensibilidad del destino para cambiar.** Esta opción crea un gráfico de tornado que muestra el efecto del destino del destino de modulación de cada entrada simulada más o menos un número especificado de desviaciones estándar de la distribución asociada con la entrada. Este tipo de gráfico de tornado no admite destinos con niveles de medición nominales u ordinales, o entradas simuladas con cualquiera de las distribuciones siguientes: categórica, Bernoulli, binomial, Poisson o binomial negativa.

Diagramas de caja de distribuciones de destino. Los diagramas de caja están disponibles para destinos continuos. Seleccione **Superponer resultados de destinos diferentes** si su modelo predictivo tiene múltiples destinos continuos y desea visualizar los diagramas de caja de todos los destinos en un gráfico único.

Diagramas de dispersión de destinos frente a entradas. Los diagramas de dispersión frente a entradas simuladas están disponibles para destinos continuos y categóricos e incluyen dispersiones del destino con entradas continuas y categóricas. Las dispersiones que incluyen un destino o una entrada categórica se muestran como un mapa de calor.

Crear una tabla de valores percentiles. En destinos continuos, puede obtener una tabla de percentiles especificados de las distribuciones de destino. Los cuartiles (los percentiles 25, 50 y 75) dividen las observaciones en cuatro grupos de igual tamaño. Si desea un número igual de grupos que no sea cuatro, seleccione **Intervalos** y especifique el número. Seleccione **Percentiles personalizados** para especificar percentiles individuales, por ejemplo, el percentil 99.

Estadísticos descriptivos de distribuciones de destino. Esta opción crea tablas de estadísticos descriptivos de destinos continuos y categóricos así como entradas continuas. En destinos continuos, la tabla incluye la media, la desviación estándar, la mediana, el mínimo y el máximo, el intervalo de confianza de la media en el nivel especificado y los percentiles 5 y 95 de la distribución de destino. En destinos categóricos, la tabla incluye el porcentaje de casos que entran en cada categoría del destino. En destinos categóricos de modelos PMML, la tabla también incluye la probabilidad media de cada categoría del destino. En entradas continuas, la tabla incluye la media, desviación estándar, mínima y máxima.

Correlaciones y tabla de contingencia para entradas. Esta opción muestra una tabla de coeficientes de correlación entre entradas simuladas. Cuando se generan entradas con distribuciones categóricas a partir de una tabla de contingencia, también se muestra la tabla de contingencia de los datos que se generan para dichas entradas.

Entradas simuladas que se incluirán en el resultado. De forma predeterminada, todas las entradas simuladas se incluyen en los resultados. Puede excluir las entradas simuladas de las salidas. Se excluirán de los gráficos de tornado, diagramas de dispersión y resultados tabulares.

Rangos de límite para destinos continuos. Puede especificar el rango de valores válidos para uno o más destinos continuos. Los valores que están fuera del rango especificado se excluyen de todos los análisis de salida asociados con los destinos. Para establecer un límite inferior, seleccione **Inferior** en la columna Límite y especifique un valor en la columna Mínimo. Para establecer un límite superior, seleccione **Superior** en la columna Límite y especifique un valor en la columna Máximo. Para establecer un límite inferior y un límite superior, seleccione **Ambos** en la columna Límite y especifique los valores en las columnas Mínimo y Máximo.

Formatos de visualización. Puede definir el formato utilizado cuando se visualizan los valores de destinos y entradas (tanto entradas fijas como simuladas).

Guardar

Guardar el plan de esta simulación. Puede guardar las especificaciones actuales de su simulación en un archivo de plan de simulación. Los archivos de plan de simulación tienen la extensión *.splan*. Puede volver a abrir el plan en el Generador de simulaciones, y también puede realizar modificaciones y ejecutar la simulación. Puede compartir el plan de simulación con otros usuarios, que pueden ejecutarlo en el cuadro de diálogo Ejecutar simulación. Los planes de simulación incluyen todas las especificaciones excepto las siguientes: ajustes de Funciones de densidad; Ajustes de resultados de gráficos y tablas; Opciones avanzadas de ajustes, Distribución empírica y Semilla aleatoria.

Guardar los datos simulados como un nuevo archivo de datos. Puede guardar entradas simuladas, entradas fijas y valores de destino pronosticados en un archivo de datos SPSS Statistics, un nuevo conjunto de datos en la sesión actual o un archivo Excel. Cada caso (o fila) del archivo de datos consta de los valores pronosticados de los objetivos junto con las entradas simuladas y las entradas fijas que generan los valores objetivo. Si se especifica el análisis de sensibilidad, cada iteración genera un conjunto contiguo de casos que se etiquetan con el número de iteración.

Cuadro de diálogo Ejecutar simulación

El cuadro de diálogo Ejecutar simulación está diseñado para usuarios que tienen un plan de simulación y que desean preferentemente ejecutar la simulación. También proporciona las características que necesita para ejecutar la simulación en condiciones diferentes. Permite ejecutar las siguientes tareas generales:

- Configurar o modificar los análisis de sensibilidad para investigar el efecto de variar el valor de una entrada fija o variar un parámetro de distribución de una entrada simulada.
- Reajustar distribuciones de probabilidad de entradas inciertas (y correlaciones entre esas entradas) a nuevos datos.
- Modificar la distribución de una entrada simulada.
- Personalizar los resultados.
- Ejecutar la simulación.

Pestaña Simulación

La pestaña Simulación permite especificar análisis de sensibilidad, reajustar distribuciones de probabilidad de entradas simuladas y correlaciones entre entradas simuladas en nuevos datos y modificar la distribución de probabilidad asociada con una entrada simulada.

La cuadrícula Entradas simuladas contiene una entrada para cada campo de entrada que se define en el plan de simulación. Cada entrada muestra el nombre y el tipo de distribución de probabilidad asociada con la entrada, junto con un gráfico de muestra de la curva de distribución asociada. Cada entrada también tiene un icono de estado asociado (un círculo de color con una marca de verificación) que es útil si está reajustando distribuciones a nuevos datos. Además, las entradas pueden incluir un icono de candado que indica que la entrada está bloqueada y no se puede modificar o reajustar a nuevos datos en el cuadro de diálogo Ejecutar simulación. Para modificar una entrada bloqueada, necesitará abrir el plan de simulación en el Generador de simulaciones.

Cada entrada puede ser simulada o fija. Las entradas simuladas son aquellas cuyos valores son inciertos y se generarán a partir de una distribución de probabilidad especificada. Las entradas fijas son aquellas cuyos valores se conocen y permanecen constantes en cada caso generado en la simulación. Para trabajar con una entrada concreta, seleccione la entrada en la cuadrícula Entradas simuladas.

Especificación de análisis de sensibilidad

Los análisis de sensibilidad permiten investigar el efecto de los cambios sistemáticos en una entrada fija o en un parámetro de distribución de una entrada estimulada generando un conjunto independiente de casos simulados (una simulación separada) para cada valor especificado. Para especificar los análisis de sensibilidad, seleccione una entrada fija o simulada y pulse en **Análisis de sensibilidad**. El análisis de sensibilidad está limitado a una única entrada fija o un parámetro de distribución único de una entrada simulada. Consulte el tema "Análisis de sensibilidad" en la página 188 para obtener más información.

