IBM SPSS Complex Samples 19

Note: Before using this information and the product it supports, read the general information under Notices el p. 282.

Prefacio

IBM® SPSS® Statistics es un sistema global para el análisis de datos. El módulo adicional opcional Muestras complejas proporciona las técnicas de análisis adicionales que se describen en este manual. El módulo adicional Muestras complejas se debe utilizar con el sistema básico de SPSS Statistics y está completamente integrado en dicho sistema.

Acerca de SPSS Inc., an IBM Company

SPSS Inc., an IBM Company, es uno de los principales proveedores globales de software y soluciones de análisis predictivo. La gama completa de productos de la empresa (recopilación de datos, análisis estadístico, modelado y distribución) capta las actitudes y opiniones de las personas, predice los resultados de las interacciones futuras con los clientes y, a continuación, actúa basándose en esta información incorporando el análisis en los procesos comerciales. Las soluciones de SPSS Inc. tratan los objetivos comerciales interconectados en toda una organización centrándose en la convergencia del análisis, la arquitectura de TI y los procesos comerciales. Los clientes comerciales, gubernamentales y académicos de todo el mundo confían en la tecnología de SPSS Inc. como ventaja ante la competencia para atraer, retener y hacer crecer los clientes, reduciendo al mismo tiempo el fraude y mitigando los riesgos. SPSS Inc. fue adquirida por IBM en octubre de 2009. Para obtener más información, visite *http://www.spss.com*.

Asistencia técnica

El servicio de asistencia técnica está a disposición de todos los clientes de mantenimiento. Los clientes podrán ponerse en contacto con este servicio de asistencia técnica si desean recibir ayuda sobre la utilización de los productos de SPSS Inc. o sobre la instalación en alguno de los entornos de hardware admitidos. Para ponerse en contacto con el servicio de asistencia técnica, consulte el sitio web de SPSS Inc. en *http://support.spss.com* o encuentre a su representante local a través del sitio web *http://support.spss.com/default.asp?refpage=contactus.asp*. Tenga a mano su identificación, la de su organización y su contrato de asistencia cuando solicite ayuda.

Servicio de atención al cliente

Si tiene cualquier duda referente a la forma de envío o pago, póngase en contacto con su oficina local, que encontrará en el sitio Web en *http://www.spss.com/worldwide*. Recuerde tener preparado su número de serie para identificarse.

Cursos de preparación

SPSS Inc. ofrece cursos de preparación, tanto públicos como in situ. Todos los cursos incluyen talleres prácticos. Los cursos tendrán lugar periódicamente en las principales ciudades. Si desea obtener más información sobre estos cursos, póngase en contacto con su oficina local que encontrará en el sitio Web en *http://www.spss.com/worldwide*.

Publicaciones adicionales

Los documentos *SPSS Statistics: Guide to Data Analysis*, *SPSS Statistics: Statistical Procedures Companion* y *SPSS Statistics: Advanced Statistical Procedures Companion*, escritos por Marija Norušis y publicados por Prentice Hall, están disponibles y se recomiendan como material adicional. Estas publicaciones cubren los procedimientos estadísticos del módulo SPSS Statistics Base, el módulo Advanced Statistics y el módulo Regression. Tanto si da sus primeros pasos en el análisis de datos como si ya está preparado para las aplicaciones más avanzadas, estos libros le ayudarán a aprovechar al máximo las funciones ofrecidas por IBM® SPSS® Statistics. Si desea información adicional sobre el contenido de la publicación o muestras de capítulos, consulte el sitio web de la autora: *http://www.norusis.com*

Contenido

Parte II: Ejemplos

13 Asistente de muestreo de la opción Muestras complejas 98

Apéndices

Parte I:
Manual del usuario

Capítulo
1

Introducción a los procedimientos de muestras complejas

Un supuesto inherente a los procedimientos de análisis en los paquetes de software tradicionales es que las observaciones de un archivo de datos representan una muestra aleatoria simple de la población de interés. Este supuesto es insostenible para un número cada vez mayor de empresas e investigadores que consideran más económico y cómodo obtener las muestras de una forma más estructurada.

La opción Muestras complejas permite seleccionar una muestra de acuerdo con un diseño complejo e incorporar las especificaciones del diseño al análisis de los datos, asegurando así que los resultados serán válidos.

Propiedades de las muestras complejas

Una muestra compleja puede ser distinta de una muestra aleatoria simple en muchos aspectos. En una muestra aleatoria simple, las unidades de muestreo individuales se seleccionan aleatoriamente con la misma probabilidad y sin reposición (SR) directamente a partir de la totalidad de la población. Por lo contrario, una muestra compleja determinada puede tener alguna o todas las características siguientes:

Estratificación. El muestreo estratificado implica seleccionar muestras independientemente dentro de los subgrupos de la población que no se solapen o estratos. Por ejemplo, los estratos pueden ser grupos socioeconómicos, categorías laborales, grupos de edad o grupos étnicos. Con la estratificación, puede asegurar que los tamaños muestrales de los subgrupos de interés son adecuados, mejorar la precisión de las estimaciones globales y utilizar distintos métodos de muestreo entre los diferentes estratos.

Conglomerados. El muestreo por conglomerados implica la selección de grupos de unidades muestrales o conglomerados. Por ejemplo, los conglomerados pueden ser escuelas, hospitales o zonas geográficas y las unidades muestrales pueden ser alumnos, pacientes o ciudadanos. El conglomerado es común en los diseños polietápicos y en las muestras de zona (geográfica).

Múltiples etapas. En el muestreo polietápico, se selecciona una muestra de primera etapa basada en conglomerados. A continuación, se crea una muestra de segunda etapa extrayendo submuestras a partir de los conglomerados seleccionados. Si la muestra de segunda etapa está basada en subconglomerados, entonces puede añadir una tercera etapa a la muestra. Por ejemplo, en la primera etapa de una encuesta, se podría extraer una muestra de ciudades. A continuación, y a partir de las ciudades seleccionadas, se podrían muestrear unidades familiares. Finalmente, a partir de las unidades familiares seleccionadas, se podría encuestar a individuos. Los Asistentes de muestreo y preparación del análisis permiten especificar tres etapas en un diseño.

Muestreo no aleatorio. Cuando es difícil obtener la muestra aleatoriamente, las unidades se pueden muestrear sistemáticamente (con un intervalo fijo) o secuencialmente.

Probabilidades de selección desiguales. Cuando se muestrean conglomerados que contienen números de unidades desiguales, puede utilizar el muestreo probabilístico proporcional al tamaño (PPS) para que la probabilidad de selección del conglomerado sea igual a la proporción de unidades que contiene. El muestreo PPS también puede utilizar esquemas de ponderación más generales para seleccionar unidades.

Muestreo no restringido. El muestreo no restringido selecciona las unidades con reposición (CR). Por lo tanto, se puede seleccionar más de una vez una unidad individual para la muestra.

Ponderaciones muestrales. Las ponderaciones muestrales se calculan automáticamente al extraer una muestra compleja y de forma ideal se corresponden con la "frecuencia" que cada unidad muestral representa en la población objetivo. Por lo tanto, la suma de las ponderaciones muestrales debe estimar el tamaño de la población. Los procedimientos de análisis de muestras complejas requieren las ponderaciones muestrales para poder analizar correctamente una muestra compleja. Tenga en cuenta que estas ponderaciones se deben utilizar exclusivamente dentro de la opción Muestras complejas y no con otros procesos analíticos a través del procedimiento Ponderar casos, el cual trata las ponderaciones como réplicas de casos.

Uso de los procedimientos de Muestras complejas

El uso de los procedimientos de Muestras complejas depende de las necesidades específicas. Los tipos fundamentales de usuarios son aquéllos que:

- Planifican y llevan a cabo encuestas de acuerdo con diseños complejos, analizando posiblemente la muestra más tarde. La herramienta principal de los encuestadores es elAsistente de muestreo.
- Analiza archivos de datos muestrales obtenidos previamente según diseños complejos. Antes de utilizar los procedimientos de análisis de muestras complejas puede que deba utilizar el Asistente de preparación del análisis.

Independientemente del tipo de usuario que sea, debe proporcionar información del diseño a los procedimientos de Muestras complejas. Esta información está almacenada en un **archivo de plan** para volver a utilizarla con mayor facilidad.

Archivos de plan

Los archivos de plan contienen especificaciones de la muestra compleja. Existen dos tipos de archivos de plan:

Plan de muestreo. Las especificaciones dadas en el Asistente de muestreo definen un diseño muestral que se utiliza para extraer una muestra compleja. El archivo del plan de muestreo contiene esas especificaciones. El archivo del plan de muestreo también contiene un plan de análisis por defecto que utiliza métodos de estimación adecuados para el diseño muestral especificado.

Plan de análisis. Este archivo de plan contiene la información necesaria en los procedimientos de análisis de Muestras complejas para calcular correctamente las estimaciones de la varianza de una muestra compleja. El plan incluye la estructura de la muestra, los métodos de estimación de cada

etapa y las referencias para variables necesarias como por ejemplo, las ponderaciones muestrales. El Asistente de preparación del análisis permite crear y editar los planes de análisis.

Existen distintas ventajas al guardar las especificaciones en un archivo de plan, por ejemplo:

- Un encuestador puede especificar la primera etapa de un plan de muestreo de varias etapas y extraer en el momento las unidades de la primera etapa, reunir información sobre las unidades muestrales para la segunda etapa y a continuación, modificar el plan de muestreo para incluir la segunda etapa.
- Un analista que no tenga acceso al archivo del plan de muestreo puede especificar un plan de análisis y hacer referencia a ese plan en cada procedimiento de análisis de Muestras complejas.
- Un diseñador de muestras a gran escala de uso público puede publicar el archivo del plan de muestreo, lo que simplifica las instrucciones para el analista y evita que cada analista deba especificar sus propios planes de análisis.

Lecturas adicionales

Si desea obtener más información sobre las técnicas de muestreo, consulte los siguientes textos:

Cochran, W. G. 1977. *Sampling Techniques*, 3rd ed. Nueva York: John Wiley and Sons.

Kish, L. 1965. *Survey Sampling*. Nueva York: John Wiley and Sons.

Kish, L. 1987. *Statistical Design for Research*. Nueva York: John Wiley and Sons.

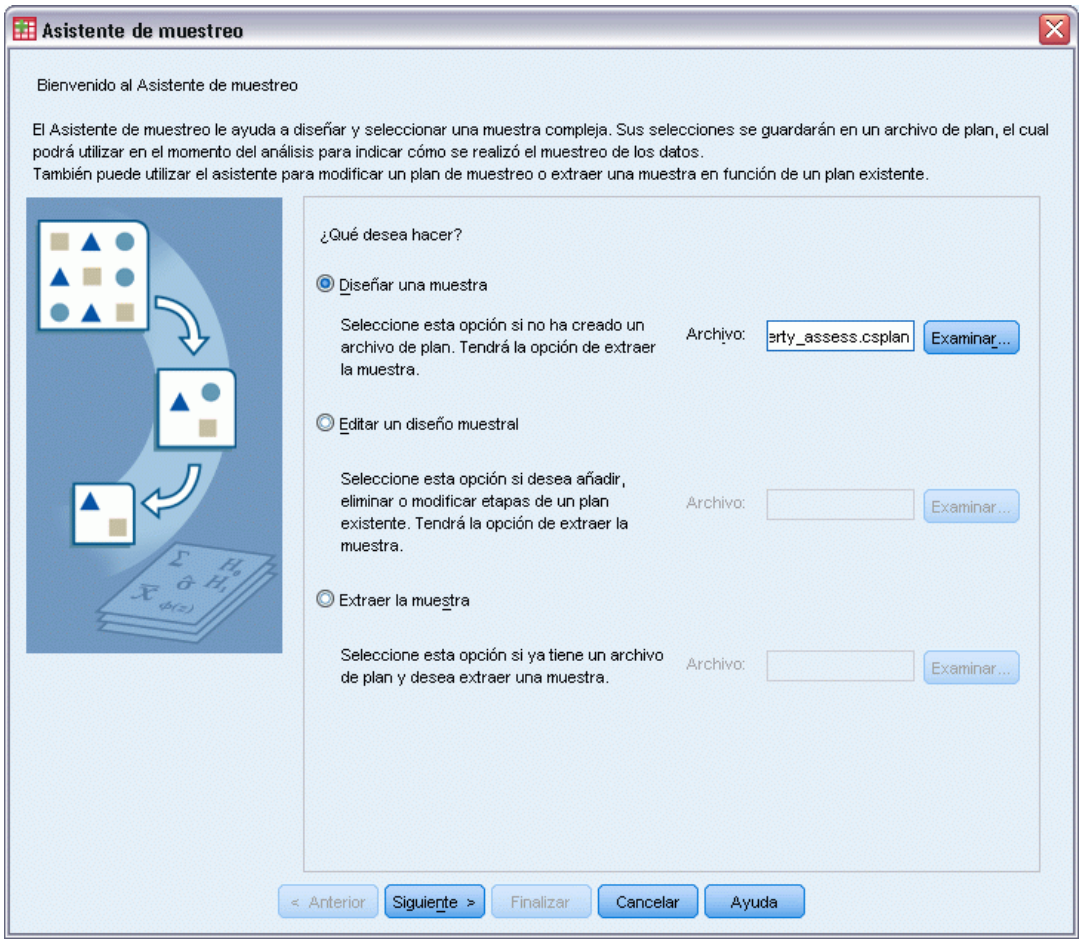
Murthy, M. N. 1967. *Sampling Theory and Methods*. Calcuta (India): Statistical Publishing Society.

Särndal, C., B. Swensson, y J. Wretman. 1992. *Model Assisted Survey Sampling*. Nueva York: Springer-Verlag.

Capítulo
2

Muestreo a partir de un diseño complejo

Figura 2-1
Asistente de muestreo: paso Bienvenida



El Asistente de muestreo le guía a través de los pasos necesarios para crear, modificar o ejecutar un archivo de plan de muestreo. Antes de utilizar el Asistente, debe tener en mente una población objetivo bien definida, una lista de las unidades muestrales y un diseño muestral adecuado.

Creación de un nuevo plan de muestreo

- En los menús, seleccione:
  Analizar > Complex Samples > Seleccionar una muestra...

- Seleccione Diseñar una muestra y elija un nombre de archivo de plan para guardar el plan de muestreo.

- Pulse Siguiente para continuar usando el Asistente.

- Si lo desea, en el paso Variables del diseño puede definir estratos, conglomerados e introducir ponderaciones muestrales. Después de definirlos, pulse Siguiente.

- Si lo desea, en el paso Método de muestreo, puede elegir un método para seleccionar los elementos.

  Si selecciona Muestreo de Brewer proporcional al tamaño o Muestreo de Murthy proporcional al tamaño, puede pulsar Finalizar para extraer la muestra. En caso contrario, pulse Siguiente y a continuación:

- En el paso Tamaño muestral, especifique el número o proporción de unidades que muestrear.

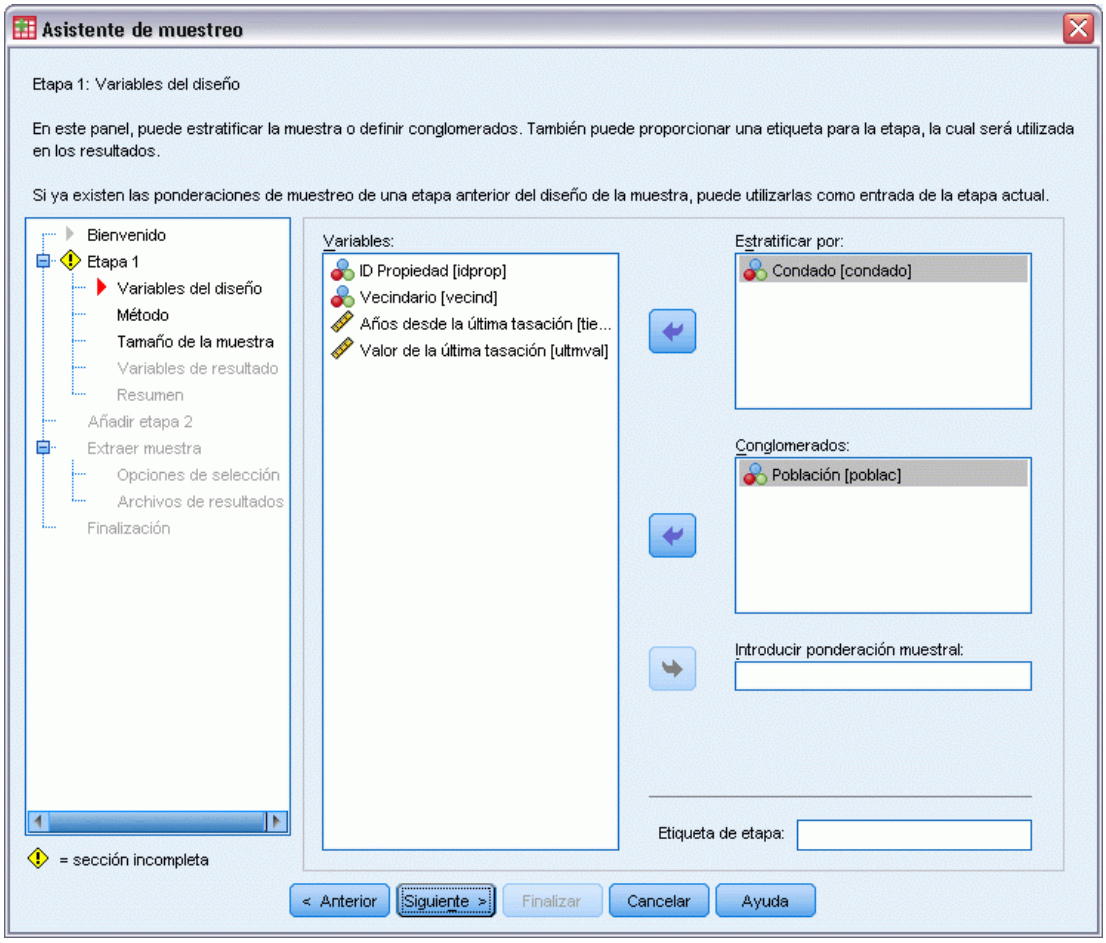
- Ahora puede pulsar Finalizar para extraer la muestra.

  Si lo desea, en los siguientes pasos puede:

  - Elegir las variables de resultado para guardar.
  - Añadir una segunda o tercera etapa al diseño.
  - Establecer varias opciones de selección, incluyendo las etapas a partir de las cuales se van a extraer las muestras, la semilla de aleatorización y si los valores perdidos definidos por el usuario se van a tratar como valores válidos de las variables del diseño.
  - Elegir dónde guardar los datos de resultado.
  - Pegar las selecciones como sintaxis de comandos.

Asistente de muestreo: Variables del diseño

Figura 2-2
Asistente de muestreo: paso Variables del diseño



Este paso permite seleccionar las variables de estratificación y conglomeración y definir unas ponderaciones muestrales de entrada. También puede especificar una etiqueta para la etapa.

Estratificar por. La clasificación conjunta por las variables de estratificación define distintas subpoblaciones o estratos. Se obtienen muestras individuales para cada estrato. Para mejorar la precisión de las estimaciones, las unidades de los estratos deben ser tan homogéneas como sea posible respecto a las características de interés.

Conglomerados. Las variables de conglomeración definen grupos de unidades de observación o conglomerados. Los conglomerados son útiles cuando es difícil o imposible realizar el muestreo de las unidades de observación directamente desde la población; en su lugar, se puede realizar el muestreo de los conglomerados a partir de la población y a continuación, realizar el muestreo de las unidades de observación a partir de los conglomerados seleccionados. Sin embargo, el uso de conglomerados puede introducir correlaciones entre las unidades muestrales, con la consiguiente pérdida de precisión. Para minimizar este efecto, las unidades de los conglomerados deben ser tan heterogéneas como sea posible respecto a las características de interés. Deberá definir

una variable de conglomeración como mínimo para planificar un diseño de varias etapas. Los conglomerados también son necesarios al utilizar distintos métodos de muestreo. Si desea obtener más información, consulte el tema Asistente de muestreo: Método de muestreo el p. 8.

Introducir ponderación muestral. Si el diseño muestral actual forma parte de un diseño muestral mayor, puede disponer de ponderaciones muestrales de una etapa anterior del diseño mayor. Puede especificar una variable numérica que contenga estas ponderaciones en la primera etapa del diseño actual. Las ponderaciones muestrales se calculan automáticamente para las etapas posteriores del diseño actual.

Etiqueta de etapa. Puede especificar una etiqueta de cadena opcional para cada etapa. Esto se utiliza en los resultados para facilitar la identificación de la información por etapas.

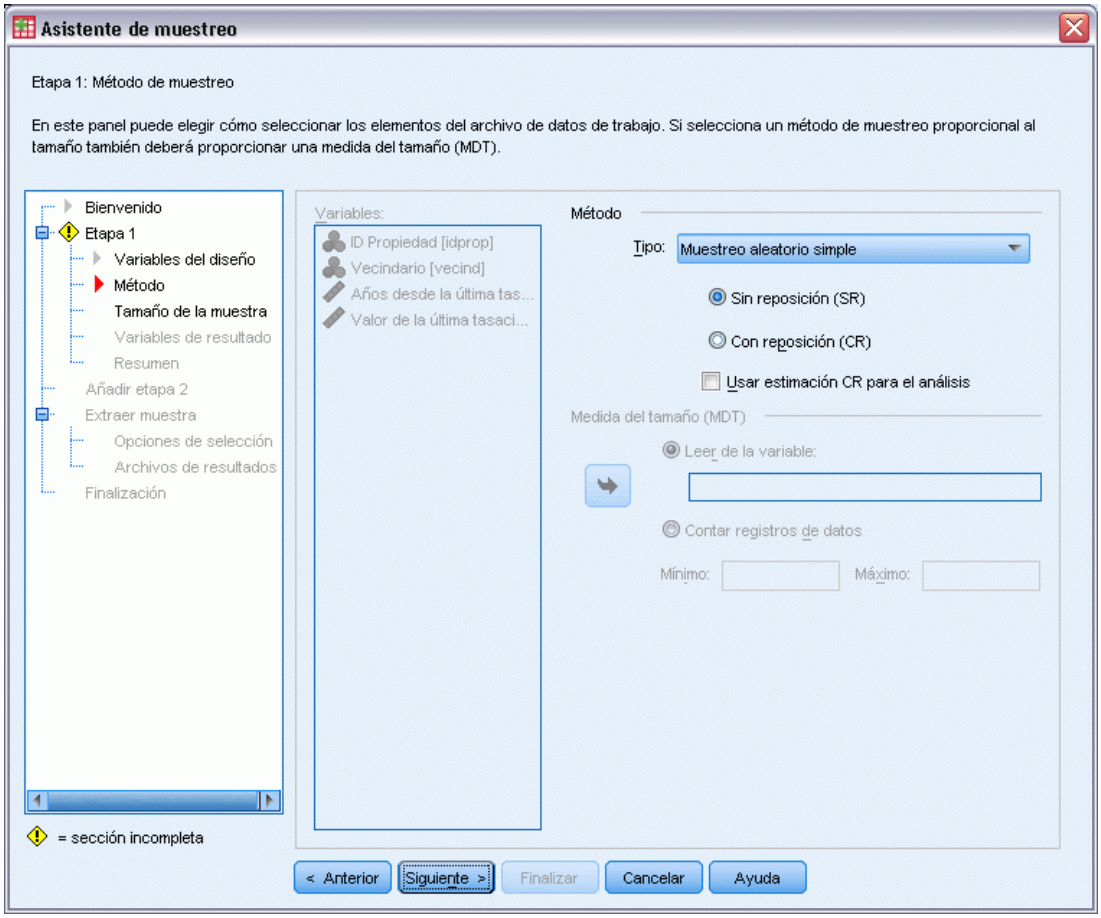
Nota: La lista de variables origen tiene el mismo contenido a lo largo de los pasos del Asistente. En otras palabras, las variables de la lista de origen eliminadas en un paso determinado se borran de la lista en todos los pasos. Las variables devueltas a la lista de origen aparecen en la lista en todos los pasos.

Controles de árbol para navegar por el Asistente de muestreo

En la parte izquierda de cada paso del Asistente de muestreo se muestra un esquema de los titulares de todos los pasos. Puede navegar por el Asistente al pulsar el nombre de uno de los pasos activados en el esquema. Los pasos están activados cuando todos los pasos anteriores sean válidos—, es decir, si cada uno de los pasos anteriores dispone de las especificaciones mínimas necesarias para ese paso. Consulte la ayuda de los pasos individuales para obtener más información sobre los motivos por los que un paso determinado puede no ser válido.

Asistente de muestreo: Método de muestreo

Figura 2-3
Asistente de muestreo: paso Método de muestreo



Este paso permite especificar cómo seleccionar los casos del conjunto de datos activo.

Método. Los controles de este grupo se utilizan para elegir un método de selección. Algunos tipos de muestreo permiten elegir entre realizar un muestreo con reposición (CR) o sin reposición (SR). Si desea obtener más información, consulte las descripciones de los tipos. Tenga en cuenta que algunos tipos de probabilidad proporcional al tamaño (PPS) están disponibles sólo cuando se han definido conglomerados y todos los tipos de PPS están disponibles sólo en la primera etapa de un diseño. Además, los métodos SR están disponibles sólo en la última etapa de un diseño.

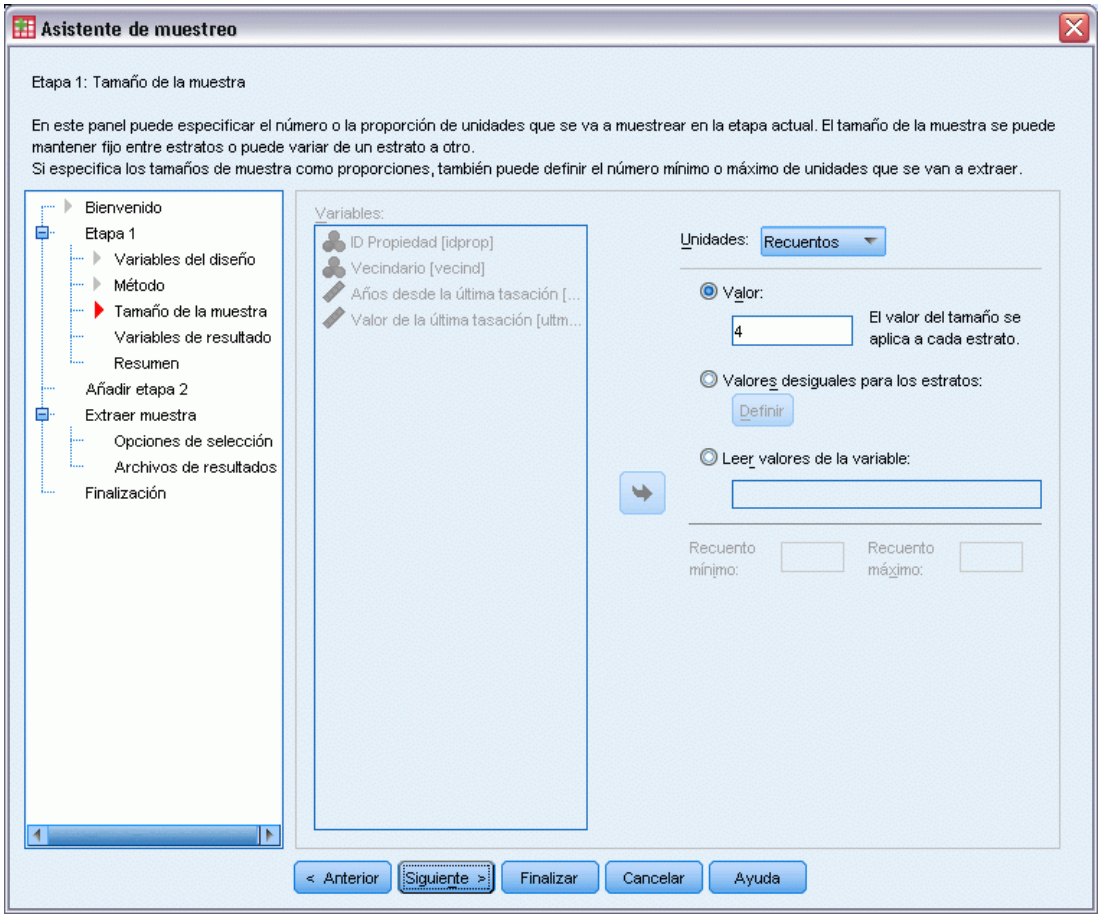
- **Muestreo aleatorio simple:** Las unidades se seleccionan con probabilidad igual. Se pueden seleccionar con o sin reposición.
- **Sistemático simple.** Las unidades se seleccionan con un intervalo fijo en todo el marco muestral (o en los estratos, si se han especificado) y se extraen sin reposición. Se selecciona una unidad aleatoriamente dentro del primer intervalo como el punto inicial.
- **Secuencial simple.** Las unidades se seleccionan de forma secuencial con probabilidad igual y sin reposición.

- **Probabilidad proporcional al tamaño.** Método de primera etapa que selecciona unidades de forma aleatoria con probabilidad proporcional al tamaño. Se puede seleccionar cualquier unidad con reposición; sólo se puede realizar muestreo sin reposición de los conglomerados.
- **Muestreo sistemático proporcional al tamaño.** Método de primera etapa que selecciona unidades de forma sistemática con probabilidad proporcional al tamaño. Se seleccionan sin reposición.
- **Muestreo secuencial proporcional al tamaño.** Método de primera etapa que selecciona unidades de forma secuencial con probabilidad proporcional al tamaño del conglomerado y sin reposición.
- **Muestreo de Brewer proporcional al tamaño.** Método de primera etapa que selecciona dos conglomerados de cada estrato con probabilidad proporcional al tamaño del conglomerado y sin reposición. Se debe especificar una variable de conglomeración para utilizar este método.
- **Muestreo de Murthy proporcional al tamaño.** Método de primera etapa que selecciona dos conglomerados de cada estrato con probabilidad proporcional al tamaño del conglomerado y sin reposición. Se debe especificar una variable de conglomeración para utilizar este método.
- **Muestreo de Sampford proporcional al tamaño.** Método de primera etapa que selecciona más de dos conglomerados de cada estrato con probabilidad proporcional al tamaño del conglomerado y sin reposición. Es una extensión del método de Brewer. Se debe especificar una variable de conglomeración para utilizar este método.
- **Usar estimación CR para el análisis.** Por defecto, el método de estimación se especifica en el archivo de plan de manera coherente con el método de muestreo seleccionado. Esta opción permite utilizar la estimación con reposición incluso si el método de muestreo implica la estimación SR. Esta opción solamente está disponible en la etapa 1.

Medida del tamaño (MDT). Si se selecciona un método PPS, deberá especificar una medida del tamaño que defina el tamaño de cada unidad. Estos tamaños pueden definirse explícitamente en una variable o se pueden calcular a partir de los datos. Opcionalmente, se pueden establecer los límites inferior y superior de la MDT, anulando cualquier valor encontrado en la variable MDT o calculado a partir de los datos. Estas opciones solamente están disponibles en la etapa 1.

Asistente de muestreo: Tamaño muestral

Figura 2-4
Asistente de muestreo: paso Tamaño muestral



Este paso permite especificar el número o la proporción de unidades que se van a muestrear dentro de la etapa actual. El tamaño muestral puede ser fijo o variar entre estratos. Para el propósito de especificar el tamaño muestral, se pueden utilizar los conglomerados elegidos en etapas anteriores para definir estratos.

Unidades. Puede especificar un tamaño muestral exacto o una proporción de unidades a muestrear.

- **Valor.** Se aplica un valor particular a todos los estratos. Si se selecciona Recuentos como la unidad métrica, deberá introducir un entero positivo. Si se selecciona Proporciones, deberá introducir un valor no negativo. A no ser que se realice una muestra con reposición, los valores de proporción no deberán ser mayores que 1.
- **Valores desiguales para estratos.** Permite introducir distintos valores de tamaño para cada estrato a través del cuadro de diálogo Definir tamaños desiguales.
- **Leer valores de la variable.** Permite seleccionar una variable numérica que contenga los valores de tamaño para los estratos.

Si se selecciona Proporciones, tiene la opción de establecer los límites inferior y superior para el número de unidades muestreadas.

Definir tamaños desiguales

Figura 2-5
Cuadro de diálogo Definir tamaños desiguales

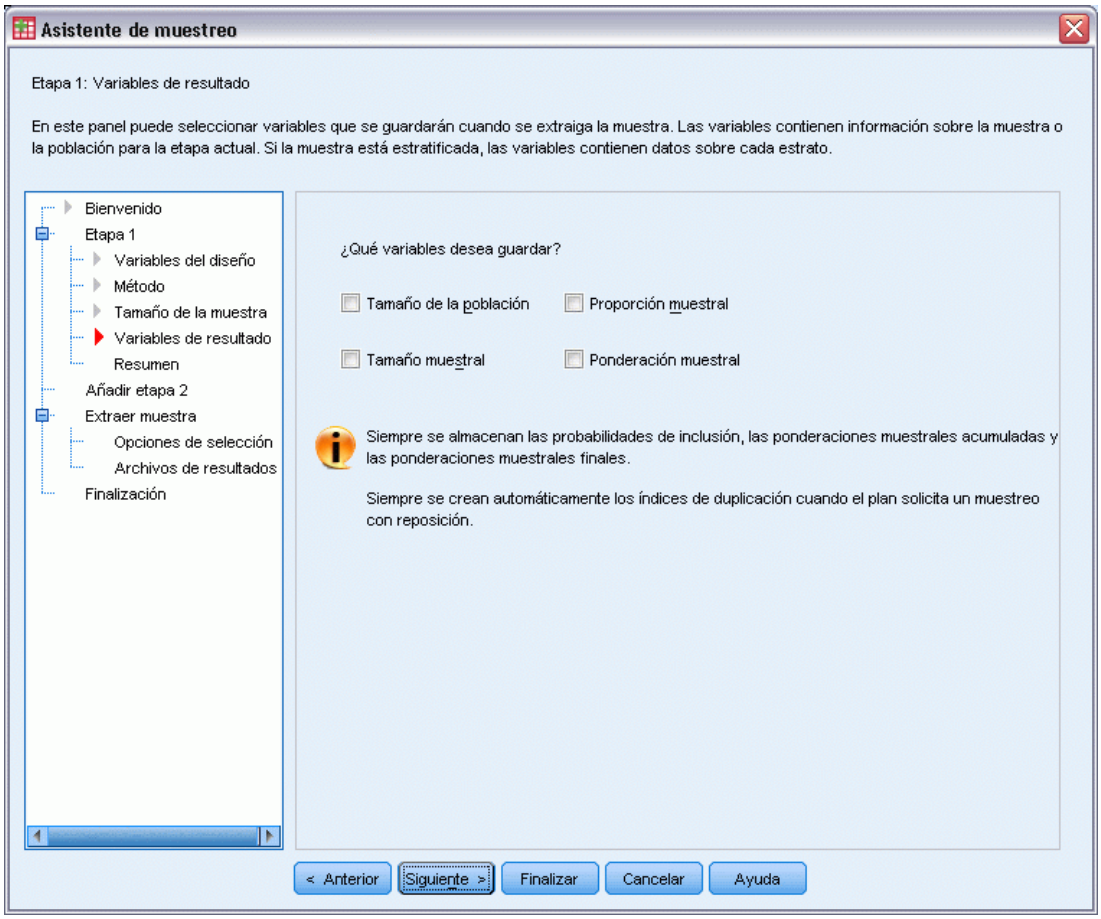
El cuadro de diálogo Definir tamaños desiguales permite introducir los tamaños para cada estrato.

Rejilla de especificaciones de tamaño. La rejilla muestra la clasificación conjunta de hasta cinco variables de conglomeración o estrato, con una combinación de estrato/conglomerado por fila. Las variables elegibles en la rejilla serán todas las variables de estratificación de las etapas anteriores y actuales además de todas las variables de conglomeración de las etapas anteriores. Las variables se pueden reordenar dentro de la rejilla o ser desplazadas a la lista Excluir. Introduzca los tamaños en la última columna de la derecha. Pulse en Etiquetas o Valores para conmutar entre la visualización de las etiquetas de valor y los valores de los datos para las variables de estratificación y de conglomeración de las casillas de la rejilla. Las casillas que contienen valores sin etiquetas siempre muestran valores. Pulse Actualizar estratos para volver a rellenar la rejilla con cada combinación de los valores de los datos etiquetados para las variables de la rejilla.

Excluir. Para especificar los tamaños de un subconjunto de combinaciones de estrato/conglomerado, desplace una o más variables a la lista Excluir. Estas variables no se utilizan para definir tamaños muestrales.

Asistente de muestreo: Variables de resultado

Figura 2-6
Asistente de muestreo: paso Variables de resultado



Este paso permite elegir las variables que desea guardar cuando se extraiga la muestra.

Tamaño poblacional. El número estimado de unidades en la población de una etapa dada. El nombre raíz de la variable guardada es *TamañoPoblación_*.

Proporción muestral. Tasa de la muestra en una etapa dada. El nombre raíz de la variable guardada es *TasaMuestreo_*.

Tamaño muestral. Número de unidades extraídas en una etapa dada. El nombre raíz de la variable guardada es *TamañoMuestral_*.

Ponderación muestral. La inversa de las probabilidades de inclusión. El nombre raíz de la variable guardada es *PonderaciónMuestral_*.

Algunas variables por etapa se generan automáticamente. Entre éstos se incluyen:

Probabilidades de inclusión. Proporción de unidades extraídas en una etapa dada. El nombre raíz de la variable guardada es *ProbabilidadInclusión_*.

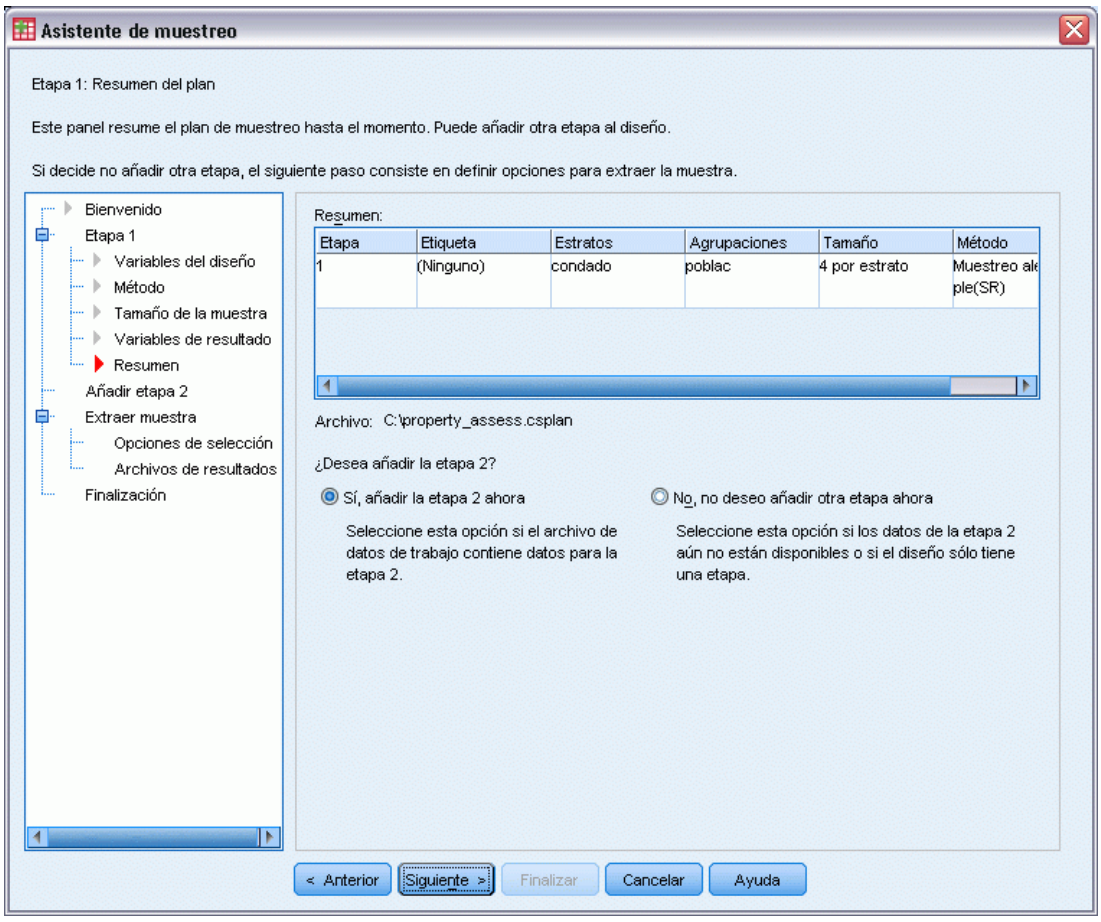
Ponderación acumulada. Ponderación de la muestra acumulada a lo largo de las etapas anteriores a la actual e incluyendo esta última. El nombre raíz de la variable guardada es *PonderaciónMuestralAcumulada_*.

Índice. Identifica las unidades seleccionadas varias veces dentro de una etapa dada. El nombre raíz de la variable guardada es *Índice_*.

Nota: Los nombres raíz de la variable guardada incluyen un sufijo entero que refleja el número de la etapa, por ejemplo, *TamañoPoblación_1_* para el tamaño de la población guardada de la etapa 1.

Asistente de muestreo: Resumen del plan

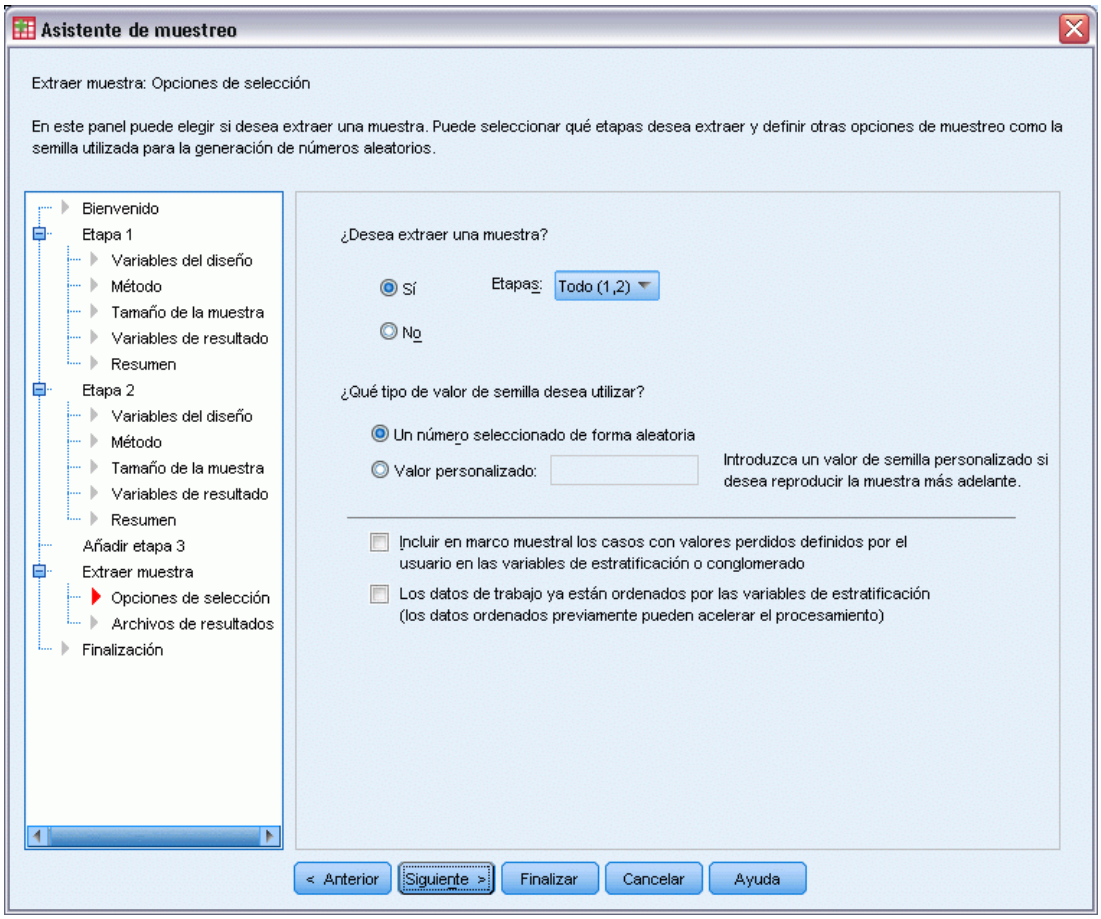
Figura 2-7
Asistente de muestreo: paso Resumen del plan



Último paso de cada etapa que proporciona un resumen de las especificaciones del diseño muestral hasta la etapa actual. A partir de aquí, puede pasar a la siguiente etapa (creándola si es necesario) o definir las opciones para extraer la muestra.

Asistente de muestreo: Extraer muestra: Opciones de selección

Figura 2-8
Asistente de muestreo: Extraer muestra: paso Opciones de selección



Este paso permite elegir si desea extraer una muestra. También puede controlar otras opciones del muestreo, como la semilla aleatoria y el tratamiento de los valores perdidos.

Extraer muestra. Además de elegir si desea extraer una muestra, también puede elegir ejecutar parte del diseño muestral. Las etapas se deben extraer en orden (es decir, la etapa 2 no se puede extraer a menos que ya se haya extraído la etapa 1). Al editar o ejecutar un plan, no puede volver a muestrear etapas bloqueadas.

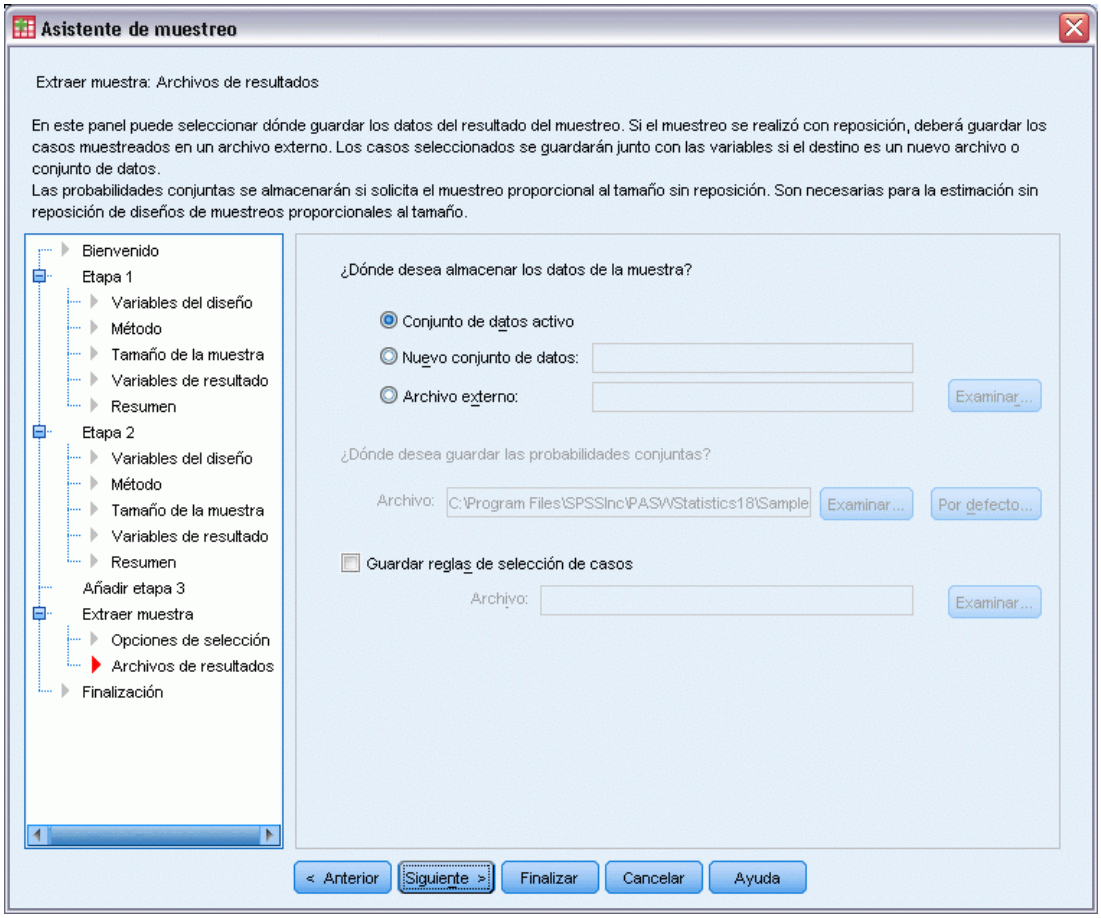
Semilla. Permite elegir un valor de semilla para la generación de números aleatorios.

Incluye los valores perdidos definidos por el usuario. Determina si los valores perdidos definidos por el usuario son tratados como válidos Si es así, los valores perdidos definidos por el usuario se tratan como una categoría diferente.

Los datos ya están ordenados. Si el marco muestral está clasificado previamente por los valores de las variables de estratificación, esta opción permite acelerar el proceso de selección.

Asistente de muestreo: Extraer muestra: Archivos de resultado

Figura 2-9
Asistente de muestreo: Extraer muestra: paso Archivos de resultado



Este paso permite elegir dónde dirigir los casos muestreados, las variables de ponderación, las probabilidades conjuntas y las reglas de selección de casos.

Datos muestrales. Estas opciones permiten determinar dónde se escribe el resultado de la muestra. Se puede añadir a un conjunto de datos activo, escribir en un nuevo conjunto de datos o guardar en un archivo de datos con formato IBM® SPSS® Statistics externo. Los conjuntos de datos están disponibles durante la sesión actual, pero no así en las sesiones posteriores, a menos que los haya guardado explícitamente como archivos de datos. El nombre de un conjunto de datos debe cumplir las normas de denominación de variables. Si se especifica un archivo externo o un nuevo conjunto de datos, se escribirán las variables de los resultados del muestreo y las variables del conjunto de datos activo para los casos seleccionados.

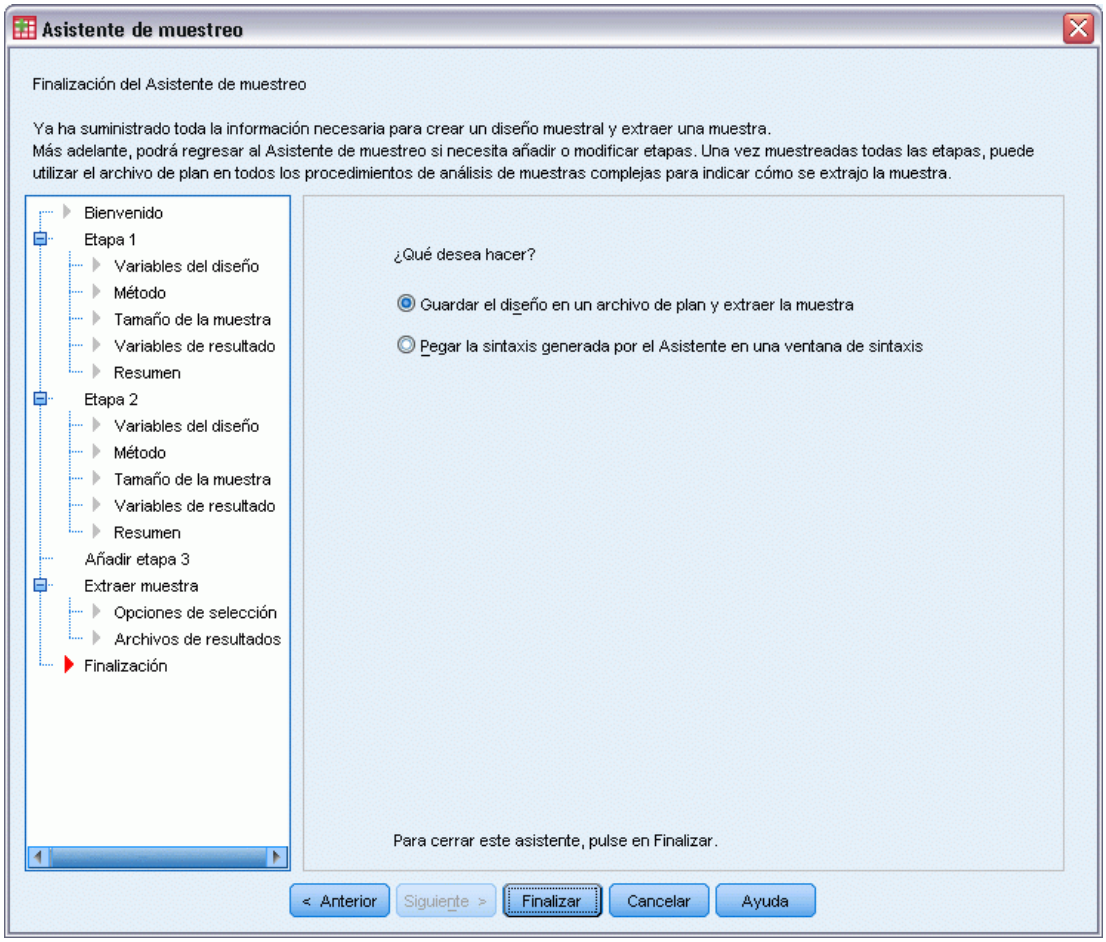
Probabilidades conjuntas. Estas opciones permiten determinar dónde se escriben las probabilidades conjuntas. Éstas se guardan en un archivo de datos con formato SPSS Statistics externo. Las probabilidades conjuntas se producen si se seleccionan la probabilidad proporcional al tamaño sin reposición, el muestreo de Brewer proporcional al tamaño, el muestreo de Sampford proporcional

al tamaño, o el muestreo de Murthy proporcional al tamaño y la estimación con reposición no se especifica.

Reglas de selección de casos. Si está construyendo la muestra por etapas, es posible que quiera guardar las reglas de selección de casos en un archivo de texto. Son útiles para construir el submarco de las etapas posteriores.

Asistente de muestreo: Finalizar

Figura 2-10
Asistente de muestreo: paso Finalizar



Este paso es el último. Puede guardar el archivo de plan y extraer la muestra ahora o pegar las selecciones en una ventana de sintaxis.

Al realizar cambios a las etapas del archivo de plan existente, puede guardar el plan editado en un archivo nuevo o sobrescribir el archivo existente. Al añadir etapas sin realizar cambios en las etapas existentes, el asistente sobrescribe de manera automática el archivo de planificación existente. Si desea guardar la planificación en un nuevo archivo, seleccione Pegar la sintaxis generada por el asistente en una ventana de sintaxis y cambie el nombre del archivo en los comandos de sintaxis.

Modificar un plan de muestreo existente

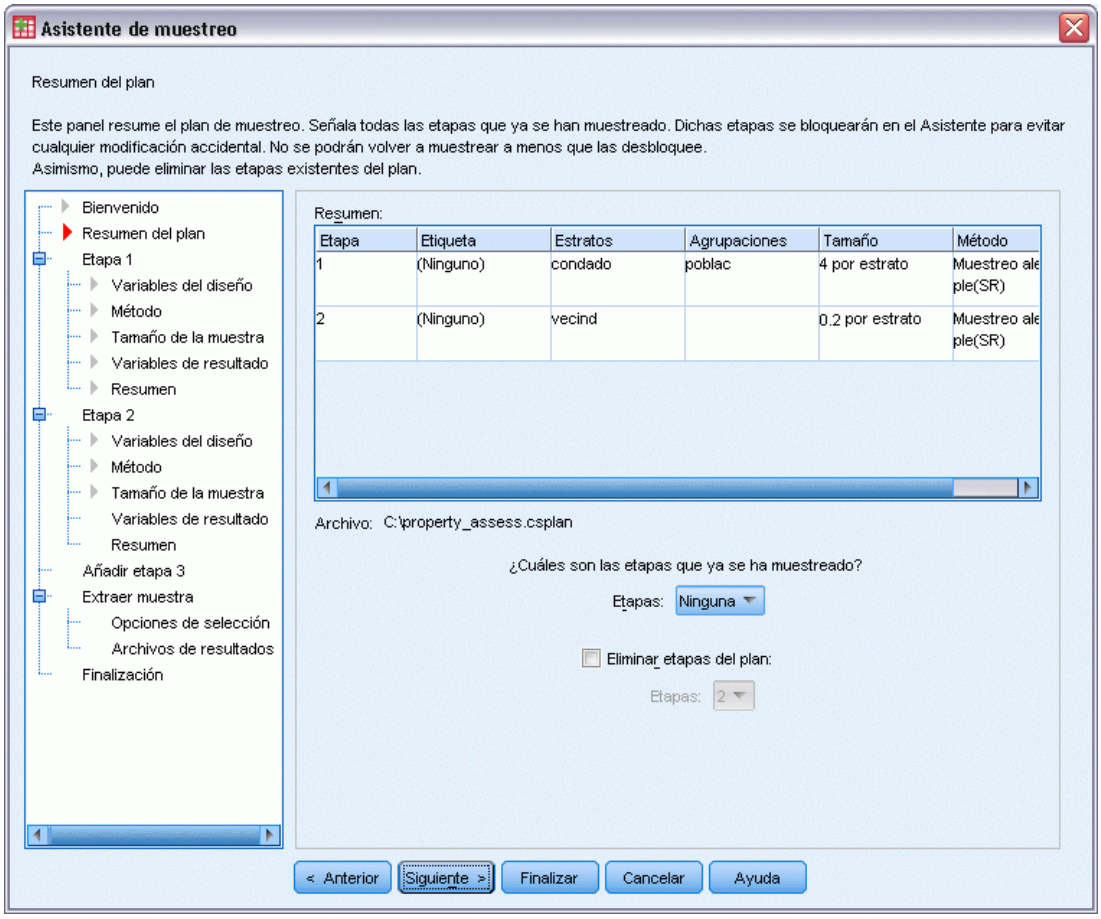
- En los menús, seleccione:
  Analizar > Complex Samples > Seleccionar una muestra...
- Seleccione Editar un diseño muestral y elegir un archivo de plan para editar.
- Pulse Siguiente para continuar usando el Asistente.
- Revise el plan de muestreo del paso Resumen del plan, y a continuación pulse Siguiente.

  Los pasos posteriores son prácticamente iguales que los de un diseño nuevo. Si desea obtener más información sobre los pasos individuales, consulte la ayuda.
- Vaya al paso final y especifique un nombre nuevo para el archivo de plan editado o sobrescriba el archivo de plan existente.

  Si lo desea, puede:
  - Especificar las etapas que ya se han muestreado.
  - Eliminar etapas del plan.

Asistente de muestreo: Resumen del plan

Figura 2-11
Asistente de muestreo: paso Resumen del plan



Este paso permite revisar el plan de muestreo e indicar las etapas que ya se han muestreado. Al editar un plan, también puede eliminar etapas del plan.

Etapas muestreadas previamente. Si un marco de muestreo ampliado no está disponible, deberá ejecutar un diseño muestral polietápico etapa por etapa. Seleccione las etapas que ya se han muestreado en la lista desplegable. Las etapas que ya se hayan ejecutado estarán bloqueadas, por lo que no estarán disponibles en el paso Extraer muestra: Opciones de selección y no se podrán modificar al editar un plan.

Eliminar etapas. Puede eliminar las etapas 2 y 3 de un diseño polietápico.

Ejecutar un plan de muestreo existente

- En los menús, seleccione:
  Analizar > Complex Samples > Seleccionar una muestra...
- Seleccione Extraer una muestra y elija un archivo de plan para ejecutar.

- Pulse Siguiente para continuar usando el Asistente.
- Revise el plan de muestreo del paso Resumen del plan, y a continuación pulse Siguiente.
- Cuando se ejecuta un plan de muestreo se omiten los pasos individuales que contienen información de la etapa. Ya puede pasar al paso de finalización.

  Si lo desea, puede especificar las etapas que ya se han muestreado.

Funciones adicionales de los comandos CSPLAN y CSSELECT

Con el lenguaje de sintaxis de comandos también podrá:

- Especificar nombres personalizados para las variables de resultado.
- Controlar los resultados en el Visor. Por ejemplo, puede suprimir el resumen por etapas del plan que se muestra si se diseña o modifica una muestra, suprimir el resumen de la distribución de los casos muestreados por etapas que se muestra si el diseño muestral se ejecuta y solicitar un resumen del procesamiento de los casos.
- Elegir un subconjunto de las variables existentes en el conjunto de datos activo para escribirlo en un archivo muestral externo o en otro conjunto de datos.

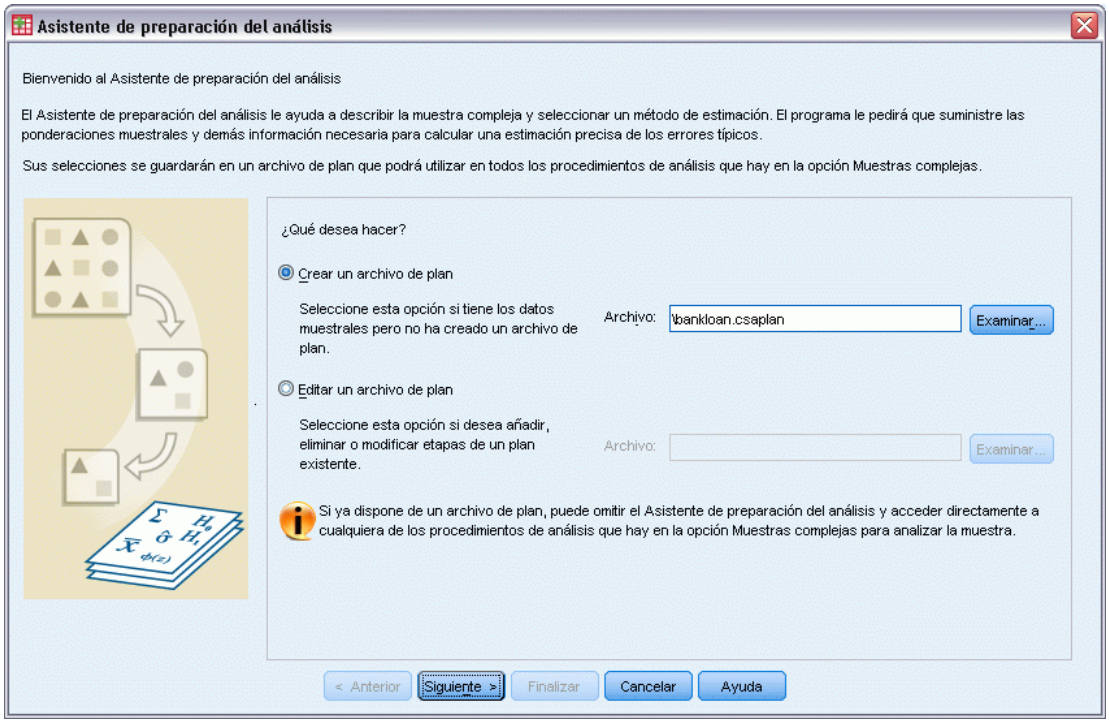
Si desea información detallada sobre la sintaxis, consulte la referencia de sintaxis de comandos (*Command Syntax Reference*).

Capítulo

3

Preparación de una muestra compleja para su análisis

Figura 3-1
Asistente de preparación del análisis: paso Bienvenida



El Asistente de preparación del análisis le guía a través de los pasos para crear o modificar un plan de análisis y utilizarlo con los distintos procedimientos de análisis de Muestras complejas. Antes de utilizar el Asistente, debe haber extraído la muestra de acuerdo con un diseño complejo.

Es más útil crear un plan nuevo cuando no se tiene acceso al archivo del plan de muestreo utilizado para extraer la muestra (recuerde que el plan de muestreo contiene un plan de análisis por defecto). Si no tiene acceso al archivo del plan de muestreo utilizado para extraer la muestra, puede utilizar el plan de análisis contenido por defecto en el archivo del plan de muestreo u omitir las especificaciones del análisis por defecto y guardar los cambios en un archivo nuevo.

Creación de un nuevo plan de análisis

- ▶ En los menús, seleccione:
  Analizar > Complex Samples > Preparar para el análisis...
- ▶ Seleccione Crear un archivo de plan, y elija un nombre de archivo de plan para guardar el plan del análisis.
- ▶ Pulse Siguiente para continuar usando el Asistente.
- ▶ Especifique la variable que contiene las ponderaciones muestrales en el paso Variables del diseño, si lo desea puede definir estratos y conglomerados.
- ▶ Ahora puede pulsar Finalizar para guardar el plan.

  Si lo desea, en los siguientes pasos puede:
  - Seleccionar el método de estimación de los errores típicos en el paso Método de estimación.
  - Especificar el número de unidades muestrales o la probabilidad de inclusión por unidad en el paso Tamaño.
  - Añadir una segunda o tercera etapa al diseño.
  - Pegar las selecciones como sintaxis de comandos.

Asistente de preparación del análisis: Variables del diseño

Figura 3-2
Asistente de preparación del análisis: paso Variables del diseño

Este paso permite identificar las variables de estratificación y conglomeración y definir las ponderaciones muestrales. También puede proporcionar una etiqueta para la etapa.

Estratos. La clasificación conjunta por las variables de estratificación define distintas subpoblaciones o estratos. El total muestral representa la combinación de las muestras independientes pertenecientes a cada estrato.

Conglomerados. Las variables de conglomeración definen grupos de unidades de observación o conglomerados. Las muestras extraídas en varias etapas seleccionan conglomerados en las etapas anteriores y, a continuación, unidades de submuestreo dentro de los conglomerados seleccionados. Al analizar un archivo de datos obtenido mediante el muestreo de conglomerados con reposición, debe incluir el índice de duplicación como una variable de conglomeración.

Ponderación muestral. Debe proporcionar ponderaciones muestrales en la primera etapa. Las ponderaciones muestrales se calculan automáticamente para las etapas posteriores del diseño actual.

Etiqueta de etapa. Puede especificar una etiqueta de cadena opcional para cada etapa. Esto se utiliza en los resultados para facilitar la identificación de la información por etapas.

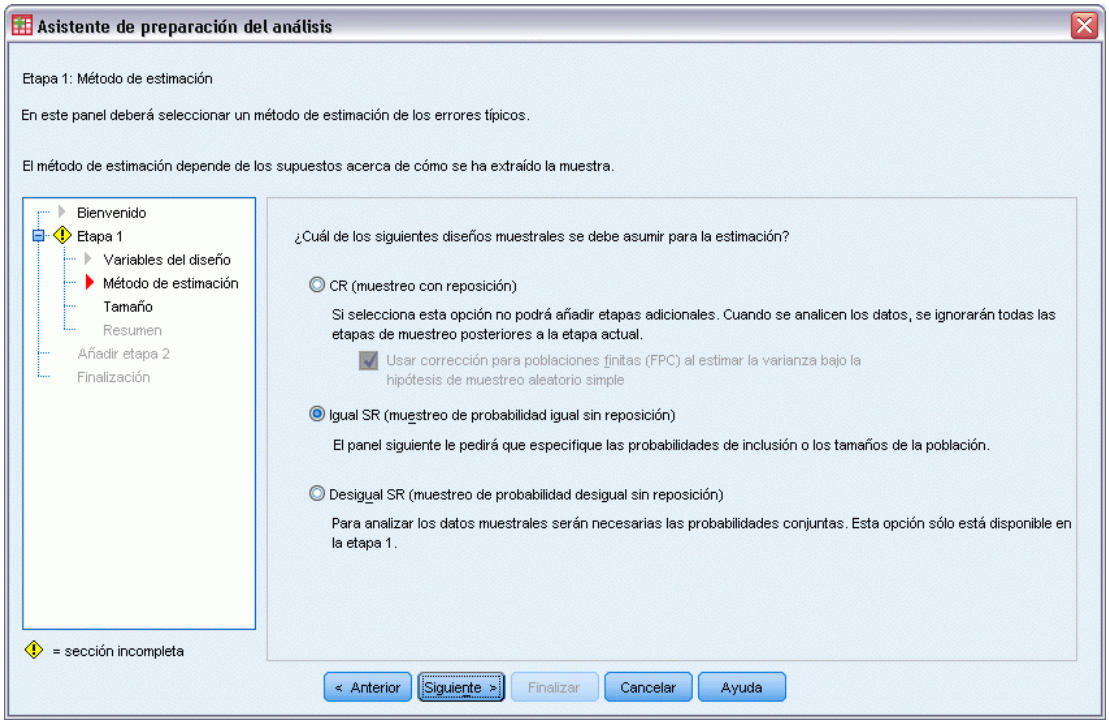
Nota: la lista de variables de origen tiene el mismo contenido a lo largo de los pasos del Asistente. En otras palabras, las variables de la lista de origen eliminadas en un paso determinado se borran de la lista en todos los pasos. Las variables devueltas a la lista de origen aparecen en todos los pasos.

Controles de árbol para desplazarse por el Asistente para el análisis

En la parte izquierda de cada paso del Asistente para el análisis se muestra un esquema con los titulares de todos los pasos. Puede navegar por el Asistente al pulsar el nombre de uno de los pasos activados en el esquema. Los pasos están activados mientras todos los pasos anteriores sean válidos, es decir, mientras cada uno de los pasos anteriores tenga las especificaciones mínimas necesarias para ese paso. Consulte la ayuda de los pasos individuales para obtener más información sobre los motivos por los que un paso dado puede no ser válido.

Asistente de preparación del análisis: Método de estimación

Figura 3-3
Asistente de preparación del análisis: paso Método de estimación



Este paso permite especificar un método de estimación para la etapa.

CR (muestreo con reposición). La estimación CR no incluye una corrección de muestreo para poblaciones finitas (FPC) al estimar la varianza bajo el diseño de muestreo complejo. Puede incluir o excluir la FPC al estima la varianza bajo muestreo aleatorio simple (SRS).

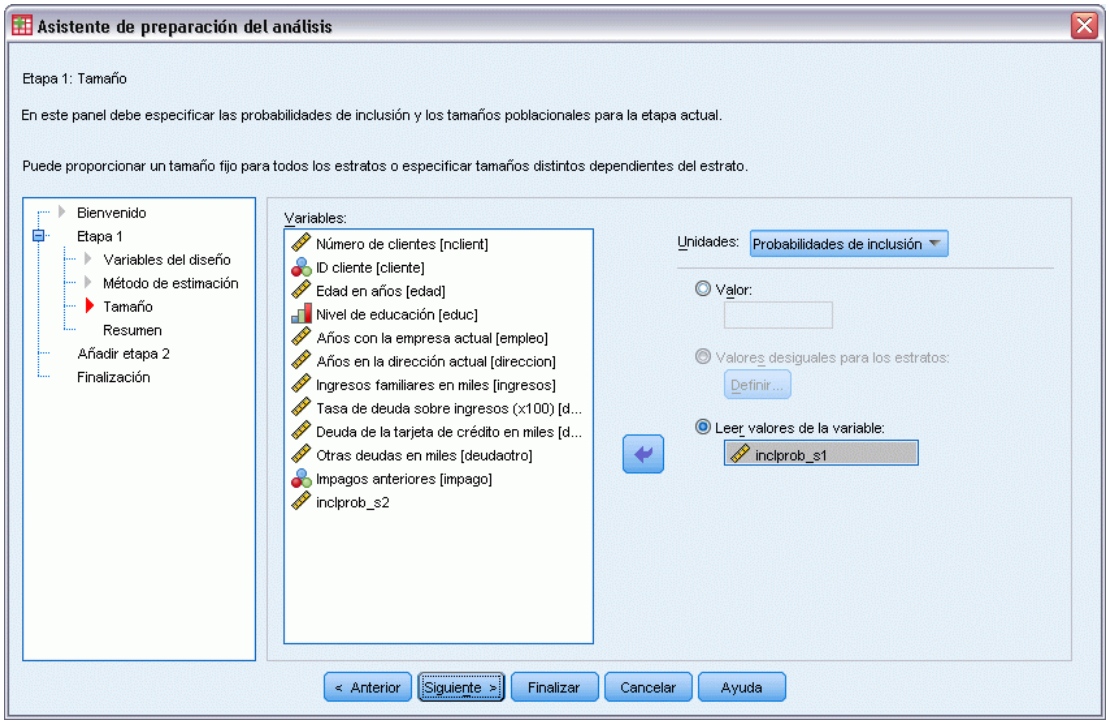
Se recomienda no incluir la FPC para la estimación de varianza SRS cuando las ponderaciones de análisis se hayan escalado de forma que no se agreguen al tamaño de la población. La estimación de varianza SRS se utiliza para calcular estadísticos como el efecto del diseño. La estimación CR sólo se puede especificar en la etapa final de un diseño; el Asistente no permitirá añadir otra etapa si se selecciona la estimación CR.

Igual SR (muestreo de igual probabilidad sin reposición). La estimación Igual SR incluye la corrección para poblaciones finitas y supone que las unidades se muestrearon con la misma probabilidad. El método Igual SR se puede especificar en cualquiera de las etapas de un diseño.

Desigual SR (muestreo de probabilidad desigual sin reposición). Además de utilizar la corrección para poblaciones finitas, el método Desigual SR tiene en cuenta las unidades muestrales (normalmente conglomerados) que han sido seleccionadas con probabilidades desiguales. Este método de estimación sólo está disponible en la primera etapa.

Asistente de preparación del análisis: Tamaño

Figura 3-4
Asistente de preparación del análisis: paso Tamaño



Este paso se utiliza para especificar las probabilidades de inclusión o los tamaños poblacionales para la etapa actual. Los tamaños pueden ser fijos o variar entre estratos. Para especificar los tamaños, los conglomerados especificados en las etapas anteriores se pueden utilizar para definir estratos. Tenga en cuenta que este paso sólo es necesario cuando se elige el método Igual SR como método de estimación.

Unidades. Puede especificar los tamaños poblacionales exactos o las probabilidades con las que se ha realizado el muestreo de las unidades.

- **Valor.** Se aplica un valor particular a todos los estratos. Si se selecciona Tamaños poblacionales como la unidad métrica, se deberá introducir un entero no negativo. Si se selecciona Probabilidades de inclusión, se deberá introducir un valor entre 0 y 1, ambos incluidos.
- **Valores desiguales para estratos.** Permite introducir distintos valores de tamaño para cada estrato a través del cuadro de diálogo Definir tamaños desiguales.
- **Leer valores de la variable.** Permite seleccionar una variable numérica que contenga los valores de tamaño para los estratos.

Definir tamaños desiguales

Figura 3-5
Cuadro de diálogo Definir tamaños desiguales

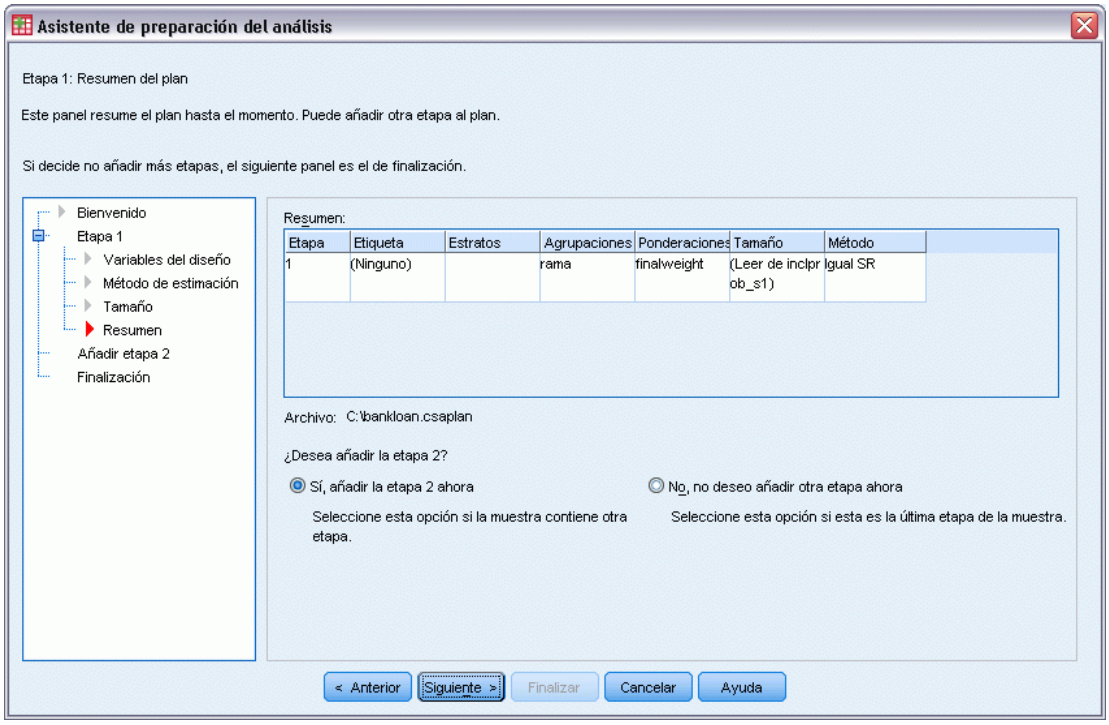
El cuadro de diálogo Definir tamaños desiguales permite introducir los tamaños para cada estrato.

Rejilla de especificaciones de tamaño. La rejilla muestra la clasificación conjunta de hasta cinco variables de conglomeración o estrato, con una combinación de estrato/conglomerado por fila. Las variables elegibles en la rejilla serán todas las variables de estratificación de las etapas anteriores y actuales además de todas las variables de conglomeración de las etapas anteriores. Las variables se pueden reordenar dentro de la rejilla o ser desplazadas a la lista Excluir. Introduzca los tamaños en la última columna de la derecha. Pulse en Etiquetas o Valores para conmutar entre la visualización de las etiquetas de valor y los valores de los datos para las variables de estratificación y de conglomeración de las casillas de la rejilla. Las casillas que contienen valores sin etiquetas siempre muestran valores. Pulse Actualizar estratos para volver a rellenar la rejilla con cada combinación de los valores de los datos etiquetados para las variables de la rejilla.

Excluir. Para especificar los tamaños de un subconjunto de combinaciones de estrato/conglomerado, desplace una o más variables a la lista Excluir. Estas variables no se utilizan para definir tamaños muestrales.

Asistente de preparación del análisis: Resumen del plan

Figura 3-6
Asistente de preparación del análisis, paso Resumen del plan



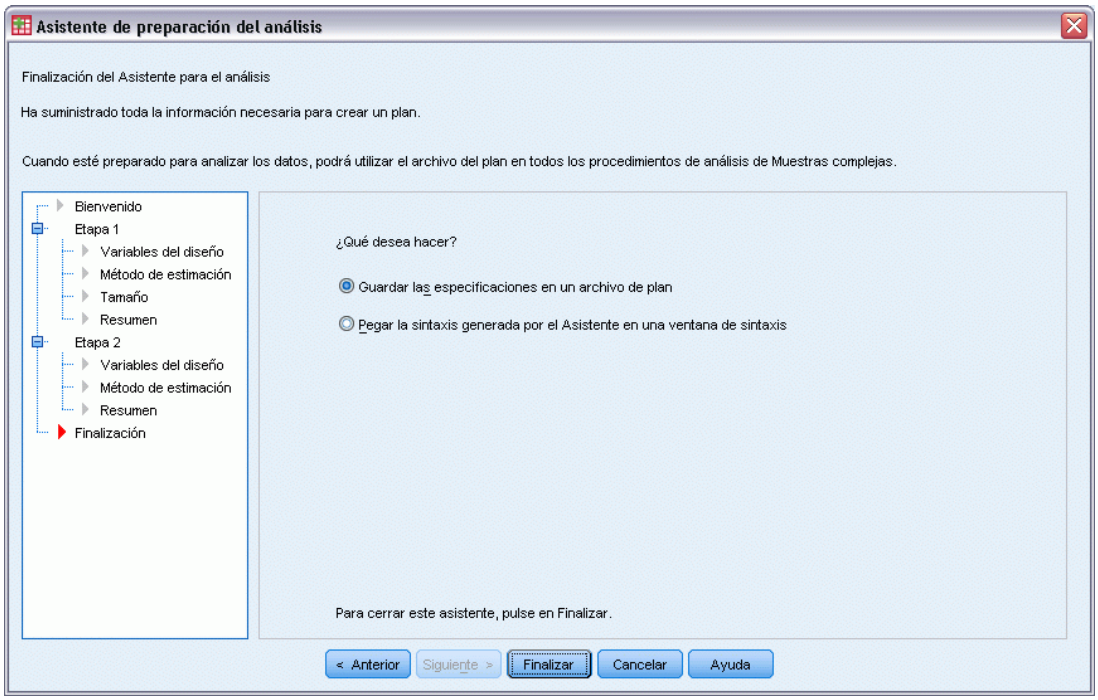
Este paso es el último de cada etapa y proporciona un resumen de las especificaciones del diseño del análisis hasta la etapa actual, ésta incluida. A partir de aquí, puede pasar a la siguiente etapa (creándola si fuera necesario) o guardar las especificaciones del análisis.

Si no puede añadir otra etapa, esto puede deberse a:

- No se especificó ninguna variable de conglomeración en el paso Variables del diseño.
- Seleccionó la estimación CR en el paso Método de estimación.
- Este paso es el tercero del análisis; el Asistente admite un máximo de tres etapas.

Asistente de preparación del análisis: Finalizar

Figura 3-7
Asistente de preparación del análisis: Finalización



Este paso es el último. Puede guardar el archivo del plan ahora o pegar las selecciones en una ventana de sintaxis.

Al realizar cambios a las etapas del archivo de plan existente, puede guardar el plan editado en un archivo nuevo o sobrescribir el archivo existente. Al añadir etapas sin realizar cambios en las etapas existentes, el asistente sobrescribe de manera automática el archivo de planificación existente. Si desea guardar la planificación en un nuevo archivo, elija Pegar la sintaxis generada por el asistente en una ventana de sintaxis y cambie el nombre del archivo en los comandos de sintaxis.

Modificar un plan de análisis existente

- En los menús, seleccione:
  Analizar > Complex Samples > Preparar para el análisis...
- Seleccione Editar un archivo de plan y elija un nombre de archivo de plan en el que se guardará el plan del análisis.
- Pulse Siguiente para continuar usando el Asistente.

- Revise el plan de análisis en el paso Resumen del plan y, a continuación, pulse Siguiente.

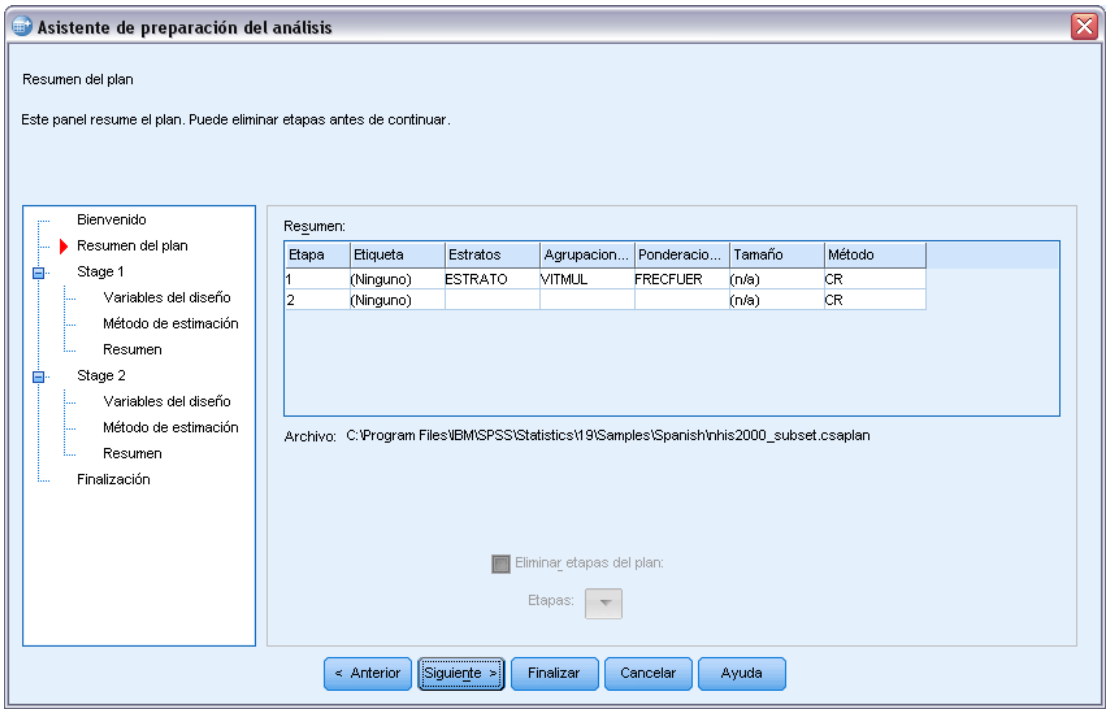
  Los pasos posteriores son prácticamente iguales que los de un diseño nuevo. Si desea obtener más información, consulte la ayuda sobre los pasos individuales.

- Vaya al paso final y especifique un nombre nuevo para el archivo de plan editado o sobrescriba el archivo de plan existente.

  Si lo desea, puede eliminar algunas etapas del plan.

Asistente de preparación del análisis: Resumen del plan

Figura 3-8
Asistente de preparación del análisis, paso Resumen del plan



Este paso permite revisar el plan de análisis y eliminar etapas del plan.

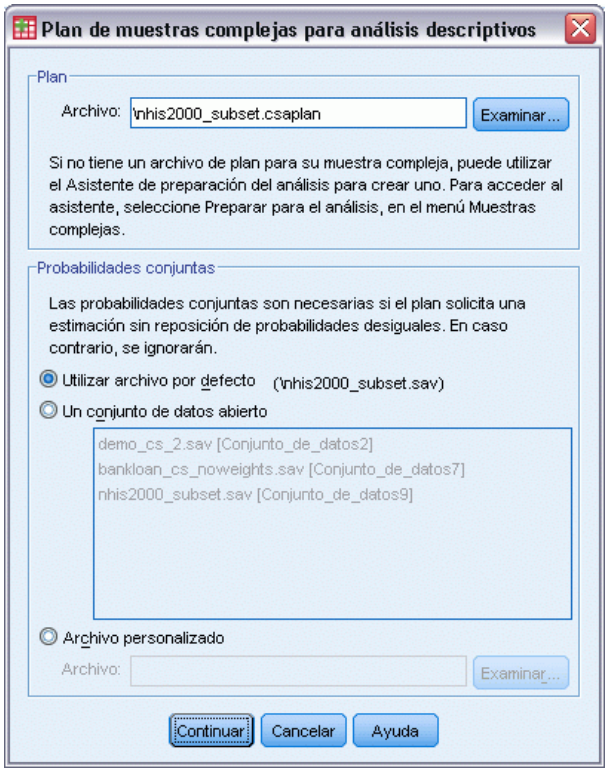
Eliminar etapas. Puede eliminar las etapas 2 y 3 de un diseño polietápico. Debido a que los planes deben tener al menos una etapa, puede editar la etapa 1 pero no eliminarla del diseño.

Capítulo
4

Plan de muestras complejas

Los procedimientos de análisis de Muestras complejas requieren las especificaciones de análisis de un archivo de plan de muestreo o un plan de análisis para poder proporcionar resultados válidos.

Figura 4-1
Cuadro de diálogo Plan de muestras complejas



Plan. Especifique la ruta de un archivo de plan de muestreo o análisis.

Probabilidades conjuntas. Para utilizar una estimación Desigual SR para los conglomerados extraídos utilizando un método PPS SR, debe especificar un archivo independiente o un conjunto de datos abierto que contenga las probabilidades conjuntas. El archivo o conjunto de datos se crea mediante el Asistente de muestreo durante el muestreo.

Capítulo
5

Frecuencias de Muestras complejas

El procedimiento Frecuencias de Muestras complejas genera tablas de frecuencias para las variables seleccionadas y muestra estadísticos univariantes. Si lo desea, puede solicitar estadísticos por subgrupos, definidos por una o más variables categóricas.

Ejemplo. Mediante el procedimiento Frecuencias de Muestras complejas, puede obtener estadísticos tabulares univariantes para el consumo de vitaminas entre los ciudadanos de EE.UU., basados en los resultados del National Health Interview Survey (NHIS, Centro Nacional de Estadísticas de Salud) y con un plan de análisis adecuado para estos datos de uso público.

Estadísticos. El procedimiento genera estimaciones de los tamaños poblacionales de las casillas, además de errores típicos, intervalos de confianza, coeficientes de variación, efectos del diseño, raíz cuadrada de los efectos del diseño, valores acumulados y recuentos no ponderados para cada estimación. Además, se calculan los estadísticos de chi-cuadrado y la razón de verosimilitud para el contraste de proporciones de casilla iguales.

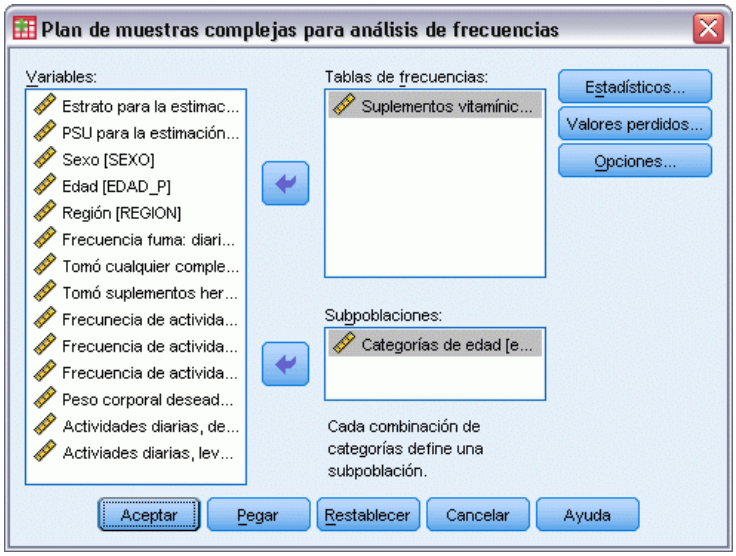
Datos. Variables para las que se generan las tablas de frecuencias deben ser categóricas. Las variables que definen las subpoblaciones pueden ser numéricas o de cadena, pero siempre deben ser categóricas.

Supuestos. Los casos del archivo de datos representan una muestra de un diseño complejo que se debe analizar según las especificaciones del archivo seleccionado en el Cuadro de diálogo Plan de muestras complejas.

Obtención de Frecuencias de Muestras complejas

- En los menús, seleccione:
  Analizar > Complex Samples > Frecuencias...
- Seleccione un archivo de plan. Si lo desea, elija un archivo de probabilidades conjuntas personalizado.
- Pulse en Continuar.

Figura 5-1
Cuadro de diálogo Frecuencias

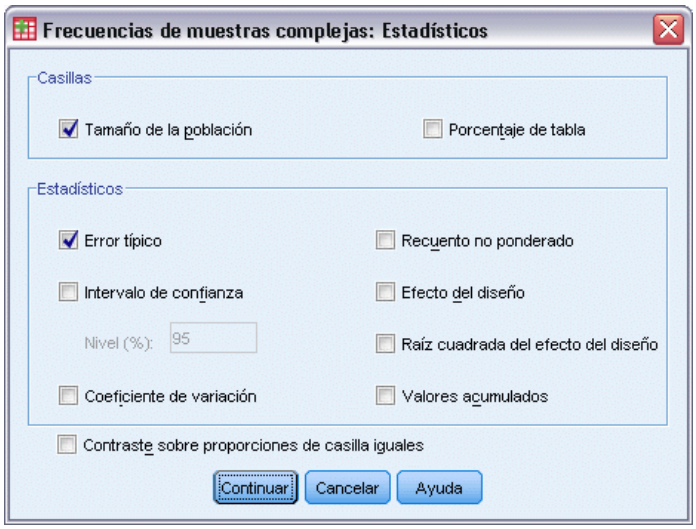


- Seleccione al menos una variable de frecuencia.

Si lo desea, puede especificar variables para definir subpoblaciones. Los estadísticos se calculan por separado para cada subpoblación.

Frecuencias de Muestras complejas: Estadísticos

Figura 5-2
Cuadro de diálogo Frecuencias: Estadísticos



Casillas. Este grupo permite solicitar estimaciones de los tamaños poblacionales de las casillas así como porcentajes de tabla.

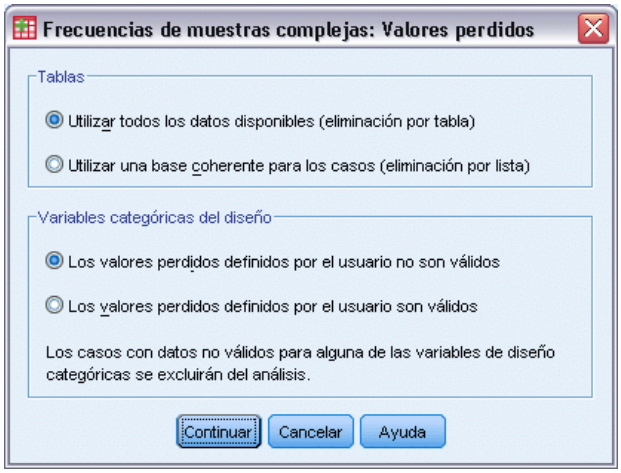
Estadísticos. Este grupo genera estadísticos asociados con el tamaño poblacional o los porcentajes de tabla.

- **Error típico.** El error típico de la estimación.
- **Intervalo de confianza.** Intervalo de confianza para la estimación, utilizando el nivel especificado.
- **Coeficiente de variación.** Cociente del error típico de la estimación dividida por la estimación.
- **Recuento no ponderado.** Número de unidades utilizadas para calcular la estimación.
- **Efecto del diseño.** Cociente de la variación de la estimación entre la variación obtenida al suponer que la muestra es una muestra aleatoria simple. Es una medida del efecto de especificar un diseño complejo donde los valores más distantes de 1 indican efectos mayores.
- **Raíz cuadrada del efecto del diseño.** Es una medida del efecto de especificar un diseño complejo donde los valores más distantes de 1 indican efectos mayores.
- **Valores acumulados.** La estimación acumulada a través de los valores de la variable.

Contraste sobre proporciones de casilla iguales. Esto genera los contrastes de chi-cuadrado y la razón de verosimilitud sobre la hipótesis de que las categorías de una variable tienen la misma frecuencia. Se realizan contrastes por separado para cada variable.

Muestras complejas: Valores perdidos

Figura 5-3
Cuadro de diálogo Valores perdidos



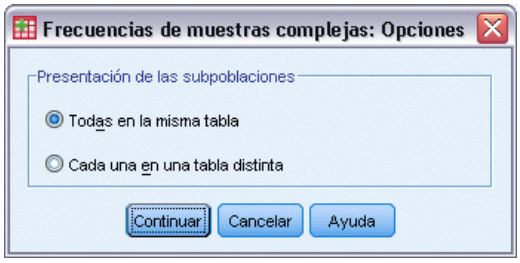
Tablas. Este grupo determina los casos que se utilizan en el análisis.

- **Utilizar todos los datos disponibles.** Los valores perdidos se determinan en base a tabla por tabla. Así, los casos utilizados para calcular los estadísticos pueden variar a través de la frecuencia o tablas de contingencia.
- **Utilizar una base coherente para los casos.** Los valores perdidos se determinan a través de todas las variables. Por lo tanto, los casos utilizados para calcular los estadísticos son coherentes con las tablas.

Variables categóricas del diseño. Este grupo determina si los valores perdidos definidos por el usuario son considerados válidos o inválidos.

Opciones de Muestras complejas

Figura 5-4
Cuadro de diálogo Opciones



Mostrar subpoblación. Puede elegir entre mostrar las subpoblaciones en la misma tabla o en tablas separadas.

Capítulo

6

Descriptivos de Muestras complejas

El procedimiento Descriptivos de Muestras complejas muestra estadísticos de resumen univariantes para distintas variables. Si lo desea, puede solicitar estadísticos por subgrupos, definidos por una o más variables categóricas.

