Miguel Martín Mateo y Natalia Romero Sandoval

Uso básico del SPSS para Ciencias de la Salud

UIDE
Powered by Arizona State University®

UAB
Universitat Autònoma de Barcelona

Uso básico del SPSS para Ciencias de la Salud

Miguel Martín Mateo y Natalia Romero Sandoval
Uso básico del SPSS para Ciencias de la Salud

Quito: Universidad Internacional del Ecuador, 2024
1.ª edición, 261 pp. Vol: 15 x 21 cm

CDU: 614 + 311 + 008.1
ISBN 978-9942-682-01-7
DOI: https://doi.org/10.33890/spssparacienciasdelasalud

1. Salud pública
2. Investigación estadística
3. Metodología de la investigación

Como citar: Martín-Mateo, M. y Romero-Sandoval, N. (2024). *Uso básico del SPSS para Ciencias de la Salud*. Universidad Internacional del Ecuador. https://doi.org/10.33890/spssparacienciasdelasalud

Av. Simón Bolívar y Av. Jorge Fernández.
(593-2) 2985-600 / (593-2) 5000-600
www.uide.edu.ec

Directora editorial: María Belén Calvache
Asistente editorial: Andrea Farfán

Diseño y corrección de estilo: La Caracola Editores

Este libro fue sometido a un proceso de revisión por pares bajo el sistema de doble ciego (*peer review*).

Uso básico del SPSS para Ciencias de la Salud

Miguel Martín Mateo
Universidad Autónoma de Barcelona

Natalia Romero Sandoval
Universidad Internacional del Ecuador

UIDE
Powered by
Arizona State University

UAB
Universitat Autònoma
de Barcelona

GRAAL

Índice

11 Capítulo uno

11 Introducción
12 ¿A quién se dirige el libro?
13 ¿Qué es un paquete estadístico?
14 Esquema didáctico
16 Enunciado de los ejemplos prácticos
19 Bibliografía

21 Capítulo dos

21 Estructura básica del paquete estadístico SPSS
22 Módulo de comunicación. Lectura y traducción de archivos de datos
24 Módulo de sintaxis
26 Normas generales de la sintaxis
29 Módulo de modificación y análisis
30 Módulo de resultados

33 Capítulo tres

33 Lectura y traducción de archivos de datos
35 Lectura y traducción de un archivo en Excel
36 Grabar un archivo de trabajo
36 Grabar en formato SPSS
36 Grabar en formato de otros programas
37 Fusión de archivos
37 Añadir casos
40 Añadir variables

45 Capítulo cuatro

45 Variables. Definiciones
49 Definición de valores desconocidos, *missing values*
52 Creación y modificación de variables

71 Capítulo cinco

71 Introducción
72 Selección de casos. Selección mediante una condición
74 Selección por muestreo aleatorio
77 Segmentación de un archivo
81 Agregación de datos