Reajuste de distribuciones a nuevos datos

Para reajustar automáticamente las distribuciones de probabilidad de entradas simuladas (y las correlaciones entre entradas simuladas) a los datos del conjunto de datos activo:

1. Compruebe que todas las entradas del modelo coinciden con el campo correcto del conjunto de datos activo. Todas las entradas simuladas se reajustan al campo del conjunto de datos activo especificado en la lista desplegable **Campo** con esa entrada. Puede identificar fácilmente las entradas no coincidentes buscando entradas con un icono de estado que incluya una marca de verificación con un signo de interrogación, tal y como se muestra a continuación.

2. Modifique cualquier campo necesario coincidente seleccionando **Ajustar a un campo en el conjunto de datos** y seleccionando el campo de la lista.
3. Pulse en **Ajustar todas**.

En cada entrada ajustada, la distribución que mejor se ajusta a los datos se muestra junto con una representación de la distribución superpuesta en un histograma (o gráfico de barras) de los datos históricos. Si no se encuentra un ajuste aceptable, se utilizará la distribución empírica. En el caso de las entradas que se ajusten a la distribución empírica, solo verá un histograma de los datos históricos porque la distribución empírica está, de hecho, representada por ese histograma.

Nota: para ver una lista completa de iconos de estado, consulte el tema "Campos simulados" en la página 185.

Modificación de las distribuciones de probabilidad

Puede modificar la distribución de probabilidad de una entrada simulada y, opcionalmente, cambiar una entrada simulada a una entrada fija o viceversa.

1. Seleccione la entrada y seleccione **Ajustar la distribución manualmente**.
2. Seleccione el tipo de distribución y especifique los parámetros de distribución. Para cambiar una entrada simulada por una entrada fija, seleccione Fija en la lista desplegable **Tipo**.

Una vez haya introducido los parámetros de una distribución, el gráfico de muestra de la distribución (se muestra en la entrada) se actualizará reflejando sus cambios. Para obtener más información sobre la especificación manual de distribuciones de probabilidad, consulte el tema "Campos simulados" en la página 185.

Incluir valores perdidos del usuario de entradas categóricas al ajustar. Esto especifica si los valores perdidos del usuario de las entradas con una distribución categórica se tratan como válidos cuando se vuelven a ajustar los datos en el conjunto de datos activo. Los valores perdidos del sistema y los valores perdidos del usuario para todos los demás tipos de entradas siempre se tratan como no válidos. Todas las entradas deben tener valores válidos para que un caso se incluya en el ajuste de distribución y el cálculo de correlaciones.

Pestaña Resultado

La pestaña Resultado permite personalizar los resultados que genera la simulación.

Funciones de densidad. Las funciones de densidad son las medias principales de la prueba del conjunto de resultados de su simulación.

- **Función de densidad de probabilidad** La función de densidad de probabilidad muestra la distribución de los valores de destino, lo que permite determinar la probabilidad de que el destino esté dentro de una región concreta. En destinos con un conjunto fijo de resultados, como "servicio deficiente", "servicio correcto", "buen servicio" y "excelente servicio", se genera un gráfico de barras que muestra el porcentaje de casos que entran en cada categoría del destino.
- **Función de distribución acumulada.** La función de distribución acumulada muestra la probabilidad de que el valor del destino sea menor o igual que un valor especificado.

Gráficos de tornado. Los gráficos de tornado son gráficos de barra que muestran relaciones entre destinos y entradas simuladas utilizando una variedad de métricas.

- **Correlación del destino con la entrada.** Esta opción crea un gráfico de tornado de los coeficientes de correlación entre un destino especificado y cada una de sus entradas simuladas.
- **Contribución a la varianza.** Esta opción crea un gráfico de tornado que muestra la contribución a la varianza de un destino de cada una de sus entradas simuladas, lo que permite evaluar el grado en el que cada entrada contribuye a la incertidumbre general del destino.
- **Sensibilidad del destino para cambiar.** Esta opción crea un gráfico de tornado que muestra el efecto del destino del destino de modulación de cada entrada simulada más o menos una desviación estándar de la distribución asociada con la entrada.

Diagramas de dispersión de destinos frente a entradas. Esta opción genera diagramas de dispersión de destinos frente a entradas simuladas.

Diagramas de caja de distribuciones de destino. Esta opción genera diagramas de caja de las distribuciones de destino.

Tabla de cuartiles. Esta opción genera una tabla de cuartiles de las distribuciones de destino. Los cuartiles de una distribución son los percentiles 25, 50 y 75 de la distribución y dividen las observaciones en cuatro grupos de igual tamaño.

Correlaciones y tabla de contingencia para entradas. Esta opción muestra una tabla de coeficientes de correlación entre entradas simuladas. Una tabla de contingencia de asociaciones entre entradas con una distribución categórica se muestra cuando el plan de simulación especifica la generación de datos categóricos a partir de una tabla de contingencia.

Superponer resultados de destinos diferentes. Si el modelo predictivo que simula contiene múltiples destinos, puede especificar si los resultados de destinos diferentes se muestran en un único gráfico. Este ajuste se aplica a los gráficos de funciones de densidad de probabilidad, funciones de distribución acumuladas y diagramas de caja. Por ejemplo, si selecciona esta opción, las funciones de densidad de probabilidad de todos los destinos se mostrarán en un gráfico único.

Guardar el plan de esta simulación. Puede guardar las modificaciones de su simulación en un archivo de plan de simulación. Los archivos de plan de simulación tienen la extensión *.splan*. Puede volver a abrir

el plan en el cuadro de diálogo Ejecutar simulación o en el Generador de simulaciones. Los planes de simulación incluyen todas las especificaciones excepto los ajustes de resultados:

Guardar los datos simulados como un nuevo archivo de datos. Puede guardar entradas simuladas, entradas fijas y valores de destino pronosticados en un archivo de datos SPSS Statistics, un nuevo conjunto de datos en la sesión actual o un archivo Excel. Cada caso (o fila) del archivo de datos consta de los valores pronosticados de los objetivos junto con las entradas simuladas y las entradas fijas que generan los valores objetivo. Si se especifica el análisis de sensibilidad, cada iteración genera un conjunto contiguo de casos que se etiquetan con el número de iteración.

Si requiere más personalizaciones de un resultado que las disponibles aquí, considere ejecutar su simulación desde el Generador de simulaciones. Consulte el tema “Para ejecutar una simulación de un plan de simulación” en la página 181 para obtener más información.

Trabajo con resultados de gráficos de simulación

A partir de una simulación se generan diferentes gráficos que tienen características interactivas que permiten personalizar la vista. Las características interactivas están disponibles activando (pulsando dos veces) el objeto de gráfico en el Visor de resultados. Todos los gráficos de simulación son visualizaciones de tablero.