Ejemplo. Mediante el procedimiento Descriptivos de Muestras complejas, puede obtener estadísticos descriptivos univariantes de los niveles de actividad de los ciudadanos de EE.UU., basados en los resultados de la National Health Interview Survey (NHIS, Centro Nacional de Estadísticas de Salud) y con un plan de análisis adecuado para estos datos de uso público.

Estadísticos. El procedimiento genera medias y sumas, además de pruebas *t*, errores típicos, intervalos de confianza, coeficientes de variación, recuentos no ponderados, efectos del diseño y la raíz cuadrada del efecto del diseño de cada estimación.

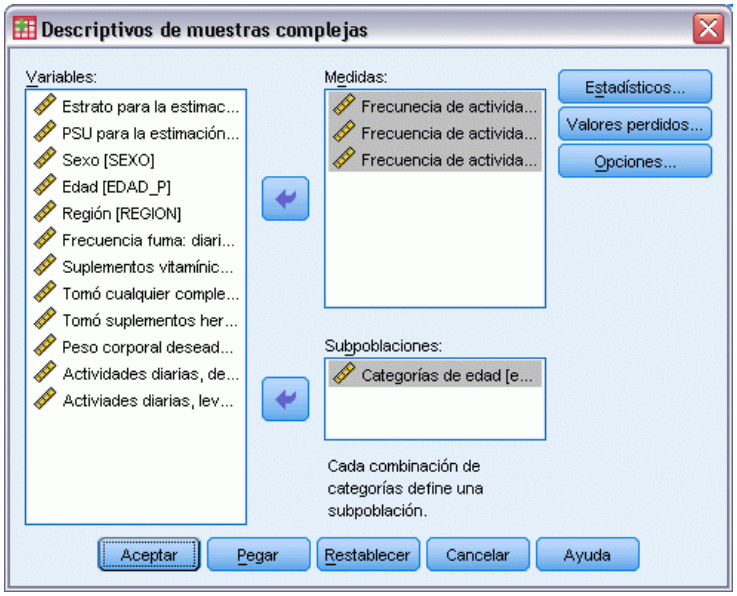
Datos. Las medidas deben ser variables de escala. Las variables que definen las subpoblaciones pueden ser numéricas o de cadena, pero siempre deben ser categóricas.

Supuestos. Los casos del archivo de datos representan una muestra de un diseño complejo que se debe analizar según las especificaciones del archivo seleccionado en el Cuadro de diálogo Plan de muestras complejas.

Obtención de Descriptivos de Muestras complejas

- En los menús, seleccione:
  Analizar > Complex Samples > Descriptivos...
- Seleccione un archivo de plan. Si lo desea, elija un archivo de probabilidades conjuntas personalizado.
- Pulse en Continuar.

Figura 6-1
Cuadro de diálogo Descriptivos

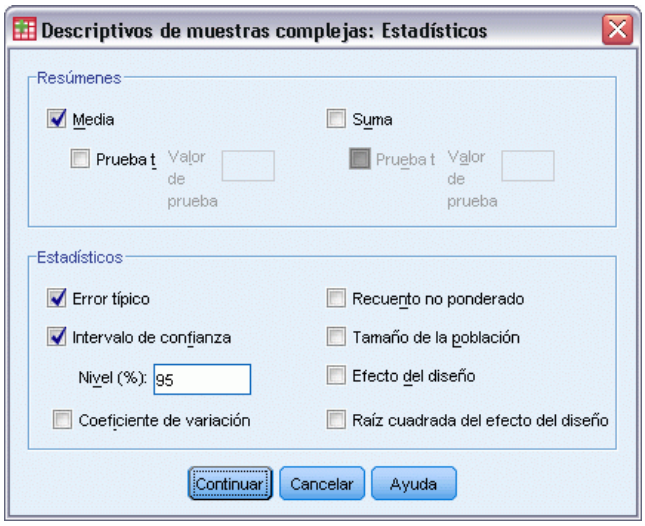


▶ Seleccione al menos una variable de medida.

Si lo desea, puede especificar variables para definir subpoblaciones. Los estadísticos se calculan por separado para cada subpoblación.

Descriptivos de Muestras complejas: Estadísticos

Figura 6-2
Cuadro de diálogo Descriptivos: Estadísticos



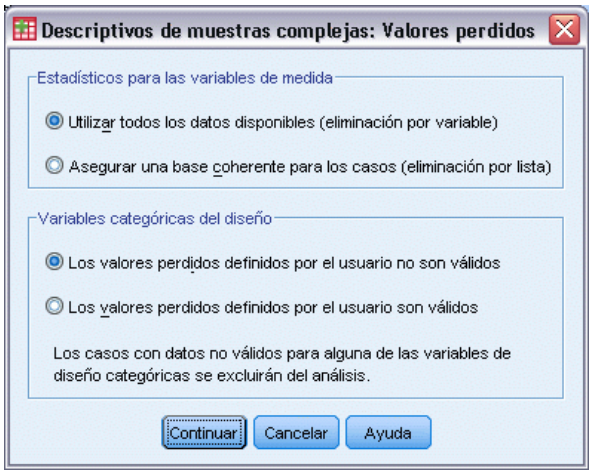
Resúmenes. Este grupo permite solicitar estimaciones de las medias y sumas de las variables de medida. Además, puede solicitar pruebas *t* de las estimaciones con respecto a un valor especificado.

Estadísticos. Este grupo genera estadísticos asociados con la media o la suma.

- **Error típico.** El error típico de la estimación.
- **Intervalo de confianza.** Intervalo de confianza para la estimación, utilizando el nivel especificado.
- **Coeficiente de variación.** Cociente del error típico de la estimación dividida por la estimación.
- **Recuento no ponderado.** Número de unidades utilizadas para calcular la estimación.
- **Tamaño poblacional.** Número estimado de unidades en la población.
- **Efecto del diseño.** Cociente de la variación de la estimación entre la variación obtenida al suponer que la muestra es una muestra aleatoria simple. Es una medida del efecto de especificar un diseño complejo donde los valores más distantes de 1 indican efectos mayores.
- **Raíz cuadrada del efecto del diseño.** Es una medida del efecto de especificar un diseño complejo donde los valores más distantes de 1 indican efectos mayores.

Valores perdidos en los descriptivos de Muestras complejas

Figura 6-3
Cuadro de diálogo Valores perdidos de descriptivos



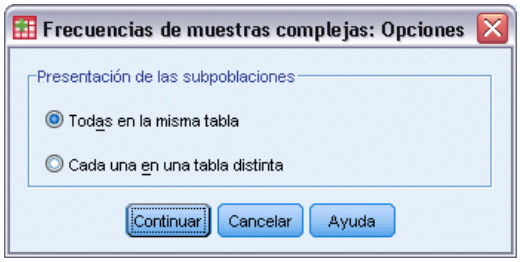
Estadísticos para variables de medida. Este grupo determina los casos que se utilizan en el análisis.

- **Utilizar todos los datos disponibles.** Los valores perdidos se determinan variable por variable; por ello los casos utilizados para calcular los estadísticos pueden variar entre las variables de medida.
- **Asegurar una base coherente para los casos.** Los valores perdidos se determinan a partir de todas las variables, así, los casos utilizados para calcular los estadísticos son coherentes.

Variables categóricas del diseño. Este grupo determina si los valores perdidos definidos por el usuario son considerados válidos o inválidos.

Opciones de Muestras complejas

Figura 6-4
Cuadro de diálogo Opciones



Mostrar subpoblación. Puede elegir entre mostrar las subpoblaciones en la misma tabla o en tablas separadas.

Capítulo
7

Tablas de contingencia de Muestras complejas

El procedimiento Tablas de contingencia de Muestras complejas genera tablas de contingencia para los pares de variables seleccionadas y muestra estadísticos sobre la clasificación bivariante. Si lo desea, puede solicitar estadísticos por subgrupos, definidos por una o más variables categóricas.

Ejemplo. Mediante el procedimiento Tablas de contingencia de Muestras complejas, se pueden obtener estadísticos de clasificaciones cruzadas de la frecuencia de consumo de tabaco por el consumo de vitaminas en los ciudadanos de EE.UU, basado en los resultados del National Health Interview Survey (NHIS, Centro Nacional de Estadísticas de Salud) y con un plan de análisis adecuado para estos datos de uso público.

Estadísticos. El procedimiento genera estimaciones de los tamaños poblacionales de las casillas, así como porcentajes de tabla, columna y fila, además de errores típicos, intervalos de confianza, coeficientes de variación, valores esperados, efectos del diseño, raíz cuadrada de los efectos del diseño, residuos, residuos corregidos y frecuencias no ponderadas para cada estimación. Para las tablas 2 por 2, se calculan la razón de ventajas, el riesgo relativo y la diferencia de riesgos. Además, para el contraste de independencia de las variables de las filas y las variables de las columnas, se calculan los estadísticos de Pearson y de la razón de verosimilitud.

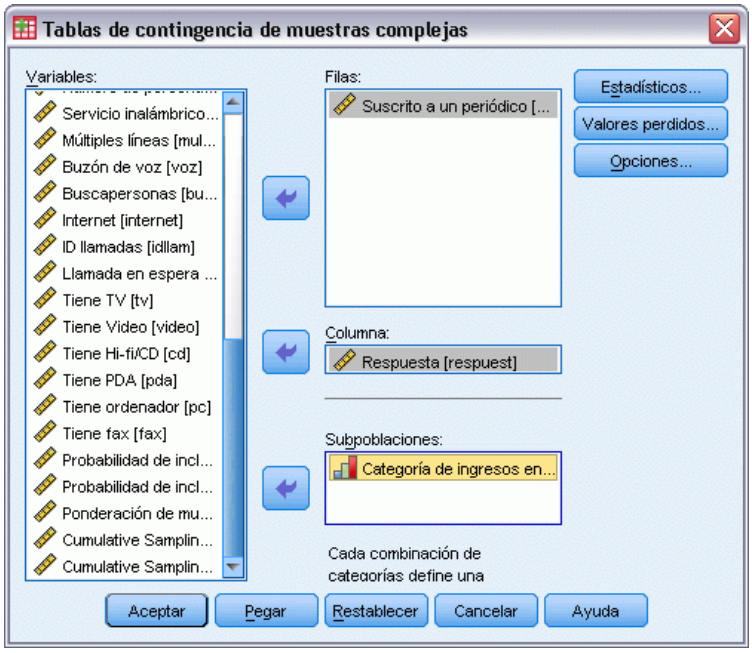
Datos. Las variables de fila y columna deben ser categóricas. Las variables que definen las subpoblaciones pueden ser numéricas o de cadena, pero siempre deben ser categóricas.

Supuestos. Los casos del archivo de datos representan una muestra de un diseño complejo que se debe analizar según las especificaciones del archivo seleccionado en el Cuadro de diálogo Plan de muestras complejas.

Obtención de Tablas de contingencia de Muestras complejas

- En los menús, seleccione:
  Analizar > Complex Samples > Tablas de contingencia...
- Seleccione un archivo de plan. Si lo desea, elija un archivo de probabilidades conjuntas personalizado.
- Pulse en Continuar.

Figura 7-1
Cuadro de diálogo Tablas de contingencia

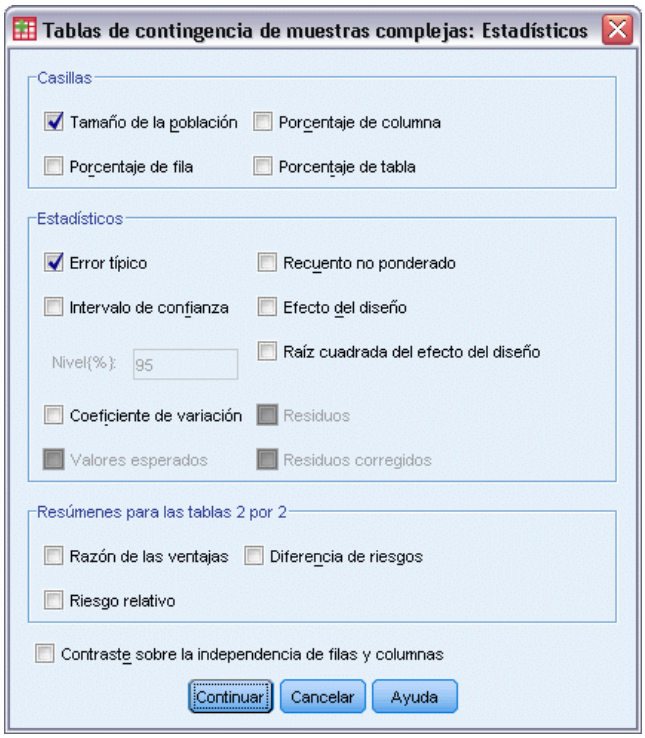


▶ Seleccione al menos una variable de fila y una variable de columna.

Si lo desea, puede especificar variables para definir subpoblaciones. Los estadísticos se calculan por separado para cada subpoblación.

Tablas de contingencia de Muestras complejas: Estadísticos

Figura 7-2
Cuadro de diálogo Tablas de contingencia: Estadísticos



Casillas. Este grupo permite solicitar estimaciones del tamaño poblacional de las casillas así como porcentajes de columna, fila y de tabla.

Estadísticos. Este grupo genera estadísticos asociados con el tamaño de la población y los porcentajes de tabla, columna y fila.

- **Error típico.** El error típico de la estimación.
- **Intervalo de confianza.** Intervalo de confianza para la estimación, utilizando el nivel especificado.
- **Coeficiente de variación.** Cociente del error típico de la estimación dividida por la estimación.
- **Valores esperados.** Valor esperado de la estimación, bajo la hipótesis de independencia de las variables de fila y columna.
- **Recuento no ponderado.** Número de unidades utilizadas para calcular la estimación.
- **Efecto del diseño.** Cociente de la variación de la estimación entre la variación obtenida al suponer que la muestra es una muestra aleatoria simple. Es una medida del efecto de especificar un diseño complejo donde los valores más distantes de 1 indican efectos mayores.
- **Raíz cuadrada del efecto del diseño.** Es una medida del efecto de especificar un diseño complejo donde los valores más distantes de 1 indican efectos mayores.

- **Residuos.** El valor pronosticado 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.
- **Residuos corregidos.** El residuo de una casilla (el valor observado menos el valor pronosticado) dividido por una estimación de su error típico. El residuo tipificado resultante viene expresado en unidades de desviación típica, por encima o por debajo de la media.

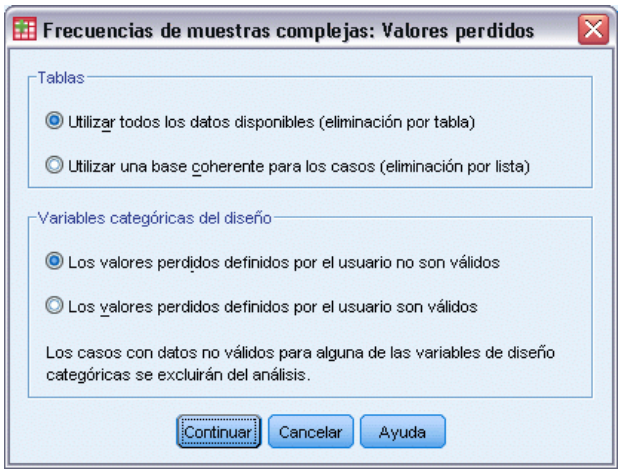
Resúmenes para las tablas 2 por 2. Este grupo genera estadísticos para las tablas en las que la variable de fila y la de columna tienen dos categorías. Cada una es una medida de la fuerza de la asociación entre la presencia de un factor y la aparición de un evento.

- **Razón de las ventajas.** Cuando la ocurrencia del factor es poco común, se puede utilizar la razón de las ventajas como estimación del riesgo relativo.
- **Riesgo relativo.** La razón del riesgo de un evento en presencia del factor respecto al riesgo del evento en ausencia del factor.
- **Diferencia de riesgos.** La diferencia entre el riesgo de un evento en presencia del factor y el riesgo del evento en ausencia del factor.

Contraste sobre la independencia de filas y columnas. Esta opción genera los contrastes de chi-cuadrado y la razón de verosimilitud sobre la hipótesis de que las variables de fila y columna son independientes. Se realizan contrastes por separado para cada pareja de variables.

Muestras complejas: Valores perdidos

Figura 7-3
Cuadro de diálogo Valores perdidos



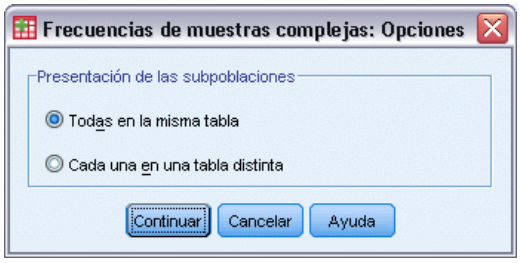
Tablas. Este grupo determina los casos que se utilizan en el análisis.

- **Utilizar todos los datos disponibles.** Los valores perdidos se determinan en base a tabla por tabla. Así, los casos utilizados para calcular los estadísticos pueden variar a través de la frecuencia o tablas de contingencia.
- **Utilizar una base coherente para los casos.** Los valores perdidos se determinan a través de todas las variables. Por lo tanto, los casos utilizados para calcular los estadísticos son coherentes con las tablas.

Variables categóricas del diseño. Este grupo determina si los valores perdidos definidos por el usuario son considerados válidos o inválidos.

Opciones de Muestras complejas

Figura 7-4
Cuadro de diálogo Opciones



Mostrar subpoblación. Puede elegir entre mostrar las subpoblaciones en la misma tabla o en tablas separadas.

Capítulo
8

Razones de Muestras complejas

El procedimiento Razones de Muestras complejas muestra estadísticos de resumen univariantes para razones de variables. Si lo desea, puede solicitar estadísticos por subgrupos, definidos por una o más variables categóricas.

Ejemplo. Mediante el procedimiento Razones de Muestras complejas, puede obtener estadísticos descriptivos para el cociente del valor de la propiedad actual sobre el último valor certificado, basado en los resultados de una encuesta a nivel estatal llevada a cabo según un diseño complejo y con un plan de análisis adecuado para los datos.

Estadísticos. El procedimiento genera estimaciones de razón, pruebas *t*, errores típicos, intervalos de confianza, coeficientes de variación, recuentos no ponderados, tamaños poblacionales, efectos del diseño y raíz cuadrada del efecto del diseño.

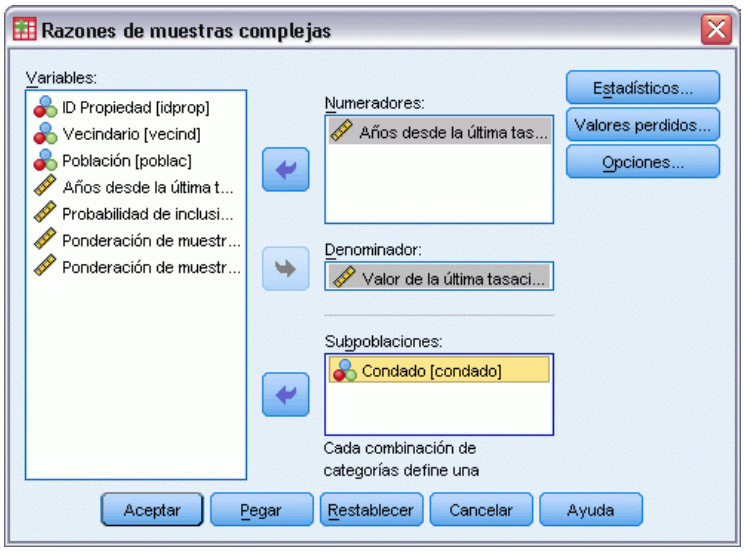
Datos. Los numeradores y los denominadores deben ser variables de escala con valores positivos. Las variables que definen las subpoblaciones pueden ser numéricas o de cadena, pero siempre deben ser categóricas.

Supuestos. Los casos del archivo de datos representan una muestra de un diseño complejo que se debe analizar según las especificaciones del archivo seleccionado en el Cuadro de diálogo Plan de muestras complejas.

Obtención de razones de Muestras complejas

- En los menús, seleccione:
  Analizar > Complex Samples > Razones...
- Seleccione un archivo de plan. Si lo desea, elija un archivo de probabilidades conjuntas personalizado.
- Pulse en Continuar.

Figura 8-1
Cuadro de diálogo Razones de Muestras complejas

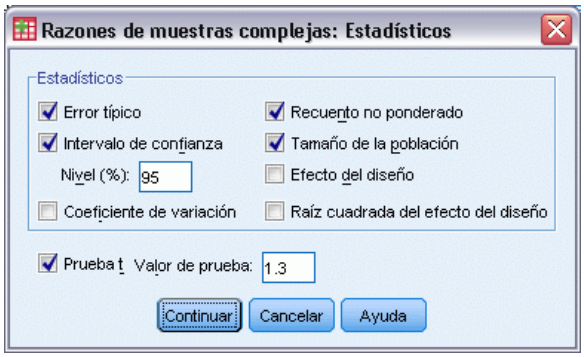


- Seleccione al menos una variable de numerador y una variable de denominador.

  Si lo desea, puede especificar variables para definir subgrupos para los que se desea generar estadísticos.

Razones de Muestras complejas. Estadísticos

Figura 8-2
Cuadro de diálogo Estadísticos de la razón



Estadísticos. Este grupo genera estadísticos asociados con la estimación de la razón.

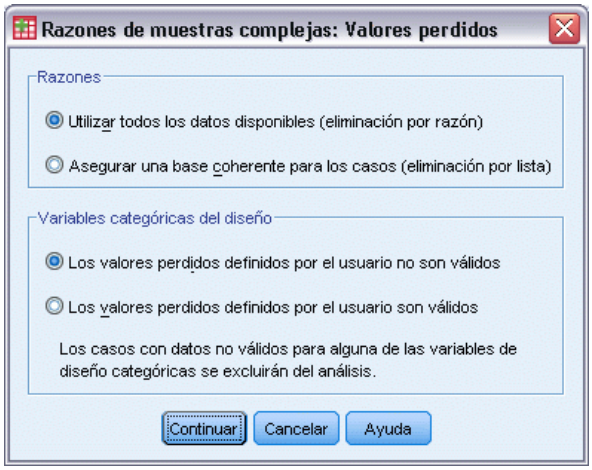
- **Error típico.** El error típico de la estimación.
- **Intervalo de confianza.** Intervalo de confianza para la estimación, utilizando el nivel especificado.
- **Coeficiente de variación.** Cociente del error típico de la estimación dividida por la estimación.
- **Recuento no ponderado.** Número de unidades utilizadas para calcular la estimación.
- **Tamaño poblacional.** Número estimado de unidades en la población.

- **Efecto del diseño.** Cociente de la variación de la estimación entre la variación obtenida al suponer que la muestra es una muestra aleatoria simple. Es una medida del efecto de especificar un diseño complejo donde los valores más distantes de 1 indican efectos mayores.
- **Raíz cuadrada del efecto del diseño.** Es una medida del efecto de especificar un diseño complejo donde los valores más distantes de 1 indican efectos mayores.

Prueba t. Puede solicitar pruebas *t* de las estimaciones con respecto a un valor especificado.

Razones de Muestras complejas: Valores perdidos

Figura 8-3
El cuadro de diálogo Razones de Muestras complejas: Valores perdidos



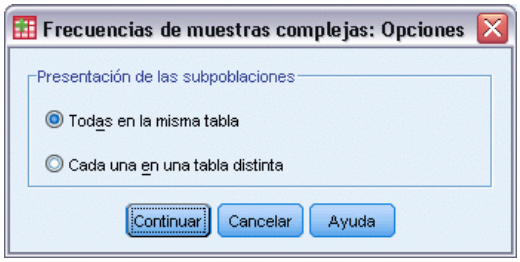
Razones. Este grupo determina los casos que se utilizan en el análisis.

- **Utilizar todos los datos disponibles.** Los valores perdidos se determinan en base a razón por razón. Así, los casos utilizados para calcular los estadísticos pueden variar a través de pares numerador-denominador.
- **Asegurar una base coherente para los casos.** Los valores perdidos se determinan a través de todas las variables. Por lo tanto, los casos utilizados para calcular los estadísticos son coherentes con las tablas.

Variables categóricas del diseño. Este grupo determina si los valores perdidos definidos por el usuario son considerados válidos o inválidos.

Opciones de Muestras complejas

Figura 8-4
Cuadro de diálogo Opciones



Mostrar subpoblación. Puede elegir entre mostrar las subpoblaciones en la misma tabla o en tablas separadas.

Capítulo
9

Modelo lineal general de muestras complejas

El procedimiento Modelo lineal general de muestras complejas (CSGLM) realiza análisis de regresión lineal y análisis de varianza y covarianza de muestras extraídas mediante métodos de muestreo complejo. Si lo desea, puede solicitar análisis de una subpoblación.

Ejemplo. Una cadena de tiendas de alimentos realiza una encuesta sobre los hábitos de compra de una serie de clientes basándose en un diseño complejo. Una vez obtenidos los resultados de la encuesta y la cantidad que cada cliente gastó el mes anterior, la cadena desea averiguar si la frecuencia con que los clientes hacen la compra está relacionada con la cantidad mensual que gastan, controlando el sexo del cliente e incorporando el diseño del muestreo.

Estadísticos. El procedimiento genera estimaciones, errores típicos, pruebas *t*, efectos del diseño, raíz cuadrada de los efectos del diseño para parámetros de modelo y las correlaciones y covarianzas entre las estimaciones de los parámetros. Las medidas de ajuste del modelo y los estadísticos descriptivos de las variables dependientes e independientes también están disponibles. Además, se pueden solicitar medias marginales estimadas para los niveles de factores de modelado u las interacciones de los factores.

Datos. La variable dependiente es cuantitativa. Los factores son categóricos; pueden tener valores numéricos o valores de cadena de hasta ocho caracteres. Las covariables son variables cuantitativas que están relacionadas con la variable dependiente. Las variables que definen las subpoblaciones pueden ser numéricas o de cadena, pero siempre deben ser categóricas.

Supuestos. Los casos del archivo de datos representan una muestra de un diseño complejo que se debe analizar según las especificaciones del archivo seleccionado en el Cuadro de diálogo Plan de muestras complejas.

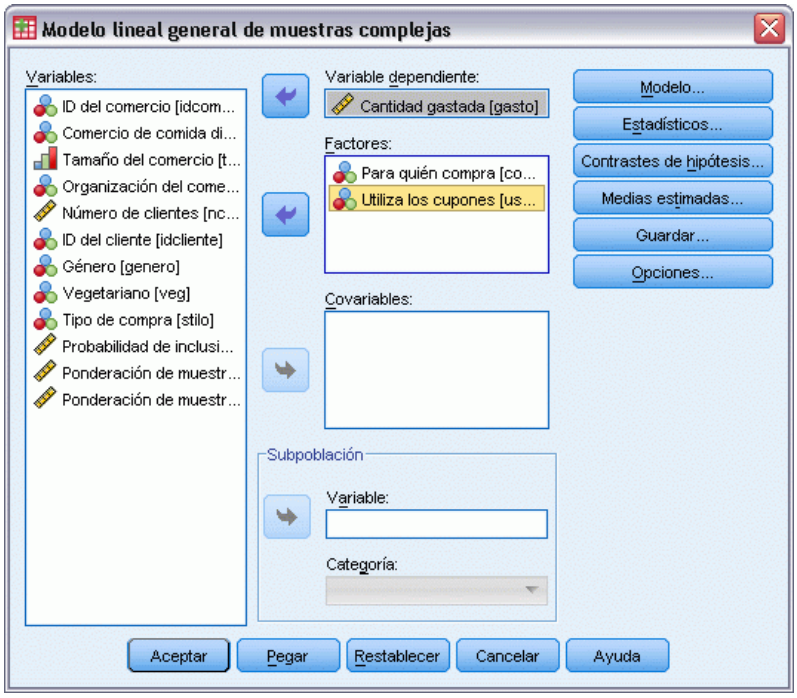
Para obtener un Modelo lineal general de muestras complejas

Seleccione en los menús:

Analizar > Muestras complejas > Modelo lineal general...

- Seleccione un archivo de plan. Si lo desea, elija un archivo de probabilidades conjuntas personalizado.
- Pulse en Continuar.

Figura 9-1
Cuadro de diálogo Modelo lineal general de muestras complejas

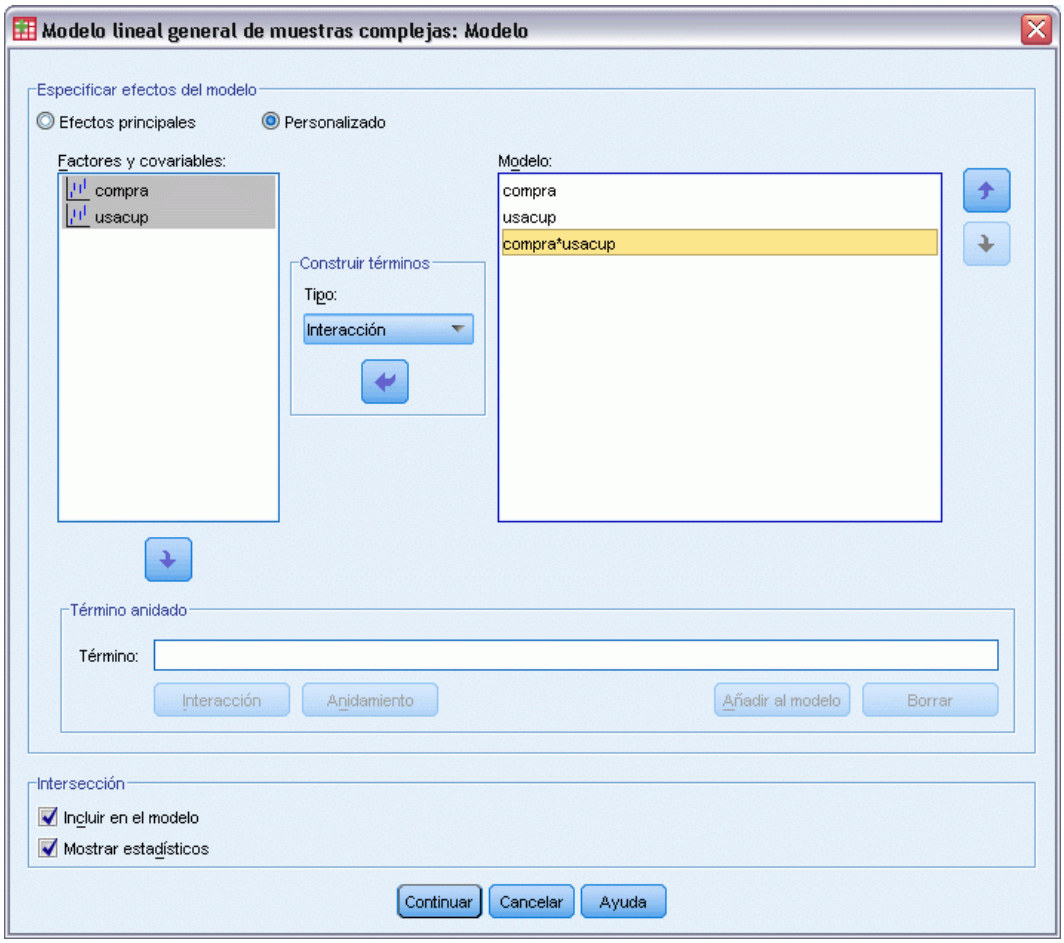


► Seleccione una variable dependiente.

Si lo desea, puede:

- Seleccione variables para factores y covariables, según corresponda a los datos.
- Especifique una variable para definir una subpoblación. El análisis se lleva a cabo únicamente en la categoría seleccionada de la variable de subpoblación.

Figura 9-2
Cuadro de diálogo Modelo



Especificar efectos del modelo. Por defecto, el procedimiento crea un modelo de efectos principales utilizando los factores y las covariables especificadas en el cuadro de diálogo principal. Si lo desea, también puede crear un modelo personalizado que contenga los efectos de la interacción y los términos anidados.

Términos no anidados

Para las covariables y los factores seleccionados:

Interacción. Crea el término de interacción de mayor nivel para todas las variables seleccionadas.

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

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

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

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

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

Términos anidados

En este procedimiento, puede construir términos anidados para el modelo. Los términos anidados resultan útiles para modelar el efecto de un factor o covariable cuyos valores no interactúan con los niveles de otro factor. Por ejemplo, una cadena de tiendas de comestibles desea realizar un seguimiento de los hábitos de gasto de los clientes en las diversas ubicaciones de sus tiendas. Dado que cada cliente frecuenta tan sólo una de estas ubicaciones, se puede decir que el efecto de *Cliente* está **anidado dentro** del efecto de *Ubicación de la tienda*.

Además, puede incluir efectos de interacción, como términos polinómicos que implican a la misma covariable, o añadir varios niveles de anidación al término anidado.

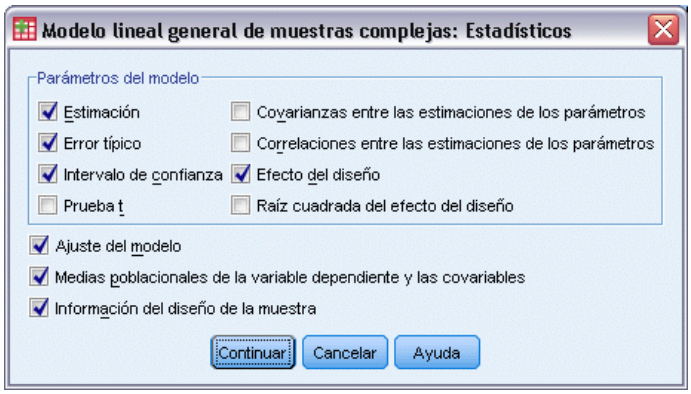
Limitaciones. Existen las siguientes restricciones para los términos anidados:

- Todos los factores incluidos en una interacción deben ser exclusivos entre sí. Por consiguiente, si *A* es un factor, no es válido especificar *A*A*.
- Todos los factores incluidos en un efecto anidado deben ser exclusivos entre sí. Por consiguiente, si *A* es un factor, no es válido especificar *A(A)*.
- No se puede anidar ningún efecto dentro de una covariable. Por consiguiente, si *A* es un factor y *X* es una covariable, no es válido especificar *A(X)*.

Intersección. La intersección se incluye normalmente en el modelo. Si asume que los datos pasan por el origen, puede excluir la intersección. Incluso aunque incluya la intersección en el modelo, puede suprimir los estadísticos relacionados con ella.

Estadísticos de Modelo lineal general de muestras complejas

Figura 9-3
Cuadro de diálogo Modelo lineal general de muestras complejas: Estadísticos



Parámetros del modelo. Este grupo permite controlar la presentación de estadísticos relacionados con los parámetros del modelo.

- **Estimación.** Muestra estimaciones de los coeficientes.
- **Error típico.** Muestra el error típico de cada estimación de los coeficientes.
- **Intervalo de confianza.** Muestra un intervalo de confianza para cada estimación de los coeficientes. El nivel de confianza de los intervalos se configura en el cuadro de diálogo Opciones.

- **Prueba t.** Muestra una prueba *t* de cada estimación de coeficientes. La hipótesis nula de cada prueba es que el valor del coeficiente sea 0.
- **Covarianzas de las estimaciones de los parámetros.** Muestra una estimación de la matriz de covarianzas de los coeficientes del modelo.
- **Correlaciones de las estimaciones de los parámetros.** Muestra una estimación de la matriz de correlaciones de los coeficientes del modelo.
- **Efecto del diseño.** Cociente de la variación de la estimación entre la variación obtenida al suponer que la muestra es una muestra aleatoria simple. Es una medida del efecto de especificar un diseño complejo donde los valores más distantes de 1 indican efectos mayores.
- **Raíz cuadrada del efecto del diseño.** Es una medida del efecto de especificar un diseño complejo donde los valores más distantes de 1 indican efectos mayores.

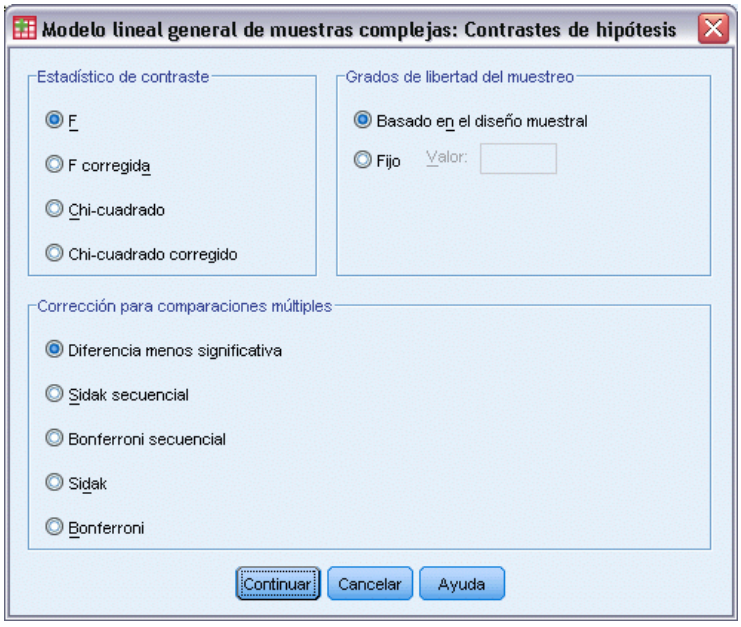
Ajuste del modelo. Muestra R^2 y estadísticos de error cuadrático medio.

Medias de población de covariables y variables dependientes. Muestra información resumida acerca de los factores, las covariables y las variables dependientes.

Información del diseño muestral. Muestra información resumida acerca de la muestra, incluidos un recuento no ponderado y el tamaño de la población.

Muestras complejas: Contrastes de hipótesis

Figura 9-4
Cuadro de diálogo Contrastes de hipótesis



Estadístico de contraste. Este grupo le permite seleccionar el tipo de estadístico utilizado para contrastar las hipótesis. Es posible elegir entre *F*, *F* corregida, chi-cuadrado y chi-cuadrado corregido.

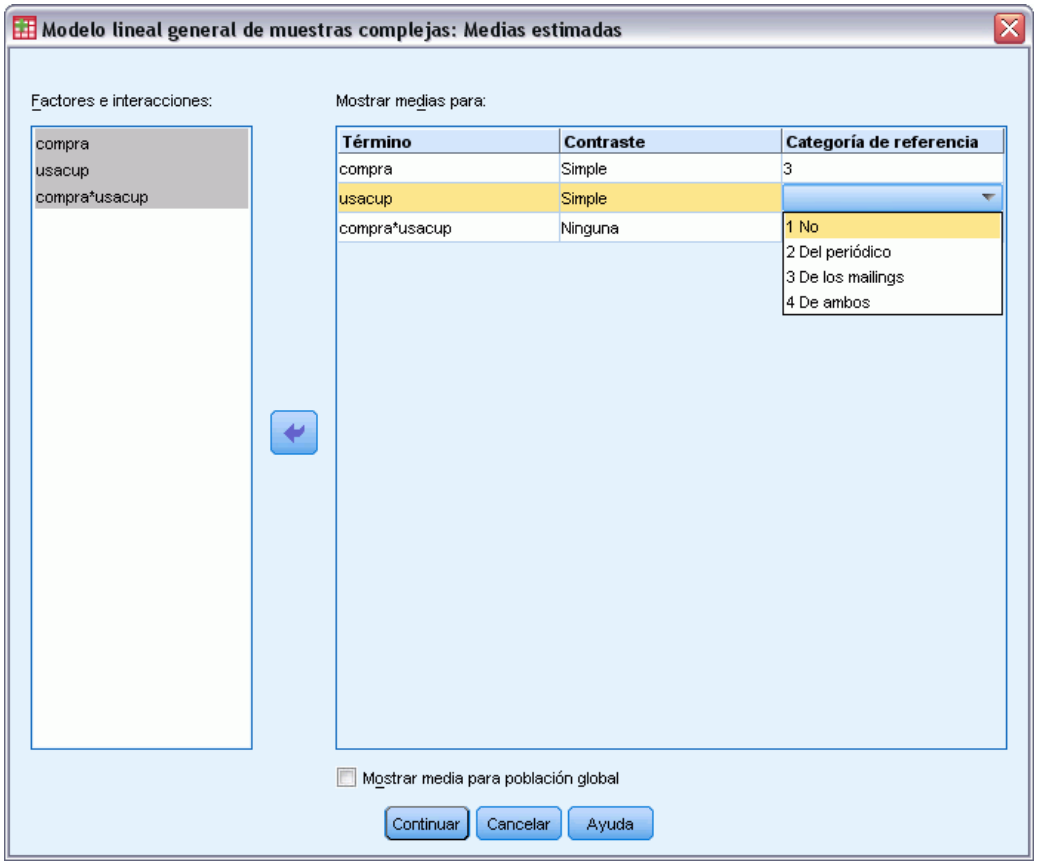
Muestreo de grados de libertad. Este grupo permite controlar los grados de libertad en el diseño de muestra usados para calcular los valores *p* de todos los estadísticos de contraste. Si se basa en el diseño muestral, el valor es la diferencia entre el número de unidades de muestra primarias y el número de estratos de la primera etapa del muestreo. Si lo desea, puede especificar los grados de libertad que desee introduciendo un número entero positivo.

Corrección para comparaciones múltiples. Al realizar contrastes de hipótesis con varios contrastes, el nivel de significación global se puede ajustar utilizando los niveles de significación de los contrastes incluidos. Este grupo permite elegir el método de ajuste.

- **Diferencia menos significativa.** Este método no controla la probabilidad general de rechazar las hipótesis de que algunos contrastes lineales son diferentes a los valores de hipótesis nula.
- **Sidak secuencial.** Este es un procedimiento de Sidak de rechazo secuencial decreciente que es mucho menos conservador en términos de rechazar las hipótesis individuales pero que mantiene el mismo nivel de significación global.
- **Bonferroni secuencial.** Este es un procedimiento de Bonferroni de rechazo secuencial decreciente que es mucho menos conservador en términos de rechazar las hipótesis individuales pero que mantiene el mismo nivel de significación global.
- **Sidak.** Este método ofrece límites más estrechos que los de la aproximación de Bonferroni.
- **Bonferroni.** Este método corrige el nivel de significación observado por el hecho de que se están poniendo a prueba múltiples contrastes.

Medias estimadas del Modelo lineal general de muestras complejas

Figura 9-5
Cuadro de diálogo Modelo lineal general de muestras complejas: Medias estimadas



En el cuadro de diálogo Medias estimadas se pueden ver las medias marginales estimadas por el modelo para los niveles de factores y las interacciones de factores especificadas en el subcuadro de diálogo Modelo. También se puede solicitar que se muestre la media de población global.

Término. Se calculan las medias estimadas de los factores seleccionados y las interacciones de los factores.

Contraste. El contraste determina como se configuran los contrastes de hipótesis para comparar las medias estimadas.

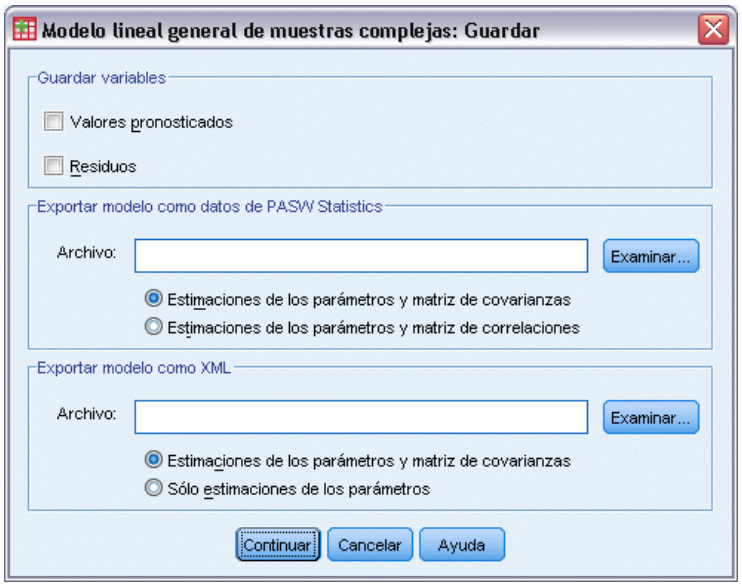
- **Simple.** Compara la media de cada nivel con la media de un nivel especificado. Este tipo de contraste resulta útil cuando existe un grupo de control.
- **Desviación.** Compara la media de cada nivel (excepto una categoría de referencia) con la media de todos los niveles (media global). Los niveles del factor pueden colocarse en cualquier orden.
- **Diferencia.** Compara la media de cada nivel (excepto el primero) con la media de los niveles anteriores. En ocasiones se les denomina contrastes de Helmert invertidos.
- **Helmert.** Compara la media de cada nivel del factor (excepto el último) con la media de los niveles siguientes.

- **Repetido.** 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.

Categoría de referencia. Los contrastes simple y de desviación requieren una categoría de referencia o un factor de nivel con que comparar los demás.

Modelo lineal general de muestras complejas: Guardar

Figura 9-6
Cuadro de diálogo Modelo lineal general de muestras complejas: Guardar



Guardar variables. Este grupo permite guardar los valores pronosticados para el modelo y los residuos como nuevas variables en el archivo de trabajo.

Exportar modelo como datos de SPSS Statistics. Escribe un conjunto de datos de formato IBM® SPSS® Statisticsque contiene la matriz de covarianzas o correlaciones de los parámetros con las estimaciones de los parámetros, errores típicos, valores de significación y grados de libertad. El orden de las variables en el archivo matricial es el siguiente.

- **rowtype_.** Toma los valores (y las etiquetas de valor), COV (covarianzas), CORR (correlaciones), EST (estimaciones de los parámetros), SE (errores típicos), SIG (niveles de significación) y DF (grados de libertad del diseño muestral). Hay un caso diferente con el tipo de fila COV (o CORR) para cada parámetro del modelo, además de un caso diferente para cada uno de los otros tipos de filas.

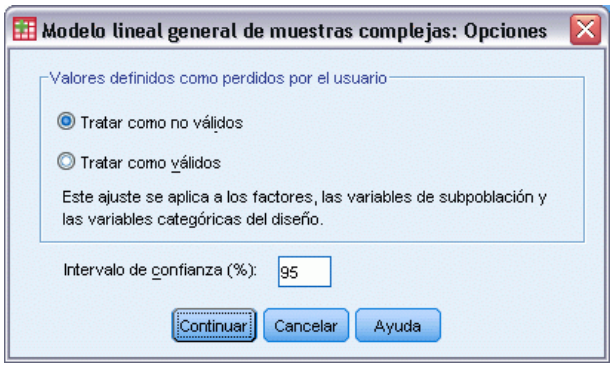
- **varname_.** Toma los valores P1, P2, ..., correspondientes a una lista ordenada de todos los parámetros del modelo para los tipos de fila COV o CORR, con las etiquetas de valor correspondientes a las cadenas de parámetros mostradas en la tabla de estimaciones de los parámetros. Las casillas están vacías para los demás tipos de filas.
- **P1, P2, ...** Estas variables corresponden a una lista ordenada de todos los parámetros del modelo con las etiquetas de variable correspondientes a las cadenas de parámetros mostradas en la tabla de estimaciones de los parámetros y toman valores según el tipo de fila. Para los parámetros redundantes, todas las covarianzas se establecen en cero, las correlaciones se establecen en el valor perdido del sistema; todas las estimaciones de los parámetros se establecen en cero; y todos los errores típicos, niveles de significación y los grados de libertad residuales se establecen en el valor perdido del sistema.

Nota: Este archivo no se puede utilizar directamente para realizar otros análisis en otros procedimientos que lean un archivo matricial a menos que dichos procedimientos acepten todos los tipos de filas que aquí se exportan.

Exportar modelo como XML. Guarda las estimaciones de los parámetros y la matriz de covarianzas de los parámetros (si se selecciona) en formato XML (PMML). Puede utilizar este archivo de modelo para aplicar la información del modelo a otros archivos de datos para puntuarlo.

Modelo lineal general de muestras complejas: Opciones

Figura 9-7
Cuadro de diálogo Modelo lineal general de muestras complejas: Opciones



Valores definidos como perdidos por el usuario. Todas las variables de diseño, así como la variable dependiente y cualquier covariable, deben contener datos válidos. Los casos con datos no válidos de cualquiera de estas variables se excluyen del análisis. Estos controles permiten decidir si los valores definidos como perdidos por el usuario se deben tratar como válidos entre las variables de estratificación, conglomeración, subpoblación y de factor.

Intervalo de confianza. Se trata del nivel de intervalo de confianza para las estimaciones de coeficiente y las medias marginales estimadas. Especifique un valor mayor o igual a 50 e inferior a 100.

Funciones adicionales del comando CSGLM

La sintaxis de comandos también le permite:

- Especificar contrastes personalizados de los efectos respecto a una combinación lineal de efectos o un valor (utilizando el subcomando `CUSTOM`).
- Fijar covariables en valores distintos los de sus medias al calcular las medias marginales estimadas (utilizando el subcomando `EMMEANS`).
- Especificar una métrica para los contrastes polinómicos (utilizando el subcomando `EMMEANS`).
- Especificar un valor de tolerancia para la comprobación de la singularidad (utilizando el subcomando `CRITERIA`).
- Crear nombres especificados por el usuario para las variables almacenadas (utilizando el subcomando `SAVE`).
- Generar una tabla de función estimable general (utilizando el subcomando `PRINT`).

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

Capítulo
10

Regresión logística de muestras complejas

El procedimiento Regresión logística de muestras complejas lleva a cabo análisis de regresión logística sobre una variable binaria o una variable dependiente multinomial para muestras extraídas mediante métodos de muestreo complejo. Si lo desea, puede solicitar análisis de una subpoblación.

Ejemplo. Un encargado de préstamos ha recopilado registros antiguos de préstamos concedidos a clientes en diversas ramas, de acuerdo con un diseño complejo. Al incorporar el diseño muestral, el encargado desea comprobar si la probabilidad con que las moras de un cliente se asocian a su edad, historial de empleo y cantidad de crédito adeudado; posteriormente.

Estadísticos. El procedimiento genera estimaciones, estimaciones exponenciadas, errores típicos, intervalos de confianza, pruebas *t*, efectos del diseño, raíz cuadrada de los efectos del diseño para parámetros de modelo y las correlaciones y covarianzas entre las estimaciones de los parámetros. También hay disponibles estadísticos pseudo R^2, tablas de clasificación y estadísticos descriptivos para las variables dependientes e independientes.

Datos. La variable dependiente es categórica. Los factores son categóricos; pueden tener valores numéricos o valores de cadena de hasta ocho caracteres. Las covariables son variables cuantitativas que están relacionadas con la variable dependiente. Las variables que definen las subpoblaciones pueden ser numéricas o de cadena, pero siempre deben ser categóricas.

Supuestos. Los casos del archivo de datos representan una muestra de un diseño complejo que se debe analizar según las especificaciones del archivo seleccionado en el Cuadro de diálogo Plan de muestras complejas.

Obtención de Regresión logística de muestras complejas

Seleccione en los menús:
Analizar > Muestras complejas > Regresión logística...

- ▶ Seleccione un archivo de plan. Si lo desea, elija un archivo de probabilidades conjuntas personalizado.
- ▶ Pulse en Continuar.

Figura 10-1
Cuadro de diálogo Regresión logística

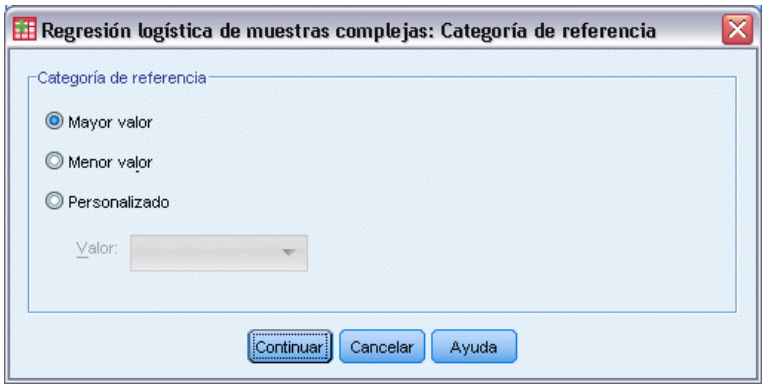
- Seleccione una variable dependiente.

Si lo desea, puede:

- Seleccione variables para factores y covariables, según corresponda a los datos.
- Especifique una variable para definir una subpoblación. El análisis se lleva a cabo únicamente en la categoría seleccionada de la variable de subpoblación.

Regresión logística de muestras complejas: Categoría de referencia

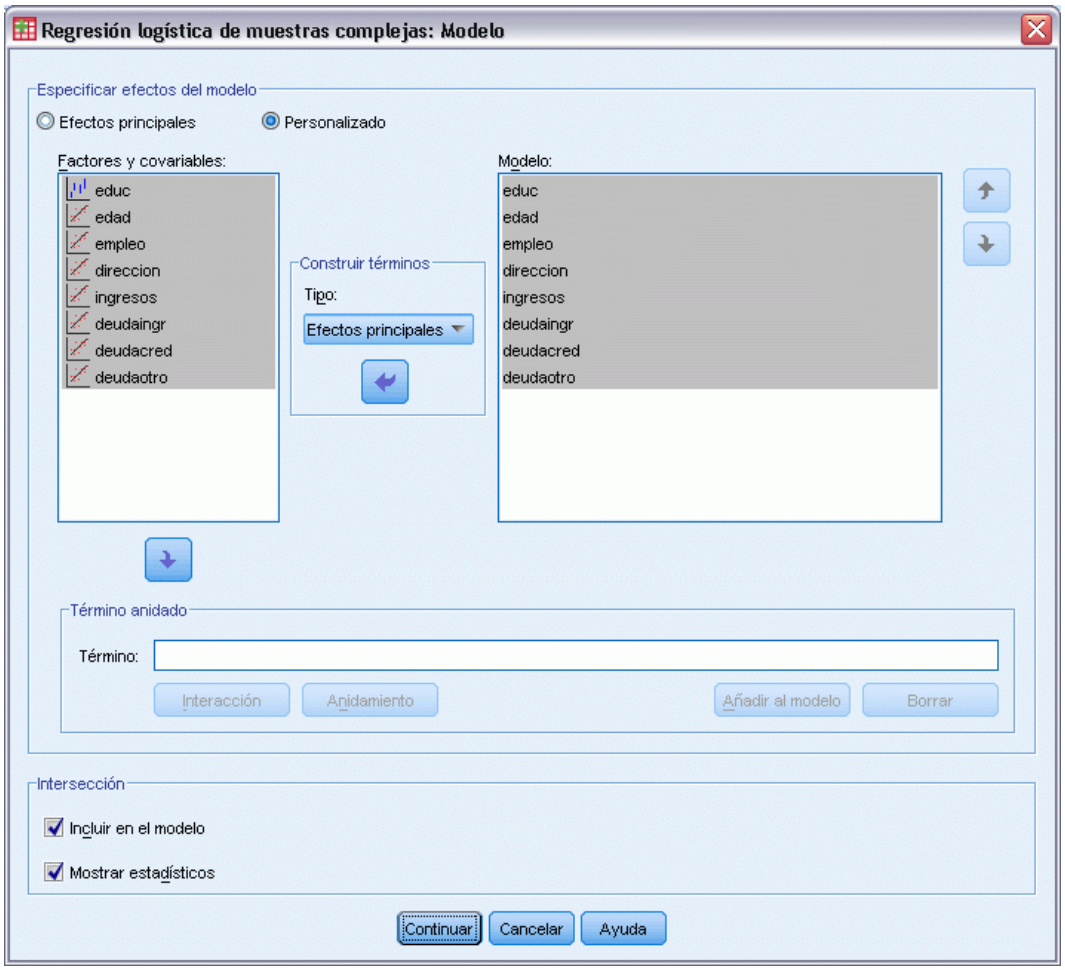
Figura 10-2
Cuadro de diálogo Regresión logística de muestras complejas: Categoría de referencia



Por defecto, el procedimiento Regresión logística de muestras complejas hace de la categoría con el valor más alto la categoría de referencia. Este cuadro de diálogo permite especificar el valor más alto, el valor más bajo o una categoría personalizada como la categoría de referencia.

Regresión logística de muestras complejas: Modelo

Figura 10-3
Cuadro de diálogo Regresión logística de muestras complejas



Especificar efectos del modelo. Por defecto, el procedimiento crea un modelo de efectos principales utilizando los factores y las covariables especificadas en el cuadro de diálogo principal. Si lo desea, también puede crear un modelo personalizado que contenga los efectos de la interacción y los términos anidados.

Términos no anidados

Para las covariables y los factores seleccionados:

Interacción. Crea el término de interacción de mayor nivel para todas las variables seleccionadas.

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

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

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

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

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

Términos anidados

En este procedimiento, puede construir términos anidados para el modelo. Los términos anidados resultan útiles para modelar el efecto de un factor o covariable cuyos valores no interactúan con los niveles de otro factor. Por ejemplo, una cadena de tiendas de comestibles desea realizar un seguimiento de los hábitos de gasto de los clientes en las diversas ubicaciones de sus tiendas. Dado que cada cliente frecuenta tan sólo una de estas ubicaciones, se puede decir que el efecto de *Cliente* está **anidado dentro** del efecto de *Ubicación de la tienda*.

Además, puede incluir efectos de interacción, como términos polinómicos que implican a la misma covariable, o añadir varios niveles de anidación al término anidado.

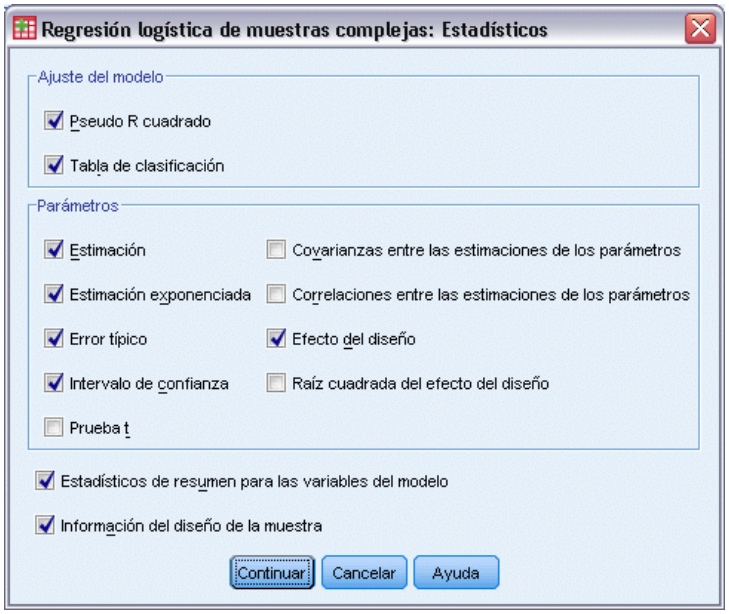
Limitaciones. Existen las siguientes restricciones para los términos anidados:

- Todos los factores incluidos en una interacción deben ser exclusivos entre sí. Por consiguiente, si *A* es un factor, no es válido especificar *A*A*.
- Todos los factores incluidos en un efecto anidado deben ser exclusivos entre sí. Por consiguiente, si *A* es un factor, no es válido especificar *A(A)*.
- No se puede anidar ningún efecto dentro de una covariable. Por consiguiente, si *A* es un factor y *X* es una covariable, no es válido especificar *A(X)*.

Intersección. La intersección se incluye normalmente en el modelo. Si asume que los datos pasan por el origen, puede excluir la intersección. Incluso aunque incluya la intersección en el modelo, puede suprimir los estadísticos relacionados con ella.

Regresión logística de muestras complejas: Estadísticos

Figura 10-4
Cuadro de diálogo Regresión logística: Estadísticos



Ajuste del modelo. Controla la presentación de estadísticos que miden el rendimiento global del proceso.

- **Pseudo R cuadrado.** El estadístico R^2 de regresión lineal no cuenta con un análogo exacto entre los modelos de regresión logística. En su lugar existen varias medidas que tratan de imitar las propiedades del estadístico R^2.
- **Tabla de clasificación.** Muestra las clasificaciones conjuntas tabuladas de la categoría observada por la categoría pronosticada por el modelo en la variable dependiente.

Parámetros. Este grupo permite controlar la presentación de estadísticos relacionados con los parámetros del modelo.

- **Estimación.** Muestra estimaciones de los coeficientes.
- **Estimación exponenciada.** Muestra la base del logaritmo natural elevada a la potencia de las estimaciones de los coeficientes. Mientras que las estimaciones tienen propiedades agradables para la comprobación estadística, la estimación exponenciada (o exp[B]) es más sencilla de interpretar.
- **Error típico.** Muestra el error típico de cada estimación de los coeficientes.
- **Intervalo de confianza.** Muestra un intervalo de confianza para cada estimación de los coeficientes. El nivel de confianza de los intervalos se configura en el cuadro de diálogo Opciones.
- **Prueba t.** Muestra una prueba *t* de cada estimación de coeficientes. La hipótesis nula de cada prueba es que el valor del coeficiente sea 0.
- **Covarianzas de las estimaciones de los parámetros.** Muestra una estimación de la matriz de covarianzas de los coeficientes del modelo.

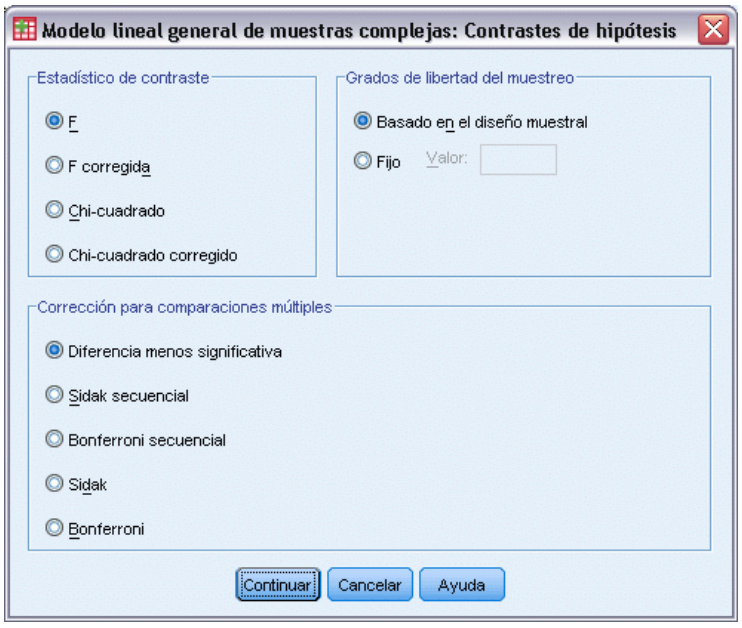
- **Correlaciones de las estimaciones de los parámetros.** Muestra una estimación de la matriz de correlaciones de los coeficientes del modelo.
- **Efecto del diseño.** Cociente de la variación de la estimación entre la variación obtenida al suponer que la muestra es una muestra aleatoria simple. Es una medida del efecto de especificar un diseño complejo donde los valores más distantes de 1 indican efectos mayores.
- **Raíz cuadrada del efecto del diseño.** Es una medida del efecto de especificar un diseño complejo donde los valores más distantes de 1 indican efectos mayores.

Estadísticos de resumen para las variables del modelo. Muestra información resumida acerca de los factores, las covariables y las variables dependientes.

Información del diseño muestral. Muestra información resumida acerca de la muestra, incluidos un recuento no ponderado y el tamaño de la población.

Muestras complejas: Contrastes de hipótesis

Figura 10-5
Cuadro de diálogo Contrastes de hipótesis



Estadístico de contraste. Este grupo le permite seleccionar el tipo de estadístico utilizado para contrastar las hipótesis. Es posible elegir entre *F*, *F* corregida, chi-cuadrado y chi-cuadrado corregido.

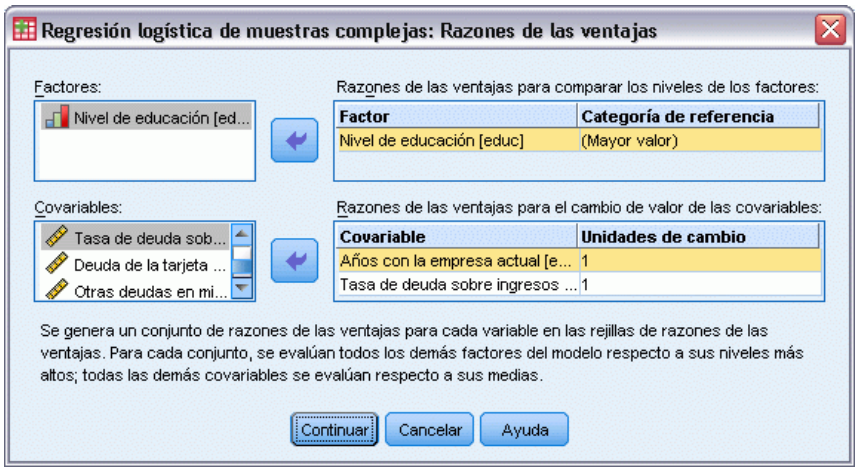
Muestreo de grados de libertad. Este grupo permite controlar los grados de libertad en el diseño de muestra usados para calcular los valores *p* de todos los estadísticos de contraste. Si se basa en el diseño muestral, el valor es la diferencia entre el número de unidades de muestra primarias y el número de estratos de la primera etapa del muestreo. Si lo desea, puede especificar los grados de libertad que desee introduciendo un número entero positivo.

Corrección para comparaciones múltiples. Al realizar contrastes de hipótesis con varios contrastes, el nivel de significación global se puede ajustar utilizando los niveles de significación de los contrastes incluidos. Este grupo permite elegir el método de ajuste.

- **Diferencia menos significativa.** Este método no controla la probabilidad general de rechazar las hipótesis de que algunos contrastes lineales son diferentes a los valores de hipótesis nula.
- **Sidak secuencial.** Este es un procedimiento de Sidak de rechazo secuencial decreciente que es mucho menos conservador en términos de rechazar las hipótesis individuales pero que mantiene el mismo nivel de significación global.
- **Bonferroni secuencial.** Este es un procedimiento de Bonferroni de rechazo secuencial decreciente que es mucho menos conservador en términos de rechazar las hipótesis individuales pero que mantiene el mismo nivel de significación global.
- **Sidak.** Este método ofrece límites más estrechos que los de la aproximación de Bonferroni.
- **Bonferroni.** Este método corrige el nivel de significación observado por el hecho de que se están poniendo a prueba múltiples contrastes.

Regresión logística de muestras complejas: Razones de las ventajas

Figura 10-6
Cuadro de diálogo Regresión logística de muestras complejas: Razones de las ventajas



El cuadro de diálogo Razones de las ventajas permite mostrar las razones de las ventajas estimadas por el modelo para los factores y las covariables que se especifican. Se calcula un conjunto independiente de razones de las ventajas para cada categoría de la variable dependiente excepto para el caso de la categoría de referencia.

Factores. En cada factor seleccionado, muestra la razón de las ventajas de cada categoría del factor hasta las ventajas en la categoría de referencia especificada.

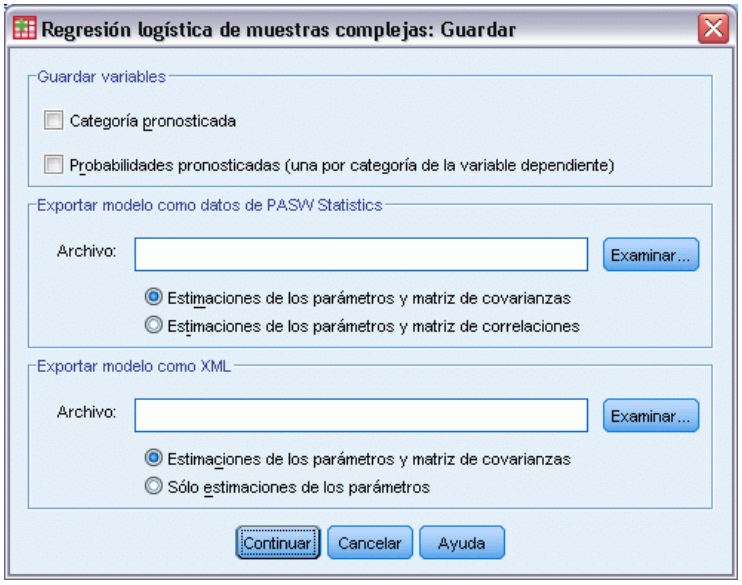
Covariables. En cada covariable seleccionada, muestra la razón de las ventajas en el valor medio de la covariable más las unidades de cambio especificadas para las ventajas de la media.

Al calcular las razones de las ventajas de un factor o una covariable, el procedimiento fija todos los demás factores en sus niveles más altos y el resto de covariables, en sus niveles medios. Si un factor o una covariable interactúan con otros predictores en el modelo, las razones de las ventajas

dependerán no sólo de la modificación en la variable especificada, sino también de los valores de las variables con las que interactúe. Si una covariable especificada interactúa consigo misma en el modelo (por ejemplo, *edad*edad*), las razones de las ventajas dependerán entonces tanto del cambio en la covariable como del valor de ésta.

Regresión logística de muestras complejas: Guardar

Figura 10-7
Cuadro de diálogo Regresión logística de muestras complejas: Guardar



Guardar variables. Este grupo permite guardar la categoría pronosticada para el modelo y las probabilidades pronosticadas como nuevas variables en el conjunto de datos activo.

Exportar modelo como datos de SPSS Statistics. Escribe un conjunto de datos de formato IBM® SPSS® Statisticsque contiene la matriz de covarianzas o correlaciones de los parámetros con las estimaciones de los parámetros, errores típicos, valores de significación y grados de libertad. El orden de las variables en el archivo matricial es el siguiente.

- **rowtype_.** Toma los valores (y las etiquetas de valor), COV (covarianzas), CORR (correlaciones), EST (estimaciones de los parámetros), SE (errores típicos), SIG (niveles de significación) y DF (grados de libertad del diseño muestral). Hay un caso diferente con el tipo de fila COV (o CORR) para cada parámetro del modelo, además de un caso diferente para cada uno de los otros tipos de filas.
- **varname_.** Toma los valores P1, P2, ..., correspondientes a una lista ordenada de todos los parámetros del modelo para los tipos de fila COV o CORR, con las etiquetas de valor correspondientes a las cadenas de parámetros mostradas en la tabla de estimaciones de los parámetros. Las casillas están vacías para los demás tipos de filas.
- **P1, P2, ...** Estas variables corresponden a una lista ordenada de todos los parámetros del modelo con las etiquetas de variable correspondientes a las cadenas de parámetros mostradas en la tabla de estimaciones de los parámetros y toman valores según el tipo de fila. Para los parámetros redundantes, todas las covarianzas se establecen en cero, las correlaciones

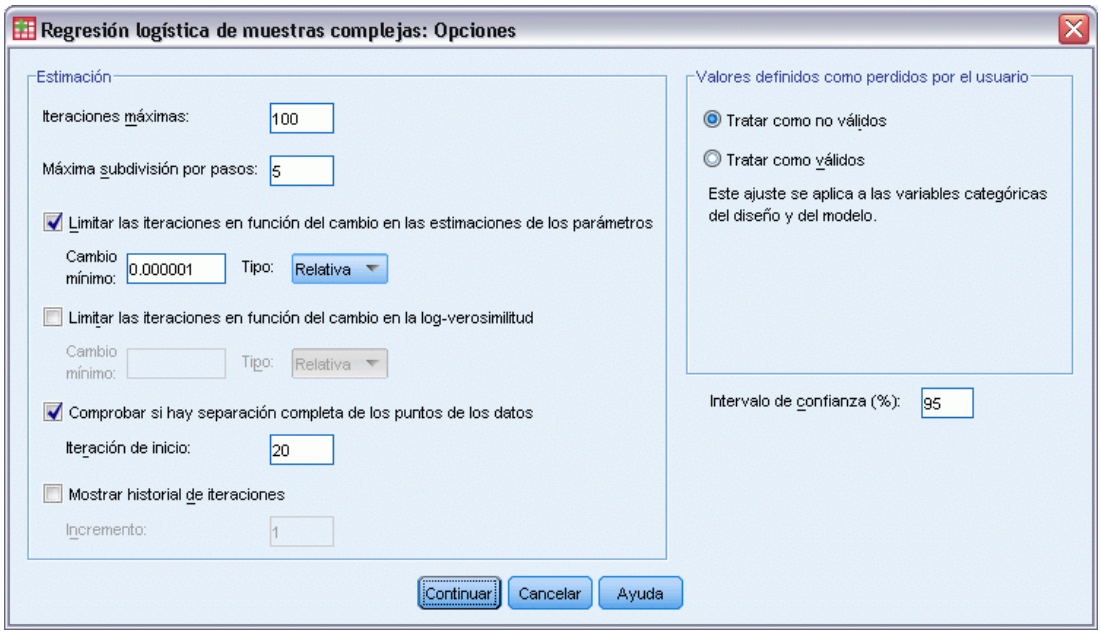
se establecen en el valor perdido del sistema; todas las estimaciones de los parámetros se establecen en cero; y todos los errores típicos, niveles de significación y los grados de libertad residuales se establecen en el valor perdido del sistema.

Nota: Este archivo no se puede utilizar directamente para realizar otros análisis en otros procedimientos que lean un archivo matricial a menos que dichos procedimientos acepten todos los tipos de filas que aquí se exportan.

Exportar modelo como XML. Guarda las estimaciones de los parámetros y la matriz de covarianzas de los parámetros (si se selecciona) en formato XML (PMML). Puede utilizar este archivo de modelo para aplicar la información del modelo a otros archivos de datos para puntuarlo.

Regresión logística de muestras complejas: Opciones

Figura 10-8
Cuadro de diálogo Regresión logística: Opciones



Estimación. Este grupo otorga el control sobre varios criterios utilizados en la estimación del modelo.

- **Nº máximo de iteraciones.** Número máximo de iteraciones que se ejecutará el algoritmo. Especifique un número entero no negativo.
- **Máxima subdivisión por pasos.** En cada iteración, se reduce el tamaño del paso mediante un factor de 0,5 hasta que aumenta el logaritmo de la verosimilitud o se alcanza la máxima subdivisión por pasos. Especifique un número entero positivo.
- **Limitar las iteraciones en función del cambio en las estimaciones de los parámetros.** Si se activa, el algoritmo se detiene tras una iteración en la que las modificaciones absolutas o relativas en las estimaciones de los parámetros sean inferiores que el valor especificado, que debe ser no negativo.

- **Limitar las iteraciones en función del cambio en la log-verosimilitud.** Si se activa, el algoritmo se detiene tras una iteración en la que las modificaciones absolutas o relativas en la función de log-verosimilitud sean inferiores que el valor especificado, que debe ser no negativo.
- **Comprobar si hay separación completa de los puntos de los datos.** Si se activa, el algoritmo realiza una prueba para garantizar que las estimaciones de los parámetros tienen valores exclusivos. Se produce una separación cuando el procedimiento pueda generar un modelo que clasifique cada caso de forma correcta.
- **Mostrar historial de iteraciones.** Muestra los estadísticos y las estimaciones de los parámetros cada *n* iteraciones, comenzando por la iteración 0 (estimaciones iniciales). Si decide imprimir el historial de iteraciones, la última iteración se imprimirá siempre independientemente del valor de *n*.

Valores definidos como perdidos por el usuario. Todas las variables de diseño, así como la variable dependiente y cualquier covariable, deben contener datos válidos. Los casos con datos no válidos de cualquiera de estas variables se excluyen del análisis. Estos controles permiten decidir si los valores definidos como perdidos por el usuario se deben tratar como válidos entre las variables de estratificación, conglomeración, subpoblación y de factor.

Intervalo de confianza. Se trata del nivel de intervalo de confianza para las estimaciones de coeficiente, las estimaciones de coeficiente exponenciadas y las razones de las ventajas. Especifique un valor mayor o igual a 50 e inferior a 100.

Funciones adicionales del comando CSLOGISTIC

La sintaxis de comandos también le permite:

- Especificar contrastes personalizados de los efectos respecto a una combinación lineal de efectos o un valor (utilizando el subcomando `CUSTOM`).
- Fijar valores de otras variables de modelo al calcular las razones de las ventajas para factores y covariables (utilizando el subcomando `ODDSRATIOS`).
- Especificar un valor de tolerancia para la comprobación de la singularidad (utilizando el subcomando `CRITERIA`).
- Crear nombres especificados por el usuario para las variables almacenadas (utilizando el subcomando `SAVE`).
- Generar una tabla de función estimable general (utilizando el subcomando `PRINT`).

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

Capítulo
11

Regresión ordinal de muestras complejas

El procedimiento Regresión ordinal de muestras complejas realiza análisis de regresión sobre una variable binaria o una variable dependiente ordinal para muestras extraídas mediante métodos de muestreo complejo. Si lo desea, puede solicitar análisis de una subpoblación.

Ejemplo. Los diputados que estudian un proyecto de ley antes de una asamblea legislativa se interesan por conocer si la opinión pública apoya dicho proyecto de ley y qué relación guarda dicho apoyo con los datos demográficos de los votantes. Los encuestadores diseñan entrevistas y las realizan siguiendo un diseño muestral complejo. Utilice la regresión ordinal de muestras complejas para ajustar un modelo acerca del nivel de apoyo a la ley de acuerdo en los datos demográficos de los votantes.

Datos. La variable dependiente es ordinal. Los factores son categóricos; pueden tener valores numéricos o valores de cadena de hasta ocho caracteres. Las covariables son variables cuantitativas que están relacionadas con la variable dependiente. Las variables que definen las subpoblaciones pueden ser numéricas o de cadena, pero siempre deben ser categóricas.

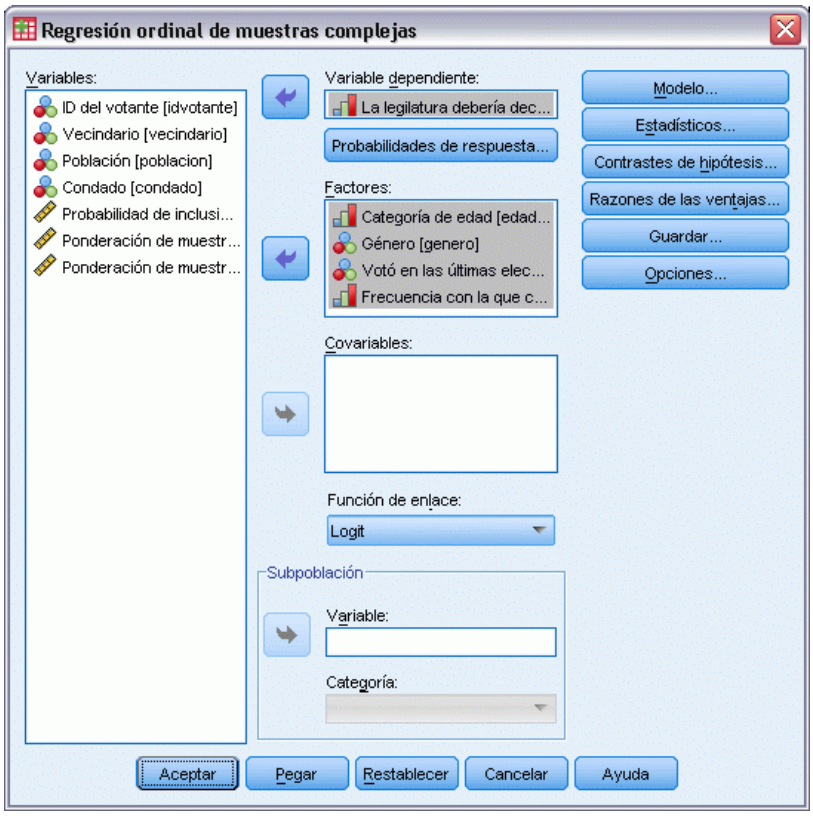
Supuestos. Los casos del archivo de datos representan una muestra de un diseño complejo que se debe analizar según las especificaciones del archivo seleccionado en el Cuadro de diálogo Plan de muestras complejas.

Obtención de regresión ordinal de muestras complejas

Seleccione en los menús:
Analizar > Muestras complejas > Regresión ordinal...

- ▶ Seleccione un archivo de plan. Si lo desea, elija un archivo de probabilidades conjuntas personalizado.
- ▶ Pulse en Continuar.

Figura 11-1
Cuadro de diálogo Regresión ordinal



▶ Seleccione una variable dependiente.

Si lo desea, puede:

- Seleccione variables para factores y covariables, según corresponda a los datos.
- Especifique una variable para definir una subpoblación. El análisis se realiza únicamente para la categoría seleccionada de la variable de subpoblación, aunque para la estimación correcta de las varianzas sigue siendo necesario basarse en el conjunto de datos completo.
- Seleccione una función de enlace.

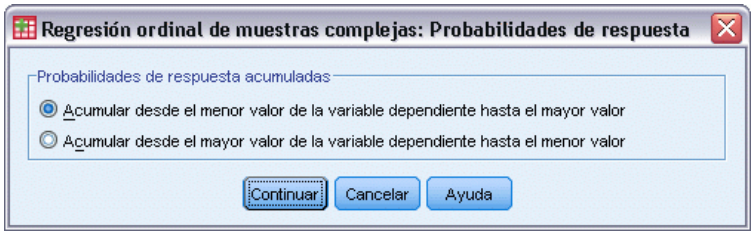
Función de enlace. La función de enlace es una transformación de las probabilidades acumuladas que permiten la estimación del modelo. Existen cinco funciones de enlace que se resumen en la siguiente tabla.

Función	Fórmula	Aplicación típica
Logit	$\log(\xi / (1-\xi))$	Categorías distribuidas de forma uniforme
Log-log complementario	$\log(-\log(1-\xi))$	Categorías más altas más probables
Log-log negativo	$-\log(-\log(\xi))$	Categorías más bajas más probables

Función	Fórmula	Aplicación típica
Probit	$\Phi^{-1}(\xi)$	La variable latente sigue una distribución normal
Cauchit (Cauchy inversa)	$\tan(\pi(\xi-0,5))$	La variable latente tiene muchos valores extremos

Regresión ordinal de muestras complejas: Probabilidades de respuesta

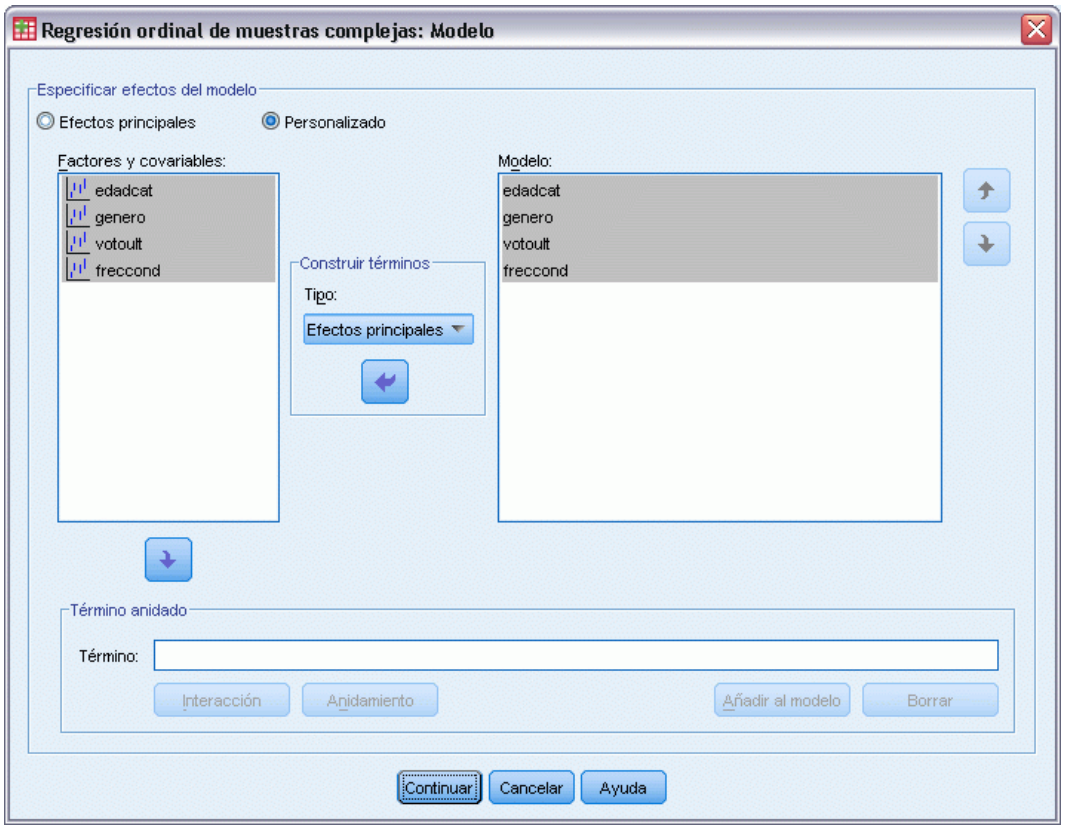
Figura 11-2
Cuadro de diálogo Regresión ordinal de muestras complejas: Probabilidades de respuesta



El cuadro de diálogo Probabilidades de respuesta permite especificar si la probabilidad acumulada de una respuesta (es decir, la probabilidad de pertenecer hasta una determinada categoría, incluida la propia categoría, de la variable dependiente) aumenta con valores de que aumentan o disminuyen de la variable dependiente.

Regresión ordinal de muestras complejas: Modelo

Figura 11-3
Cuadro de diálogo Regresión ordinal de muestras complejas: Modelo



Especificar efectos del modelo. Por defecto, el procedimiento crea un modelo de efectos principales utilizando los factores y las covariables especificadas en el cuadro de diálogo principal. Si lo desea, también puede crear un modelo personalizado que contenga los efectos de la interacción y los términos anidados.

Términos no anidados

Para las covariables y los factores seleccionados:

Interacción. Crea el término de interacción de mayor nivel para todas las variables seleccionadas.

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

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

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

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

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

Términos anidados

En este procedimiento, puede construir términos anidados para el modelo. Los términos anidados resultan útiles para modelar el efecto de un factor o covariable cuyos valores no interactúan con los niveles de otro factor. Por ejemplo, una cadena de tiendas de comestibles desea realizar un seguimiento de los hábitos de gasto de los clientes en las diversas ubicaciones de sus tiendas. Dado que cada cliente frecuenta tan sólo una de estas ubicaciones, se puede decir que el efecto de *Cliente* está **anidado dentro** del efecto de *Ubicación de la tienda*.

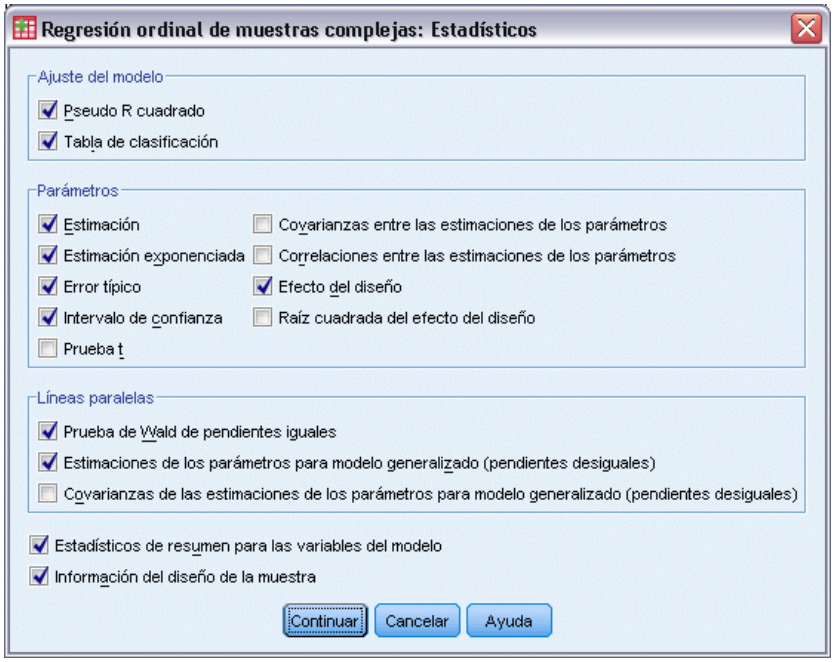
Además, puede incluir efectos de interacción, como términos polinómicos que implican a la misma covariable, o añadir varios niveles de anidación al término anidado.

Limitaciones. Existen las siguientes restricciones para los términos anidados:

- Todos los factores incluidos en una interacción deben ser exclusivos entre sí. Por consiguiente, si *A* es un factor, no es válido especificar *A*A*.
- Todos los factores incluidos en un efecto anidado deben ser exclusivos entre sí. Por consiguiente, si *A* es un factor, no es válido especificar *A(A)*.
- No se puede anidar ningún efecto dentro de una covariable. Por consiguiente, si *A* es un factor y *X* es una covariable, no es válido especificar *A(X)*.

Regresión ordinal de muestras complejas: Estadísticos

Figura 11-4
Cuadro de diálogo Regresión ordinal de muestras complejas: Estadísticos



Ajuste del modelo. Controla la presentación de estadísticos que miden el rendimiento global del proceso.

- **Pseudo R cuadrado.** El estadístico R^2 de regresión lineal no cuenta con un análogo exacto entre los modelos de regresión ordinal. En su lugar existen varias medidas que tratan de imitar las propiedades del estadístico R^2.
- **Tabla de clasificación.** Muestra las clasificaciones conjuntas tabuladas de la categoría observada por la categoría pronosticada por el modelo en la variable dependiente.

Parámetros. Este grupo permite controlar la presentación de estadísticos relacionados con los parámetros del modelo.

- **Estimación.** Muestra estimaciones de los coeficientes.
- **Estimación exponenciada.** Muestra la base del logaritmo natural elevada a la potencia de las estimaciones de los coeficientes. Mientras que las estimaciones tienen propiedades agradables para la comprobación estadística, la estimación exponenciada (o exp[B]) es más sencilla de interpretar.
- **Error típico.** Muestra el error típico de cada estimación de los coeficientes.
- **Intervalo de confianza.** Muestra un intervalo de confianza para cada estimación de los coeficientes. El nivel de confianza de los intervalos se configura en el cuadro de diálogo Opciones.
- **Prueba t.** Muestra una prueba *t* de cada estimación de coeficientes. La hipótesis nula de cada prueba es que el valor del coeficiente sea 0.
- **Covarianzas de las estimaciones de los parámetros.** Muestra una estimación de la matriz de covarianzas de los coeficientes del modelo.
- **Correlaciones de las estimaciones de los parámetros.** Muestra una estimación de la matriz de correlaciones de los coeficientes del modelo.
- **Efecto del diseño.** Cociente de la variación de la estimación entre la variación obtenida al suponer que la muestra es una muestra aleatoria simple. Es una medida del efecto de especificar un diseño complejo donde los valores más distantes de 1 indican efectos mayores.
- **Raíz cuadrada del efecto del diseño.** Es una medida, expresada en unidades y comparable a las de los errores típicos, resultado de especificar un diseño complejo, donde los valores más distantes de 1 indican mayores efectos.

Líneas paralelas. Este grupo permite solicitar estadísticos asociados a un modelo con líneas no paralelas, donde se ajusta una línea de regresión distinta para cada categoría de respuesta (excepto la última).

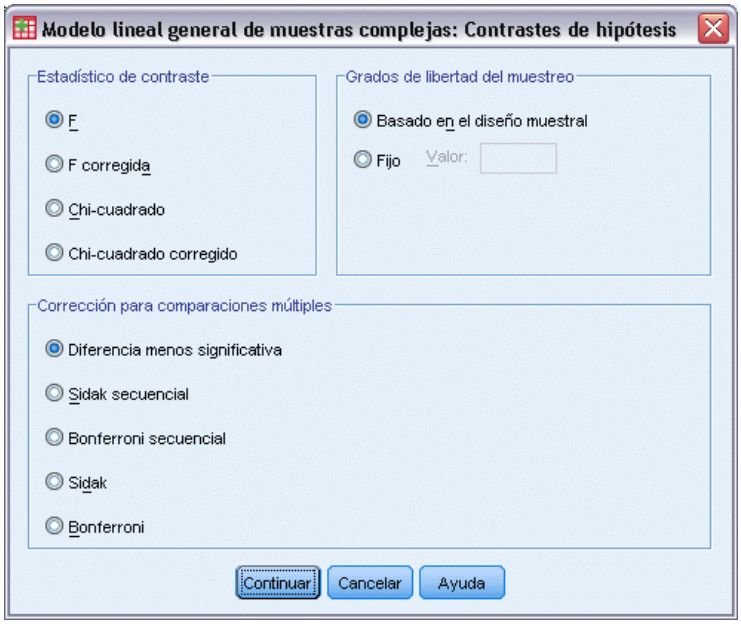
- **Prueba de Wald.** Produce una prueba de la hipótesis nula de que los parámetros de regresión son iguales para todas las respuestas acumuladas. Se estima el modelo con líneas no paralelas y se aplica la prueba de Wald de parámetros iguales.
- **Estimaciones de los parámetros.** Muestra las estimaciones de los coeficientes y errores típicos para el modelo con líneas no paralelas.
- **Covarianzas de las estimaciones de los parámetros.** Muestra una estimación de la matriz de covarianza para los coeficientes del modelo con líneas no paralelas.

Estadísticos de resumen para las variables del modelo. Muestra información resumida acerca de los factores, las covariables y las variables dependientes.

Información del diseño muestral. Muestra información resumida acerca de la muestra, incluidos un recuento no ponderado y el tamaño de la población.

Muestras complejas: Contrastes de hipótesis

Figura 11-5
Cuadro de diálogo Contrastes de hipótesis



Estadístico de contraste. Este grupo le permite seleccionar el tipo de estadístico utilizado para contrastar las hipótesis. Es posible elegir entre *F*, *F* corregida, chi-cuadrado y chi-cuadrado corregido.

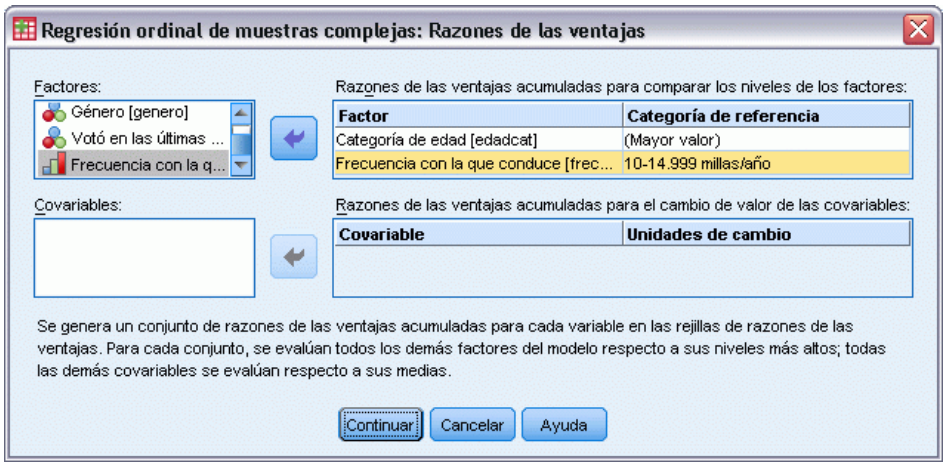
Muestreo de grados de libertad. Este grupo permite controlar los grados de libertad en el diseño de muestra usados para calcular los valores *p* de todos los estadísticos de contraste. Si se basa en el diseño muestral, el valor es la diferencia entre el número de unidades de muestra primarias y el número de estratos de la primera etapa del muestreo. Si lo desea, puede especificar los grados de libertad que desee introduciendo un número entero positivo.