85 Capítulo seis

85 Introducción

86 Control de calidad de la base de datos

92 Descripción univariada de variables

101 Descripción bivariada

117 Descripción de los resultados en forma gráfica

123 Capítulo siete

123 Introducción

124 Comparación de dos porcentajes o dos distribuciones de frecuencias

125 Comparación de dos medias muestrales

126 Comparación de dos medias pertenecientes a dos muestras independientes

130 Comparación de dos medias de muestras dependientes

131 Comparación de más de dos medias muestrales

131 Análisis de la varianza. ANOVA

135 Análisis de la dependencia lineal entre dos variables continuas

135 Correlación

137 Regresión lineal entre dos variables continuas

141 Capítulo ocho

141 Presentación

141 Pirámide demográfica poblacional

148 Estudio de proporciones en estudios no dependientes del tiempo: cálculo de tasas y su estandarización

161 Análisis de proporciones en estudios dependientes del tiempo: cálculo de tasas de densidad de incidencia

171 Capítulo nueve

171 Introducción

172 El estudio

172 El ejercicio

175 Capítulo diez

175 Ejercicios del Capítulo 2

177 Ejercicios del Capítulo 3

183 Ejercicios del Capítulo 4

192 Ejercicios del Capítulo 5

200 Ejercicios del Capítulo 6

222 Ejercicios del Capítulo 7

231 Ejercicio del Capítulo 9

239 Capítulo once

239 Presentación

240 Capítulos 1 a 3

244 Capítulo 4 a 7

246 Capítulo 8

248 Capítulo 9

Capítulo uno
Presentación

1.1 Introducción

El grupo de investigación en Ciencias de la Salud, GRAAL (*Grups de Recerca d'Amèrica i Àfrica Llatines*) por sus siglas en catalán, está formado por un colectivo de investigadores extendido por América Latina. GRAAL tiene, entre sus objetivos, la formación de investigadores en el área de Metodología en Ciencias de la Salud.

Hace nueve años, el primer autor de esta obra y un colega de GRAAL en la Universidad Autónoma de Barcelona crearon un texto de ayuda para el aprendizaje de las utilidades del programa SPSS en su décima versión. Este manuscrito abordó las diferentes facetas en las que, a partir de casos reales, se inicia al alumno no únicamente en el uso del programa por ventanas, sino también en la utilización de la sintaxis de cada tipo de acción, con el fin doble de no solo entender mejor el proceso de análisis, sino también crear una constancia de las acciones efectuadas y facilitar su reproducción en situaciones futuras o por otros investigadores.

El libro ha sido una herramienta muy apreciada por los estudiantes, pero la continua actualización del programa SPSS, sumada a la experiencia adquirida en estos años, aconseja la redacción de un nuevo libro. En los últimos años, la actividad de los investigadores de GRAAL se ha concretado en la elaboración de un programa de maestría, en el cual se enfatiza el análisis de

datos del entorno del programa SPSS. Así, en esta nueva obra, se incorporan aquellos aspectos que, en la experiencia de estos años, han ido acumulando los profesores de esta materia.

Para esta nueva obra, se emplean casos de nuevos estudios lo que permite al estudiante desarrollar nóveles conocimientos, competencias y habilidades en las nuevas versiones del mencionado programa, sin dejar de emplear las herramientas de la sintaxis. "Uso básico del SPSS para Ciencias de la Salud" llevará al estudiante, a manera de un relato descriptivo y crítico de estudios reales, por las diversas acciones que ofrece el programa SPSS y proporcionará varios ejercicios, de los que encontrará la resolución en el último capítulo, además de que se enriquecerá con los resultados publicados en los diversos artículos en revistas científicas.

1.2 ¿A quién se dirige el libro?

Este libro está dirigido a diferentes tipos de profesionales, no solo del área de las Ciencias Médicas ni de la Estadística Aplicada en Ciencias de la Salud, sino a cualquier investigador de campos que utilicen la estadística como herramienta del conocimiento, como por ejemplo, las Ciencias Sociales, la Demografía y cualquier otra actividad que componga el concepto indicado de Ciencias de la Salud en su sentido más amplio, pero especialmente a los estudiantes de maestría que requieran un análisis de sus datos controlando todo el proceso.

Así, un estudiante de Metodología en Ciencias de la Salud debe poseer una formación que le capacite en el análisis de datos, empezando desde los aspectos más básicos —como son la definición de la naturaleza de las variables de un estudio, el formato de estas— así como explicitar desde un buen comienzo las etiquetas de los valores posibles si la variable es de tipo categórico. Por otro lado, dado que, en la mayoría de los estudios, especialmente los implicados en estimar la incidencia de problemas de salud, deben explicitarse los tiempos en que se observan los sucesos, debe tener muy clara la definición del formato de ese tiempo.

Una vez definidos esos aspectos, el responsable de la explotación de la información, que llamaremos a partir de ahora, por simplicidad, el estadístico, no debería dar por supuesto ningún tipo de garantías ofrecidas acerca de la

calidad de los datos, especialmente si la adquisición y digitalización de estos los ha llevado a cabo otra persona. Es por ello que forzosamente debe efectuar un análisis previo de aspectos como la información incompleta e incluso de la información errónea detectable.