Gráficos de funciones de densidad de probabilidad de destinos continuos. Este gráfico tiene dos líneas de referencia verticales deslizantes que dividen el gráfico en regiones diferentes. La tabla bajo el gráfico muestra la probabilidad de que el destino esté en cada una de las regiones. Si se muestran múltiples funciones de densidad en el mismo gráfico, la tabla tiene una fila separada para las probabilidades asociadas con cada función de densidad. Cada una de las líneas de referencia tiene un deslizador (un triángulo invertido) que permite mover fácilmente la línea. Existe un número adicional de características disponibles pulsando en el botón **Opciones de gráfico**. En concreto, puede definir explícitamente las posiciones de los deslizadores, añadir líneas de referencia fijas y cambiar la vista del gráfico de una curva continua a un histograma o viceversa. Consulte el tema “Opciones de gráfico” en la página 197 para obtener más información.

Gráficos de funciones de distribución acumulada de destinos continuos. Este gráfico tiene las dos mismas líneas de referencia verticales móviles y la tabla asociada que se describe en el gráfico de funciones de densidad de probabilidad anterior. También proporciona acceso al cuadro de diálogo Opciones de gráfico, que permite definir explícitamente las posiciones de los deslizadores, añadir líneas de referencia fijas y especificar si la función de distribución acumulada se muestra como una función creciente (la opción predeterminada) o una función decreciente. Consulte el tema “Opciones de gráfico” en la página 197 para obtener más información.

Gráficos de barras de destinos categóricos con iteraciones del análisis de sensibilidad. En los destinos categóricos con iteraciones de análisis de sensibilidad, los resultados de la categoría de destino pronosticada se muestran como un gráfico de barras agrupadas que incluye los resultados de todas las iteraciones. El gráfico incluye una lista desplegable que permite agrupar según la categoría o la iteración. En modelos de clúster de dos fases y modelos de clúster de K-medias, puede seleccionar agrupar según el número o iteración de clústeres.

Diagramas de caja de múltiples destinos con iteraciones del análisis de sensibilidad. En modelos predictivos con múltiples destinos continuos e iteraciones del análisis de sensibilidad, si selecciona mostrar diagramas de caja de todos los destinos en un único gráfico, se producirá un diagrama de caja agrupado. El gráfico incluye una lista desplegable que permite agrupar según el destino o la iteración.

Opciones de gráfico

El cuadro de diálogo Opciones de gráfico permite personalizar la vista de los gráficos activados de funciones de densidad de probabilidad y funciones de distribución acumuladas generadas a partir de una simulación.

Ver. La lista desplegable **Ver** solo se aplica al gráfico de funciones de densidad de probabilidad. Permite cambiar la vista del gráfico de una curva continua a un histograma. Esta característica no está disponible si se muestran múltiples funciones de densidad en el mismo gráfico. En ese caso, las funciones de densidad solo se pueden visualizar como curvas continuas.

Ordenar. La lista desplegable **Ordenar** solo se aplica al gráfico de función de distribución acumulada. Especifica si la función de distribución acumulada se muestra como una función ascendente (la opción predeterminada) o una función descendente. Si se muestra como una función descendente, el valor de la función en un punto concreto del eje horizontal es la probabilidad de que el destino se encuentre a la derecha de ese punto.

Posiciones del deslizador. Puede definir explícitamente las posiciones de las líneas de referencia de los deslizadores introduciendo valores en los cuadros de texto Superior e Inferior. Puede eliminar la línea de la izquierda seleccionando **-Infinito**, definiendo la posición al infinito negativo, y puede eliminar la línea de la derecha seleccionando **Infinito**, definiendo la posición al infinito.

Líneas de referencia. Puede añadir varias líneas de referencia verticales fijas various funciones de densidad de probabilidad y funciones de distribución acumuladas. Cuando se muestran varias funciones en un único gráfico (debido a varios destinos o resultados de iteraciones del análisis de sensibilidad), puede especificar las funciones concretas a las que se aplican las líneas.

- **Sigmas.** Puede añadir líneas de referencia por encima o por debajo de un número especificado de desviaciones estándar desde la media de un objetivo.
- **Percentiles.** Ahora puede añadir líneas de referencia a uno o dos valores de los percentiles de la distribución de un objetivo introduciendo los valores en los cuadros de texto inferior y superior. Por ejemplo, un valor de 95 en el cuadro de texto superior representa el percentil 95, que es el valor por debajo del cual cae el 95% de las observaciones. Del mismo modo, un valor de 5 en el cuadro de texto inferior representa el percentil 5, que es el valor por debajo del cual cae el 5% de las observaciones.
- **Posiciones personalizadas.** Puede añadir líneas de referencia a los valores especificados en el eje horizontal.

Líneas de referencia de etiquetas. Esta opción contrla si las etiquetas se aplican a las líneas de referencia seleccionadas.

Las líneas de referencia se eliminan anulando la selección asociada en el diálogo Opciones de gráfico y pulsando **Continuar**.

Capítulo 35. Modelado geoespacial

Las técnicas de modelado geoespacial están diseñadas para descubrir patrones de datos que incluyen un componente geoespacial (mapa). El sistema de modelado geoespacial proporciona métodos para analizar datos geoespaciales con y sin un componente de tiempo.

Busque asociaciones basadas en datos de evento y geoespaciales (reglas de asociación geoespacial)
Mediante reglas de asociación geoespacial, puede encontrar patrones en datos que se basan en las propiedades espaciales y no espaciales. Por ejemplo, podría identificar patrones en datos de delincuencia por ubicación y atributos demográficos. A partir de estos patrones, podrá generar reglas que predicen dónde es más probable que se vayan a producir determinados tipos de delitos.

Realice predicciones utilizando series temporales y datos geoespaciales (predicción espacio-temporal)
La predicción espacio-temporal utiliza datos que contienen datos de ubicación, campos de entrada para la predicción (predictores), uno o varios campos de hora y un campo de objetivo. Cada ubicación tiene muchas filas en los datos que representan los valores de cada predictor y el objetivo en cada intervalo de tiempo.

Utilización del asistente de modelado geoespacial

1. Desde los menús, elija:

   Analizar > **Modelado espacial y temporal** > **Modelado espacial**
2. Siga los pasos del asistente.

Ejemplos

Hay ejemplos detallados disponibles en el sistema de ayuda.

- Reglas de asociación geoespacial: **Ayuda** > **Temas** > **Estudios de casos** > **Statistics Base** > **Reglas de asociación espacial**
- Predicción temporal espacial: **Ayuda** > **Temas** > **Estudios de casos** > **Statistics Base** > **Predicción temporal espacial**

Selección de mapas

El modelado geoespacial puede utilizar uno o más orígenes de datos de mapas. Los orígenes de datos de mapas contienen información que define áreas geográficas y otras características geográficas como, por ejemplo, carreteras o ríos. Muchos orígenes de mapas también contienen datos demográficos u otros datos descriptivos y datos de evento como, por ejemplo, informes de delitos o tasas de desempleo. Puede utilizar un archivo de especificación de mapa definido previamente o definir especificaciones de mapas aquí y guardar estas especificaciones para su uso posterior.