Corrección para comparaciones múltiples. Al realizar contrastes de hipótesis con varios contrastes, el nivel de significación global se puede ajustar utilizando los niveles de significación de los contrastes incluidos. Este grupo permite elegir el método de ajuste.

- **Diferencia menos significativa.** Este método no controla la probabilidad general de rechazar las hipótesis de que algunos contrastes lineales son diferentes a los valores de hipótesis nula.
- **Sidak secuencial.** Este es un procedimiento de Sidak de rechazo secuencial decreciente que es mucho menos conservador en términos de rechazar las hipótesis individuales pero que mantiene el mismo nivel de significación global.
- **Bonferroni secuencial.** Este es un procedimiento de Bonferroni de rechazo secuencial decreciente que es mucho menos conservador en términos de rechazar las hipótesis individuales pero que mantiene el mismo nivel de significación global.
- **Sidak.** Este método ofrece límites más estrechos que los de la aproximación de Bonferroni.
- **Bonferroni.** Este método corrige el nivel de significación observado por el hecho de que se están poniendo a prueba múltiples contrastes.

Regresión ordinal de muestras complejas: Razones de las ventajas

Figura 11-6
Cuadro de diálogo Regresión ordinal de muestras complejas: Razones de las ventajas



El cuadro de diálogo Razones de las ventajas permite mostrar las razones de las ventajas acumuladas estimadas por el modelo para los factores y las covariables que se especifican. Esta característica está disponible únicamente para modelos que utilizan la función de enlace Logit. Se calcula una sola razón de ventajas acumuladas para todas las categorías de la variante dependiente, excepto la última; el modelo de razones proporcionales postula que son todas iguales.

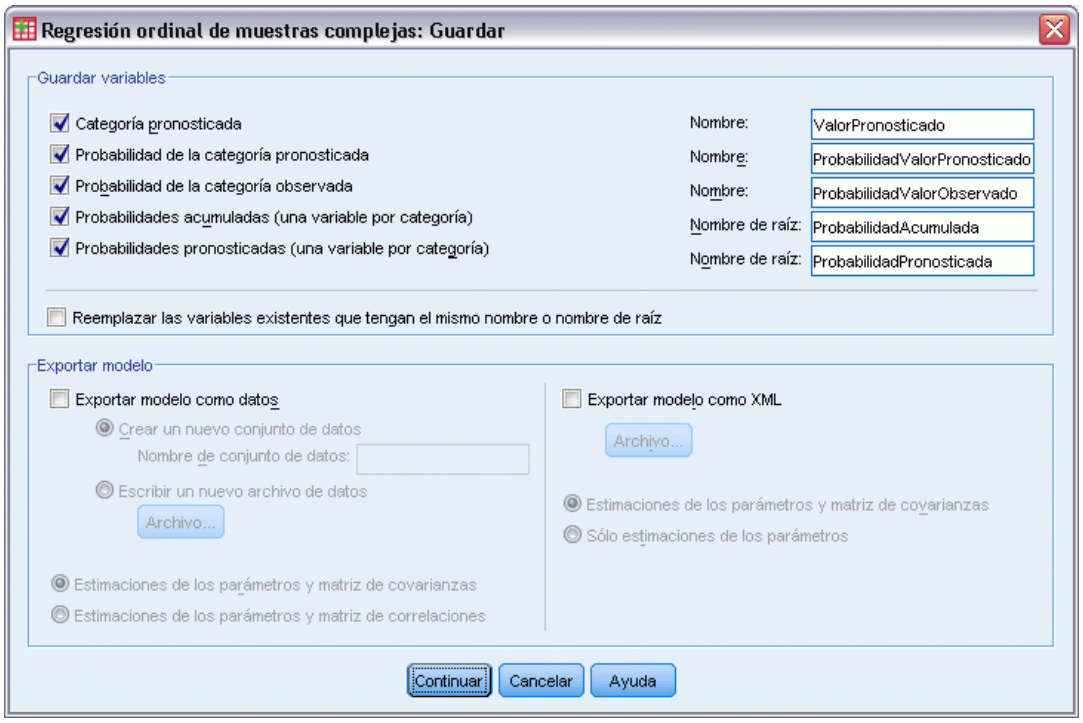
Factores. En cada factor seleccionado, muestra la razón de las ventajas acumuladas de cada categoría del factor hasta las ventajas en la categoría de referencia especificada.

Covariables. En cada covariable seleccionada, muestra la razón de las ventajas acumuladas en el valor medio de la covariable más las unidades de cambio especificadas para las ventajas de la media.

Al calcular las razones de las ventajas de un factor o una covariable, el procedimiento fija todos los demás factores en sus niveles más altos y el resto de covariables, en sus niveles medios. Si un factor o una covariable interactúan con otros predictores en el modelo, las razones de las ventajas dependerán no sólo de la modificación en la variable especificada, sino también de los valores de las variables con las que interactúe. Si una covariable especificada interactúa consigo misma en el modelo (por ejemplo, *edad*edad*), las razones de las ventajas dependerán entonces tanto del cambio en la covariable como del valor de ésta.

Regresión ordinal de muestras complejas: Guardar

Figura 11-7
Cuadro de diálogo Regresión ordinal de muestras complejas: Guardar



Guardar variables. Este grupo permite guardar la categoría pronosticada para el modelo, la probabilidad de la categoría pronosticada, la probabilidad de la categoría observada, las probabilidades acumuladas y las probabilidades pronosticadas como nuevas variables en el conjunto de datos activo.

Exportar modelo como datos de SPSS Statistics. Escribe un conjunto de datos de formato IBM® SPSS® Statisticsque contiene la matriz de covarianzas o correlaciones de los parámetros con las estimaciones de los parámetros, errores típicos, valores de significación y grados de libertad. El orden de las variables en el archivo matricial es el siguiente.

- **rowtype_.** Toma los valores (y las etiquetas de valor), COV (covarianzas), CORR (correlaciones), EST (estimaciones de los parámetros), SE (errores típicos), SIG (niveles de significación) y DF (grados de libertad del diseño muestral). Hay un caso diferente con el tipo de fila COV (o CORR) para cada parámetro del modelo, además de un caso diferente para cada uno de los otros tipos de filas.
- **varname_.** Toma los valores P1, P2, ..., correspondientes a una lista ordenada de todos los parámetros del modelo para los tipos de fila COV o CORR, con las etiquetas de valor correspondientes a las cadenas de parámetros mostradas en la tabla de estimaciones de los parámetros. Las casillas están vacías para los demás tipos de filas.
- **P1, P2, ...** Estas variables corresponden a una lista ordenada de todos los parámetros del modelo con las etiquetas de variable correspondientes a las cadenas de parámetros mostradas en la tabla de estimaciones de los parámetros y toman valores según el tipo de fila. Para

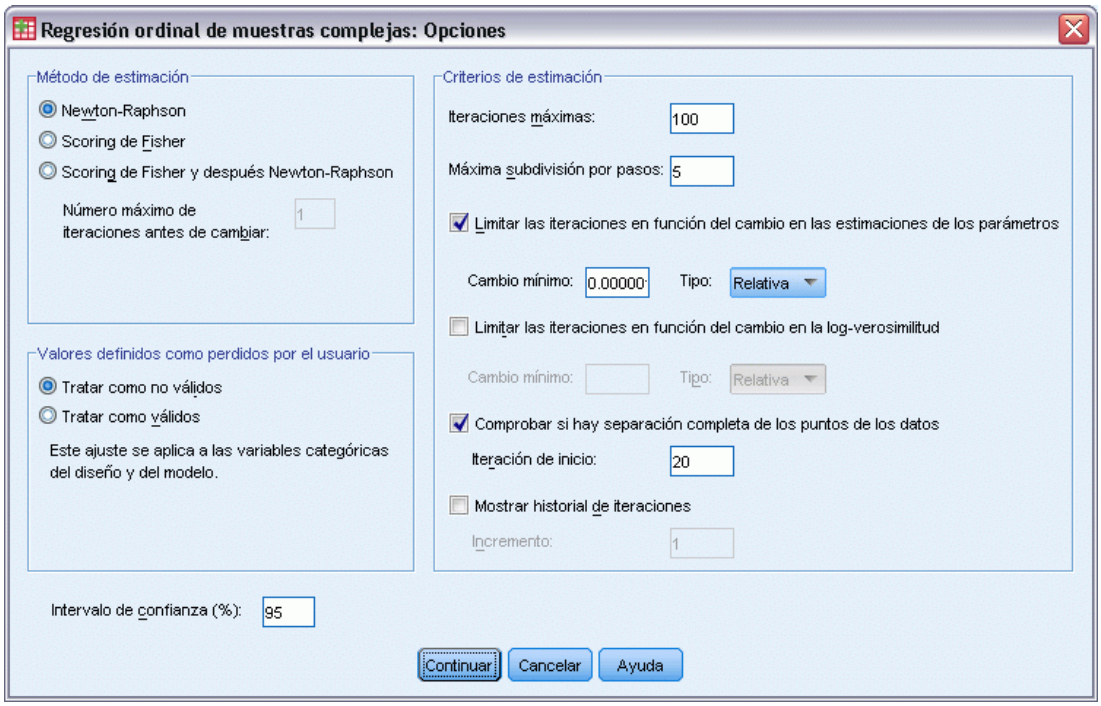
los parámetros redundantes, todas las covarianzas se establecen en cero, las correlaciones se establecen en el valor perdido del sistema; todas las estimaciones de los parámetros se establecen en cero; y todos los errores típicos, niveles de significación y los grados de libertad residuales se establecen en el valor perdido del sistema.

Nota: Este archivo no se puede utilizar directamente para realizar otros análisis en otros procedimientos que lean un archivo matricial a menos que dichos procedimientos acepten todos los tipos de filas que aquí se exportan.

Exportar modelo como XML. Guarda las estimaciones de los parámetros y la matriz de covarianzas de los parámetros (si se selecciona) en formato XML (PMML). Puede utilizar este archivo de modelo para aplicar la información del modelo a otros archivos de datos para puntuarlo.

Regresión ordinal de muestras complejas: Opciones

Figura 11-8
Cuadro de diálogo Regresión ordinal: Opciones



Método de estimación. Puede seleccionar un método de estimación de parámetros. Los métodos disponibles son Newton-Raphson, Scoring de Fisher o un método híbrido en el que las iteraciones de Scoring de Fisher se realizan antes de cambiar al método de Newton-Raphson. Si se logra la convergencia durante la fase de Scoring de Fisher del método híbrido antes de que se lleven a cabo el número máximo de iteraciones de Fisher, el algoritmo continúa con el método de Newton-Raphson.

Estimación. Este grupo otorga el control sobre varios criterios utilizados en la estimación del modelo.

- **Nº máximo de iteraciones.** Número máximo de iteraciones que se ejecutará el algoritmo. Especifique un número entero no negativo.
- **Máxima subdivisión por pasos.** En cada iteración, se reduce el tamaño del paso mediante un factor de 0,5 hasta que aumenta el logaritmo de la verosimilitud o se alcanza la máxima subdivisión por pasos. Especifique un número entero positivo.
- **Limitar las iteraciones en función del cambio en las estimaciones de los parámetros.** Si se activa, el algoritmo se detiene tras una iteración en la que las modificaciones absolutas o relativas en las estimaciones de los parámetros sean inferiores que el valor especificado, que debe ser no negativo.
- **Limitar las iteraciones en función del cambio en la log-verosimilitud.** Si se activa, el algoritmo se detiene tras una iteración en la que las modificaciones absolutas o relativas en la función de log-verosimilitud sean inferiores que el valor especificado, que debe ser no negativo.
- **Comprobar si hay separación completa de los puntos de los datos.** Si se activa, el algoritmo realiza una prueba para garantizar que las estimaciones de los parámetros tienen valores exclusivos. Se produce una separación cuando el procedimiento pueda generar un modelo que clasifique cada caso de forma correcta.
- **Mostrar historial de iteraciones.** Muestra los estadísticos y las estimaciones de los parámetros cada *n* iteraciones, comenzando por la iteración 0 (estimaciones iniciales). Si decide imprimir el historial de iteraciones, la última iteración se imprimirá siempre independientemente del valor de *n*.

Valores definidos como perdidos por el usuario. Las variables de diseño de escala, así como la variable dependiente y cualquier covariable, deben contener datos válidos. Los casos con datos no válidos de cualquiera de estas variables se excluyen del análisis. Estos controles permiten decidir si los valores definidos como perdidos por el usuario se deben tratar como válidos entre las variables de estratificación, conglomeración, subpoblación y de factor.

Intervalo de confianza. Se trata del nivel de intervalo de confianza para las estimaciones de coeficiente, las estimaciones de coeficiente exponenciadas y las razones de las ventajas. Especifique un valor mayor o igual a 50 e inferior a 100.

Funciones adicionales del comando CSORDINAL

La sintaxis de comandos también le permite:

- Especificar contrastes personalizados de los efectos respecto a una combinación lineal de efectos o un valor (utilizando el subcomando `CUSTOM`).
- Fijar valores de otras variables de modelo en valores distintos de sus medias al calcular las razones de las ventajas para factores y covariables (utilizando el subcomando `ODDSRATIOS`).
- Utilice valores sin etiquetar como categorías de referencia personalizadas para los factores cuando se soliciten razones de las ventajas (usando el subcomando `ODDSRATIOS`).
- Especificar un valor de tolerancia para la comprobación de la singularidad (utilizando el subcomando `CRITERIA`).
- Generar una tabla de función estimable general (utilizando el subcomando `PRINT`).
- Guarde más de 25 variables de probabilidad (usando el subcomando `SAVE`).

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

Capítulo
12

Regresión de Cox de muestras complejas

El procedimiento Regresión de Cox de muestras complejas realiza análisis de supervivencias para muestras extraídas mediante métodos de muestreo complejo. Si lo desea, puede solicitar análisis de una subpoblación.

Ejemplos. Un organismo de orden público está preocupado por los índices de reincidencia en su área de jurisdicción. Una de las medidas de la reincidencia es el tiempo que transcurre antes del segundo arresto de los delincuentes. El organismo desea crear un modelo que refleje el tiempo que transcurre antes de un nuevo arresto utilizando la regresión de Cox, pero les preocupa que el supuesto de proporcionalidad de los impactos no sea válido en todas las diferentes categorías de edad.

Un grupo de investigadores estudia los tiempos de supervivencia de los pacientes que finalizan un programa de rehabilitación tras un ataque isquémico. Pueden existir varios casos por sujeto, ya que los historiales de los pacientes cambian cuando se registra la ocurrencia de eventos diferentes de la muerte y el momento en que ocurren dichos eventos. La muestra también está truncada a la izquierda en el sentido de que los tiempos de supervivencia observados se han "inflado" con la duración de la rehabilitación, ya que el comienzo del riesgo comienza en el momento del ataque isquémico, dado que la muestra incluye únicamente a los pacientes que han sobrevivido al programa de rehabilitación.

Tiempo de supervivencia. El procedimiento aplica la regresión de Cox al análisis de los tiempos de supervivencia, es decir, el tiempo que transcurre hasta que se produzca un evento. Hay dos maneras de especificar el tiempo de supervivencia, dependiendo del momento de inicio del intervalo:

- **Tiempo=0.** Normalmente, dispondrá de información completa acerca del inicio del intervalo para cada sujeto y sencillamente tendrá una variable que contenga los momentos de finalización (o creará una única variable con los momentos de finalización a partir de variables de fecha y hora, como verá a continuación).
- **Varía según el sujeto.** Esta opción es adecuada cuando tiene **truncación a la izquierda**, también denominada **entrada retardada**; por ejemplo, si desea analizar los tiempos de supervivencia de los pacientes que finalizan un programa de rehabilitación tras un ataque, tal vez prefiera considerar que el comienzo del riesgo comienza en el momento en que se produjo el ataque. No obstante, si la muestra incluye únicamente a aquellos pacientes que han sobrevivido al programa de rehabilitación, la muestra está truncada a la izquierda en el sentido de que los tiempos de supervivencia observados se han "inflado" con la duración de la rehabilitación. Puede tener este hecho en cuenta si especifica el momento en el que abandonaron la rehabilitación como el momento de entrada en el estudio.

Variables de fecha y hora. No se pueden utilizar directamente variables de fecha y hora para definir el inicio y la finalización del intervalo. Si tiene variables de fecha y hora, deberá utilizarlas para crear variables que contengan tiempos de supervivencia. Si no hay ninguna truncación a la izquierda, basta con crear una variable que contenga las horas de finalización basadas en la diferencia entre la fecha de entrada en el estudio y la fecha de la observación. Si existe truncación a la izquierda, puede crear una variable que contenga las horas de inicio, basadas en la diferencia entre la fecha de inicio del estudio y la fecha de entrada, y otra variable que contenga las horas de finalización, basadas en la diferencia entre la fecha de inicio del estudio y la fecha de la observación.

Estado del evento. Necesita una variable que registre si el sujeto ha experimentado el evento de interés dentro del intervalo. Los sujetos para los que no se ha producido el evento se censuran a la derecha.

Identificador de sujetos. Puede incorporar fácilmente predictores dependientes del tiempo y constantes por tramos dividiendo las observaciones de un único sujeto en varios casos. Por ejemplo, si desea analizar los tiempos de supervivencia de los pacientes tras un ataque, las variables que representan su historial médico probablemente sean útiles como predictores. Con el tiempo, pueden sufrir eventos médicos importantes que alteren su historial médico. La siguiente tabla muestra cómo estructurar un conjunto de datos de este tipo: *ID de paciente* es el identificador de los sujetos, *Hora de finalización* define los intervalos observados, *Estado* registra los eventos médicos importantes; *Historial anterior de ataques al corazón* e *Historial previo de hemorragias* son predictores dependientes del tiempo y constantes por tramos.

ID de paciente	*Hora de finalización*	*Estado*	*Historial anterior de ataques al corazón*	*Historial previo de hemorragias*
1	5	Ataque al corazón	No	No
1	7	Hemorragia	Sí	No
1	8	Fallecimiento	Sí	Sí
2	24	Fallecimiento	No	No
3	8	Ataque al corazón	No	No
3	15	Fallecimiento	Sí	No

Supuestos. Los casos del archivo de datos representan una muestra de un diseño complejo que se debe analizar según las especificaciones del archivo seleccionado en el Cuadro de diálogo Plan de muestras complejas.

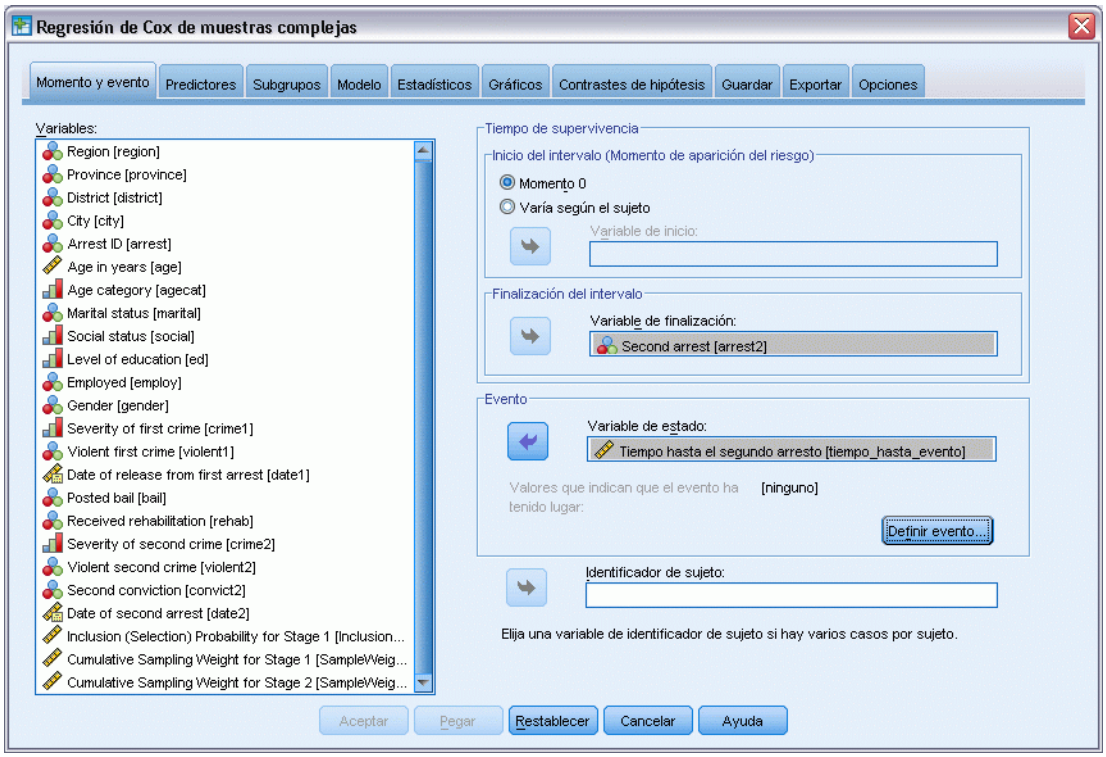
Normalmente, los modelos de regresión de Cox suponen que los impactos son proporcionales, es decir, que la proporción de impactos de un caso a otro no cambia con el tiempo. Si no se cumple este supuesto, tal vez sea necesario añadir predictores dependientes del tiempo al modelo.

Análisis Kaplan-Meier. Si no selecciona ningún predictor (o no introduce ninguno de los predictores seleccionados en el modelo) y elige el método de límite de producto para calcular la curva de supervivencia basal en la pestaña Opciones, el procedimiento realizará un análisis de supervivencia de tipo Kaplan-Meier.

Para obtener la regresión de Cox de muestras complejas

- Seleccione en los menús:
  Analizar > Muestras complejas > Regresión de Cox...
- Seleccione un archivo de plan. Si lo desea, elija un archivo de probabilidades conjuntas personalizado.
- Pulse en Continuar.

Figura 12-1
Cuadro de diálogo Regresión de Cox, pestaña Momento y evento

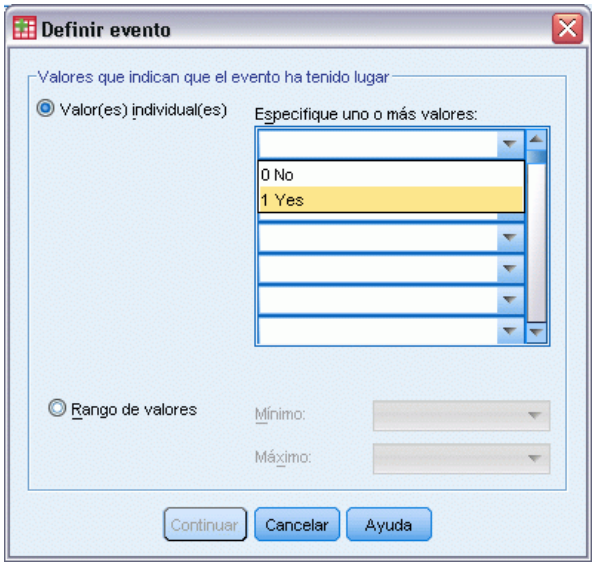


- Especifique el tiempo de supervivencia seleccionando los momentos de entrada y salida del estudio.
- Seleccione una variable de estado del evento.
- Pulse en Definir evento y defina al menos un valor de evento.

  Si lo desea, puede seleccionar un identificador de sujetos.

Definir evento

Figura 12-2
Cuadro de diálogo Definir evento

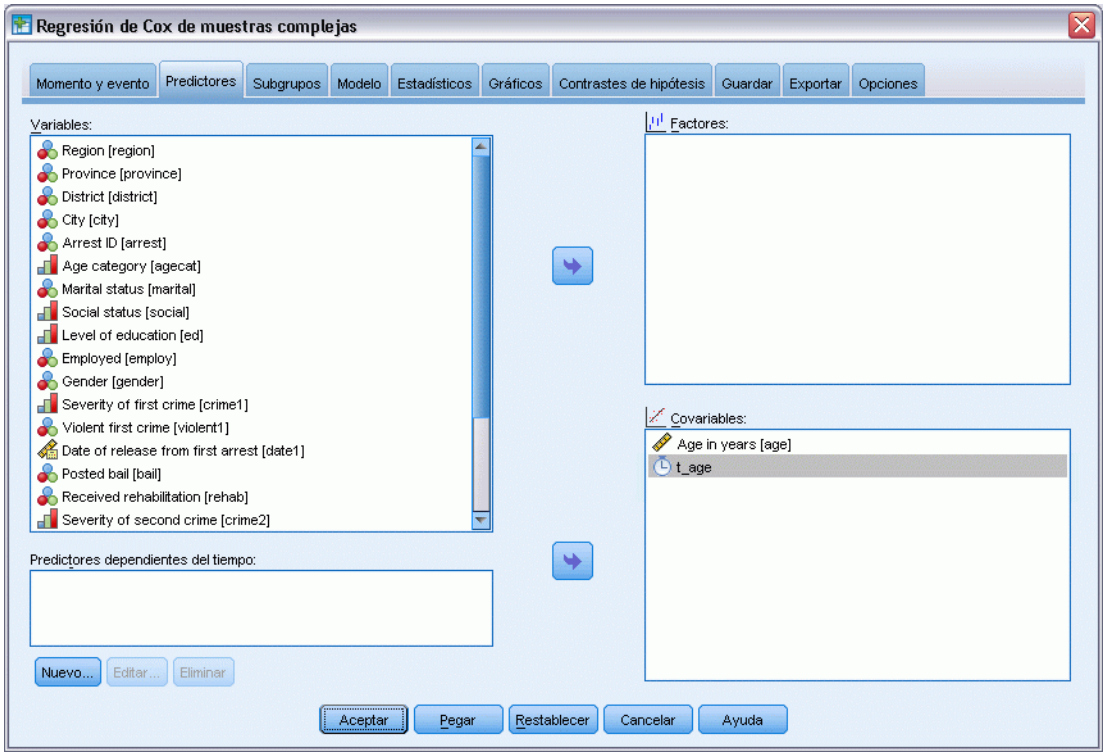


Especifique los valores que indican que se ha producido un evento terminal.

- **Valores individuales.** Especifique uno o más valores introduciéndolos en la cuadrícula o seleccionándolos en una lista de valores con etiquetas de valor definidas.
- **Rango de valores.** Especifique un rango de valores introduciendo los valores mínimo y máximo o seleccionando los valores en una lista con etiquetas de valor definidas.

Predictores

Figura 12-3
Cuadro de diálogo Regresión de Cox, pestaña Predictores



La pestaña Predictores permite especificar los factores y las covariables que se utilizarán para crear los efectos del modelo.

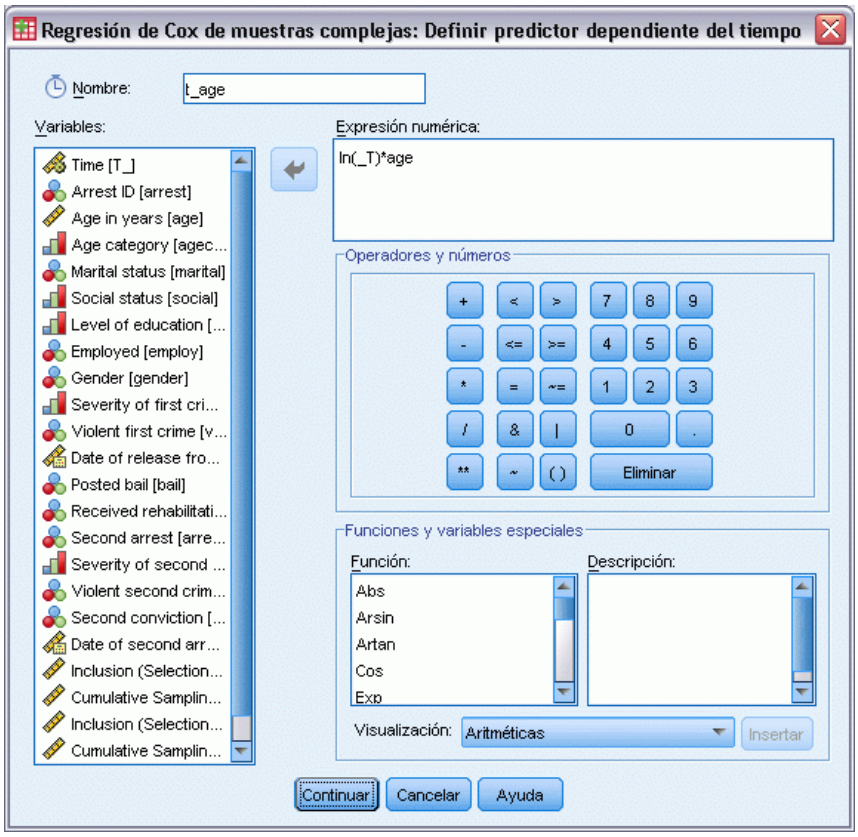
Factores. Los factores son predictores categóricos y pueden ser numéricos o de cadena.

Covariables. Las covariables son predictores de escala y deben ser numéricas.

Predictores dependientes del tiempo. Hay ciertas situaciones en las que no se cumple el supuesto de proporcionalidad de los impactos. Es decir, las tasas de impacto cambian con el tiempo y los valores de uno (o más) de los predictores son diferentes en los distintos puntos temporales. En tales casos, es necesario especificar predictores dependientes del tiempo. Si desea obtener más información, consulte el tema Definir predictor dependiente del tiempo el p. 83. Los predictores dependientes del tiempo se pueden seleccionar como factores o como covariables.

Definir predictor dependiente del tiempo

Figura 12-4
Cuadro de diálogo Regresión de Cox: Definir predictor dependiente del tiempo



El cuadro de diálogo Definir predictor dependiente del tiempo le permite crear un predictor que dependa de la variable de tiempo preincorporada, *T_*. Puede utilizar esta variable para definir covariables dependientes del tiempo empleando dos métodos generales:

- Si desea estimar un modelo de regresión de Cox extendido que permita impactos no proporcionales, defina el predictor dependiente del tiempo como una función de la variable de tiempo *T_* y la covariable en cuestión. Un ejemplo habitual sería el simple producto de la variable de tiempo y el predictor, pero también se pueden especificar funciones más complejas.
- Algunas variables pueden tener valores distintos en períodos diferentes del tiempo, pero no están sistemáticamente relacionadas con el tiempo. En tales casos es necesario definir un **predictor dependiente del tiempo segmentado**, lo cual puede llevarse a cabo usando expresiones lógicas. Las expresiones lógicas toman el valor 1 cuando son verdaderas y el valor 0 cuando son falsas. Es posible crear un predictor dependiente del tiempo a partir de un conjunto de medidas, usando una serie de expresiones lógicas. Por ejemplo, si se toma la tensión una vez a la semana durante cuatro semanas (identificadas como *BP1* a *BP4*), puede definir el predictor dependiente del tiempo como (*T_* < 1) * *BP1* + (*T_* >= 1 & *T_* < 2) * *BP2* + (*T_* >= 2 & *T_* < 3) * *BP3* + (*T_* >= 3 & *T_* < 4) * *BP4*. Tenga en cuenta que exactamente uno de los términos entre paréntesis será igual a uno para cualquier caso dado y el resto serán todos 0. En otras palabras, esta función se puede interpretar diciendo que “Si

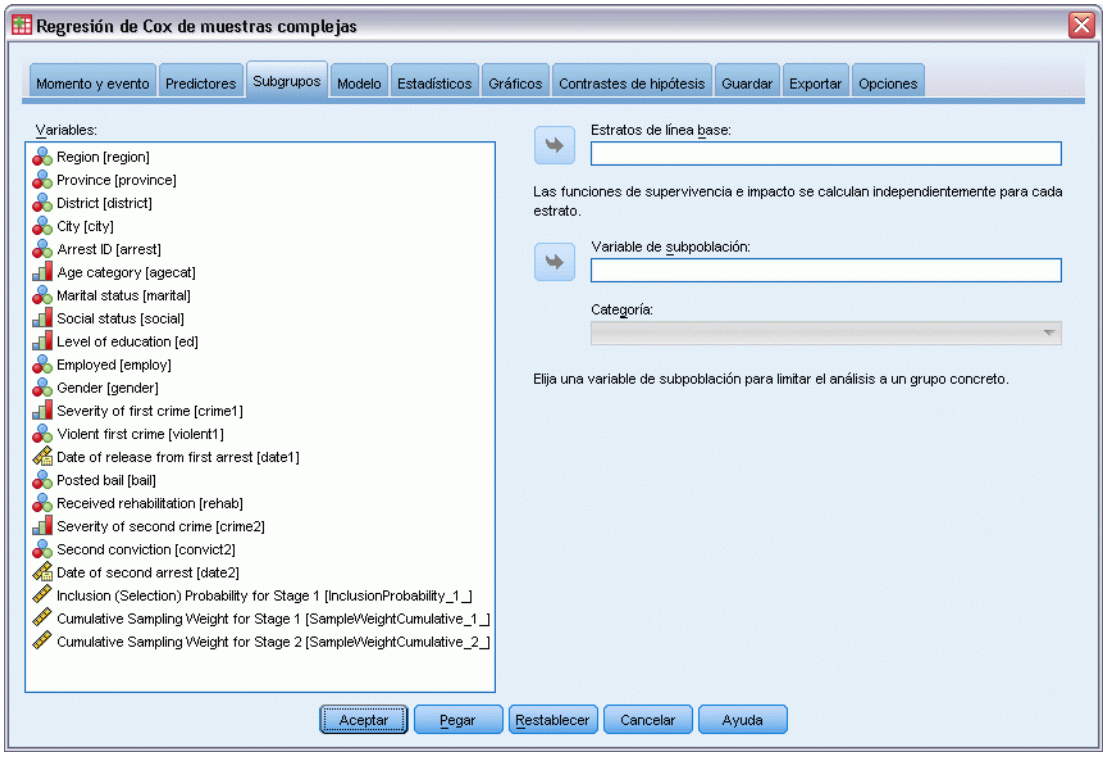
el tiempo es inferior a una semana, use *BP1*; si es más de una semana pero menos de dos, utilice *BP2*; y así sucesivamente".

Nota: si el predictor dependiente del tiempo segmentado es constante en los segmentos, como ocurre en el ejemplo de la tensión arterial explicado anteriormente, es posible que sea más sencillo especificar el predictor dependiente del tiempo y constante por tramos dividiendo los sujetos en varios casos. Consulte la explicación acerca de los identificadores de los sujetos en Regresión de Cox de muestras complejas el p. 78 para obtener más información.

En el cuadro de diálogo Definir predictor dependiente del tiempo, puede utilizar los controles de generación de funciones para crear la expresión para la covariable dependiente del tiempo o bien introducirla directamente en el área de texto Expresión numérica. Tenga en cuenta que las constantes de cadena deben ir entre comillas o apóstrofes y que las constantes numéricas se deben escribir en formato americano, con el punto como separador de la parte decimal. A la variable resultante se le asignará el nombre especificado y deberá incluirse como factor o covariable en la pestaña Predictores.

Subgrupos

Figura 12-5
Cuadro de diálogo Regresión de Cox, pestaña Subgrupos

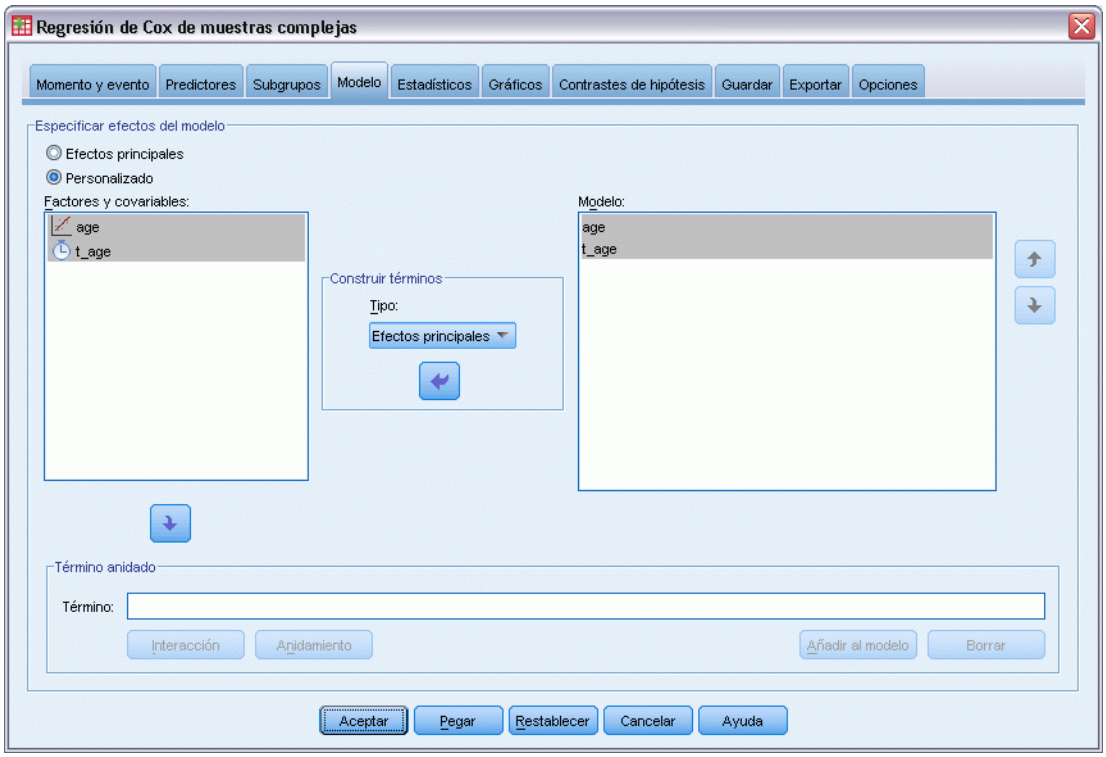


Estratos de línea base. Se calcula una función de supervivencia y de impacto basal diferente para cada valor de esta variable, a la vez que se estima un único conjunto de coeficientes del modelo en todos los estratos.

Variable de subpoblación. Especifique una variable para definir una subpoblación. El análisis se lleva a cabo únicamente en la categoría seleccionada de la variable de subpoblación.

Modelo

Figura 12-6
Cuadro de diálogo Regresión de Cox, pestaña Modelo



Especificar efectos del modelo. Por defecto, el procedimiento crea un modelo de efectos principales utilizando los factores y las covariables especificadas en el cuadro de diálogo principal. Si lo desea, también puede crear un modelo personalizado que contenga los efectos de la interacción y los términos anidados.

Términos no anidados

Para las covariables y los factores seleccionados:

Interacción. Crea el término de interacción de mayor nivel para todas las variables seleccionadas.

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

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

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

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

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

Términos anidados

En este procedimiento, puede construir términos anidados para el modelo. Los términos anidados resultan útiles para modelar el efecto de un factor o covariable cuyos valores no interactúan con los niveles de otro factor. Por ejemplo, una cadena de tiendas de comestibles desea realizar un seguimiento de los hábitos de gasto de los clientes en las diversas ubicaciones de sus tiendas. Dado que cada cliente frecuenta tan sólo una de estas ubicaciones, se puede decir que el efecto de *Cliente* está **anidado dentro** del efecto de *Ubicación de la tienda*.

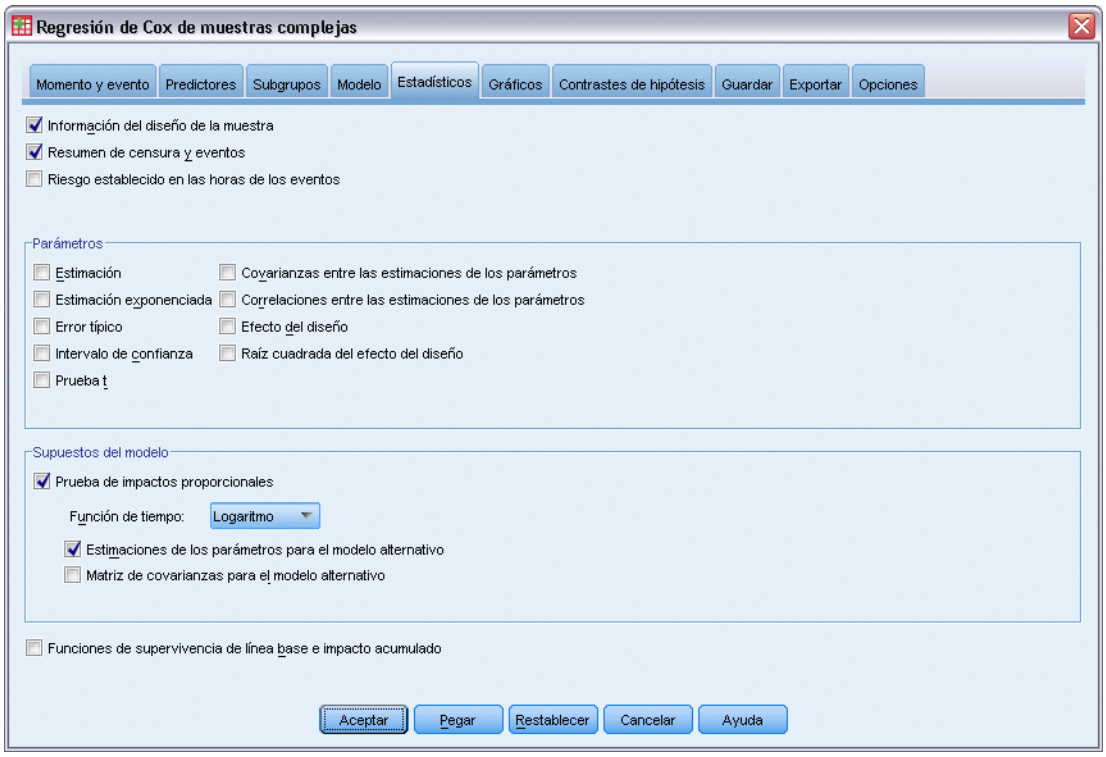
Además, puede incluir efectos de interacción, como términos polinómicos que implican a la misma covariable, o añadir varios niveles de anidación al término anidado.

Limitaciones. Existen las siguientes restricciones para los términos anidados:

- Todos los factores incluidos en una interacción deben ser exclusivos entre sí. Por consiguiente, si *A* es un factor, no es válido especificar *A*A*.
- Todos los factores incluidos en un efecto anidado deben ser exclusivos entre sí. Por consiguiente, si *A* es un factor, no es válido especificar *A(A)*.
- No se puede anidar ningún efecto dentro de una covariable. Por consiguiente, si *A* es un factor y *X* es una covariable, no es válido especificar *A(X)*.

Estadísticas

Figura 12-7
Cuadro de diálogo Regresión de Cox, pestaña Estadísticos



Información del diseño muestral. Muestra información resumida acerca de la muestra, incluidos un recuento no ponderado y el tamaño de la población.

Resumen de censura y eventos. Muestra información resumida acerca del número y el porcentaje de los casos censurados.

Riesgo establecido en las horas de los eventos. Muestra el número de eventos y el número bajo riesgo para cada momento de evento en cada estrato de línea base.

Parámetros. Este grupo permite controlar la presentación de estadísticos relacionados con los parámetros del modelo.

- **Estimación.** Muestra estimaciones de los coeficientes.
- **Estimación exponenciada.** Muestra la base del logaritmo natural elevada a la potencia de las estimaciones de los coeficientes. Mientras que las estimaciones tienen propiedades agradables para la comprobación estadística, la estimación exponenciada (o exp[B]) es más sencilla de interpretar.
- **Error típico.** Muestra el error típico de cada estimación de los coeficientes.
- **Intervalo de confianza.** Muestra un intervalo de confianza para cada estimación de los coeficientes. El nivel de confianza de los intervalos se configura en el cuadro de diálogo Opciones.
- **Prueba t.** Muestra una prueba *t* de cada estimación de coeficientes. La hipótesis nula de cada prueba es que el valor del coeficiente sea 0.
- **Covarianzas de las estimaciones de los parámetros.** Muestra una estimación de la matriz de covarianzas de los coeficientes del modelo.
- **Correlaciones de las estimaciones de los parámetros.** Muestra una estimación de la matriz de correlaciones de los coeficientes del modelo.
- **Efecto del diseño.** Cociente de la variación de la estimación entre la variación obtenida al suponer que la muestra es una muestra aleatoria simple. Es una medida del efecto de especificar un diseño complejo donde los valores más distantes de 1 indican efectos mayores.
- **Raíz cuadrada del efecto del diseño.** Es una medida del efecto de especificar un diseño complejo donde los valores más distantes de 1 indican efectos mayores.

Supuestos del modelo. Este grupo le permite generar un contraste del supuesto de proporcionalidad de los impactos. El contraste compara el modelo ajustado con un modelo alternativo que incluye los predictores dependientes del tiempo *x**_TF* para cada predictor *x*, donde *_TF* es la función de tiempo especificada.

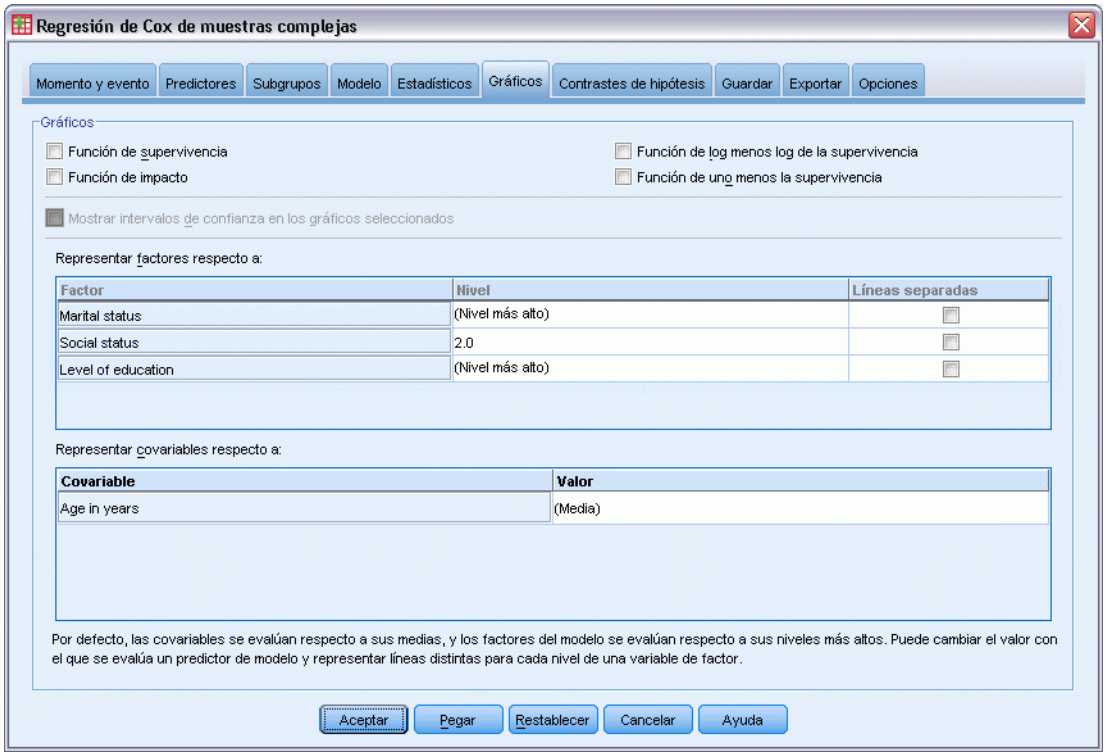
- **Función de tiempo.** Especifica la forma de *_TF* para el modelo alternativo. Para la función **identidad**, *_TF*=*T_*. Para la función **log**, *_TF*=log(*T_*). Para la función **Kaplan-Meier**, *_TF*=1−S_{KM}(*T_*), donde S_{KM}(.) es la estimación de Kaplan-Meier de la función de supervivencia. Para **rango**, *_TF* es el orden de rango de *T_* entre los momentos de finalización observados.
- **Estimaciones de los parámetros para el modelo alternativo.** Muestra la estimación, el error típico y el intervalo de confianza de cada parámetro del modelo alternativo.
- **Matriz de covarianzas del modelo alternativo.** Muestra la matriz de las covarianzas estimadas entre los parámetros del modelo alternativo.

Funciones de supervivencia de línea base e impacto acumulado. Muestra la función de supervivencia de línea base y la función de impactos acumulados de línea base junto con sus errores típicos.

Nota: si se han incluido en el modelo los predictores dependientes del tiempo definidos en la pestaña Predictores, esta opción no está disponible.

Gráficos

Figura 12-8
Cuadro de diálogo Regresión de Cox, pestaña Gráficos



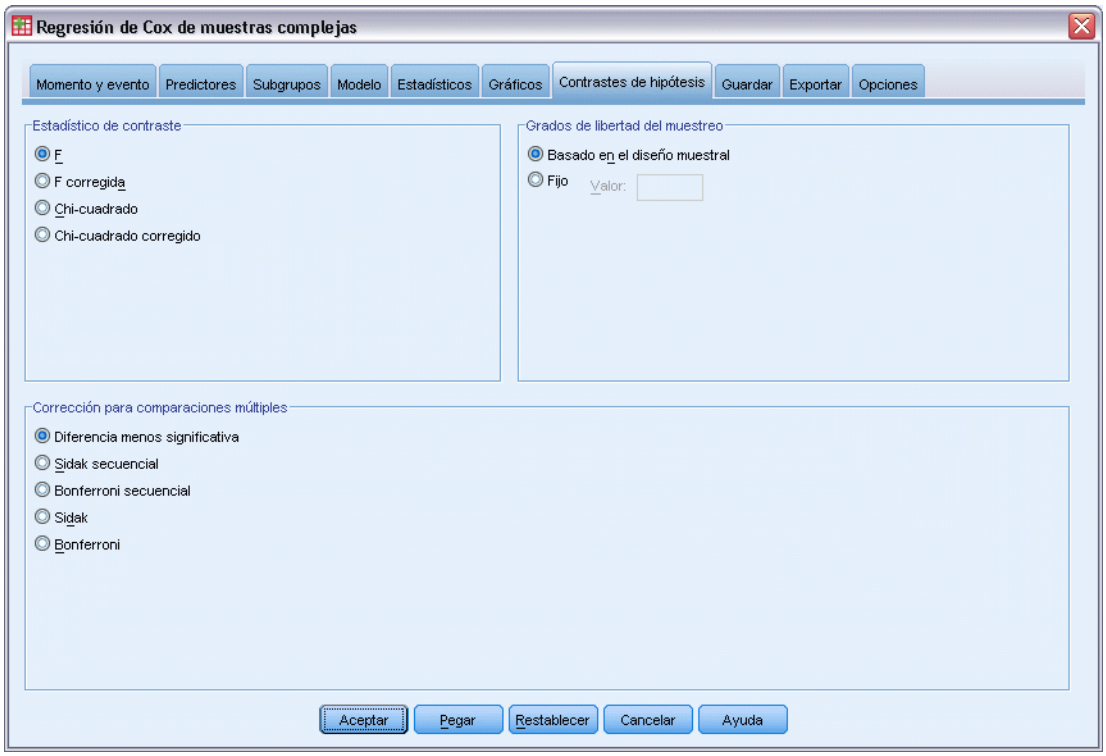
La pestaña Gráficos le permite solicitar gráficos de la función de impacto, la función de supervivencia, la función log menos log de la supervivencia y la función uno menos la supervivencia. También puede solicitar que se representen los intervalos de confianza junto con las funciones especificadas; el nivel de confianza se establece en la pestaña Opciones.

Patrones de predictores. Puede especificar que se utilice un patrón de valores de predictores para los gráficos solicitados y el archivo de supervivencia exportado en la pestaña Exportar. Tenga en cuenta que estas opciones no están disponibles si se han incluido en el modelo los predictores dependientes del tiempo definidos en la pestaña Predictores.

- **Representar factores respecto a.** Por defecto, cada factor se evalúa respecto a su nivel superior. Si lo desea, introduzca o seleccione otro nivel. También puede solicitar que se representen líneas distintas para cada nivel de un factor individual activando la casilla de verificación correspondiente a dicho factor.
- **Representar covariables respecto a.** Cada covariable se evalúa respecto a su media. Si lo desea, introduzca o seleccione otro valor.

Contrastes de hipótesis

Figura 12-9
Cuadro de diálogo Regresión de Cox, pestaña Contrastes de hipótesis



Estadístico de contraste. Este grupo le permite seleccionar el tipo de estadístico utilizado para contrastar las hipótesis. Es posible elegir entre *F*, *F* corregida, chi-cuadrado y chi-cuadrado corregido.

Muestreo de grados de libertad. Este grupo permite controlar los grados de libertad en el diseño de muestra usados para calcular los valores *p* de todos los estadísticos de contraste. Si se basa en el diseño muestral, el valor es la diferencia entre el número de unidades de muestra primarias y el número de estratos de la primera etapa del muestreo. Si lo desea, puede especificar los grados de libertad que desee introduciendo un número entero positivo.

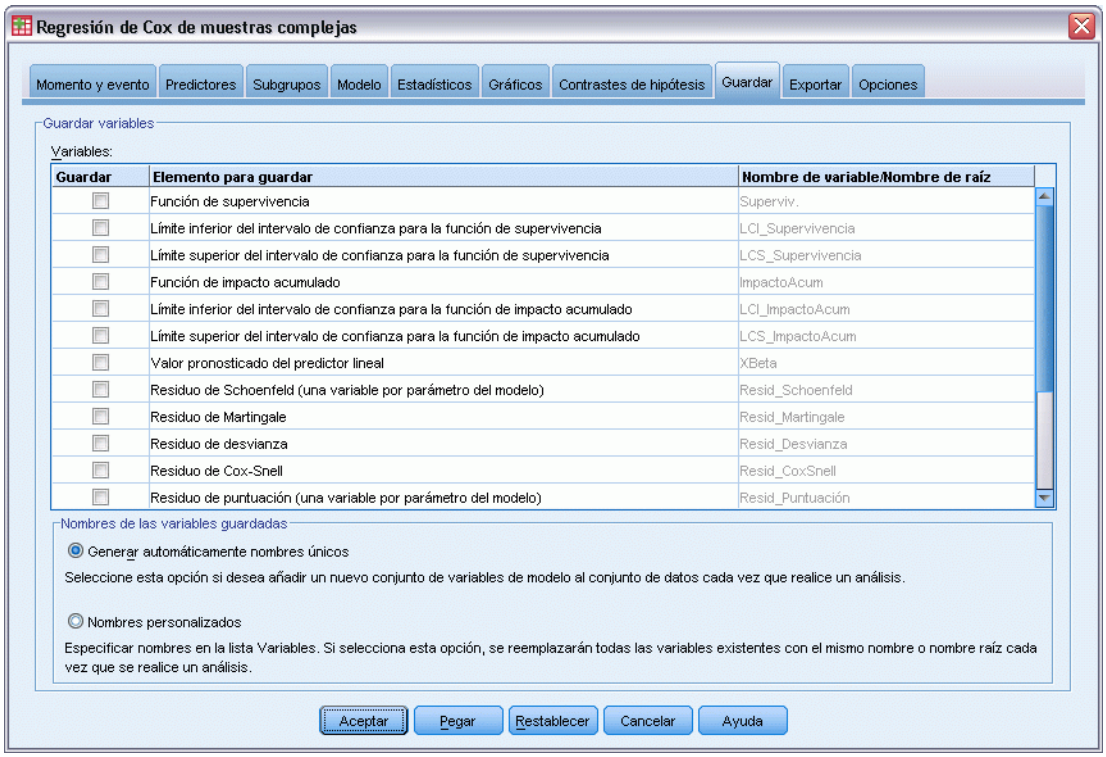
Corrección para comparaciones múltiples. Al realizar contrastes de hipótesis con varios contrastes, el nivel de significación global se puede ajustar utilizando los niveles de significación de los contrastes incluidos. Este grupo permite elegir el método de ajuste.

- **Diferencia menos significativa.** Este método no controla la probabilidad general de rechazar las hipótesis de que algunos contrastes lineales son diferentes a los valores de hipótesis nula.
- **Sidak secuencial.** Este es un procedimiento de Sidak de rechazo secuencial decreciente que es mucho menos conservador en términos de rechazar las hipótesis individuales pero que mantiene el mismo nivel de significación global.
- **Bonferroni secuencial.** Este es un procedimiento de Bonferroni de rechazo secuencial decreciente que es mucho menos conservador en términos de rechazar las hipótesis individuales pero que mantiene el mismo nivel de significación global.

- **Sidak.** Este método ofrece límites más estrechos que los de la aproximación de Bonferroni.
- **Bonferroni.** Este método corrige el nivel de significación observado por el hecho de que se están poniendo a prueba múltiples contrastes.

Guardar

Figura 12-10
Cuadro de diálogo Regresión de Cox, pestaña Guardar



Guardar variables. Este grupo permite guardar las variables relacionadas con el modelo en el conjunto de datos activo para utilizarlas posteriormente en otros diagnósticos y los informes sobre los resultados. Ninguna de estas opciones estará disponible si se incluyen en el modelo predictores dependientes del tiempo.

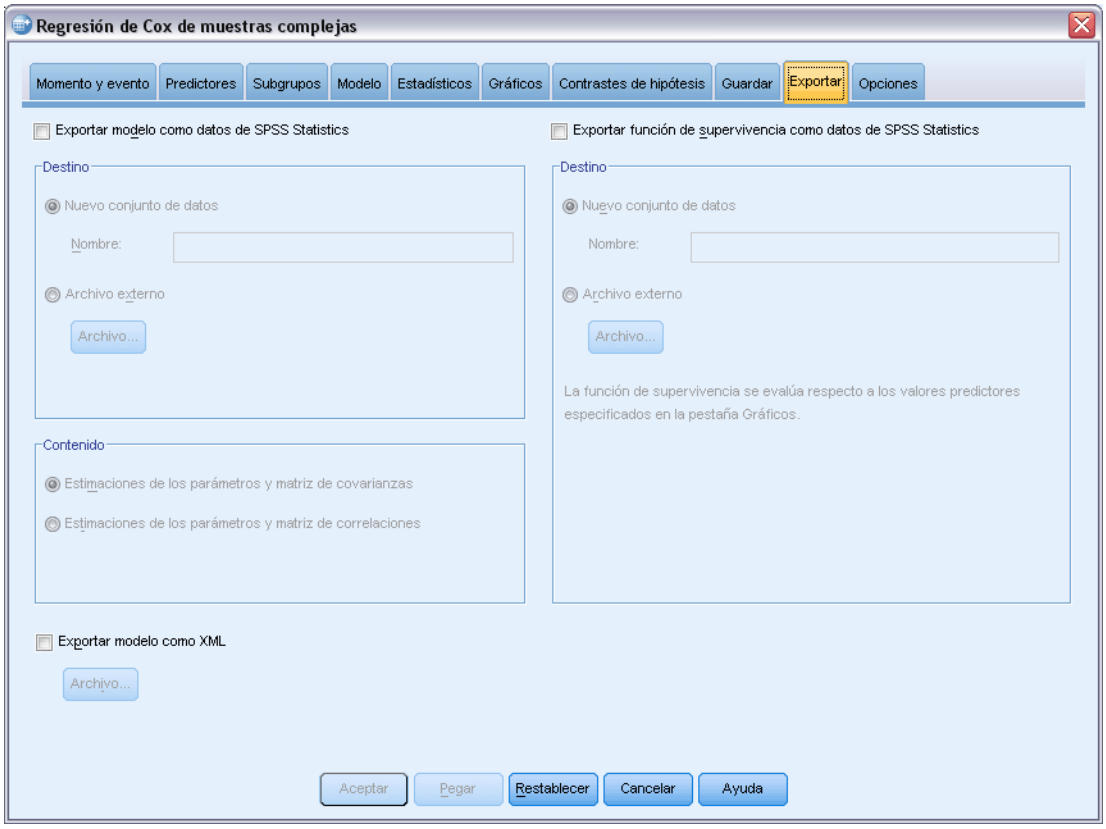
- **Función de supervivencia.** Guarda la probabilidad de supervivencia (el valor de la función de supervivencia) en el momento observado y los valores de los predictores para cada caso.
- **Límite inferior del intervalo de confianza para la función de supervivencia.** Guarda el límite inferior del intervalo de confianza de la función de supervivencia en el momento observado y los valores de los predictores para cada caso.
- **Límite superior del intervalo de confianza para la función de supervivencia.** Guarda el límite superior del intervalo de confianza de la función de supervivencia en el momento observado y los valores de los predictores para cada caso.
- **Función de impacto acumulado.** Guarda el impacto acumulado, o −ln(supervivencia), en el momento observado y los valores de los predictores para cada caso.

- **Límite inferior del intervalo de confianza para la función de impacto acumulado.** Guarda el límite superior del intervalo de confianza de la función de impacto acumulado en el momento observado y los valores de los predictores para cada caso.
- **Límite superior del intervalo de confianza para la función de impacto acumulado.** Guarda el límite inferior del intervalo de confianza de la función de impacto acumulado en el momento observado y los valores de los predictores para cada caso.
- **Valor pronosticado del predictor lineal.** Guarda la combinación lineal de los predictores corregidos de los valores de referencia por los coeficientes de regresión. El predictor lineal es la proporción de la función de impacto respecto al impacto de línea base. En el modelo de impactos proporcionales, este valor es constante respecto al tiempo.
- **Residuo de Schoenfeld.** Para cada caso no censurado y cada parámetro no redundante del modelo, el residuo de Schoenfeld es la diferencia entre el valor observado del predictor asociado al parámetro del modelo y el valor esperado del predictor para los casos con el riesgo establecido en el momento del evento observado. Los residuos de Schoenfeld se pueden utilizar para evaluar el supuesto de proporcionalidad de los impactos; por ejemplo, para un predictor *x*, los gráficos de los residuos de Schoenfeld de un predictor dependiente del tiempo *x**ln(*T_*) respecto al tiempo mostrarían una línea horizontal en 0 si se cumple el supuesto de proporcionalidad de los impactos. Se guarda una variable diferente para cada parámetro no redundante del modelo. Sólo se calculan los residuos de Schoenfeld para los casos no censurados.
- **Residuo de Martingale.** Para cada caso, el residuo de Martingale es la diferencia entre la censura observada (0 si está censurado, 1 si no lo está) y la esperanza de un evento durante el tiempo de observación.
- **Residuo de desvianza.** Los residuos de desvianza son residuos de Martingale "corregidos" para aparecer más simétricos alrededor de 0. Los gráficos de los residuos de desvianza respecto a los predictores no deben mostrar ningún patrón.
- **Residuo de Cox-Snell.** Para cada caso, el residuo de Cox-Snell es la esperanza de que se produzca un evento durante el tiempo de observación o la censura observada menos el residuo de Martingale.
- **Residuo de puntuación.** Para cada caso y cada parámetro no redundante del modelo, el residuo de puntuación es la contribución del caso a la primera derivada de la pseudo-verosimilitud. Se guarda una variable diferente para cada parámetro no redundante del modelo.
- **Residuo de DFBeta.** Para cada caso y cada parámetro no redundante del modelo, el residuo de DFBeta es una aproximación del cambio en el valor de la estimación del parámetro cuando se elimina el caso del modelo. Los casos con residuos de DFBeta relativamente grandes pueden estar ejerciendo una influencia indebida sobre el análisis. Se guarda una variable diferente para cada parámetro no redundante del modelo.
- **Residuos agregados.** Cuando varios casos representan a un único sujeto, el residuo agregado de un sujeto es sencillamente la suma de los residuos de todos los casos que corresponden al mismo sujeto. Para el residuo de Schoenfeld, la versión agregada es la misma que la versión no agregada, ya que el residuo de Schoenfeld sólo se define para los casos no censurados. Estos residuos sólo están disponibles cuando se ha especificado un identificador de sujetos en la pestaña Momento y evento.

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.

Exportar

Figura 12-11
Cuadro de diálogo Regresión de Cox, pestaña Exportar



Exportar modelo como datos de SPSS Statistics. Escribe un conjunto de datos de formato IBM® SPSS® Statisticsque contiene la matriz de covarianzas o correlaciones de los parámetros con las estimaciones de los parámetros, errores típicos, valores de significación y grados de libertad. El orden de las variables en el archivo matricial es el siguiente.

- **rowtype_.** Toma los valores (y las etiquetas de valor), COV (covarianzas), CORR (correlaciones), EST (estimaciones de los parámetros), SE (errores típicos), SIG (niveles de significación) y DF (grados de libertad del diseño muestral). Hay un caso diferente con el tipo de fila COV (o CORR) para cada parámetro del modelo, además de un caso diferente para cada uno de los otros tipos de filas.

- **varname_.** Toma los valores P1, P2, ..., correspondientes a una lista ordenada de todos los parámetros del modelo para los tipos de fila COV o CORR, con las etiquetas de valor correspondientes a las cadenas de parámetros mostradas en la tabla de estimaciones de los parámetros. Las casillas están vacías para los demás tipos de filas.
- **P1, P2, ...** Estas variables corresponden a una lista ordenada de todos los parámetros del modelo con las etiquetas de variable correspondientes a las cadenas de parámetros mostradas en la tabla de estimaciones de los parámetros y toman valores según el tipo de fila. Para los parámetros redundantes, todas las covarianzas se establecen en cero, las correlaciones se establecen en el valor perdido del sistema; todas las estimaciones de los parámetros se establecen en cero; y todos los errores típicos, niveles de significación y los grados de libertad residuales se establecen en el valor perdido del sistema.

Nota: Este archivo no se puede utilizar directamente para realizar otros análisis en otros procedimientos que lean un archivo matricial a menos que dichos procedimientos acepten todos los tipos de filas que aquí se exportan.

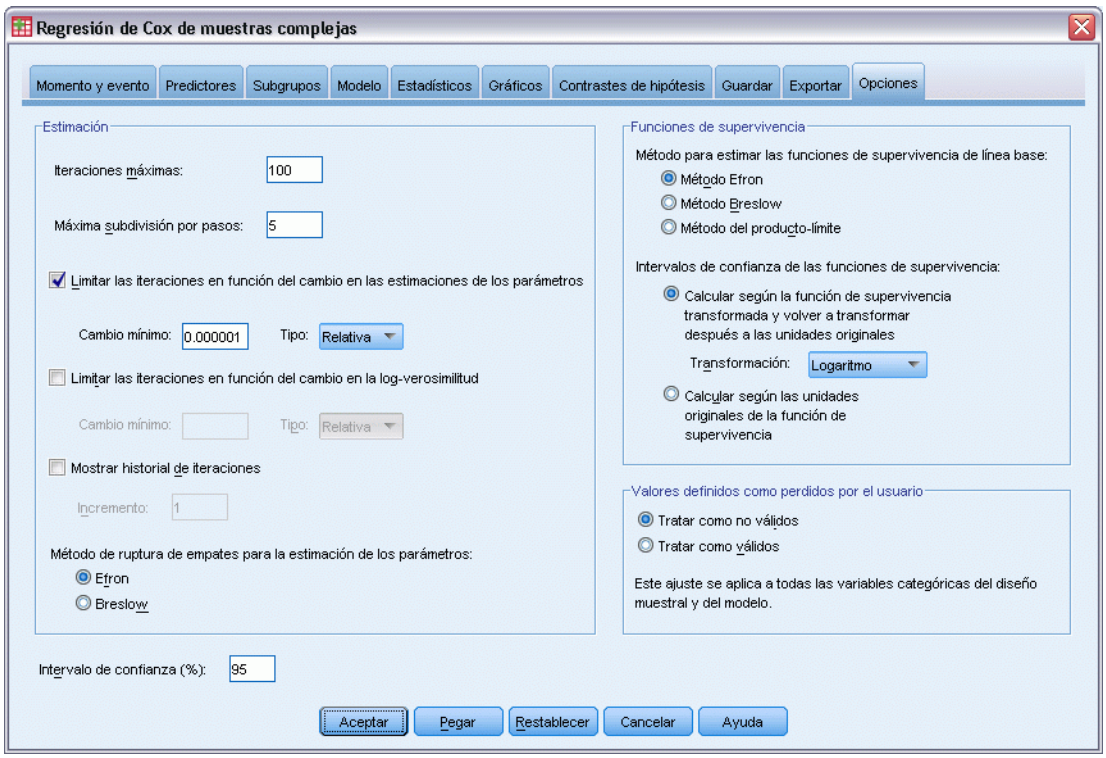
Exportar función de supervivencia como datos de SPSS Statistics. Escribe un conjunto de datos de formato SPSS Statistics que contiene la función de supervivencia; el error típico de la función de supervivencia; los límites superior e inferior del intervalo de confianza de la función de supervivencia; y la función de impactos acumulados de cada momento de evento o fallo, evaluada en la línea base y en los patrones de predictores especificados en la pestaña Gráfico. El orden de las variables en el archivo matricial es el siguiente.

- **Variable de estratos de línea base.** Se generan una tabla de supervivencia diferente para cada valor de la variable de los estratos.
- **Variable de tiempo de supervivencia.** Hora del evento; se crea un caso distinto para cada hora de evento única.
- **Sur_0, LCL_Sur_0, UCL_Sur_0.** Función de supervivencia de línea base y límites inferior y superior de su intervalo de confianza.
- **Sur_R, LCL_Sur_R, UCL_Sur_R.** Función de supervivencia evaluada en el patrón de "referencia" (consulte la tabla de valores de patrones que aparece en los resultados) y los límites superior e inferior de su intervalo de confianza.
- **Sur_#.#, LCL_Sur_#.#, UCL_Sur_#.#, …** Función de supervivencia evaluada en cada uno de los patrones de predictores especificados en la pestaña Gráficos y los límites superior e inferior de sus intervalos de confianza. Consulte la tabla de valores de los patrones que aparece en los resultados para ver la correspondencia entre los patrones y el número #.#.
- **Haz_0, LCL_Haz_0, UCL_Haz_0.** Función de impacto acumulado de línea base y límites inferior y superior de su intervalo de confianza.
- **Haz_R, LCL_Haz_R, UCL_Haz_R.** Función de impacto acumulado evaluada en el patrón de "referencia" (consulte la tabla de valores de patrones que aparece en los resultados) y los límites superior e inferior de su intervalo de confianza.
- **Haz_#.#, LCL_Haz_#.#, UCL_Haz_#.#, …** Función de impacto acumulado evaluada en cada uno de los patrones de predictores especificados en la pestaña Gráficos y los límites superior e inferior de sus intervalos de confianza. Consulte la tabla de valores de los patrones que aparece en los resultados para ver la correspondencia entre los patrones y el número #.#.

Exportar modelo como XML. Guarda toda la información necesaria para pronosticar la función de supervivencia, incluidas las estimaciones de los parámetros y la función de supervivencia de línea base, en formato XML (PMML). Puede utilizar este archivo de modelo para aplicar la información del modelo a otros archivos de datos para puntuarlo.

Opciones

Figura 12-12
Cuadro de diálogo Regresión de Cox, pestaña Opciones



Estimación. Estos controles especifican los criterios que se utilizan para estimar los coeficientes de regresión.

- **Nº máximo de iteraciones.** Número máximo de iteraciones que se ejecutará el algoritmo. Especifique un número entero no negativo.
- **Máxima subdivisión por pasos.** En cada iteración, se reduce el tamaño del paso mediante un factor de 0,5 hasta que aumenta el logaritmo de la verosimilitud o se alcanza la máxima subdivisión por pasos. Especifique un número entero positivo.
- **Limitar las iteraciones en función del cambio en las estimaciones de los parámetros.** Si se activa, el algoritmo se detiene tras una iteración en la que las modificaciones absolutas o relativas en las estimaciones de los parámetros son inferiores al valor especificado, que debe ser positivo.
- **Limitar las iteraciones en función del cambio en la log-verosimilitud.** Si se activa, el algoritmo se detiene tras una iteración en la que las modificaciones absolutas o relativas en la función de log-verosimilitud sean inferiores que el valor especificado, que debe ser positivo.

- **Mostrar historial de iteraciones.** Muestra el historial de iteraciones de las estimaciones de los parámetros y el pseudo log-verosimilitud. Además, imprime la última evaluación del cambio en las estimaciones de los parámetros y el pseudo log-verosimilitud. La tabla del historial de iteraciones se imprime cada *n* iteraciones a partir de la iteración *0* (las estimaciones iniciales), donde *n* es el valor del incremento. Si se solicita el historial de iteraciones, la última iteración siempre se muestra independientemente de *n*.
- **Método de ruptura de empates para la estimación de los parámetros.** Cuando hay momentos de fallo observados empatados, se utiliza uno de los siguientes métodos para romper los empates. El método de Efron requiere un proceso de cálculo más extenso.

Funciones de supervivencia. Estos controles especifican los criterios de los cálculos que implican a la función de supervivencia.

- **Método de estimación de las funciones de supervivencia de línea base.** El método de **Breslow** (o de Nelson-Aalan o empírico) estima el impacto acumulado de línea base mediante una función de paso no decreciente con pasos en los momentos de fallo observados. A continuación, calcula la supervivencia de línea base mediante la relación supervivencia=exp(−impacto acumulado). El método de **Efron** requiere un proceso de cálculo más extenso y se reduce al método de Breslow cuando no hay ningún empate. El método de **límite del producto** estima la supervivencia de línea base mediante una función continua a la derecha no creciente; cuando no hay ningún predictor en el modelo, este método se reduce a la estimación de Kaplan-Meier.
- **Intervalos de confianza de las funciones de supervivencia.** El intervalo de confianza se puede calcular de tres maneras: en unidades originales, mediante una transformación logarítmica o mediante una transformación log menos log. Sólo la transformación log menos log garantiza que los límites del intervalo de confianza estarán comprendidos entre 0 y 1, pero la transformación logarítmica suele funcionar "mejor".

Valores definidos como perdidos por el usuario. Para que un caso se incluya en el análisis, todas las variables deben tener valores válidos para dicho caso. Estos controles permiten decidir si los valores definidos como perdidos por el usuario se tratan como válidos en los modelos categóricos (incluidas las variables de subpoblación, estratos, evento y factores) y las variables del diseño muestral.

Intervalo de confianza (%). Nivel del intervalo de confianza utilizado para las estimaciones de los coeficientes, estimaciones de los coeficientes exponenciadas, estimaciones de la función de supervivencia y estimaciones de la función de impacto acumulado. Especifique un valor mayor o igual a 0 y menor que 100.

Funciones adicionales del comando CSCOXREG

Con el lenguaje de comandos también podrá:

- Realizar contrastes de hipótesis personalizados (utilizando el subcomando `CUSTOM` y `/PRINT LMATRIX`).
- Especificar la tolerancia (utilizando `/CRITERIA SINGULAR`).
- Tabla de función estimable general (utilizando `/PRINT GEF`).

- Varios patrones de predictores (utilizando varios subcomandos PATTERN).
- Número máximo de variables guardadas cuando se especifica un nombre de raíz (utilizando el subcomando SAVE). El cuadro de diálogo respeta el valor por defecto de CSCOXREG de 25 variables.

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

Parte II: Ejemplos

Capítulo
13

Asistente de muestreo de la opción Muestras complejas

El Asistente de muestreo le guía a través de los pasos necesarios para crear, modificar o ejecutar un archivo de plan de muestreo. Antes de utilizar el asistente, debe tener en mente una población objetivo bien definida, una lista de las unidades muestrales y un diseño muestral adecuado.

Obtención de una muestra a partir de un marco de muestreo completo

A una agencia inmobiliaria se le asigna la tarea de asegurarse de que los impuestos sobre las propiedades se aplican de manera justa en todos los condados. Los impuestos se basan en el valor tasado de la propiedad, por lo que la agencia quiere realizar una encuesta a una muestra de propiedades de los condados para asegurarse de que los registros de todos los condados están igualmente actualizados. Sin embargo, los recursos para obtener las tasaciones actuales son limitados, por lo que es importante que se utilicen prudentemente los recursos disponibles. La agencia decide utilizar una metodología de muestreo complejo para seleccionar una muestra de propiedades.

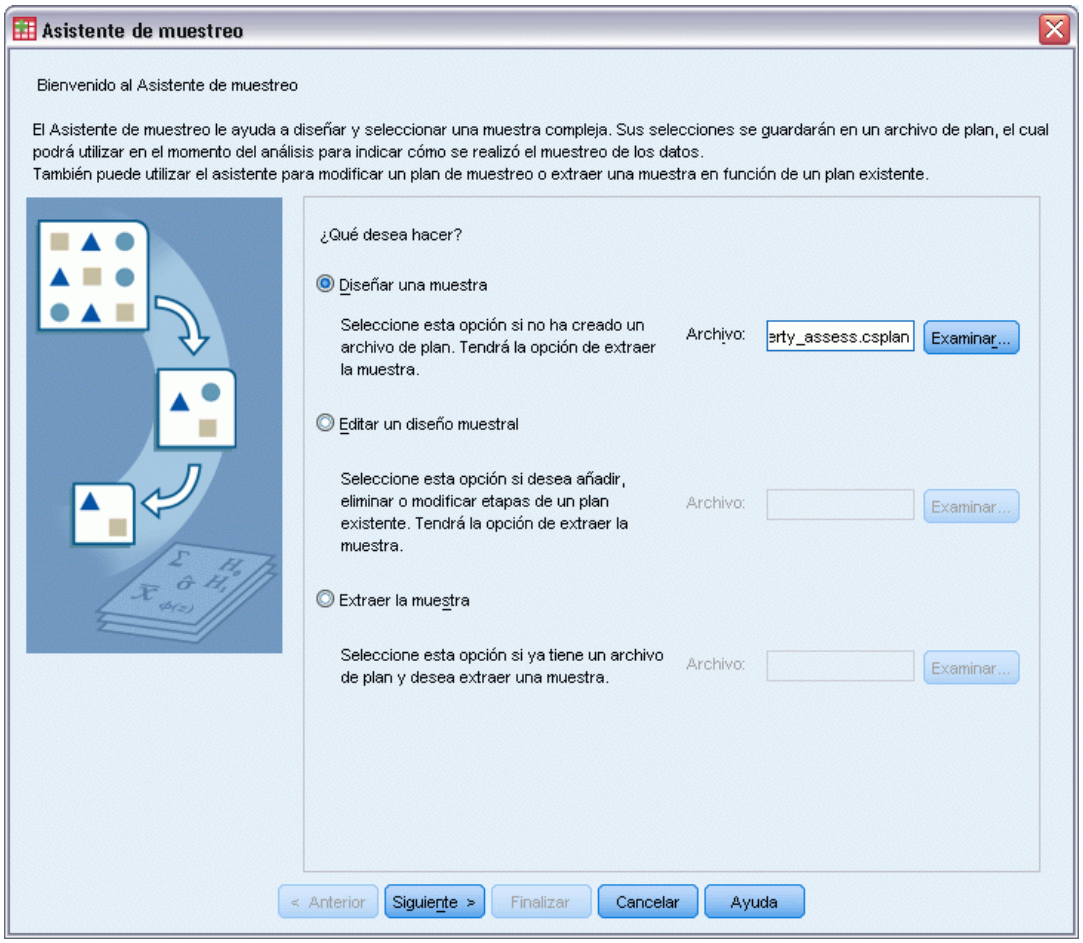
Se incluye una lista de propiedades en *property_assess_cs.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*. Utilice el Asistente de muestreo de la opción Muestras complejas para seleccionar una muestra.

Uso del asistente

- Para ejecutar el Asistente de muestreo de la opción de Muestras complejas, seleccione en los menús:

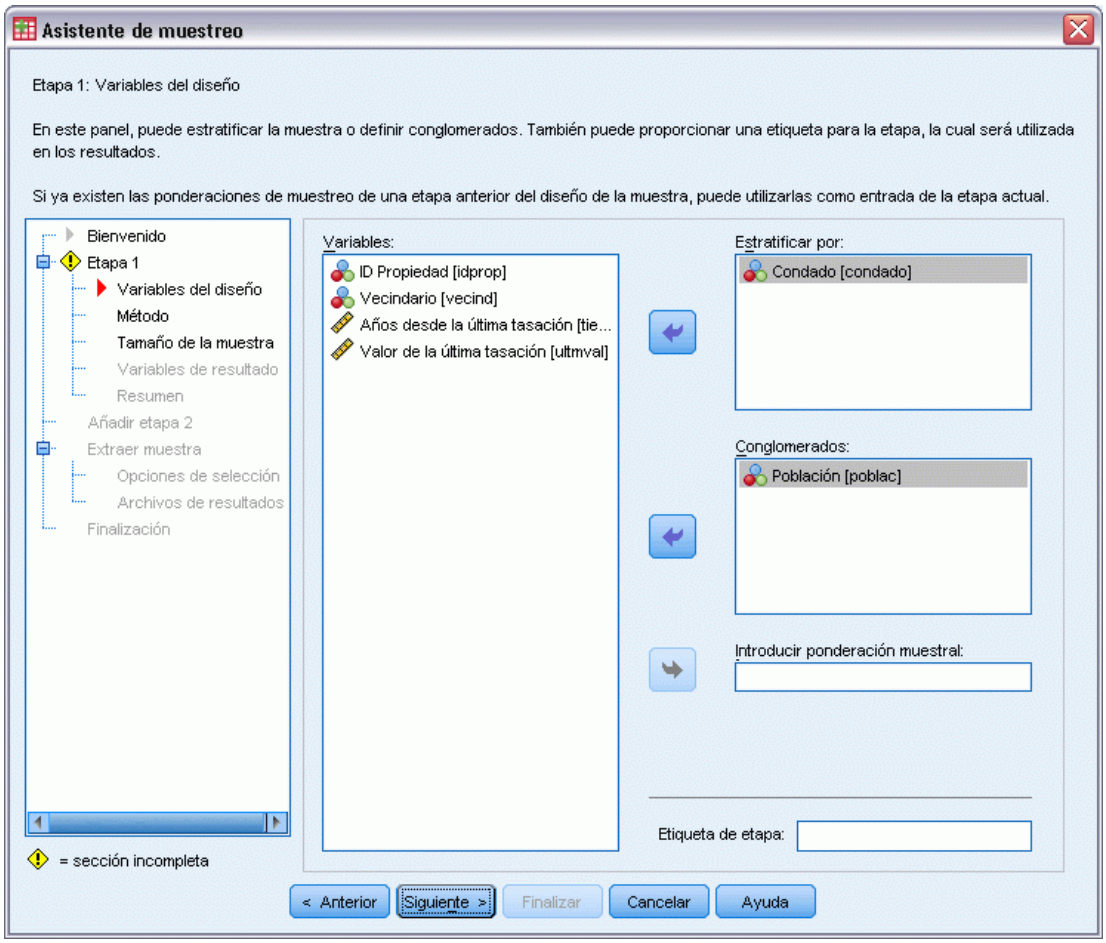
  Analizar > Muestras complejas > Seleccionar una muestra...

Figura 13-1
Asistente de muestreo: paso Bienvenida



- Seleccione Diseñar una muestra, navegue hasta la ubicación en la que desee guardar el archivo e introduzca property_assess.csplan como el nombre del archivo de plan.
- Pulse en Siguiente.

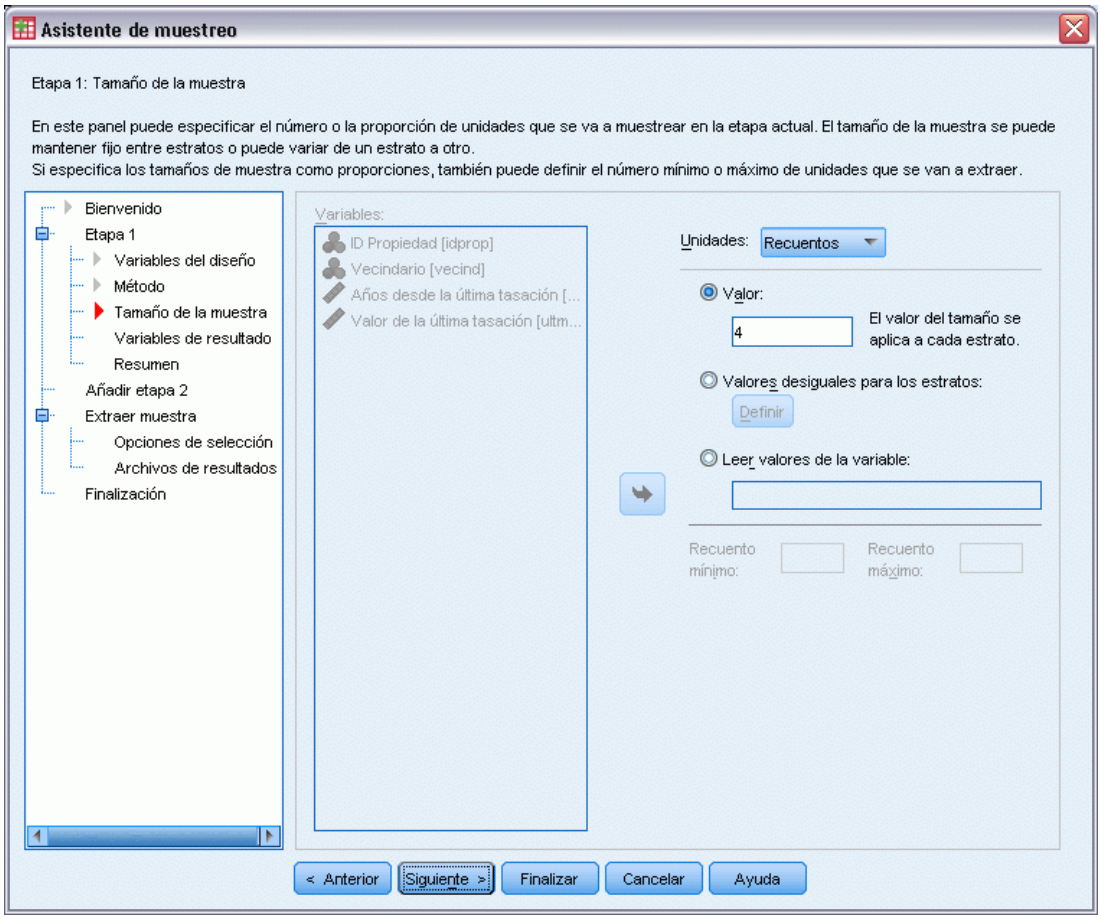
Figura 13-2
Asistente de muestreo: paso Variables del diseño (etapa 1)



- Seleccione *Condado* como variable de estratificación.
- Seleccione *Población* como variable de conglomeración.
- Pulse en Siguiente y, a continuación, pulse en Siguiente en el paso Método de muestreo.

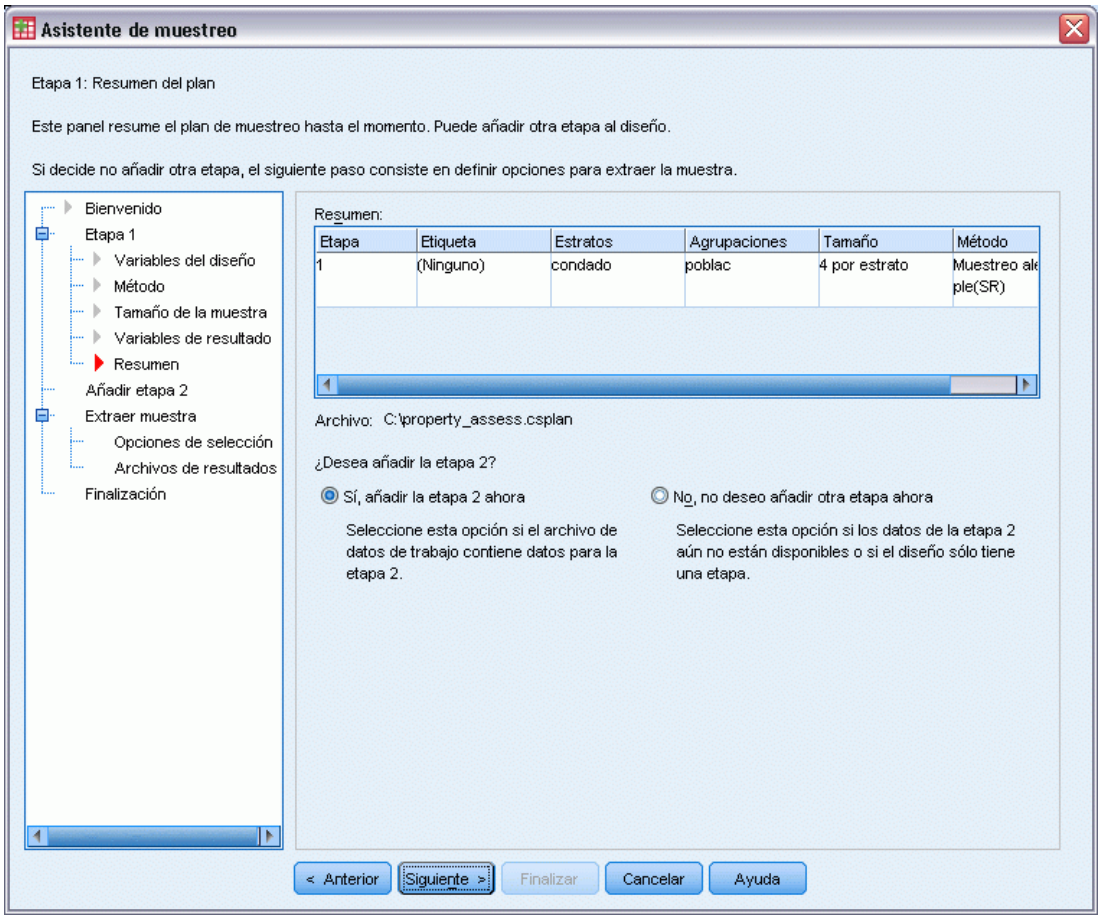
  Esta estructura de diseño indica que se extraen muestras independientes para cada condado. En esta etapa, se extraen las poblaciones como unidad muestral primaria mediante el método por defecto: Muestreo aleatorio simple.

Figura 13-3
Asistente de muestreo: paso Tamaño muestral (etapa 1)



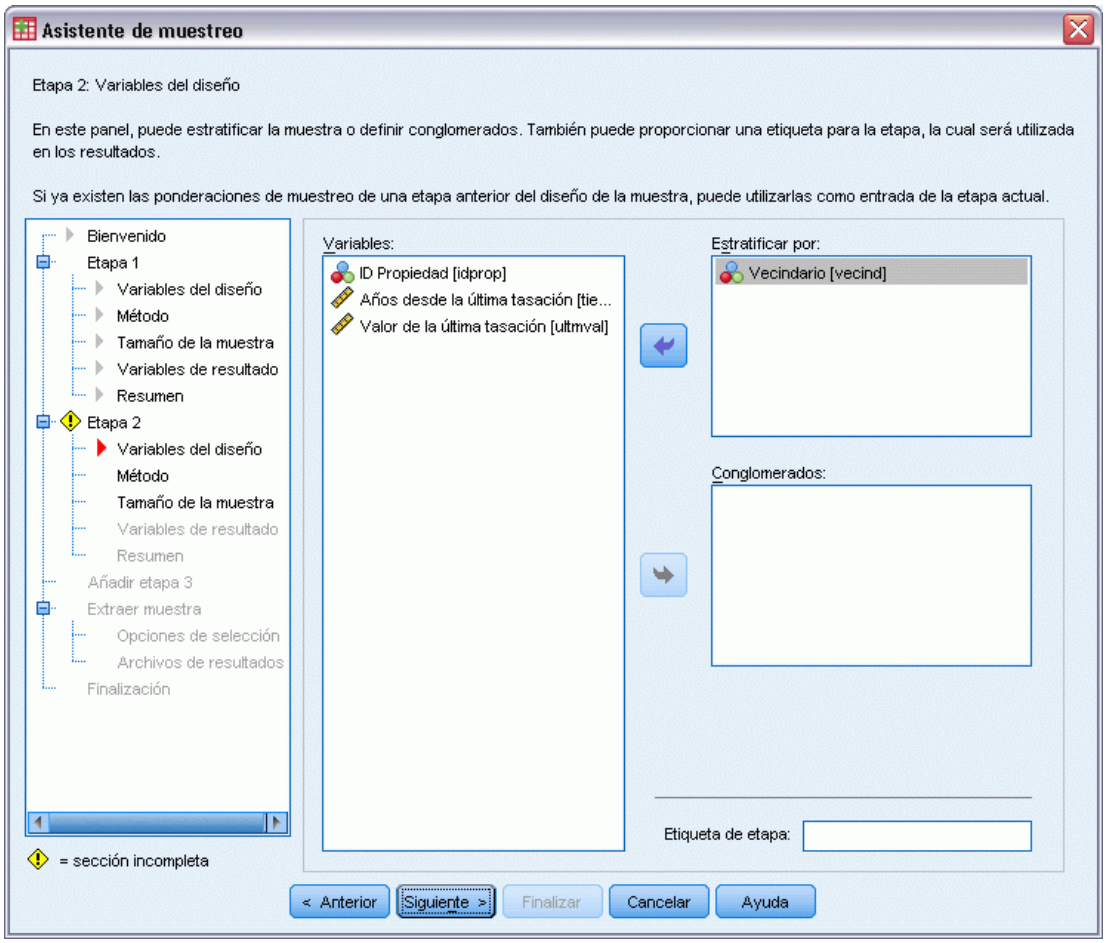
- Seleccione Recuentos en la lista desplegable Unidades.
- Escriba 4 como el valor del número de unidades que se van a seleccionar en esta etapa.
- Pulse en Siguiente y, a continuación, pulse en Siguiente en el paso Variables de resultado.

Figura 13-4
Asistente de muestreo: paso Resumen del plan (etapa 1)



- Seleccione Sí, añadir la etapa 2 ahora.
- Pulse en Siguiente.

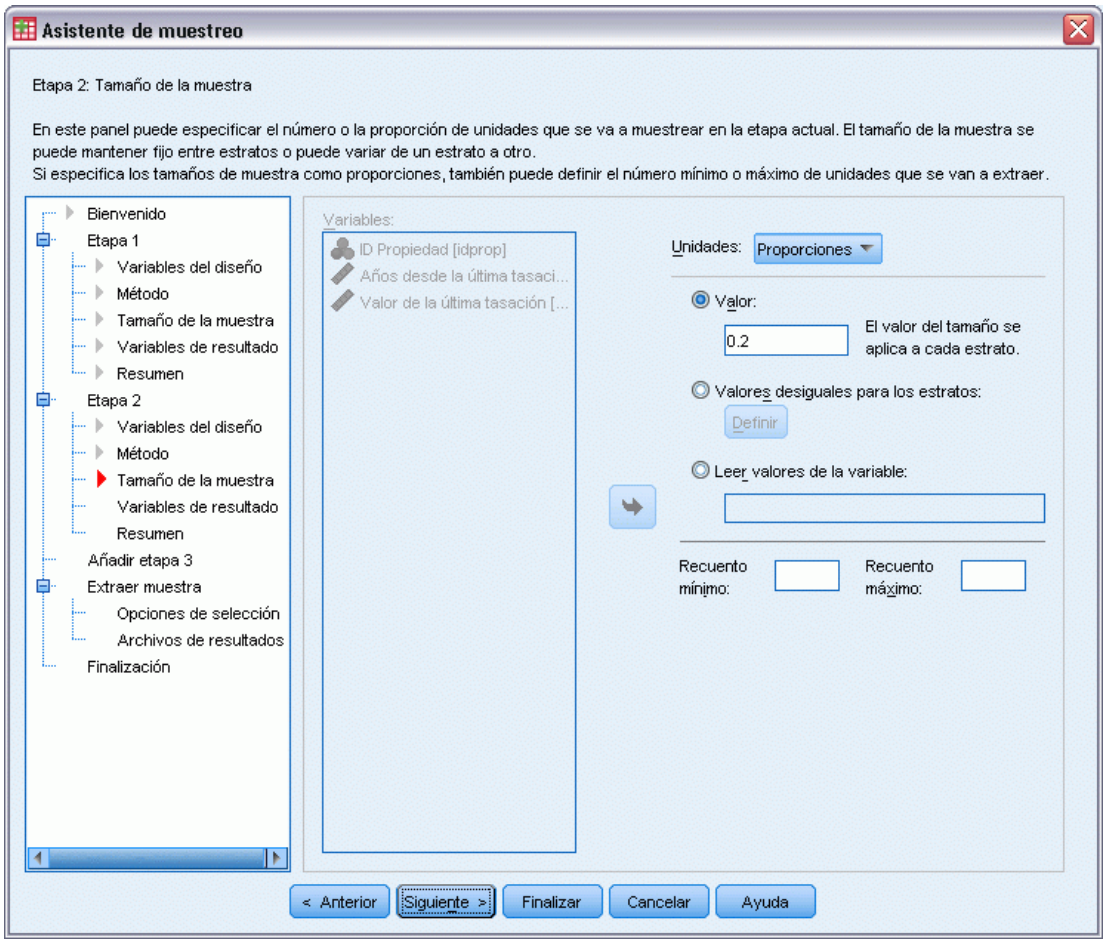
Figura 13-5
Asistente de muestreo: paso Variables del diseño (etapa 2)



- Seleccione *Vecindario* como variable de estratificación.
- Pulse en Siguiente y, a continuación, pulse en Siguiente en el paso Método de muestreo.