Nunca debe abordarse el análisis estadístico de la información recogida en una matriz de datos sin haber realizado el control de calidad de los datos disponibles. De no hacerlo así, la probabilidad de tener que corregirla a posteriori y repetir los análisis es muy elevada. Este es un aspecto en el que se insistirá en este libro, ya que frecuentemente se inicia el análisis de la información sin haber cumplido con esta fase.

1.3 ¿Qué es un paquete estadístico?

Un paquete estadístico consiste en un conjunto de programas creado en un lenguaje común que pretende dar respuesta a todos los aspectos enumerados anteriormente. El analista debe recurrir al uso de programas o paquetes estadísticos, preferentemente homologados y de distribución amplia, de forma que los resultados obtenidos sean siempre comprobables y comparables por cualquier otro investigador.

Los paquetes estadísticos que permiten efectuar un análisis de datos, al igual que manipular y gestionar las matrices de datos, son muy variados y su utilización está muy sometida a corrientes muchas veces influidas por el mercado de *software* y modas entre los grupos de investigación.

En este libro, se muestra cómo utilizar el paquete SPSS en el entorno Windows. Esta elección está motivada por dos aspectos. El primero es su gran difusión en la mayoría de las universidades y, en segundo lugar, porque su aprendizaje es, sin duda, mucho más rápido, ya que su lenguaje o sintaxis está más próxima al lenguaje natural en lengua inglesa.

Otros paquetes estadísticos —especialmente los que resuelven los problemas de análisis en lenguaje R, popular por ser de acceso libre— requieren, en la experiencia de los autores, un mayor esfuerzo en la programación, así como unos conocimientos más elevados de Estadística, nivel que no se le presupone al lector de este libro.

1.4 Esquema didáctico

El abordaje de esta obra intenta acercarse al uso del paquete estadístico SPSS de una forma profesional, es decir, explicando la sintaxis de las aplicaciones y no solo el manejo de las acciones preprogramadas que se muestran en los menús desplegables por el sistema de ventanas que ofrece el paquete. Este esquema permite una utilización consciente de los análisis que se realizan, con todas las posibilidades que precisamente son las que distinguen a un profesional de un conocedor superficial de paquetes estadísticos.

Esta forma de trabajar da acceso a otras posibilidades como son la utilización de recursos de análisis o de descripción que no existen de forma preprogramada, así como también la de crear programas aplicables en futuras ocasiones sin necesidad de repetir el proceso de generación del análisis, asegurándose por lo tanto de que el análisis es siempre el mismo.

El sistema de aprendizaje se basa en la resolución de problemas a partir de un estudio real, con lo cual se genera la posibilidad de enfrentarse a problemas muy frecuentes en la realidad, que el alumno desconoce, tanto su existencia como su resolución, huyendo de la resolución de problemas o entrar directamente en el análisis estadístico a partir de archivos de interés exclusivamente académicos, los cuales, en general, están formado por pocos casos y número limitado de variables.

Los ejemplos que se desarrollan a lo largo de este estudio están centrados en el ámbito de la Epidemiología, campo en el que los autores han desarrollado la mayoría de su actividad profesional. No obstante, la complejidad de situaciones considerada a la hora de manipular archivos complejos hace que el interés sea inmediato para cualquier profesional que requiera la combinación de diferentes archivos de datos.

El libro está estructurado de forma que el lector vaya adquiriendo los conocimientos generales de uso de un paquete estadístico a partir, como ya se ha indicado, de un ejemplo de análisis de los datos de un estudio real.

No se busca la redacción de un manual simplificado, ni suplir a los sistemas de ayuda que le ofrece el programa SPSS, sino facilitar el seguimiento del curso, así como proveer el trabajo autónomo de los estudiantes. El proceso recomendado es, por lo tanto, el seguimiento ordenado de los capítulos del libro, si bien existen dos recorridos diferenciados que, según sea el conocimiento y nivel de práctica del lector, podrían superponerse.