Cargar una especificación de mapa
Carga un archivo (.mplan) de especificación de mapa definido previamente. Los orígenes de datos de mapa que defina aquí se pueden guardar en un archivo de especificación de mapa. Para la predicción temporal espacial, si selecciona un archivo de especificación de mapa que identifica más de un mapa, se le solicitará que seleccione un mapa del archivo.

Añadir archivo de mapa
Añada un archivo de forma ESRI (.shp) o un archivo .zip que contenga un archivo de forma ESRI.

- Debe haber un archivo .dbf correspondiente en la misma ubicación que el archivo .shp, y dicho archivo debe tener el mismo nombre raíz que el archivo .shp.

- Si el archivo es un archivo .zip, los archivos .shp y .dbf deben tener el mismo nombre raíz que el archivo .zip.
- Si no hay ningún archivo (.prj) de proyección correspondiente, se le solicitará que seleccione un sistema de proyección.

Relación
Para las reglas de asociación geoespacial, esta columna define cómo están relacionados los eventos con las características en el mapa. Este valor no está disponible para la predicción temporal espacial.

Subir, Bajar
El orden de la capa de los elementos del mapa se determina mediante el orden en el cual aparecen en la lista. El primer mapa de la lista es la capa inferior.

Selección de un mapa

Para la predicción temporal espacial, si selecciona un archivo de especificación de mapa que identifica más de un mapa, se le solicitará que seleccione un mapa del archivo. La predicción temporal espacial no soporta varios mapas.

Relación geoespacial

Para las reglas de asociación geoespacial, el diálogo Relación geoespacial define cómo se relacionan los eventos con las características en el mapa.

- Este ajuste se aplica sólo a las reglas de asociación geoespacial.
- Este valor sólo afecta a orígenes de datos asociados con mapas especificados como datos de contexto en el paso de selección de orígenes de datos.

Relación

Cerrar El evento se produce cerca de una área o punto especificados en el mapa.

Dentro
El evento se produce dentro de un área especificada en el mapa.

Contiene
El área de evento contiene un objeto de contexto de mapa.

Intersecta
Las ubicaciones donde las líneas o regiones de los distintos mapas interseccionan entre sí.

Cruz Para varios mapas, las ubicaciones donde las líneas (para carreteras, ríos, ferrocarriles) de distintas líneas se cruzan entre sí.

Norte de, Sur de, Este de, Oeste de
El evento se produce dentro de un área que está al norte, sur, este o oeste de un punto especificado en el mapa.

Establecer sistema de coordenadas

Si no hay ningún archivo de proyección (.prj) con el mapa o define dos campos en un origen de datos como un conjunto de coordenadas, debe definir el sistema de coordenadas.

Valor geográfico predeterminado (longitud y latitud)
El sistema de coordenadas es de longitud y latitud.

Cartesiano simple (X e Y)
El sistema de coordenadas es de coordenadas simples X e Y.

Utilizar un ID conocido (WKID)
Un "ID conocido" para proyecciones comunes.

Utilizar un nombre de sistema de coordenadas
El sistema de coordenadas se basa en la proyección denominada. El nombre se especifica entre paréntesis.

Definición de la proyección

Si no se puede determinar el sistema de proyección a partir de la información proporcionada con el mapa, debe especificar el sistema de proyección. La causa más habitual de esta condición es que falte un archivo de proyecto (.prj) asociado con el mapa o exista un archivo de proyección que no se pueda utilizar.

- **Una ciudad, región o país (Mercator)**
- **Un país grande, varios países o continentes (Winkel Tripel)**
- **Un área muy cercana al ecuador (Mercator)**
- **Un área cercana a uno de los polos (Stereographic)**

La proyección de Mercator es una proyección común utilizada en muchos mapas. Esta proyección trata el globo como un cilindro desplegado en una superficie plana. La proyección de Mercator distorsiona el tamaño y forma de los objetos grandes. Esta distorsión aumenta a medida que se aleja del ecuador y se acerca a los polos. Las proyecciones de Winkel Tripel y Stereographic realizan ajustes al hecho de que un mapa representa una parte de una esfera tridimensional que se visualiza en dos dimensiones.

Sistema de proyección y de coordenadas

Si selecciona más de una correlación y las correlaciones tienen distintos sistemas de proyección y coordenadas, debe seleccionar la correlación con el sistema de proyección que desea utilizar. Dicho sistema de proyección se utilizará para todas las correlaciones cuando se combinan juntos en el resultado.

Orígenes de datos

Un origen de datos puede ser un archivo dBase que se proporciona con el archivo de forma, un archivo de datos de IBM SPSS Statistics o un conjunto de datos abierto en la sesión actual.

Datos de contexto. Los datos de contexto identifican características en el mapa. Los datos de contexto también pueden contener campos que se pueden utilizar como entradas para el modelo. Para utilizar un archivo dBase (.dbf) de contexto asociado con un archivo de forma (.shp) de mapa, el archivo dBase de contexto debe estar en la misma ubicación que el archivo de forma y debe tener el mismo nombre raíz. Por ejemplo, si el archivo de forma es `geodata.shp`, el archivo dBase se debe llamar `geodata.dbf`

Datos de evento. Los datos de evento contienen información sobre eventos que se producen como, por ejemplo, delitos o accidentes. Esta opción está disponible solo para reglas de asociación geoespacial.

Densidad de puntos. El intervalo de tiempo y los datos de coordenada para las estimaciones de la densidad de kernel. Esta opción sólo está disponible para la predicción temporal espacial.

Añadir. Abre un diálogo para añadir orígenes de datos. Un origen de datos puede ser un archivo dBase que se proporciona con el archivo de forma, un archivo de datos de IBM SPSS Statistics o un conjunto de datos abierto en la sesión actual.

Asociar. Abre un diálogo para especificar los identificadores (coordenadas o claves) utilizadas para asociar datos con mapas. Cada origen de datos debe contener uno o más identificadores que asocian los datos con el mapa. Los archivos dBase que vienen con un archivo de forma normalmente contienen un campo que se utiliza automáticamente como el identificador predeterminado. Para otros orígenes de datos, debe especificar los campos que se utilizan como identificadores.

Validar clave. Abre un diálogo para validar la coincidencia de claves entre el mapa y el origen de datos.

Reglas de asociación geoespacial

- Al menos, un origen de datos debe ser un origen de datos de evento.
- Todos los orígenes de datos de evento deben utilizar el mismo formato de identificadores de asociación de mapa: coordenadas o valores de clave.
- Si los orígenes de datos de evento se asocian a los mapas con valores de clave, todos los orígenes deben utilizar el mismo tipo de característica (por ejemplo, polígonos, puntos, líneas).

Predicción temporal espacial

- Debe haber un origen de datos de contexto.
- Si solo hay un origen de datos (un archivo de datos sin ningún mapa asociado), debe incluir valores de coordenadas.
- Si tiene dos orígenes de datos, un origen de datos debe ser datos de contexto y el otro origen de datos debe ser datos de densidad de puntos.
- No puede incluir más de dos orígenes de datos.

Añadir un origen de datos

Un origen de datos puede ser un archivo dBase que se proporciona con el archivo de forma y el archivo de contexto, un archivo de datos de IBM SPSS Statistics o un conjunto de datos abierto en la sesión actual.