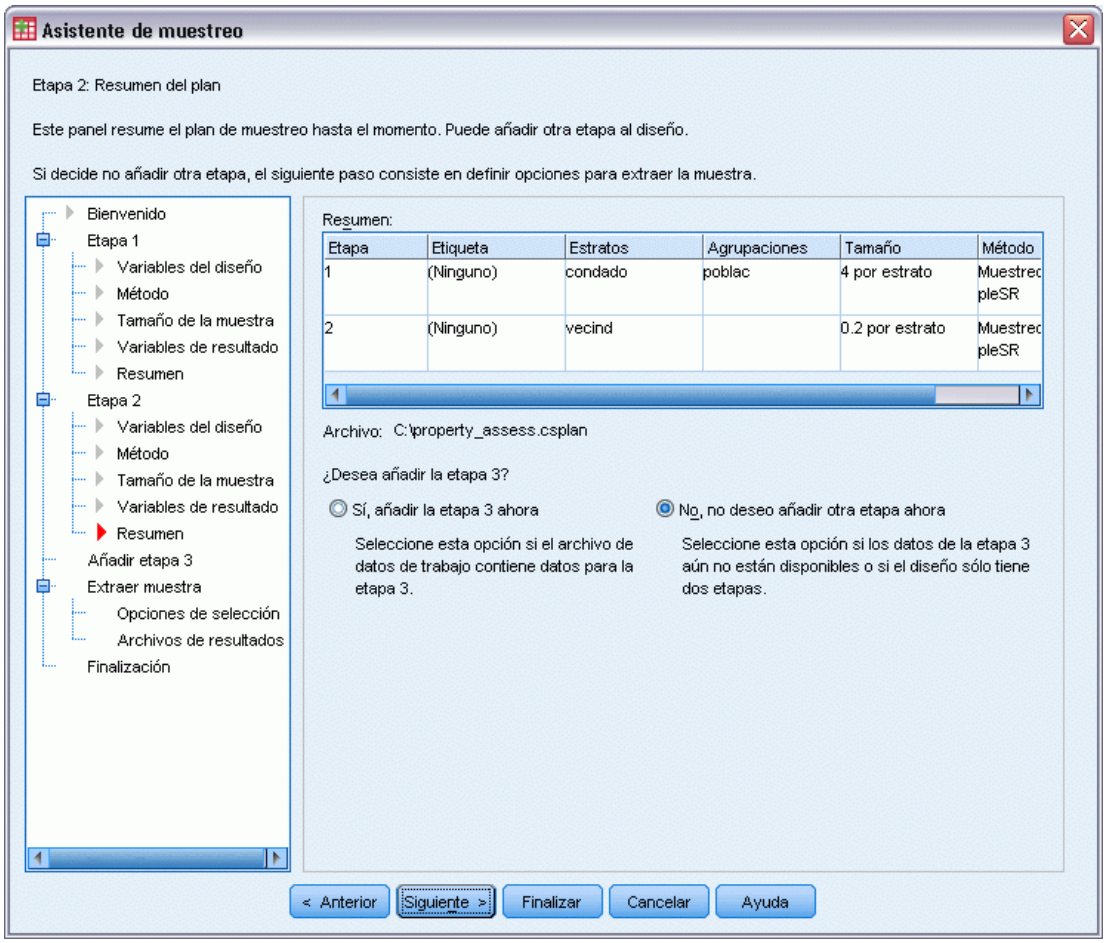
  Esta estructura de diseño indica que se extraen muestras independientes para cada vecindario de las poblaciones extraídas en la etapa 1. En esta etapa, se extraen las propiedades como unidad muestral primaria utilizando el muestreo aleatorio simple.

Figura 13-6
Asistente de muestreo: paso Tamaño muestral (etapa 2)



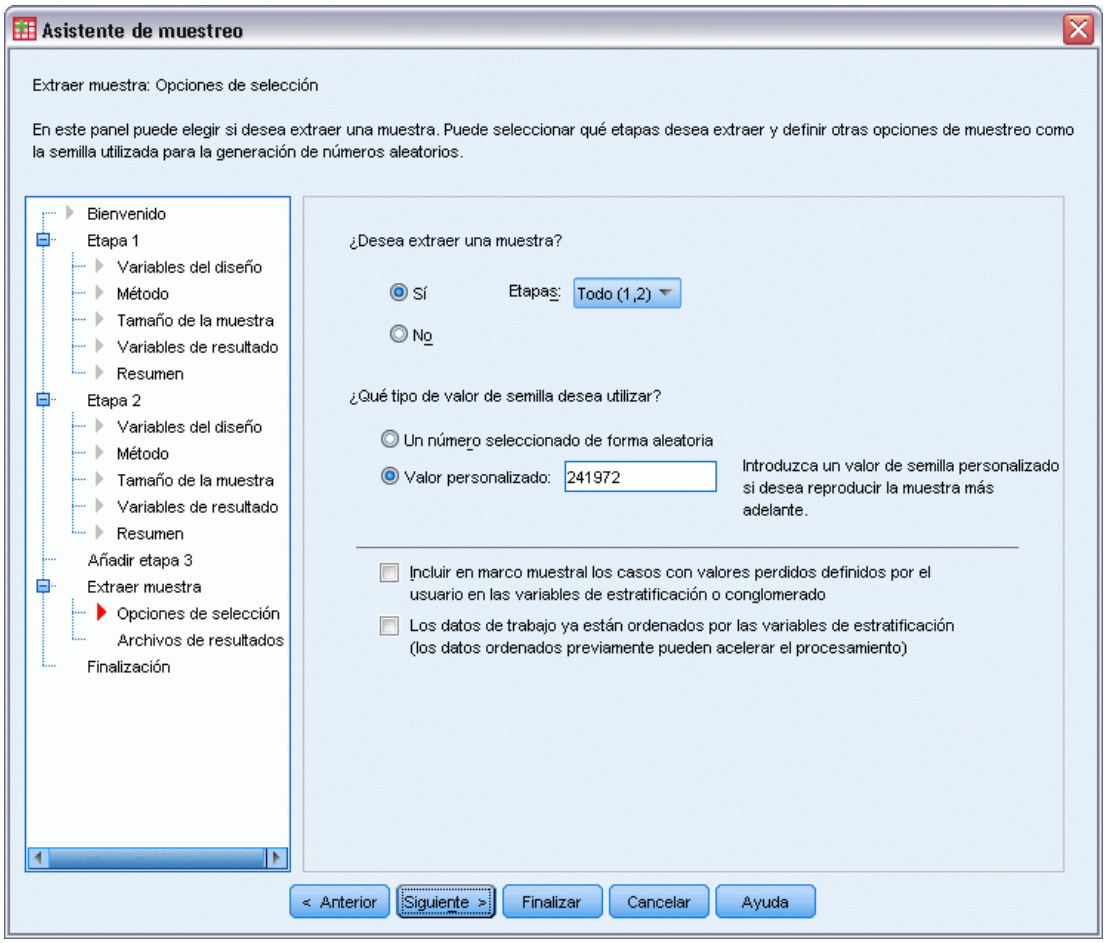
- Seleccione Proporciones en la lista desplegable Unidades.
- Escriba 0,2 como valor de la proporción de unidades que se van a extraer como muestra de cada estrato.
- Pulse en Siguiente y, a continuación, pulse en Siguiente en el paso Variables de resultado.

Figura 13-7
Asistente de muestreo: paso Resumen del plan (etapa 2)



▶ Revise el diseño muestral y, a continuación, pulse en Siguiente.

Figura 13-8
Asistente de muestreo: Extraer muestra: paso Opciones de selección

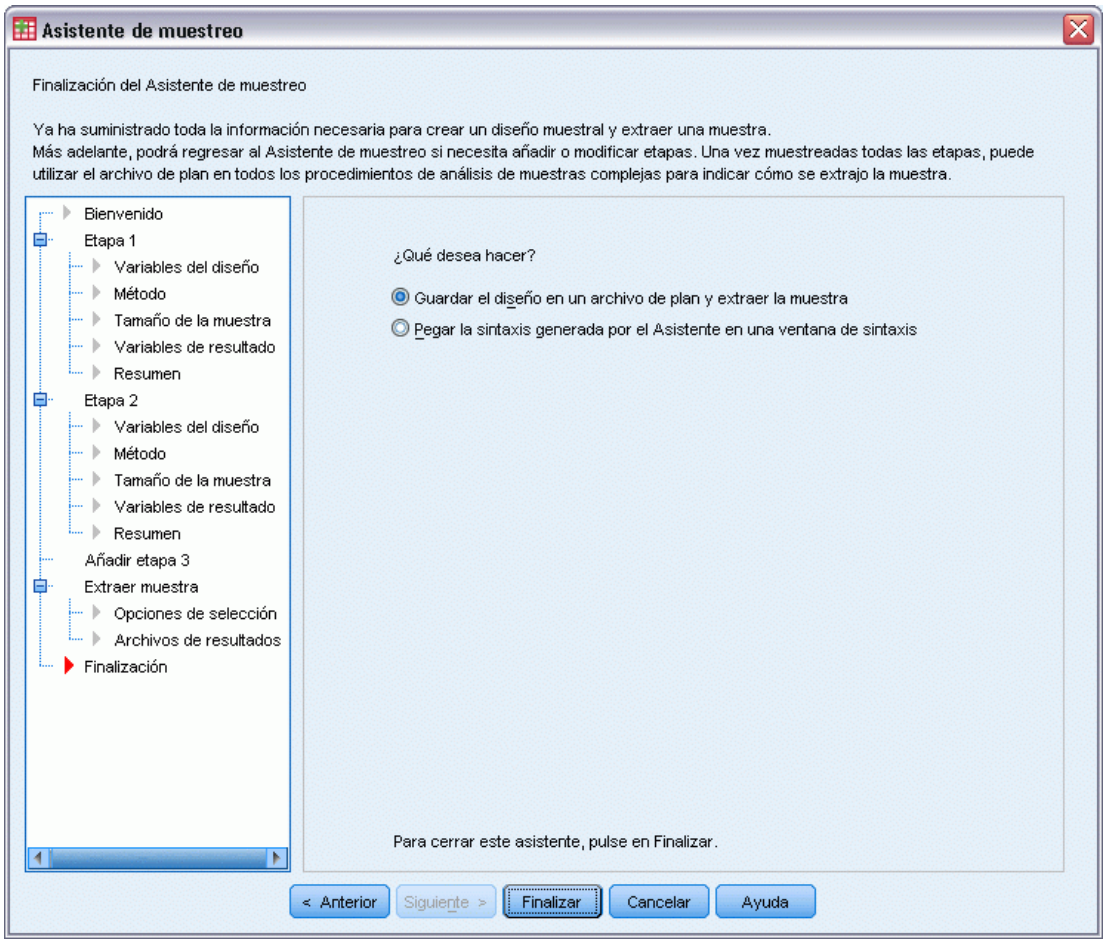


- Seleccione Valor personalizado como tipo de semilla aleatoria que se va a utilizar y escriba 241972 como valor.

  Al utilizar un valor personalizado, es posible replicar los resultados de este ejemplo de manera exacta.

- Pulse en Siguiente y, a continuación, pulse en Siguiente en Extraer muestra: paso Archivos de resultados.

Figura 13-9
Asistente de muestreo: paso Finalizar



▶ Pulse en Finalizar.

Estas selecciones generan el archivo de plan de muestreo *property_assess.csplan* y extraen una muestra de acuerdo con dicho plan.

Resumen del plan

Figura 13-10
Resumen del plan

			Etapa 1	Etapa 2
Variables del diseño	Estratificación	1	Población	Vecindario
	Conglomerado	1	Condado	
Información de la muestra	Método de selección		Muestreo aleatorio simple sin reposición	Muestreo aleatorio simple sin reposición
	Número de unidades muestreadas		4	
	Variables creadas o modificadas	Probabilidad de inclusión (selección) según etapa	Probabilidad Inclusión_1_	Probabilidad Inclusión_2_
		Ponderación de muestreo acumulada según etapa	Ponderación Muestral Acumulada_1_	Ponderación Muestral Acumulada_2_
	Proporción de unidades muestreadas			,2
Información sobre el análisis	Supuestos del estimador		Muestreo de probabilidad igual sin reposición	Muestreo de probabilidad igual sin reposición
	Probabilidad de inclusión		A partir de la variable Probabilidad Inclusión_1_	A partir de la variable Probabilidad Inclusión_2_

Archivo del plan: C:\property_assess2.csplan
Variable de ponderación: PonderaciónMuestral_Final_

La tabla de resumen muestra el plan de muestreo y resulta útil para asegurarse de que el plan corresponde a sus intenciones.

Resumen de muestreo

Figura 13-11
Resumen de las etapas

	Número de unidades muestreadas		Proporción de unidades muestreadas	
Condado	Solicitados	Reales	Solicitados	Reales
Este	4	4	44,4%	44,4%
Central	4	4	57,1%	57,1%
Oeste	4	4	25,0%	25,0%
Norte	4	4	44,4%	44,4%
Sur	4	4	50,0%	50,0%

Archivo del plan: C:\property_assess.csplan

Esta tabla de resumen muestra la primera etapa del muestreo y resulta útil para comprobar que el muestreo se ha realizado de acuerdo con el plan. Tal como se solicitó, se tomaron muestras de cuatro poblaciones de cada condado.

Figura 13-12
Resumen de las etapas

			Número de unidades muestreadas		Proporción de unidades muestreadas	
Condado	Población	Vecindario	Solicitados	Reales	Solicitados	Reales
Este	2	8	4	4	20,0%	19,0%
		9	14	14	20,0%	20,6%
		10	7	7	20,0%	18,9%
		11	14	14	20,0%	20,0%
	6	36	13	13	20,0%	20,3%
		37	14	14	20,0%	20,6%
		38	13	13	20,0%	20,6%
	7	43	12	12	20,0%	20,7%
		44	11	11	20,0%	19,6%
		45	11	11	20,0%	20,8%
		46	13	13	20,0%	20,0%
	9	57	13	13	20,0%	20,6%
		58	5	5	20,0%	18,5%
		59	11	11	20,0%	19,3%
		60	13	13	20,0%	19,4%
Central	22	148	9	9	20,0%	19,6%
		149	8	8	20,0%	20,0%

Esta tabla de resumen (de la cual se muestra aquí la parte superior) muestra la segunda etapa del muestreo. También resulta útil para comprobar que el muestreo se ha realizado de acuerdo con el plan. Como se solicitó, se muestreó aproximadamente el 20% de las propiedades de cada vecindario de cada una de las poblaciones muestreadas en la primera etapa.

Resultados de la muestra

Figura 13-13
Editor de datos con los resultados de la muestra

	idprop	vecind	poblac	condado	tiempo	ultmval	Probabilidadlnclusión_1_	PonderaciónMuestralAcumulada_1_	Probabilidadlnclusión_2_	PonderaciónMuestralAcumulada_2_	PonderaciónMuestral_Final_
273	13661,00	171	25	2	7	152,20	0,57	1,75	0,20	8,75	8,75
274	13668,00	171	25	2	6	104,30	0,57	1,75	0,20	8,75	8,75
275	13688,00	172	25	2	6	203,00	0,57	1,75	0,19	9,19	9,19
276	13690,00	172	25	2	7	201,40	0,57	1,75	0,19	9,19	9,19
277	13691,00	172	25	2	6	188,30	0,57	1,75	0,19	9,19	9,19
278	13699,00	172	25	2	6	207,00	0,57	1,75	0,19	9,19	9,19
279	13703,00	172	25	2	6	179,90	0,57	1,75	0,19	9,19	9,19
280	13709,00	172	25	2	6	85,30	0,57	1,75	0,19	9,19	9,19
281	13712,00	172	25	2	6	162,30	0,57	1,75	0,19	9,19	9,19
282	13719,00	172	25	2	7	154,50	0,57	1,75	0,19	9,19	9,19
283	14563,00	183	27	2	5	158,10	0,57	1,75	0,20	8,62	8,62
284	14566,00	183	27	2	6	104,60	0,57	1,75	0,20	8,62	8,62

Vista de datos | Vista de variables

SPSS El procesador está listo

Puede ver los resultados del muestreo en el Editor de datos. Se han guardado cinco nuevas variables en el archivo de trabajo, que representan las probabilidades de inclusión y las ponderaciones muestrales acumuladas para cada etapa, además de las ponderaciones muestrales finales.

- Los casos con valores para estas variables se seleccionaron para la muestra.
- Los casos con valores perdidos del sistema para las variables no se seleccionaron.

La agencia ahora utilizará sus recursos para reunir las tasaciones actuales de las propiedades seleccionadas en la muestra. Una vez que estas tasaciones estén disponibles, puede procesar la muestra con los procedimientos de análisis de Muestras complejas, utilizando el plan de muestreo *property_assess.csplan* para proporcionar las especificaciones de muestreo.

Obtención de una muestra a partir de un marco de muestreo parcial

Una compañía está interesada en recopilar y vender una base de datos con información de encuestas de alta calidad. La muestra de la encuesta debe ser representativa, pero ha de llevarse a cabo de manera eficiente, por lo que se utilizan métodos de muestreo complejo. El diseño de muestreo completo requiere la siguiente estructura:

Etapa	Estratos	Agrupaciones
1	Región	Provincia
2	Distrito	Ciudad
3	Subdivisión	

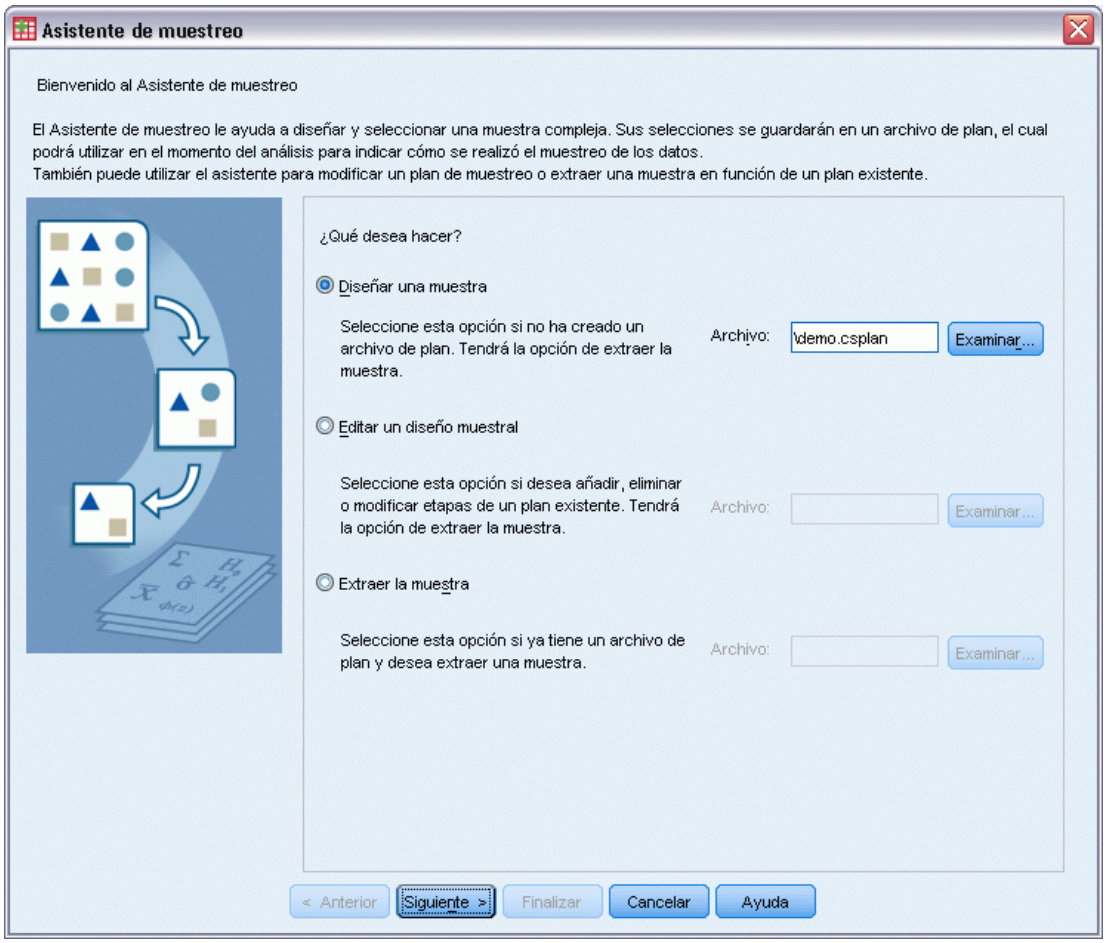
En la tercera etapa, las unidades familiares son la unidad muestral primaria y se realizarán encuestas a las unidades familiares seleccionadas. Sin embargo, dado que sólo se puede disponer con facilidad de la información de ciudad, la compañía tiene pensado llevar a cabo las dos primeras etapas del diseño ahora y, a continuación, recopilar la información sobre el número de subdivisiones y unidades familiares de las ciudades muestreadas. La información disponible acerca de las ciudades se incluye en *demo_cs_1.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*. Tenga en cuenta que este archivo contiene una variable *Subdivisión* que sólo contiene el valor 1. Es un marcador de posición para la variable “verdadera”, cuyos valores se recopilan después de ejecutar las dos primeras etapas del diseño, que permite especificar ahora el diseño de muestreo de tres etapas completo. Utilice el Asistente de muestreo de la opción Muestras complejas para especificar el diseño de muestreo complejo completo y, a continuación, extraiga las dos primeras etapas.

Uso del asistente para extraer la muestra del primer marco parcial

▶ Para ejecutar el Asistente de muestreo de la opción de Muestras complejas, seleccione en los menús:

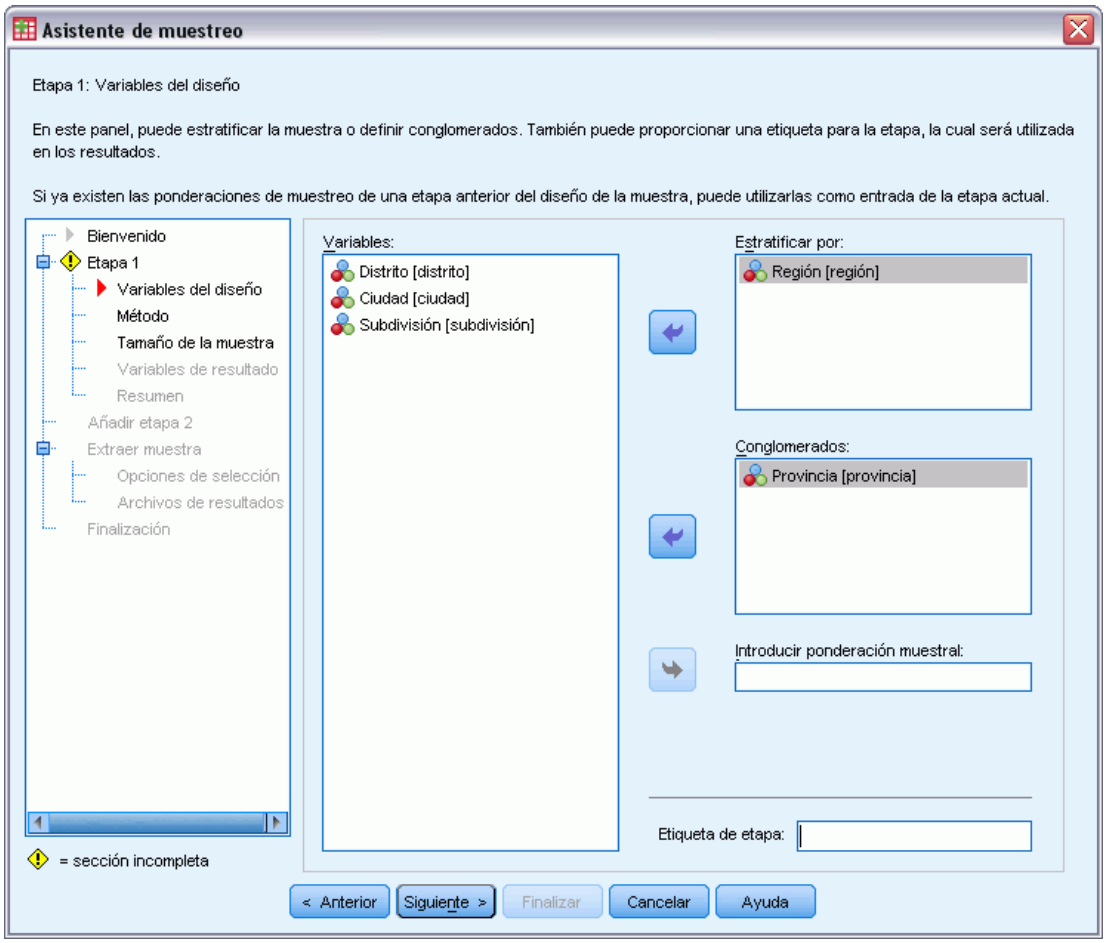
Analizar > Muestras complejas > Seleccionar una muestra...

Figura 13-14
Asistente de muestreo: paso Bienvenida



- Seleccione Diseñar una muestra, navegue hasta la ubicación en la que desee guardar el archivo e introduzca demo.csplan como el nombre del archivo de plan.
- Pulse en Siguiente.

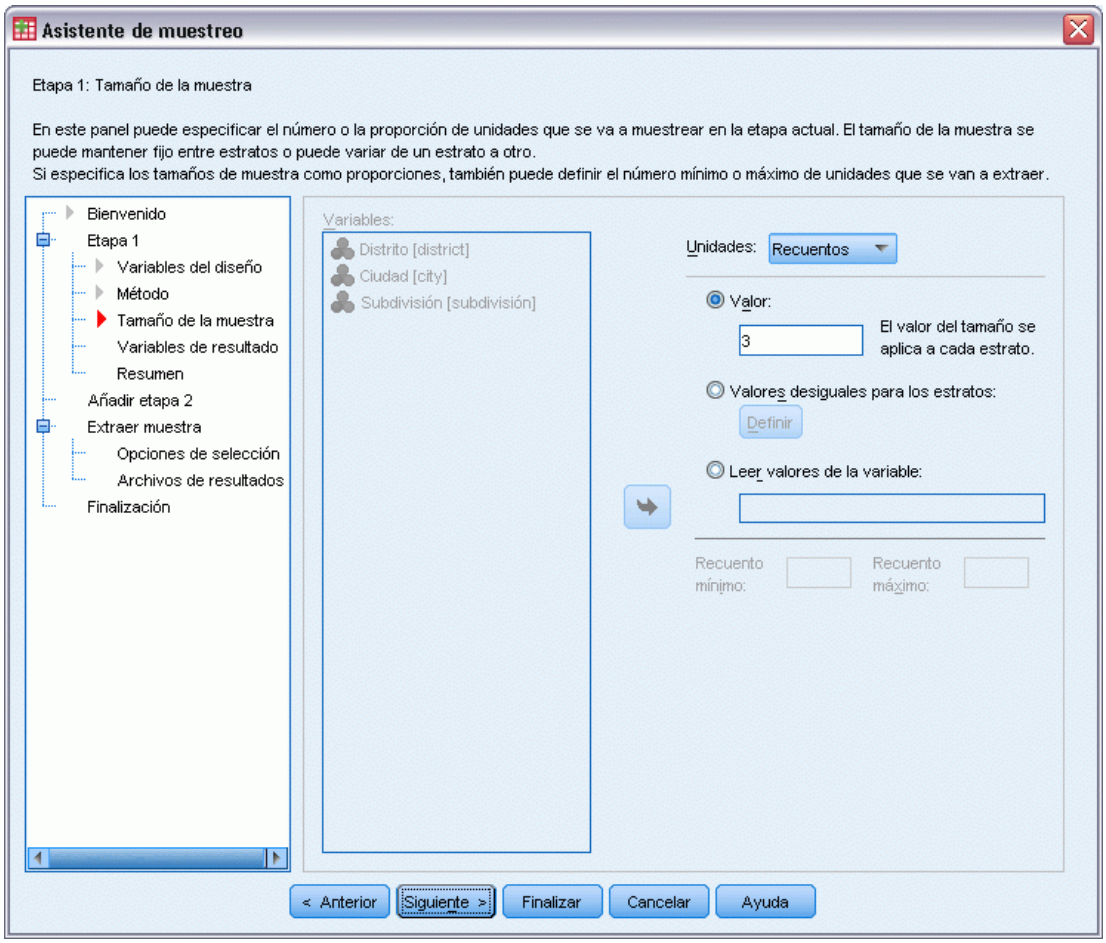
Figura 13-15
Asistente de muestreo: paso Variables del diseño (etapa 1)



- Seleccione *Región* como variable de estratificación.
- Seleccione *Provincia* como variable de conglomeración.
- Pulse en Siguiente y, a continuación, pulse en Siguiente en el paso Método de muestreo.

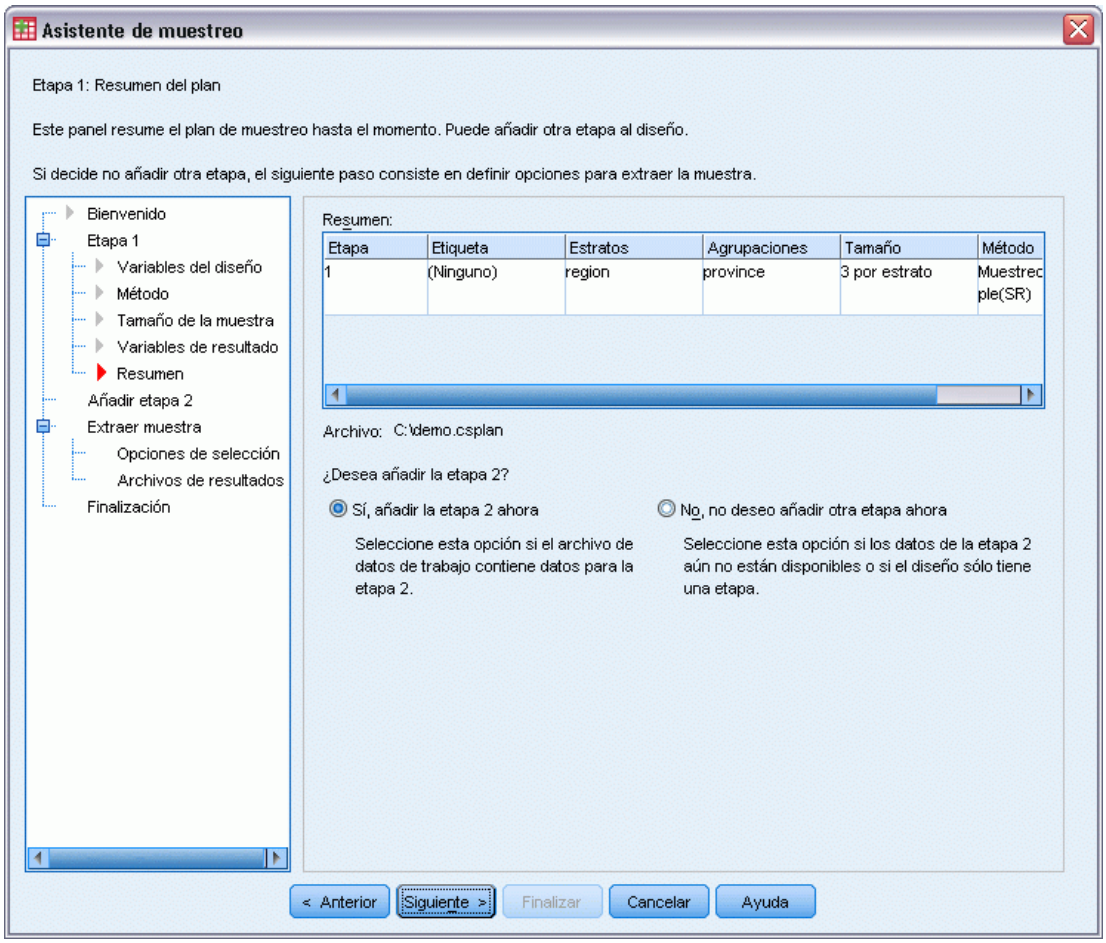
  Esta estructura de diseño indica que se extraen muestras independientes para cada región. En esta etapa, se extraen las provincias como unidad muestral primaria mediante el método por defecto: Muestreo aleatorio simple.

Figura 13-16
Asistente de muestreo: paso Tamaño muestral (etapa 1)



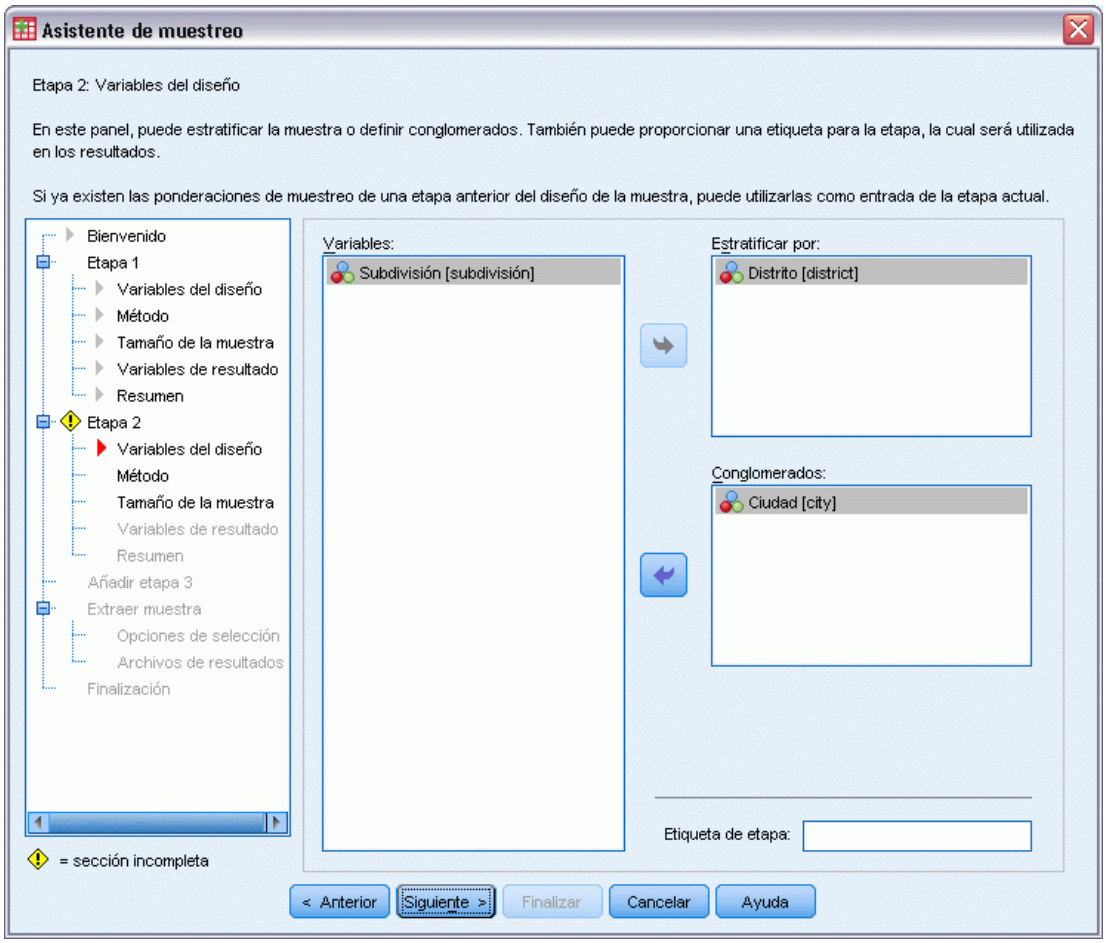
- Seleccione Recuentos en la lista desplegable Unidades.
- Escriba 3 como el valor del número de unidades que se van a seleccionar en esta etapa.
- Pulse en Siguiente y, a continuación, pulse en Siguiente en el paso Variables de resultado.

Figura 13-17
Asistente de muestreo: paso Resumen del plan (etapa 1)



- Seleccione Sí, añadir la etapa 2 ahora.
- Pulse en Siguiente.

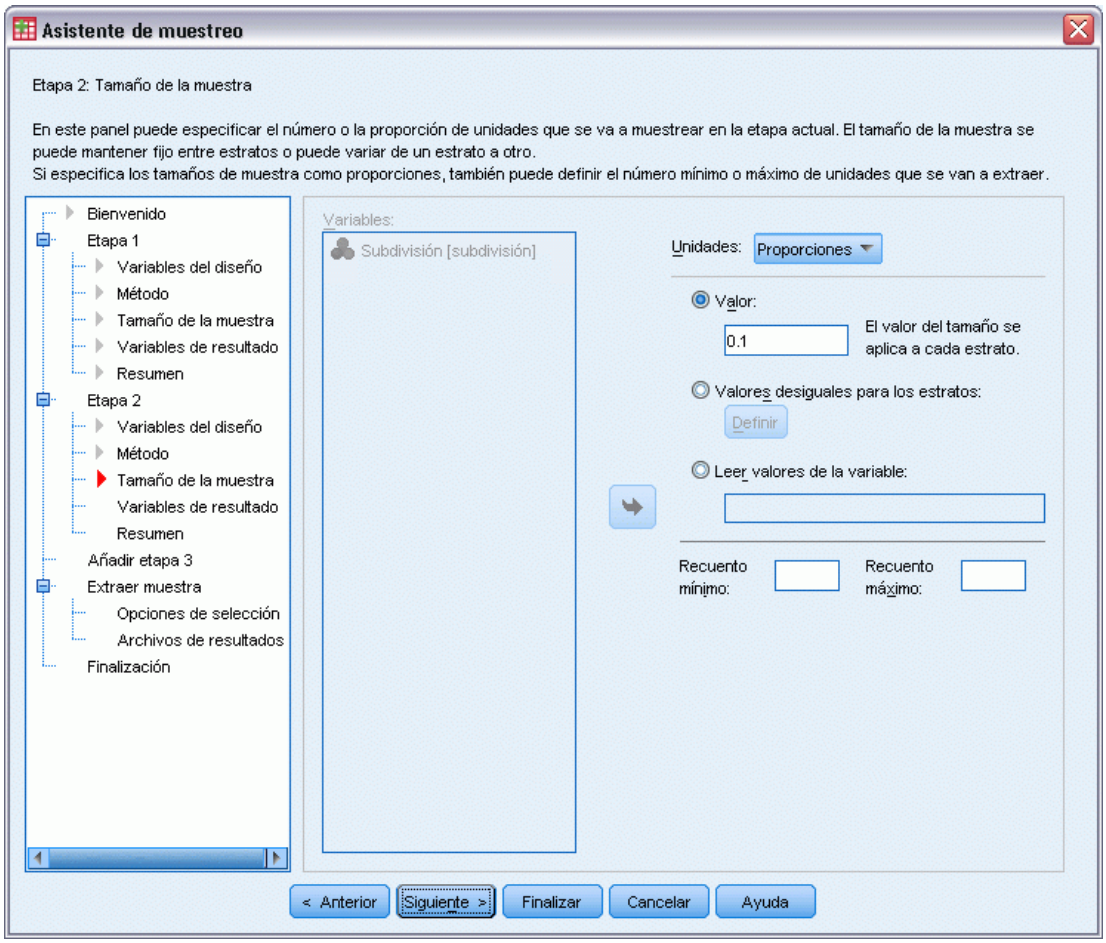
Figura 13-18
Asistente de muestreo: paso Variables del diseño (etapa 2)



- Seleccione *Distrito* como variable de estratificación.
- Seleccione *Ciudad* como variable de conglomeración.
- Pulse en Siguiente y, a continuación, pulse en Siguiente en el paso Método de muestreo.

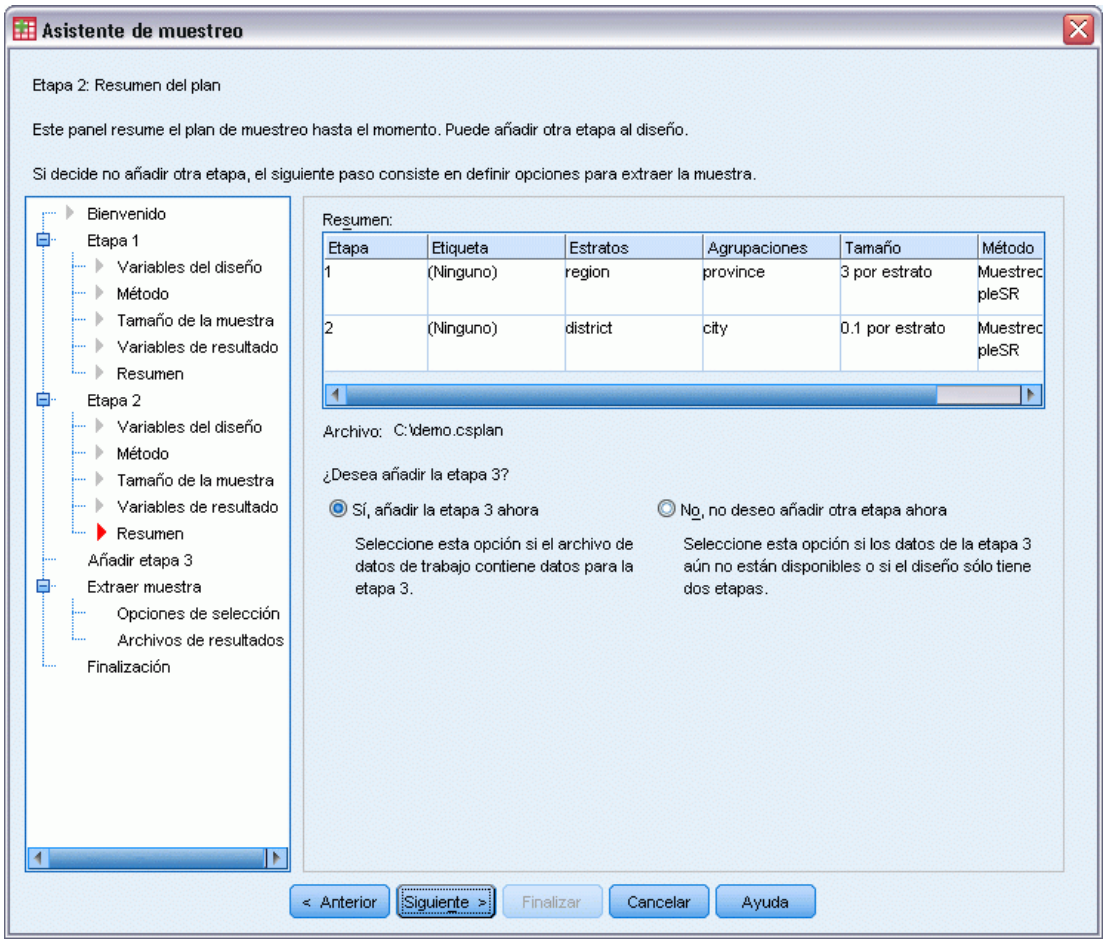
  Esta estructura de diseño indica que se extraen muestras independientes para cada distrito. En esta etapa, se extraen las ciudades como unidad muestral primaria mediante el método por defecto: Muestreo aleatorio simple.

Figura 13-19
Asistente de muestreo: paso Tamaño muestral (etapa 2)



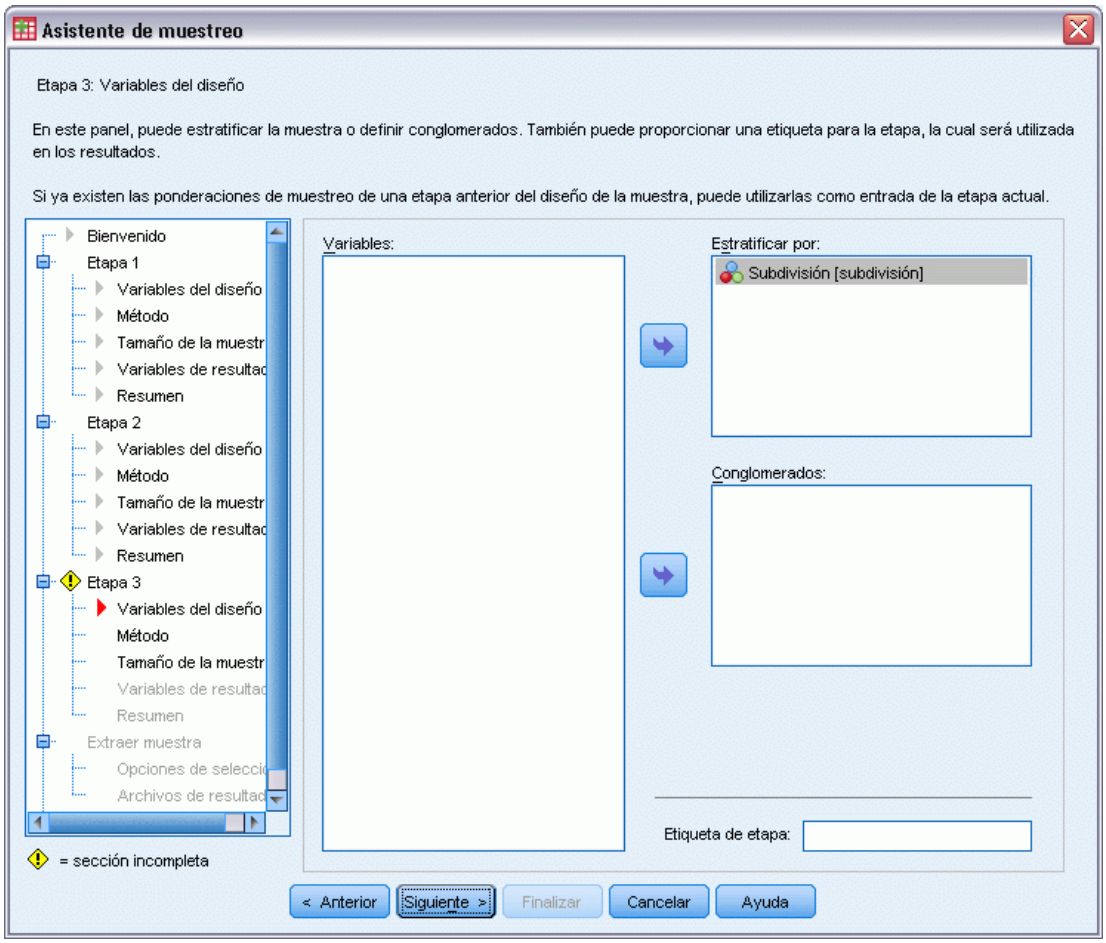
- Seleccione Proporciones en la lista desplegable Unidades.
- Escriba 0,1 como valor de la proporción de unidades que se van a extraer como muestra de cada estrato.
- Pulse en Siguiente y, a continuación, pulse en Siguiente en el paso Variables de resultado.

Figura 13-20
Asistente de muestreo: paso Resumen del plan (etapa 2)



- Seleccione Sí, añadir la etapa 3 ahora.
- Pulse en Siguiente.

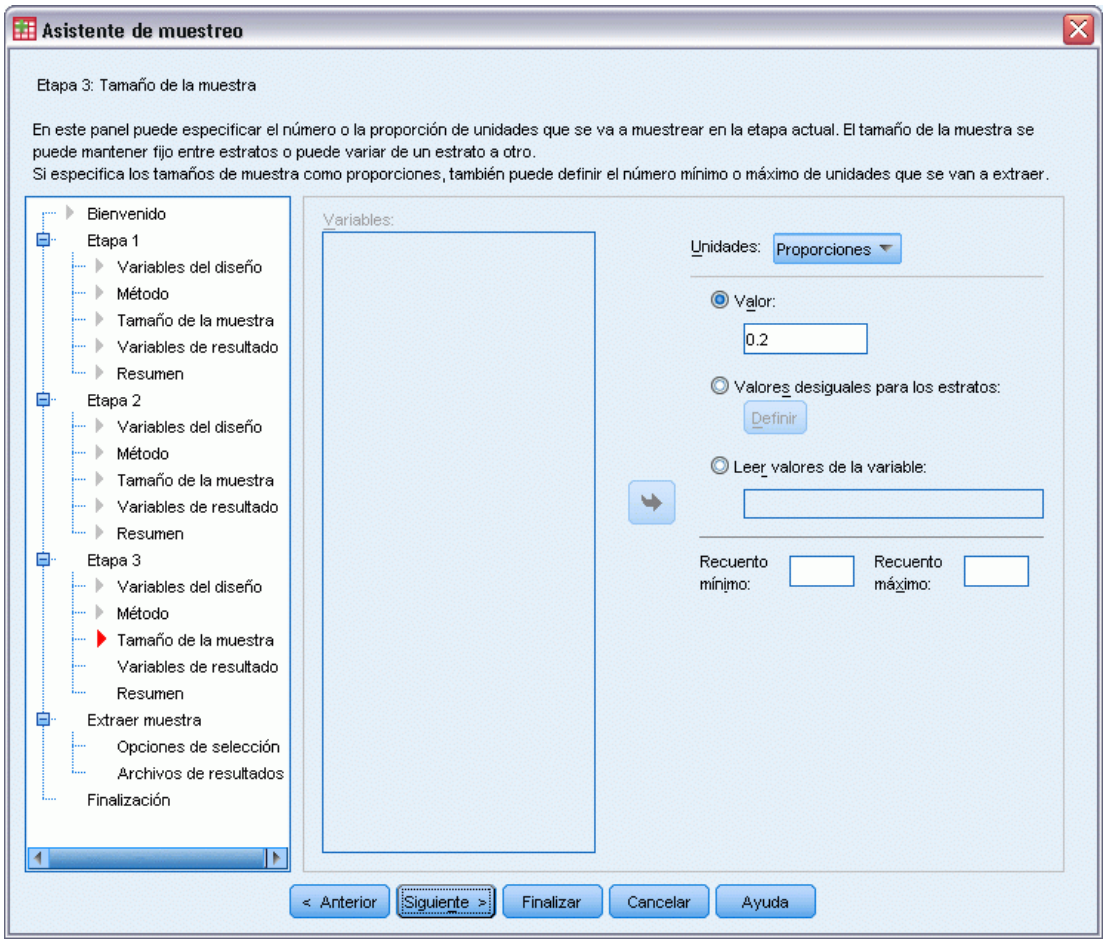
Figura 13-21
Asistente de muestreo: paso Variables del diseño (etapa 3)



- Seleccione *Subdivisión* como variable de estratificación.
- Pulse en Siguiente y, a continuación, pulse en Siguiente en el paso Método de muestreo.

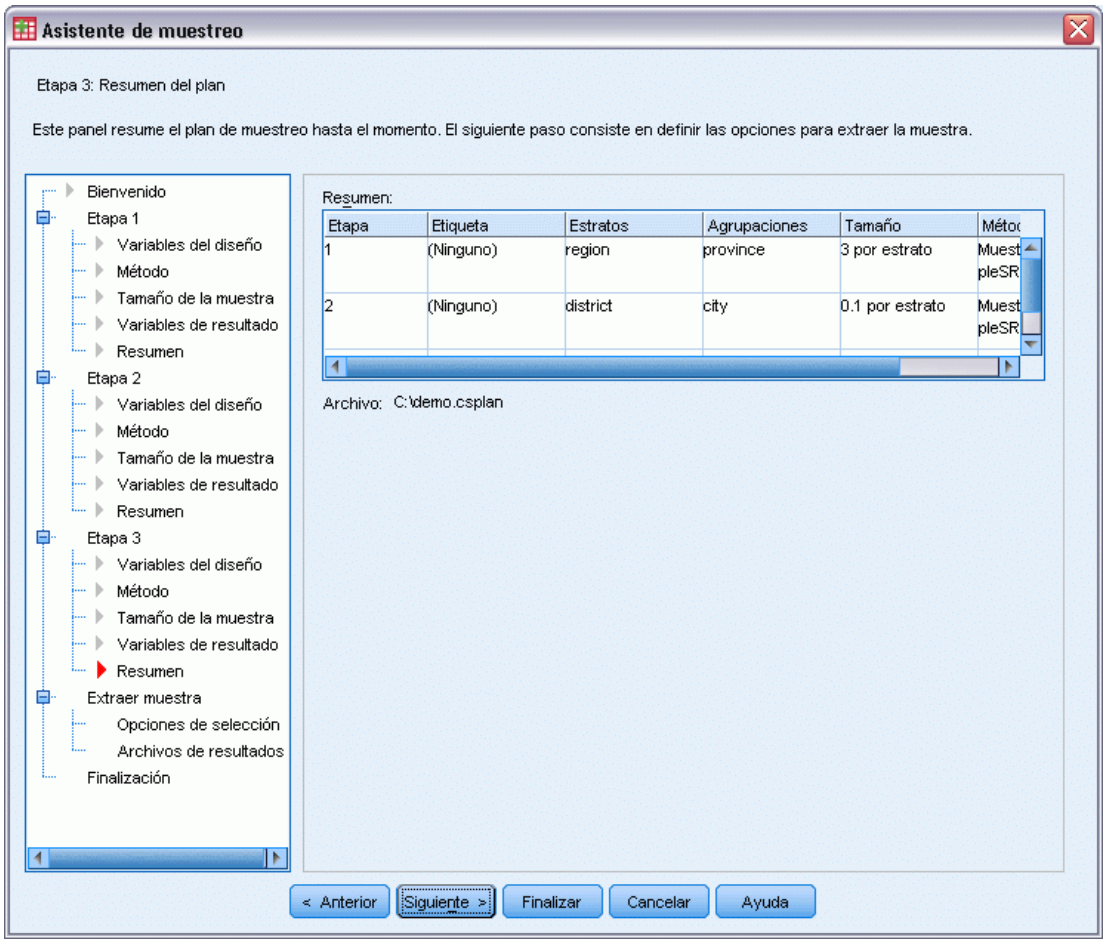
  Esta estructura de diseño indica que se extraen muestras independientes para cada subdivisión. En esta etapa, se extraen las unidades familiares como unidad muestral primaria mediante el método por defecto: Muestreo aleatorio simple.

Figura 13-22
Asistente de muestreo: paso Tamaño muestral (etapa 3)



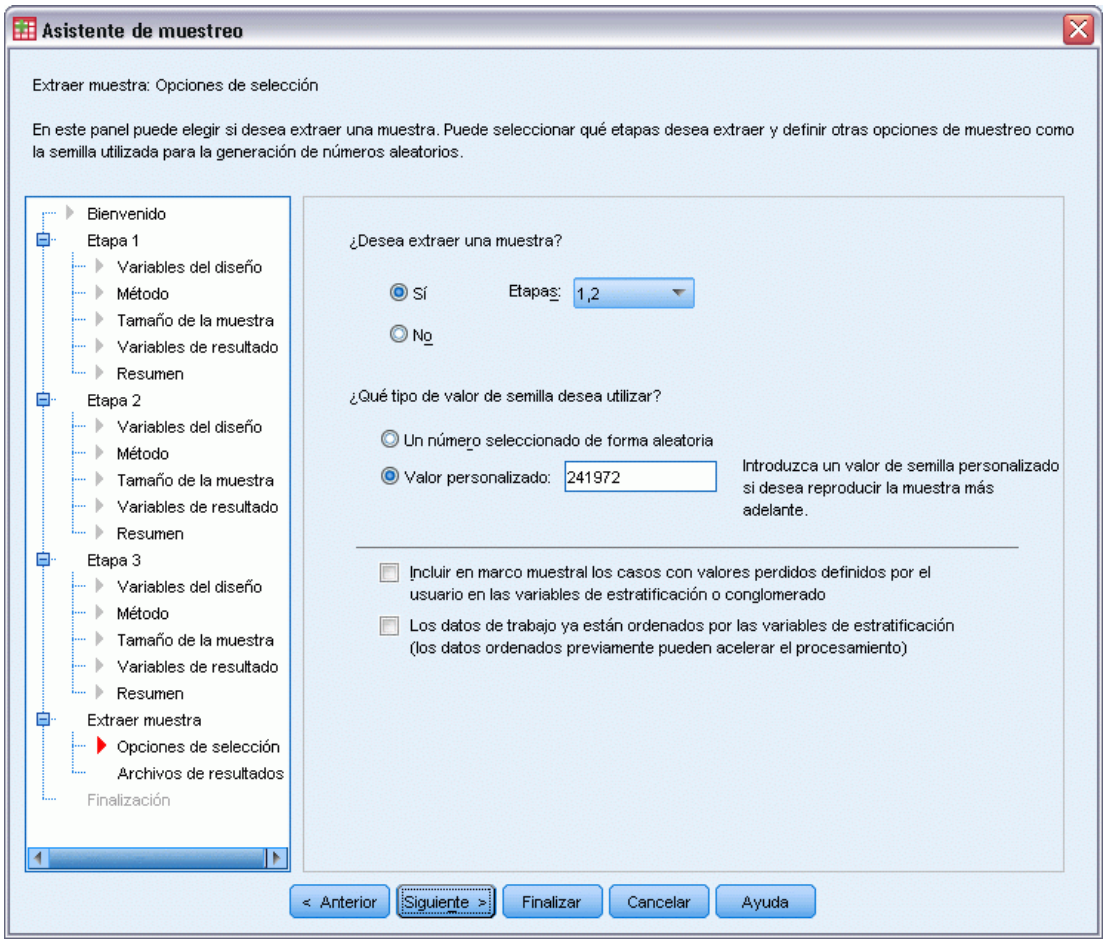
- Seleccione Proporciones en la lista desplegable Unidades.
- Escriba 0,2 como el valor de la proporción de unidades que se van a seleccionar en esta etapa.
- Pulse en Siguiente y, a continuación, pulse en Siguiente en el paso Variables de resultado.

Figura 13-23
Asistente de muestreo: paso Resumen del plan (etapa 3)



▶ Revise el diseño muestral y, a continuación, pulse en Siguiente.

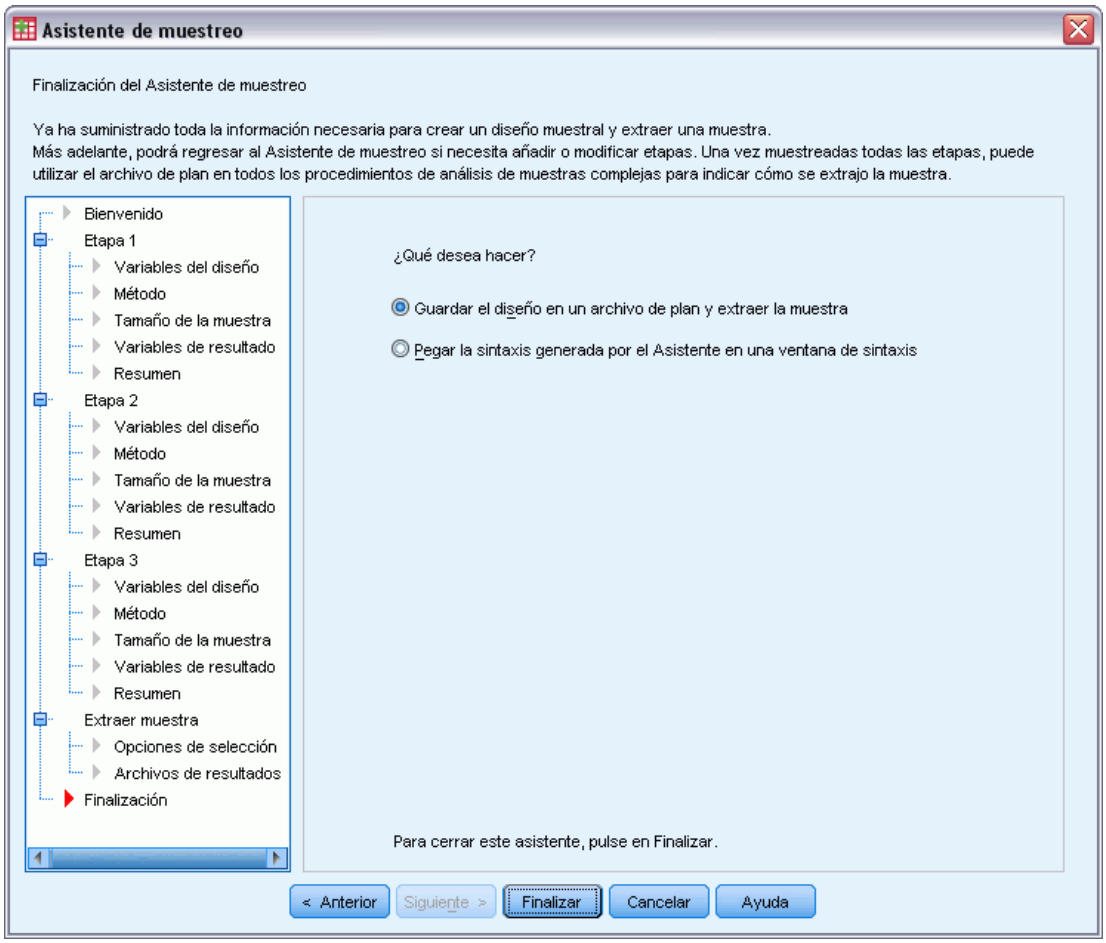
Figura 13-24
Asistente de muestreo: Extraer muestra: paso Opciones de selección



- Seleccione 1, 2 como las etapas que se van a extraer como muestra ahora.
- Seleccione Valor personalizado como tipo de semilla aleatoria que se va a utilizar y escriba 241972 como valor.

  Al utilizar un valor personalizado, es posible replicar los resultados de este ejemplo de manera exacta.
- Pulse en Siguiente y, a continuación, pulse en Siguiente en Extraer muestra: paso Archivos de resultados.

Figura 13-25
Asistente de muestreo: paso Finalizar

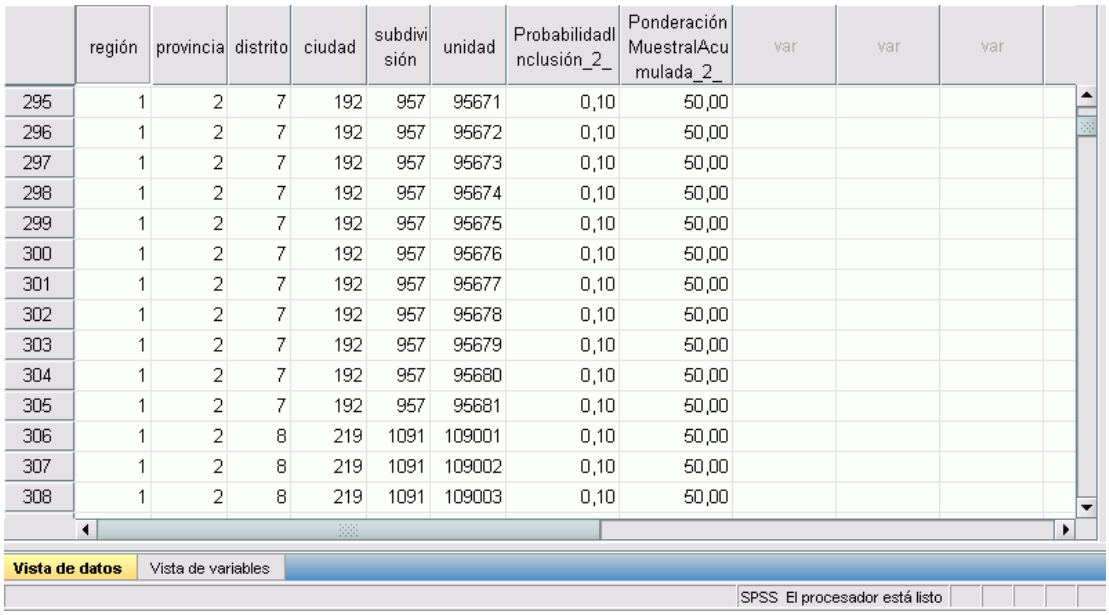


▶ Pulse en Finalizar.

Estas selecciones generan el archivo de plan de muestreo *demo.csplan* y extraen una muestra de acuerdo con las primeras dos etapas del plan.

Resultados de la muestra

Figura 13-26
Editor de datos con los resultados de la muestra



	región	provincia	distrito	ciudad	subdivisión	unidad	Probabilidadlnclusión_2_	PonderaciónMuestralAcumulada_2_	var	var	var	
295	1	2	7	192	957	95671	0,10	50,00				
296	1	2	7	192	957	95672	0,10	50,00				
297	1	2	7	192	957	95673	0,10	50,00				
298	1	2	7	192	957	95674	0,10	50,00				
299	1	2	7	192	957	95675	0,10	50,00				
300	1	2	7	192	957	95676	0,10	50,00				
301	1	2	7	192	957	95677	0,10	50,00				
302	1	2	7	192	957	95678	0,10	50,00				
303	1	2	7	192	957	95679	0,10	50,00				
304	1	2	7	192	957	95680	0,10	50,00				
305	1	2	7	192	957	95681	0,10	50,00				
306	1	2	8	219	1091	109001	0,10	50,00				
307	1	2	8	219	1091	109002	0,10	50,00				
308	1	2	8	219	1091	109003	0,10	50,00				

Puede ver los resultados del muestreo en el Editor de datos. Se han guardado cinco nuevas variables en el archivo de trabajo, que representan las probabilidades de inclusión y las ponderaciones muestrales acumuladas para cada etapa, además de las ponderaciones muestrales "finales" de las dos primeras etapas.

- Las ciudades con valores para estas variables se seleccionaron para la muestra.
- Las ciudades con valores perdidos del sistema para las variables no se seleccionaron.

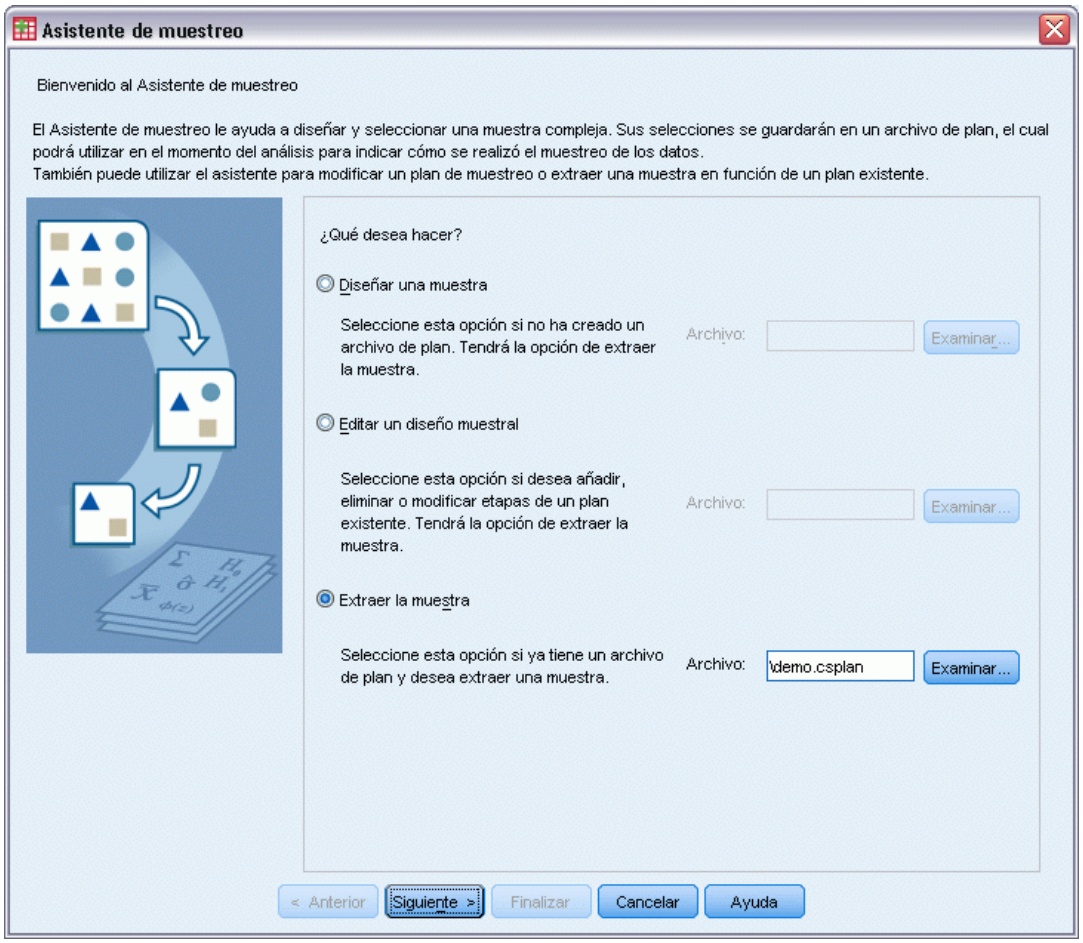
Para cada ciudad seleccionada, la compañía adquirió información sobre subdivisiones y unidades familiares y la colocó en *demo_cs_2.sav*. Utilice este archivo y el Asistente de muestreo para extraer la muestra de la tercera etapa de este diseño.

Uso del asistente para extraer la muestra del segundo marco parcial

► Para ejecutar el Asistente de muestreo de la opción de Muestras complejas, seleccione en los menús:

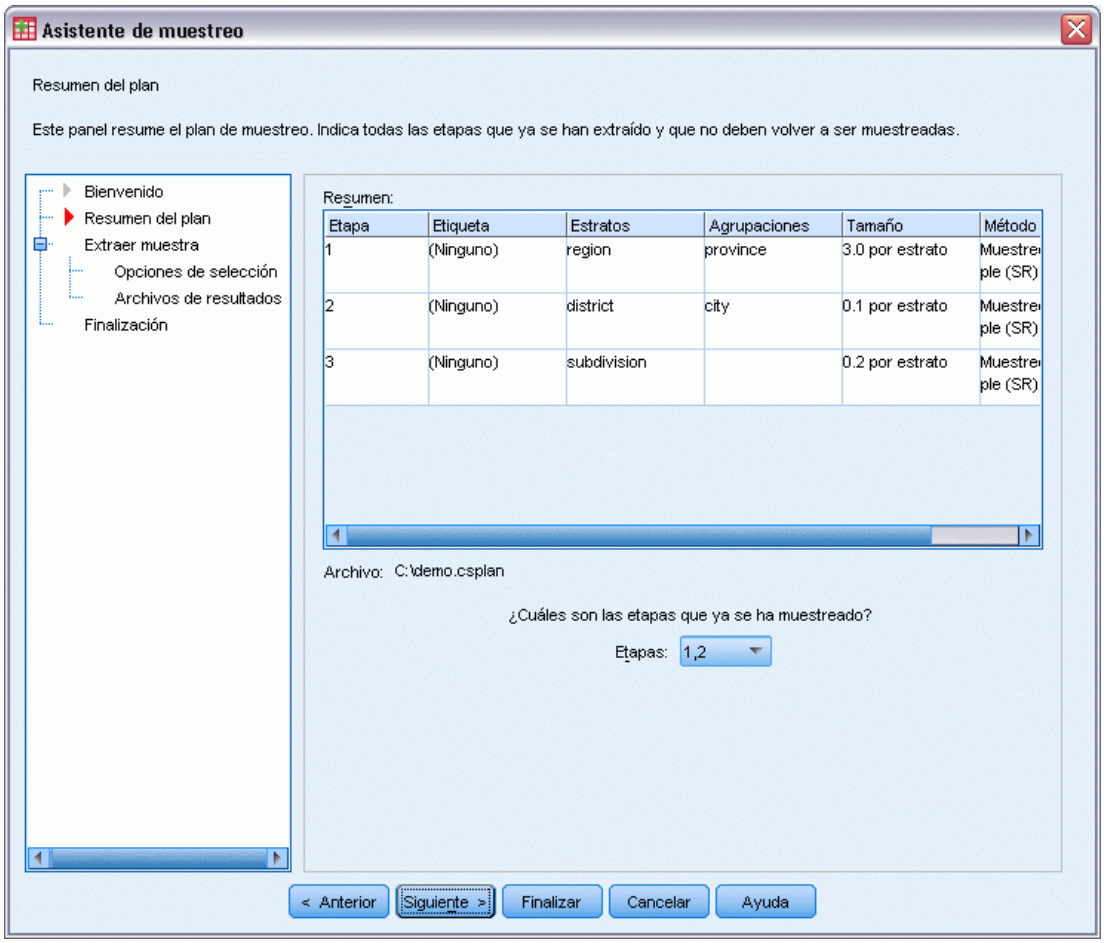
Analizar > Muestras complejas > Seleccionar una muestra...

Figura 13-27
Asistente de muestreo: paso Bienvenida



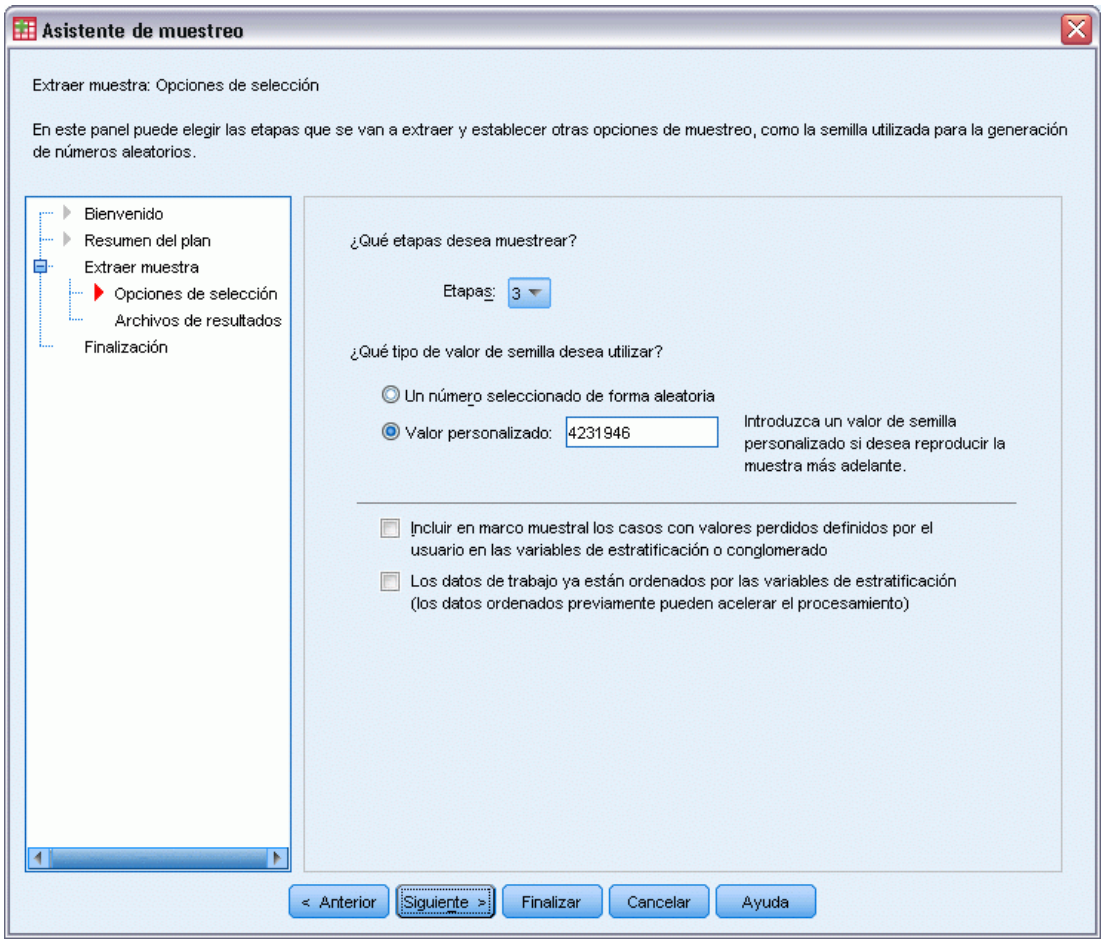
- Seleccione Extraer una muestra, desplácese hasta la ubicación en la que ha guardado el archivo de plan y seleccione el archivo de plan demo.csplan que ha creado.
- Pulse en Siguiente.

Figura 13-28
Asistente de muestreo: paso Resumen del plan (etapa 3)



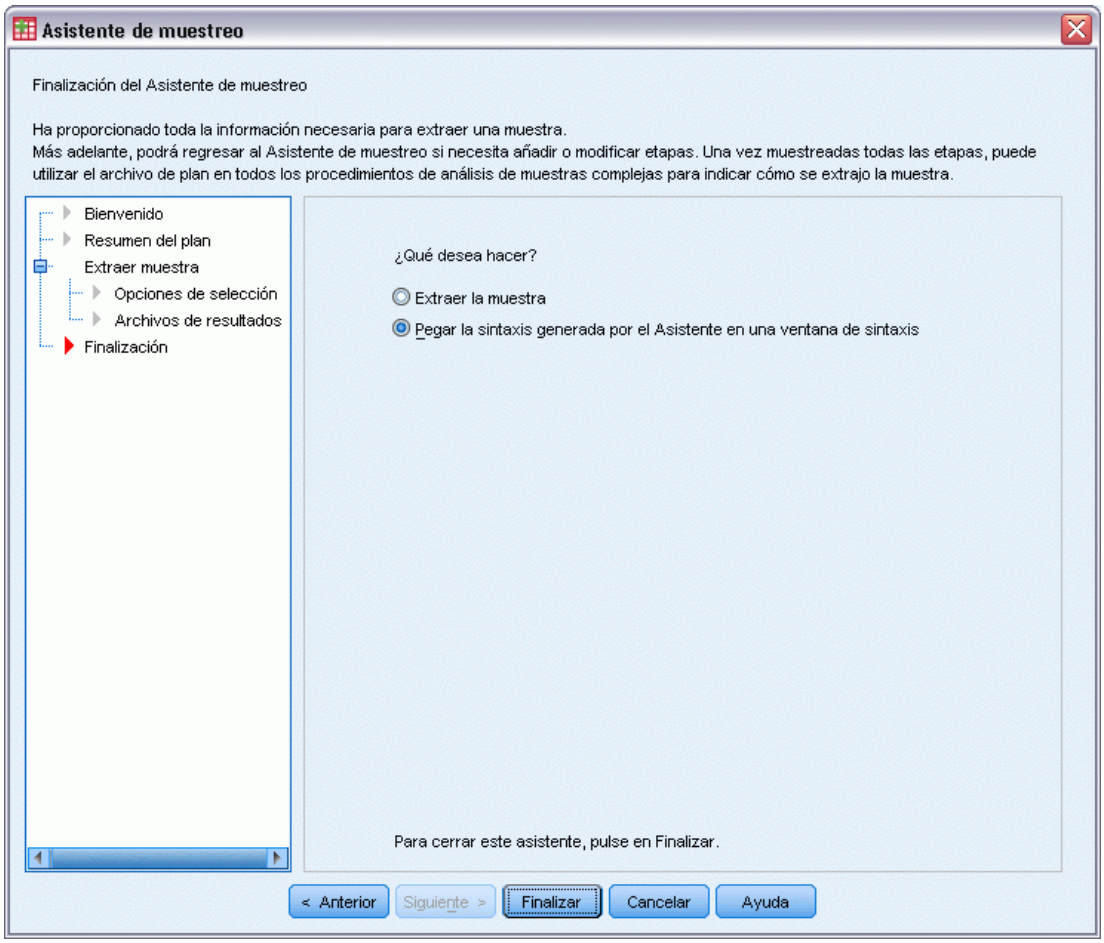
- Seleccione 1, 2 como las etapas que ya se han muestreado.
- Pulse en Siguiente.

Figura 13-29
Asistente de muestreo: Extraer muestra: paso Opciones de selección



- Seleccione Valor personalizado como tipo de semilla aleatoria que se va a utilizar y escriba 4231946 como valor.
- Pulse en Siguiente y, a continuación, pulse en Siguiente en Extraer muestra: paso Archivos de resultados.

Figura 13-30
Asistente de muestreo: paso Finalizar



- Seleccione Pegar la sintaxis generada por el Asistente en una ventana de sintaxis.
- Pulse en Finalizar.

Se genera la siguiente sintaxis:

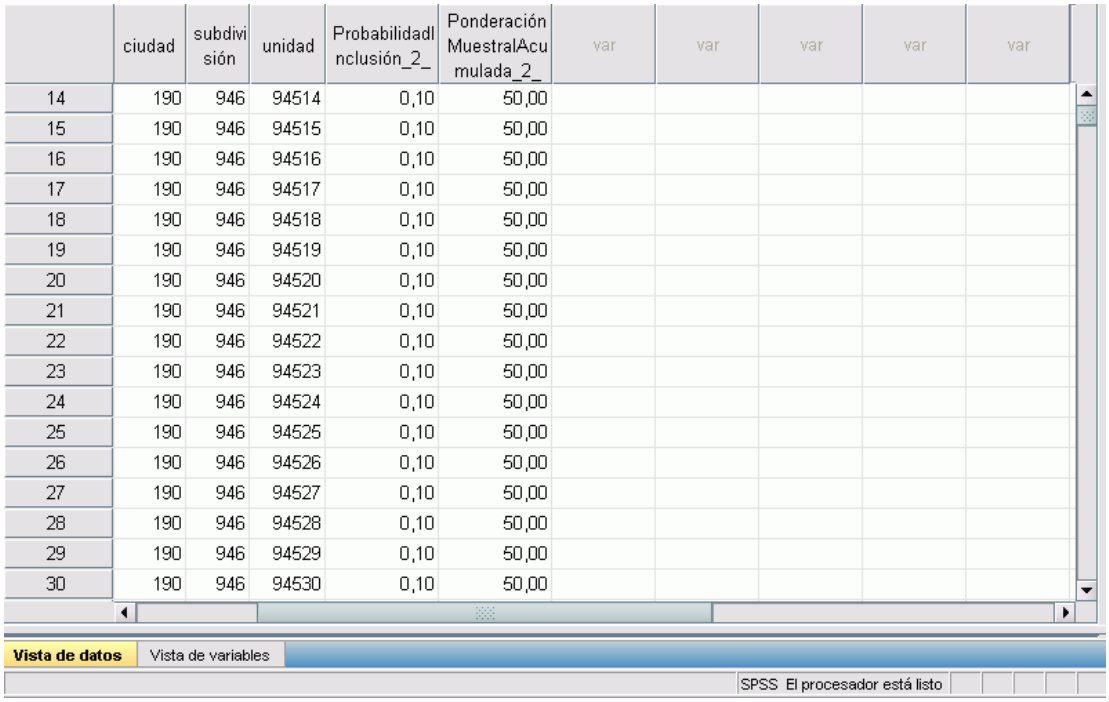
```
* Asistente de muestreo.
CSSELECT
  /PLAN FILE='demo.csplan'
  /CRITERIA STAGES = 3 SEED = 4231946
  /CLASSMISSING EXCLUDE
  /DATA RENAMEVARS
  /PRINT SELECTION.
```

La impresión del resumen de muestreo en este caso produce una tabla confusa que provoca problemas en el Visor de resultados. Para desactivar la presentación del resumen de muestreo, reemplace `SELECTION` por `CPS` en el subcomando `PRINT`. A continuación, ejecute la sintaxis en la ventana de sintaxis.

Estas selecciones extraen una muestra de acuerdo con la tercera etapa del plan de muestreo *demo.csplan*.

Resultados de la muestra

Figura 13-31
Editor de datos con los resultados de la muestra



	ciudad	subdivisión	unidad	ProbabilidadInclusión_2_	PonderaciónMuestralAcumulada_2_	var	var	var	var	var
14	190	946	94514	0,10	50,00					
15	190	946	94515	0,10	50,00					
16	190	946	94516	0,10	50,00					
17	190	946	94517	0,10	50,00					
18	190	946	94518	0,10	50,00					
19	190	946	94519	0,10	50,00					
20	190	946	94520	0,10	50,00					
21	190	946	94521	0,10	50,00					
22	190	946	94522	0,10	50,00					
23	190	946	94523	0,10	50,00					
24	190	946	94524	0,10	50,00					
25	190	946	94525	0,10	50,00					
26	190	946	94526	0,10	50,00					
27	190	946	94527	0,10	50,00					
28	190	946	94528	0,10	50,00					
29	190	946	94529	0,10	50,00					
30	190	946	94530	0,10	50,00					

Vista de datos | Vista de variables

SPSS El procesador está listo

Puede ver los resultados del muestreo en el Editor de datos. Se han guardado tres nuevas variables en el archivo de trabajo, que representan las probabilidades de inclusión y las ponderaciones muestrales acumuladas de la tercera etapa, además de las ponderaciones muestrales finales. Estas nuevas ponderaciones tienen en cuenta los pesos calculados durante el muestreo de las dos primeras etapas.

- Las unidades con valores para estas variables se seleccionaron para la muestra.
- Las unidades con valores perdidos del sistema para estas variables no se seleccionaron.

La compañía utilizará ahora sus recursos para obtener información mediante encuestas acerca de las unidades familiares seleccionadas en la muestra. Una vez que se recopilen estas encuestas, puede procesar la muestra con los procedimientos de análisis de Muestras complejas, utilizando el plan de muestreo *demo.csplan* para proporcionar las especificaciones de muestreo.

Muestreo con probabilidad proporcional al tamaño (PPS)

Los diputados que estudian un proyecto de ley antes de una asamblea legislativa se interesan por conocer si la opinión pública apoya dicho proyecto de ley y qué relación guarda dicho apoyo con los datos demográficos de los votantes. Los encuestadores diseñan entrevistas y las realizan siguiendo un diseño muestral complejo.

Se incluye una lista de votantes registrados en *poll_cs.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*. Utilice el Asistente de muestreo de la opción Muestras complejas para seleccionar una muestra y llevar a cabo su posterior análisis.

Uso del asistente

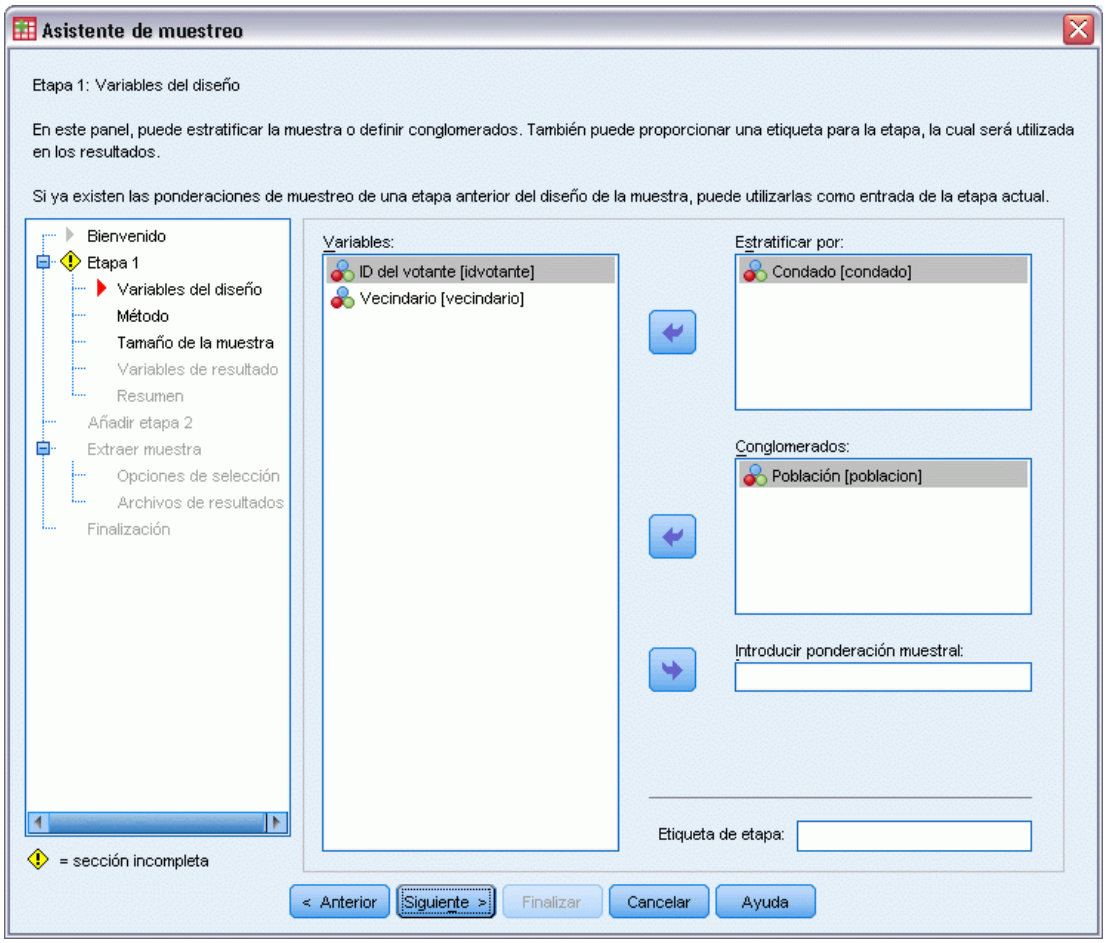
- Para ejecutar el Asistente de muestreo de la opción de Muestras complejas, seleccione en los menús:
  Analizar > Muestras complejas > Seleccionar una muestra...

Figura 13-32
Asistente de muestreo: paso Bienvenida

- Seleccione Diseñar una muestra, navegue hasta la ubicación en la que desee guardar el archivo e introduzca poll.csplan como el nombre del archivo de plan.

- Pulse en Siguiente.

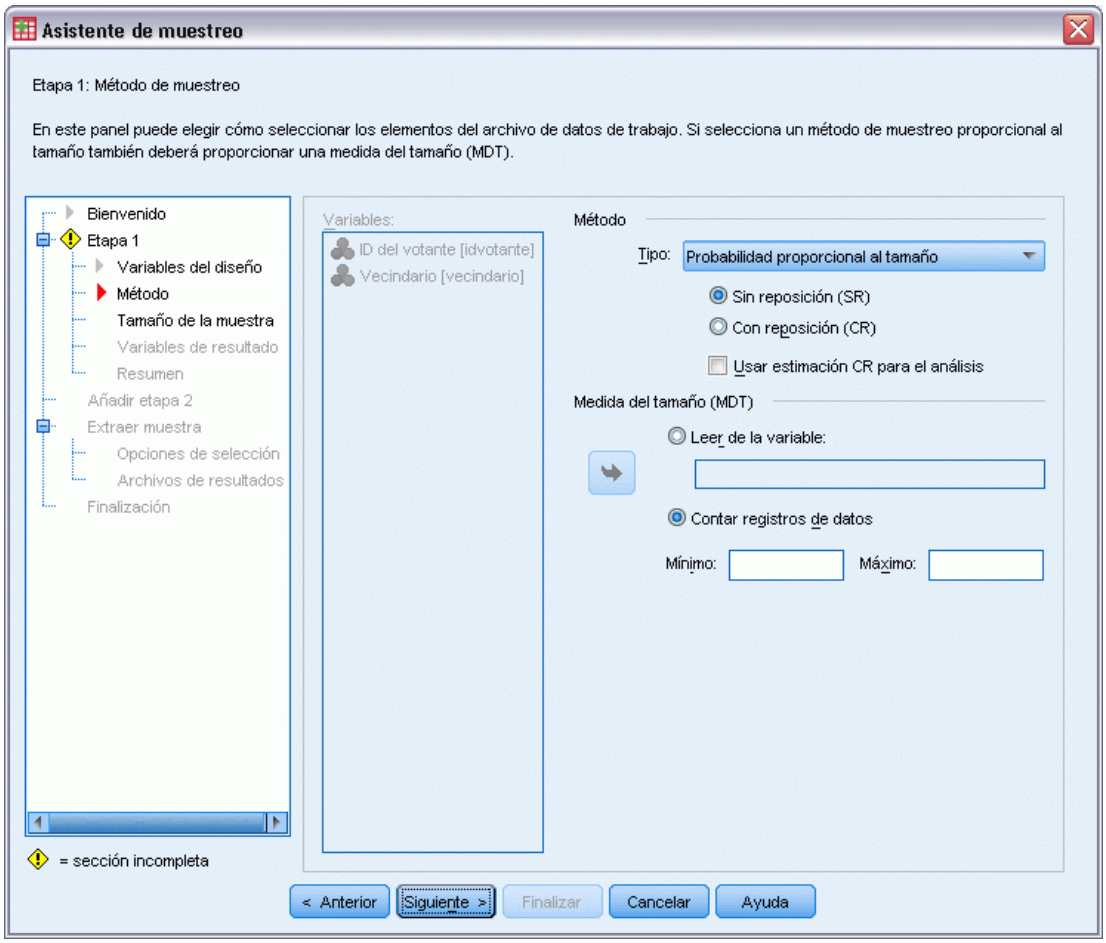
Figura 13-33
Asistente de muestreo: paso Variables del diseño (etapa 1)



- Seleccione *Condado* como variable de estratificación.
- Seleccione *Población* como variable de conglomeración.
- Pulse en Siguiente.

Esta estructura de diseño indica que se extraen muestras independientes para cada condado. En esta etapa, las poblaciones se extraen como la unidad muestral primaria.

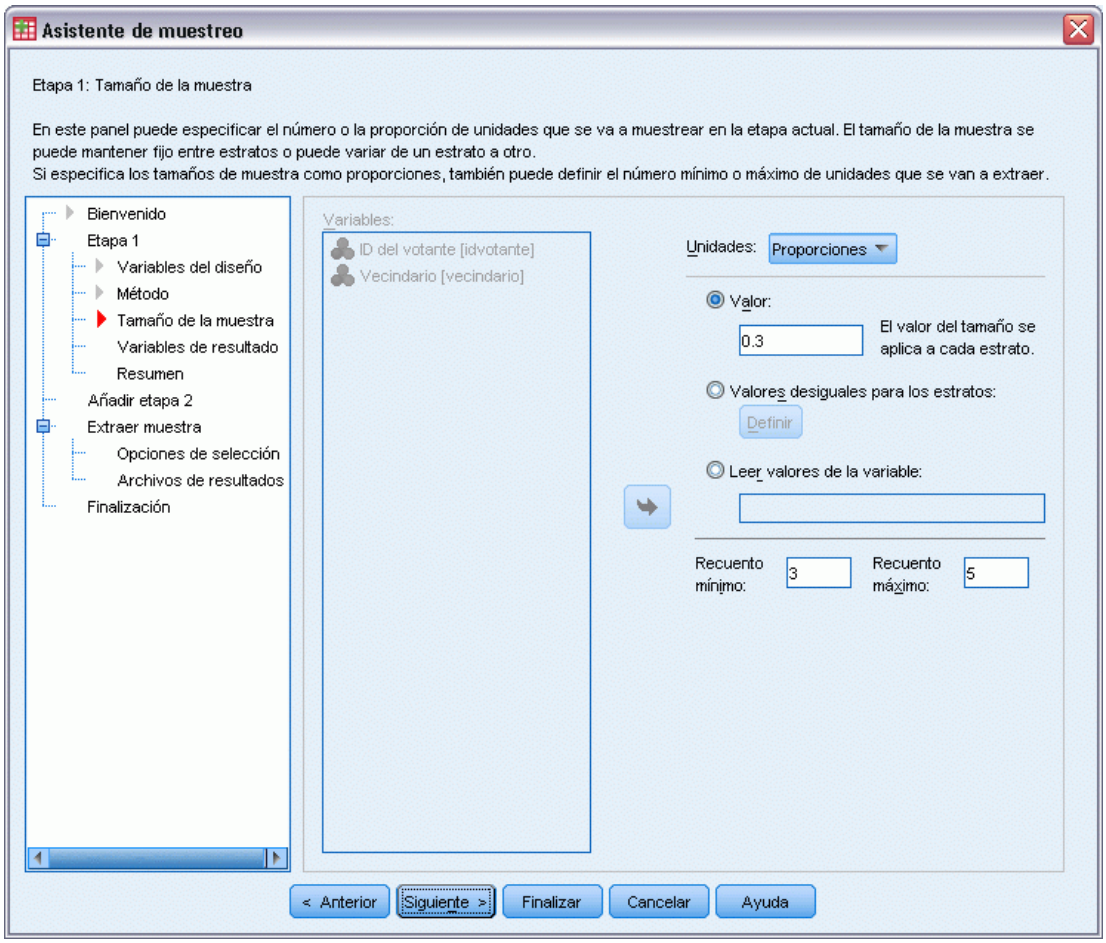
Figura 13-34
Asistente de muestreo: paso Método de muestreo (etapa 1)



- Seleccione PPS como método de muestreo.
- Seleccione Contar registros de datos como medida de tamaño.
- Pulse en Siguiente.