- El primer recorrido, indicado con el epígrafe de **Ventanas**, introduce al lector en el uso de los menús desplegables y en el trabajo clásico del entorno Windows.
- El segundo recorrido, más profesional, sería el que va indicando las distintas instrucciones de **Sintaxis** de cada apartado, acompañado de la propuesta de ejercicios de los que se ofrece una solución de sintaxis para el logro de los objetivos descritos, al final de cada capítulo.

La conjunción de los dos recorridos o técnicas se lleva a cabo insistiendo en la generación automática de un cuadro de sintaxis a partir de la opción **Pegar** presente en casi todas las ventanas descritas. Se aconseja al estudiante efectuar el recorrido de forma paralela no solo por mejorar la comprensión de sus acciones en los menús desplegables, sino también facilitar la corrección de errores que se pueden producir al utilizarlos.

A partir de este doble esquema, los procedimientos se presentarán según el siguiente esquema:

- En primer lugar, se describen las diferentes formas de definir la matriz de datos y su exportación e importación a otros sistemas de análisis.
- En segundo lugar, se detallan los procedimientos para la definición de variables, su nombre, etiqueta y, en el caso de variables categóricas, la etiqueta de cada categoría. En este mismo apartado se definen los valores perdidos o sin información, así como su manipulación.
- Una vez definida la información que se va a utilizar en el o los archivos de datos, se describen los pasos necesarios para combinar casos y variables de distintos ficheros, así como para la selección temporal o definitiva de casos y las opciones para efectuar el mismo análisis diversas veces en función de un factor.
- Un cuarto bloque hace referencia a cómo efectuar la creación de nuevas variables a partir de las existentes, así como modificarlas mediante recodificaciones creando nuevas variables.
- El quinto paso incluye la aproximación, introduce las técnicas para realizar los análisis descriptivos univariados y bivariados más frecuentes, destacando también la fase previa del control de calidad de los datos.

Se pretende que, con este esquema, la resolución de todas las fases descritas en el caso real, que se utiliza de ejemplo, muestre todos aquellos problemas y dificultades presentes en el trabajo que un analista de datos desarrolla cotidianamente.

El control profesional de las actuaciones que hay que realizar para resolver estos problemas es otro de los objetivos de este libro. Por lo que en todos los ejemplos se muestran, tal y como se ha comentado anteriormente, las acciones que se deben tomar bajo dos puntos de vista, el automático mediante el uso de ventanas y el consistente en la utilización de la sintaxis.

1.5 Enunciado de los ejemplos prácticos

El ejercicio práctico fundamental a partir del cual se desarrolla el aprendizaje se basa en un estudio realizado en Ecuador, en las escuelas municipales de la ciudad de Quito efectuado durante el curso lectivo 2010-2011. En ese período, el número de estudiantes matriculados entre 8 y 18 años fue de 7365 en las 23 escuelas. Este estudio contó con la aprobación de un Comité de Bioética y la base de datos anonimizada está cedida a la Red GRAAL para uso académico exclusivamente. Las edades de los estudiantes del estudio fluctúan entre los 9 y 17 años y estuvieron entre el quinto y décimo año de educación básica. El esquema general se describe en el artículo:

> Romero-Sandoval, N., Recalde, R., Quizanga, J., Anchali, E., Ruiz, V., Falconí, J., Flores, O., y Martín, M. (2012). "Quito Municipal Schools" Cohort Study. Baseline results. *Open Journal of Epidemiology, 2(3), 70 - 74*.

Las escuelas se agruparon en cuatro regiones sanitarias. Luego de la firma del consentimiento informado junto con el cumplimiento de las condiciones y criterios de admisión, se tuvo el resultado de las medidas antropométricas de los estudiantes que asistieron a clase el día de la medición. El cuestionario que se administró era acerca de hábitos asociados al sedentarismo, sus pautas alimentarias, entre otras variables.