Puede añadir el mismo origen de datos varias veces si desea utilizar una asociación espacial diferente con cada uno.

Datos y asociación de mapas

Cada origen de datos debe contener uno o más identificadores que asocian los datos con el mapa.

Coordenadas
: El origen de datos contiene campos que representan coordenadas cartesianas, seleccione los campos que representan dichas coordenadas X e Y. Para las reglas de asociación geoespacial, puede haber también una coordenada Z.

Valores de clave
: Los valores de clave en los campos del origen de datos corresponden a las claves de mapa seleccionadas. Por ejemplo, un mapa de regiones puede tener un identificador de nombre (clave de mapa) que etiqueta cada región. Dicho identificador se corresponde con un campo en los datos que también contienen los nombres de las regiones (clave de datos). Los campos se comparan con las claves de mapa basándose en el orden en el que se visualizan en las dos listas.

Validar claves

El diálogo Validar claves proporciona un resumen de coincidencias de registro entre la correlación y la fuente de datos, basándose en las claves de identificador seleccionadas. Si hay valores de clave de datos que no coinciden, puede hacer manualmente que coincidan con los valores de clave de correlación.

Reglas asociación geoespacial

Para las reglas de asociación geoespacial, tras definir mapas y orígenes de datos, los pasos restantes en el asistente son:

- Si hay varios orígenes de datos de evento, defina cómo se fusionan orígenes de datos de evento.
- Seleccione campos para utilizar como condiciones y predicciones en el análisis.

Si lo desea, tiene la posibilidad de:

- Seleccione distintas opciones de resultado.
- Guarde un archivo de modelo de puntuación.

- Cree campos nuevos para valores pronosticados y las reglas en los orígenes de datos utilizadas en el modelo.
- Personalice valores para crear regla de asociación.
- Personalice los valores de agrupación y agregación.

Definir campos de datos de eventos

Para las reglas de asociación geoespacial, si hay más de un origen de datos de evento, los orígenes de datos de evento se fusionan.

- De forma predeterminada, solo se incluyen los campos que son comunes para todos los orígenes de datos de evento.
- Puede visualizar una lista de campos comunes, campos para un origen de datos específico o campos de todos los orígenes de datos y seleccionar los campos que desea incluir.
- Para los campos comunes, los campos **Tipo** y **Medición** deben ser iguales para todos los orígenes de datos. Si hay conflictos, puede especificar el tipo y el nivel de medición para utilizar para cada campo común.

Seleccionar campos

La lista de campos disponibles incluye campos de los orígenes de datos de eventos y campos de los orígenes de datos de contexto.

- Puede controlar la lista de campos visualizados seleccionando un origen de datos de la lista **Orígenes de datos**.
- Debe seleccionar al menos dos campos. Al menos uno debe ser una condición y una debe ser, al menos, una predicción. Existen varias maneras de cumplir este requisito, incluyendo seleccionar dos campos para la lista **Ambas (condición y predicción)**.
- Las reglas de asociación predicen valores de los campos de predicción que se basan en valores de los campos de condición. Por ejemplo, en la regla "Si x=1 e y=2, entonces z=3", los valores de x e y son condiciones y el valor de z es la predicción.

Resultado

Tablas de reglas
: Cada tabla de reglas muestra las reglas superiores y los valores para la confianza, el soporte para regla, la elevación, el soporte de condición y la capacidad de despliegue. Cada tabla se clasifica por los valores del criterio seleccionado. Puede mostrar todas las reglas o el **Número** superior de reglas, basándose en el criterio seleccionado.

Nube de palabras clasificables
: Una lista de las reglas superiores, basándose en los valores del criterio seleccionado. El tamaño del texto indica la importancia relativa de la regla. El objeto de resultados interactivos contiene las reglas superiores para la confianza, el soporte de regla, la elevación, el soporte de condición y la capacidad de despliegue. El criterio seleccionado determina qué lista de reglas se visualiza de forma predeterminada. Puede seleccionar un criterio diferente de forma interactiva en el resultado. **Máx. de reglas que se va a mostrar** determina el número de reglas que se visualizan en el resultado.

Mapa
: Gráfico de barras interactivo y mapa de las reglas superiores, basándose en el criterio seleccionado. Cada objeto de resultados interactivos contiene las reglas superiores para la confianza, el soporte de regla, la elevación, el soporte de condición y la capacidad de despliegue. El criterio seleccionado determina qué lista de reglas se visualiza de forma predeterminada. Puede seleccionar un criterio diferente de forma interactiva en el resultado. **Máx. de reglas que se va a mostrar** determina el número de reglas que se visualizan en el resultado.

Tablas de información de modelos

Transformaciones de campo
Describe las transformaciones que se aplican a campos utilizados en el análisis.

Resumen de registros
Número y porcentaje de registros incluidos y excluidos

Estadísticos de regla
Estadísticas de resumen para el soporte de condición, la confianza, el soporte de regla, la elevación y la capacidad de despliegue. Las estadísticas incluyen media, mínimo, máximo y desviación estándar.

Elementos más frecuentes
Los elementos que aparecen con más frecuencia. Un elemento se incluye en una condición o una predicción en una regla. Por ejemplo, edad < 18 o sexo=mujer.

Campos más frecuentes
Los campos que aparecen con más frecuencia en las reglas.

Entradas excluidas
Los campos que se excluyen del análisis y la razón por la cual se ha excluido cada campo.

Criterio para tablas de reglas, nube de palabras y mapas

Confianza.
EL porcentaje de las predicciones de reglas correctas.

Soporte de regla
El porcentaje de casos para los cuales la regla es verdadera. Por ejemplo, si la regla es "si x=1 e y=2, z=3," el soporte de la regla es el porcentaje real de casos en los datos para los cuales es correcto x=1, y=2 y z=3.

Elevación
La elevación es una medida que indica el grado de mejora que aplica la regla a la predicción en comparación con una predicción al azar. Es la proporción de predicciones correctas en la aparición global del valor pronosticado. El valor debe ser mayor que 1. Por ejemplo, si el valor pronosticado se produce un 20 % del tiempo y la confianza en la predicción es del 80 %, el valor de elevación es 4.

Soporte de condición
El porcentaje de casos para los cuales existe la condición de la regla. Por ejemplo, si la regla es "si x=1 e y=2, z=3," el soporte de condición es la proporción de casos en los cuales los datos son x=1 e y=2.

Capacidad de despliegue
El porcentaje de predicciones incorrectas cuando las condiciones son verdaderas. La capacidad de despliegue es igual a (1-confianza) multiplicado por el soporte de condición o soporte de condición menos soporte de regla.

Guardar

Guarde el mapa y los datos de contexto como una especificación de mapa
Guarde las especificaciones del mapa en un archivo externo(.mplan). Puede cargar este archivo de especificación de mapa en el asistente para análisis posteriores. También puede utilizar el archivo de especificación de mapa con el comando `SPATIAL ASSOCIATION RULES`.

Copie cualquier archivo de mapa y datos en la especificación
Los datos de los archivos de forma de mapa, archivos de datos externos y conjuntos de datos utilizados en la especificación de mapa se guardan en el archivo de especificación de mapa.