En cada condado, las poblaciones se extraen sin reposición con una probabilidad proporcional al número de registros para cada población. El método PPS genera probabilidades de muestreo conjuntas para las poblaciones; el paso Archivos de resultado permite especificar dónde se van a guardar estos valores.

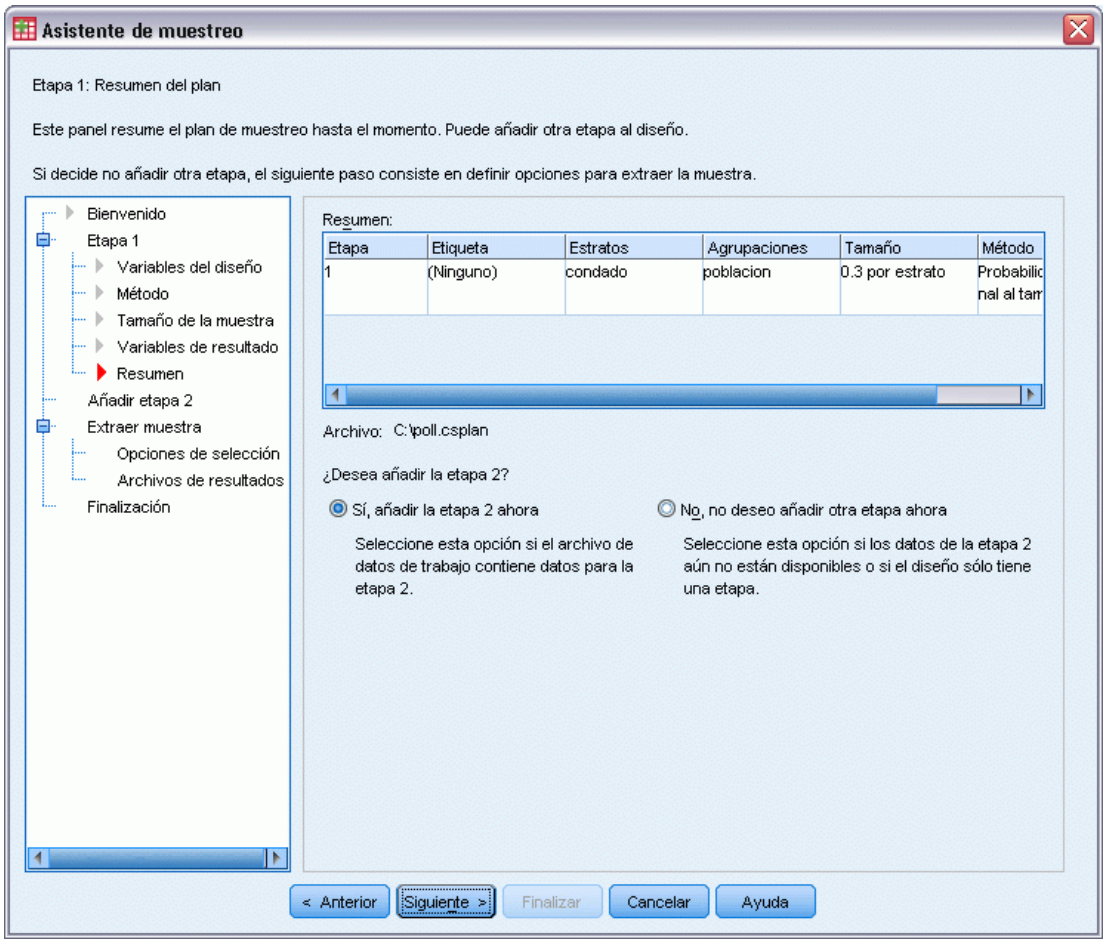
Figura 13-35
Asistente de muestreo: paso Tamaño muestral (etapa 1)



- Seleccione Proporciones en la lista desplegable Unidades.
- Escriba 0,3 como el valor de la proporción de poblaciones que se van a seleccionar por condado en esta etapa.

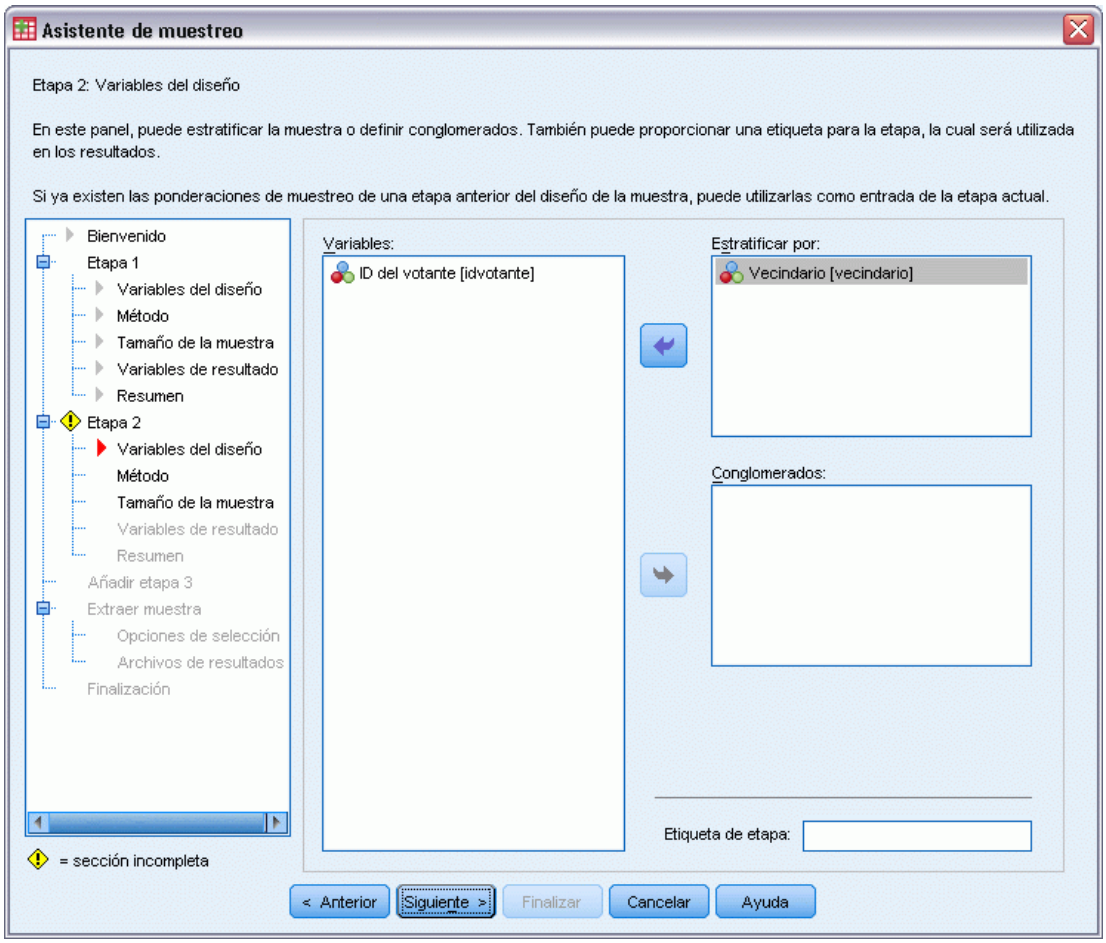
  Los legisladores del condado del oeste señalan que hay menos poblaciones en su condado que en otros. Para asegurar una representación adecuada, desean establecer un mínimo de 3 poblaciones muestreadas de cada condado.
- Escriba 3 como el número de poblaciones mínimo para seleccionar y 5 como el número máximo.
- Pulse en Siguiente y, a continuación, pulse en Siguiente en el paso Variables de resultado.

Figura 13-36
Asistente de muestreo: paso Resumen del plan (etapa 1)



- Seleccione Sí, añadir la etapa 2 ahora.
- Pulse en Siguiente.

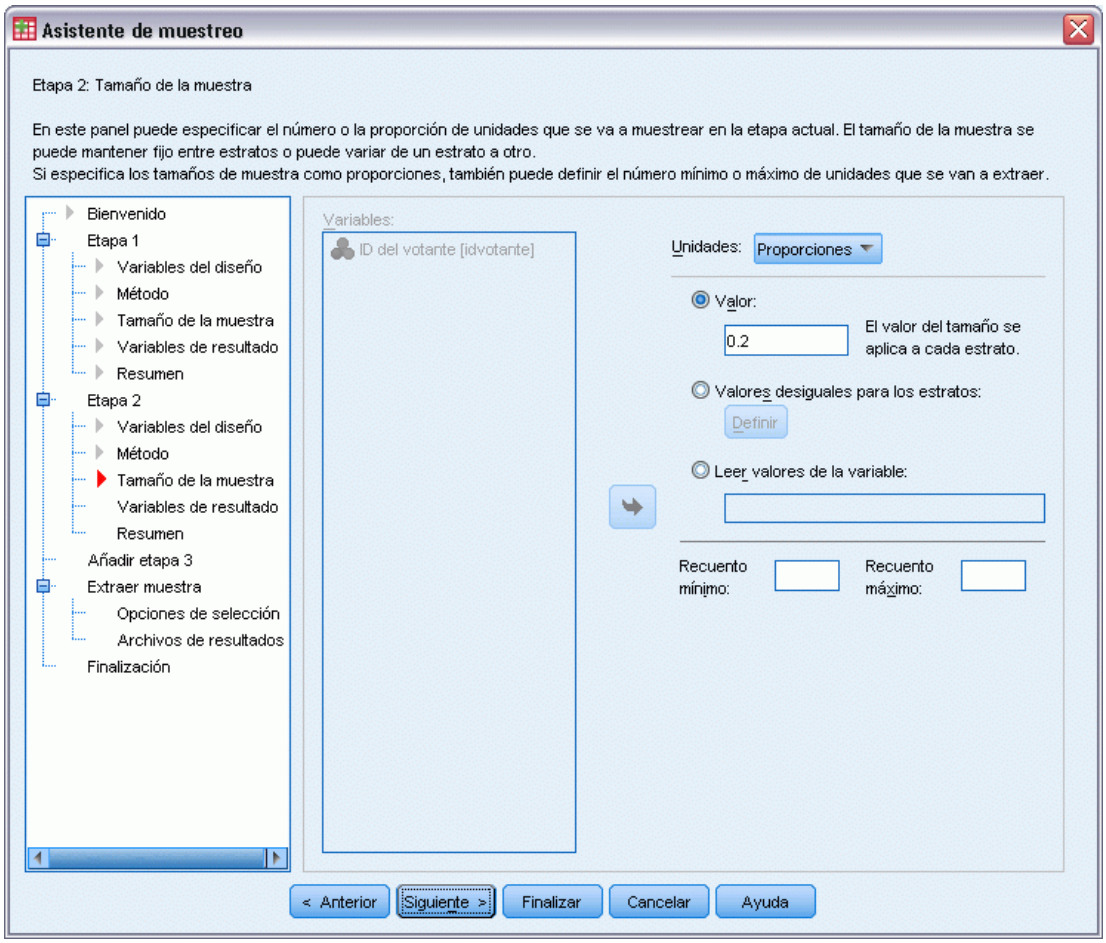
Figura 13-37
Asistente de muestreo: paso Variables del diseño (etapa 2)



- Seleccione *Vecindario* como variable de estratificación.
- Pulse en Siguiente y, a continuación, pulse en Siguiente en el paso Método de muestreo.

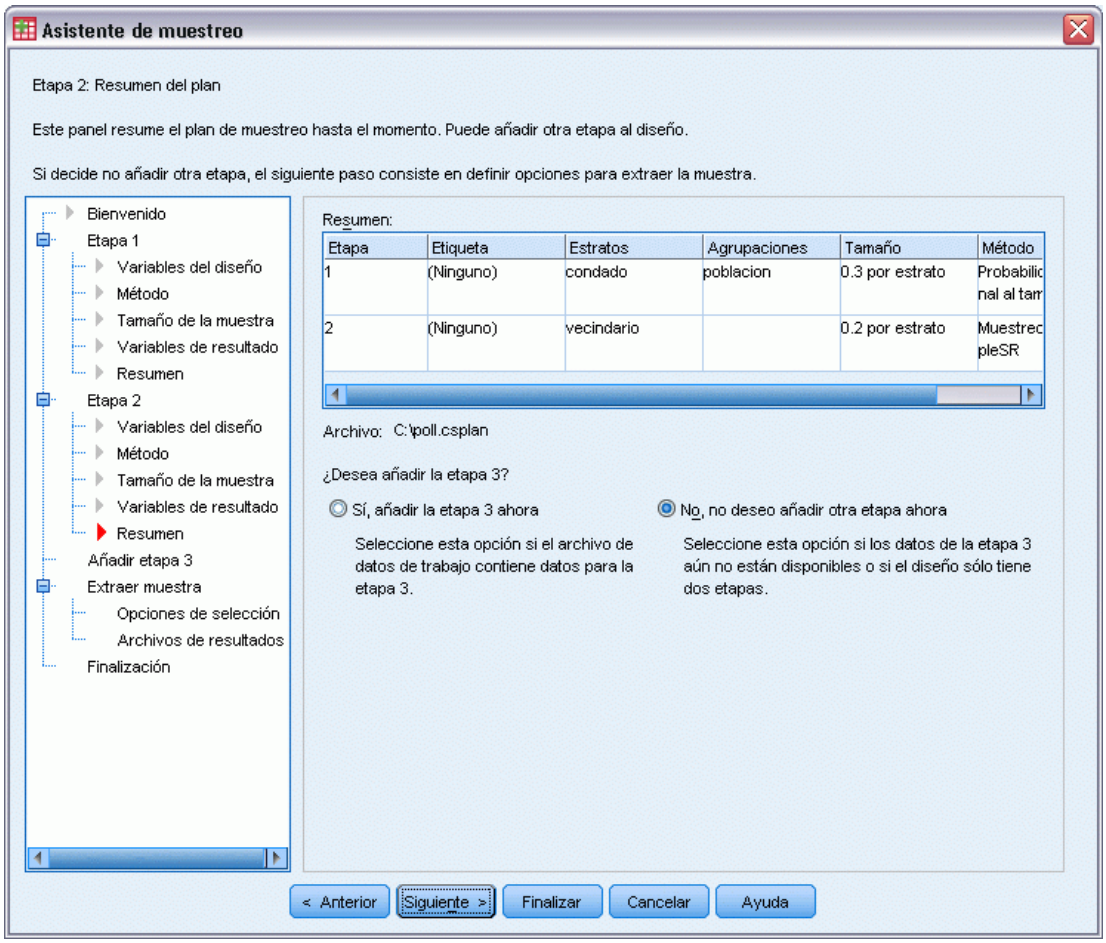
  Esta estructura de diseño indica que se extraen muestras independientes para cada vecindario de las poblaciones extraídas en la etapa 1. En esta etapa, se extraen los votantes como unidad muestral primaria utilizando el muestreo aleatorio simple sin reposición.

Figura 13-38
Asistente de muestreo: paso Tamaño muestral (etapa 2)



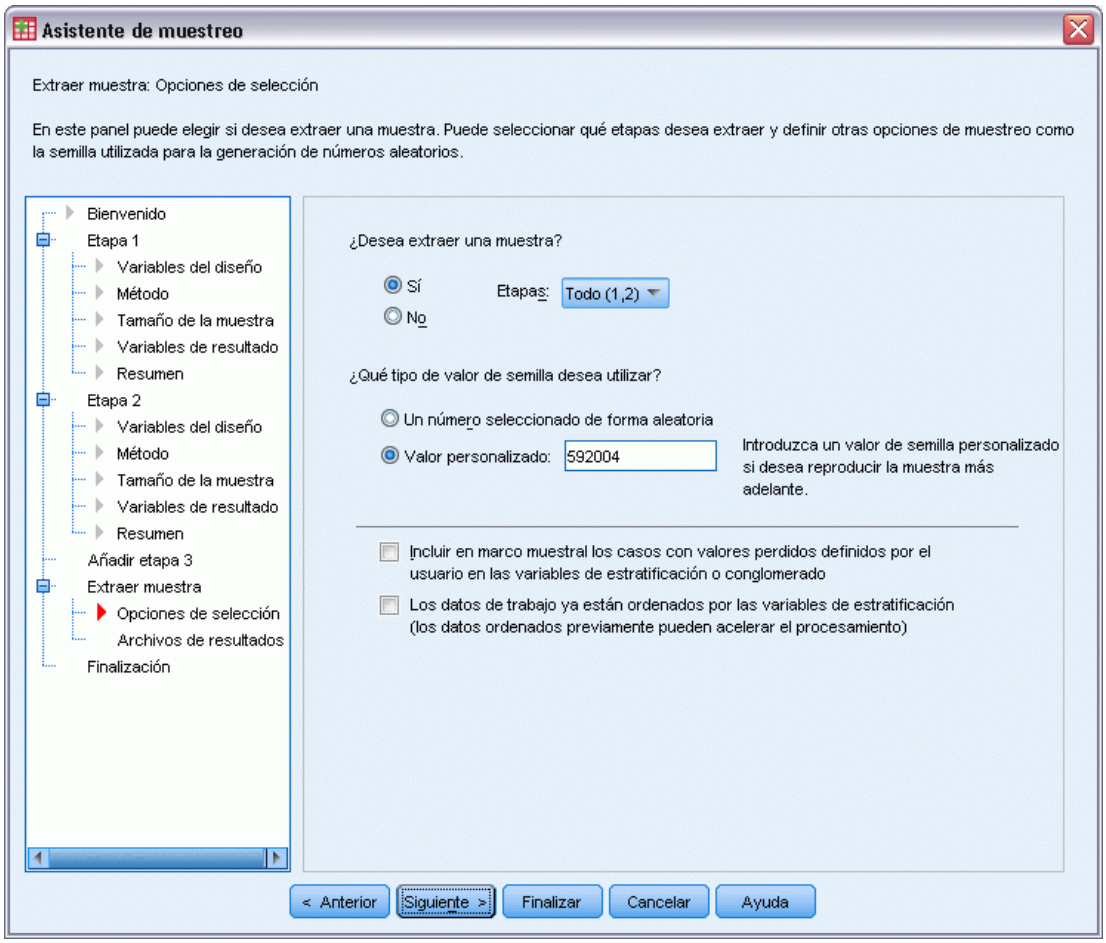
- Seleccione Proporciones en la lista desplegable Unidades.
- Escriba 0,2 como valor de la proporción de unidades que se van a extraer como muestra de cada estrato.
- Pulse en Siguiente y, a continuación, pulse en Siguiente en el paso Variables de resultado.

Figura 13-39
Asistente de muestreo: paso Resumen del plan (etapa 2)



▶ Revise el diseño muestral y, a continuación, pulse en Siguiente.

Figura 13-40
Asistente de muestreo: Extraer muestra: paso Opciones de selección

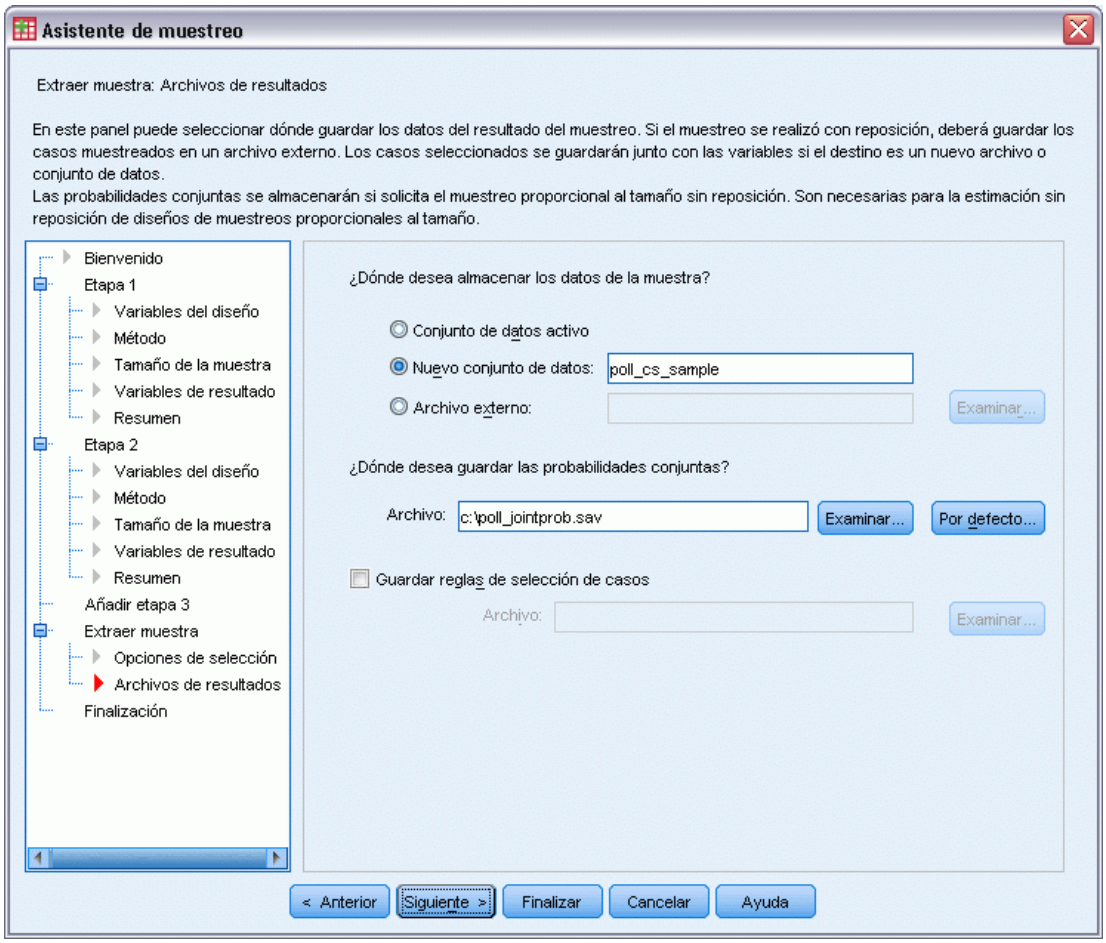


▶ Seleccione Valor personalizado como tipo de semilla aleatoria que se va a utilizar y escriba 592004 como valor.

Al utilizar un valor personalizado, es posible replicar los resultados de este ejemplo de manera exacta.

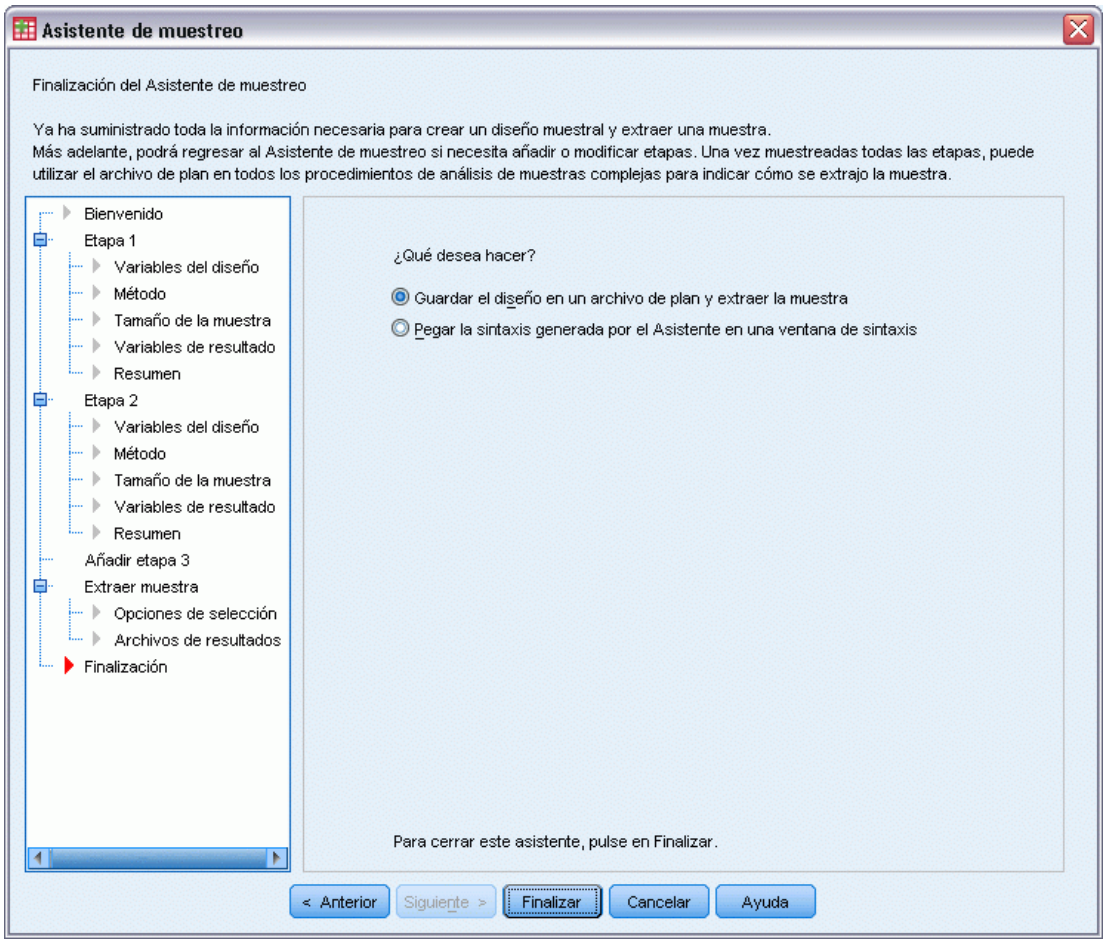
▶ Pulse en Siguiente.

Figura 13-41
Asistente de muestreo: Extraer muestra: paso Opciones de selección



- Seleccione guardar la muestra en un nuevo conjunto de datos y escriba poll_cs_sample como nombre del conjunto de datos.
- Desplácese hasta la ubicación en la que desea guardar las probabilidades conjuntas e introduzca poll_jointprob.sav como el nombre del archivo de probabilidades conjuntas.
- Pulse en Siguiente.

Figura 13-42
Asistente de muestreo: paso Finalizar



▶ Pulse en Finalizar.

Estas selecciones producen el archivo de plan de muestreo *poll.csplan* y extraen una muestra de acuerdo con dicho plan, guardan los resultados de la muestra en un nuevo conjunto de datos *poll_cs_sample*, y guardan el archivo de probabilidades conjuntas en el archivo de datos externo *poll_jointprob.sav*.

Resumen del plan

Figura 13-43
Resumen del plan

			Etapa 1	Etapa 2
Variables del diseño	Estratificación	1	Condado	Vecindario
	Conglomerado	1	Población	
Información de la muestra	Método de selección		Muestreo proporcional al tamaño sin reposición	Muestreo aleatorio simple sin reposición
	Medida del tamaño		A partir de los datos	
	Proporción de unidades muestreadas		,3	,2
	Número mínimo de unidades muestreadas		3	
	Número máximo de unidades muestreadas		5	
	Variables creadas o modificadas	Probabilidad de inclusión (selección) según etapa	ProbabilidadInclusión_1_	ProbabilidadInclusión_2_
		Ponderación de muestreo acumulada según etapa	PonderaciónMuestral Acumulada_1_	PonderaciónMuestral Acumulada_2_
Información sobre el análisis	Supuestos del estimador		Muestreo de probabilidad desigual sin reposición (mediante probabilidades conjuntas de inclusión)	Muestreo de probabilidad igual sin reposición
	Probabilidad de inclusión		A partir de la variable ProbabilidadInclusión_1_	A partir de la variable ProbabilidadInclusión_2_

Archivo del plan: C:\poll.csplan
Variable de ponderación: PonderaciónMuestral_Final_

La tabla de resumen muestra el plan de muestreo y resulta útil para asegurarse de que el plan corresponde a sus intenciones.

Resumen de muestreo

Figura 13-44
Resumen de las etapas

	Número de unidades muestreadas		Proporción de unidades muestreadas	
Condado	Solicitados	Reales	Solicitados	Reales
Este	4	4	30,0%	30,8%
Central	4	4	30,0%	30,8%
Oeste	3	3	30,0%	50,0%
Norte	5	5	30,0%	33,3%
Sur	3	3	30,0%	50,0%

Archivo del plan: C:\poll.csplan

Esta tabla de resumen muestra la primera etapa del muestreo y resulta útil para comprobar que el muestreo se ha realizado de acuerdo con el plan. Recuerde que solicitó una muestra del 30% de las poblaciones por condado; las proporciones reales muestreadas son cercanas al 30%, excepto en los condados del oeste y del sur. Esto se debe a que cada uno de estos condados sólo tiene seis poblaciones y se ha especificado que se debe seleccionar un mínimo de tres poblaciones por condado.

Figura 13-45
Resumen de las etapas

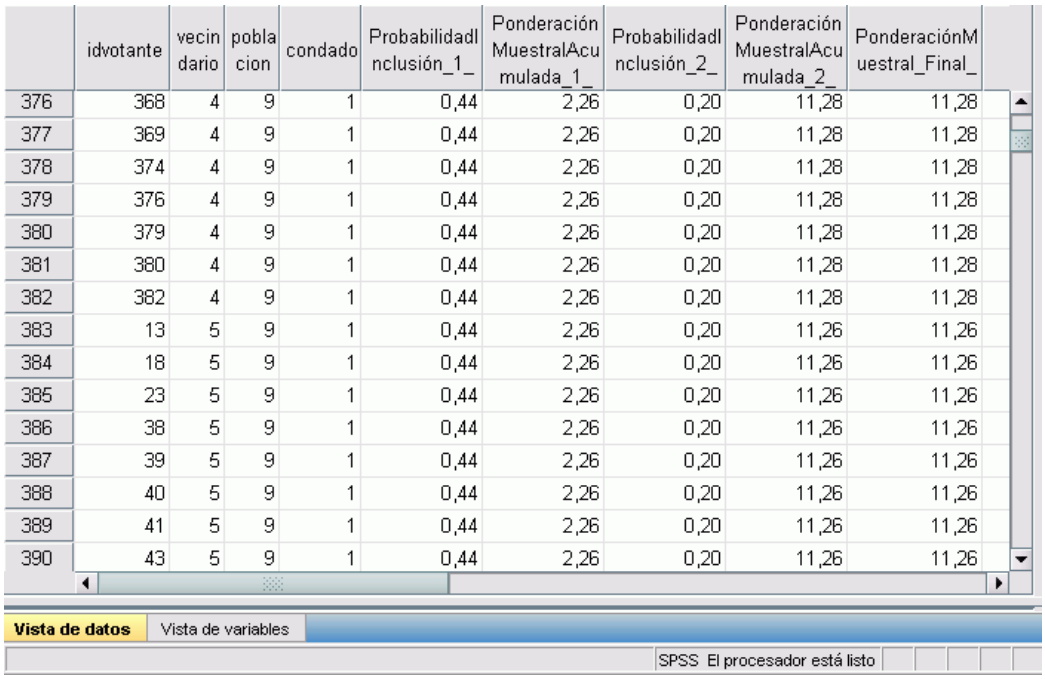
			Número de unidades muestreadas		Proporción de unidades muestreadas	
Condado	Población	Vecindario	Solicitados	Reales	Solicitados	Reales
Este	9	1	49	49	20,0%	19,9%
		2	143	143	20,0%	20,0%
		3	113	113	20,0%	20,0%
		4	77	77	20,0%	20,0%
		5	139	139	20,0%	20,0%
		6	120	120	20,0%	20,0%
	10	1	149	149	20,0%	20,1%
		2	117	117	20,0%	20,0%
		3	116	116	20,0%	20,0%
		4	69	69	20,0%	19,9%
	11	1	65	65	20,0%	19,9%
		2	72	72	20,0%	19,9%
		3	109	109	20,0%	20,0%
		4	140	140	20,0%	20,0%
		5	42	42	20,0%	19,8%
		6	142	142	20,0%	20,0%
	12	1	145	145	20,0%	20,1%
		2	69	69	20,0%	20,1%
		3	98	98	20,0%	20,1%
		4	134	134	20,0%	20,0%
		5	114	114	20,0%	20,0%
		6	137	137	20,0%	19,9%
Central	2	1	119	119	20,0%	20,1%
		2	153	153	20,0%	19,9%
		3	101	101	20,0%	20,0%
		4	52	52	20,0%	19,8%
		5	144	144	20,0%	20,0%

Plan File: c:\poll.csplan

Esta tabla de resumen (de la cual se muestra aquí la parte superior) muestra la segunda etapa del muestreo. También resulta útil para comprobar que el muestreo se ha realizado de acuerdo con el plan. Como se solicitó, se muestreó aproximadamente el 20% de los votantes de cada vecindario de cada una de las poblaciones muestreadas en la primera etapa.

Resultados de la muestra

Figura 13-46
Editor de datos con los resultados de la muestra



	idvotante	vecindario	poblacion	condado	ProbabilidadInclusión_1_	PonderaciónMuestralAcumulada_1_	ProbabilidadInclusión_2_	PonderaciónMuestralAcumulada_2_	PonderaciónMuestral_Final_
376	368	4	9	1	0,44	2,26	0,20	11,28	11,28
377	369	4	9	1	0,44	2,26	0,20	11,28	11,28
378	374	4	9	1	0,44	2,26	0,20	11,28	11,28
379	376	4	9	1	0,44	2,26	0,20	11,28	11,28
380	379	4	9	1	0,44	2,26	0,20	11,28	11,28
381	380	4	9	1	0,44	2,26	0,20	11,28	11,28
382	382	4	9	1	0,44	2,26	0,20	11,28	11,28
383	13	5	9	1	0,44	2,26	0,20	11,26	11,26
384	18	5	9	1	0,44	2,26	0,20	11,26	11,26
385	23	5	9	1	0,44	2,26	0,20	11,26	11,26
386	38	5	9	1	0,44	2,26	0,20	11,26	11,26
387	39	5	9	1	0,44	2,26	0,20	11,26	11,26
388	40	5	9	1	0,44	2,26	0,20	11,26	11,26
389	41	5	9	1	0,44	2,26	0,20	11,26	11,26
390	43	5	9	1	0,44	2,26	0,20	11,26	11,26

Puede ver los resultados del muestreo en el conjunto de datos recién creado. Se han guardado cinco nuevas variables en el archivo de trabajo, que representan las probabilidades de inclusión y las ponderaciones muestrales acumuladas para cada etapa, además de las ponderaciones muestrales finales. Los votantes que no se han seleccionado para la muestra se excluyen de este conjunto de datos.

Las ponderaciones muestrales finales son idénticas para los votantes de algunos vecindarios ya que están seleccionados de acuerdo con un método de muestreo aleatorio simple de los vecindarios. Sin embargo, son distintos entre vecindarios de la misma población ya que las proporciones muestreadas no son exactamente el 20% en todos los vecindarios.

Figura 13-47
Editor de datos con los resultados de la muestra

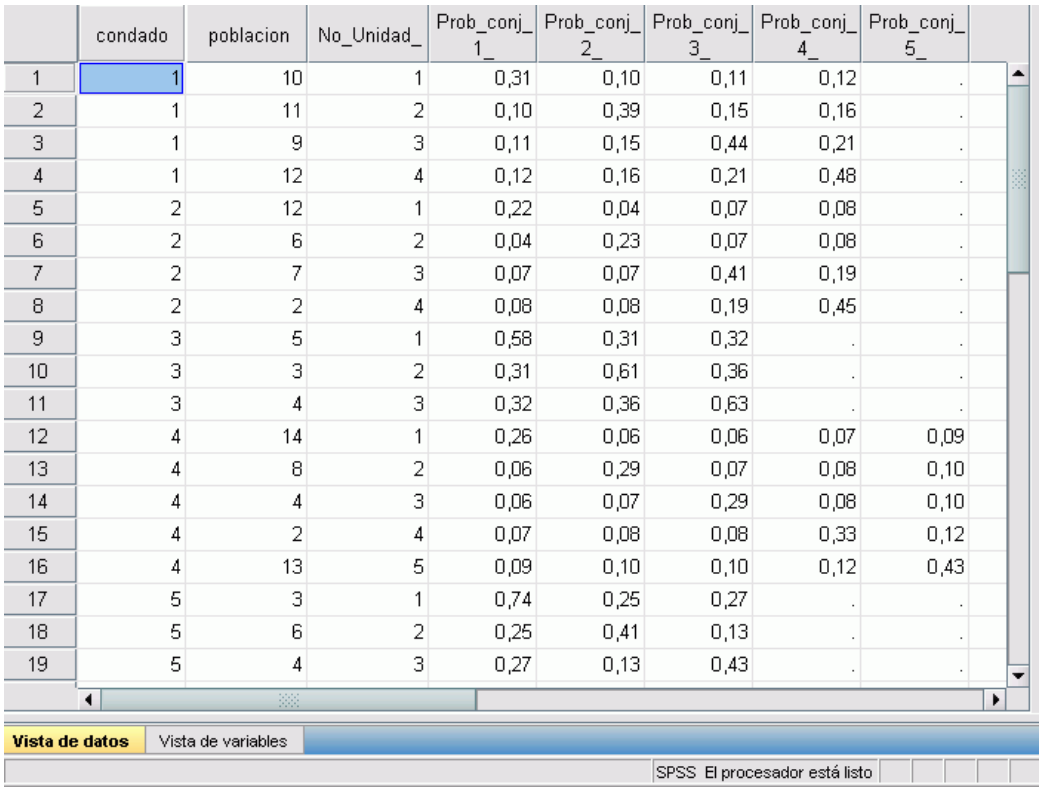
	idvotante	vecindario	poblacion	condado	ProbabilidadInclusión_1_	PonderaciónMuestralAcumulada_1_	ProbabilidadInclusión_2_	PonderaciónMuestralAcumulada_2_	PonderaciónMuestral_Final_
635	577	6	9	1	0,44	2,26	0,20	11,30	11,30
636	578	6	9	1	0,44	2,26	0,20	11,30	11,30
637	582	6	9	1	0,44	2,26	0,20	11,30	11,30
638	590	6	9	1	0,44	2,26	0,20	11,30	11,30
639	594	6	9	1	0,44	2,26	0,20	11,30	11,30
640	597	6	9	1	0,44	2,26	0,20	11,30	11,30
641	600	6	9	1	0,44	2,26	0,20	11,30	11,30
642	4	1	10	1	0,31	3,21	0,20	16,00	16,00
643	5	1	10	1	0,31	3,21	0,20	16,00	16,00
644	9	1	10	1	0,31	3,21	0,20	16,00	16,00
645	10	1	10	1	0,31	3,21	0,20	16,00	16,00
646	12	1	10	1	0,31	3,21	0,20	16,00	16,00
647	16	1	10	1	0,31	3,21	0,20	16,00	16,00
648	17	1	10	1	0,31	3,21	0,20	16,00	16,00
649	19	1	10	1	0,31	3,21	0,20	16,00	16,00

Vista de datos | Vista de variables

SPSS El procesador está listo

A diferencia de los votantes de la segunda etapa, las ponderaciones muestrales de la primera etapa no son idénticas para las poblaciones del mismo condado porque se han seleccionado con probabilidad proporcional al tamaño.

Figura 13-48
Archivo de probabilidades conjuntas



	condado	poblacion	No_Unidad_	Prob_conj_1_	Prob_conj_2_	Prob_conj_3_	Prob_conj_4_	Prob_conj_5_
1	1	10	1	0,31	0,10	0,11	0,12	.
2	1	11	2	0,10	0,39	0,15	0,16	.
3	1	9	3	0,11	0,15	0,44	0,21	.
4	1	12	4	0,12	0,16	0,21	0,48	.
5	2	12	1	0,22	0,04	0,07	0,08	.
6	2	6	2	0,04	0,23	0,07	0,08	.
7	2	7	3	0,07	0,07	0,41	0,19	.
8	2	2	4	0,08	0,08	0,19	0,45	.
9	3	5	1	0,58	0,31	0,32	.	.
10	3	3	2	0,31	0,61	0,36	.	.
11	3	4	3	0,32	0,36	0,63	.	.
12	4	14	1	0,26	0,06	0,06	0,07	0,09
13	4	8	2	0,06	0,29	0,07	0,08	0,10
14	4	4	3	0,06	0,07	0,29	0,08	0,10
15	4	2	4	0,07	0,08	0,08	0,33	0,12
16	4	13	5	0,09	0,10	0,10	0,12	0,43
17	5	3	1	0,74	0,25	0,27	.	.
18	5	6	2	0,25	0,41	0,13	.	.
19	5	4	3	0,27	0,13	0,43	.	.

El archivo *poll_jointprob.sav* contiene las probabilidades conjuntas en la primera etapa para las poblaciones seleccionadas dentro de condados. *Condado* es una variable de estratificación de primera etapa y *Población* es una variable de aglomeración. Las combinaciones de estas variables identifican de forma única todas las PSU de primera etapa. *No_Unidad_* etiqueta las PSU dentro de cada estrato y se utiliza para que coincida con *Prob_conj_1_*, *Prob_conj_2_*, *Prob_conj_3_*, *Prob_conj_4_* y *Prob_conj_5_*. Los dos primeros estratos tienen 4 PSU, por lo que las matrices de probabilidad de inclusión conjunta son 4×4 para estos estratos y la columna *Prob_conj_5_* está vacía a la izquierda para estas filas. Del mismo modo, los estratos 3 y 5 tienen matrices de probabilidad de inclusión conjunta 3×3 y el estrato 4 tiene una matriz de probabilidad de inclusión conjunta 5×5.

Si se examinan los valores de las matrices de probabilidad de inclusión conjunta se puede determinar la necesidad de un archivo de probabilidades conjuntas. Cuando el método de muestreo no es un método PPS SR, la selección de una PSU es independiente de la selección de otra PSU y la probabilidad de inclusión conjunta es simplemente el producto de sus probabilidades de inclusión. Por el contrario, la probabilidad de inclusión conjunta de las poblaciones 9 y 10 del condado 1 es aproximadamente 0,11 (consulte el primer caso de *Prob_conj_3_* o el tercer caso de *Prob_conj_1_*) o menor que el producto de sus probabilidades de inclusión individuales (el producto del primer caso de *Prob_conj_1_* y el tercer caso de *Prob_conj_3_* es 0,31×0,44=0,1364).

Los encuestadores ahora llevarán a cabo entrevistas para la muestra seleccionada. Una vez que los resultados están disponibles, puede procesar la muestra con procedimientos de análisis de Muestras complejas mediante el plan de muestreo *poll.csplan* para proporcionar

las especificaciones de muestreo y *poll_jointprob.sav* para proporcionar las probabilidades de inclusión conjunta necesarias.

Procedimientos relacionados

El procedimiento Asistente de muestreo de la opción Muestras complejas es una herramienta útil para crear un archivo de plan de muestreo y extraer una muestra.

- Para preparar una muestra para su análisis cuando no puede acceder al archivo de plan de muestreo, utilice el Asistente de preparación del análisis.

Capítulo
14

Asistente de preparación del análisis de la opción Muestras complejas

El Asistente de preparación del análisis le guía a través de los pasos para crear o modificar un plan de análisis y utilizarlo con los distintos procedimientos de análisis de Muestras complejas. Resulta especialmente útil cuando no se puede acceder al archivo del plan de muestreo que se utilizó para extraer la muestra.

Uso del Asistente de preparación del análisis de la opción Muestras complejas para preparar los datos de uso público de la NHIS

La National Health Interview Survey (NHIS, encuesta del Centro Nacional de Estadísticas de Salud de EE.UU.) es una encuesta muy detallada realizada entre la población civil de Estados Unidos. Las encuestas se realizaron en persona a una muestra representativa de las unidades familiares del país. Se recogió tanto la información demográfica como las observaciones acerca del estado y los hábitos de salud de los integrantes de cada unidad familiar.

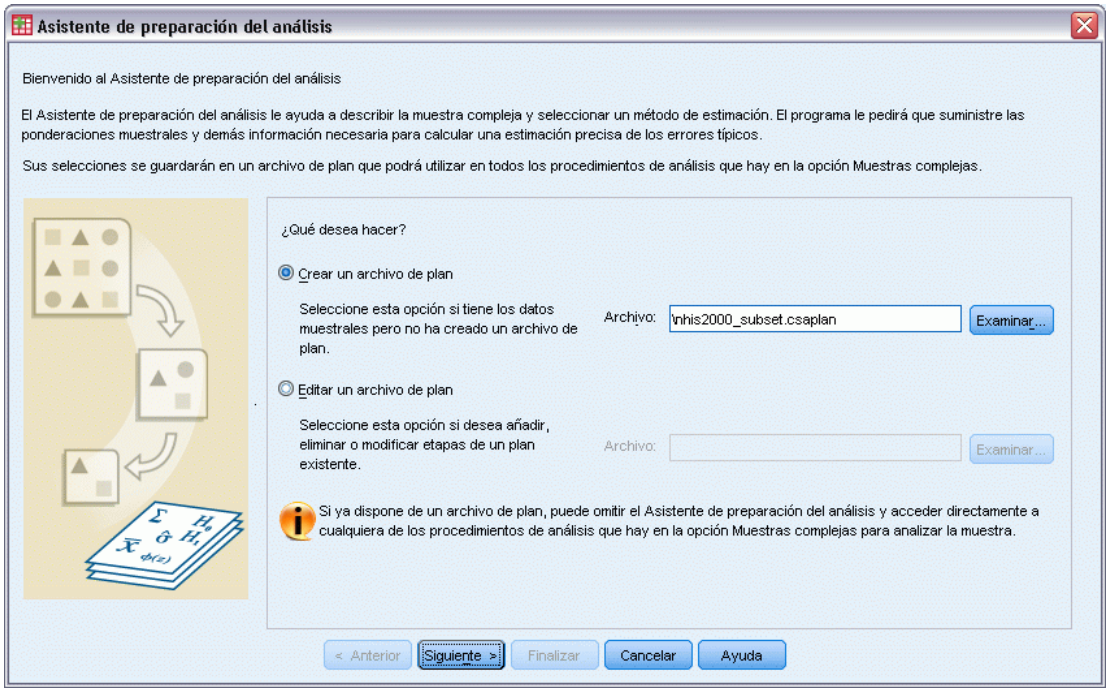
Un subconjunto de la encuesta de 2000 se incluye en *nhis2000_subset.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*. Utilice el Asistente de preparación del análisis de la opción Muestras complejas para crear un plan de análisis para este archivo de datos de manera que se pueda procesar mediante los procedimientos de análisis de Muestras complejas.

Uso del asistente

- Para preparar una muestra mediante el Asistente de preparación del análisis de la opción Muestras complejas, seleccione en los menús:

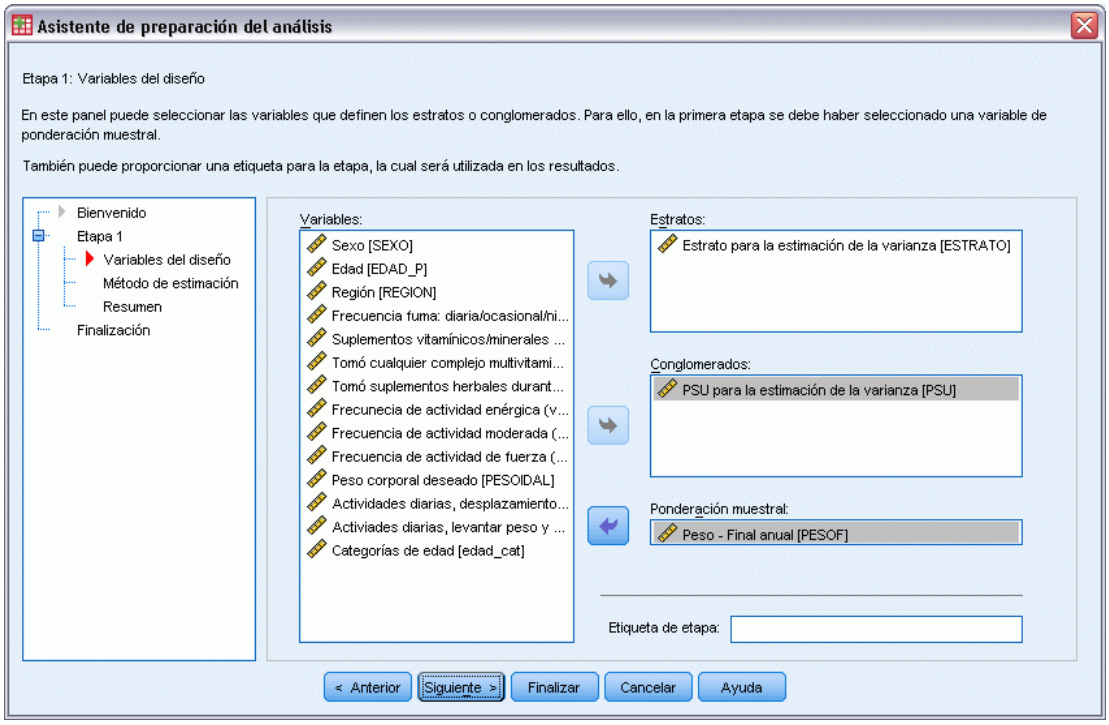
  Analizar > Muestras complejas > Preparar para el análisis...

Figura 14-1
Asistente de preparación del análisis: paso Bienvenida



- Desplácese hasta la ubicación en la que desea guardar el archivo de plan e introduzca nhis2000_subset.csaplan como el nombre del archivo de plan de análisis.
- Pulse en Siguiente.

Figura 14-2
Asistente de preparación del análisis: paso Variables del diseño (etapa 1)



Los datos se obtuvieron utilizando una muestra compleja polietápica. No obstante, para los usuarios finales, las variables de diseño originales de la NHIS se transformaron en un conjunto simplificado de variables de diseño y de ponderación cuyos resultados se aproximan a los de las estructuras de diseño originales.

- Seleccione *Estrato para la estimación de la varianza* como variable de estrato.
- Seleccione *PSU para la estimación de la varianza* como variable de conglomerado.
- Seleccione *Peso - Final anual* como variable de ponderación muestral.
- Pulse en Finalizar.

Resumen

Figura 14-3
Resumen

			Etapa 1
Variables del diseño	Estratificación	1	Estrato para la estimación de la varianza
	Conglomerado	1	PSU para la estimación de la varianza
Información sobre el análisis	Supuestos del estimador		Muestreo con reposición

Archivo de plan: C:\nhis2000_subset.csaplan
Variable de ponderación: Peso - Final anual
Estimador SRS: Muestreo sin reposición

La tabla de resumen permite revisar el plan de análisis. El plan se compone de una etapa cuyo diseño se compone de una variable de estratificación y una variable de conglomerado. Se utiliza estimación con reposición (CR) y el plan se almacena en el archivo *c:\nhis2000_subset.csaplan*. Ahora puede utilizar este archivo de plan para procesar *nhis2000_subset.sav* con los procedimientos de análisis de Muestras complejas.

Preparación del análisis cuando las ponderaciones muestrales no se encuentran en el archivo de datos

Un encargado de préstamos tiene un conjunto de registros de clientes que se han realizado siguiendo un diseño complejo. Sin embargo, las ponderaciones muestrales no se incluyen en el archivo. Esta información se recoge en *bankloan_cs_noweights.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*. Basándose en sus conocimientos sobre el diseño muestral, el encargado desea utilizar el Asistente de preparación del análisis de la opción Muestras complejas para crear un plan de análisis para este archivo de datos con el fin de procesarlo mediante los procedimientos de análisis de Muestras complejas.

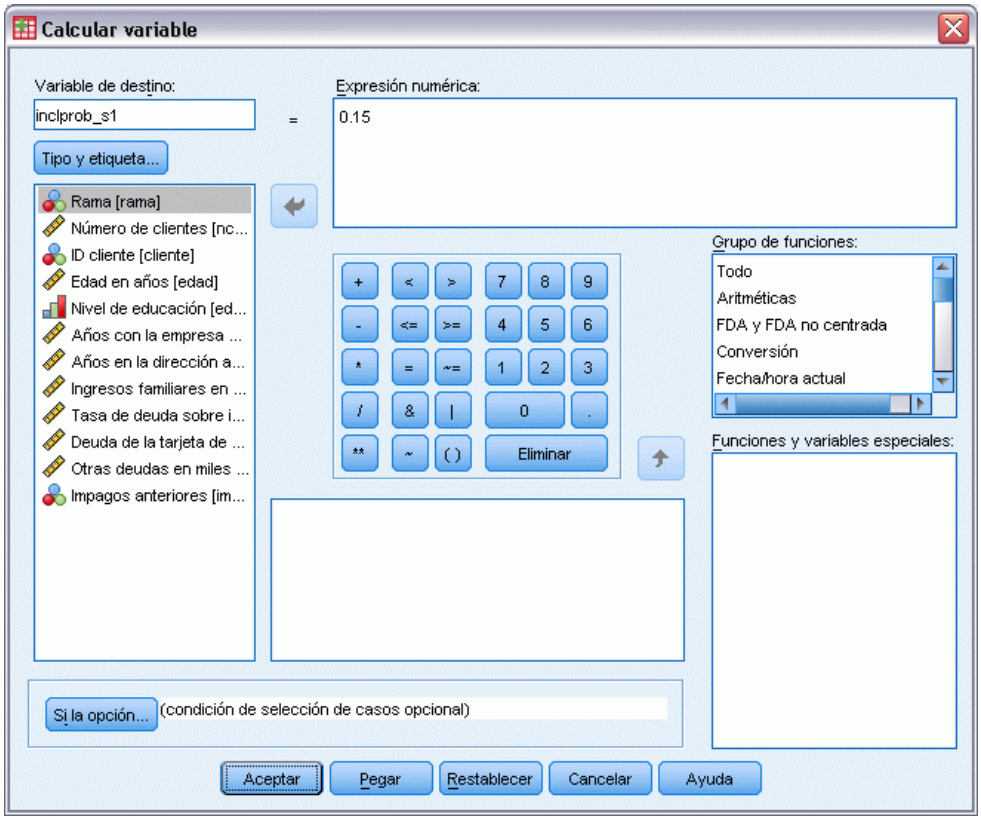
El encargado de préstamos sabe que los registros se seleccionaron en dos etapas, con 15 sucursales bancarias seleccionadas de un total de 100, con probabilidad igual y sin reposición en la primera etapa. Se seleccionaron cien clientes de cada una de esas sucursales con probabilidad igual y sin reposición en la segunda etapa, incluyéndose en el archivo de datos la información del número de clientes de cada sucursal. El primer paso para crear un plan de análisis consiste en calcular las probabilidades de inclusión según etapa y las ponderaciones muestrales finales.

Cálculo de las probabilidades de inclusión y las ponderaciones muestrales

- Para calcular las probabilidades de inclusión de la primera etapa, seleccione en el menú las siguientes opciones:

  Transformar > Calcular variable...

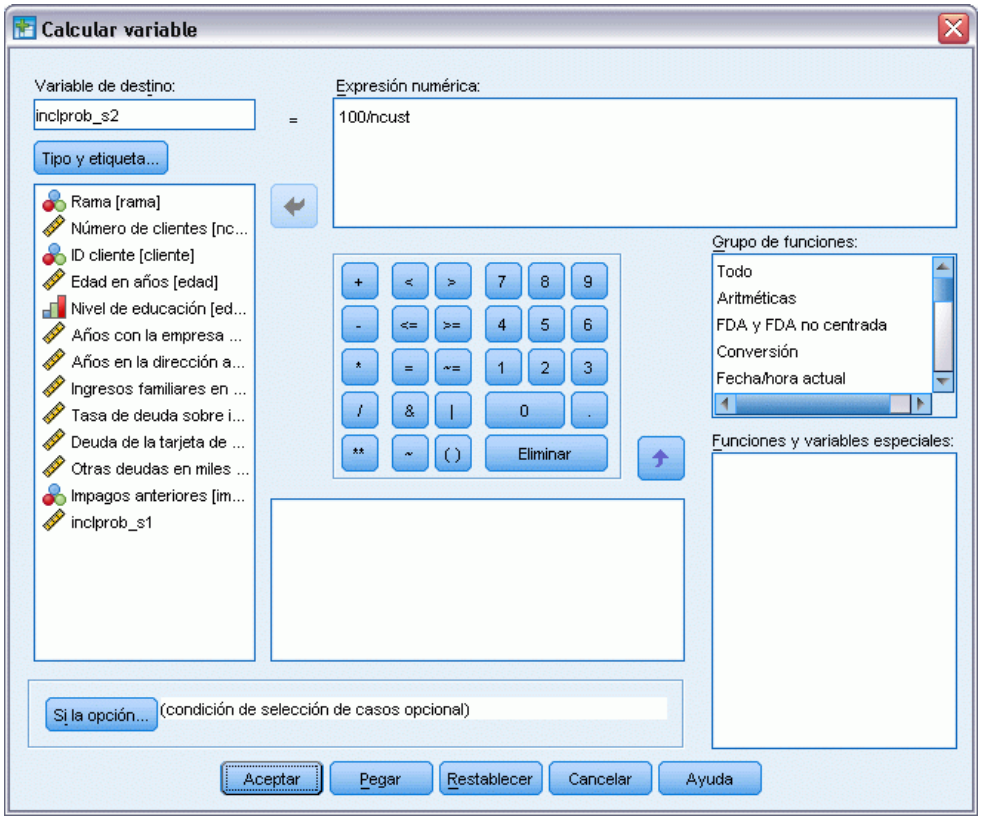
Figura 14-4
Cuadro de diálogo Calcular variable



En la primera etapa se han seleccionado quince de las cien sucursales sin sustitución. Por consiguiente, la probabilidad de que un banco determinado se seleccionara es de 15/100 = 0,15.

- Escriba inclprob_s1 como variable de destino.
- Escriba 0,15 como expresión numérica.
- Pulse en Aceptar.

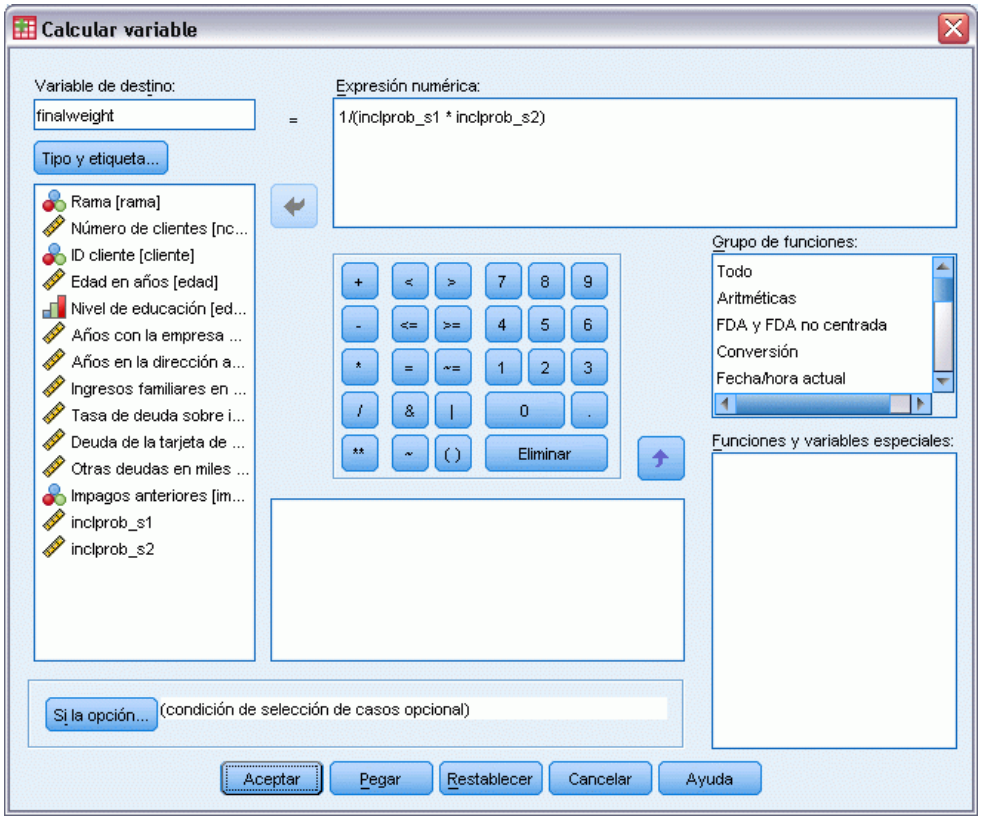
Figura 14-5
Cuadro de diálogo Calcular variable



En la segunda etapa se han seleccionado cien clientes de cada sucursal. Por consiguiente, la probabilidad de inclusión de la segunda etapa para un cliente determinado de una sucursal determinada es de 100/número de clientes de esa sucursal.

- Vuelva a abrir el cuadro de diálogo Calcular variable.
- Escriba inclprob_s2 como variable de destino.
- Escriba 100/ncust como expresión numérica.
- Pulse en Aceptar.

Figura 14-6
Cuadro de diálogo Calcular variable



Ahora que ha obtenido las probabilidades de inclusión de cada etapa, es muy sencillo calcular las ponderaciones muestrales finales.

- Vuelva a abrir el cuadro de diálogo Calcular variable.
- Escriba finalweight como variable de destino.
- Escriba 1/(inclprob_s1 * inclprob_s2) como expresión numérica.
- Pulse en Aceptar.

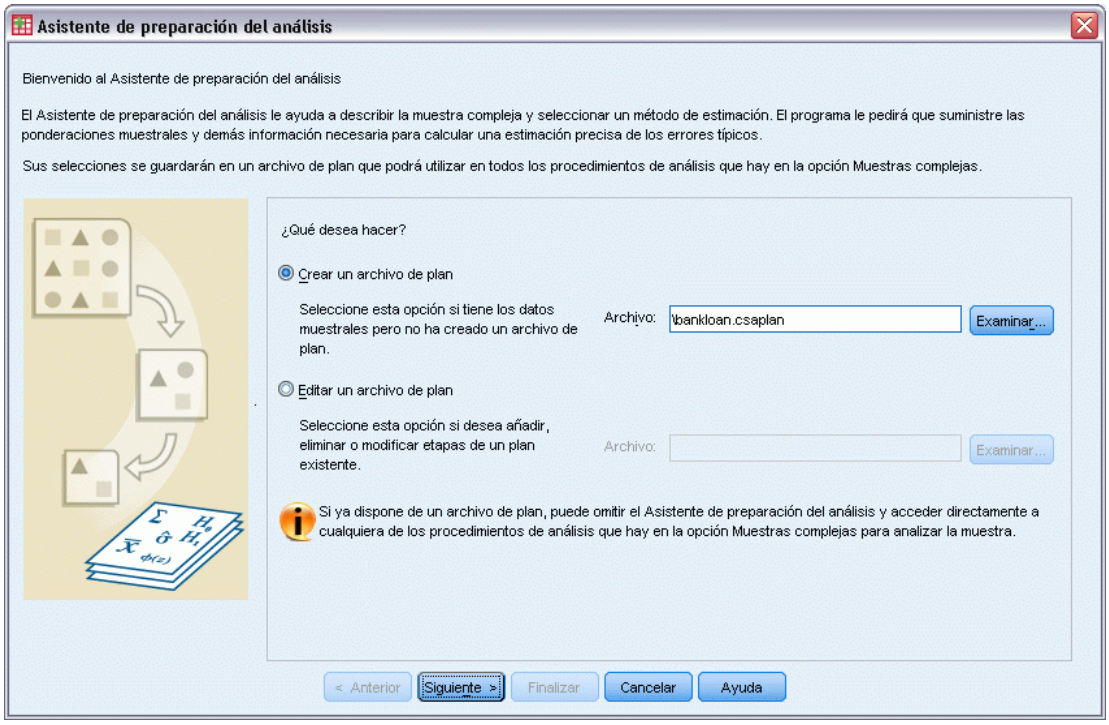
Ya puede crear el plan de análisis.

Uso del asistente

- Para preparar una muestra mediante el Asistente de preparación del análisis de la opción Muestras complejas, seleccione en los menús:

  Analizar > Muestras complejas > Preparar para el análisis...

Figura 14-7
Asistente de preparación del análisis: paso Bienvenida

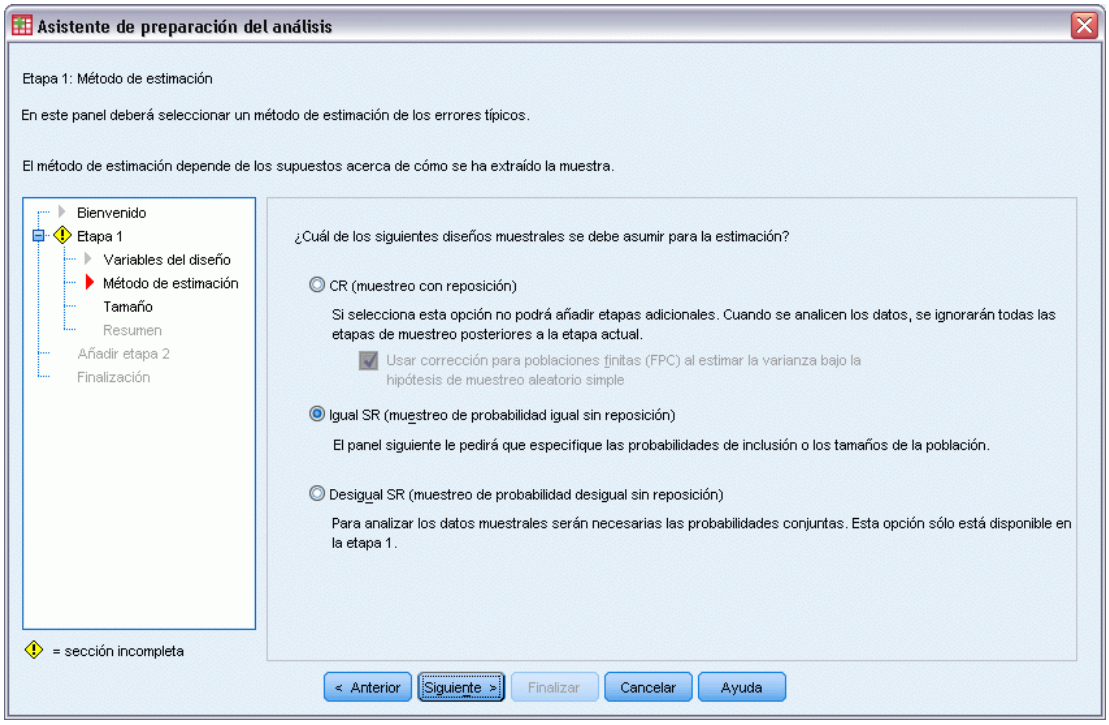


- Desplácese hasta la ubicación en la que desea guardar el archivo de plan e introduzca bankloan.csaplan como el nombre del archivo de plan de análisis.
- Pulse en Siguiente.

Figura 14-8
Asistente de preparación del análisis: paso Variables del diseño (etapa 1)

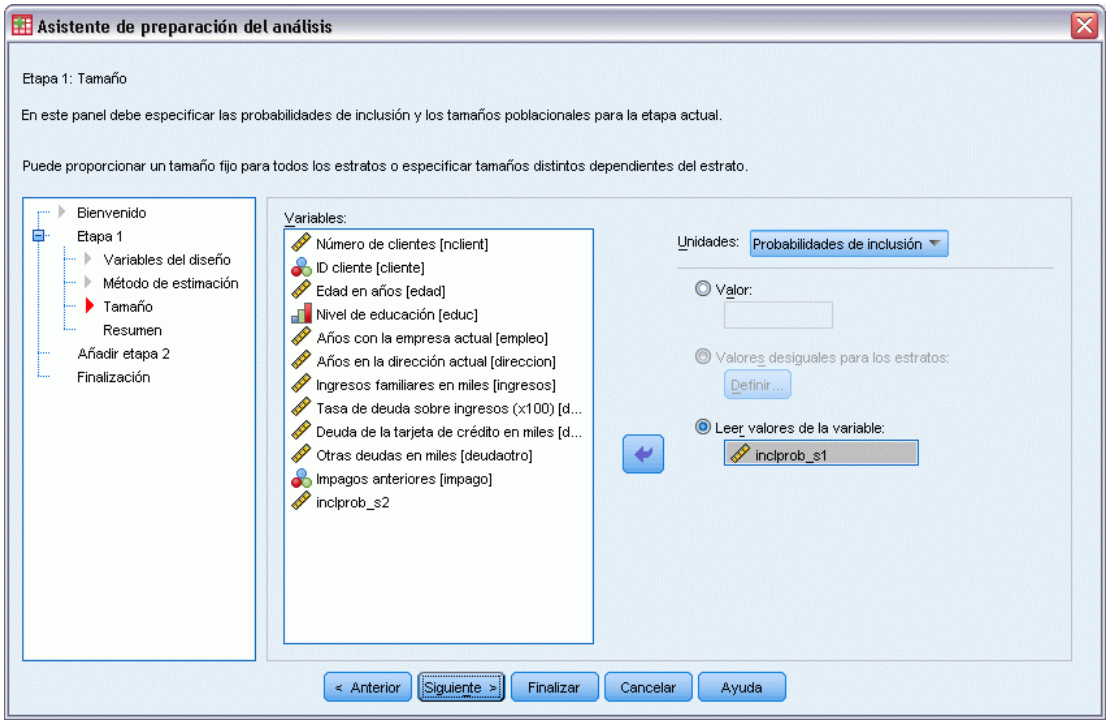
- Seleccione *Rama* como variable de aglomeración.
- Seleccione *finalweight* como variable de ponderación muestral.
- Pulse en Siguiente.

Figura 14-9
Asistente de preparación del análisis: paso Método de estimación (etapa 1)



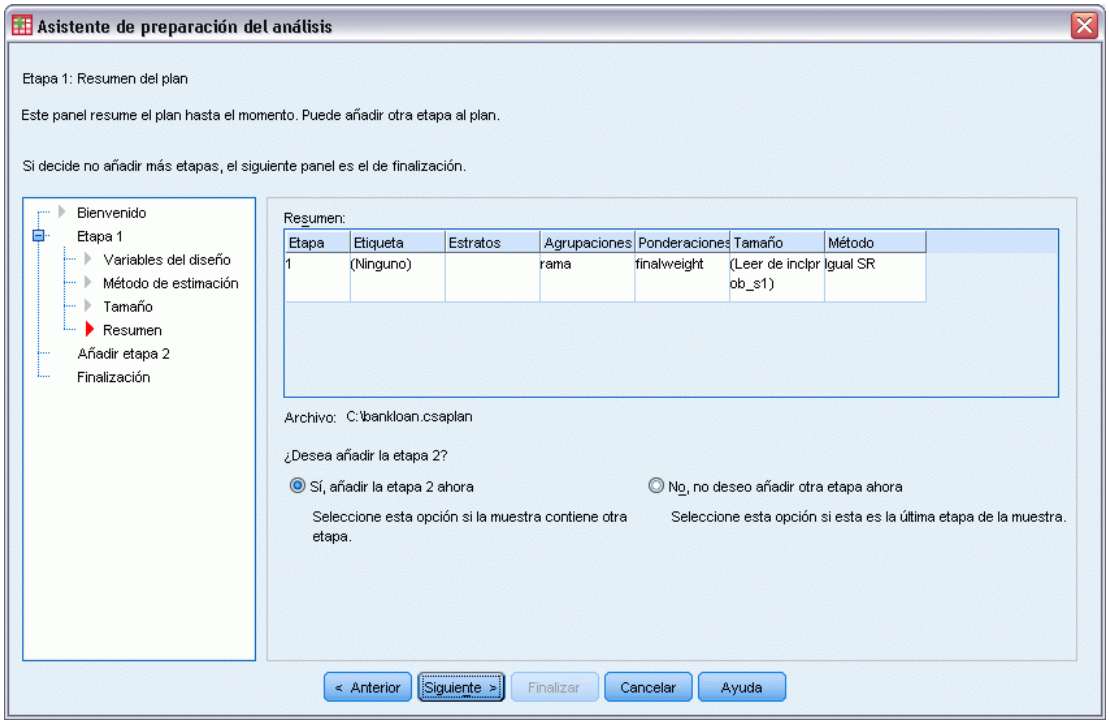
- Seleccione Igual SR como el método de estimación de la primera etapa.
- Pulse en Siguiente.

Figura 14-10
Asistente de preparación del análisis: paso Tamaño (etapa 1)



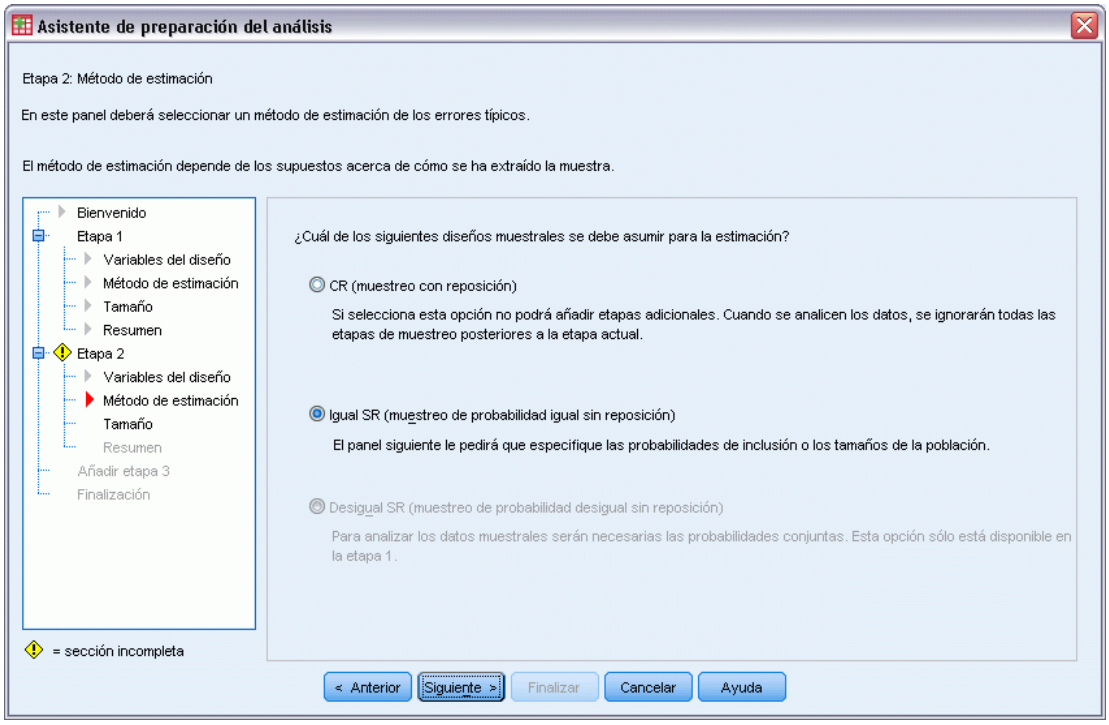
- Seleccione Leer valores de la variable y elija *inclprob_s1* como la variable que contiene las probabilidades de inclusión de la primera etapa.
- Pulse en Siguiente.

Figura 14-11
Asistente de preparación del análisis: paso Resumen del plan (etapa 1)



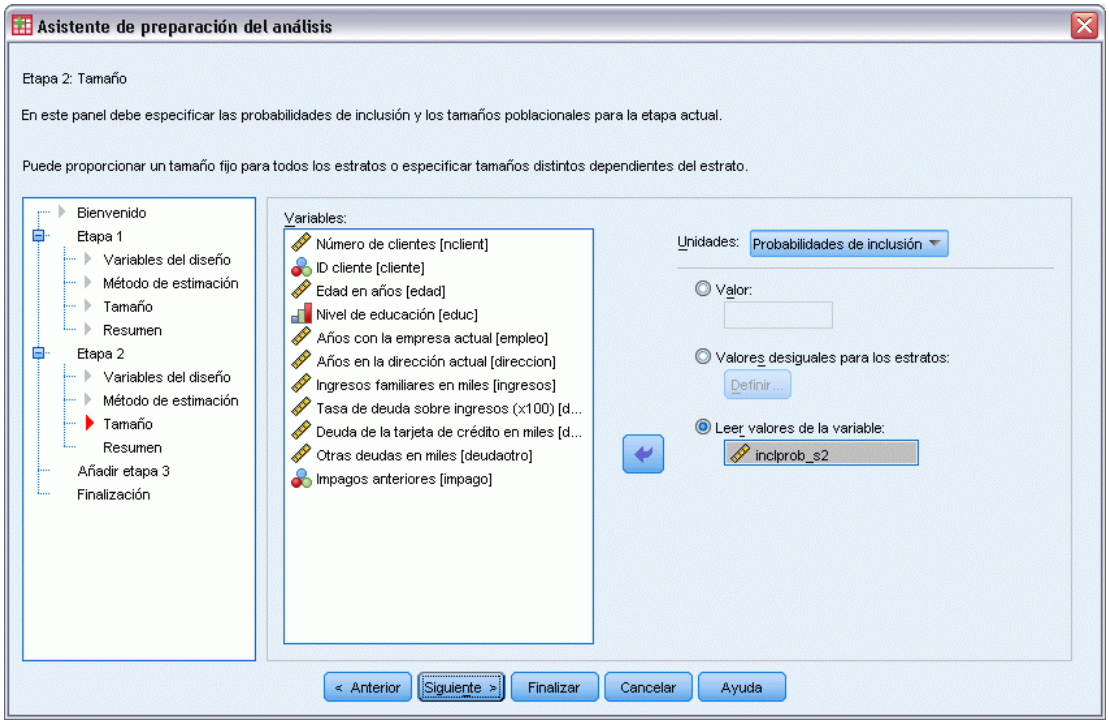
- Seleccione Sí, añadir la etapa 2 ahora.
- Pulse en Siguiente y, a continuación, pulse en Siguiente en el paso Variables del diseño.

Figura 14-12
Asistente de preparación del análisis: paso Método de estimación (etapa 2)



- Seleccione Igual SR como el método de estimación de la segunda etapa.
- Pulse en Siguiente.

Figura 14-13
Asistente de preparación del análisis: paso Tamaño (etapa 2)



- Seleccione Leer valores de la variable y seleccione *inclprob_s2* como la variable que contiene las probabilidades de inclusión de la segunda etapa.
- Pulse en Finalizar.

Resumen

Figura 14-14
Tabla de resumen

			Etapa 1	Etapa 2
Variables del diseño	Estratificación	1	Rama	
Información sobre el análisis	Supuestos del estimador		Muestreo de probabilidad igual sin reposición	Muestreo de probabilidad igual sin reposición
	Probabilidad de inclusión		A partir de la variable inclprob_s2	A partir de la variable inclprob_s2

Archivo de plan: C:\bankloan.csaplan
Variable de ponderación: finalweight
Estimador SRS: Muestreo sin reposición

La tabla de resumen permite revisar el plan de análisis. El plan está formado por dos etapas con un diseño de una variable de agrupación. Se utiliza la estimación de probabilidad igual sin reposición (CR) y el plan se almacena en el archivo *c:\bankloan.csaplan*. Ya puede utilizar este archivo del plan para procesar *bankloan_noweights.sav* (con las probabilidades de inclusión y las ponderaciones muestrales que ha calculado) con los procedimientos de análisis de Muestras complejas.

Procedimientos relacionados

El procedimiento del Asistente de preparación del análisis de la opción Muestras complejas es una herramienta útil para preparar una muestra para su análisis cuando no puede acceder al archivo del plan de muestreo.

- Para crear un archivo del plan de muestreo y extraer una muestra, utilice el Asistente de muestreo.

Capítulo
15

Frecuencias de Muestras complejas

El procedimiento Frecuencias de Muestras complejas genera tablas de frecuencias para las variables seleccionadas y muestra estadísticos univariantes. Si lo desea, puede solicitar estadísticos por subgrupos, definidos por una o más variables categóricas.

Uso de Frecuencias de muestras complejas para analizar el consumo de suplementos nutritivos

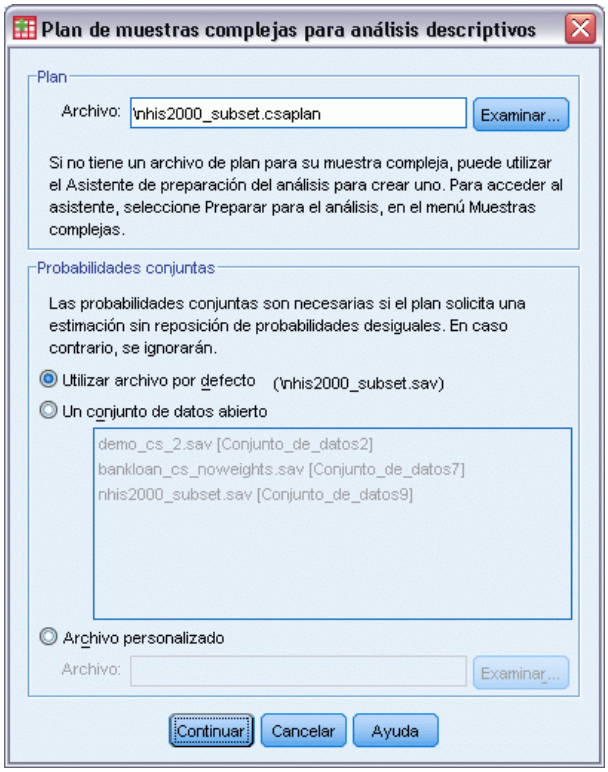
Un investigador desea estudiar el uso de suplementos nutritivos de los ciudadanos de EE.UU., utilizando los resultados de la National Health Interview Survey (NHIS, Centro Nacional de Estadísticas de Salud) y un plan de análisis anteriormente creado. Si desea obtener más información, consulte el tema Uso del Asistente de preparación del análisis de la opción Muestras complejas para preparar los datos de uso público de la NHIS en el capítulo 14 el p. 146.

Un subconjunto de la encuesta de 2000 se incluye en *nhis2000_subset.sav*. El plan del análisis se guarda en *nhis2000_subset.csaplan*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*. Uso de Frecuencias de muestras complejas para generar estadísticos acerca del consumo de suplementos nutritivos.

Ejecución del análisis

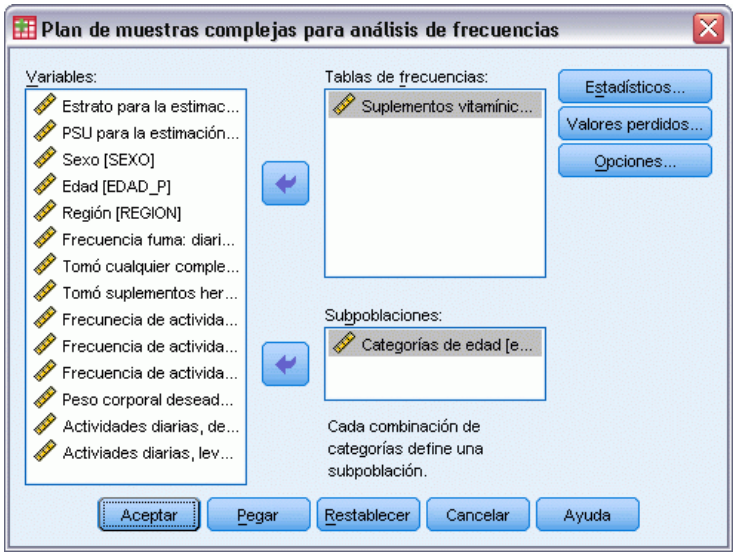
- Para ejecutar un análisis de Frecuencias de muestras complejas, seleccione en los menús:
  Analizar > Complex Samples > Frecuencias...

Figura 15-1
Cuadro de diálogo Plan de muestras complejas



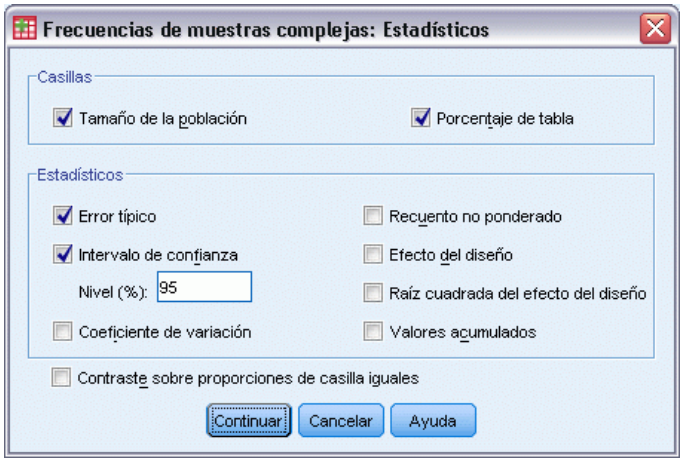
- Acceda al archivo *nhis2000_subset.csaplan* y selecciónelo. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*.
- Pulse en Continuar.

Figura 15-2
Cuadro de diálogo Frecuencias



- Seleccione *Suplementos vitamínicos/minerales últimos 12 m* como variable de frecuencia.
- Seleccione *Categorías de edad* como una variable de subpoblación.
- Pulse en Estadísticos.

Figura 15-3
Cuadro de diálogo Frecuencias: Estadísticos



- Seleccione Porcentaje de tabla en el grupo Casillas.
- Seleccione Intervalo de confianza en el grupo Estadísticos.
- Pulse en Continuar.
- Pulse en Aceptar en el cuadro de diálogo Frecuencias.

Tabla de frecuencia

Figura 15-4
Tabla de frecuencia para variable/situación

		Estimación	Error típico	Intervalo de confianza al 95%	
				Inferior	Superior
Tamaño de la población	Sí	102767095	1185126,709	100435967	105098223
	No	90794234	1094401,949	88641560	92946908
	Total	193561329	1789098,713	190042196	197080462
% del total	Sí	53,1%	,4%	52,4%	53,8%
	No	46,9%	,4%	46,2%	47,6%
	Total	100,0%	,0%	100,0%	100,0%

Se calcula cada estadístico seleccionado para cada medida de casilla seleccionada. La primera columna contiene estimaciones del número y el porcentaje de la población que toma o no toma suplementos vitamínicos/minerales. Los intervalos de confianza no se solapan; por tanto, se puede concluir que, en general, hay más americanos que toman suplementos vitamínicos/minerales que los que no los toman.

Frecuencia por subpoblación

Figura 15-5
Tabla de frecuencia por subpoblación

Categorías de edad			Estimación	Error típico	Intervalo de confianza al 95%	
					Inferior	Superior
18-24	Tamaño de la población	Sí	10018312	350602,35	9328682	10707942
		No	15472368	499182,39	14490483	16454253
		Total	25490680	680732,81	24151688	26829672
	% del total	Sí	39,3%	1,0%	37,4%	41,2%
		No	60,7%	1,0%	58,8%	62,6%
		Total	100,0%	,0%	100,0%	100,0%
25-44	Tamaño de la población	Sí	39163840	660855,72	37863946	40463734
		No	39503150	645934,19	382322606	40773694
		Total	78666990	961114,33	76776491	80557489
	% del total	Sí	49,8%	,6%	48,7%	50,9%
		No	50,2%	,6%	49,1%	51,3%
		Total	100,0%	,0%	100,0%	100,0%
45-64	Tamaño de la población	Sí	34154952	598603,73	32977507	35332397
		No	24005512	497723,83	23026496	24984528
		Total	58160464	814680,41	56557999	59762929
	% del total	Sí	58,7%	,6%	57,5%	60,0%
		No	41,3%	,6%	40,0%	42,5%
		Total	100,0%	,0%	100,0%	100,0%
65+	Tamaño de la población	Sí	19429991	439459,79	18565580	20294402
		No	11813204	314238,08	11195102	12431306
		Total	31243195	587623,44	30087348	32399042
	% del total	Sí	62,2%	,7%	60,7%	63,6%
		No	37,8%	,7%	36,4%	39,3%
		Total	100,0%	,0%	100,0%	100,0%

Al calcular los estadísticos por subpoblación, se calcula cada estadístico seleccionado para cada una de las medidas de las casillas seleccionadas por el valor de *Categorías de edad*. La primera columna contiene estimaciones del número y el porcentaje de la población de cada categoría que toma o no toma suplementos vitamínicos/minerales. Los intervalos de confianza para los porcentajes de la tabla no se solapan; por lo tanto, se puede concluir que el uso de los suplementos vitamínicos/minerales aumenta con la edad.

Resumen

Mediante el procedimiento Frecuencias de muestras complejas, ha obtenido los estadísticos acerca del consumo de suplementos nutritivos de los ciudadanos de EE.UU.

- En general, hay más americanos que toman suplementos vitamínicos/minerales que los que no los toman.
- Una vez desglosados por categoría de edad, una mayor proporción de americanos toman suplementos vitamínicos/minerales al aumentar la edad.

Procedimientos relacionados

El procedimiento Frecuencias de muestras complejas es una herramienta útil para obtener estadísticos descriptivos univariantes de variables categóricas de las observaciones obtenidas mediante un diseño muestral complejo.

- El Asistente de muestreo de la opción Muestras complejas se utiliza para definir las especificaciones de diseño de las muestras complejas y obtener una muestra. El archivo del plan de muestreo creado por el Asistente de muestreo contiene un plan de análisis por defecto que se puede especificar en el cuadro de diálogo Plan cuando se analiza la muestra obtenida de acuerdo con dicho plan.
- El Asistente de preparación del análisis de la opción Muestras complejas se utiliza para configurar las especificaciones de análisis para una muestra compleja existente. El archivo del plan de muestreo creado por el Asistente de muestreo se puede especificar en el cuadro de diálogo Plan cuando se analiza la muestra correspondiente a dicho plan.
- El procedimiento Tablas de contingencia de muestras complejas proporciona estadísticos descriptivos de las tablas de contingencia de variables categóricas.
- El procedimiento Descriptivos de muestras complejas proporciona estadísticos descriptivos univariantes para variables de escala.

Capítulo
16

Descriptivos de Muestras complejas

El procedimiento Descriptivos de Muestras complejas muestra estadísticos de resumen univariantes para distintas variables. Si lo desea, puede solicitar estadísticos por subgrupos, definidos por una o más variables categóricas.

Uso de los descriptivos de Muestras complejas para analizar los niveles de actividad

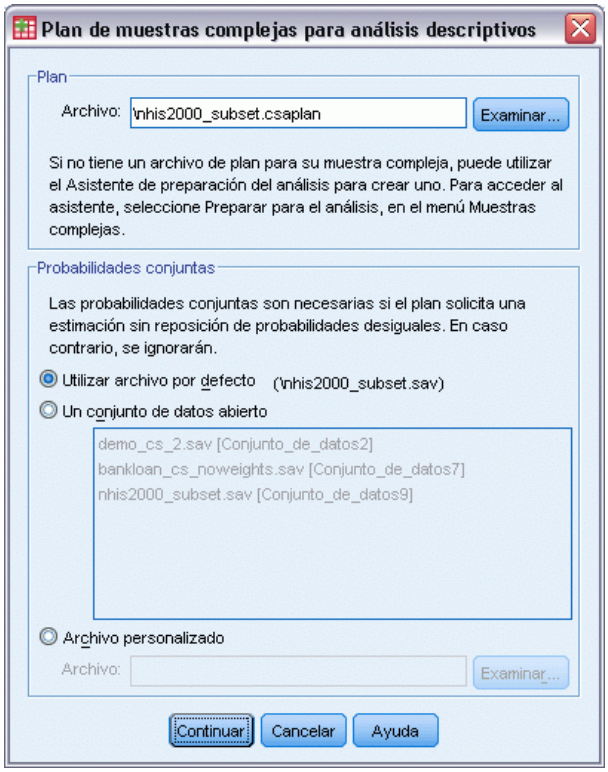
Un investigador desea estudiar los niveles de actividad de los ciudadanos de EE.UU., utilizando los resultados de la National Health Interview Survey (NHIS, Centro Nacional de Estadísticas de Salud) y un plan de análisis anteriormente creado. Si desea obtener más información, consulte el tema Uso del Asistente de preparación del análisis de la opción Muestras complejas para preparar los datos de uso público de la NHIS en el capítulo 14 el p. 146.

Un subconjunto de la encuesta de 2000 se incluye en *nhis2000_subset.sav*. El plan del análisis se guarda en *nhis2000_subset.csaplan*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*. Puede utilizar los descriptivos de Muestras complejas para generar estadísticos descriptivos univariantes para niveles de actividad.

Ejecución del análisis

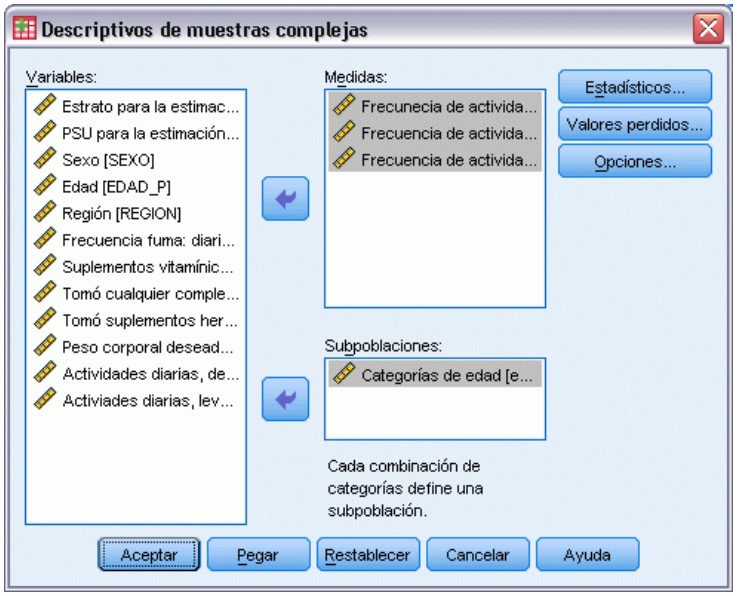
- Para ejecutar un análisis de Descriptivos de Muestras complejas, seleccione en los menús:
  Analizar > Complex Samples > Descriptivos...