El resultado de los mismos se encuentra en el archivo escolares.dat, el cual está dividido en dos: **escolareszonas1y2.sav** y **escolareszonas3y4.sav** ya en formato directamente legible por el programa SPSS. En la tabla 1.1, se describe la información recogida en estos archivos, los cuales tienen una estructura idéntica.

Tabla 1.1 Variables de los estudiantes del ejemplo práctico

Variable	Etiqueta	Naturaleza	Tipo
codigo	Código Alumno	Escala	Numérico
talla	Talla (m)	Escala	Numérico
peso	Peso (Kg)	Escala	Numérico
p_grasa	Proporción de grasa	Escala	Numérico
sexo	Sexo	Nominal	Numérico
ed	Edad	Escala	Numérico
status_valoración_IMC	categoría de IMC según WHO 2007	Nominal	Numérico
cua_cami	cuadras que caminas	Ordinal	Numérico
autoper	autopercepción de la imagen	Nominal	Numérico
tip_flia	tipologia familiar	Nominal	Numérico
desayuno	toma desayuno?	Nominal	Numérico
com_comp	Come acompañado	Nominal	Numérico
com_tres	Come tres comida al día	Nominal	Numérico
diet	Está siguiendo alguna dieta	Nominal	Numérico
frut	fruta	Ordinal	Numérico
com_rap	comida rápida	Ordinal	Numérico
dulce	dulces	Ordinal	Numérico
refresc	refrescos	Ordinal	Numérico
zona_sanit	red sanitaria	Nominal	Numérico
status2	clasificación según percentiles IMC-OMS	Escala	Numérico

A todos los escolares se adjudicó un código de referencia con el fin de asignar de forma inequívoca otros tipos de información. De todos los aspectos que se investigaban, en estos archivos junto a la información acerca de sus padres, la cual se encuentra en el archivo Excel **padres.xls**, se han escogido las variables adecuadas para efectuar el curso con cierta agilidad.

En la tabla 1.2 se describe la información recogida por el cuestionario aplicado a los padres de los alumnos y se refieren al conocimiento que los mismos poseen sobre los hábitos dietéticos y de ejercicio de sus hijos.

Tabla 1.2 Variables de los padres del ejemplo práctico

Variable	Etiqueta	Naturaleza	Tipo
código	Código Alumno	Escala	Numérico
participa	Participa	Nominal	Numérico
psexo	Sexo del padre/madre/tutor	Nominal	Numérico
pescola	Escolaridad representante	Nominal	Numérico
esc_madre	Escolaridad de la madre	Nominal	Numérico
pcua_cami	Cuadras que caminas	Ordinal	Numérico
pfruta	Fruta	Ordinal	Numérico
pc_rap	Comida rápida	Ordinal	Numérico
pdul	Dulces	Ordinal	Numérico
prefr	Refrescos	Ordinal	Numérico
psna	Snacks	Ordinal	Numérico
pc_sol	Come acompañado	Ordinal	Numérico
pdiet	¿Su hijo/a está realizando dieta?	Nominal	Numérico
autoper_pad	Autopercepción de la imagen	Nominal	Numérico
pedad	Edad del padre/madre/tutor	Escala	Numérico
pdesay	¿Toma desayuno su hijo/a?	Nominal	Numérico

En los capítulos 8, 9 y 10, se utilizan otros archivos, los cuales se enuncian en los mismos con tal de poder ejemplarizar otras utilidades y que, con frecuencia, no se encuentran descritas en los manuales más comunes. Los archivos están disponibles para los usuarios de este libro en los siguientes enlaces:

Bases_grupo 1: archivos para el estudio de escolares (tres archivos *xls y cinco *sav). Inicie con los archivos en Excel. Usted podrá tener el apoyo para su trabajo con los archivos en SPSS. En esta carpeta también encontrará la información de las etiquetas de valor de las variables de cada archivo. Este archivo será usado en el capítulo 4, etiquetas de las variables y sus categorías. *https://n9.cl/8cv7h*