Puntuación

Guarda los mejores valores de regla, los valores de confianza para las reglas y los valores de ID numérico para las reglas como campos nuevos en el origen de datos especificado.

Origen de datos para puntuar

El origen u orígenes de datos donde se crean los campos nuevos. Si el origen de datos no está abierto en la sesión actual, se abre en la sesión actual. Debe guardar explícitamente el archivo modificado para guardar los campos nuevos.

Valores de objetivo

Cree campos nuevos para los campos de objetivo seleccionado (predicción).

- Se crean dos campos nuevos para cada campo de objetivo: valor pronosticado y valor de confianza.
- Para los campos de objetivos continuos (escala), el valor pronosticado es una cadena que describe un rango de valores. Un valor de la forma "(value1, value2]" significa "mayor que el value1 y menor o igual que value2."

Número de mejores reglas

Cree campos nuevos para el número de mejores reglas especificado. Se crean tres campos nuevos para cada regla: valor de regla, valor de confianza y un valor de ID numérico para la regla.

Prefijo de nombre

Prefijo para utilizar para los nombres de campo nuevos.

Creación de reglas

Los parámetros de creación de reglas definen los criterios para las reglas de asociación generadas.

Elementos por regla

Número de valores de campo que se pueden incluir en las condiciones y predicciones de regla. EL número total de elemento no puede superar el 10. Por ejemplo, en la regla "si x=1 e y=2, z=3", hay dos elementos de condición y un elemento de predicción.

Número máximo de predicciones

El número máximo de valores de campo que se pueden producir en las predicciones para una regla.

Número máximo de condiciones

El número máximo de valores de campo que se pueden producir en las condiciones para una regla.

Excluir para

Excluye los pares de campos especificados de incluirse en la misma regla.

Criterios de regla

Confianza.

La confianza mínima que debe tener una regla para incluirse en el resultado. La confianza es el porcentaje de predicciones correctas.

Soporte de regla

El soporte mínimo de regla que debe tener para que se incluya en el resultado. El valor representa el porcentaje de casos para los cuales la regla es verdadera en los datos observados. Por ejemplo, si la regla es "si x=1 e y=2, z=3," el soporte de la regla es el porcentaje real de casos en los datos para los cuales es correcto x=1, y=2 y z=3.

Soporte de condición

Soporte de condición mínima que debe tener una regla para que se incluya en el resultado. EL valor representa el porcentaje de casos para los cuales existe la condición. Por ejemplo, si la regla es "si x=1 e y=2, z=3," el soporte de condición es el porcentaje de casos en los datos para los cuales x=1 e y=2.

Elevación
La elevación mínima que debe tener una regla para que se incluya en el resultado. La elevación es una medida que indica el grado de mejora que aplica la regla a la predicción con una predicción al azar. Es la proporción de predicciones correctas en la aparición global del valor pronosticado. Por ejemplo, si el valor pronosticado se produce un 20 % del tiempo y la confianza en la predicción es del 80 %, el valor de elevación es 4.

Tratar como igual
Identifica pares de campos, que se deben tratar como el mismo campo.

Agrupación y agregación

- La agregación es necesaria cuando hay más registros en los datos que características en el mapa. Por ejemplo, tiene registros de datos para países individuales pero tiene un mapa de estados.
- Puede especificar el método de medida de resumen de agregado para campos continuos y ordinales. Los campos nominales se agregan en función del valor modal.

Continuo
Para campos continuos (escala), la medida del resumen puede ser media, mediana o suma.

Ordinal
Para campos ordinales, la medida de resumen puede ser mediana, moda, mayor o menor.

Número de intervalos
Define el número máximo de intervalos para campos continuos (escala). Los campos continuos siempre se agrupan en rangos de valores. Por ejemplo, menor o igual que 5, mayor que 5 y menor o igual que 10, o mayor que 10.

Agregar el mapa
Aplique la agregación a datos y mapas.

Valores personalizados para campos específicos
Puede alterar temporalmente la medida de resumen predeterminada y el número de intervalos para campos específicos.

- Pulse el icono para abrir el diálogo **Selector de campos** y seleccione un campo para añadir a la lista.
- En la columna **Agregación**, seleccione una medida de resumen.
- Para campos continuos, pulse el botón en la columna **Intervalos** para especificar un número personalizado de intervalos para el campo en el diálogo **Intervalos**.

Predicción temporal espacial

Para la predicción temporal espacial, tras definir mapas y orígenes de datos, los pasos restantes del asistente son:

- Especifique el campo objetivo, los campos de tiempo y predictores opcionales.
- Defina los intervalos de tiempo o periodos cíclicos para los campos de tiempo.

Si lo desea, tiene la posibilidad de:

- Seleccione distintas opciones de resultado.
- Personalice parámetros de generación de modelos.
- Personalice valores de agregación.
- Guarde los valores pronosticados en un conjunto de datos en la sesión actual o en un archivo de datos de formato de IBM SPSS Statistics.

Seleccionar campos

La lista de campos disponibles incluye campos de los orígenes de datos seleccionados. Puede controlar la lista de campos visualizados seleccionando un origen de datos en la lista **Orígenes de datos**.

Objetivo

Es necesario un campo objetivo. El objetivo es el campo para el cual se han pronosticado valores.

- El campo objetivo debe ser un campo número continuo (escala).
- Si hay dos orígenes de datos, el objetivo es la estimación de densidad del kernel y la "densidad" se visualiza en el nombre del objetivo. No puede cambiar esta selección.

Predictores

Se pueden especificar uno o más campos de predictor. Este valor es opcional.

Campos de tiempo

Debe seleccionar uno o más campos que representan periodos de tiempo o seleccionar **Periodos cíclicos**.

- Si hay dos orígenes de datos, debe seleccionar campos de tiempo de ambos orígenes de datos. Ambos campos de tiempo deben representar el mismo intervalo.
- Para los periodos cíclicos, debe especificar los campos que definen ciclos de periodicidad en el panel Intervalos de tiempo del asistente.

Intervalos de tiempo

Las opciones de este panel se basan en la selección de **Campos de hora** o **Periodo cíclico** en el paso de selección de campos.

Campos de hora

Campos de hora seleccionados. Si selecciona uno o más campos de hora en el paso para seleccionar campos, estos campos se visualizan en esta lista.

Intervalo de tiempo. Seleccione el intervalo de tiempo apropiado en la lista. En función del intervalo de tiempo, también puede especificar otros valores como, por ejemplo, el intervalo entre observaciones (incremento) y el valor de inicio. Este intervalo de tiempo se utiliza para todos los campos de hora seleccionados.