Figura 16-1
Cuadro de diálogo Plan de muestras complejas



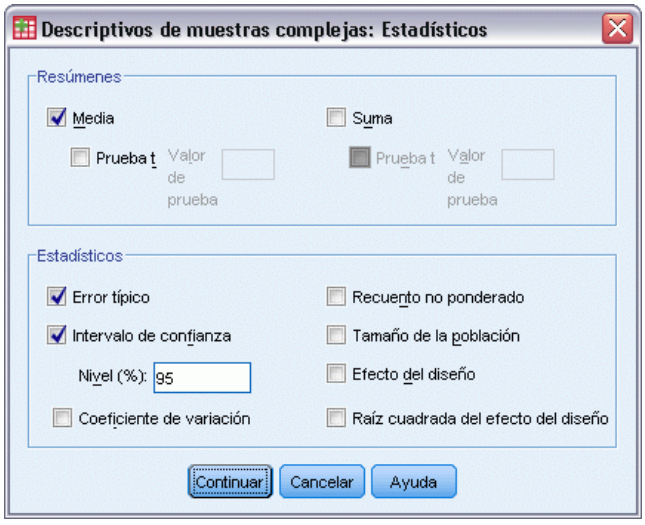
- Acceda al archivo *nhis2000_subset.csaplan* y selecciónelo. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*.
- Pulse en Continuar.

Figura 16-2
Cuadro de diálogo Descriptivos



- Seleccione desde *Frecuencia de actividad vigorosa (veces por semana)* hasta *Frecuencia de actividad de fuerza (veces por semana)* como variables de medida.
- Seleccione *Categorías de edad* como una variable de subpoblación.
- Pulse en Estadísticos.

Figura 16-3
Cuadro de diálogo Descriptivos: Estadísticos



- Seleccione Intervalo de confianza en el grupo Estadísticos.
- Pulse en Continuar.
- Pulse en Aceptar en el cuadro de diálogo Descriptivos de Muestras complejas.

Estadísticos univariantes

Figura 16-4
Estadísticos univariantes

				Intervalo de confianza al 95%	
		Estimación	Error típico	Inferior	Superior
Media	Frecunecia de actividad enérgica (veces por semana)	3,73	,033	3,66	3,79
	Frecuencia de actividad moderada (veces por semana)	4,90	,041	4,82	4,98
	Frecuencia de actividad de fuerza (veces por semana)	3,52	,042	3,43	3,60

Cada estadístico seleccionado se calcula para cada variable de medida. La primera columna contiene estimaciones del número medio de veces a la semana que una persona realiza determinado tipo de actividad. Los intervalos de confianza para las medias no se solapan. Por lo tanto, se puede concluir que, globalmente, los americanos realizan actividades de fuerza con menos frecuencia que actividades vigorosas y que realizan actividades vigorosas con menos frecuencia que actividades moderadas.

Estadísticos univariantes por subpoblación

Figura 16-5
Estadísticos univariantes por subpoblación

Categorías de edad			Estimación	Error típico	Intervalo de confianza al 95%	
					Inferior	Superior
18-24	Media	Frecunecia de actividad enérgica (veces por semana)	3,92	,087	3,75	4,09
		Frecuencia de actividad moderada (veces por semana)	5,18	,137	4,91	5,45
		Frecuencia de actividad de fuerza (veces por semana)	3,45	,085	3,28	3,62
25-44	Media	Frecunecia de actividad enérgica (veces por semana)	3,55	,048	3,46	3,65
		Frecuencia de actividad moderada (veces por semana)	4,73	,056	4,62	4,84
		Frecuencia de actividad de fuerza (veces por semana)	3,28	,052	3,18	3,38
45-64	Media	Frecunecia de actividad enérgica (veces por semana)	3,79	,063	3,66	3,91
		Frecuencia de actividad moderada (veces por semana)	4,88	,070	4,74	5,02
		Frecuencia de actividad de fuerza (veces por semana)	3,65	,092	3,47	3,84
65+	Media	Frecunecia de actividad enérgica (veces por semana)	4,18	,111	3,96	4,39
		Frecuencia de actividad moderada (veces por semana)	5,22	,084	5,06	5,39
		Frecuencia de actividad de fuerza (veces por semana)	4,66	,155	4,36	4,97

Cada estadístico seleccionado se calcula para cada variable de medida según los valores de *Categorías de edad*. La primera columna contiene estimaciones del número medio de veces a la semana que las personas de cada categoría realizan un determinado tipo de actividad. Los intervalos de confianza de las medias permiten extraer ciertas interesantes conclusiones.

- En lo que se refiere a las actividades vigorosas y moderadas, las personas de 25–44 años son menos activos que las de 18–24 y las de 45–64, mientras que las personas de 45–64 años son menos activas que las de 65 o mayores.
- En lo que se refiere a las actividades de fuerza, las personas de 25–44 años son menos activas que las de 45–64, mientras que las personas de 18–24 y 45–64 años son menos activas que las de 65 o mayores.

Resumen

Mediante el procedimiento Descriptivos de Muestras complejas, ha obtenido los estadísticos de los niveles de actividad de los ciudadanos de EE.UU.

- En general, los americanos pasan diferentes intervalos de tiempo realizando diferentes tipos de actividades.
- Una vez desglosados por edades, los datos parecen indicar que los americanos que han finalizado sus estudios universitarios son en principio menos activos que cuando estaban estudiando, pero conforme envejecen vez más pasan a ser más conscientes de la necesidad de hacer ejercicio.

Procedimientos relacionados

El procedimiento Descriptivos de Muestras complejas es una herramienta útil para obtener estadísticos descriptivos univariantes de las medidas de escala de las observaciones obtenidas mediante un diseño muestral complejo.

- El Asistente de muestreo de la opción Muestras complejas se utiliza para definir las especificaciones de diseño de las muestras complejas y obtener una muestra. El archivo del plan de muestreo creado por el Asistente de muestreo contiene un plan de análisis por defecto que se puede especificar en el cuadro de diálogo Plan cuando se analiza la muestra obtenida de acuerdo con dicho plan.
- El Asistente de preparación del análisis de la opción Muestras complejas se utiliza para configurar las especificaciones de análisis para una muestra compleja existente. El archivo del plan de muestreo creado por el Asistente de muestreo se puede especificar en el cuadro de diálogo Plan cuando se analiza la muestra correspondiente a dicho plan.
- El procedimiento Razones de muestras complejas proporciona estadísticos descriptivos para razones de medidas de escala.
- El procedimiento Frecuencias de muestras complejas proporciona estadísticos descriptivos univariantes para variables categóricas.

Capítulo
17

Tablas de contingencia de Muestras complejas

El procedimiento Tablas de contingencia de Muestras complejas genera tablas de contingencia para los pares de variables seleccionadas y muestra estadísticos sobre la clasificación bivariante. Si lo desea, puede solicitar estadísticos por subgrupos, definidos por una o más variables categóricas.

Uso de muestras complejas de tablas de contingencia para medir el riesgo relativo de un evento

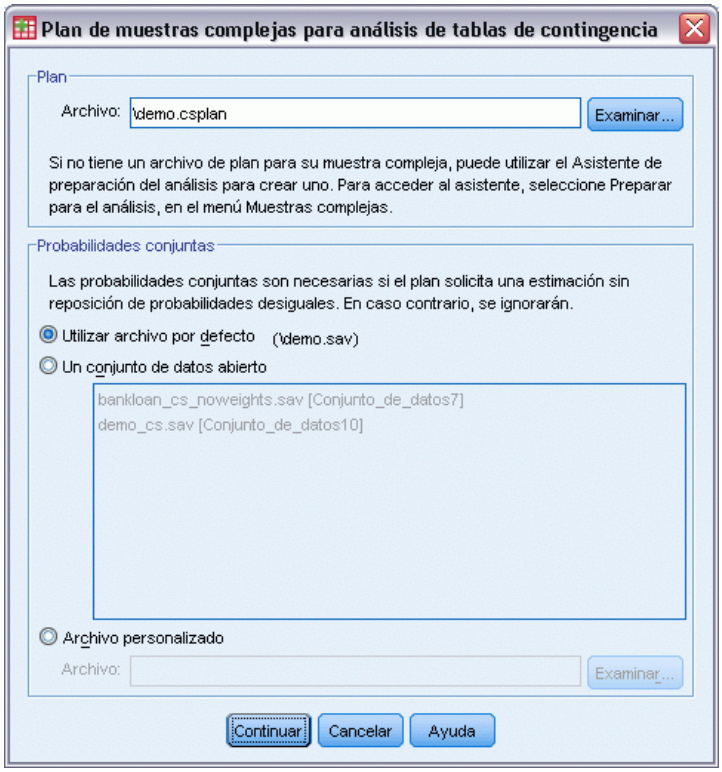
Una compañía que vende suscripciones a revistas suele enviar todos los meses mailings a los nombres que aparecen en una base de datos que ha adquirido. La tasa de respuesta normalmente es muy baja, por lo que necesita encontrar una manera de dirigirse mejor a los posibles clientes. Una sugerencia consiste en concentrar el envío de mailings a aquellas personas que ya están suscritas a periódicos, basándose en el supuesto de que las personas que leen periódicos tienen mayor propensión a suscribirse a revistas.

Se puede utilizar el procedimiento Tablas de contingencia de Muestras complejas para probar esta teoría construyendo una tabla de dos filas por dos columnas de *Suscrito a un periódico* por *Responde* y calcular el riesgo relativo de que una persona que esté suscrita a un periódico responda al mailing. Esta información se recoge en el archivo *demo_cs.sav* y deberán analizarse utilizando el archivo del plan de muestreo *demo.csplan*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*.

Ejecución del análisis

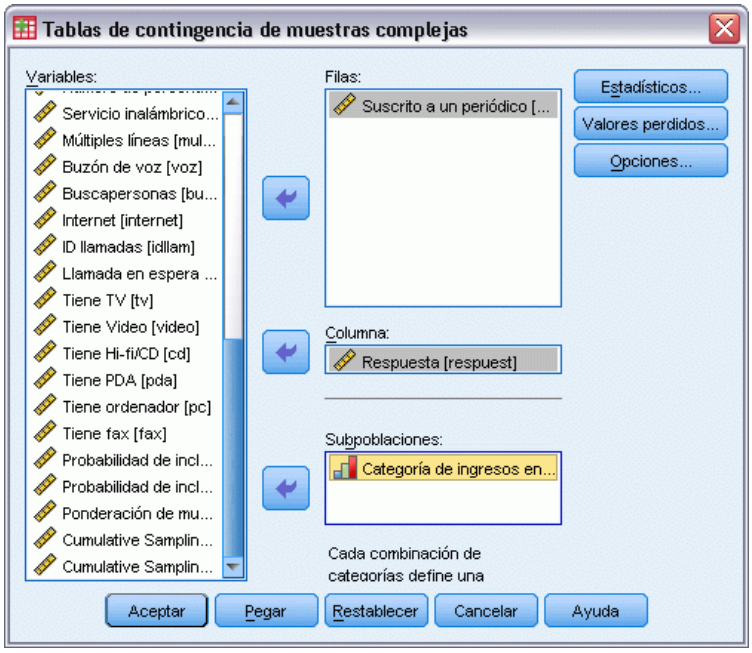
- Para ejecutar un análisis de tablas de contingencia de Muestras complejas, seleccione en los menús:
  Analizar > Complex Samples > Tablas de contingencia...

Figura 17-1
Cuadro de diálogo Plan de muestras complejas



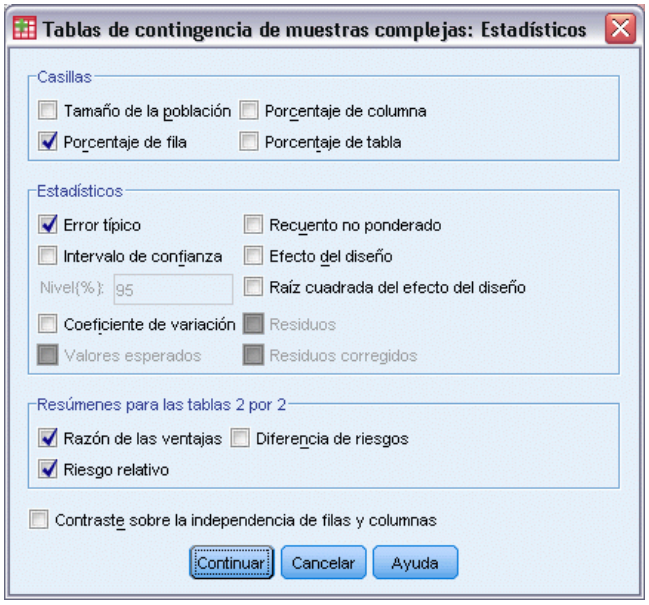
- Acceda al archivo *demo.csplan* y selecciónelo. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*.
- Pulse en Continuar.

Figura 17-2
Cuadro de diálogo Tablas de contingencia



- Seleccione *Suscrito a un periódico* como variable de fila.
- Seleccione *Responde* como una variable de columna.
- También resulta interesante ver los resultados desglosados por categorías de ingresos, así que seleccione *Categoría de ingresos en miles* como variable de subpoblación.
- Pulse en Estadísticos.

Figura 17-3
Cuadro de diálogo Tablas de contingencia: Estadísticos



- Anule la selección de Tamaño de la población y seleccione Porcentaje de fila en el grupo Casillas.
- Seleccione Razón de las ventajas y Riesgo relativo en el grupo Resúmenes para las tablas 2 por 2.
- Pulse en Continuar.
- Pulse en Aceptar en el cuadro de diálogo Tablas de contingencia de Muestras complejas.

  Estas selecciones generarán una tabla de contingencia y una estimación del riesgo de *Suscrito a un periódico* por *Responde*. También se crean tablas diferentes con los resultados divididos por *Categoría de ingresos en miles*.

Tabla de contingencia

Figura 17-4
Tabla de contingencia de suscripción a un periódico por respuesta

Suscrito a un periódico			Respuesta		
			Sí	No	Total
Sí	% de Suscrito a un periódico	Estimación	17,2%	82,8%	100,0%
		Error típico	1,0%	1,0%	,0%
No	% de Suscrito a un periódico	Estimación	10,3%	89,7%	100,0%
		Error típico	,7%	,7%	,0%
Total	% de Suscrito a un periódico	Estimación	12,8%	87,2%	100,0%
		Error típico	,7%	,7%	,0%

La tabla de contingencia muestra que, en general, pocas personas respondieron al mailing. No obstante, respondió una mayor proporción de personas suscritas a periódicos.

Estimación de riesgo

Figura 17-5
Estimación del riesgo para la suscripción a un periódico por respuesta

			Estimación
Suscrito a un periódico * Respuesta	Razón de ventajas		1,812
	Riesgo relativo	Para la cohorte Respuesta = Sí	1,673
		Para la cohorte Respuesta = No	,923

Sólo se computan los estadísticos para las tablas 2 por 2 con todas las casillas observadas.

El riesgo relativo es una razón de probabilidades de eventos. El riesgo relativo de una respuesta al mailing es la razón de la probabilidad de que una persona suscrita a un periódico responda, respecto a la probabilidad de que una persona que no está suscrita lo haga. Por tanto, la estimación el riesgo relativo es sencillamente 17,2%/10,3% = 1.673. Igualmente, el riesgo relativo de que no haya respuesta es la razón de la probabilidad de que una persona suscrita no responda, respecto a la probabilidad de que una persona no suscrita no responda. La estimación de este riesgo relativo es 0.923. Con estos resultados, puede estimar que es 1.673 veces más probable que una persona suscrita a un periódico responda al mailing que una persona que no lo está, o 0.923 veces tan probable que no responda como una persona que no está suscrita.

La razón de las ventajas es la razón de las ventajas de los eventos. Las ventajas de un evento es la razón de la probabilidad de que ocurra el evento, respecto a la probabilidad de que no ocurra el evento. Por tanto, la estimación de las ventajas de que una persona suscrita a un periódico responda al mailing es de 17.2%/82.8% = 0.208. Igualmente, la estimación de las ventajas de que una persona no suscrita responda es de 10.3%/89.7% = 0.115. La estimación de la ocurrencia del factor es por tanto 0.208/0.115 = 1.812 (tenga en cuenta que existe un error de redondeo en los paso intercalados). La razón de las ventajas es la razón del riesgo relativo de responder, respecto al riesgo relativo de no responder, o sea 1.673/0.923 = 1.812.

Razón de las ventajas respecto al riesgo relativo

Ya que se trata de una razón de razones, la razón de las ventajas es muy difícil de interpretar. El riesgo relativo es más fácil de interpretar, por lo que la razón de las ventajas por sí sola no resulta muy útil. Sin embargo, hay determinadas situaciones muy habituales en las que la estimación del riesgo relativo no es muy buena y la razón de las ventajas se puede utilizar para calcular una aproximación del riesgo relativo del evento de interés. La razón de las ventajas se puede utilizar como aproximación del riesgo relativo del evento de interés cuando se cumplen las dos siguientes condiciones:

- La probabilidad del evento de interés es pequeña (<0,1). Esta condición garantiza que la razón de las ventajas será una buena aproximación del riesgo relativo. En este ejemplo, el evento de interés es una respuesta al mailing.
- El diseño del estudio es un control de casos. Esta condición indica que la estimación habitual del riesgo relativo probablemente no sea buena. Un estudio de control de casos es retrospectivo, se utiliza sobre todo cuando el evento de interés es poco probable o cuando el diseño de un futuro experimento es poco práctico o poco ético.

Ninguna de estas condiciones se cumple en este ejemplo, ya que la proporción global de personas que respondieron fue del 12.8% y el diseño del estudio no fue un control de casos, por lo que resulta más seguro tomar 1.673 como el riesgo relativo, en vez del valor de la razón de las ventajas.

Estimación del riesgo por subpoblación

Figura 17-6
Estimación del riesgo para la suscripción a un periódico por respuesta, con control de la categoría de ingresos

Categoría de				Estimación
menos de $25	Suscrito a un periódico * Respuesta	Razón de ventajas		2,712
		Riesgo relativo	Para la cohorte Respuesta = Sí	2,241
			Para la cohorte Respuesta = No	,826
$25 - $49	Suscrito a un periódico * Respuesta	Razón de ventajas		1,794
		Riesgo relativo	Para la cohorte Respuesta = Sí	1,645
			Para la cohorte Respuesta = No	,917
$50 - $74	Suscrito a un periódico * Respuesta	Razón de ventajas		1,168
		Riesgo relativo	Para la cohorte Respuesta = Sí	1,152
			Para la cohorte Respuesta = No	,986
$75+	Suscrito a un periódico * Respuesta	Razón de ventajas		1,242
		Riesgo relativo	Para la cohorte Respuesta = Sí	1,227
			Para la cohorte Respuesta = No	,988

Sólo se computan los estadísticos para las tablas 2 por 2 con todas las casillas observadas.

Las estimaciones del riesgo relativo se calculan por separado para cada categoría de ingresos. Observe que el riesgo relativo de una respuesta positiva de las personas suscritas a un periódico parece disminuir gradualmente al aumentar los ingresos, lo que indica que es posible limitar aún más los destinatarios del mailing.

Resumen

Mediante las estimaciones del riesgo de las tablas de contingencia de Muestras complejas, ha descubierto que puede aumentar la tasa de respuesta a los mailings directos dirigiéndose a personas suscritas a periódicos. Además, encuentra cierta evidencia de que las estimaciones de riesgo puede que no sean constantes dependiendo de la *Categoría de ingresos*, por lo que puede aumentar aún más la tasa de respuesta si se dirige a las personas suscritas a periódicos que tienen menores ingresos.

Procedimientos relacionados

El procedimiento Tablas de contingencia de Muestras complejas es una herramienta útil para obtener estadísticos descriptivos de las tablas de contingencia de variables categóricas de observaciones obtenidas mediante un diseño muestral complejo.

- El Asistente de muestreo de la opción Muestras complejas se utiliza para definir las especificaciones de diseño de las muestras complejas y obtener una muestra. El archivo del plan de muestreo creado por el Asistente de muestreo contiene un plan de análisis por defecto que se puede especificar en el cuadro de diálogo Plan cuando se analiza la muestra obtenida de acuerdo con dicho plan.
- El Asistente de preparación del análisis de la opción Muestras complejas se utiliza para configurar las especificaciones de análisis para una muestra compleja existente. El archivo del plan de muestreo creado por el Asistente de muestreo se puede especificar en el cuadro de diálogo Plan cuando se analiza la muestra correspondiente a dicho plan.
- El procedimiento Frecuencias de muestras complejas proporciona estadísticos descriptivos univariantes para variables categóricas.

Capítulo
18

Razones de Muestras complejas

El procedimiento Razones de Muestras complejas muestra estadísticos de resumen univariantes para razones de variables. Si lo desea, puede solicitar estadísticos por subgrupos, definidos por una o más variables categóricas.

Uso de razones de Muestras complejas como ayuda en la evaluación de los valores de las propiedades

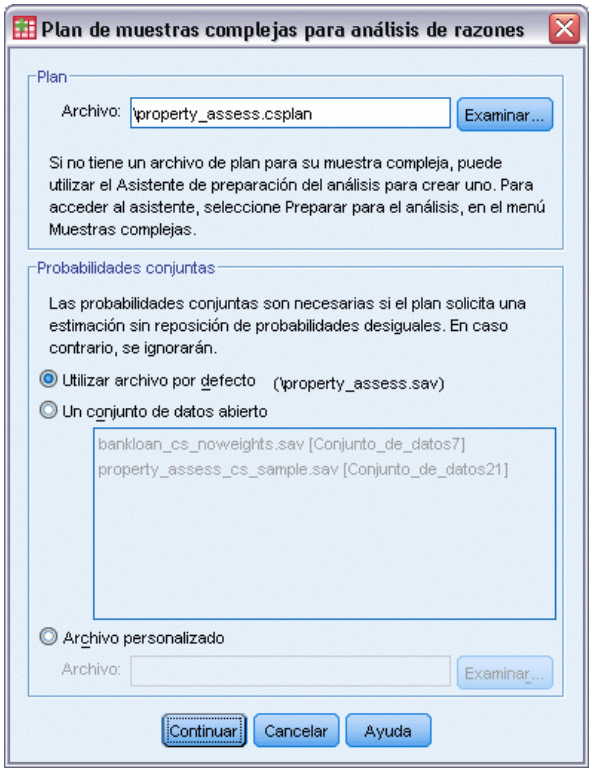
Una agencia inmobiliaria se encarga que asegurar que los impuestos sobre las propiedades se evalúan de la misma manera en los diferentes condados. Los impuestos se basan en el valor tasado de la propiedad, por lo que la agencia desea realizar un seguimiento de los valores de las propiedades en diferentes condados para asegurarse de que los registros de todos los condados están igualmente actualizados. Ya que los recursos necesarios para obtener las tasaciones actuales son limitados, la agencia decide utilizar una metodología de muestreo complejo para seleccionar las propiedades.

La muestra de propiedades seleccionadas y su información de tasación actual se recoge en *property_assess_cs_sample.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*. Uso de razones de Muestras complejas para evaluar el cambio de los valores de las propiedades desde la última tasación en cinco condados.

Ejecución del análisis

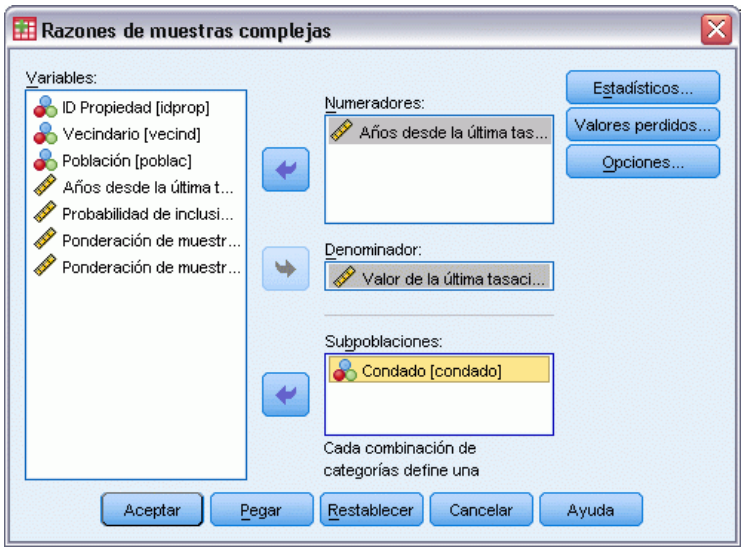
- Para ejecutar un análisis de razones de Muestras complejas, seleccione en los menús:
  Analizar > Complex Samples > Razones...

Figura 18-1
Cuadro de diálogo Plan de muestras complejas



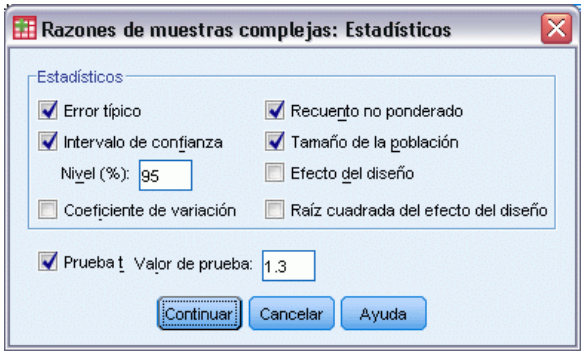
- Acceda al archivo *property_assess.csplan* y selecciónelo. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*.
- Pulse en Continuar.

Figura 18-2
Cuadro de diálogo Razones de Muestras complejas



- Seleccione *Valor actual* como variable de numerador.
- Seleccione *Valor de la última tasación* como la variable de denominador.
- Seleccione *Condado* como variable de subpoblación.
- Pulse en Estadísticos.

Figura 18-3
Cuadro de diálogo Estadísticos de la razón



- Seleccione Intervalo de confianza, Recuento no ponderado y Tamaño de la población en el grupo Estadísticos.
- Seleccione Prueba t y escriba 1,3 como valor de prueba.
- Pulse en Continuar.
- Pulse en Aceptar en el cuadro de diálogo Razones de Muestras complejas.

Razones

Figura 18-4
Tabla de razones

Condado	Numerador	Denominador	Estimación de la razón	Error típico	Intervalo de confianza al 95%	
					Inferior	Superior
Este	Current value	Valor de la última tasación	1,381	,068	1,236	1,525
Central	Current value	Valor de la última tasación	1,364	,064	1,227	1,502
Oeste	Current value	Valor de la última tasación	1,524	,053	1,410	1,638
Norte	Current value	Valor de la última tasación	1,277	,032	1,208	1,346
Sur	Current value	Valor de la última tasación	1,195	,029	1,134	1,256

La presentación por defecto de la tabla es muy ancha, por lo que deberá pivotarla para poder verla con mayor claridad.

Pivotado de la tabla de razones

- ▶ Pulse dos veces en la tabla para activarla.
- ▶ Seleccione en los menús del Visor:
  Pivotar > Paneles de pivotado
- ▶ Arrastre *Numerador* y, a continuación, *Denominador* desde la fila a la capa.
- ▶ Arrastre *Condado* desde la fila a la columna.
- ▶ Arrastre *Estadísticos* desde la columna a la fila.
- ▶ Cierre la ventana Paneles de pivotado.

Tabla de razones pivotada

Figura 18-5
Tabla de razones pivotada

Denominador Valor de la última tasación

Numerador Años desde la última tasación

		Condado				
		Este	Central	Oeste	Norte	Sur
Estimación de la razón		1,381	1,364	1,524	1,277	1,195
Error típico		,068	,064	,053	,032	,029
Intervalo de confianza al 95%	Inferior	1,236	1,227	1,410	1,208	1,134
	Superior	1,525	1,502	1,638	1,346	1,256
Contraste de hipótesis	Valor de contraste	1,3	1,3	1,3	1,3	1,3
	t	1,191	,997	4,201	-,702	-3,646
	gl	15	15	15	15	15
	Sig.	,252	,334	,001	,493	,002
Recuento no ponderado		168	179	202	205	220

La tabla de razones ahora está pivotada de manera que resulta más fácil comparar los estadísticos correspondientes a los diferentes condados.

- Las estimaciones de las razones varían desde un mínimo de 1,195 en el condado del sur hasta un máximo de 1,524 en el condado del oeste.
- También hay bastante variación en los errores típicos, que oscilan desde un mínimo de 0,029 en el condado del sur hasta un máximo de 0,068 en el condado del este.
- Algunos de los intervalos de confianza no se solapan; por tanto, se puede concluir que las razones del condado del oeste son mayores que las razones de los condados del norte y del sur.
- Por último, como medida más objetiva, observe que los valores de significación de las pruebas *t* de los condados del oeste y del sur son menores de 0,05. Por tanto, se puede concluir que la razón del condado del oeste es mayor que 1,3 y la razón del condado del sur es menor que 1,3.

Resumen

Mediante el procedimiento Razones de Muestras complejas, hemos obtenido varios estadísticos para las razones del *Valor actual* respecto al *Valor de la última tasación*. Los resultados sugieren que tal vez existan cierta falta de armonización en la evaluación de los impuestos sobre las propiedades en los diferentes condados, concretamente:

- Las razones del condado del oeste son altas, lo que indica que sus registros no están tan actualizados como los de otros condados en lo que se refiere a la apreciación de los valores de las propiedades. Los impuestos sobre las propiedades son probablemente demasiado bajos en este condado.

- Las razones del condado del sur son bajas, lo que indica que sus registros son más actualizados que los de los otros condados en lo que se refiere a la apreciación de los valores de las propiedades. Los impuestos sobre las propiedades son probablemente demasiado altos en este condado.
- Las razones del condado del sur son inferiores que las del condado del oeste, pero se mantienen dentro del objetivo de 1,3.

Los recursos utilizados para realizar el seguimiento de los valores en el condado del sur se asignarán al condado del sur para armonizar las razones de estos condados con los demás y con el objetivo de 1,3.

Procedimientos relacionados

El procedimiento Razones de Muestras complejas es una herramienta útil para obtener estadísticos descriptivos univariantes de la razón de las medidas de escala de las observaciones obtenidas mediante un diseño muestral complejo.

- El Asistente de muestreo de la opción Muestras complejas se utiliza para definir las especificaciones de diseño de las muestras complejas y obtener una muestra. El archivo del plan de muestreo creado por el Asistente de muestreo contiene un plan de análisis por defecto que se puede especificar en el cuadro de diálogo Plan cuando se analiza la muestra obtenida de acuerdo con dicho plan.
- El Asistente de preparación del análisis de la opción Muestras complejas se utiliza para configurar las especificaciones de análisis para una muestra compleja existente. El archivo del plan de muestreo creado por el Asistente de muestreo se puede especificar en el cuadro de diálogo Plan cuando se analiza la muestra correspondiente a dicho plan.
- El procedimiento Descriptivos de muestras complejas proporciona estadísticos descriptivos univariantes para variables de escala.

Capítulo
19

Modelo lineal general de muestras complejas

El procedimiento Modelo lineal general de muestras complejas (CSGLM) realiza análisis de regresión lineal y análisis de varianza y covarianza de muestras extraídas mediante métodos de muestreo complejo. Si lo desea, puede solicitar análisis de una subpoblación.

Uso del Modelo lineal general de muestras complejas para ajustar ANOVA de dos factores

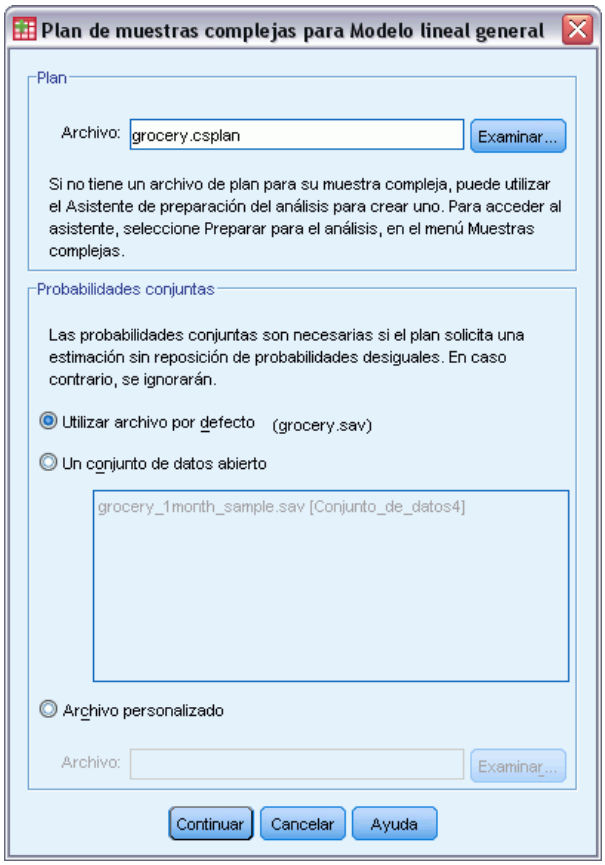
Una cadena de tiendas de alimentación realiza una encuesta sobre los hábitos de compra de una serie de clientes basándose en un diseño complejo. Una vez obtenidos los resultados de la encuesta y la cantidad que cada cliente gastó el mes anterior, la cadena desea averiguar si la frecuencia con que los clientes hacen la compra está relacionada con la cantidad mensual que gastan, controlando el sexo del cliente e incorporando el diseño del muestreo.

Esta información se recoge en el archivo *grocery_1month_sample.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*. Utilice el procedimiento Modelo lineal general de muestras complejas para realizar un análisis ANOVA de dos factores de las cantidades gastadas.

Ejecución del análisis

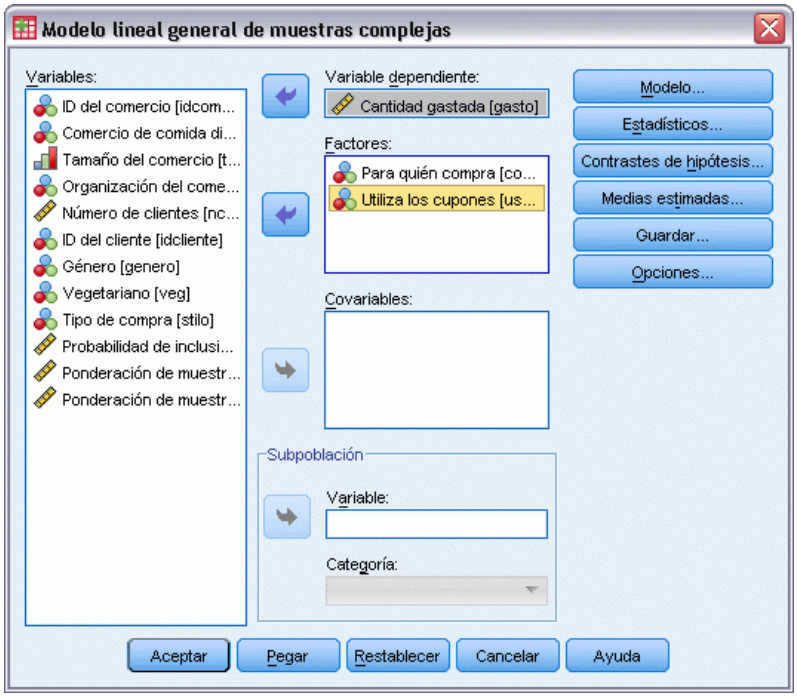
- Para ejecutar un análisis de Modelo lineal general de muestras complejas, seleccione en los menús:
  Analizar > Complex Samples > Modelo lineal general...

Figura 19-1
Cuadro de diálogo Plan de muestras complejas



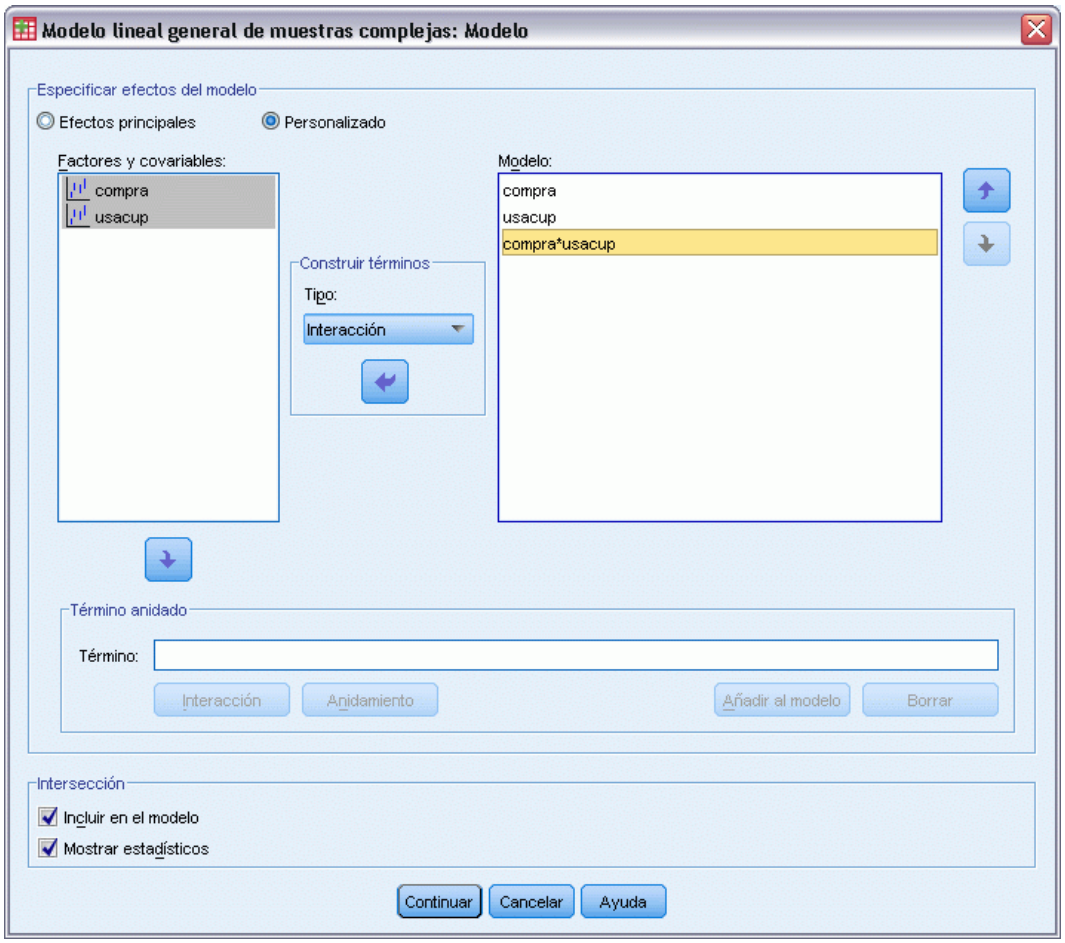
- Acceda al archivo *grocery.csplan* y selecciónelo. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*.
- Pulse en Continuar.

Figura 19-2
Cuadro de diálogo Modelo lineal general de muestras complejas



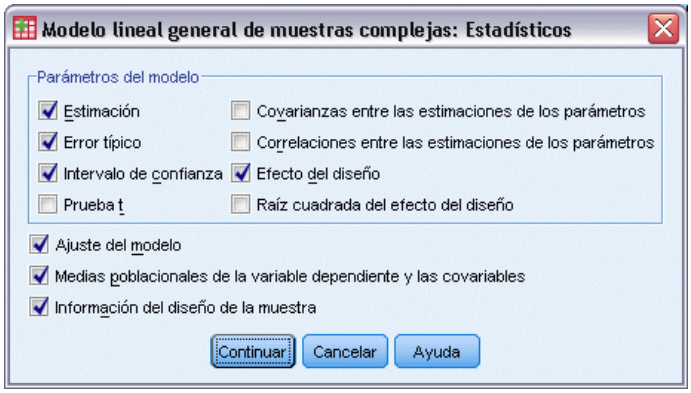
- Seleccione *Amount spent* como la variable dependiente.
- Seleccione *Who shopping for* y *Use coupons* como factores.
- Pulse en Modelo.

Figura 19-3
Cuadro de diálogo Modelo



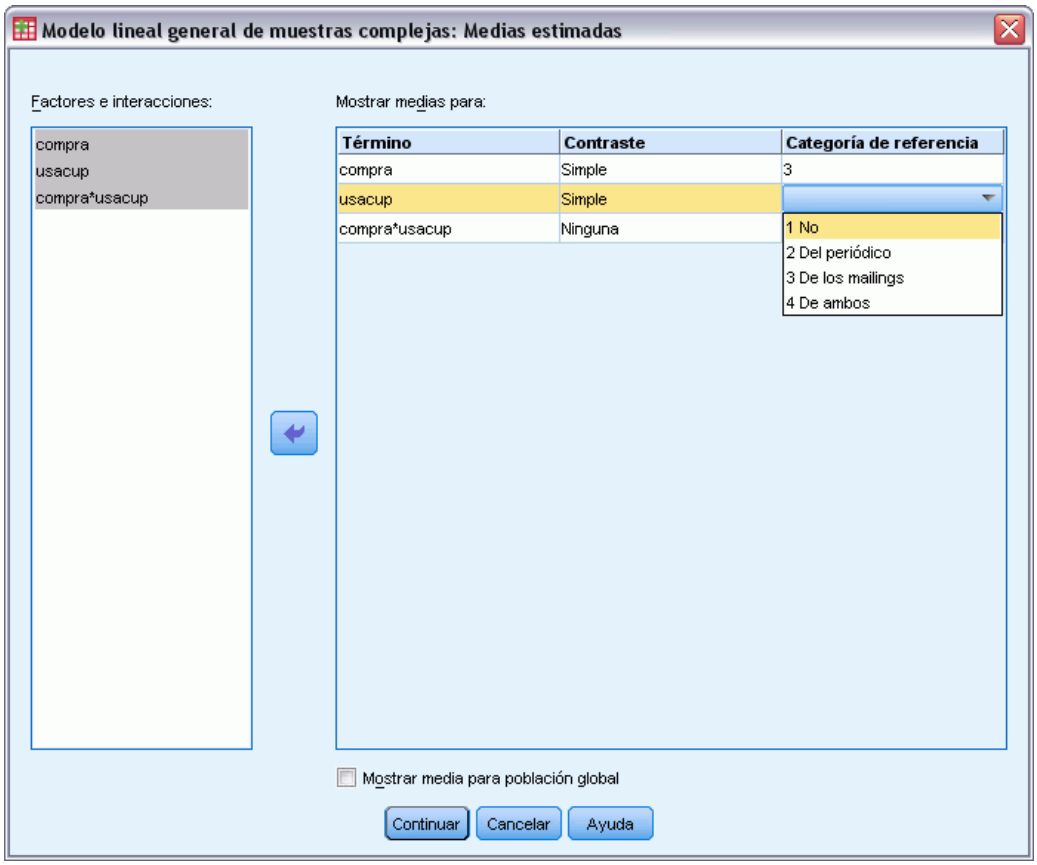
- Elija crear un modelo Personalizado.
- Seleccione Efectos principales como tipo de término que se va a crear y seleccione *shopfor* y *usecoup* como términos del modelo.
- Seleccione Interacción como tipo de término que se va a crear y añada la interacción *shopfor*usecoup* como un término del modelo.
- Pulse en Continuar.
- En el Cuadro de diálogo Modelo lineal general de muestras complejas, pulse en Estadísticos.

Figura 19-4
Cuadro de diálogo Modelo lineal general de muestras complejas: Estadísticos



- Seleccione Estimación, Error típico, Intervalo de confianza y Efecto del diseño en el grupo Parámetros del modelo.
- Pulse en Continuar.
- En el Cuadro de diálogo Modelo lineal general de muestras complejas, pulse en Medias estimadas.

Figura 19-5
Cuadro de diálogo Modelo lineal general de muestras complejas: Medias estimadas



- Elija mostrar las medias para *shopfor*, *usecoup* y la interacción *shopfor*usecoup*.
- Seleccione un contraste Simple y 3 Self and family como la categoría de referencia para *shopfor*. Observe que, una vez seleccionada, la categoría aparece como "3" en el cuadro de diálogo.
- Seleccione un contraste Simple y 1 No como la categoría de referencia para *usecoup*.
- Pulse en Continuar.
- En el Cuadro de diálogo Modelo lineal general de muestras complejas, pulse en Aceptar.

Resumen del modelo

Figura 19-6
estadístico R cuadrado

R cuadrado	.601

a. Modelo: Cantidad gastada = (Intersección) + compra + usacup + compra * usacup

R cuadrado, el coeficiente de determinación, es una medida de la fuerza del ajuste del modelo. Muestra que el modelo explica cerca del 60% de la variación en *Amount spent*, lo que ofrece una buena capacidad explicativa. Es posible que desee añadir otros predictores al modelo para mejorar aún más el ajuste.

Pruebas de efectos del modelo

Figura 19-7
Pruebas de los efectos inter-sujetos

Origen	gl1	gl2	F de Wald	Sig.
(Modelo corregido)	11,000	3,000	127,231	,001
(Intersección)	1,000	13,000	6321,597	,000
compra	2,000	12,000	643,593	,000
usacup	3,000	11,000	87,453	,000
compra * usacup	6,000	8,000	10,688	,002

a. Modelo: Cantidad gastada = (Intersección) + compra + usacup + compra * usacup

Cada término del modelo, además del propio modelo, se prueba para comprobar si el valor de su efecto es igual a 0. Los términos con valores de significación inferiores a 0,05 tienen algún efecto perceptible. Por lo tanto, todos los términos del modelo contribuyen a él.

Estimaciones de los parámetros

Figura 19-8
Estimaciones de los parámetros

Parámetro	Estimación	Error típico	Intervalo de confianza al 95%		Efecto del diseño
			Lower	Upper	
(Intersección)	518,249	11,731	492,905	543,592	1,387
[compra=1]	-174,757	10,762	-198,0	-151,51	,950
[compra=2]	-129,443	11,455	-154,2	-104,70	,925
[compra=3]	,000[a]	.	.	.	.
[usacup=1]	-140,838	10,180	-162,8	-118,85	,649
[usacup=2]	-63,026	13,195	-91,531	-34,520	,940
[usacup=3]	-31,375	9,726	-52,387	-10,363	,564
[usacup=4]	,000[a]	.	.	.	.
[compra=1] * [usacup=1]	41,693	11,170	17,562	65,824	,606
[compra=1] * [usacup=2]	44,505	18,068	5,471	83,539	1,413
[compra=1] * [usacup=3]	9,204	11,057	-14,684	33,092	,594
[compra=1] * [usacup=4]	,000[a]	.	.	.	.
[compra=2] * [usacup=1]	89,211	10,967	65,518	112,903	,533
[compra=2] * [usacup=2]	54,267	14,949	21,972	86,562	,836
[compra=2] * [usacup=3]	17,884	13,753	-11,828	47,595	,797
[compra=2] * [usacup=4]	,000[a]	.	.	.	.
[compra=3] * [usacup=1]	,000[a]	.	.	.	.
[compra=3] * [usacup=2]	,000[a]	.	.	.	.
[compra=3] * [usacup=3]	,000[a]	.	.	.	.
[compra=3] * [usacup=4]	,000[a]	.	.	.	.

a. Establecido en cero porque este parámetro es redundante.
b. Modelo: Cantidad gastada = (Intersección) + compra + usacup + compra * usacup

Las estimaciones de los parámetros muestran los efectos de cada predictor en *Amount spent*. El valor 518,249 del término de intersección indica que la cadena de productos alimenticios puede esperar que un comprador con familia que utiliza cupones de los periódicos y mailings dirigidos se gaste 518,25 dólares de media. Se puede decir que la intersección está asociada con dichos niveles de factor porque esos son los niveles de factor cuyos parámetros son redundantes.

- Los coeficientes de *shopfor* sugieren que, entre los clientes que utilizan tanto los cupones de los periódicos como los recibidos por mailing, aquellos que no tienen familia tienden a gastar menos que los clientes con cónyuge, quienes a su vez gastan menos que los clientes que vivan con personas a su cargo. Como las pruebas de los efectos del modelo demostraron que este término contribuía al modelo, estas diferencias no se deben a la casualidad.
- Los coeficientes *usecoup* sugieren que el gasto entre los clientes con personas a su cargo desciende con el menor uso de cupones. Existe una moderada cantidad de incertidumbre en las estimaciones, pero los intervalos de confianza no incluyen el 0.
- Los coeficientes de interacción sugieren que los clientes que no usan cupones o sólo recortes del periódico y no tienen personas a su cargo tienden a gastar más de lo que se podría esperar. Si alguna parte de un parámetro de interacción es redundante, el parámetro de interacción será redundante.
- La desviación del 1 en los valores de los efectos del diseño indica que algunos de los errores típicos calculados para estas estimaciones de parámetros son mayores que los que se obtendrían si se supone que dichas observaciones proceden de una muestra aleatoria simple, mientras que los demás son más pequeños. Es de vital importancia incorporar la información sobre el diseño muestral al análisis porque, en caso contrario, se podría inferir, por ejemplo, que el coeficiente *usecoup=3* no es distinto de 0.

Las estimaciones de los parámetros son útiles para cuantificar el efecto de cada uno de los términos del modelo, pero las tablas de medias marginales estimadas pueden simplificar la interpretación de los resultados del modelo.

Medias marginales estimadas

Figura 19-9
Medias marginales estimadas por niveles de Who shopping for

Para quién compra	Media	Error típico	Intervalo de confianza al 95%	
			Inferior	Superior
Él mismo	308,5326	3,94286	300,0145	317,0506
Él mismo y esposa	370,3361	4,87908	359,7955	380,8767
Él mismo y familia	459,4392	7,19769	443,8895	474,9888

Esta tabla muestra las medias marginales estimadas por el modelo y los errores típicos de *Amount spent* en los niveles de factor de *Who shopping for*. Esta tabla es útil para explorar las diferencias entre los niveles de este factor. En este ejemplo, un cliente que compra para sí mismo se espera que gaste cerca de 308,53 dólares, mientras que un cliente casado se espera que gaste unos 370,34 dólares y un cliente con personas a su cargo gastará unos 459,44 dólares. Para comprobar si esto representa una diferencia real o puede deberse a una variación debida al azar, examine los resultados de la prueba.

Figura 19-10
Resultados de las pruebas individuales para medias marginales estimadas de sexo

Contrastes simples Para quién compra[a]	Estimación del contraste	Valor hipotetizado	Diferencia (estimación - hipotetizado)	Error típico	gl1	gl2	F de Wald	Sig.
Nivel Él mismo frente a nivel Él mismo y familia	-150,907	,000	-150,907	4,903	1,000	13,00	947,41	.
Nivel Él mismo y esposa frente a nivel Él mismo y familia	-89,103	,000	-89,103	5,903	1,000	13,00	227,84	.

a. Categoría de referencia = Él mismo y familia

La tabla de las pruebas individuales muestra dos contrastes simples en el gasto.

- La estimación del contraste es la diferencia en el gasto para los niveles de *Who shopping for*.
- El valor hipotetizado de 0,00 representa la creencia de que no hay diferencia en el gasto.
- El estadístico *F* de Wald, con los grados de libertad que se muestran, se utiliza para probar si la diferencia entre una estimación de contraste y el valor hipotetizado es por una variación debida al azar.
- Como los valores de significación son inferiores a 0,05, se puede concluir que existen diferencias en el gasto.

Los valores de las estimaciones de los contrastes son distintos a los de las estimaciones de los parámetros. Esto se debe a que hay un término de interacción que contiene el efecto de *Who shopping for*. Como resultado, la estimación de los parámetros para *shopfor=1* es un contraste simple entre los niveles *Self* y *Self and Family* en el nivel *From both* de la variable *Use coupons*. La estimación del contraste en esta tabla se promedia sobre los niveles de *Use coupons*.

Figura 19-11
Resultados de las pruebas globales para medias marginales estimadas de sexo

gl1	gl2	F de Wald	Sig.
2,000	12,000	643,593	,000

La tabla de pruebas globales informa de los resultados de una prueba de todos los contrastes de la tabla de pruebas individuales. Su valor de significación menor que 0,05 confirma que existe una diferencia en el gasto entre los niveles de *Who shopping for*.

Figura 19-12
Medias marginales estimadas por niveles de estilo de compra

			Intervalo de confianza al 95%	
Utiliza los cupones	Media	Error típico	Inferior	Superior
No	319,6455	6,51429	305,5722	333,7188
Del periódico	386,7469	4,32295	377,4077	396,0861
De los mailings	394,5028	5,54218	382,5297	406,4760
De ambos	416,8486	6,51260	402,7790	430,9182

Esta tabla muestra las medias marginales estimadas por el modelo y los errores típicos de *Amount spent* en los niveles de factor de *Use coupons*. Esta tabla es útil para explorar las diferencias entre los niveles de este factor. En este ejemplo, un cliente que no utiliza cupones se espera que se gaste unos 319,65 dólares, mientras que aquellos que sí usan cupones se espera que gasten considerablemente más.

Figura 19-13
Resultados de las pruebas individuales para medias marginales estimadas de estilo de compra

Contrastes simples Utiliza los cupones[a]	Diferencia (estimación - hipotetizado)	Valor hipotetizado	Error típico	gl1	gl2	F de Wald	Sig.
Nivel Del periódico frente a nivel No	67,101	,000	6,537	1,000	13,000	105,350	,000
Nivel De los mailings frente a nivel No	74,857	,000	5,875	1,000	13,000	162,330	,000
Nivel De ambos frente a nivel No	97,203	,000	5,603	1,000	13,000	300,920	,000

a. Categoría de referencia = No

La tabla de pruebas individuales muestra tres contrastes simples, en los que se comparan los gastos de los clientes que no usan cupones frente a los que sí los usan.

Como los valores de significación de las pruebas son menores que 0,05, se puede concluir que los clientes que usan cupones tienden a gastar más que los que no usan cupones.

Figura 19-14
Resultados de las pruebas globales para medias marginales estimadas de estilo de compra

gl1	gl2	F de Wald	Sig.
3,000	11,000	87,453	,000

La tabla de pruebas globales informa de los resultados de una prueba de todos los contrastes de la tabla de pruebas individuales. Su valor de significación menor que 0,05 confirma que existe una diferencia en el gasto entre los niveles de *Use coupons*. Observe que las pruebas globales para *Use coupons* y *Who shopping for* son equivalentes a las pruebas de los efectos del modelo ya que los valores de contraste hipotetizados son iguales a 0.

Figura 19-15
Medias marginales estimadas por niveles de sexo por estilo de compra

Para quién compra	Utiliza los cupones	Media	Error típico	Intervalo de confianza al 95%	
				Inferior	Superior
Él mismo	No	244,3471	6,00949	231,3644	257,3298
	Del periódico	324,9708	5,94134	312,1353	337,8063
	De los mailings	321,3207	4,11028	312,4410	330,2005
	De ambos	343,4916	6,57845	329,2797	357,7034
Él mismo y esposa	No	337,1783	7,12181	321,7925	352,5640
	Del periódico	380,0468	7,91038	362,9574	397,1361
	De los mailings	375,3141	6,22468	361,8665	388,7617
	De ambos	388,8054	7,12101	373,4214	404,1894
Él mismo y familia	No	377,4111	11,58215	352,3894	402,4328
	Del periódico	455,2232	6,14420	441,9494	468,4969
	De los mailings	486,8736	10,76529	463,6166	510,1306
	De ambos	518,2488	11,73120	492,9050	543,5925

Esta tabla muestra las medias marginales estimadas por el modelo, los errores típicos y los intervalos de confianza de *Amount spent* en las combinaciones de factores de *Who shopping for* y *Use coupons*. Esta tabla es útil para explorar el efecto de la interacción entre estos dos factores detectada en las pruebas de los efectos del modelo.

Resumen

En este ejemplo, las medias marginales estimadas han revelado diferencias en el gasto entre clientes a distintos niveles de *Who shopping for* y *Use coupons*. Las pruebas de los efectos del modelo confirmaron la existencia de dicha diferencia, así como el hecho de que parece ser producto de un efecto de la interacción *Who shopping for*Use coupons*. La tabla de resumen del modelo reveló que el modelo actual explica algo más de la mitad de la variación hallada en los datos, y se podría mejorar añadiendo más predictores.

Procedimientos relacionados

El procedimiento Modelo lineal general de muestras complejas es una herramienta útil para crear modelos de una variable de escala cuando los casos se han extraído siguiendo un esquema de muestreo complejo.

- El Asistente de muestreo de la opción Muestras complejas se utiliza para definir las especificaciones de diseño de las muestras complejas y obtener una muestra. El archivo del plan de muestreo creado por el Asistente de muestreo contiene un plan de análisis por defecto que se puede especificar en el cuadro de diálogo Plan cuando se analiza la muestra obtenida de acuerdo con dicho plan.
- El Asistente de preparación del análisis de la opción Muestras complejas se utiliza para configurar las especificaciones de análisis para una muestra compleja existente. El archivo del plan de muestreo creado por el Asistente de muestreo se puede especificar en el cuadro de diálogo Plan cuando se analiza la muestra correspondiente a dicho plan.
- El procedimiento Regresión logística de muestras complejas permite crear un modelo de una respuesta categórica.
- El procedimiento Regresión ordinal de muestras complejas permite crear un modelo de una respuesta ordinal.

Capítulo
20

Regresión logística de muestras complejas

El procedimiento Regresión logística de muestras complejas lleva a cabo análisis de regresión logística sobre una variable binaria o una variable dependiente multinomial para muestras extraídas mediante métodos de muestreo complejo. Si lo desea, puede solicitar análisis de una subpoblación.

Uso del procedimiento Regresión logística de muestras complejas para evaluar riesgos de crédito

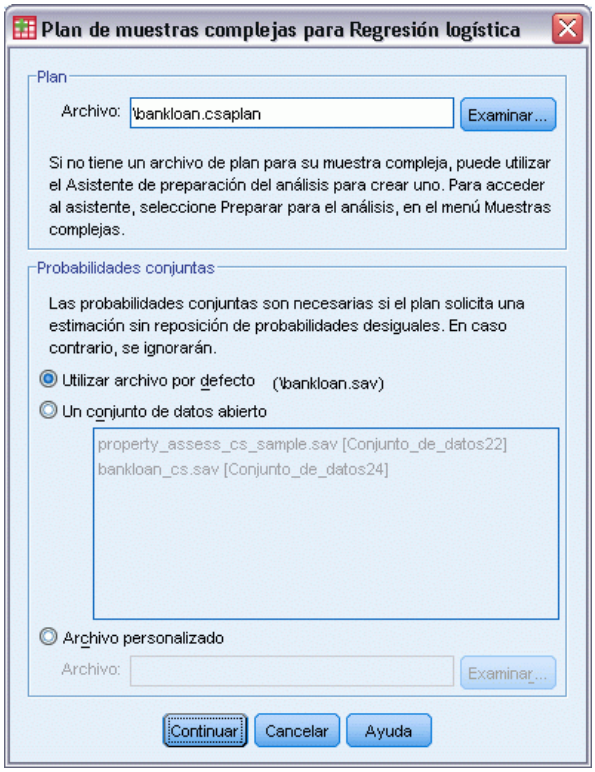
Si es el encargado de préstamos en un banco, deseará poder identificar características que sean indicativas de personas que puedan causar mora en los créditos y utilizar dichas características para identificar riesgos de crédito positivos y negativos.

Suponga que un encargado de préstamos ha recopilado registros antiguos de préstamos concedidos a clientes en diversas ramas, de acuerdo con un diseño complejo. Esta información se recoge en *bankloan_cs.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*. El encargado desea comprobar si la probabilidad con que las moras de un cliente se asocian a su edad, historial de empleo y cantidad de crédito adeudado; posteriormente, incorporará el diseño muestral.

Ejecución del análisis

▶ Para crear un modelo de regresión logística, elija en los menús:
Analizar > Complex Samples > Regresión logística...

Figura 20-1
Cuadro de diálogo Plan de muestras complejas

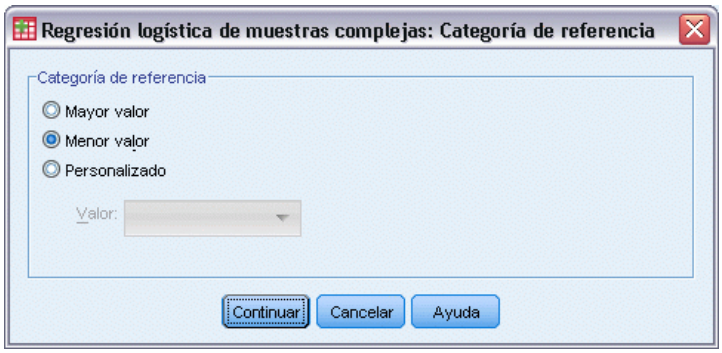


- Acceda al archivo *bankloan.csaplan* y selecciónelo. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*.
- Pulse en Continuar.

Figura 20-2
Cuadro de diálogo Regresión logística

- Seleccione *Previously defaulted* como la variable dependiente.
- Seleccione *Level of education* como un factor.
- Seleccione *Age in years* y *Other debt in thousands* como covariables.
- Seleccione *Previously defaulted* y pulse en Reference Category.

Figura 20-3
Cuadro de diálogo Regresión logística de muestras complejas: Categoría de referencia



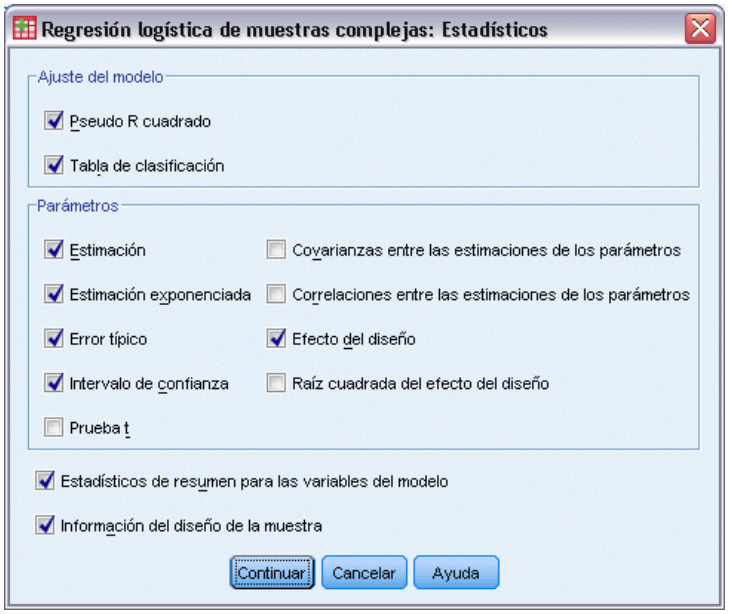
- Seleccione Lowest value como la categoría de referencia.

  Esto definirá la categoría "did not default" como la categoría de referencia; por lo tanto, las razones de las ventajas que aparecen en el resultado tendrán la propiedad de que cuanto mayores sean las razones de las ventajas mayor será la probabilidad de mora.

- Pulse en Continuar.

- En el cuadro de diálogo Regresión logística, pulse en Estadísticos.

Figura 20-4
Cuadro de diálogo Regresión logística: Estadísticos



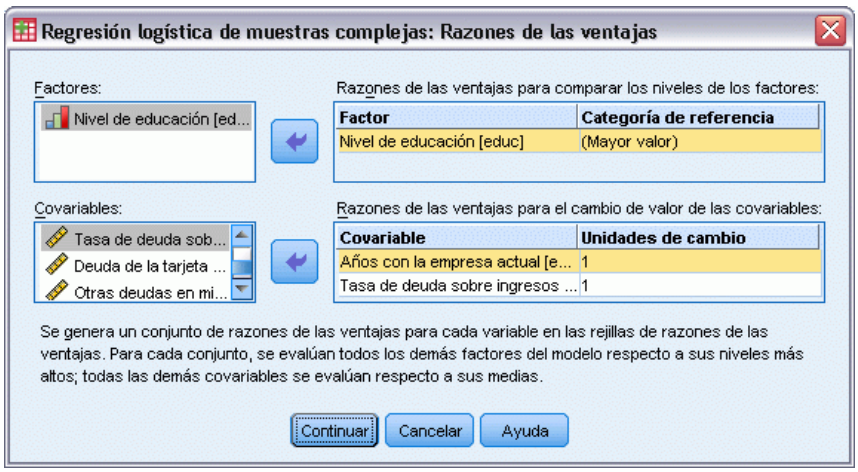
- Seleccione Tabla de clasificación en el grupo Ajuste del modelo.

- Seleccione Estimación, Estimación exponenciada, Error típico, Intervalo de confianza y Efecto del diseño en el grupo Parámetros.

- Pulse en Continuar.

- En el cuadro de diálogo Regresión logística, pulse en Razones de las ventajas.

Figura 20-5
Cuadro de diálogo Regresión logística de muestras complejas: Razones de las ventajas



- Seleccione para crear las razones de las ventajas para el factor *ed* y las covariables *employ* y *debtinc*.
- Pulse en Continuar.
- En el cuadro de diálogo Regresión logística, pulse en Aceptar.

Pseudo R cuadrado

Figura 20-6
Estadísticos pseudo R cuadrado

Cox y Snell	,330
Nagelkerke	,451
McFadden	,304

Variable dependiente: Impagos anteriores (categoría de referencia = No)
Modelo: (Intersección), educ, edad, empleo, direccion, ingresos, deudacred, deudaotro

En el modelo de regresión lineal, el coeficiente de determinación, R^2 resume la proporción de la varianza de la variable dependiente asociada con las variables predictoras (independientes), con valores R^2 mayores, indicando que el aumento de la variación se explica por el modelo hasta un máximo de 1. Para los modelos de regresión con una variable dependiente categórica, no es posible calcular un único estadístico R^2 que tenga todas las características de R^2 en el modelo de regresión lineal, por lo que en su lugar, se calculan estas aproximaciones. Los siguientes métodos se utilizan para realizar una estimación del coeficiente de determinación.

- R^2(Cox y Snell, 1989) de Cox y Snell está basado en el logaritmo de verosimilitud del modelo comparado con el logaritmo de verosimilitud de un modelo de línea base. Sin embargo tiene un valor máximo teórico menor que 1 con resultados categóricos, incluso para un modelo "perfecto".

- R^2(Nagelkerke, 1991) de Nagelkerke es una versión ajustada de *R*-cuadrado de Cox y Snell que ajusta la escala del estadístico para cubrir todo el rango de 0 a 1.
- R^2(McFadden, 1974) de McFadden es otra versión basada en los kernel del logaritmo de verosimilitud para el modelo de sólo intersección y el modelo estimado completo.

Los factores que constituyen un "buen " valor de R^2 varían entre las distintas áreas de aplicación. Mientras que estos estadísticos pueden ser indicativos por sí solos, son más útiles para comparar modelos que compiten con los mismos datos. El modelo con el mayor R^2 es "el mejor" según esta medida.

Classification

Figura 20-7
Tabla de clasificación

	Pronosticado		
Observado	No	Sí	Porcentaje correcto
No	188289,667	31871,267	85,5%
Sí	49970,600	77675,133	60,9%
Porcentaje global	68,5%	31,5%	76,5%

Variable dependiente: Impagos anteriores (categoría de referencia = No)
Modelo: (Intersección), educ, edad, empleo, direccion, ingresos, deudacred, deudaotro, pesofinal

La tabla de clasificación muestra los resultados prácticos de la utilización del modelo de regresión logística. Para cada caso, la respuesta pronosticada es *Sí* si el valor del logit pronosticado por el modelo de dicho caso es mayor que 0. Los casos se ponderan mediante *finalweight*, de manera que la tabla de clasificación informa del rendimiento esperado del modelo en la población.

- Las casillas de la diagonal son los pronósticos correctos.
- Las casillas fuera de la diagonal son los pronósticos incorrectos.

Según los casos utilizados para crear el modelo, se puede esperar, mediante la utilización de este modelo, clasificar correctamente el 85,5% de las personas que no causan mora en la población. De igual manera, se puede esperar clasificar correctamente el 60,9% de las personas que puedan causar mora. En general, se puede esperar que la clasificación del 76,5% de los casos se realice correctamente; sin embargo, debido a que esta tabla se creó con los casos utilizados para crear el modelo, es bastante probable que estas estimaciones sean excesivamente optimistas.

Pruebas de efectos del modelo

Figura 20-8
Pruebas de los efectos inter-sujetos

Origen	gl1	gl2	F de Wald	Sig.
(Modelo corregido)	11,000	4,000	14669,000	,010
(Intersección)	1,000	14,000	5,777	,031
educ	4,000	11,000	1,683	,224
edad	1,000	14,000	5,352	,036
empleo	1,000	14,000	88,244	,000
direccion	1,000	14,000	1,123	,307
ingresos	1,000	14,000	,007	,932
deudacred	1,000	14,000	27,632	,000
deudaotro	1,000	14,000	33,402	,000
pesofinal	1,000	14,000	,709	,414

Variable dependiente: Impagos anteriores (categoría de referencia = No)
Modelo: (Intersección), educ, edad, empleo, direccion, ingresos, deudacred, deudaotro, pesofinal

Cada término del modelo, además del propio modelo, se prueba para comprobar si su efecto es igual a 0. Los términos con valores de significación inferiores a 0,05 tienen algún efecto perceptible. Por consiguiente, *age*, *employ*, *debtinc* y *creddebt* contribuyen al modelo, mientras que los demás efectos principales no. En un análisis más detallado de los datos, es probable que se pudiera quitar *ed*, *address*, *income* y *othdebt* de la consideración del modelo.

Estimaciones de los parámetros

Figura 20-9
Estimaciones de los parámetros

Impagos anteriores	Parámetro	B	Error típico	Intervalo de confianza al 95%		Efecto del diseño	Exp(B)	Intervalo de confianza al 95% Exp(B)	
				Inferior	Superior			Inferior	Superior
Yes	(Intersección)	-1,140	,399	-1,995	-,284	,665	,320	,136	,753
	[educ=1]	,720	,340	-,010	1,449	,862	2,054	,990	4,259
	[educ=2]	,684	,371	-,112	1,481	1,247	1,983	,894	4,397
	[educ=3]	,518	,307	-,140	1,177	,813	1,679	,869	3,244
	[educ=4]	,789	,302	,142	1,437	,817	2,202	1,152	4,208
	[educ=5]	,000[a]	.	.	.	.	1,000	.	.
	edad	-,023	,010	-,043	-,002	,418	,978	,958	,998
	empleo	-,225	,024	-,277	-,174	1,200	,798	,758	,840
	direccion	-,028	,026	-,085	,029	,651	,972	,919	1,029
	ingresos	,000	,003	-,007	,006	1,410	1,000	,993	1,006
	deudacred	,095	,018	,056	,134	1,222	1,100	1,058	1,143
	deudaotro	,493	,085	,310	,676	1,373	1,637	1,363	1,966
	pesofinal	,026	,031	-,041	,094	1,219	1,027	,960	1,098

Variable dependiente: Impagos anteriores (categoría de referencia = No)
Modelo: (Intersección), educ, edad, empleo, direccion, ingresos, deudacred, deudaotro

a. Establecido en cero porque este parámetro es redundante.

La tabla de estimaciones de los parámetros resume el efecto de cada predictor. Observe que los valores de los parámetros afectan a la verosimilitud de la categoría “did default” relacionada con la categoría “did not default”. Por consiguiente, los parámetros con coeficientes positivos

aumentan la verosimilitud de la mora, mientras que los parámetros con coeficientes negativos disminuyen la verosimilitud de la mora.

El significado de un coeficiente de una regresión logística es más complejo que el de un coeficiente de una regresión lineal. Mientras que *B* es adecuado para probar los efectos del modelo, *Exp(B)* es más fácil de interpretar. *Exp(B)* representa el cambio en las razones de las ventajas del evento de interés atribuible a un aumento de una unidad en el predictor, para predictores que no formen parte de términos de interacción. Por ejemplo, *Exp(B)* para *employ* es igual a 0,798, lo que significa que las ventajas de la mora para personas cuya antigüedad en la empresa actual sea de dos años son 0,798 veces las ventajas de la mora de aquellas personas cuya antigüedad en la empresa actual sea de un año, siendo todo lo demás exactamente igual.

Los efectos del diseño indican que algunos de los errores típicos calculados para estas estimaciones de parámetros son mayores que los que se obtendrían si se supone que dichas observaciones proceden de una muestra aleatoria simple, mientras que los demás son más pequeños. Es de vital importancia incorporar la información sobre el diseño muestral al análisis porque, en caso contrario, se podría inferir, por ejemplo, que el coeficiente edad no es distinto de 0.

Razones de las ventajas

Figura 20-10
Razones de las ventajas para el nivel educativo

		Impagos anteriores	Razón de ventajas	Intervalo de confianza al 95%	
				Inferior	Superior
Nivel de educación	No completó el bachillerato vs. Título de Post-grado	Sí	2,054	,990	4,259
	Título de Bachiller vs.	Sí	1,983	,894	4,397
	Superiores iniciados vs.	Sí	1,679	,869	3,244
	Título Superior vs. Título	Sí	2,202	1,152	4,208

Variable dependiente: Impagos anteriores (categoría de referencia = No)
Modelo: (Intersección), educ, edad, empleo, direccion, ingresos, deudacred, deudaotro, inclprob_s2, pesofinal

a. Los factores y las covariables utilizadas en el cálculo se fijan en los siguientes valores: Nivel de educación=Título de Post-grado; Edad en años=34,19; Años con la empresa actual=6,99; Años en la dirección actual=6,32; Ingresos familiares en miles=60,1581; Deuda de la tarjeta de crédito en miles=1,9764; Otras deudas en miles=3,9164;

Esta tabla muestra las razones de las ventajas de *Previously defaulted* en los niveles de factor de *Level of education*. Los valores indicados son las razones de las ventajas de mora para *Did not complete high school* hasta *College degree*, comparadas a las razones de las ventajas para *Post-undergraduate degree*. Por consiguiente, la razón de las ventajas de 2,054 en la primera fila de la tabla significa que las ventajas de mora de una persona que no tiene estudios secundarios son 2,054 veces las ventajas de mora de una persona con una titulación de postgraduado.

Figura 20-11
Razones de las ventajas para años con la empresa actual

Unidades de cambio		Impagos anteriores	Razón de ventajas	Intervalo de confianza al 95%	
				Inferior	Superior
Años con la empresa actual	1,000	Sí	,798	,758	,840

Variable dependiente: Impagos anteriores (categoría de referencia = No)
Modelo: (Intersección), educ, edad, empleo, direccion, ingresos, deudacred, deudaotro, inclprob_s2, pesofinal

a. Los factores y las covariables utilizadas en el cálculo se fijan en los siguientes valores: Nivel de educación=Título de Post-grado; Edad en años=34,19; Años con la empresa actual=6,99; Años en la dirección actual=6,32; Ingresos familiares en miles=60,1581; Deuda de la tarjeta de crédito en miles=1,9764; Otras deudas en miles=3,9164;

Esta tabla muestra la razón de las ventajas de *Previously defaulted* para un cambio de unidad en la covariable *Years with current employer*. El valor indicado es la razón de las ventajas de mora de una persona con 7,99 años en la empresa actual comparada con las ventajas de mora de una persona con 6,99 años (la media).

Figura 20-12
Razones de las ventajas para la razón entre el endeudamiento y los ingresos

Unidades de cambio		Impagos anteriores	Razón de ventajas	Intervalo de confianza al 95%	
				Inferior	Superior
Deuda de la tarjeta de crédito en miles	1,000	Sí	1,100	1,058	1,143

Variable dependiente: Impagos anteriores (categoría de referencia = No)
Modelo: (Intersección), educ, edad, empleo, direccion, ingresos, deudacred, deudaotro,

a. Los factores y las covariables utilizadas en el cálculo se fijan en los siguientes valores: Nivel de educación=Título de Post-grado;Edad en años=34,19; Años con la empresa actual=6,99; Años en la dirección actual=6,32; Ingresos familiares en miles=60,1581; Deuda de la tarjeta de crédito en miles=1,9764; Otras deudas en miles=3,9164

Esta tabla muestra la razón de las ventajas de *Previously defaulted* para un cambio de unidad en la covariable *Debt to income ratio*. El valor indicado es la razón de las ventajas de mora de una persona con una razón de endeudamiento/ingresos de 10,9341 comparada con las ventajas de mora de una persona con una razón de endeudamiento/ingresos de 9,9341 (la media).

Observe que debido a que ninguno de estos predictores forman parte de los términos de interacción, los valores de las razones de las ventajas indicados en estas tablas son iguales a los valores de las estimaciones exponenciadas de los parámetros. Cuando un predictor forma parte de un término de interacción, su razón de las ventajas en estas tablas también dependerá de los valores de los demás predictores que componen la interacción.

Resumen

Mediante el procedimiento de Regresión logística de muestras complejas, se ha construido un modelo para pronosticar la probabilidad de que un cliente dado cause mora en un crédito.

Un problema crítico para los encargados de los créditos es el coste de los errores de Tipo I y Tipo II. Es decir, ¿cuál es el coste de clasificar una persona susceptible de causar mora como una persona que no va a causar mora (Tipo I)? ¿Cuál es el coste de clasificar una persona que no va a causar mora como una persona susceptible de causar mora (Tipo II)? Si la principal preocupación es la concesión de mal crédito, entonces será deseable reducir el error de Tipo I y maximizar la

sensitivity. Si la prioridad es aumentar la base de clientes, entonces será deseable reducir el error de Tipo II y maximizar la **specificity**. Normalmente, ambas son cuestiones importantes, así que se deberá elegir una regla de decisión para clasificar los clientes que ofrezcan la mejor combinación de susceptibilidad y especificidad.

Procedimientos relacionados

El procedimiento Regresión logística de muestras complejas es una herramienta útil para crear modelos de una variable categórica cuando los casos se han extraído siguiendo un esquema de muestreo complejo.

- El Asistente de muestreo de la opción Muestras complejas se utiliza para definir las especificaciones de diseño de las muestras complejas y obtener una muestra. El archivo del plan de muestreo creado por el Asistente de muestreo contiene un plan de análisis por defecto que se puede especificar en el cuadro de diálogo Plan cuando se analiza la muestra obtenida de acuerdo con dicho plan.
- El Asistente de preparación del análisis de la opción Muestras complejas se utiliza para configurar las especificaciones de análisis para una muestra compleja existente. El archivo del plan de muestreo creado por el Asistente de muestreo se puede especificar en el cuadro de diálogo Plan cuando se analiza la muestra correspondiente a dicho plan.
- El procedimiento Modelo lineal general de muestras complejas permite crear un modelo de una respuesta de escala.
- El procedimiento Regresión ordinal de muestras complejas permite crear un modelo de una respuesta ordinal.

Capítulo
21

Regresión ordinal de muestras complejas

El procedimiento Regresión ordinal de muestras complejas crea un modelo predictivo de una variable dependiente ordinal para muestras extraídas mediante métodos de muestreo complejo. Si lo desea, puede solicitar análisis de una subpoblación.

Uso de la regresión ordinal de muestras complejas para analizar los resultados de encuestas

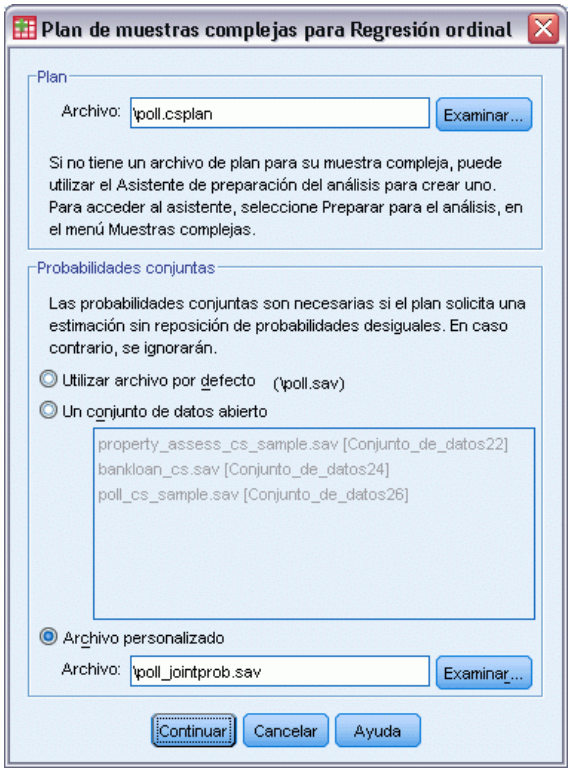
Los diputados que estudian un proyecto de ley antes de una asamblea legislativa se interesan por conocer si la opinión pública apoya dicho proyecto de ley y qué relación guarda dicho apoyo con los datos demográficos de los votantes. Los encuestadores diseñan entrevistas y las realizan siguiendo un diseño muestral complejo.

Los resultados de las encuestas se recopilan en *poll_cs_sample.sav*. El plan de muestreo utilizado por los encuestadores se incluye en *poll.csplan*. Como utiliza un método de probabilidad proporcional al tamaño (PPS), también hay un archivo que contiene las probabilidades de selección conjunta (*poll_jointprob.sav*). Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*. Utilice la regresión ordinal de muestras complejas para ajustar un modelo acerca del nivel de apoyo a la ley de acuerdo con los datos demográficos de los votantes.

Ejecución del análisis

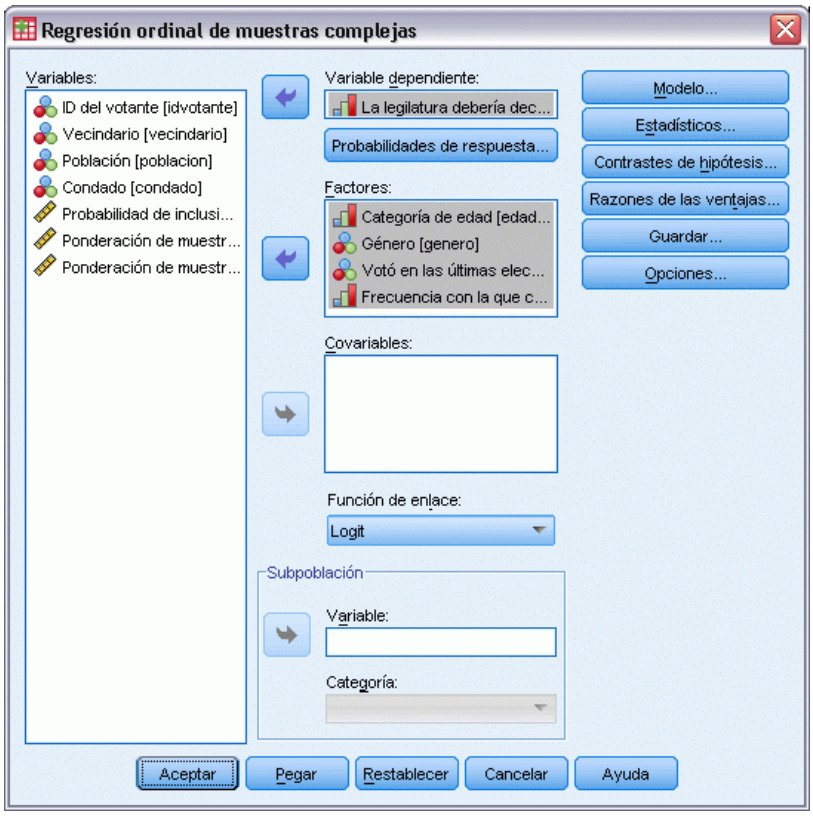
- Para ejecutar un análisis de Regresión ordinal de muestras complejas, seleccione en los menús:
  Analizar > Complex Samples > Regresión ordinal...

Figura 21-1
Cuadro de diálogo Plan de muestras complejas



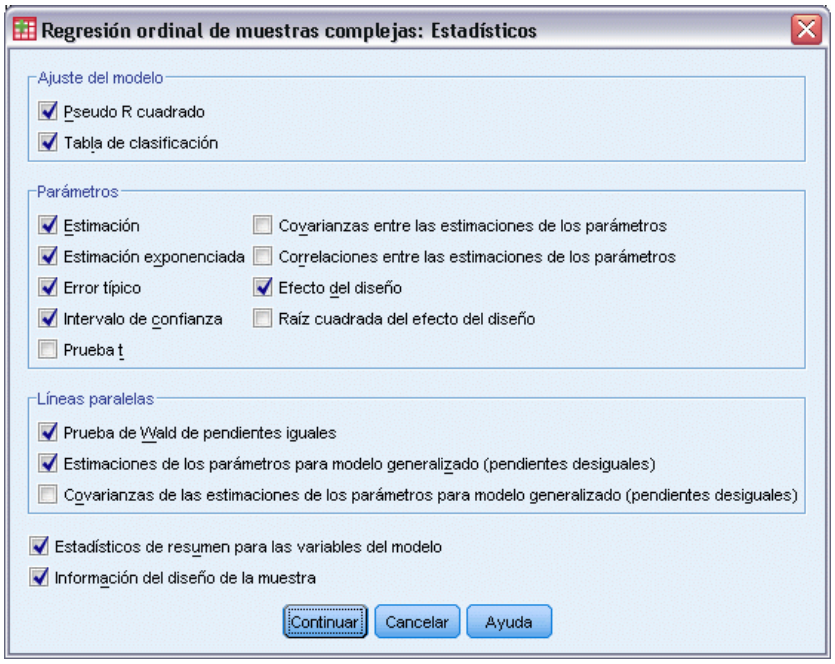
- Acceda al archivo *poll.csplan* y selecciónelo como el archivo del plan. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*.
- Seleccione *poll_jointprob.sav* como archivo de las probabilidades conjuntas.
- Pulse en Continuar.

Figura 21-2
Cuadro de diálogo Regresión ordinal



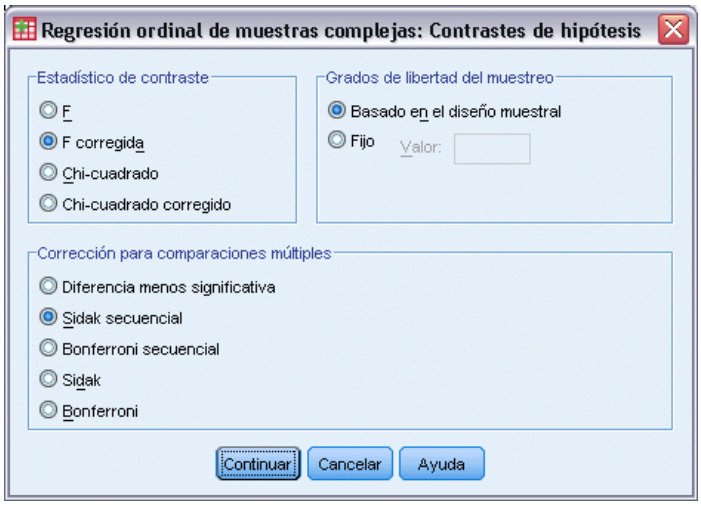
- Seleccione *La legsilatura debería decretar un impuesto sobre la gasolina* como la variable dependiente.
- Seleccione desde *Categoría de edad* hasta *Frecuencia con la que conduce* como factores.
- Pulse en Estadísticos.

Figura 21-3
Cuadro de diálogo Regresión ordinal de muestras complejas: Estadísticos



- Seleccione Tabla de clasificación en el grupo Ajuste del modelo.
- Seleccione Estimación, Estimación exponenciada, Error típico, Intervalo de confianza y Efecto del diseño en el grupo Parámetros.
- Seleccione Prueba de Wald de pendientes iguales y Estimaciones de los parámetros para modelo generalizado (pendientes desiguales).
- Pulse en Continuar.
- Pulse en Contrastes de hipótesis en el cuadro de diálogo Regresión ordinal de muestras complejas.

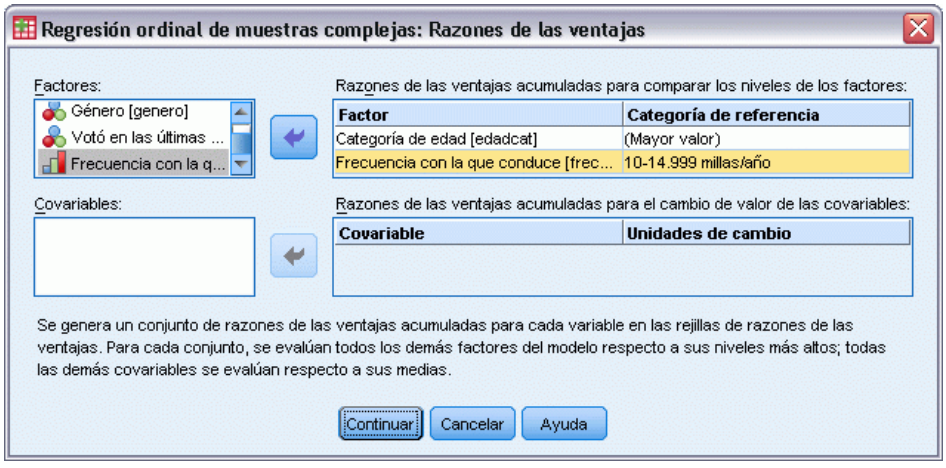
Figura 21-4
Cuadro de diálogo Contrastes de hipótesis



Incluso para un número moderado de predictores y categorías de respuesta, el estadístico de contraste de la *F* de Wald es posible que no se pueda estimar para la prueba de líneas paralelas.

- Seleccione F corregida en el grupo Estadístico de contraste.
- Seleccione Sidak secuencial como método de ajuste para comparaciones múltiples.
- Pulse en Continuar.
- Pulse en Razones de las ventajas en el cuadro de diálogo Regresión ordinal de muestras complejas.

Figura 21-5
Cuadro de diálogo Regresión ordinal de muestras complejas: Razones de las ventajas



- Seleccione generar razones de las ventajas acumulativas para *Categoría de edad* y *Frecuencia con la que conduce*.
- Seleccione 10-14.999 millas/año, un kilometraje anual más "habitual" que el máximo, como categoría de referencia de *Frecuencia con la que conduce*.

- ▶ Pulse en Continuar.
- ▶ Pulse en Aceptar en el cuadro de diálogo Regresión ordinal de muestras complejas.

Pseudo R cuadrado

Figura 21-6
Pseudo R cuadrado

Cox y Snell	,179
Nagelkerke	,191
McFadden	,071

Variable dependiente: La legilatura debería decretar un impuesto sobre la gasolina (Ascendente)
Modelo: (Umbral), edadcat, genero, votoult, freccond
Función de vínculo: Logit

En el modelo de regresión lineal, el coeficiente de determinación, R^2 resume la proporción de la varianza de la variable dependiente asociada con las variables predictoras (independientes), con valores R^2 mayores, indicando que el aumento de la variación se explica por el modelo hasta un máximo de 1. Para los modelos de regresión con una variable dependiente categórica, no es posible calcular un único estadístico R^2 que tenga todas las características de R^2 en el modelo de regresión lineal, por lo que en su lugar, se calculan estas aproximaciones. Los siguientes métodos se utilizan para realizar una estimación del coeficiente de determinación.

- R^2(Cox y Snell, 1989) de Cox y Snell está basado en el logaritmo de verosimilitud del modelo comparado con el logaritmo de verosimilitud de un modelo de línea base. Sin embargo tiene un valor máximo teórico menor que 1 con resultados categóricos, incluso para un modelo "perfecto".
- R^2(Nagelkerke, 1991) de Nagelkerke es una versión ajustada de *R*-cuadrado de Cox y Snell que ajusta la escala del estadístico para cubrir todo el rango de 0 a 1.
- R^2(McFadden, 1974) de McFadden es otra versión basada en los kernel del logaritmo de verosimilitud para el modelo de sólo intersección y el modelo estimado completo.

Los factores que constituyen un "buen " valor de R^2 varían entre las distintas áreas de aplicación. Mientras que estos estadísticos pueden ser indicativos por sí solos, son más útiles para comparar modelos que compiten con los mismos datos. El modelo con el mayor R^2 es "el mejor" según esta medida.

Pruebas de efectos del modelo

Figura 21-7
Contrastes de los efectos del modelo

Origen	gl1	gl2	F de Wald corregida	Sig.	Sig. de Sidak secuencial
edadcat	2,283	31,966	6,215	,004	,003
genero	1,000	14,000	,046	,834	,834
votoult	1,000	14,000	,076	,787	,787
freccond	3,785	52,987	228,015	,000	,000

Variable dependiente: La legilatura debería decretar un impuesto sobre la gasolina (Ascendente)
Modelo: (Umbral), edadcat, genero, votoult, freccond
Función de vínculo: Logit

Cada término del modelo se prueba para comprobar si su efecto es igual a 0. Los términos con valores de significación inferiores a 0,05 tienen algún efecto perceptible. Por consiguiente, *edadcat* y *freccond* contribuyen al modelo, mientras que los demás efectos principales no. En los análisis posteriores de los datos, puede considerar quitar *genero* y *votoult* del modelo.

Estimaciones de los parámetros

La tabla de estimaciones de los parámetros resume el efecto de cada predictor. Mientras que es difícil interpretar los coeficientes de este modelo debido a la naturaleza de la función de enlace, los signos de los coeficientes para las covariables y los valores relativos de los coeficientes para los niveles de factores pueden proporcionar información relevante de los predictores del modelo.

- Para las covariables, los coeficientes positivos (negativos) indican relaciones positivas (negativas) entre predictores y resultados. Un valor mayor de una covariable con un coeficiente positivo corresponde a una mayor probabilidad de estar en una de las categorías de resultados acumulados "superiores".
- Para los factores, un nivel de factor con un mayor coeficiente indica una mayor probabilidad de ser una de las categorías de resultados acumulados "superiores". El signo de un coeficiente para un nivel de factor depende del efecto del nivel de factor relativo a la categoría de referencia.

Figura 21-8
Estimaciones de los parámetros

Parámetro		B	Error típico	Intervalo de confianza al 95%		Efecto del diseño	Exp(B)	95% Intervalo de confianza para Exp(B)	
				Inferior	Superior			Inferior	Superior
Umbral	[opinion_impgas=1]	-3,343	,104	-3,566	-3,120	1,132	,035	,028	,044
	[opinion_impgas=2]	-1,910	,098	-2,120	-1,700	1,058	,148	,120	,183
	[opinion_impgas=3]	-,674	,090	-,866	-,482	,915	,510	,421	,618
Regresión	[edadcat=1]	-,324	,079	-,494	-,154	1,793	,723	,610	,858
	[edadcat=2]	-,138	,054	-,255	-,022	1,158	,871	,775	,978
	[edadcat=3]	-,095	,076	-,257	,068	2,206	,909	,773	1,070
	[edadcat=4]	,000[a]	.	.	.	.	1,000	.	.
	[genero=0]	-,008	,035	-,084	,068	,949	,992	,920	1,071
	[genero=1]	,000[a]	.	.	.	.	1,000	.	.
	[votoult=0]	-,011	,039	-,095	,073	1,103	,989	,909	1,076
	[votoult=1]	,000[a]	.	.	.	.	1,000	.	.
	[freccond=1]	-3,751	,153	-4,079	-3,423	1,117	,023	,017	,033
	[freccond=2]	-3,003	,116	-3,251	-2,755	1,226	,050	,039	,064
	[freccond=3]	-2,295	,114	-2,540	-2,050	1,585	,101	,079	,129
	[freccond=4]	-1,570	,092	-1,769	-1,372	1,078	,208	,171	,254
	[freccond=5]	-,812	,089	-1,003	-,621	,941	,444	,367	,537
	[freccond=6]	,000[a]	.	.	.	.	1,000	.	.

Variable dependiente: La legilatura debería decretar un impuesto sobre la gasolina (Ascendente)
Modelo: (Umbral), edadcat, genero, votoult, freccond
Función de vínculo: Logit

a. Establecido en cero porque este parámetro es redundante.

Puede realizar las siguientes interpretaciones a partir de las estimaciones de los parámetros:

- Las personas incluidas en las categorías de edad inferiores muestran un mayor apoyo al proyecto de ley que las que se encuentran en la categoría de edad superior.
- Las personas que conducen con menor frecuencia muestran un mayor apoyo al proyecto de ley que las que conducen con mayor frecuencia.
- Los coeficientes de las variables *genero* y *votoult*, además de no ser estadísticamente significativos, parecen ser pequeños en comparación con los otros coeficientes.

Los efectos del diseño indican que algunos de los errores típicos calculados para estas estimaciones de los parámetros son mayores que los que se obtendrían si se utilizara una muestra aleatoria simple, mientras que otros son más pequeños. Es de vital importancia incorporar la información sobre el diseño muestral al análisis porque, en caso contrario, se podría inferir, por ejemplo, que el coeficiente del tercer nivel de *Categoría de edad*, *[edadcat=3]*, es significativamente distinto de 0.