- El procedimiento presupone que todos los casos (registros) representan intervalos espaciados de forma uniforme.
- Basándose en el intervalo de tiempo seleccionado, el procedimiento puede detectar observaciones que faltan o varias observaciones en el mismo intervalo de tiempo que se deben agregar de forma conjunta. Por ejemplo, si el intervalo de tiempo es días y la fecha 2014-10-27 viene seguida por 2014-10-29, hay una observación que falta para el 2014-10-28. Si el intervalo de tiempo es mes, varias fechas del mismo mes se agregan juntas.
- Para algunos intervalos de tiempo, el valor adicional puede definir saltos en los intervalos normales con un espacio uniforme. Por ejemplo, si el intervalo de tiempo es días, pero solo son válidos los fines de semana, puede especificar que hay cinco días en una semana y que la semana empieza el lunes.
- Si el campo de hora seleccionado no es un campo de formato de fecha ni de hora, el intervalo de tiempo se define automáticamente en **Periodos** y no se puede cambiar.

Campos de ciclo

Si seleccione el **Periodo cíclico** en el paso para seleccionar campos, debe especificar los campos que definen los periodos cíclicos. Un periodo cíclico identifica una variación cíclica repetitiva como, por ejemplo, el número de meses de un año o el número de días de una semana.

- Puede especificar hasta tres campos que definen periodos cíclicos.

- El primer campo de ciclo representa el nivel más alto del ciclo. Por ejemplo, si hay una variación cíclica por año, trimestre y mes, el campo que representa el año es el primer campo de ciclo.
- La longitud del ciclo para el primer y segundo campo del ciclo es la periodicidad en el nivel siguiente. Por ejemplo, si los campos de ciclo son año, trimestre y mes, la longitud del primer ciclo es 4 y la longitud del segundo ciclo es 3.
- El valor de inicio para los campos del segundo y el tercer ciclo es el primer valor de cada uno de estos periodos cíclicos.
- Los valores de longitud del ciclo y de inicio deben ser enteros positivos.

Agregación

- Si selecciona algún **Predictor** en el paso para seleccionar campos, puede seleccionar el método de resumen de agregación para los predictores.
- La agregación es necesaria cuando hay más de un registro en un intervalo de tiempo definido. Por ejemplo, si el intervalo de tiempo es Mes, varias fechas del mismo mes se agregan juntas.
- Puede especificar el método de medida de resumen de agregación para los campos continuos y ordinales. Los campos nominales se agregan en función del valor modal.

Continuo
Para campos continuos (escala), la medida del resumen puede ser media, mediana o suma.

Ordinal
Para campos ordinales, la medida de resumen puede ser mediana, moda, mayor o menor.

Valores personalizados para campos específicos
Puede alterar temporalmente la medida de resumen de agregación predeterminada para predictores específicos.

- Pulse el icono para abrir el diálogo **Selector de campos** y seleccione un campo para añadir a la lista.
- En la columna **Agregación**, seleccione una medida de resumen.

Resultado

Mapas

Valores de destino
Mapa de valores para el campo objetivo seleccionado.

Correlación
Mapa de correlaciones.

Agrupaciones
Mapa que resalta los clústeres de ubicaciones que son similares entre sí. Hay mapas de clústeres disponibles sólo para modelos empíricos.

Umbral de similitud de ubicación
La similitud necesaria para crear clústeres. El valor debe ser un número mayor que cero y menor que 1.

Especifique el número máximo de clústeres
El número máximo de clústeres para visualizar.

Tablas de evaluación de modelos

Especificaciones de modelos
Resumen de especificaciones utilizadas para ejecutar el análisis, incluidos campos objetivo, de entrada y ubicación.

Resumen de información temporal
Identifica los campos de tiempo y los intervalos de tiempo utilizados en el modelo.

Pruebas de efectos en estructura media
El resultado incluye el valor de estadístico de prueba, grados de libertad y el nivel de significación para el modelo y cada efecto.

Estructura media de los coeficientes de modelo
El resultado incluye el valor coeficiente, error estándar, valor de estadísticas de prueba, nivel de significación e intervalos de confianza para cada término del modelo.

Coeficientes autorregresivos
El resultado incluye el valor de coeficiente, error estándar, valor de estadísticas de prueba, nivel de significación e intervalos de confianza para cada retardo.

Pruebas de covarianza espacial
Para los modelos paramétricos basados en variograma, se muestran los resultados de la prueba de bondad de ajuste para la estructura de covarianza espacial. Los resultados de la prueba pueden determinar si se debe modelar la estructura de covarianza espacial de forma paramétrica o utilizar un modelo no paramétrico.

Covarianza espacial paramétrica
Para los modelos paramétricos basados en variograma, se muestran las estimaciones de parámetro para la covarianza espacial paramétrica.

Opciones de modelo

Configuración del modelo

Incluir automáticamente una interceptación
Incluir la interceptación en el modelo.

Retardo máximo de autorregresión
El retardo máximo de autorregresión. El valor debe ser un entero entre 1 y 5.

Covarianza espacial
Especifica el método de estimación para la covarianza espacial.

Paramétrico
EL método de estimación es paramétrico. El método puede ser **Gauss**, **Exponencial** o **Potencia exponencial**. Para la potencia exponencial, puede especificar el valor **Potencia**.

No paramétrica
El método de estimación no es paramétrico.

Guardar

Guarde el mapa y los datos de contexto como una especificación de mapa
Guarde las especificaciones de mapa en un archivo externo (.mplan). Puede cargar este archivo de especificación de mapa en el asistente para un análisis posterior. También puede utilizar el archivo de especificación de mapa con el comando `SPATIAL TEMPORAL PREDICTION`.

Copie cualquier archivo de mapa y datos en la especificación
Los datos de archivos de forma de mapa, archivos de datos externos y conjuntos de datos utilizados en la especificación del mapa se guardan en el archivo de especificación de mapa.

Puntuación
Guarda valores pronosticados, varianza y los límites de confianza superior e inferior para el campo objetivo en el archivo de datos seleccionado.

- Puede guardar valores pronosticados en un conjunto de datos abierto en la sesión actual o en archivo de datos de formato de IBM SPSS Statistics.
- El archivo de datos no puede ser un origen de datos utilizado en el modelo.

- El archivo de datos debe contener todos los campos de tiempo y predictores utilizados en el modelo.
- Los valores de tiempo deben ser mayores que los valores de tiempo utilizados en el modelo.

Avanzado

Número máximo de casos con valores perdidos (%)
El porcentaje máximo de casos con valores perdidos.

Nivel de significación
El nivel de significación para determinar si un modelo paramétrico basado en variograma es apropiado. El valor debe ser mayor que 0 y menor que 1. El valor predeterminado es 0,05. El nivel de significación se utiliza en la prueba de bondad de ajuste para la estructura de covarianza espacial. La estadística de bondad de ajuste se utiliza para determinar un modelo paramétrico o no paramétrico.

Factor de incertidumbre (%)
El factor de incertidumbre es un valor de porcentaje que representa el crecimiento de la incertidumbre al realizar predicciones de futuro. Los límites superior e inferior de la incertidumbre de predicción aumentan en el porcentaje especificado para cada paso en el futuro.

Finalizar

En el último paso del asistente de modelado geoespacial, puede ejecutar el modelo o pegar la sintaxis del comando generada en una ventana de sintaxis. Puede modificar y guardar la sintaxis generada para su uso posterior.

Avisos

Esta información se ha desarrollado para productos y servicios ofrecidos en los EE.UU. Este material puede estar disponible en IBM en otros idiomas. Sin embargo, es posible que deba ser propietario de una copia del producto o de la versión del producto en dicho idioma para acceder a él.