Classification

Figura 21-9
Información sobre la variable categórica

		Recuento ponderado	Porcentaje ponderado
La legilatura debería decretar un impuesto sobre la gasolina[a]	Muy de acuerdo	25132,955	21,3%
	De acuerdo	32261,425	27,3%
	En desacuerdo	29477,417	24,9%
	Muy en desacuerdo	31314,203	26,5%
Categoría de edad	18-30	20509,504	17,4%
	31-45	35380,506	29,9%
	46-60	34865,792	29,5%
	>60	27430,198	23,2%
Género	Hombre	61424,547	52,0%
	Mujer	56761,453	48,0%
Votó en las últimas elecciones	No	70607,216	59,7%
	Sí	47578,784	40,3%
Frecuencia con la que conduce	No tiene coche	3437,137	2,9%
	<10.000 millas/año	10816,349	9,2%
	10-14.999 millas/año	32539,364	27,5%
	15-19.999 millas/año	39179,814	33,2%
	20-29.999 millas/año	25617,804	21,7%
	>=30.000 millas/año	6595,532	5,6%
Tamaño de la población		118186,000	100,0%

a. Los valores de variables dependientes se ordenan en orden ascendente.

Según los datos observados, el modelo "nulo" (es decir, el que no incluye ningún predictor) clasificaría a todos los clientes en el grupo modal, *De acuerdo*. Por tanto, el modelo nulo sería correcto 27,3% de las veces.

Figura 21-10
Tabla de clasificación

	Pronosticado				
Observado	Muy de acuerdo	De acuerdo	En desacuerdo	Muy en desacuerdo	Porcentaje correcto
Muy de acuerdo	7067,567	12130,814	3875,825	2058,750	28,1%
De acuerdo	4271,234	14464,286	7320,767	6205,137	44,8%
En desacuerdo	2024,816	11703,368	7108,487	8640,746	24,1%
Muy en desacuerdo	889,869	8169,109	6946,522	15308,703	48,9%
Porcentaje global	12,1%	39,3%	21,4%	27,3%	37,2%

Variable dependiente: La legilatura debería decretar un impuesto sobre la gasolina (Ascendente)
Modelo: (Umbral), edadcat, genero, votoult, freccond
Función de vínculo: Logit

La tabla de clasificación muestra los resultados prácticos de la utilización del modelo. Para cada caso, la respuesta pronosticada es la categoría de respuesta con la mayor probabilidad pronosticada por el modelo. Los casos se ponderan mediante *Ponderaciones muestrales finales*, de manera que la tabla de clasificación informa del rendimiento del modelo esperado en la población.

- Las casillas de la diagonal son los pronósticos correctos.
- Las casillas fuera de la diagonal son los pronósticos incorrectos.

El modelo clasifica correctamente un 9,9% más, es decir, el 37,2% de los casos. En concreto, el modelo funciona considerablemente mejor al clasificar a las personas con *De acuerdo* o *Muy en desacuerdo* y ligeramente peor a las personas con *En desacuerdo*.

Razones de las ventajas

Las **ventajas acumuladas** se definen como la razón de la probabilidad de que la variable dependiente tome un valor menor o igual que una determinada categoría de respuesta respecto a la probabilidad de que tome un valor mayor que la categoría de respuesta. La **razón de las ventajas acumuladas** es la razón de las ventajas acumuladas para diferentes valores de los predictores y está estrechamente relacionada con las estimaciones exponenciadas de los parámetros. Curiosamente, la razón de las ventajas acumuladas propiamente no depende de la categoría de respuesta.

Figura 21-11
Razones de las ventajas acumuladas para Categoría de edad

		Razón de las ventajas acumulada	Intervalo de confianza al 95%		Efecto del diseño	Raíz cuadrada del efecto del diseño
			Inferior	Superior		
Categoría de edad	18-30 vs. >60	1,383	1,166	1,639	1,793	1,339
	31-45 vs. >60	1,148	1,022	1,290	1,158	1,076
	46-60 vs. >60	1,100	,935	1,294	2,206	1,485

Variable dependiente: La legilatura debería decretar un impuesto sobre la gasolina (Ascendente)
Modelo: (Umbral), edadcat, genero, votoult, freccond
Función de vínculo: Logit

a. Los factores y las covariables utilizadas en el cálculo se fijan en los siguientes valores: Categoría de edad=>60; Género=Mujer; Votó en las últimas elecciones=Sí; Frecuencia con la que conduce=>=30.000 millas/año

Esta tabla muestra las razones de las ventajas acumuladas para los niveles de factor de *Categoría de edad*. Los valores mostrados son las razones de las ventajas acumuladas para *18–30* hasta *46–60*, comparadas con las ventajas acumuladas para *>60*. Por tanto, la razón de las ventajas de 1,383 de la primera fila de la tabla indica que las ventajas acumuladas para una persona con una edad entre 18 y 30 años son 1,383 veces las ventajas acumuladas para una persona con más de 60 años. Tenga en cuenta que como *Categoría de edad* no figura en ningún término de interacción, las razones de las ventajas son meramente las razones de las estimaciones exponenciadas de los parámetros. Por ejemplo, la razón de las ventajas acumuladas para *18–30* respecto a *>60* es 1,00 / 0,723 = 1,383.

Figura 21-12
Razones de las ventajas para Frecuencia con la que conduce

		Razón de las ventajas acumulada	Intervalo de confianza al 95%		Efecto del diseño	Raíz cuadrada del efecto del diseño
			Inferior	Superior		
Frecuencia con la que conduce	No tiene coche vs. 10-14.999 millas/año	4,288	2,878	6,390	2,345	1,531
	<10.000 millas/año vs. 10-14.999 millas/año	2,030	1,656	2,488	1,838	1,356
	15-19.999 millas/año vs. 10-14.999 millas/año	,484	,430	,546	1,450	1,204
	20-29.999 millas/año vs. 10-14.999 millas/año	,227	,193	,267	2,095	1,448
	>=30.000 millas/año vs. 10-14.999 millas/año	,101	,079	,129	1,585	1,259

Variable dependiente: La legilatura debería decretar un impuesto sobre la gasolina (Ascendente)
Modelo: (Umbral), edadcat, genero, votoult, freccond
Función de vínculo: Logit

a. Los factores y las covariables utilizadas en el cálculo se fijan en los siguientes valores: Categoría de edad=>60; Género=Mujer; Votó en las últimas elecciones=Sí; Frecuencia con la que conduce=>=30.000 millas/año

Esta tabla muestra las razones de las ventajas acumuladas para los niveles de factor de *Frecuencia con la que conduce*, utilizando *10–14.999 millas/año* como categoría de referencia. Como *Frecuencia con la que conduce* no figura en ningún término de interacción, las razones de las ventajas son meramente las razones de las estimaciones exponenciadas de los parámetros. Por ejemplo, la razón de las ventajas acumuladas para *20–29.999 millas/año* respecto a *10–14.999 millas/año* es 0,101 / 0,444 = 0,227.

Modelo acumulado generalizado

Figura 21-13
Prueba de líneas paralelas

gl1	gl2	F de Wald corregida	Sig.	Sig. de Sidak secuencial
8,769	122,767	1,894	,061	,392

Variable dependiente: La legilatura debería decretar un impuesto sobre la gasolina (Ascendente)
Modelo: (Umbral), edadcat, genero, votoult, freccond
Función de vínculo: Logit

La prueba de líneas paralelas puede ayudarle a evaluar si el supuesto de que los parámetros son los mismos para todas las categorías de respuesta es razonable. Esta prueba compara el modelo estimado con el mismo conjunto de coeficientes para todas las categorías con un modelo generalizado con un conjunto diferente de coeficientes para cada categoría.

El contraste de la *F* de Wald es un contraste ómnibus de la matriz de contrastes para el supuesto de líneas paralelas que proporciona valores *p* asintóticamente correctos. Para muestras de tamaño pequeño a medio, el estadístico de la *F* de Wald corregida funciona bien. El valor de significación es cercano a 0,05, lo que sugiere que el modelo generalizado puede mejorar el ajuste del modelo. No obstante, el contraste corregido de Sidak secuencial indica un valor de significación suficientemente alto (0,392) por lo que, en general, no hay ninguna evidencia clara para rechazar el supuesto de líneas paralelas. El contraste de Sidak secuencial comienza con pruebas de Wald

de contrastes individuales que proporcionan un valor *p* global. Estos resultados deben ser comparables con el resultado del contraste ómnibus de Wald. El hecho de que sean tan diferentes en este ejemplo resulta un tanto sorprendente, pero puede deberse a la existencia de muchos contrastes en la prueba y un número relativamente pequeño de grados de libertad del diseño.

Figura 21-14
Estimaciones de los parámetros para el modelo acumulado generalizado (sólo se muestra una parte)

La legilatura debería decretar un impuesto sobre la gasolina	Parámetro	B	Error típico	Intervalo de confianza al 95%	
				Inferior	Superior
Muy de acuerdo	(Umbral)	-3,681	,221	-4,155	-3,207
	[edadcat=1]	-,320	,096	-,525	-,115
	[edadcat=2]	-,075	,071	-,227	,077
	[edadcat=3]	-,022	,073	-,180	,135
	[edadcat=4]	,000[a]	.		
	[genero=0]	-,082	,054	-,197	,033
	[genero=1]	,000[a]	.		
	[votoult=0]	,008	,052	-,104	,120
	[votoult=1]	,000[a]	.		
	[freccond=1]	-4,096	,267	-4,669	-3,523
	[freccond=2]	-3,367	,237	-3,876	-2,857
	[freccond=3]	-2,678	,224	-3,158	-2,199
	[freccond=4]	-1,928	,213	-2,384	-1,471
	[freccond=5]	-1,015	,252	-1,555	-,476
	[freccond=6]	,000[a]	.		
De acuerdo	(Umbral)	-1,963	,153	-2,291	-1,635
	[edadcat=1]	-,385	,095	-,587	-,182
	[edadcat=2]	-,130	,069	-,279	,018
	[edadcat=3]	-,139	,101	-,356	,077
	[edadcat=4]	,000[a]	.		
	[genero=0]	-,004	,040	-,090	,082
	[genero=1]	,000[a]	.		
	[votoult=0]	,009	,059	-,117	,135
	[votoult=1]	,000[a]	.		
	[freccond=1]	-3,867	,318	-4,549	-3,185
	[freccond=2]	-3,005	,175	-3,380	-2,630
	[freccond=3]	-2,290	,187	-2,691	-1,888
	[freccond=4]	-1,633	,166	-1,988	-1,278
	[freccond=5]	-,909	,137	-1,204	-,615
	[freccond=6]	,000[a]	.		

Además, los valores estimados de los coeficientes del modelo generalizado no parecen ser muy diferentes de las estimaciones obtenidas con el supuesto de líneas paralelas.

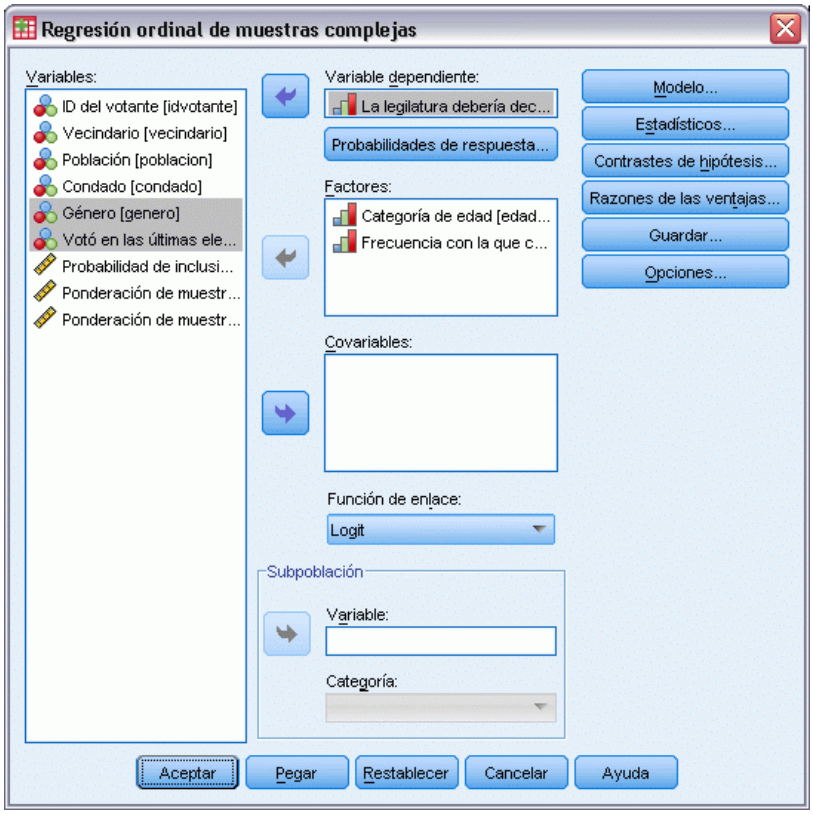
Exclusión de los predictores no significativos

Las pruebas de los efectos del modelo han mostrado que los coeficientes del modelo para *Genero* y *Votó en las últimas elecciones* no son estadísticamente distintos de 0.

- Para generar un modelo reducido, recupere el cuadro de diálogo Regresión ordinal de muestras complejas.

▶ Pulse en Continuar en el cuadro de diálogo Plan.

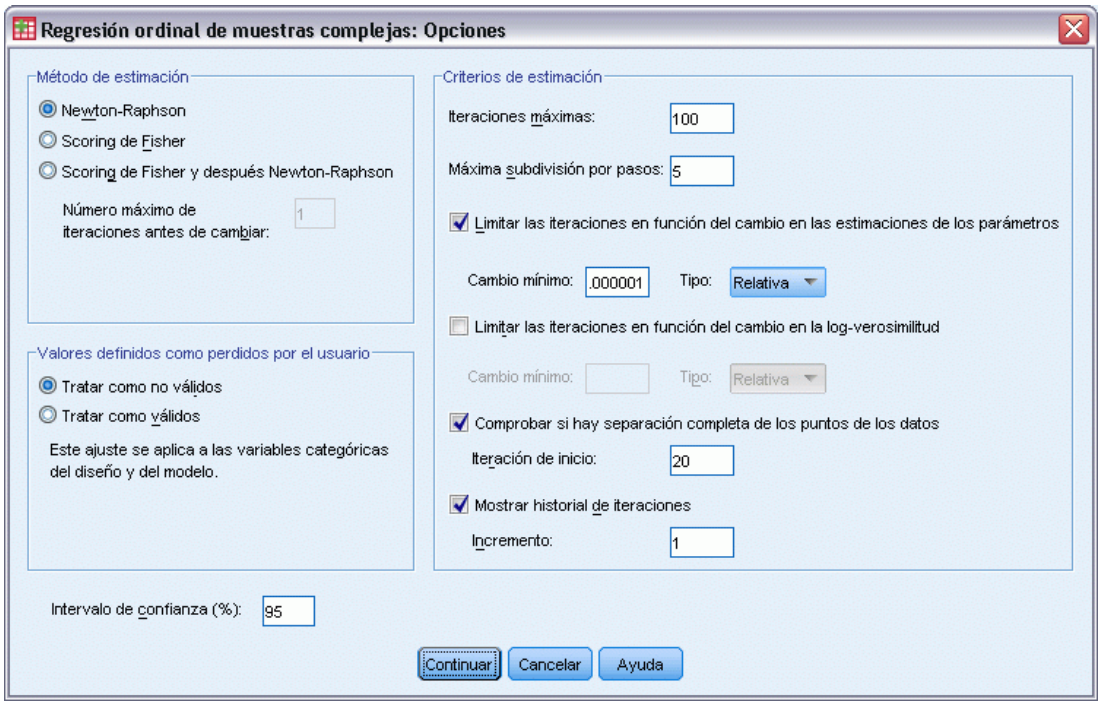
Figura 21-15
Cuadro de diálogo Regresión ordinal



▶ Anule la selección de *Genero* y *Votó en las últimas elecciones* como factores.

▶ Pulse en Opciones.

Figura 21-16
Cuadro de diálogo Regresión ordinal: Opciones



- Seleccione Mostrar historial de iteraciones.

  El historial de iteraciones es útil para diagnosticar los problemas que encuentra el algoritmo de estimación.

- Pulse en Continuar.
- Pulse en Aceptar en el cuadro de diálogo Regresión ordinal de muestras complejas.

Advertencias

Figura 21-17
Advertencias para el modelo reducido

El valor del log-verosimilitud no puede aumentar tras el número máximo de pasos en el método de subdivisión por pasos.

El procedimiento CSORDINAL continúa a pesar de la advertencia anterior. Los resultados posteriores que se muestran se basan en la última iteración. La validez del ajuste del modelo es incierta.

El siguiente mensaje se aplica al modelo acumulado generalizado.

El valor del log-verosimilitud no puede aumentar tras el número máximo de pasos en el método de subdivisión por pasos.

Las advertencias indican que la estimación del modelo reducido finalizó antes de que las estimaciones de los parámetros alcanzaran la convergencia, ya que la log-verosimilitud no pudo aumentarse con cada cambio (o “paso”) en los valores actuales de las estimaciones de los parámetros.

Figura 21-18
Advertencias para el modelo reducido

Número de iteraciones[b]	Subdivisión por N pasos	Pseudo -2 log de la verosimilitud	Umbral			Regresión							
			[opinion_impgas=1]	[opinion_impgas=2]	[opinion_impgas=3]	[edadcat=1]	[edadcat=2]	[edadcat=3]	[freccond=1]	[freccond=2]	[freccond=3]	[freccond=4]	[freccond=5]
0	0	326640,341	-1,309	-,058	1,020	,000	,000	,000	,000	,000	,000	,000	,000
1	0	303567,549	-3,242	-1,881	-,704	-,323	-,137	-,094	-3,84	-2,970	-2,25	-1,563	-,835
2	0	303336,336	-3,327	-1,897	-,664	-,325	-,139	-,095	-3,74	-2,998	-2,29	-1,568	-,811
3	0	303335,933	-3,333	-1,900	-,664	-,326	-,139	-,096	-3,75	-3,003	-2,29	-1,570	-,812
4	0	303335,933	-3,333	-1,900	-,664	-,326	-,139	-,096	-3,75	-3,003	-2,29	-1,570	-,812
5[a]	5	303335,933	-3,333	-1,900	-,664	-,326	-,139	-,096	-3,75	-3,003	-2,29	-1,570	-,812

No se muestran los parámetros redundantes. Sus valores son siempre cero en todas las iteraciones.
Variable dependiente: La legilatura debería decretar un impuesto sobre la gasolina (Ascendente)
Modelo: (Umbral), edadcat, freccond
Función de vínculo: Logit

a. El valor del log-verosimilitud no puede aumentar tras el número máximo de pasos en el método de subdivisión por pasos.

b. Se utilizó el método de Newton-Raphson para estimar los parámetros.

Mirando el historial de iteraciones, los cambios de las estimaciones de los parámetros en las últimas iteraciones son suficientemente pequeños como para no tener que preocuparse seriamente acerca del mensaje de advertencia.

Comparación de los modelos

Figura 21-19
Pseudo R cuadrado para el modelo reducido

Cox y Snell	,179
Nagelkerke	,191
McFadden	,071

Variable dependiente: La legilatura debería decretar un impuesto sobre la gasolina (Ascendente)
Modelo: (Umbral), edadcat, genero, votoult, freccond
Función de vínculo: Logit

Los valores de R^2 del modelo reducido son idénticos a los del modelo original. Esto constituye una evidencia a favor del modelo reducido.

Figura 21-20
Tabla de clasificación para el modelo reducido

Observado	Pronosticado				
	Muy de acuerdo	De acuerdo	En desacuerdo	Muy en desacuerdo	Porcentaje correcto
Muy de acuerdo	7067,567	12823,258	3183,380	2058,750	28,1%
De acuerdo	4271,234	15684,090	6100,963	6205,137	48,6%
En desacuerdo	2024,816	13157,809	5654,047	8640,746	19,2%
Muy en desacuerdo	889,869	9226,578	5889,053	15308,703	48,9%
Porcentaje global	12,1%	43,1%	17,6%	27,3%	37,0%

Variable dependiente: La legilatura debería decretar un impuesto sobre la gasolina (Ascendente)
Modelo: (Umbral), edadcat, freccond
Función de vínculo: Logit

La tabla de clasificación complica un tanto las cosas. La tasa de clasificación global de 37,0% para el modelo reducido es comparable a la del modelo original, lo que constituye una evidencia a favor del modelo reducido. No obstante, el modelo reducido cambia la respuesta pronosticada del 3,8% de los votantes de *En desacuerdo* a *De acuerdo*, más de la mitad de los cuales se observó que respondían *En desacuerdo* o *Muy en desacuerdo*. Esta diferencia es muy importante y es necesario realizar un estudio cuidadoso antes de optar por el modelo reducido.

Resumen

Mediante el procedimiento Regresión ordinal de muestras complejas, ha generado varios posibles modelos del nivel de apoyo al proyecto de ley basados en los datos demográficos de los votantes. La prueba de las líneas paralelas ha mostrado que no es necesario recurrir a un modelo acumulado generalizado. Las pruebas de los efectos del modelo sugieren que *Genero* y *Votó en las últimas elecciones* pueden eliminarse del modelo y este modelo reducido funciona bien en lo que se refiere al valor de pseudo R^2 y la tasa de clasificación global en comparación con el modelo original. No obstante, el modelo reducido clasifica incorrectamente más votantes entre la división *De acuerdo*/*En desacuerdo*, por lo que los legisladores prefieren seguir utilizando por ahora el modelo original.

Procedimientos relacionados

El procedimiento Regresión ordinal de muestras complejas es una herramienta útil para crear modelos de una variable ordinal cuando los casos se han extraído siguiendo un esquema de muestreo complejo.

- El Asistente de muestreo de la opción Muestras complejas se utiliza para definir las especificaciones de diseño de las muestras complejas y obtener una muestra. El archivo del plan de muestreo creado por el Asistente de muestreo contiene un plan de análisis por defecto que se puede especificar en el cuadro de diálogo Plan cuando se analiza la muestra obtenida de acuerdo con dicho plan.

- El Asistente de preparación del análisis de la opción Muestras complejas se utiliza para configurar las especificaciones de análisis para una muestra compleja existente. El archivo del plan de muestreo creado por el Asistente de muestreo se puede especificar en el cuadro de diálogo Plan cuando se analiza la muestra correspondiente a dicho plan.
- El procedimiento Modelo lineal general de muestras complejas permite crear un modelo de una respuesta de escala.
- El procedimiento Regresión logística de muestras complejas permite crear un modelo de una respuesta categórica.

Capítulo
22

Regresión de Cox de muestras complejas

El procedimiento Regresión de Cox de muestras complejas realiza análisis de supervivencias para muestras extraídas mediante métodos de muestreo complejo.

Uso de un predictor dependiente del tiempo en la regresión de Cox de muestras complejas

Un organismo de orden público está preocupado por los índices de reincidencia en su área de jurisdicción. Una de las medidas de la reincidencia es el tiempo que transcurre antes del segundo arresto de los delincuentes. El organismo desea crear un modelo que refleje el tiempo que transcurre antes de un nuevo arresto utilizando la regresión de Cox en una muestra extraída mediante métodos de muestreo complejo, pero les preocupa que el supuesto de proporcionalidad de los impactos no sea válido en todas las diferentes categorías de edad.

Se seleccionaron las personas puestas en libertad tras su primer arresto durante el mes de junio de 2003 de una muestra de departamentos y se examinó su historial de casos hasta el final de junio de 2006. Esta muestra se incluye en *recidivism_cs_sample.sav*. El plan de muestreo utilizado se incluye en *recidivism_cs.csplan*. Como utiliza un método de probabilidad proporcional al tamaño (PPS), también hay un archivo que contiene las probabilidades de selección conjunta (*recidivism_cs_jointprob.sav*). Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*. Utilice la regresión de Cox de muestras complejas para evaluar la validez del supuesto de proporcionalidad de los impactos y ajuste un modelo con predictores dependientes del tiempo, si es adecuado.

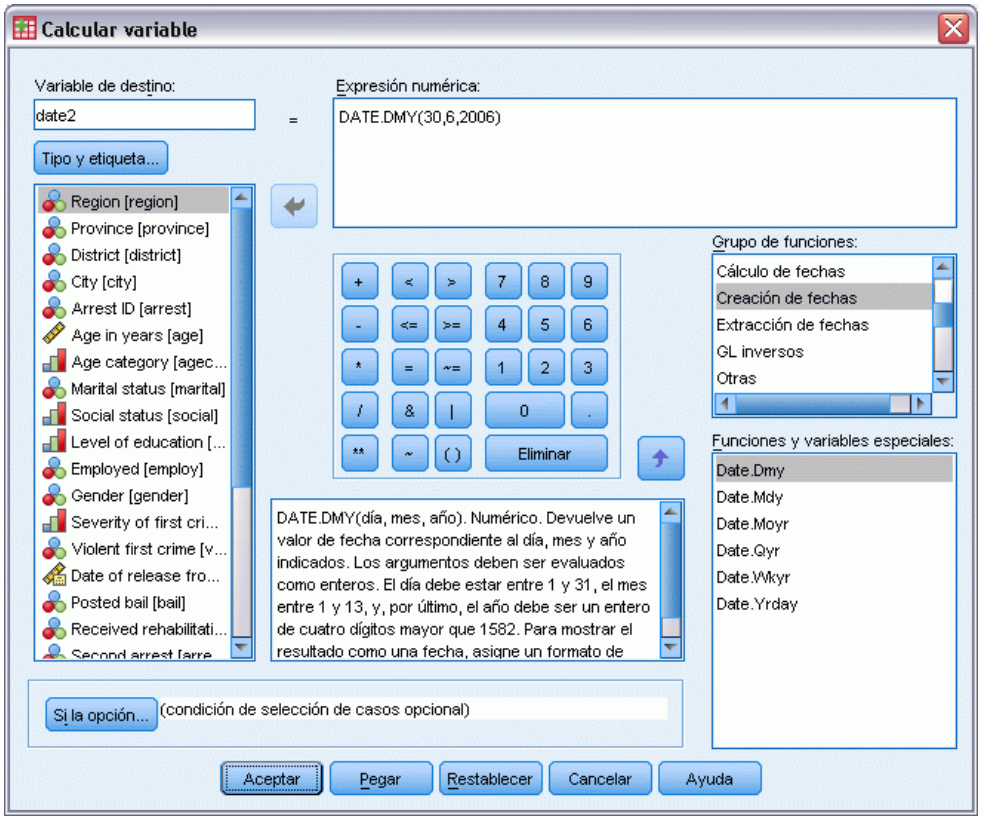
Preparación de los datos

El conjunto de datos que contiene las fechas de puesta en libertad del primer arresto y del segundo arresto. Como la regresión de Cox analiza los tiempos de supervivencia, será necesario calcular el intervalo de tiempo transcurrido entre estas fechas.

No obstante, *Date of second arrest [date2]* contiene casos con el valor 10/03/1582, un valor perdido para las variables de fecha. Estos casos corresponden a personas que no han cometido un segundo delito y, sin duda alguna, deseamos incluirlos como casos correctamente censurados a la derecha en el modelo. El final del período de seguimiento fue el 30 de junio de 2006, por lo que vamos a recodificar 10/03/1582 como 06/30/2006.

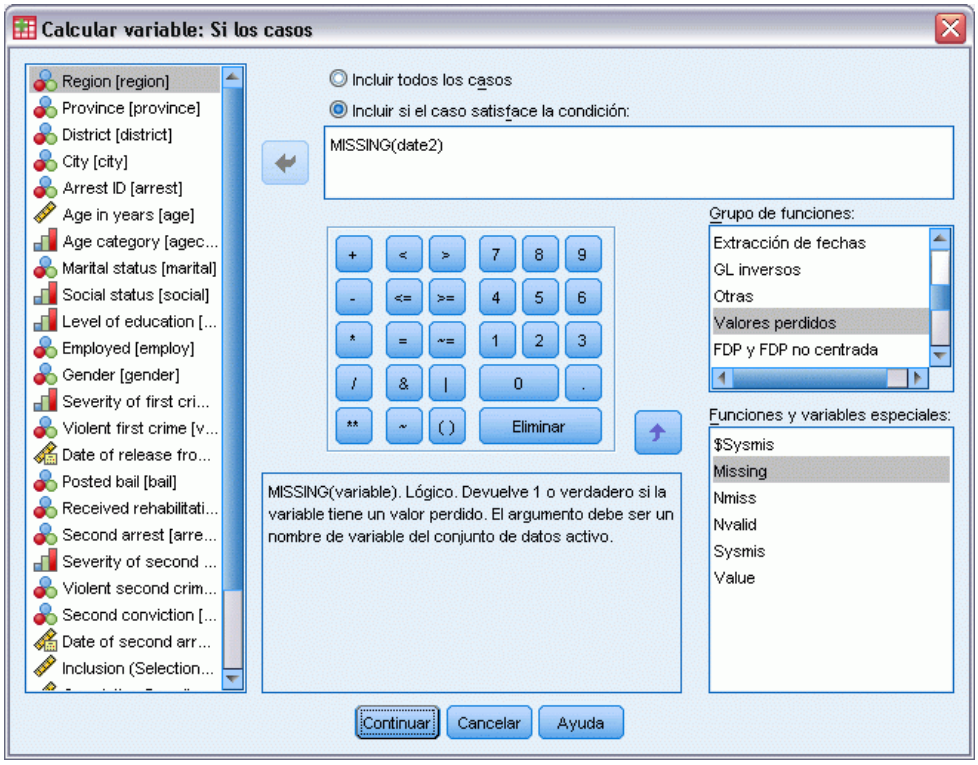
- Para recodificar estos valores, elija en los menús:
  Transformar > Calcular variable...

Figura 22-1
Cuadro de diálogo Calcular variable



- Escriba date2 como la variable de destino.
- Escriba DATE.DMY(30,6,2006) como la expresión.
- Pulse en Si.

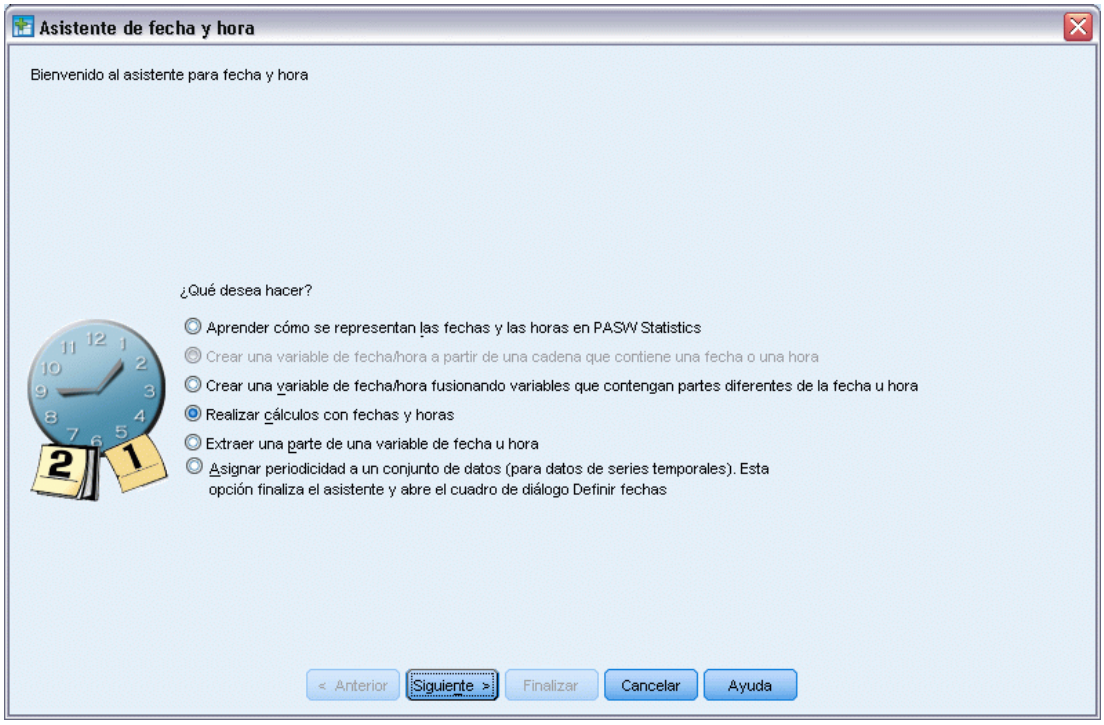
Figura 22-2
Cuadro de diálogo Calcular variable: Si los casos



- Seleccione Incluir si el caso satisface la condición.
- Escriba MISSING(date2) como la expresión.
- Pulse en Continuar.
- Pulse Aceptar en el cuadro de diálogo Calcular variable.
- A continuación, para calcular el tiempo transcurrido entre el primer arresto y el segundo, elija en los menús:

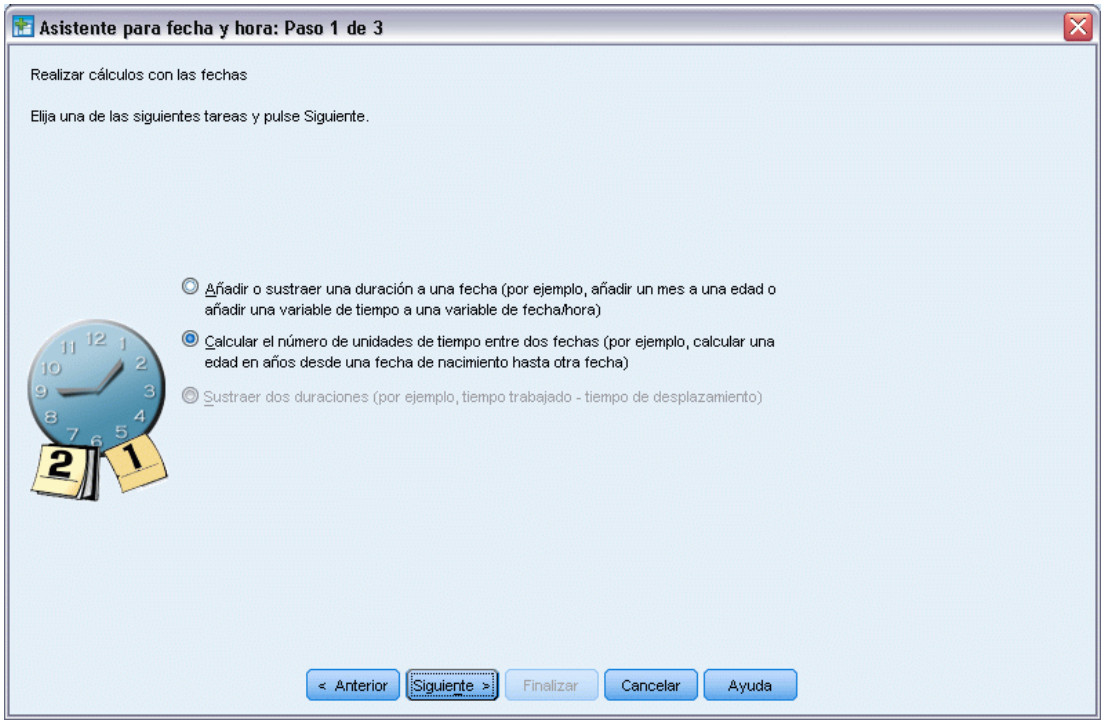
  Transformar > Asistente para fecha y hora...

Figura 22-3
Asistente para fecha y hora, paso Bienvenida



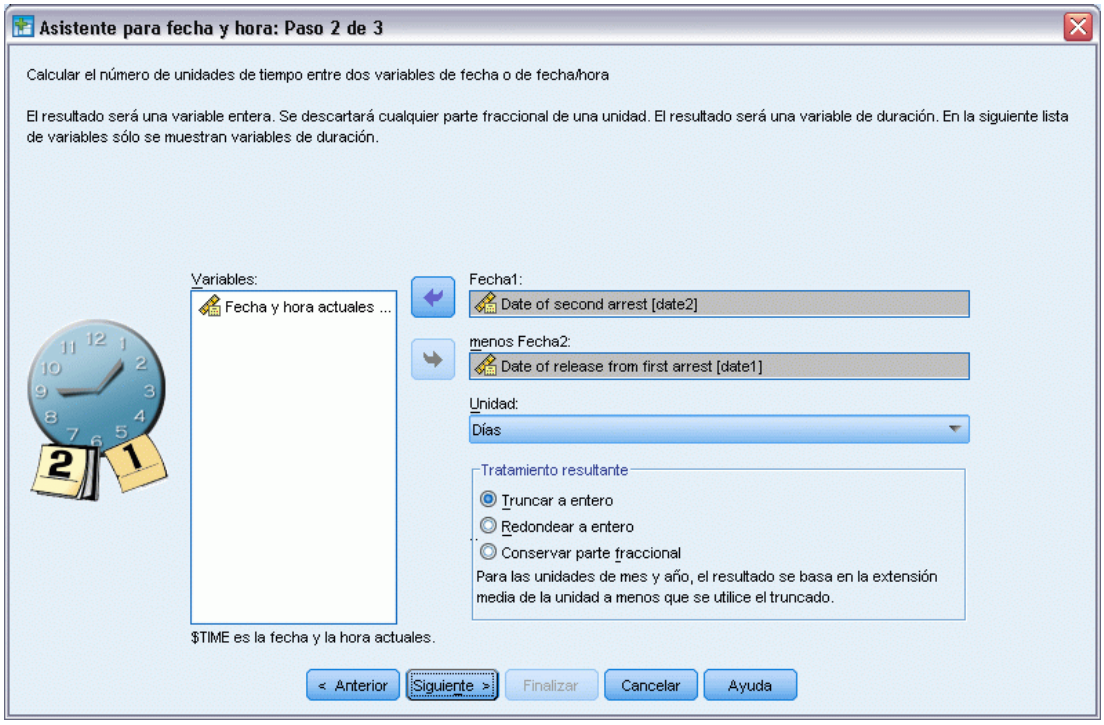
- Seleccione Realizar cálculos con fechas y horas.
- Pulse en Siguiente.

Figura 22-4
Asistente para fecha y hora, paso Realizar cálculos con las fechas



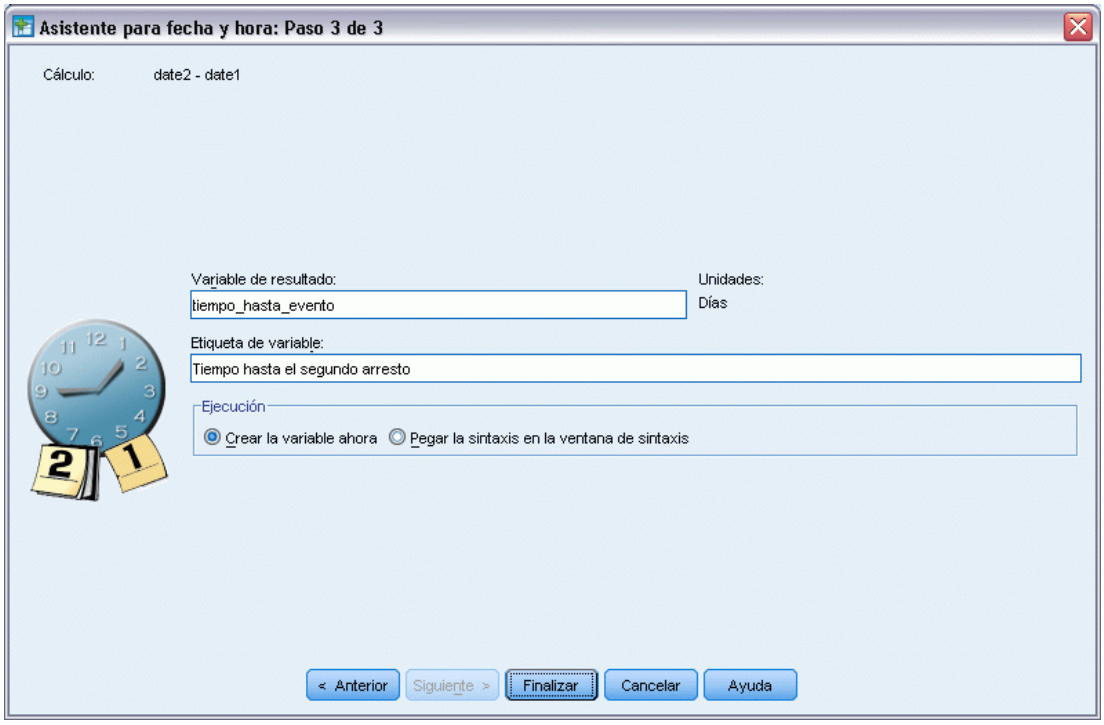
- Seleccione Calcular el número de unidades de tiempo entre dos fechas.
- Pulse en Siguiente.

Figura 22-5
Asistente para fecha y hora, paso Calcular el número de unidades de tiempo entre dos fechas



- Seleccione *Date of second arrest [date2]* como primera fecha.
- Seleccione *Date of release from first arrest [date1]* como la fecha que se restará a la primera fecha.
- Seleccione Días como unidad.
- Pulse en Siguiente.

Figura 22-6
Asistente para fecha y hora, paso Cálculo

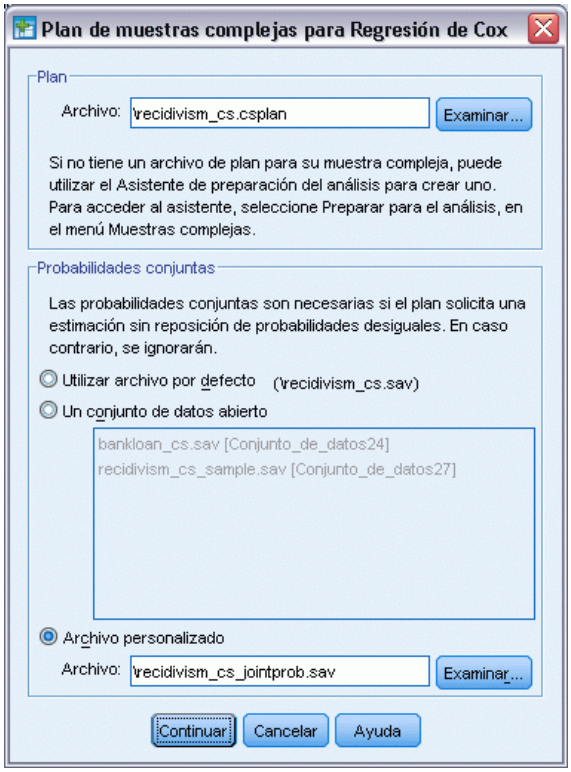


- Escriba *time_to_event* como nombre de la variable que representa el tiempo transcurrido entre las dos fechas.
- Escriba *Time to second arrest* como etiqueta de variable.
- Pulse en Finalizar.

Ejecución del análisis

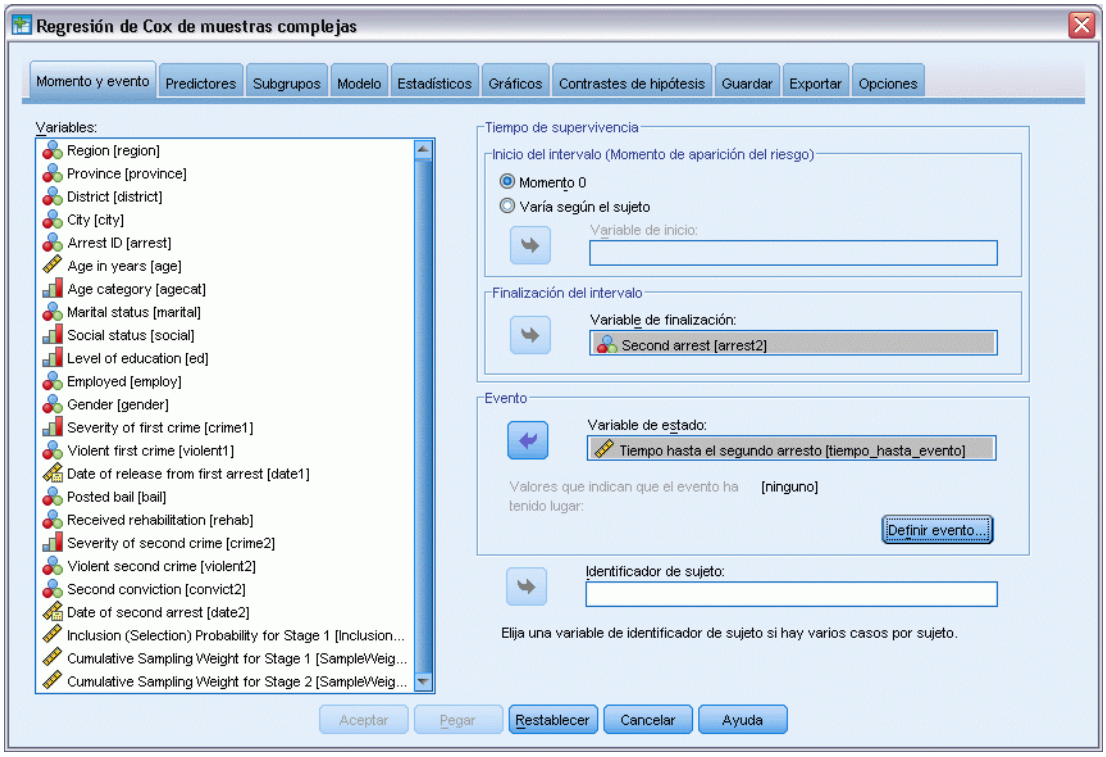
- Para ejecutar un análisis de regresión de Cox de muestras complejas, seleccione en los menús:
  Analizar > Complex Samples > Regresión de Cox...

Figura 22-7
Cuadro de diálogo Plan de muestras complejas para Regresión de Cox



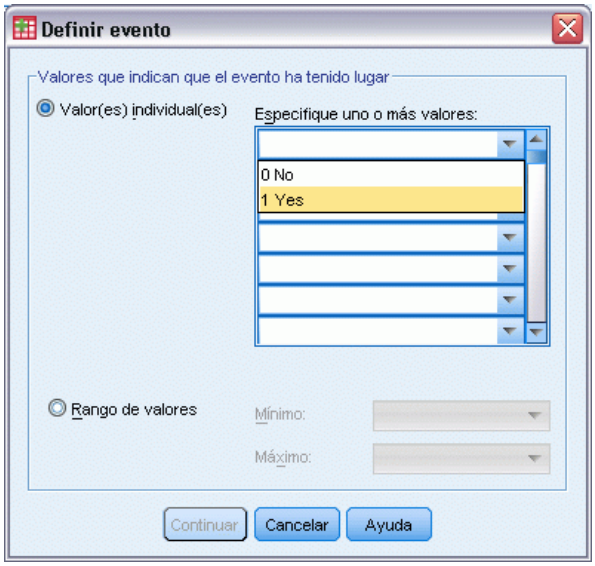
- Busque el directorio de archivos de ejemplo y seleccione *recidivism_cs.csplan* como archivo de plan.
- Seleccione Archivo personalizado en el grupo Probabilidades conjuntas, busque el directorio de archivos de ejemplo y seleccione *recidivism_cs_jointprob.sav*.
- Pulse en Continuar.

Figura 22-8
Cuadro de diálogo Regresión de Cox, pestaña Momento y evento



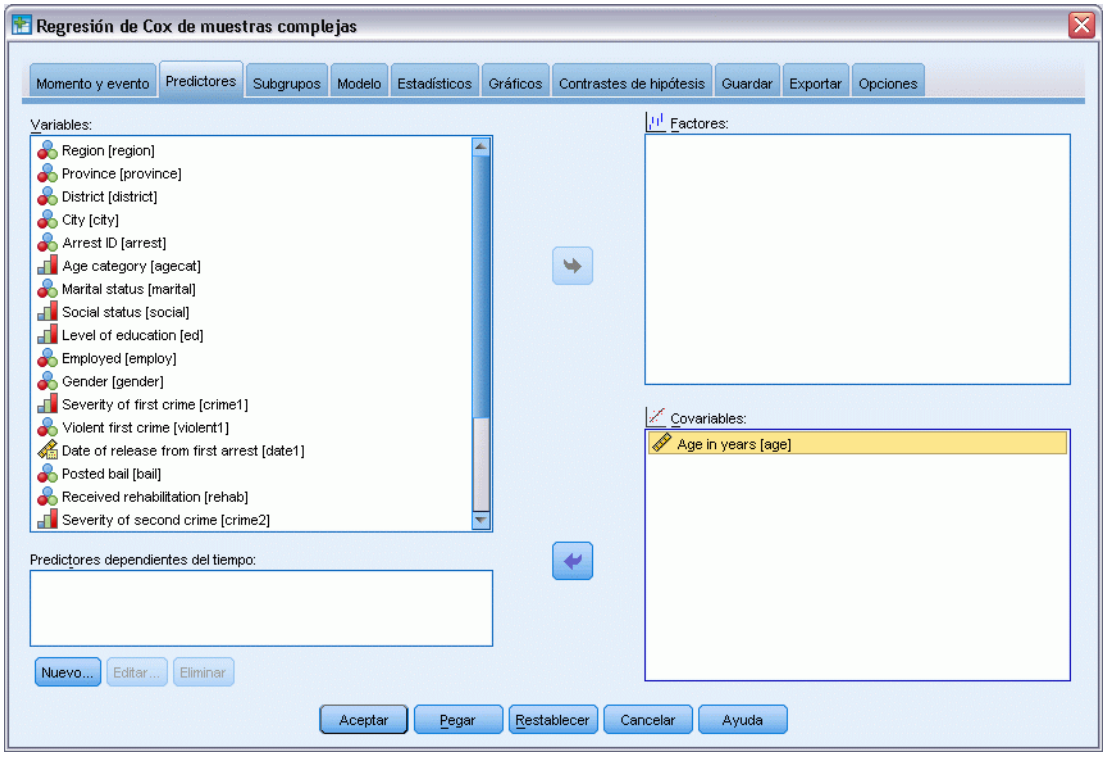
- Seleccione *Time to second arrest [time_to_event]* como la variable que define el final del intervalo.
- Seleccione *Second arrest [arrest2]* como la variable que define cuándo se ha producido el evento.
- Pulse en Definir evento.

Figura 22-9
Cuadro de diálogo Definir evento



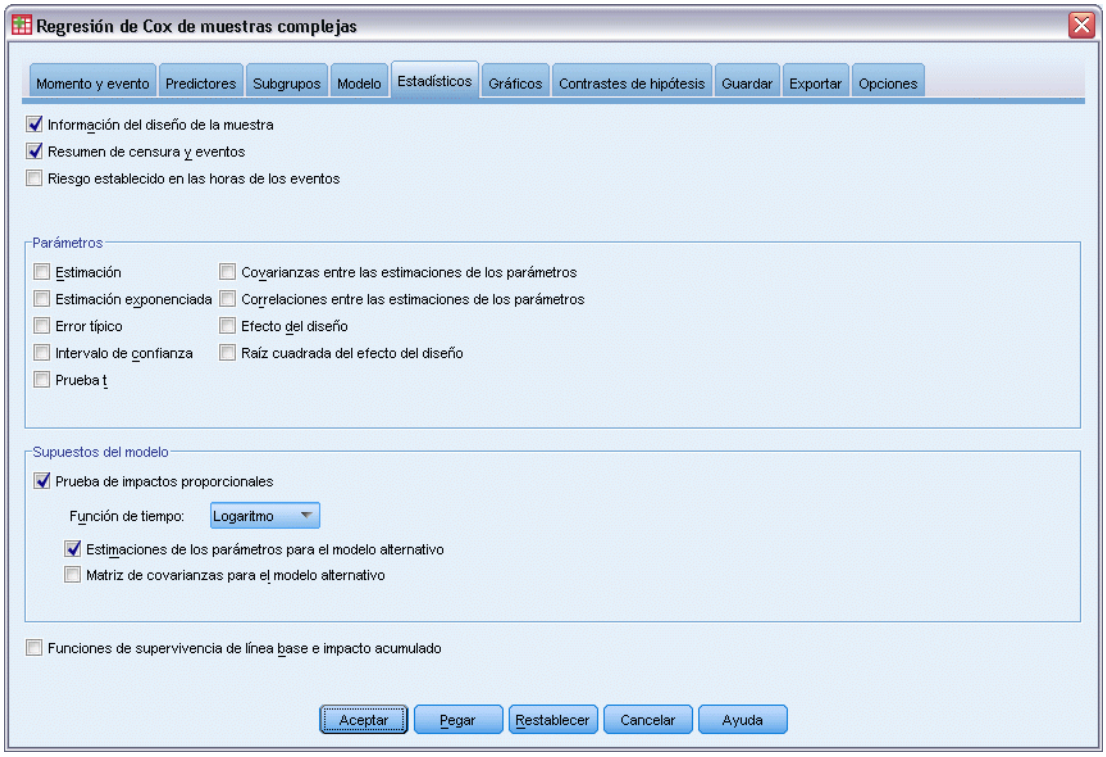
- Seleccione 1 Yes como el valor que indica que se ha producido el evento de interés (el nuevo arresto).
- Pulse en Continuar.
- Pulse en la pestaña Predictores.

Figura 22-10
Cuadro de diálogo Regresión de Cox, pestaña Predictores



- Seleccione *Age in years [age]* como covariable.
- Pulse en la pestaña Estadísticos.

Figura 22-11
Cuadro de diálogo Regresión de Cox, pestaña Estadísticos



- Seleccione Prueba de impactos proporcionales y, a continuación, seleccione Log como función de tiempo en el grupo Supuestos del modelo.
- Seleccione Estimaciones de los parámetros para el modelo alternativo.
- Pulse en Aceptar.

Información de diseño de la muestra

Figura 22-12
Información del diseño muestral

			N
Recuentos no ponderados	Válidos	Sujetos	5687
		Casos	5687
	Casos no válidos		0
	Casos totales		5687
Tamaño de sujetos de la población			307583,898
Etapa 1	Estratos		4
	Unidades		20
Grados de libertad del diseño del muestreo			16

Esta tabla contiene información sobre el diseño muestral relevante para la estimación del modelo.

- Hay un caso por sujeto y todos los 5.687 casos se han utilizado en el análisis.

- La muestra representa menos del 2% de la totalidad de la población estimada.
- El diseño requiere 4 estratos y 5 unidades por estrato para un total de 20 unidades en la primera etapa del diseño. Los grados de libertad del diseño muestral se estiman mediante 20−4=16.

Pruebas de efectos del modelo

Figura 22-13
Contrastes de los efectos del modelo

Origen	gl1	gl2	F de Wald	Sig.
age	1,000	16	504,787	1,580E-13

Variable de tiempo de supervivencia: Time to second arrest
Variable de estado de evento: Second arrest = 1,0
Modelo: age

En el modelo de impactos proporcionales, el valor de significación del predictor *age* es inferior a 0,05 y, por tanto, parece contribuir al modelo.

Prueba de impactos proporcionales

Figura 22-14
Prueba global de impactos proporcionales

gl1	gl2	F de Wald	Sig.
1,000	16,000	29,924	5,136E-5

Variable de tiempo de supervivencia: Time to second arrest
Variable de estado de evento: Second arrest = 1,0
Modelo: age, age*_TF

Figura 22-15
Estimaciones de los parámetros para el modelo alternativo

			Intervalo de confianza al 90%	
Parámetro	B	Error típico	Inferior	Superior
age	-,002	,014	-,025	0,02
age*_TF*	-,012	,002	-,016	-,009

Variable de tiempo de supervivencia: Time to second arrest
Variable de estado de evento: Second arrest = 1,0
Modelo: age, age*_TF
a. Función de tiempo: Log

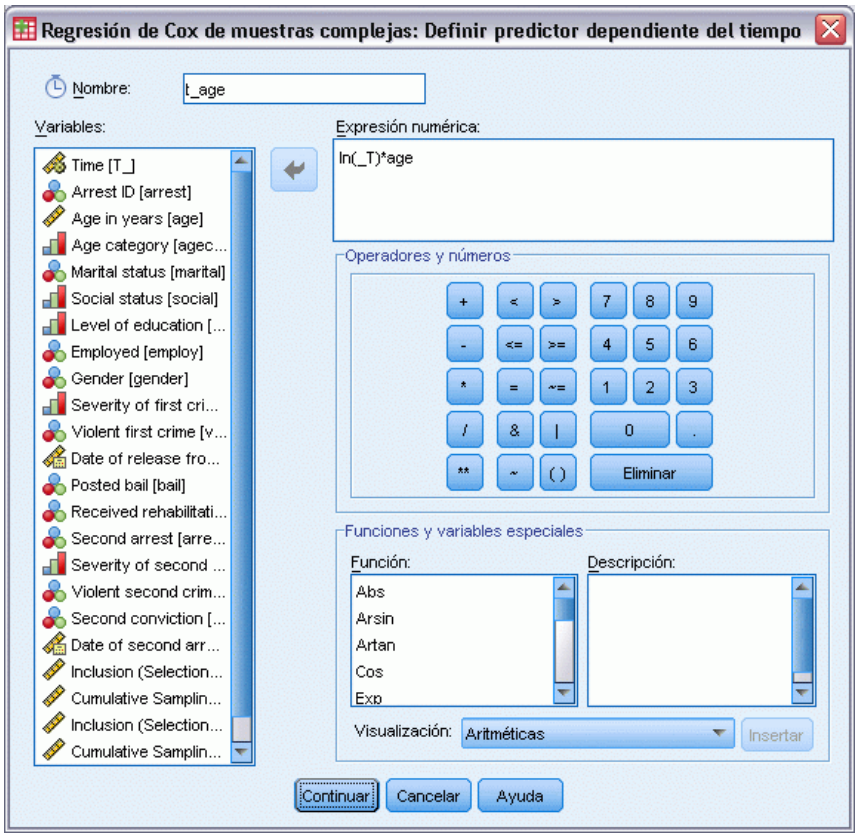
El valor de significación de la prueba global de impactos proporcionales es inferior a 0,05, lo que indica que se viola el supuesto de proporcionalidad de los impactos. En el modelo alternativo se utiliza la función de logaritmo del tiempo, por lo que será fácil replicar este predictor dependiente del tiempo.

Adición de un predictor dependiente del tiempo

► Recupere el cuadro de diálogo Regresión de Cox de muestras complejas y pulse en la pestaña Predictores.

▶ Pulse en Nuevo.

Figura 22-16
Cuadro de diálogo Regresión de Cox: Definir predictor dependiente del tiempo

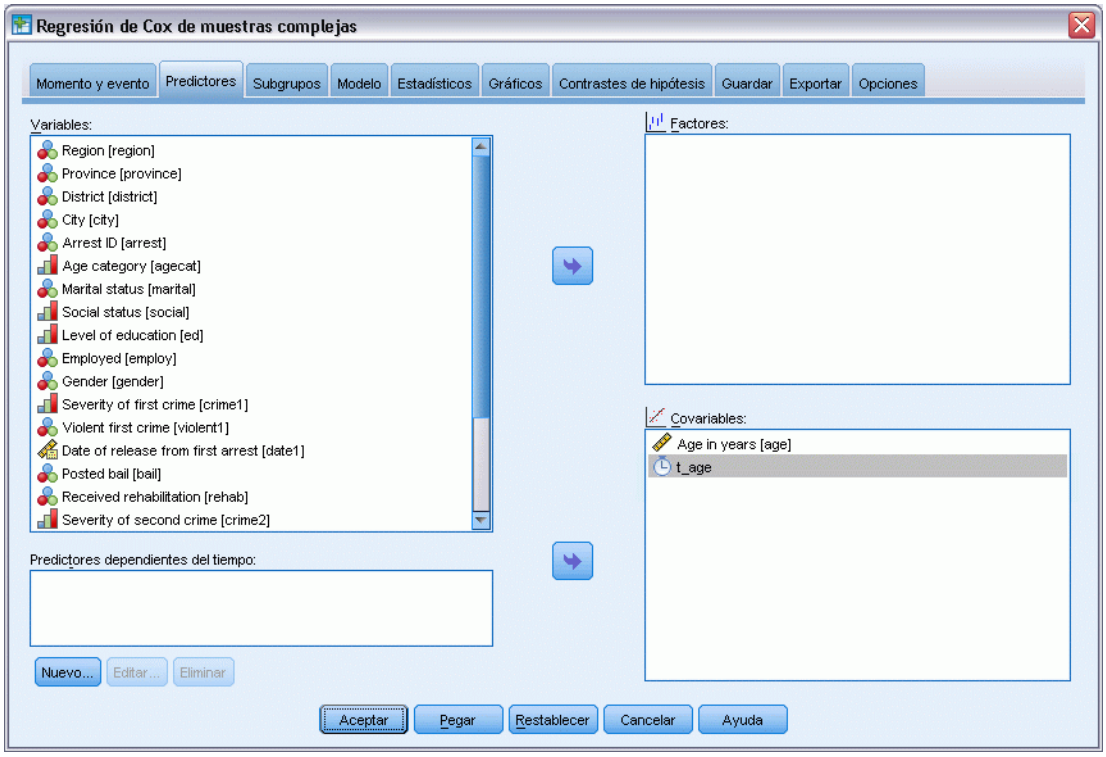


▶ Escriba t_age como nombre del predictor dependiente del tiempo que desea definir.

▶ Escriba ln(T_)*age como expresión numérica.

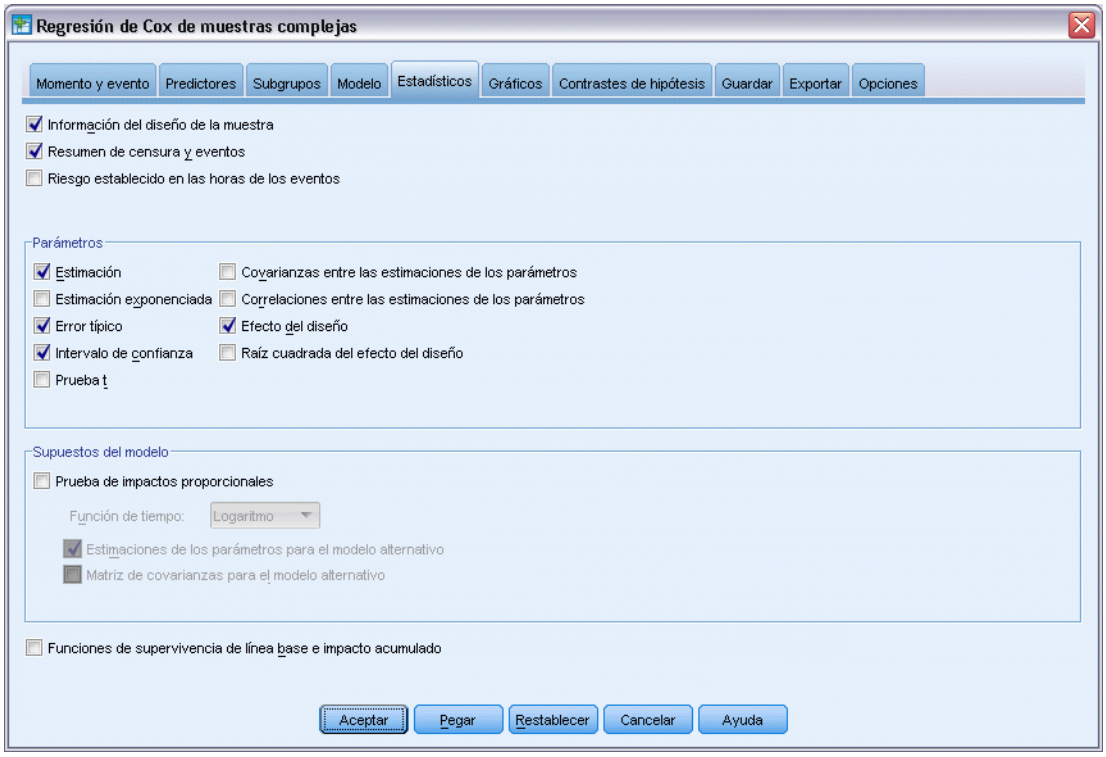
▶ Pulse en Continuar.

Figura 22-17
Cuadro de diálogo Regresión de Cox, pestaña Predictores



- Seleccione *t_age* como covariable.
- Pulse en la pestaña Estadísticos.

Figura 22-18
Cuadro de diálogo Regresión de Cox, pestaña Predictores



- Seleccione Estimación, Error típico, Intervalo de confianza y Efecto del diseño en el grupo Parámetros.
- Anule la selección de Prueba de impactos proporcionales y Estimaciones de los parámetros para el modelo alternativo en el grupo Supuestos del modelo.
- Pulse en Aceptar.

Pruebas de efectos del modelo

Figura 22-19
Contrastes de los efectos del modelo

Origen	gl1	gl2	F de Wald	Sig.
age	1,000	16,000	,015	0,91
t_age	1,000	16,000	29,924	5,136E-5

Variable de tiempo de supervivencia: Time to second arrest
Variable de estado de evento: Second arrest = 1
Modelo: age, t_age

Tras añadir el predictor dependiente del tiempo, el valor de significación de *age* es 0,91, lo que indica que su contribución al modelo queda superada por la de *t_age*.

Estimaciones de los parámetros

Figura 22-20
Estimaciones de los parámetros

Parámetro	B	Error típico	Intervalo de confianza al 95%		Efecto del diseño
			Inferior	Superior	
age	-,002	0,01	-,030	,027	,702
t_age	-,012	,002	-,017	-,008	,666

Variable de tiempo de supervivencia: Time to second arrest
Variable de estado de evento: Second arrest = 1
Modelo: age, t_age

Si examina las estimaciones de los parámetros y los errores típicos, pude ver que ha replicado el modelo alternativo de la prueba de los impactos proporcionales. Al haber especificado explícitamente el modelo, puede solicitar gráficos y estadísticos de los parámetros adicionales. Aquí hemos solicitado el efecto del diseño; el valor de *t_age* inferior a 1 indica que el error típico de *t_age* es menor que el que se obtendría si se supusiese que el conjunto de datos era una muestra aleatoria simple. En este caso, el efecto de *t_age* seguiría siendo estadísticamente significativo, pero los intervalos de confianza serían más anchos.

Varios casos por sujeto en la regresión de Cox de muestras complejas

Los investigadores que estudian los tiempos de supervivencia de los pacientes que finalizan un programa de rehabilitación tras un ataque isquémico se enfrentan a varios retos.

Varios casos por sujeto. Las variables que representan el historial médico del paciente deben ser útiles como predictores. Con el tiempo, los pacientes pueden sufrir eventos médicos importantes que alteren su historial médico. En este conjunto de datos, la ocurrencia de infarto de miocardio, ataque isquémico o ataque hemorrágico se anotan junto con el momento en el que se produce el evento registrado. Puede crear covariables dependientes del tiempo calculables dentro del procedimiento para incluir esta información en el modelo, pero probablemente sea más cómodo utilizar varios casos por sujeto. Observe que las variables se habían codificado en un principio de manera que el historial del paciente quedaba registrado en varias variables, por lo que será necesario reestructurar el conjunto de datos.

Truncación a la izquierda. El comienzo del riesgo comienza en el momento del ataque isquémico. No obstante, la muestra incluye únicamente a aquellos pacientes que han sobrevivido al programa de rehabilitación, por lo que la muestra está truncada a la izquierda en el sentido de que los tiempos de supervivencia observados se han "inflado" con la duración de la rehabilitación. Puede tener este hecho en cuenta si especifica el momento en el que abandonaron la rehabilitación como el momento de entrada en el estudio.

No hay plan de muestreo. El conjunto de datos no se ha recopilado mediante un plan de muestreo complejo y se considera una muestra aleatoria simple. Será necesario crear un plan de análisis para utilizar la regresión de Cox de muestras complejas.

El conjunto de datos se recoge en el archivo *stroke_survival.sav*. Si desea obtener más información, consulte el tema Archivos muestrales en el apéndice A en *IBM SPSS Complex Samples 19*. Utilice el Asistente de reestructuración de datos para preparar los datos para el análisis. A continuación, utilice el Asistente de preparación del análisis para crear un plan de

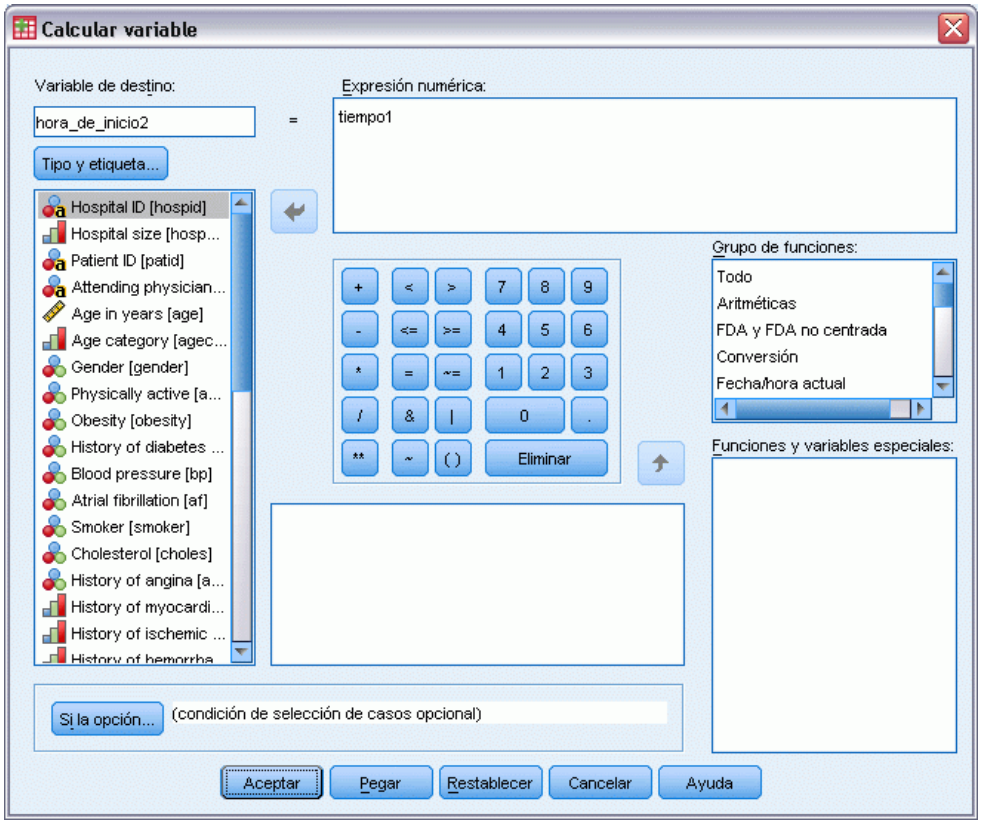
muestreo aleatorio simple y, por último, la regresión de Cox de muestras complejas para generar un modelo de los tiempos de supervivencia.

Preparación de los datos para su análisis

Antes de reestructurar los datos, deberá crear dos variables auxiliares que le ayuden en la reestructuración.

- Para calcular una nueva variable, elija en los menús:
  Transformar > Calcular variable...

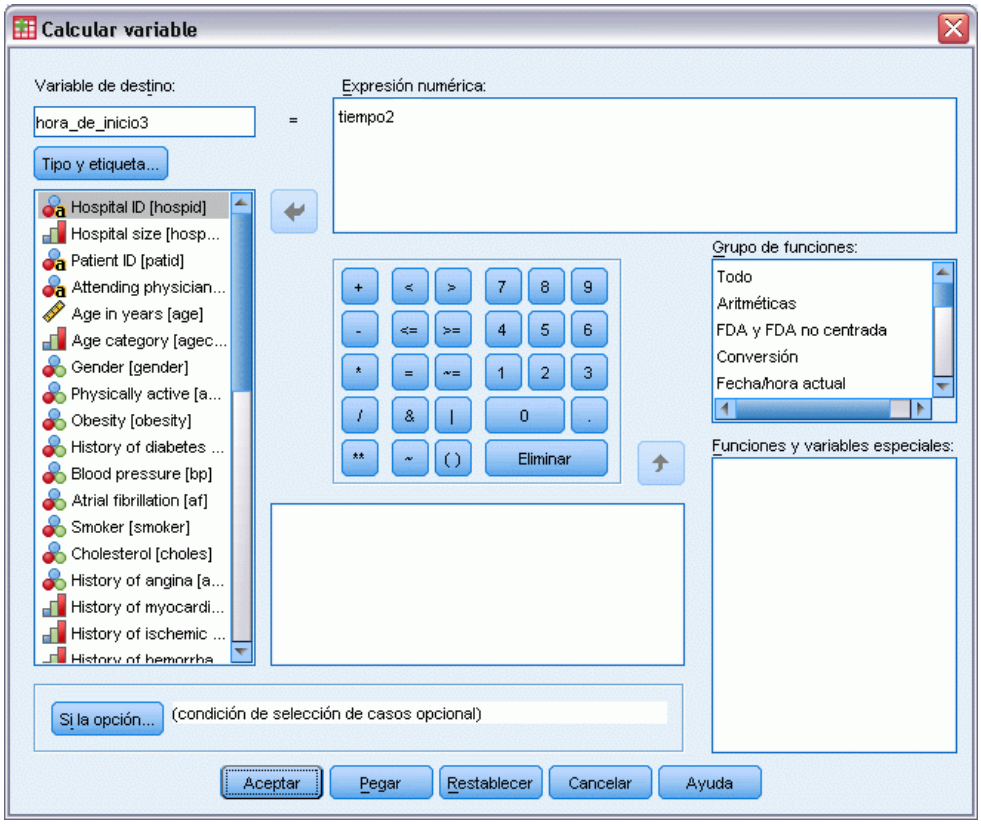
Figura 22-21
Cuadro de diálogo Calcular variable



- Escriba *start_time2* as como la variable de destino.
- Escriba *time1* como expresión numérica.
- Pulse en Aceptar.

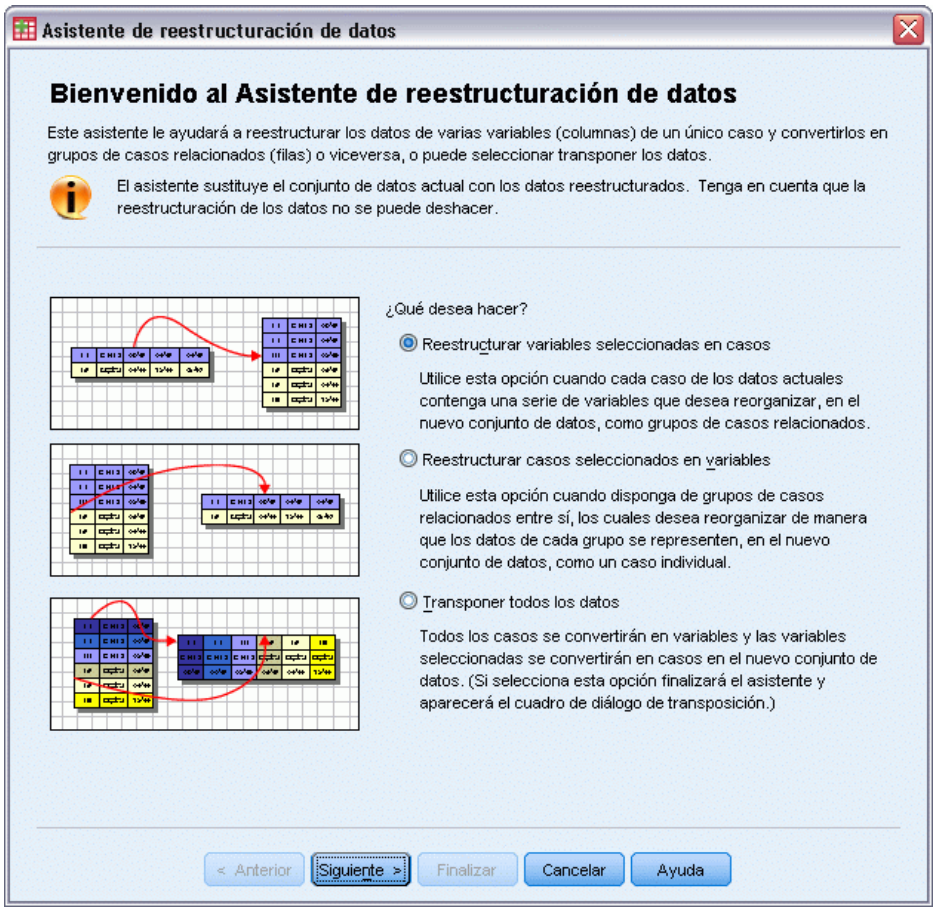
- Vuelva a abrir el cuadro de diálogo Calcular variable.

Figura 22-22
Cuadro de diálogo Calcular variable



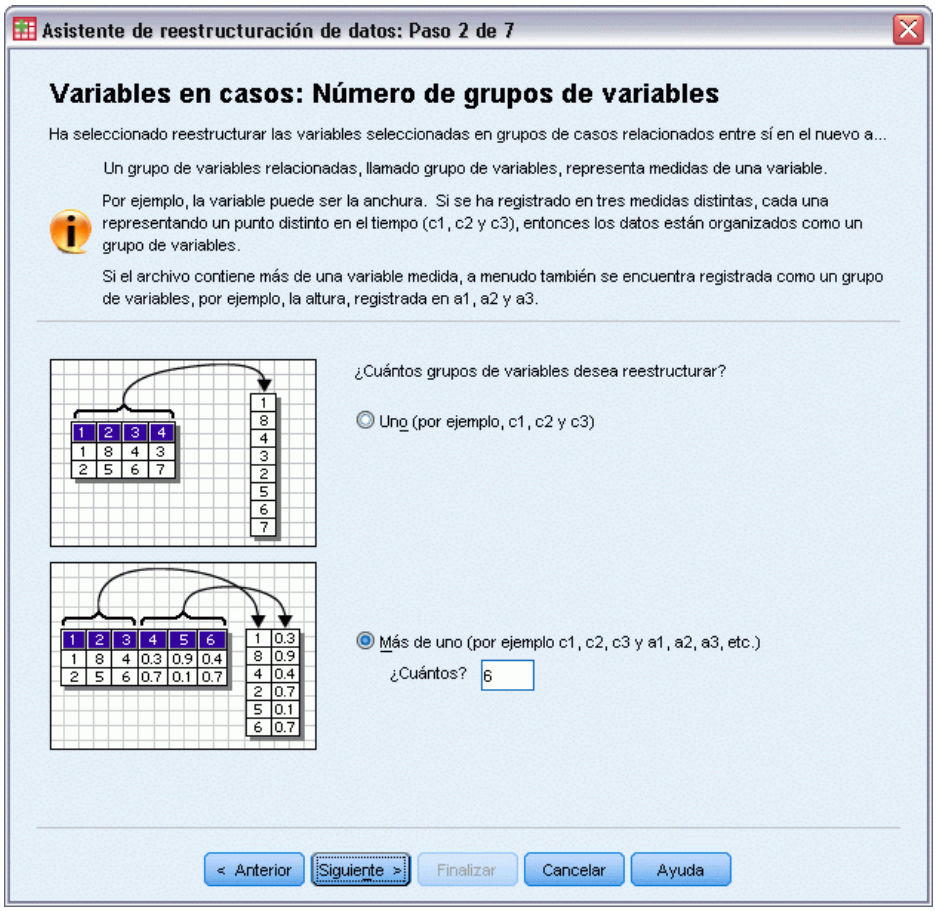
- Escriba *start_time3* como la variable de destino.
- Escriba *time2* como expresión numérica.
- Pulse en Aceptar.
- Para reestructurar los datos de variables a casos, elija en los menús:
  Datos > Reestructurar...

Figura 22-23
Asistente de reestructuración de datos, paso Bienvenida



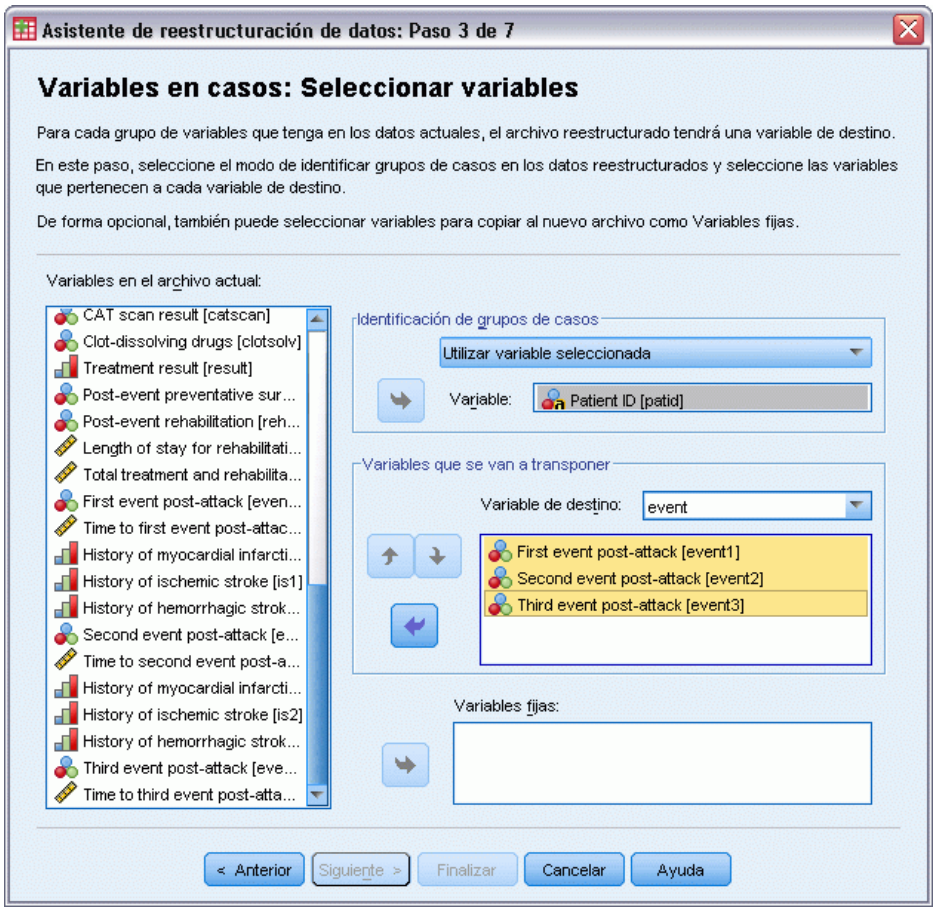
- Asegúrese de que está seleccionado Reestructurar variables seleccionadas en casos.
- Pulse en Siguiente.

Figura 22-24
Asistente de reestructuración de datos (variables a casos), paso Número de grupos de variables



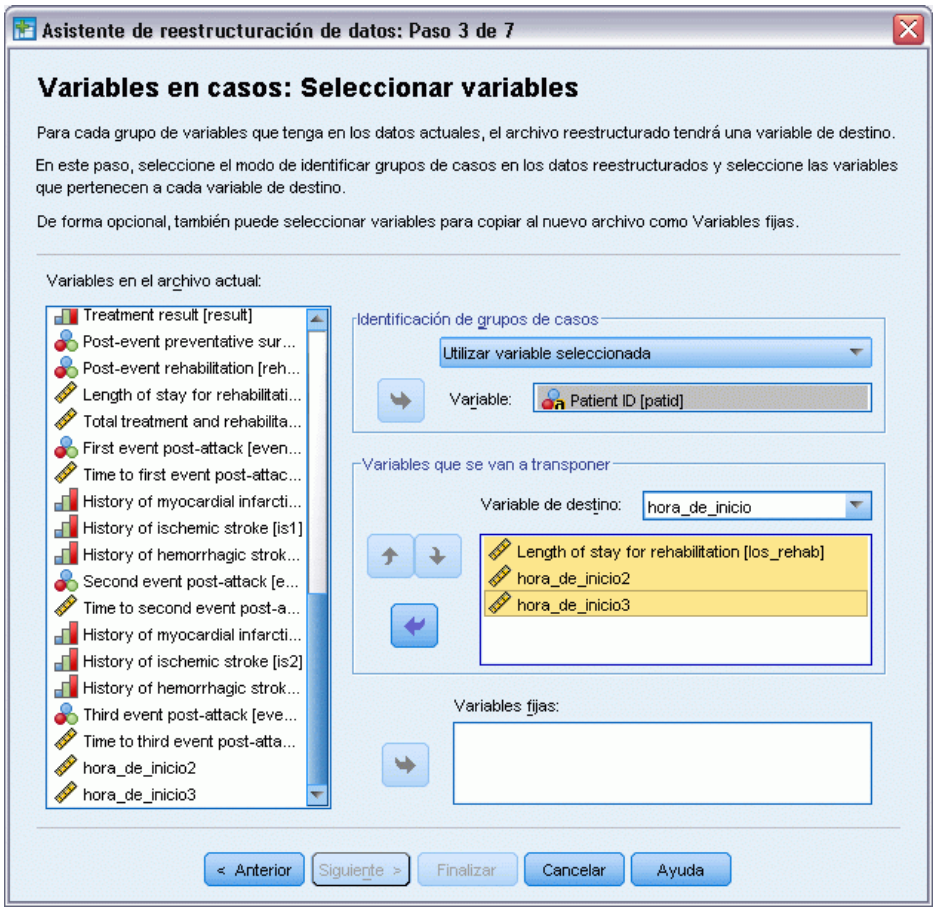
- Seleccione Más de una como grupos de variables que se van a reestructurar.
- Escriba 6 como número de grupos.
- Pulse en Siguiente.

Figura 22-25
Asistente de reestructuración de datos (variables a casos), paso Seleccionar variables



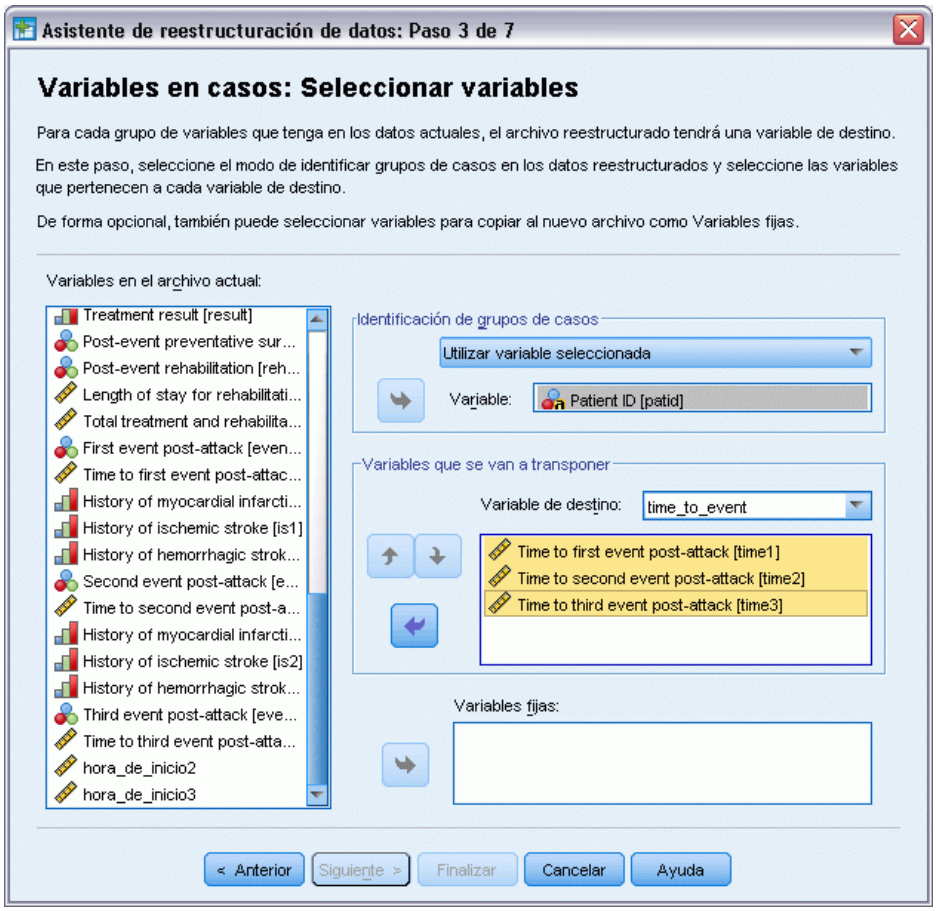
- En el grupo de Identificación de grupos de casos, seleccione Utilizar variable seleccionada y elija *Patient ID [patid]* como identificador del sujeto.
- Escriba event como la primera variable de destino.
- Seleccione *First event post-attack [event1]*, *Second event post-attack [event2]* y *Third event post-attack [event3]* como las variables que se van a transponer.
- Seleccione *trans2* en la lista de variables de destino.

Figura 22-26
Asistente de reestructuración de datos (variables a casos), paso Seleccionar variables



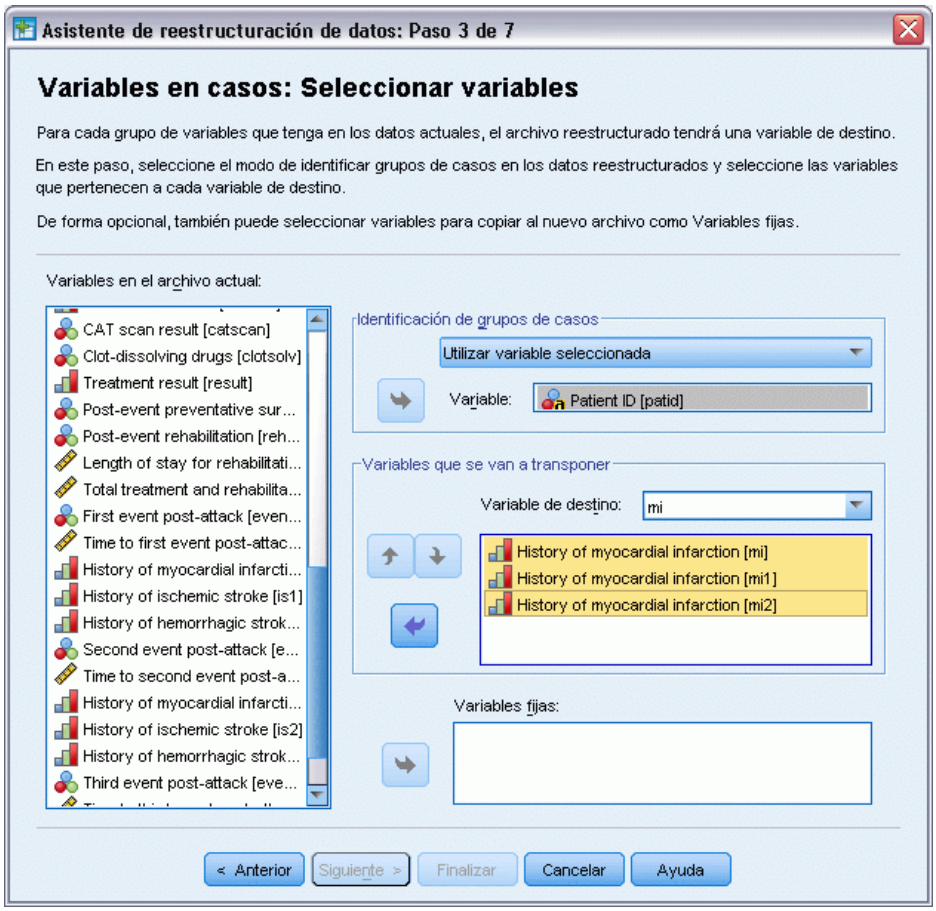
- Escriba start_time como la variable de destino.
- Seleccione *Length of stay for rehabilitation [los_rehab]*, *start_time2* y *start_time3* como las variables que se van a transponer. Se utilizarán *Time to first event post-attack [time1]* y *Time to second event post-attack [time2]* para crear las horas finales y cada variable sólo puede aparecer en una lista de variables que se van transponer, por lo que *start_time2* y *start_time3* eran necesarias.
- Seleccione *trans3* en la lista de variables de destino.

Figura 22-27
Asistente de reestructuración de datos (variables a casos), paso Seleccionar variables



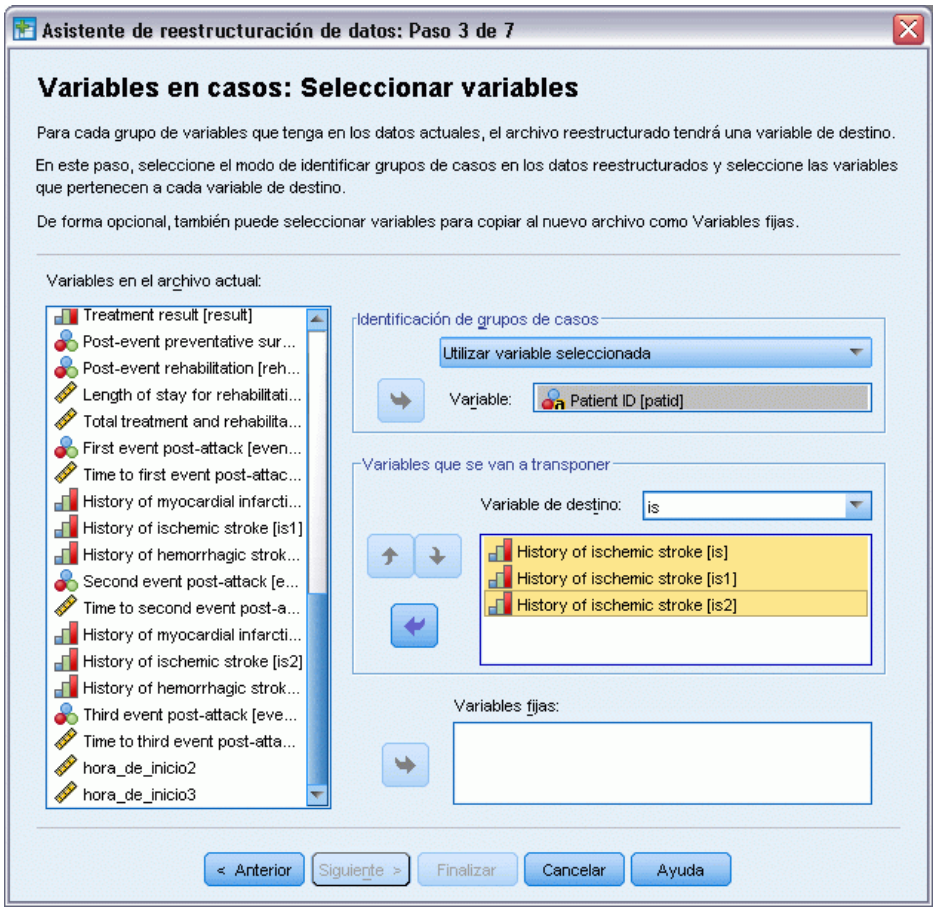
- Escriba time_to_event como la variable de destino.
- Seleccione *Time to first event post-attack [time1]*, *Time to second event post-attack [time2]* y *Time to third event post-attack [time3]* como las variables que se van a transponer.
- Seleccione *trans4* en la lista de variables de destino.