Es posible que IBM no ofrezca los productos, servicios o características que se tratan en este documento en otros países. El representante local de IBM le puede informar sobre los productos y servicios que están actualmente disponibles en su localidad. Cualquier referencia a un producto, programa o servicio de IBM no pretende afirmar ni implicar que solamente se pueda utilizar ese producto, programa o servicio de IBM. En su lugar, se puede utilizar cualquier producto, programa o servicio funcionalmente equivalente que no infrinja los derechos de propiedad intelectual de IBM. Sin embargo, es responsabilidad del usuario evaluar y comprobar el funcionamiento de todo producto, programa o servicio que no sea de IBM.

IBM puede tener patentes o solicitudes de patente en tramitación que cubran la materia descrita en este documento. Este documento no le otorga ninguna licencia para estas patentes. Puede enviar preguntas acerca de las licencias, por escrito, a:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
EE.UU.

Para consultas sobre licencias relacionadas con información de doble byte (DBCS), póngase en contacto con el departamento de propiedad intelectual de IBM de su país o envíe sus consultas, por escrito, a:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokio 103-8510, Japón

INTERNATIONAL BUSINESS MACHINES CORPORATION PROPORCIONA ESTA PUBLICACIÓN "TAL CUAL", SIN GARANTÍAS DE NINGUNA CLASE, NI EXPLÍCITAS NI IMPLÍCITAS, INCLUYENDO, PERO SIN LIMITARSE A, LAS GARANTÍAS IMPLÍCITAS DE NO VULNERACIÓN, COMERCIALIZACIÓN O ADECUACIÓN A UN PROPÓSITO DETERMINADO. Algunas jurisdicciones no permiten la renuncia a las garantías explícitas o implícitas en determinadas transacciones; por lo tanto, es posible que esta declaración no sea aplicable a su caso.

Esta información puede incluir imprecisiones técnicas o errores tipográficos. Periódicamente, se efectúan cambios en la información aquí y estos cambios se incorporarán en nuevas ediciones de la publicación. IBM puede realizar en cualquier momento mejoras o cambios en los productos o programas descritos en esta publicación sin previo aviso.

Las referencias hechas en esta publicación a sitios web que no son de IBM se proporcionan sólo para la comodidad del usuario y no constituyen de modo alguno un aval de esos sitios web. La información de esos sitios web no forma parte de la información de este producto de IBM y la utilización de esos sitios web se realiza bajo la responsabilidad del usuario.

IBM puede utilizar o distribuir la información que se le proporcione del modo que considere adecuado sin incurrir por ello en ninguna obligación con el remitente.

Los titulares de licencias de este programa que deseen tener información sobre el mismo con el fin de permitir: (i) el intercambio de información entre programas creados independientemente y otros programas (incluido este) y (ii) el uso mutuo de la información que se ha intercambiado, deberán ponerse en contacto con:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
EE.UU.

Esta información estará disponible, bajo las condiciones adecuadas, incluyendo en algunos casos el pago de una cuota.

El programa bajo licencia que se describe en este documento y todo el material bajo licencia disponible los proporciona IBM bajo los términos de las Condiciones Generales de IBM, Acuerdo Internacional de Programas Bajo Licencia de IBM o cualquier acuerdo equivalente entre las partes.

Los ejemplos de datos de rendimiento y de clientes citados se presentan solamente a efectos ilustrativos. Los resultados reales de rendimiento pueden variar en función de las configuraciones específicas y condiciones de operación.

La información relacionada con productos no IBM se ha obtenido de los proveedores de esos productos, de sus anuncios publicados o de otras fuentes disponibles públicamente. IBM no ha probado esos productos y no puede confirmar la exactitud del rendimiento, la compatibilidad ni ninguna otra afirmación relacionada con productos no IBM. Las preguntas sobre las posibilidades de productos que no son de IBM deben dirigirse a los proveedores de esos productos.

Las declaraciones sobre el futuro rumbo o intención de IBM están sujetas a cambio o retirada sin previo aviso y representan únicamente metas y objetivos.

Esta información contiene ejemplos de datos e informes utilizados en operaciones comerciales diarias. Para ilustrarlos lo máximo posible, los ejemplos incluyen los nombres de las personas, empresas, marcas y productos. Todos estos nombres son ficticios y cualquier parecido con personas o empresas comerciales reales es pura coincidencia.

LICENCIA DE DERECHOS DE AUTOR:

Esta información contiene programas de aplicación de muestra escritos en lenguaje fuente, los cuales muestran técnicas de programación en diversas plataformas operativas. Puede copiar, modificar y distribuir estos programas de muestra de cualquier modo sin realizar ningún pago a IBM, con el fin de desarrollar, utilizar, comercializar o distribuir programas de aplicación que se ajusten a la interfaz de programación de aplicaciones para la plataforma operativa para la que se han escrito los programas de muestra. Estos ejemplos no se han probado exhaustivamente en todas las condiciones. Por lo tanto, IBM no puede garantizar ni dar por supuesta la fiabilidad, la capacidad de servicio ni la funcionalidad de estos programas. Los programas de muestra se proporcionan "TAL CUAL" sin garantía de ningún tipo. IBM no será responsable de ningún daño derivado del uso de los programas de muestra.

Cada copia o fragmento de estos programas de ejemplo o de cualquier trabajo derivado de ellos, debe incluir el siguiente aviso de copyright:

© (nombre de la compañía) (año). Algunas partes de este código procede de los programas de ejemplo de IBM Corp.

© Copyright IBM Corp. _especificar el año o años_. Reservados todos los derechos.

Marcas comerciales

IBM, el logotipo de IBM e ibm.com son marcas registradas o marcas comerciales de International Business Machines Corp., registradas en muchas jurisdicciones en todo el mundo. Otros nombres de productos y servicios podrían ser marcas registradas de IBM u otras compañías. En Internet hay disponible una lista actualizada de las marcas registradas de IBM, en "Copyright and trademark information", en www.ibm.com/legal/copytrade.shtml.

Adobe, el logotipo Adobe, PostScript y el logotipo PostScript son marcas registradas o marcas comerciales de Adobe Systems Incorporated en Estados Unidos y/o otros países.

Intel, el logotipo de Intel, Intel Inside, el logotipo de Intel Inside, Intel Centrino, el logotipo de Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium y Pentium son marcas comerciales o marcas registradas de Intel Corporation o sus filiales en Estados Unidos y otros países.

Linux es una marca registrada de Linus Torvalds en Estados Unidos, otros países o ambos.

Microsoft, Windows, Windows NT, y el logotipo de Windows son marcas comerciales de Microsoft Corporation en Estados Unidos, otros países o ambos.

UNIX es una marca registrada de The Open Group en Estados Unidos y otros países.

Java y todas las marcas comerciales y los logotipos basados en Java son marcas comerciales o registradas de Oracle y/o sus afiliados.

Índice

A

E

F

G

H

I

K

L

M

N

P

Q

R

S

T

U

V

W

Z

Impreso en España