Figura 22-28
Asistente de reestructuración de datos (variables a casos), paso Seleccionar variables



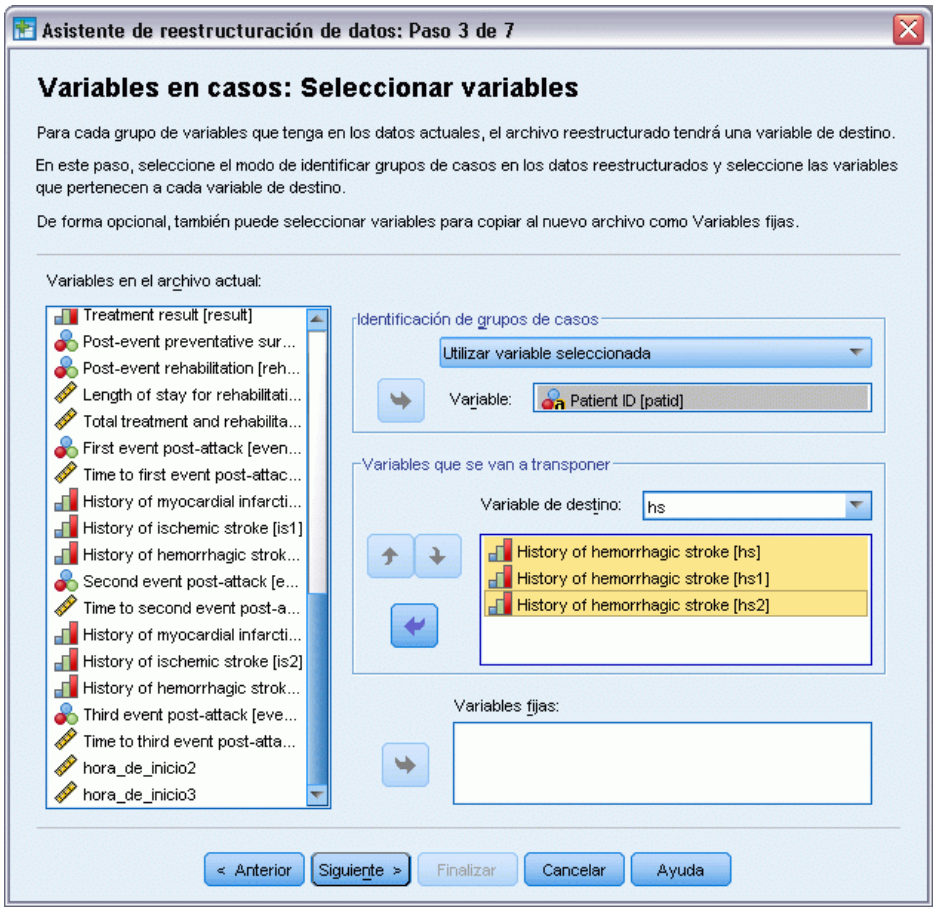
- Escriba mi como la variable de destino.
- Seleccione *History of myocardial infarction [mi]*, *History of myocardial infarction [mi1]* y *History of myocardial infarction [mi2]* como las variables que se van a transponer.
- Seleccione *trans5* en la lista de variables de destino.

Figura 22-29
Asistente de reestructuración de datos (variables a casos), paso Seleccionar variables



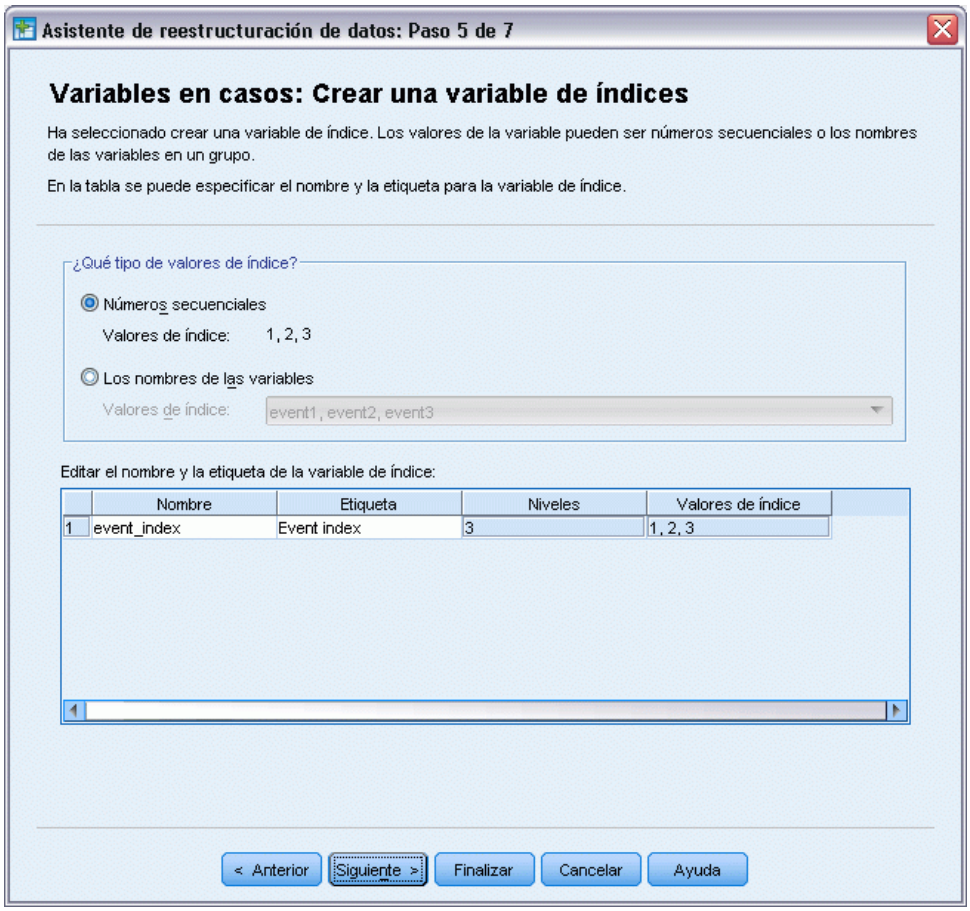
- Escriba is como la variable de destino.
- Seleccione *History of ischemic stroke [is]*, *History of ischemic stroke [is1]* y *History of ischemic stroke [is2]* como las variables que se van a transponer.
- Seleccione *trans6* en la lista de variables de destino.

Figura 22-30
Asistente de reestructuración de datos (variables a casos), paso Seleccionar variables



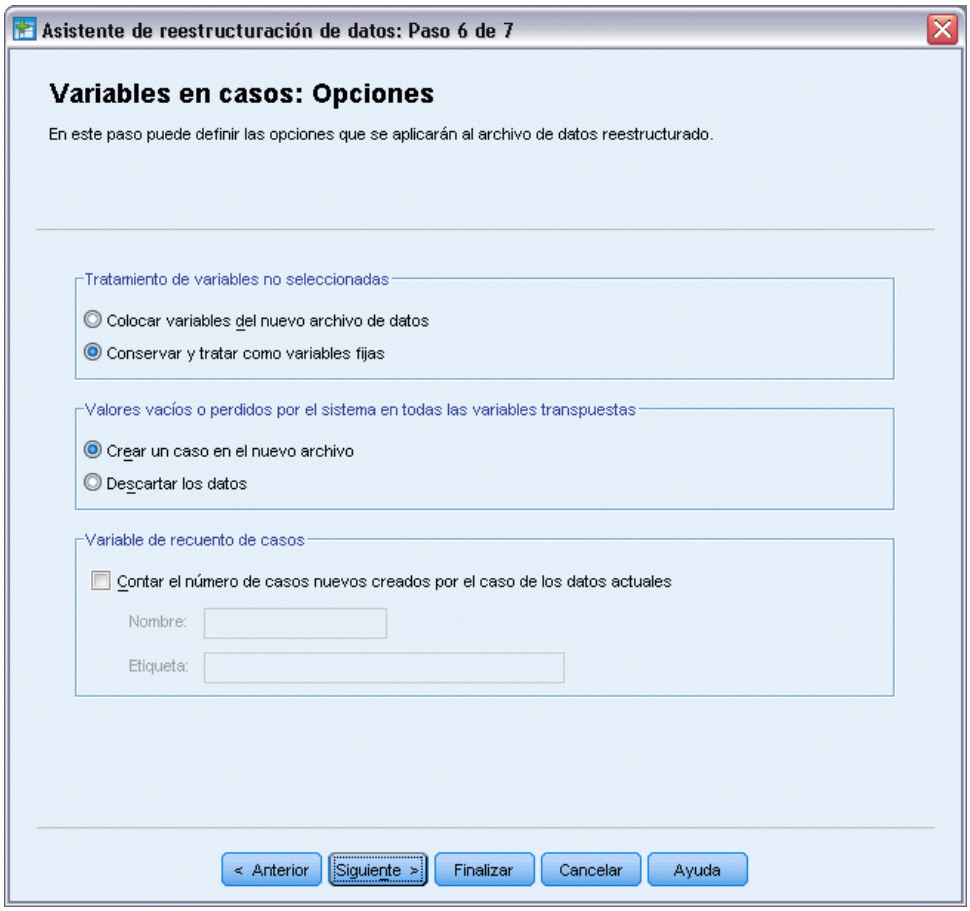
- Escriba hs como variable de destino.
- Seleccione *History of hemorrhagic stroke [hs]*, *History of hemorrhagic stroke [hs1]* y *History of hemorrhagic stroke [hs2]* como las variables que se van a transponer.
- Pulse en Siguiente y, a continuación, en Siguiente en el paso Crear variables de índice.

Figura 22-31
Asistente de reestructuración de datos (variables a casos), paso Crear una variable de índice



- Escriba event_index como nombre de la variable de índice y escriba Event index como etiqueta de variable.
- Pulse en Siguiente.

Figura 22-32
Asistente de reestructuración de datos (variables a casos), paso Crear una variable de índice



- Asegúrese de que se ha seleccionado la opción Conservar y tratar como variables fijas.
- Pulse en Finalizar.

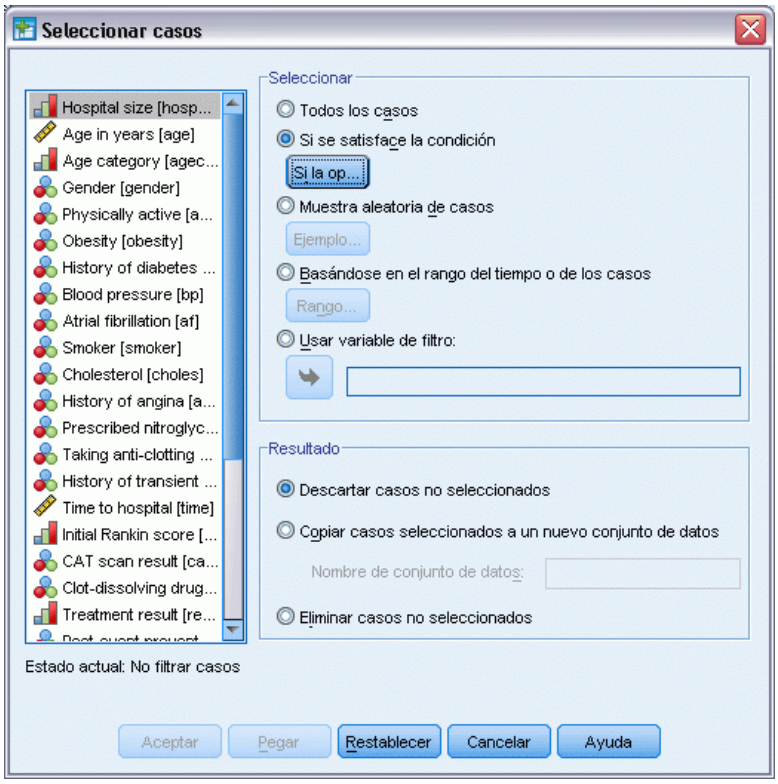
Figura 22-33
Datos reestructurados

event_index	event	start_time	time_to_event	mi	is	hs
1	0	3	1500	0	1	0
2	-4	1500	-4	-4	-4	-4
3	-4	.	-4	-4	-4	-4
1	1	33	1311	0	1	0
2	4	1311	1325	1	1	0
3	-3	1325	-3	-3	-3	-3
1	4	12	1098	1	1	0
2	-3	1098	-3	-3	-3	-3
3	-3	.	-3	-3	-3	-3
1	4	4	1356	0	1	0
2	-3	1356	-3	-3	-3	-3
3	-3	.	-3	-3	-3	-3

Los datos reestructurados contienen tres casos para cada paciente. No obstante, como muchos pacientes han sufrido menos de tres eventos, hay muchos casos con valores negativos (perdidos) para *event*. Puede filtrar el conjunto de datos para eliminar estos casos.

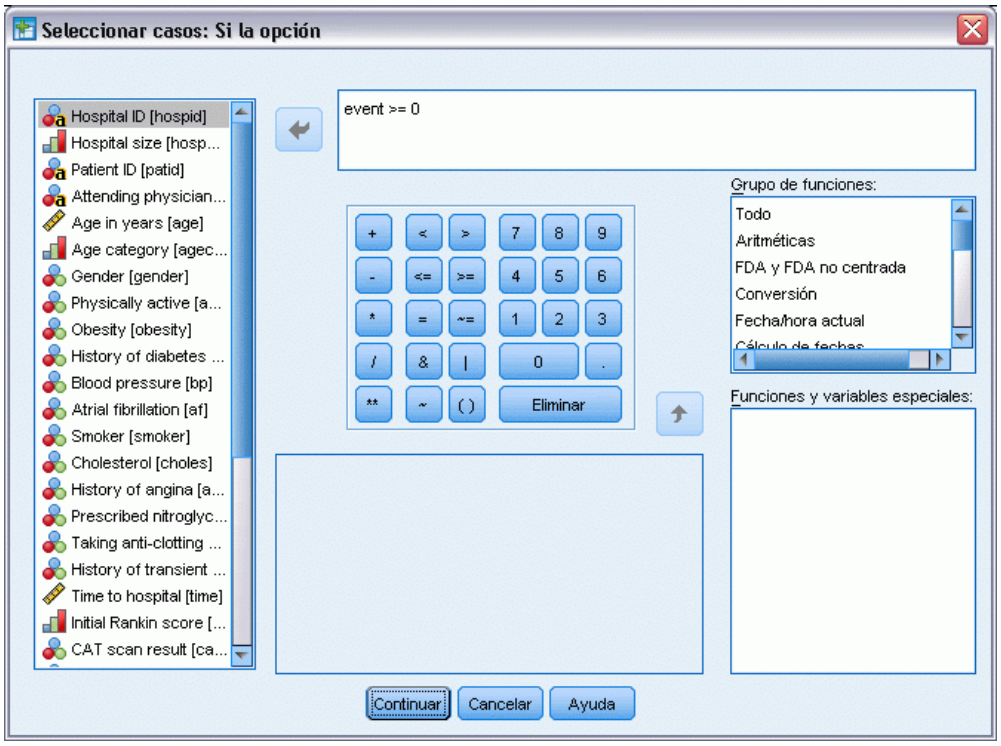
► Para ello, elija en los menús:
Datos > Seleccionar casos...

Figura 22-34
Cuadro de diálogo Seleccionar casos



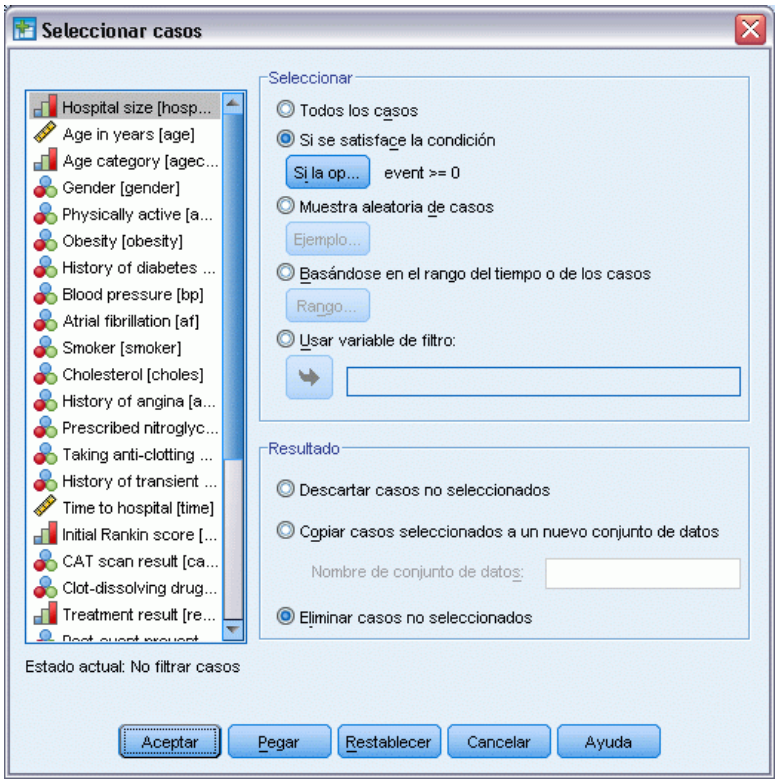
- Seleccione Si se satisface la condición.
- Pulse en Si.

Figura 22-35
Cuadro de diálogo Seleccionar casos: Si



- Escriba event >= 0 como expresión condicional.
- Pulse en Continuar.

Figura 22-36
Cuadro de diálogo Seleccionar casos



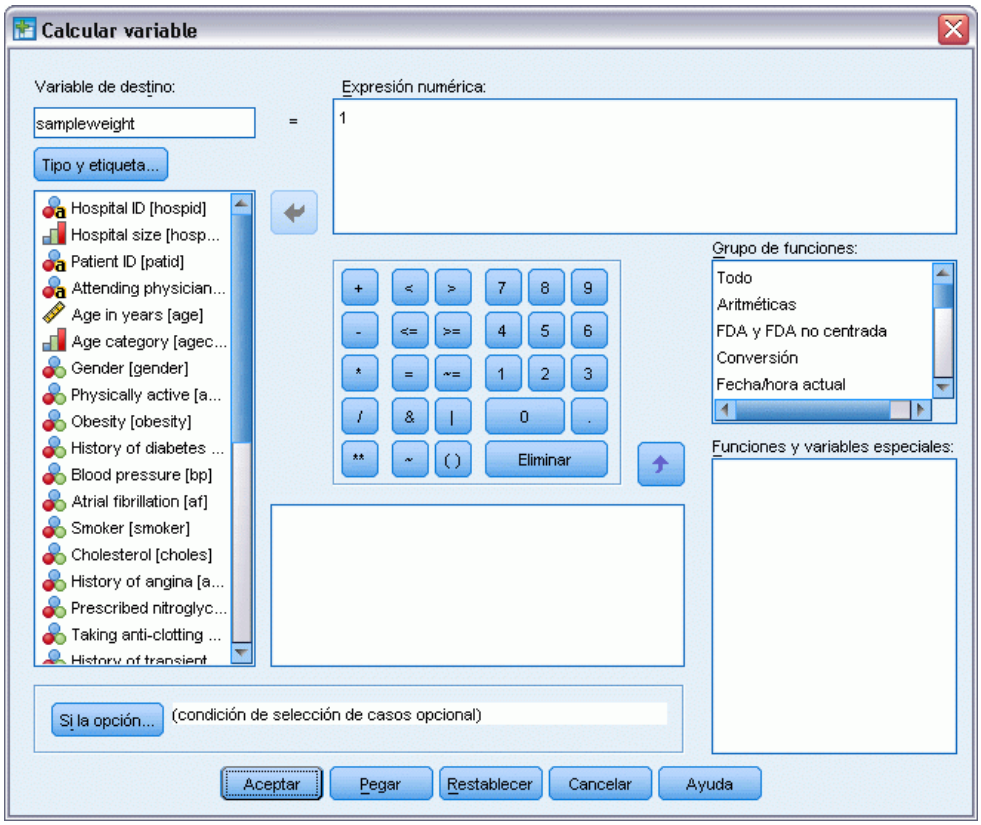
- Seleccione Eliminar casos no seleccionados.
- Pulse en Aceptar.

Creación de un plan de análisis de muestreo aleatorio simple

Ahora, ya está preparado para crear el plan de análisis de muestreo aleatorio simple.

- En primer lugar, necesita crear una variable de ponderación muestral. En los menús, seleccione: Transformar > Calcular variable...

Figura 22-37
Cuadro de diálogo principal Regresión de Cox



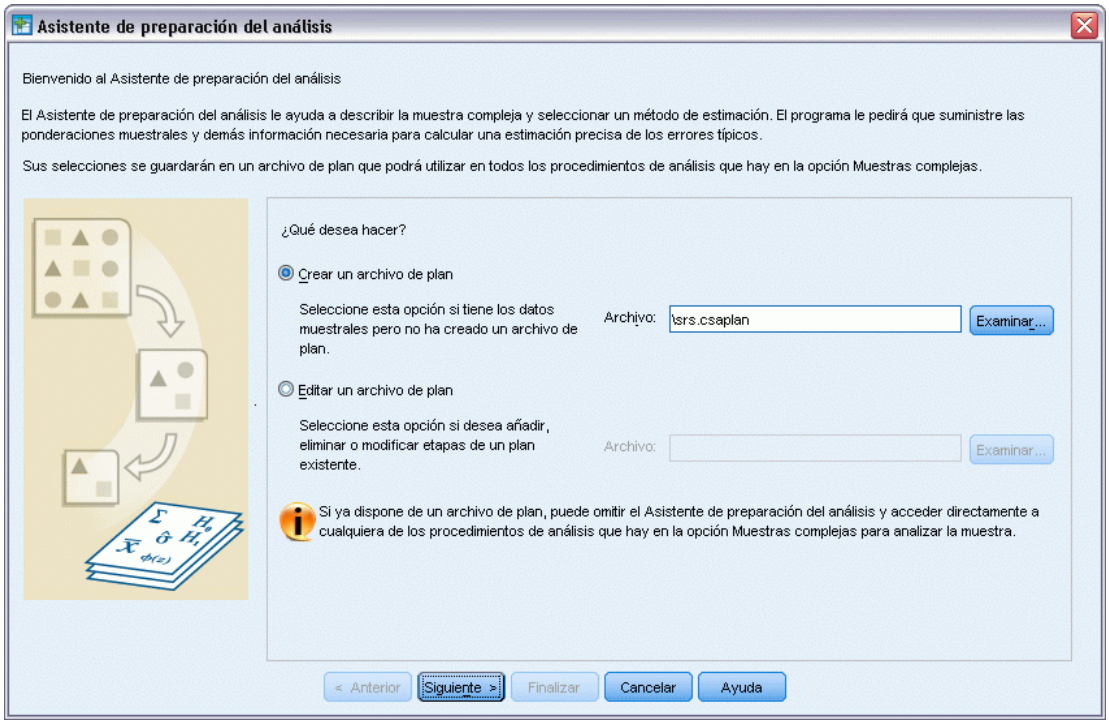
- Escriba sampleweight como variable de destino.
- Escriba 1 como expresión numérica.
- Pulse en Aceptar.

  Ya puede crear el plan de análisis.

  Nota: Hay un archivo de plan existente, *srs.csaplan*, en el directorio de archivos de ejemplo que puede utilizar si prefiere omitir las siguientes instrucciones y continuar directamente con el análisis de los datos.

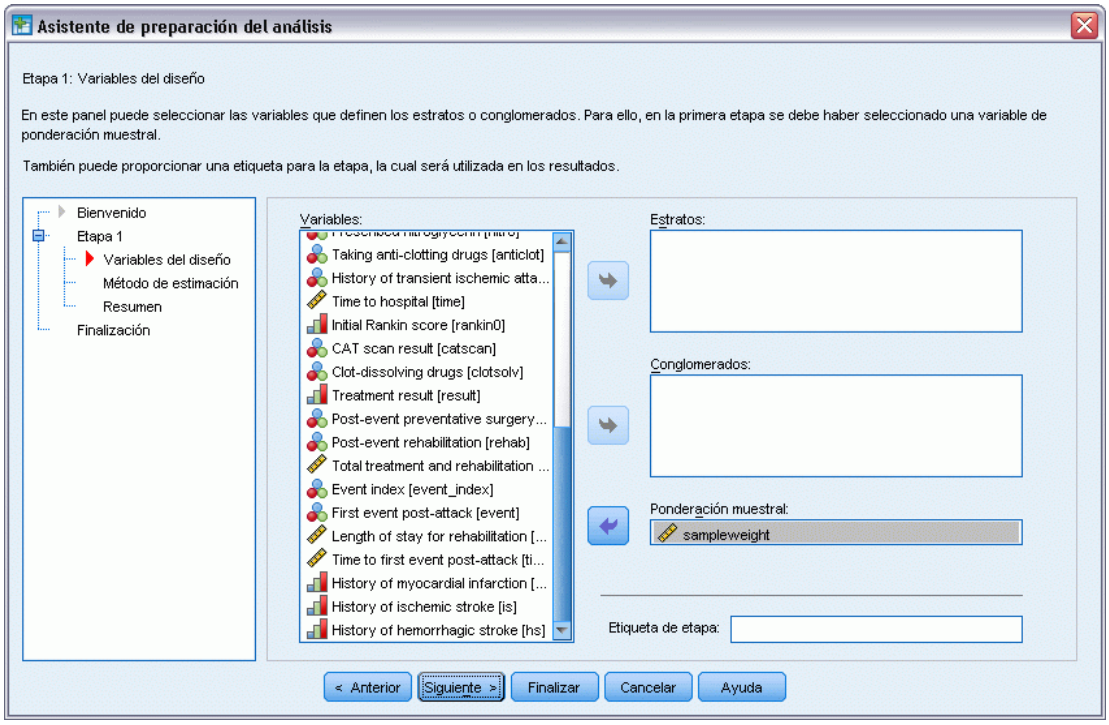
- Para crear el plan de análisis, elija en los menús:
  Analizar > Complex Samples > Preparar para el análisis...

Figura 22-38
Asistente de preparación del análisis: paso Bienvenida



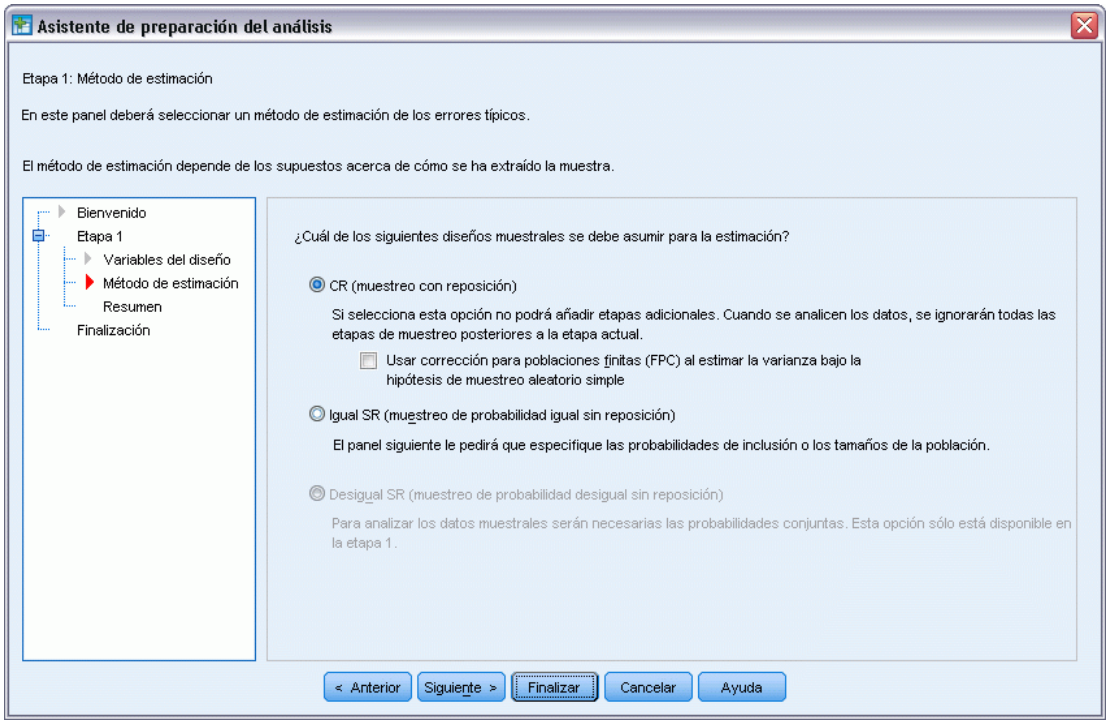
- Seleccione Crear un archivo de plan y escriba srs.csaplan como nombre del archivo. Si lo prefiere, también puede desplazarse hasta la ubicación en la que desea guardarlo.
- Pulse en Siguiente.

Figura 22-39
Asistente de preparación del análisis, Variables del diseño



- Seleccione *sampleweight* como variable de ponderación muestral.
- Pulse en Siguiente.

Figura 22-40
Asistente de preparación del análisis, Método de estimación



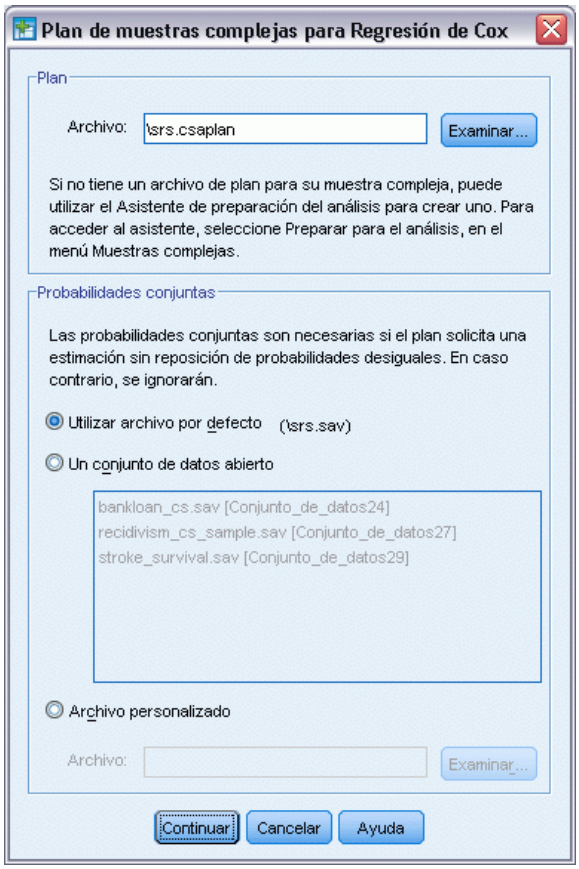
- Anule la selección de Usar corrección para poblaciones finitas.
- Pulse en Finalizar.

Ahora, ya está preparado para ejecutar el análisis.

Ejecución del análisis

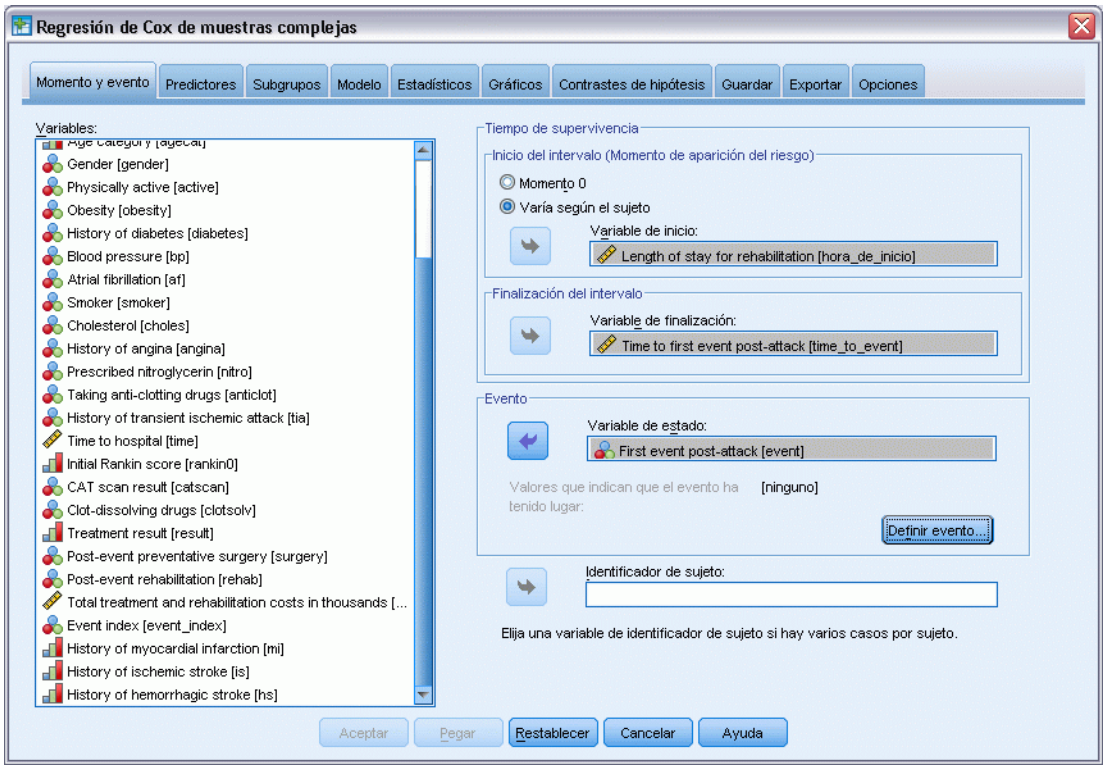
- Para ejecutar un análisis de regresión de Cox de muestras complejas, seleccione en los menús:
  Analizar > Complex Samples > Regresión de Cox...

Figura 22-41
Cuadro de diálogo Plan para Regresión de Cox



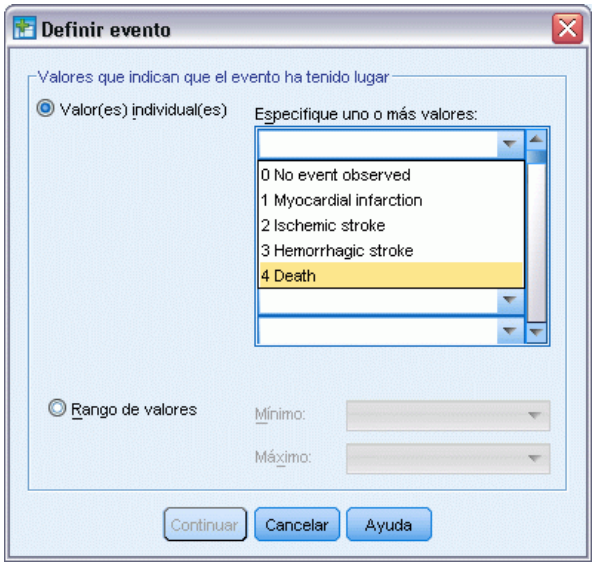
- Busque la ubicación en la que ha guardado el plan de análisis de muestreo simple (o vaya al directorio de archivos de ejemplo) y seleccione *srs.csaplan*.
- Pulse en Continuar.

Figura 22-42
Cuadro de diálogo Regresión de Cox, pestaña Momento y evento



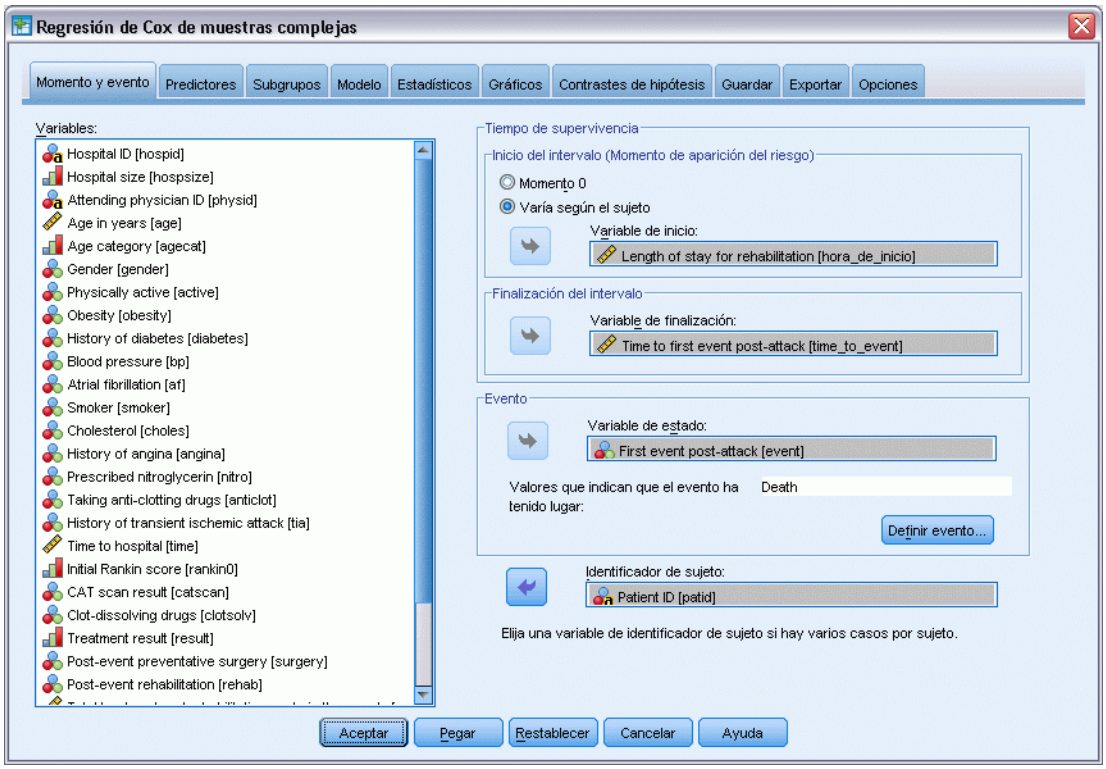
- Seleccione Varía según el sujeto y elija *Length of stay for rehabilitation [los_rehab]* como variable de inicio. Observe que la variable reestructurada ha tomado la etiqueta de variable de la primera variable utilizada para crearla, a pesar de que dicha etiqueta no es necesariamente adecuada para la variable creada.
- Seleccione *Time to first event post-attack [time_to_event]* como variable de finalización.
- Seleccione *First event post-attack [event]* como variable de estado.
- Pulse en Definir evento.

Figura 22-43
Cuadro de diálogo Definir evento



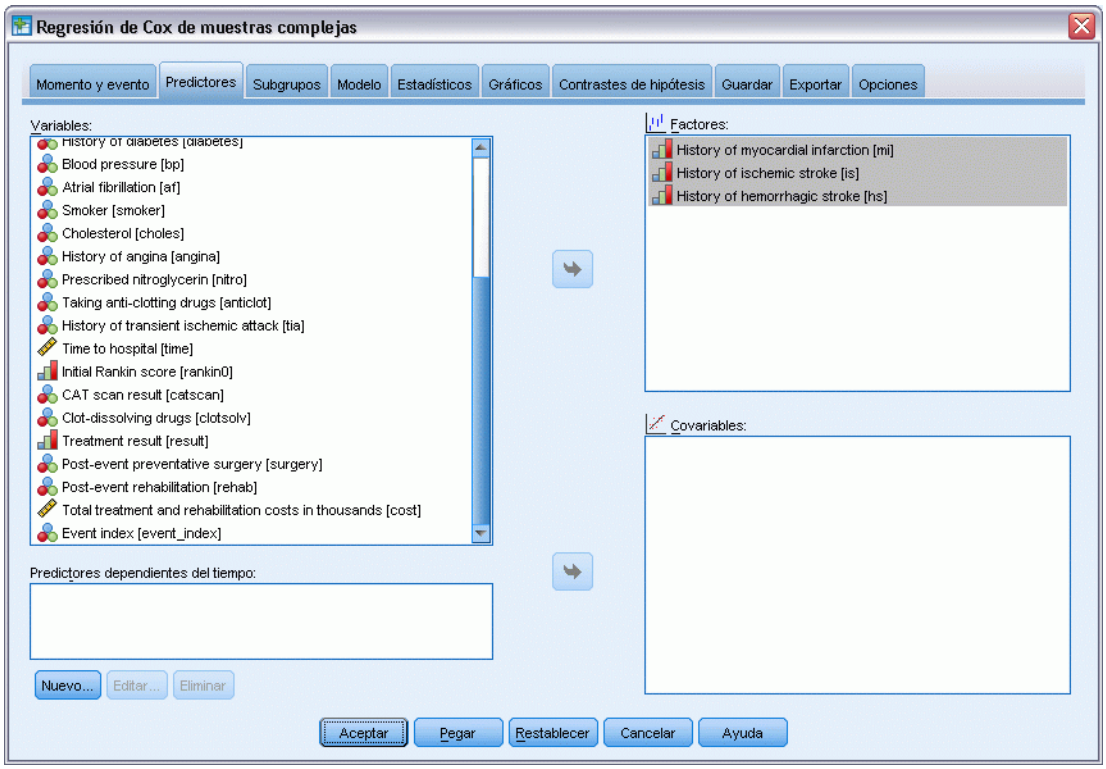
- Seleccione 4 Death como valor que indica que se ha producido el evento terminal.
- Pulse en Continuar.

Figura 22-44
Cuadro de diálogo Regresión de Cox, pestaña Momento y evento



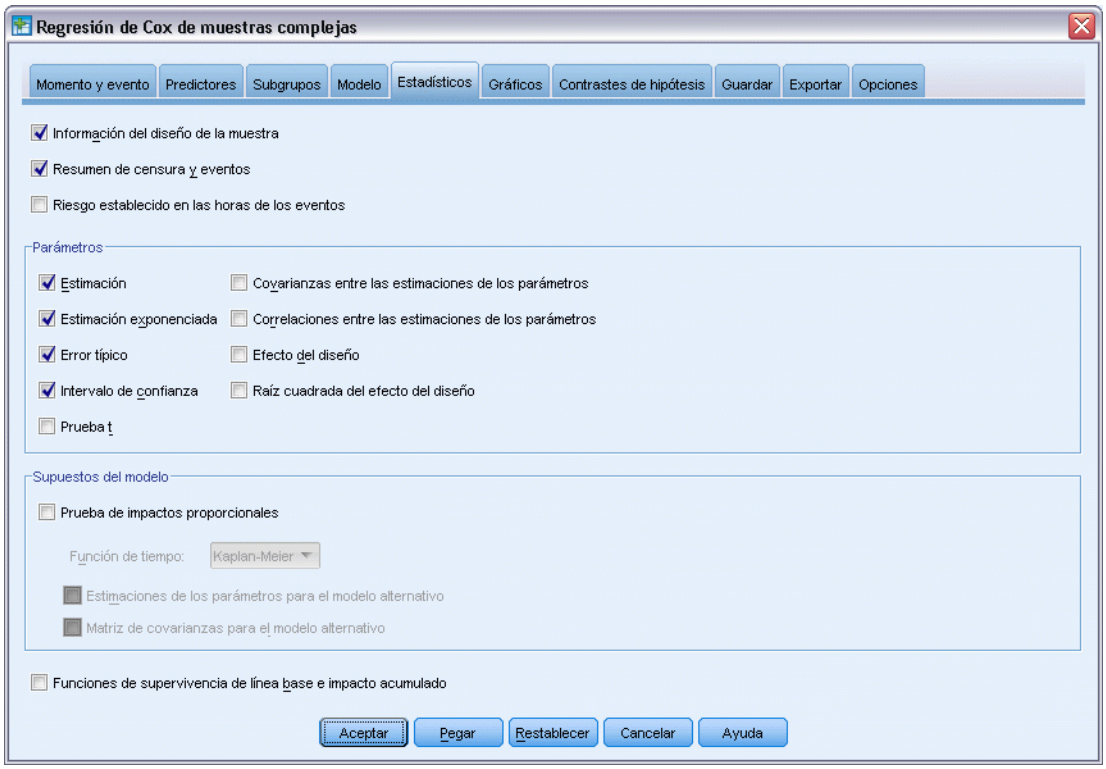
- Seleccione *Patient ID [patid]* como identificador del sujeto.
- Pulse en la pestaña Predictores.

Figura 22-45
Cuadro de diálogo Regresión de Cox, pestaña Predictores



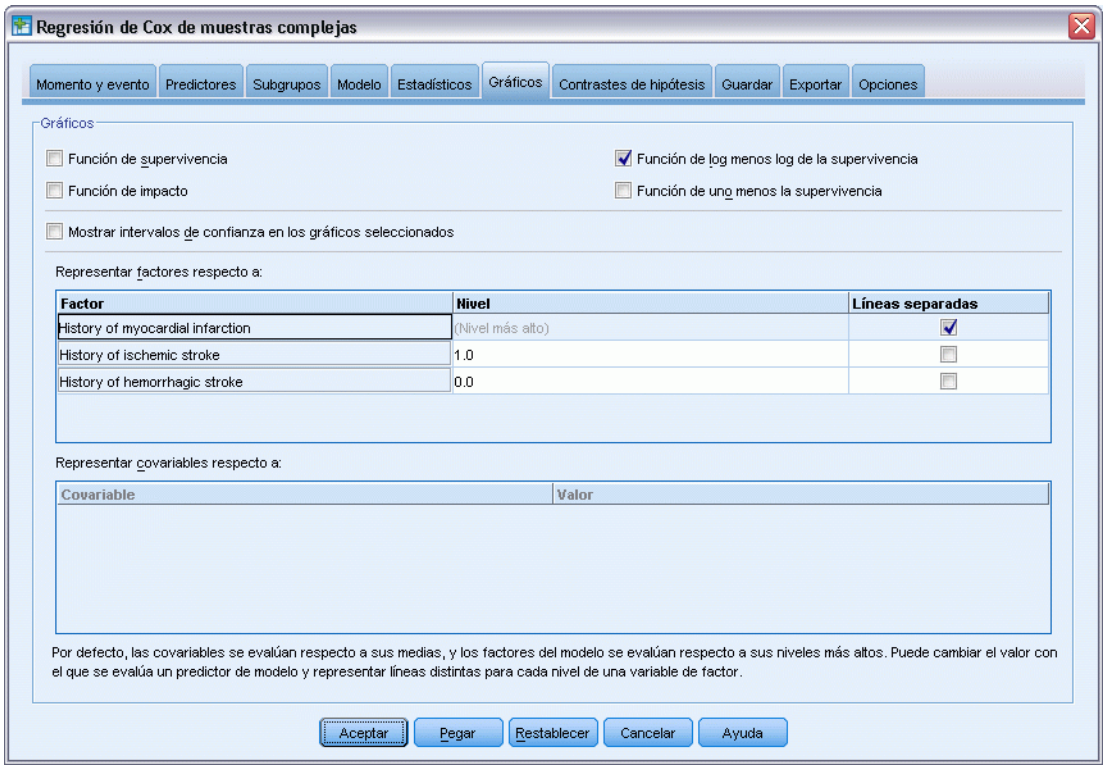
- Seleccione desde *History of myocardial infarction [mi]* hasta *History of hemorrhagic stroke [hs]* como factores.
- Pulse en la pestaña Estadísticos.

Figura 22-46
Cuadro de diálogo Regresión de Cox, pestaña Estadísticos



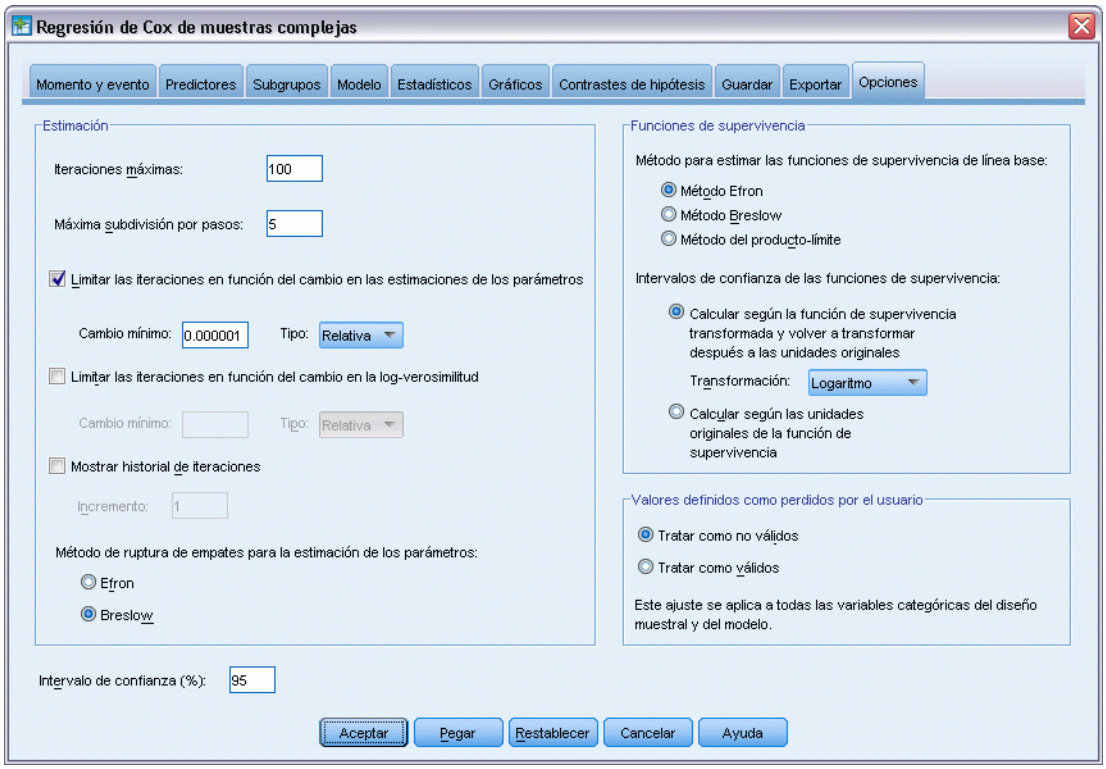
- Seleccione Estimación, Estimación exponenciada, Error típico e Intervalo de confianza en el grupo Parámetros.
- Pulse en la pestaña Gráficos.

Figura 22-47
Cuadro de diálogo Regresión de Cox, pestaña Estadísticos



- Seleccione Función de log menos log de la supervivencia.
- Active Líneas distintas para *History of myocardial infarction*.
- Seleccione 1,0 como nivel de *History of ischemic stroke*.
- Seleccione 0,0 como nivel de *History of hemorrhagic stroke*.
- Pulse en la pestaña Opciones.

Figura 22-48
Cuadro de diálogo Regresión de Cox, pestaña Opciones



- Seleccione Breslow como método de ruptura de empates en el grupo Estimación.
- Pulse en Aceptar.

Información de diseño de la muestra

Figura 22-49
Información del diseño muestral

			N
Recuentos no ponderados	Válidos	Sujetos	2421
		Casos	3310
	Casos no válidos		0
	Casos totales		3310
Tamaño de sujetos de la población			2421.000
Etapa 1	Estratos		1
	Unidades		2421
Grados de libertad del diseño del muestreo			2420

Esta tabla contiene información sobre el diseño muestral relevante para la estimación del modelo.

- Hay varios casos para algunos sujetos y se utilizan todos los 3.310 casos en el análisis.
- El diseño tiene un único estrato y 2.421 unidades (una para cada sujeto). Los grados de libertad del diseño muestral se estiman mediante 2421−1=2420.

Pruebas de efectos del modelo

Figura 22-50
Contrastes de los efectos del modelo

Origen	gl1	gl2	F de Wald	Sig.
mi	3.000	2418.000	452.873	.000
is	2.000	2419.000	1064.936	.000
hs	2.000	2419.000	739.197	.000

Variable de tiempo de supervivencia: Length of stay for rehabilitation, Time to first event post-attack
Variable de estrato de evento: First event post-attack = 4
Variable de ID de sujeto: Patient ID
Modelo: mi, is, hs

El valor de significación de todos los efectos es cercano a 0, lo que sugiere que todos ellos contribuyen al modelo.

Estimaciones de los parámetros

Figura 22-51
Estimaciones de los parámetros

Parámetro	B	Error típico	Intervalo de confianza al 95%		Exp(B)	95% Intervalo de confianza para Exp(B)	
			Inferior	Superior		Inferior	Superior
[mi=0]	-6.381	.283	-6.935	-5.827	.002	.001	.003
[mi=1]	-5.589	.284	-6.147	-5.032	.004	.002	.007
[mi=2]	-2.119	.344	-2.794	-1.445	.120	.061	.236
[mi=3]	.000[a]	.	.	.	1.000	.	.
[is=1]	-6.421	.202	-6.817	-6.024	.002	.001	.002
[is=2]	-2.803	.222	-3.239	-2.366	.061	.039	.094
[is=3]	.000[a]	.	.	.	1.000	.	.
[hs=0]	-6.148	.355	-6.844	-5.453	.002	.001	.004
[hs=1]	-2.232	.373	-2.963	-1.502	.107	.052	.223
[hs=2]	.000[a]	.	.	.	1.000	.	.

Variable de tiempo de supervivencia: Length of stay for rehabilitation, Time to first event post-attack
Variable de estrato de evento: First event post-attack = 4
Variable de ID de sujeto: Patient ID
Modelo: mi, is, hs

a. Establecido en cero porque este parámetro es redu

b. Método de ruptura de empates: Breslow

El procedimiento utiliza la última categoría de cada factor como la categoría de referencia y el efecto de las demás categorías es relativo a la categoría de referencia. Observe que mientras que la estimación es útil como contraste estadístico, la estimación exponenciada, Exp(B), puede interpretarse con mayor facilidad como el cambio pronosticado en el riesgo respecto a la categoría de referencia.

- El valor de Exp(B) de *[mi=0]* indica que el riesgo de muerte de un paciente que no ha sufrido ningún infarto de miocardio previo (mi) es 0,002 veces el de un paciente con tres infartos de miocardio previos.

- Los intervalos de confianza de *[mi=1]* y *[mi=0]* se solapan, lo que indica que el riesgo de un paciente con un único infarto de miocardio previo no puede distinguirse estadísticamente del de un paciente que no ha sufrido anteriormente ningún infarto de miocardio.
- Los intervalos de confianza de *[mi=0]* y *[mi=1]* no se solapan con el intervalo de *[mi=2]* y ninguno de ellos incluye 0. Por tanto, parece que el riesgo de los pacientes que no han sufrido ningún infarto de miocardio o sólo han sufrido uno puede distinguirse del riesgo de los pacientes que han sufrido previamente dos infartos de miocardio, lo que a su vez puede distinguirse del riesgo de los pacientes que han sufrido previamente tres infartos de miocardio.

También existen relaciones similares para los niveles de *is* y *hs*, en los que al aumentar el número de incidentes previos aumenta el riesgo de muerte.

Valores de patrón

Figura 22-52
Valores de los patrones

		Intervalo de tiempo de supervivencia				
		Inicio	Fin	History of myocardial infarction	History of ischemic stroke	History of hemorrhagic stroke
Patrón de referencia	1	.000		Three	Three	Two
Patrón 1.1	1	.000		None	One	None
Patrón 1.2	1	.000		One	One	None
Patrón 1.3	1	.000		Two	One	None
Patrón 1.4	1	.000		Three	One	None

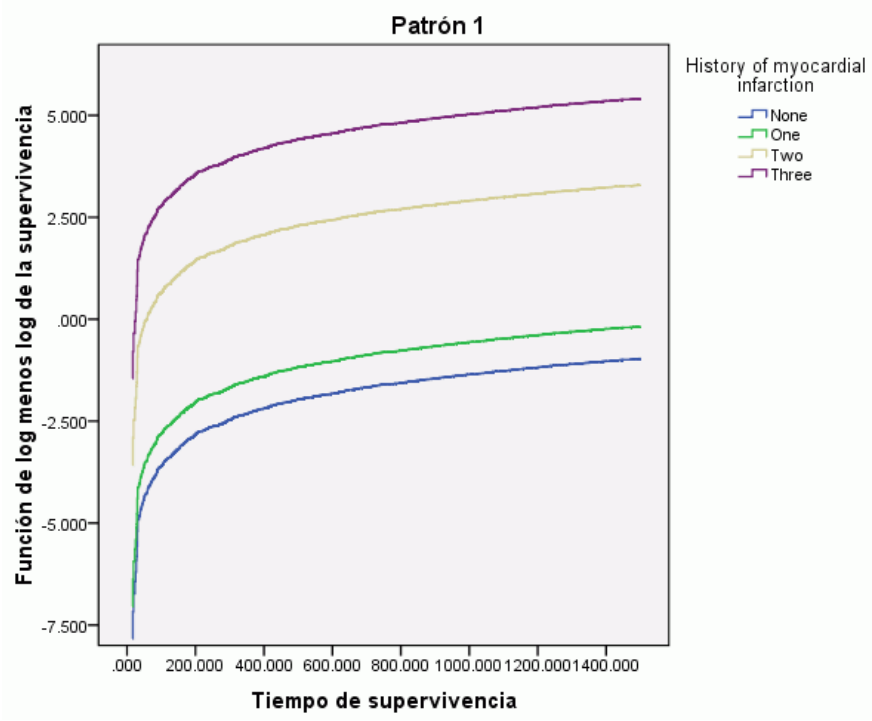
Se ha asignado a un predictor no especificado el valor de este predictor en el patrón de referencia .
Cada intervalo de tiempo de supervivencia se define como Inicio < Tiempo de supervivencia <= Fin.
Modelo: mi, is, hs.

La tabla de valores de los patrones muestra los valores que definen cada patrón de predictores. Además de los predictores del modelo, también se muestran los momentos de inicio y finalización del intervalo de supervivencia. Para los análisis ejecutados mediante los cuadros de diálogo, los momentos de inicio y finalización siempre serán 0 y sin límites, respectivamente; aunque mediante la sintaxis de comandos puede especificar rutas de predictores constantes por tramos.

- El patrón de referencia se establece en la categoría de referencia para cada factor y el valor medio de cada covariable (en este modelo no hay ninguna covariable). Para este conjunto de datos, la combinación de factores que se muestra para el modelo de referencia no se puede producir, por lo que ignoraremos el gráfico de log menos log del patrón de referencia.
- Los patrones del 1.1 al 1.4 se diferencian únicamente en el valor de *History of myocardial infarction*. Se crea un patrón distinto (y una línea distinta en el gráfico solicitado) para cada valor de *History of myocardial infarction* mientras que las demás variables permanecen constantes.

Gráfico de log menos log

Figura 22-53
Gráfico de log menos log



Este gráfico muestra el log menos log de la función de supervivencia, ln(−ln(supervivencia)), respecto al tiempo de supervivencia. Este gráfico concreto muestra una curva distinta para cada categoría de *History of myocardial infarction*, con *History of ischemic stroke* fijo en el valor *One* y *History of hemorrhagic stroke* fijo en el valor *None*. Este gráfico resulta útil como visualización del efecto de *History of myocardial infarction* de la función de supervivencia. Tal como hemos visto en la tabla de estimaciones de los parámetros, parece que la supervivencia de los pacientes que no han sufrido ningún infarto de miocardio o sólo han sufrido uno se puede distinguir de la supervivencia de los pacientes que han sufrido previamente dos infartos de miocardio, que a su vez puede distinguirse de la supervivencia de los pacientes que han sufrido previamente tres infartos de miocardio.

Resumen

Ha ajustado un modelo de regresión de Cox a la supervivencia tras un ataque que estima los efectos de los cambios registrados en el historial del paciente tras el ataque. Este análisis es únicamente un punto de partida, ya que sin duda los investigadores desearán incluir otros predictores potenciales en el modelo. Además, en posteriores análisis de este conjunto de datos tal vez quiera realizar cambios de mayor importancia a la estructura del modelo. Por ejemplo, el modelo actual supone que el efecto de un evento que afecta al historial del paciente puede ser cuantificado mediante un multiplicador del impacto basal. En su lugar, puede ser razonable

suponer que la forma del impacto basal resulta alterada por la ocurrencia de un evento diferente de la muerte. Para ello, podría estratificar el análisis basándose en *Event index*.

Apéndice

A

Archivos muestrales

Los archivos muestrales instalados con el producto se encuentran en el subdirectorio *Samples* del directorio de instalación. Hay una carpeta independiente dentro del subdirectorio Samples para cada uno de los siguientes idiomas: Inglés, francés, alemán, italiano, japonés, coreano, polaco, ruso, chino simplificado, español y chino tradicional.

No todos los archivos muestrales están disponibles en todos los idiomas. Si un archivo muestral no está disponible en un idioma, esa carpeta de idioma contendrá una versión en inglés del archivo muestral.

Descripciones

A continuación, se describen brevemente los archivos muestrales usados en varios ejemplos que aparecen a lo largo de la documentación.

- **accidents.sav.**Archivo de datos hipotéticos sobre una compañía de seguros que estudia los factores de riesgo de edad y género que influyen en los accidentes de automóviles de una región determinada. Cada caso corresponde a una clasificación cruzada de categoría de edad y género.
- **adl.sav.**Archivo de datos hipotéticos relativo a los esfuerzos para determinar las ventajas de un tipo propuesto de tratamiento para pacientes que han sufrido un derrame cerebral. Los médicos dividieron de manera aleatoria a pacientes (mujeres) que habían sufrido un derrame cerebral en dos grupos. El primer grupo recibió el tratamiento físico estándar y el segundo recibió un tratamiento emocional adicional. Tres meses después de los tratamientos, se puntuaron las capacidades de cada paciente para realizar actividades cotidianas como variables ordinales.
- **advert.sav.** Archivo de datos hipotéticos sobre las iniciativas de un minorista para examinar la relación entre el dinero invertido en publicidad y las ventas resultantes. Para ello, se recopilaron las cifras de ventas anteriores y los costes de publicidad asociados.
- **aflatoxin.sav.** Archivo de datos hipotéticos sobre las pruebas realizadas en las cosechas de maíz con relación a la aflatoxina, un veneno cuya concentración varía ampliamente en los rendimientos de cultivo y entre los mismos. Un procesador de grano ha recibido 16 muestras de cada uno de los 8 rendimientos de cultivo y ha medido los niveles de aflatoxinas en partes por millón (PPM).
- **aflatoxin20.sav.** Este archivo de datos contiene las medidas de aflatoxina de cada una de las 16 muestras de los rendimientos 4 y 8 procedentes del archivo de datos *aflatoxin.sav*.
- **anorectic.sav.**Mientras trabajaban en una sintomatología estandarizada del comportamiento anoréxico/bulímico, los investigadores (Van der Ham, Meulman, Van Strien, y Van Engeland, 1997) realizaron un estudio de 55 adolescentes con trastornos de la alimentación conocidos. Cada paciente fue examinado cuatro veces durante cuatro años, lo que representa un total

de 220 observaciones. En cada observación, se puntuó a los pacientes por cada uno de los 16 síntomas. Faltan las puntuaciones de los síntomas para el paciente 71 en el tiempo 2, el paciente 76 en el tiempo 2 y el paciente 47 en el tiempo 3, lo que nos deja 217 observaciones válidas.

- **autoaccidents.sav.** Archivo de datos hipotéticos sobre las iniciativas de un analista de seguros para elaborar un modelo del número de accidentes de automóvil por conductor teniendo en cuenta la edad y el género del conductor. Cada caso representa un conductor diferente y registra el sexo, la edad en años y el número de accidentes de automóvil del conductor en los últimos cinco años.
- **band.sav** Este archivo de datos contiene las cifras de ventas semanales hipotéticas de CD de música de una banda. También se incluyen datos para tres variables predictoras posibles.
- **bankloan.sav.**Archivo de datos hipotéticos sobre las iniciativas de un banco para reducir la tasa de moras de créditos. El archivo contiene información financiera y demográfica de 850 clientes anteriores y posibles clientes. Los primeros 700 casos son clientes a los que anteriormente se les ha concedido un préstamo. Al menos 150 casos son posibles clientes cuyos riesgos de crédito el banco necesita clasificar como positivos o negativos.
- **bankloan_binning.sav.** Archivo de datos hipotéticos que contiene información financiera y demográfica sobre 5.000 clientes anteriores.
- **behavior.sav.** En un ejemplo clásico (Price y Bouffard, 1974), se pidió a 52 estudiantes que valoraran las combinaciones de 15 situaciones y 15 comportamientos en una escala de 10 puntos que oscilaba entre 0 =“extremadamente apropiado” y 9=“extremadamente inapropiado”. Los valores promediados respecto a los individuos se toman como disimilaridades.
- **behavior_ini.sav.** Este archivo de datos contiene una configuración inicial para una solución bidimensional de *behavior.sav*.
- **brakes.sav.** Archivo de datos hipotéticos sobre el control de calidad de una fábrica que produce frenos de disco para automóviles de alto rendimiento. El archivo de datos contiene las medidas del diámetro de 16 discos de cada una de las 8 máquinas de producción. El diámetro objetivo para los frenos es de 322 milímetros.
- **breakfast.sav.**En un estudio clásico (Green y Rao, 1972), se pidió a 21 estudiantes de administración de empresas de la Wharton School y sus cónyuges que ordenaran 15 elementos de desayuno por orden de preferencia, de 1=“más preferido” a 15=“menos preferido”. Sus preferencias se registraron en seis escenarios distintos, de “Preferencia global” a “Aperitivo, con bebida sólo”.
- **breakfast-overall.sav.** Este archivo de datos sólo contiene las preferencias de elementos de desayuno para el primer escenario, “Preferencia global”.
- **broadband_1.sav** Archivo de datos hipotéticos que contiene el número de suscriptores, por región, a un servicio de banda ancha nacional. El archivo de datos contiene números de suscriptores mensuales para 85 regiones durante un período de cuatro años.
- **broadband_2.sav** Este archivo de datos es idéntico a *broadband_1.sav* pero contiene datos para tres meses adicionales.
- **car_insurance_claims.sav.** Un conjunto de datos presentados y analizados en otro lugar (McCullagh y Nelder, 1989) estudia las reclamaciones por daños en vehículos. La cantidad de reclamaciones media se puede modelar como si tuviera una distribución Gamma, mediante

una función de enlace inversa para relacionar la media de la variable dependiente con una combinación lineal de la edad del asegurado, el tipo de vehículo y la antigüedad del vehículo. El número de reclamaciones presentadas se puede utilizar como una ponderación de escalamiento.

- **car_sales.sav.** Este archivo de datos contiene estimaciones de ventas, precios de lista y especificaciones físicas hipotéticas de varias marcas y modelos de vehículos. Los precios de lista y las especificaciones físicas se han obtenido de *edmunds.com* y de sitios de fabricantes.
- **car_sales_uprepared.sav.**Ésta es una versión modificada de *car_sales.sav* que no incluye ninguna versión transformada de los campos.
- **carpet.sav** En un ejemplo muy conocido (Green y Wind, 1973), una compañía interesada en sacar al mercado un nuevo limpiador de alfombras desea examinar la influencia de cinco factores sobre la preferencia del consumidor: diseño del producto, marca comercial, precio, sello de *buen producto para el hogar* y garantía de devolución del importe. Hay tres niveles de factores para el diseño del producto, cada uno con una diferente colocación del cepillo del aplicador; tres nombres comerciales (*K2R*, *Glory* y *Bissell*); tres niveles de precios; y dos niveles (no o sí) para los dos últimos factores. Diez consumidores clasificaron 22 perfiles definidos por estos factores. La variable *Preferencia* contiene el rango de las clasificaciones medias de cada perfil. Las clasificaciones inferiores corresponden a preferencias elevadas. Esta variable refleja una medida global de la preferencia de cada perfil.
- **carpet_prefs.sav** Este archivo de datos se basa en el mismo ejemplo que el descrito para *carpet.sav*, pero contiene las clasificaciones reales recogidas de cada uno de los 10 consumidores. Se pidió a los consumidores que clasificaran los 22 perfiles de los productos empezando por el menos preferido. Las variables desde *PREF1* hasta *PREF22* contienen los ID de los perfiles asociados, como se definen en *carpet_plan.sav*.
- **catalog.sav**Este archivo de datos contiene cifras de ventas mensuales hipotéticas de tres productos vendidos por una compañía de venta por catálogo. También se incluyen datos para cinco variables predictoras posibles.
- **catalog_seasfac.sav**Este archivo de datos es igual que *catalog.sav*, con la excepción de que incluye un conjunto de factores estacionales calculados a partir del procedimiento Descomposición estacional junto con las variables de fecha que lo acompañan.
- **cellular.sav.** Archivo de datos hipotéticos sobre las iniciativas de una compañía de telefonía móvil para reducir el abandono de clientes. Las puntuaciones de propensión al abandono de clientes se aplican a las cuentas, oscilando de 0 a 100. Las cuentas con una puntuación de 50 o superior pueden estar buscando otros proveedores.
- **ceramics.sav.**Archivo de datos hipotéticos sobre las iniciativas de un fabricante para determinar si una nueva aleación de calidad tiene una mayor resistencia al calor que una aleación estándar. Cada caso representa una prueba independiente de una de las aleaciones; la temperatura a la que registró el fallo del rodamiento.
- **cereal.sav.** Archivo de datos hipotéticos sobre una encuesta realizada a 880 personas sobre sus preferencias en el desayuno, teniendo también en cuenta su edad, sexo, estado civil y si tienen un estilo de vida activo o no (en función de si practican ejercicio al menos dos veces a la semana). Cada caso representa un encuestado diferente.
- **clothing_defects.sav.** Archivo de datos hipotéticos sobre el proceso de control de calidad en una fábrica de prendas. Los inspectores toman una muestra de prendas de cada lote producido en la fábrica, y cuentan el número de prendas que no son aceptables.

- **coffee.sav.** Este archivo de datos pertenece a las imágenes percibidas de seis marcas de café helado (Kennedy, Riquier, y Sharp, 1996). Para cada uno de los 23 atributos de imagen de café helado, los encuestados seleccionaron todas las marcas que quedaban descritas por el atributo. Las seis marcas se denotan AA, BB, CC, DD, EE y FF para mantener la confidencialidad.
- **contacts.sav.**Archivo de datos hipotéticos sobre las listas de contactos de un grupo de representantes de ventas de ordenadores de empresa. Cada uno de los contactos está categorizado por el departamento de la compañía en el que trabaja y su categoría en la compañía. Además, también se registran los importes de la última venta realizada, el tiempo transcurrido desde la última venta y el tamaño de la compañía del contacto.
- **creditpromo.sav.** Archivo de datos hipotéticos sobre las iniciativas de unos almacenes para evaluar la eficacia de una promoción de tarjetas de crédito reciente. Para este fin, se seleccionaron aleatoriamente 500 titulares. La mitad recibieron un anuncio promocionando una tasa de interés reducida sobre las ventas realizadas en los siguientes tres meses. La otra mitad recibió un anuncio estacional estándar.
- **customer_dbase.sav.** Archivo de datos hipotéticos sobre las iniciativas de una compañía para usar la información de su almacén de datos para realizar ofertas especiales a los clientes con más probabilidades de responder. Se seleccionó un subconjunto de la base de clientes aleatoriamente a quienes se ofrecieron las ofertas especiales y sus respuestas se registraron.
- **customer_information.sav.** Archivo de datos hipotéticos que contiene la información de correo del cliente, como el nombre y la dirección.
- **customer_subset.sav.** Un subconjunto de 80 casos de *customer_dbase.sav*.
- **customers_model.sav.** Este archivo contiene datos hipotéticos sobre los individuos a los que va dirigida una campaña de marketing. Estos datos incluyen información demográfica, un resumen del historial de compras y si cada individuo respondió a la campaña. Cada caso representa un individuo diferente.
- **customers_new.sav.** Este archivo contiene datos hipotéticos sobre los individuos que son candidatos potenciales para una campaña de marketing. Estos datos incluyen información demográfica y un resumen del historial de compras de cada individuo. Cada caso representa un individuo diferente.
- **debate.sav.** Archivos de datos hipotéticos sobre las respuestas emparejadas de una encuesta realizada a los asistentes a un debate político antes y después del debate. Cada caso corresponde a un encuestado diferente.
- **debate_aggregate.sav.** Archivo de datos hipotéticos que agrega las respuestas de *debate.sav*. Cada caso corresponde a una clasificación cruzada de preferencias antes y después del debate.
- **demo.sav.** Archivos de datos hipotéticos sobre una base de datos de clientes adquirida con el fin de enviar por correo ofertas mensuales. Se registra si el cliente respondió a la oferta, junto con información demográfica diversa.
- **demo_cs_1.sav.**Archivo de datos hipotéticos sobre el primer paso de las iniciativas de una compañía para recopilar una base de datos de información de encuestas. Cada caso corresponde a una ciudad diferente, y se registra la identificación de la ciudad, la región, la provincia y el distrito.
- **demo_cs_2.sav.**Archivo de datos hipotéticos sobre el segundo paso de las iniciativas de una compañía para recopilar una base de datos de información de encuestas. Cada caso corresponde a una unidad familiar diferente de las ciudades seleccionadas en el primer paso, y

se registra la identificación de la unidad, la subdivisión, la ciudad, el distrito, la provincia y la región. También se incluye la información de muestreo de las primeras dos etapas del diseño.

- **demo_cs.sav.**Archivo de datos hipotéticos que contiene información de encuestas recopilada mediante un diseño de muestreo complejo. Cada caso corresponde a una unidad familiar distinta, y se recopila información demográfica y de muestreo diversa.
- **dmdata.sav.** Éste es un archivo de datos hipotéticos que contiene información demográfica y de compras para una empresa de marketing directo. *dmdata2.sav* contiene información para un subconjunto de contactos que recibió un envío de prueba, y *dmdata3.sav* contiene información sobre el resto de contactos que no recibieron el envío de prueba.
- **dietstudy.sav.**Este archivo de datos hipotéticos contiene los resultados de un estudio sobre la "dieta Stillman" (Rickman, Mitchell, Dingman, y Dalen, 1974). Cada caso corresponde a un sujeto distinto y registra sus pesos antes y después de la dieta en libras y niveles de triglicéridos en mg/100 ml.
- **dvdplayer.sav.** Archivo de datos hipotéticos sobre el desarrollo de un nuevo reproductor de DVD. El equipo de marketing ha recopilado datos de grupo de enfoque mediante un prototipo. Cada caso corresponde a un usuario encuestado diferente y registra información demográfica sobre los encuestados y sus respuestas a preguntas acerca del prototipo.
- **german_credit.sav.**Este archivo de datos se toma del conjunto de datos "German credit" de las Repository of Machine Learning Databases (Blake y Merz, 1998) de la Universidad de California, Irvine.
- **grocery_1month.sav.** Este archivo de datos hipotéticos es el archivo de datos *grocery_coupons.sav* con las compras semanales "acumuladas" para que cada caso corresponda a un cliente diferente. Algunas de las variables que cambiaban semanalmente desaparecen de los resultados, y la cantidad gastada registrada se convierte ahora en la suma de las cantidades gastadas durante las cuatro semanas del estudio.
- **grocery_coupons.sav.** Archivo de datos hipotéticos que contiene datos de encuestas recopilados por una cadena de tiendas de alimentación interesada en los hábitos de compra de sus clientes. Se sigue a cada cliente durante cuatro semanas, y cada caso corresponde a un cliente-semana distinto y registra información sobre dónde y cómo compran los clientes, incluida la cantidad que invierten en comestibles durante esa semana.
- **guttman.sav.**Bell (Bell, 1961) presentó una tabla para ilustrar posibles grupos sociales. Guttman (Guttman, 1968) utilizó parte de esta tabla, en la que se cruzaron cinco variables que describían elementos como la interacción social, sentimientos de pertenencia a un grupo, proximidad física de los miembros y grado de formalización de la relación con siete grupos sociales teóricos, incluidos multitudes (por ejemplo, las personas que acuden a un partido de fútbol), espectadores (por ejemplo, las personas que acuden a un teatro o de una conferencia), públicos (por ejemplo, los lectores de periódicos o los espectadores de televisión), muchedumbres (como una multitud pero con una interacción mucho más intensa), grupos primarios (íntimos), grupos secundarios (voluntarios) y la comunidad moderna (confederación débil que resulta de la proximidad cercana física y de la necesidad de servicios especializados).
- **health_funding.sav.** Archivo de datos hipotéticos que contiene datos sobre inversión en sanidad (cantidad por 100 personas), tasas de enfermedad (índice por 10.000 personas) y visitas a centros de salud (índice por 10.000 personas). Cada caso representa una ciudad diferente.

- **hivassay.sav.** Archivo de datos hipotéticos sobre las iniciativas de un laboratorio farmacéutico para desarrollar un ensayo rápido para detectar la infección por VIH. Los resultados del ensayo son ocho tonos de rojo con diferentes intensidades, donde los tonos más oscuros indican una mayor probabilidad de infección. Se llevó a cabo una prueba de laboratorio de 2.000 muestras de sangre, de las cuales una mitad estaba infectada con el VIH y la otra estaba limpia.
- **hourlywagedata.sav.** Archivo de datos hipotéticos sobre los salarios por horas de enfermeras de puestos de oficina y hospitales y con niveles distintos de experiencia.
- **insurance_claims.sav.** Éste es un archivo de datos hipotéticos sobre una compañía de seguros que desee generar un modelo para etiquetar las reclamaciones sospechosas y potencialmente fraudulentas. Cada caso representa una reclamación diferente.
- **insure.sav.** Archivo de datos hipotéticos sobre una compañía de seguros que estudia los factores de riesgo que indican si un cliente tendrá que hacer una reclamación a lo largo de un contrato de seguro de vida de 10 años. Cada caso del archivo de datos representa un par de contratos (de los que uno registró una reclamación y el otro no), agrupados por edad y sexo.
- **judges.sav.** Archivo de datos hipotéticos sobre las puntuaciones concedidas por jueces cualificados (y un aficionado) a 300 actuaciones gimnásticas. Cada fila representa una actuación diferente; los jueces vieron las mismas actuaciones.
- **kinship_dat.sav.** Rosenberg y Kim (Rosenberg y Kim, 1975) comenzaron a analizar 15 términos de parentesco [tía, hermano, primos, hija, padre, nieta, abuelo, abuela, nieto, madre, sobrino, sobrina, hermana, hijo, tío]. Le pidieron a cuatro grupos de estudiantes universitarios (dos masculinos y dos femeninos) que ordenaran estos grupos según las similitudes. A dos grupos (uno masculino y otro femenino) se les pidió que realizaran la ordenación dos veces, pero que la segunda ordenación la hicieran según criterios distintos a los de la primera. Así, se obtuvo un total de seis "fuentes". Cada fuente se corresponde con una matriz de proximidades de 15×15 cuyas casillas son iguales al número de personas de una fuente menos el número de veces que se particionaron los objetos en esa fuente.
- **kinship_ini.sav.** Este archivo de datos contiene una configuración inicial para una solución tridimensional de *kinship_dat.sav*.
- **kinship_var.sav.** Este archivo de datos contiene variables independientes *sexo*, *gener*(ación), y *grado* (de separación) que se pueden usar para interpretar las dimensiones de una solución para *kinship_dat.sav*. Concretamente, se pueden usar para restringir el espacio de la solución a una combinación lineal de estas variables.
- **marketvalues.sav.** Archivo de datos sobre las ventas de casas en una nueva urbanización de Algonquin, Ill., durante los años 1999 y 2000. Los datos de estas ventas son públicos.
- **nhis2000_subset.sav.** La National Health Interview Survey (NHIS, encuesta del Centro Nacional de Estadísticas de Salud de EE.UU.) es una encuesta detallada realizada entre la población civil de Estados Unidos. Las encuestas se realizaron en persona a una muestra representativa de las unidades familiares del país. Se recogió tanto la información demográfica como las observaciones acerca del estado y los hábitos de salud de los integrantes de cada unidad familiar. Este archivo de datos contiene un subconjunto de información de la encuesta de 2000. National Center for Health Statistics. National Health Interview Survey, 2000. Archivo de datos y documentación de uso público. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Fecha de acceso: 2003.

- **ozono.sav.** Los datos incluyen 330 observaciones de seis variables meteorológicas para pronosticar la concentración de ozono a partir del resto de variables. Los investigadores anteriores(Breiman y Friedman, 1985), (Hastie y Tibshirani, 1990) han encontrado que no hay linealidad entre estas variables, lo que dificulta los métodos de regresión típica.
- **pain_medication.sav.** Este archivo de datos hipotéticos contiene los resultados de una prueba clínica sobre medicación antiinflamatoria para tratar el dolor artrítico crónico. Resulta de particular interés el tiempo que tarda el fármaco en hacer efecto y cómo se compara con una medicación existente.
- **patient_los.sav.** Este archivo de datos hipotéticos contiene los registros de tratamiento de pacientes que fueron admitidos en el hospital ante la posibilidad de sufrir un infarto de miocardio (IM o "ataque al corazón"). Cada caso corresponde a un paciente distinto y registra diversas variables relacionadas con su estancia hospitalaria.
- **patlos_sample.sav.** Este archivo de datos hipotéticos contiene los registros de tratamiento de una muestra de pacientes que recibieron trombolíticos durante el tratamiento del infarto de miocardio (IM o "ataque al corazón"). Cada caso corresponde a un paciente distinto y registra diversas variables relacionadas con su estancia hospitalaria.
- **polishing.sav.** Archivo de datos "Nambeware Polishing Times" (Tiempo de pulido de metal) de la biblioteca de datos e historiales. Contiene datos sobre las iniciativas de un fabricante de cuberterías de metal (Nambe Mills, Santa Fe, N. M.) para planificar su programa de producción. Cada caso representa un artículo distinto de la línea de productos. Se registra el diámetro, el tiempo de pulido, el precio y el tipo de producto de cada artículo.
- **poll_cs.sav.** Archivo de datos hipotéticos sobre las iniciativas de los encuestadores para determinar el nivel de apoyo público a una ley antes de una asamblea legislativa. Los casos corresponden a votantes registrados. Cada caso registra el condado, la población y el vecindario en el que vive el votante.
- **poll_cs_sample.sav.** Este archivo de datos hipotéticos contiene una muestra de los votantes enumerados en *poll_cs.sav*. La muestra se tomó según el diseño especificado en el archivo de plan *poll.csplan* y este archivo de datos registra las probabilidades de inclusión y las ponderaciones muestrales. Sin embargo, tenga en cuenta que debido a que el plan muestral hace uso de un método de probabilidad proporcional al tamaño (PPS), también existe un archivo que contiene las probabilidades de selección conjunta (*poll_jointprob.sav*). Las variables adicionales que corresponden a los datos demográficos de los votantes y sus opiniones sobre la propuesta de ley se recopilaron y añadieron al archivo de datos después de tomar la muestra.
- **property_assess.sav.** Archivo de datos hipotéticos sobre las iniciativas de un asesor del condado para mantener actualizada la evaluación de los valores de las propiedades utilizando recursos limitados. Los casos corresponden a las propiedades vendidas en el condado el año anterior. Cada caso del archivo de datos registra la población en que se encuentra la propiedad, el último asesor que visitó la propiedad, el tiempo transcurrido desde la última evaluación, la valoración realizada en ese momento y el valor de venta de la propiedad.
- **property_assess_cs.sav.** Archivo de datos hipotéticos sobre las iniciativas de un asesor de un estado para mantener actualizada la evaluación de los valores de las propiedades utilizando recursos limitados. Los casos corresponden a propiedades del estado. Cada caso del archivo de datos registra el condado, la población y el vecindario en el que se encuentra la propiedad, el tiempo transcurrido desde la última evaluación y la valoración realizada en ese momento.

- **property_assess_cs_sample.sav**Este archivo de datos hipotéticos contiene una muestra de las propiedades recogidas en *property_assess_cs.sav*. La muestra se tomó en función del diseño especificado en el archivo de plan *property_assess.csplan*, y este archivo de datos registra las probabilidades de inclusión y las ponderaciones muestrales. La variable adicional *Valor actual* se recopiló y añadió al archivo de datos después de tomar la muestra.
- **recidivism.sav.** Archivo de datos hipotéticos sobre las iniciativas de una agencia de orden público para comprender los índices de reincidencia en su área de jurisdicción. Cada caso corresponde a un infractor anterior y registra su información demográfica, algunos detalles de su primer delito y, a continuación, el tiempo transcurrido desde su segundo arresto, si ocurrió en los dos años posteriores al primer arresto.
- **recidivism_cs_sample.sav.** Archivo de datos hipotéticos sobre las iniciativas de una agencia de orden público para comprender los índices de reincidencia en su área de jurisdicción. Cada caso corresponde a un delincuente anterior, puesto en libertad tras su primer arresto durante el mes de junio de 2003 y registra su información demográfica, algunos detalles de su primer delito y los datos de su segundo arresto, si se produjo antes de finales de junio de 2006. Los delincuentes se seleccionaron de una muestra de departamentos según el plan de muestreo especificado en *recidivism_cs.csplan*. Como este plan utiliza un método de probabilidad proporcional al tamaño (PPS), también existe un archivo que contiene las probabilidades de selección conjunta (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.** Archivo de datos hipotéticos que contiene datos de transacciones de compra, incluida la fecha de compra, los artículos adquiridos y el importe de cada transacción.
- **salesperformance.sav.** Archivo de datos hipotéticos sobre la evaluación de dos nuevos cursos de formación de ventas. Sesenta empleados, divididos en tres grupos, reciben formación estándar. Además, el grupo 2 recibe formación técnica; el grupo 3, un tutorial práctico. Cada empleado se sometió a un examen al final del curso de formación y se registró su puntuación. Cada caso del archivo de datos representa a un alumno distinto y registra el grupo al que fue asignado y la puntuación que obtuvo en el examen.
- **satisf.sav.** Archivo de datos hipotéticos sobre una encuesta de satisfacción llevada a cabo por una empresa minorista en cuatro tiendas. Se encuestó a 582 clientes en total y cada caso representa las respuestas de un único cliente.
- **screws.sav** Este archivo de datos contiene información acerca de las características de tornillos, pernos, clavos y tacos (Hartigan, 1975).
- **shampoo_ph.sav.**Archivo de datos hipotéticos sobre el control de calidad en una fábrica de productos para el cabello. Se midieron seis lotes de resultados distintos en intervalos regulares y se registró su pH. El intervalo objetivo es de 4,5 a 5,5.
- **ships.sav.** Un conjunto de datos presentados y analizados en otro lugar (McCullagh et al., 1989) sobre los daños en los cargueros producidos por las olas. Los recuentos de incidentes se pueden modelar como si ocurrieran con una tasa de Poisson dado el tipo de barco, el período de construcción y el período de servicio. Los meses de servicio agregados para cada casilla de la tabla formados por la clasificación cruzada de factores proporcionan valores para la exposición al riesgo.
- **site.sav.**Archivo de datos hipotéticos sobre las iniciativas de una compañía para seleccionar sitios nuevos para sus negocios en expansión. Se ha contratado a dos consultores para evaluar los sitios de forma independiente, quienes, además de un informe completo, han resumido cada sitio como una posibilidad “buena”, “media” o “baja”.

- **smokers.sav.**Este archivo de datos es un resumen de la encuesta sobre toxicomanía 1998 National Household Survey of Drug Abuse y es una muestra de probabilidad de unidades familiares americanas. (*http://dx.doi.org/10.3886/ICPSR02934*) Así, el primer paso de un análisis de este archivo de datos debe ser ponderar los datos para reflejar las tendencias de población.
- **stroke_clean.sav.** Este archivo de datos hipotéticos contiene el estado de una base de datos médica después de haberla limpiado mediante los procedimientos de la opción Preparación de datos.
- **stroke_invalid.sav.** Este archivo de datos hipotéticos contiene el estado inicial de una base de datos médica que incluye contiene varios errores de entrada de datos.
- **stroke_survival.** Este archivo de datos hipotéticos registra los tiempos de supervivencia de los pacientes que finalizan un programa de rehabilitación tras un ataque isquémico. Tras el ataque, la ocurrencia de infarto de miocardio, ataque isquémico o ataque hemorrágico se anotan junto con el momento en el que se produce el evento registrado. La muestra está truncada a la izquierda ya que únicamente incluye a los pacientes que han sobrevivido al final del programa de rehabilitación administrado tras el ataque.
- **stroke_valid.sav.** Este archivo de datos hipotéticos contiene el estado de una base de datos médica después de haber comprobado los valores mediante el procedimiento Validar datos. Sigue conteniendo casos potencialmente anómalos.
- **survey_sample.sav.** Este archivo de datos contiene datos de encuestas, incluyendo datos demográficos y diferentes medidas de actitud. Se basa en un subconjunto de variables de NORC General Social Survey de 1998, aunque algunos valores de datos se han modificado y que existen variables ficticias adicionales se han añadido para demostraciones.
- **telco.sav.** Archivo de datos hipotéticos sobre las iniciativas de una compañía de telecomunicaciones para reducir el abandono de clientes en su base de clientes. Cada caso corresponde a un cliente distinto y registra diversa información demográfica y de uso del servicio.
- **telco_extra.sav.** Este archivo de datos es similar al archivo de datos *telco.sav*, pero las variables de meses con servicio y gasto de clientes transformadas logarítmicamente se han eliminado y sustituido por variables de gasto del cliente transformadas logarítmicamente tipificadas.
- **telco_missing.sav.** Este archivo de datos es un subconjunto del archivo de datos *telco.sav*, pero algunos valores de datos demográficos se han sustituido con valores perdidos.
- **testmarket.sav.** Archivo de datos hipotéticos sobre los planes de una cadena de comida rápida para añadir un nuevo artículo a su menú. Hay tres campañas posibles para promocionar el nuevo producto, por lo que el artículo se presenta en ubicaciones de varios mercados seleccionados aleatoriamente. Se utiliza una promoción diferente en cada ubicación y se registran las ventas semanales del nuevo artículo durante las primeras cuatro semanas. Cada caso corresponde a una ubicación semanal diferente.
- **testmarket_1month.sav.** Este archivo de datos hipotéticos es el archivo de datos *testmarket.sav* con las ventas semanales "acumuladas" para que cada caso corresponda a una ubicación diferente. Como resultado, algunas de las variables que cambiaban semanalmente desaparecen y las ventas registradas se convierten en la suma de las ventas realizadas durante las cuatro semanas del estudio.
- **tree_car.sav.** Archivo de datos hipotéticos que contiene datos demográficos y de precios de compra de vehículos.

- **tree_credit.sav** Archivo de datos hipotéticos que contiene datos demográficos y de historial de créditos bancarios.
- **tree_missing_data.sav** Archivo de datos hipotéticos que contiene datos demográficos y de historial de créditos bancarios con un elevado número de valores perdidos.
- **tree_score_car.sav.** Archivo de datos hipotéticos que contiene datos demográficos y de precios de compra de vehículos.
- **tree_textdata.sav.** Archivo de datos sencillos con dos variables diseñadas principalmente para mostrar el estado por defecto de las variables antes de realizar la asignación de nivel de medida y etiquetas de valor.
- **tv-survey.sav.** Archivo de datos hipotéticos sobre una encuesta dirigida por un estudio de TV que está considerando la posibilidad de ampliar la emisión de un programa de éxito. Se preguntó a 906 encuestados si verían el programa en distintas condiciones. Cada fila representa un encuestado diferente; cada columna es una condición diferente.
- **ulcer_recurrence.sav.** Este archivo contiene información parcial de un estudio diseñado para comparar la eficacia de dos tratamientos para prevenir la reaparición de úlceras. Constituye un buen ejemplo de datos censurados por intervalos y se ha presentado y analizado en otro lugar (Collett, 2003).
- **ulcer_recurrence_recoded.sav.** Este archivo reorganiza la información de *ulcer_recurrence.sav* para permitir modelar la probabilidad de eventos de cada intervalo del estudio en lugar de sólo la probabilidad de eventos al final del estudio. Se ha presentado y analizado en otro lugar (Collett et al., 2003).
- **verd1985.sav.** Archivo de datos sobre una encuesta (Verdegaal, 1985). Se han registrado las respuestas de 15 sujetos a 8 variables. Se han dividido las variables de interés en tres grupos. El conjunto 1 incluye *edad* y *ecivil*, el conjunto 2 incluye *mascota* y *noticia*, mientras que el conjunto 3 incluye *música* y *vivir*. Se escala *mascota* como nominal múltiple y *edad* como ordinal; el resto de variables se escalan como nominal simple.
- **virus.sav.**Archivo de datos hipotéticos sobre las iniciativas de un proveedor de servicios de Internet (ISP) para determinar los efectos de un virus en sus redes. Se ha realizado un seguimiento (aproximado) del porcentaje de tráfico de correos electrónicos infectados en sus redes a lo largo del tiempo, desde el momento en que se descubre hasta que la amenaza se contiene.
- **wheeze_steubenville.sav.** Subconjunto de un estudio longitudinal de los efectos sobre la salud de la polución del aire en los niños (Ware, Dockery, Spiro III, Speizer, y Ferris Jr., 1984). Los datos contienen medidas binarias repetidas del estado de las sibilancias en niños de Steubenville, Ohio, con edades de 7, 8, 9 y 10 años, junto con un registro fijo de si la madre era fumadora durante el primer año del estudio.
- **workprog.sav.** Archivo de datos hipotéticos sobre un programa de obras del gobierno que intenta colocar a personas desfavorecidas en mejores trabajos. Se siguió una muestra de participantes potenciales del programa, algunos de los cuales se seleccionaron aleatoriamente para entrar en el programa, mientras que otros no siguieron esta selección aleatoria. Cada caso representa un participante del programa diferente.

Apéndice
B

Notices

Patent No. 7,023,453

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: SPSS INC., AN IBM COMPANY, PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. SPSS Inc. may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-SPSS and non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this SPSS Inc. product and use of those Web sites is at your own risk.

When you send information to IBM or SPSS, you grant IBM and SPSS a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Information concerning non-SPSS products was obtained from the suppliers of those products, their published announcements or other publicly available sources. SPSS has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-SPSS products. Questions on the capabilities of non-SPSS products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to SPSS Inc., for the purposes of developing,

using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. SPSS Inc., therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. SPSS Inc. shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks of IBM Corporation, registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at *http://www.ibm.com/legal/copytrade.shmtl*.

SPSS is a trademark of SPSS Inc., an IBM Company, registered in many jurisdictions worldwide.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

This product uses WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, *http://www.winwrap.com*.

Other product and service names might be trademarks of IBM, SPSS, or other companies.

Adobe product screenshot(s) reprinted with permission from Adobe Systems Incorporated.

Microsoft product screenshot(s) reprinted with permission from Microsoft Corporation.

Bibliografía

Bell, E. H. 1961. *Social foundations of human behavior: Introduction to the study of sociology*. Nueva York: Harper & Row.

Blake, C. L., y C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., y J. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, .

Cochran, W. G. 1977. *Sampling Techniques*, 3rd ed. Nueva York: John Wiley and Sons.

Collett, D. 2003. *Modelling survival data in medical research*, 2 ed. Boca Raton: Chapman & Hall/CRC.

Cox, D. R., y E. J. Snell. 1989. *The Analysis of Binary Data*, 2nd ed. Londres: Chapman and Hall.

Green, P. E., y V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

Green, P. E., y Y. Wind. 1973. *Multiattribute decisions in marketing: A measurement approach*. Hinsdale, Ill.: Dryden Press.

Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. *Psychometrika*, 33, .

Hartigan, J. A. 1975. *Clustering algorithms*. Nueva York: John Wiley and Sons.

Hastie, T., y R. Tibshirani. 1990. *Generalized additive models*. Londres: Chapman and Hall.

Kennedy, R., C. Riquier, y B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement, and Analysis for Marketing*, 5, .

Kish, L. 1965. *Survey Sampling*. Nueva York: John Wiley and Sons.

Kish, L. 1987. *Statistical Design for Research*. Nueva York: John Wiley and Sons.

McCullagh, P., y J. A. Nelder. 1989. *Modelos lineales generalizados*, 2nd ed. Londres: Chapman & Hall.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. En: *Frontiers in Economics*, P. Zarembka, ed. Nueva York: Academic Press.

Murthy, M. N. 1967. *Sampling Theory and Methods*. Calcuta (India): Statistical Publishing Society.

Nagelkerke, N. J. D. 1991. A note on the general definition of the coefficient of determination. *Biometrika*, 78:3, .

Price, R. H., y D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, .

Rickman, R., N. Mitchell, J. Dingman, y J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association*, 228, .

Rosenberg, S., y M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research*, 10, .

Särndal, C., B. Swensson, y J. Wretman. 1992. *Model Assisted Survey Sampling*. Nueva York: Springer-Verlag.

Van der Ham, T., J. J. Meulman, D. C. Van Strien, y H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective. *British Journal of Psychiatry*, 170, .

Verdegaal, R. 1985. *Meer sets analyse voor kwalitatieve gegevens (en neerlandés)*. Leiden: Department of Data Theory, University of Leiden.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, y B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases*, 129, .

Índice