Bases_grupo 2: Archivos (*sav) para el estudio consultas del Hospital de Belo Horizonte (consultas) y para la tarea de evaluación, el estudio de mordeduras de murciélagos hematófagos y rabia selvática en la Amazonía Ecuatoriana. *https://n9.cl/1gavz*

1.6 Bibliografía

Romero-Sandoval, N., Guanopatin, A., Gallegos, G., Collaguazo, A., Sáenz, P., Latorre, V., Egas, V., Flores, O., Utzet, M. y Martín, M. (2013). Breakfast Habits and Family Structure Associated with Overweight and Obesity in General Basic Students, Ecuador. *British Journal of Medicine & Medical Research, 3*(1): 128-139.

Carvajal, D., Martin, M., Romero-Sandoval, N. (2013). Modelo explicativo del efecto de la talla y grasa corporal en el peso de escolares de 9 a 17 años. *Rev. Med.Vozandes*. 24: 9–18.

Romero-Sandoval, N., Flores, O., Egas, C., Villamar, G., Larrea, Z., Cruz, M., Martín, M. (2014). Quito Municipal Schools – Cohort study: Self-perception of body image and factors related with it. *Open Journal of Epidemiology, 4*(3), 122-128. doi: 10.4236/ojepi.2014.43017.

Cevallos-Salazar, J., Flores-Carrera, O., Lozano-Ruiz, P., Cruz-Mariño, A., Martín-Mateo, M., y Romero-Sandoval, N. (2015). Glucemia y lipemia en escolares con obesidad en el distrito metropolitano de Quito, Ecuador. *Duazary, 12*(1), 7-14. *http://dx.doi.org/10.21676/2389783X.1392*.

Romero-Sandoval, N., Robles, J., Cisneros, M., Ruiz, V., & Martín, M. (2017). Relationship between excess weight and discrepancies in reporting of food habits between Ecuadorian school children and their parents. *Revista Brasileira de Saúde Materno Infantil*, *17*(3), 615-622.
http://dx.doi.org/10.1590/1806- 93042017000300011.

Capítulo dos

Estructura básica del paquete estadístico SPSS

2.1 Estructura básica del paquete estadístico SPSS

El paquete estadístico SPSS —al igual que la mayoría de los paquetes estadísticos— está constituido de forma modular, de manera que cada fase del trabajo es realizada por uno de los módulos. De forma esquemática, se compone de cuatro módulos, los cuales tienen diferentes funciones:

Un primer módulo se dedica a la conexión con el exterior del entorno del programa, es decir, permite acceder a diversos archivos externos y facilita su traducción a un lenguaje o gramática propia del programa.

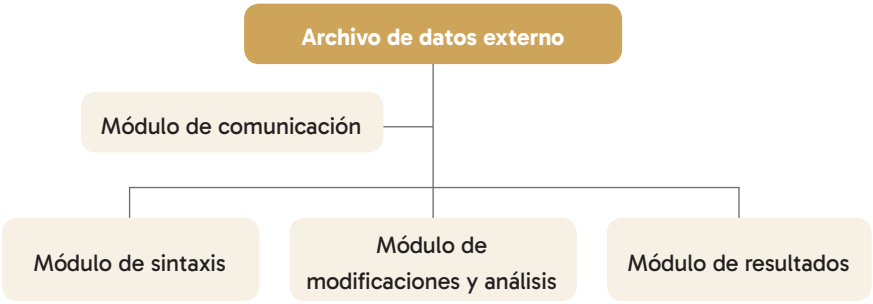


Figura 2.1 Esquema modular del programa SPSS

El módulo de comunicación permite la importación y lectura de archivos externos y produce un archivo en lenguaje SPSS, el cual se visualiza en una ventana doble llamada archivo activo y sobre la que se efectuarán los análisis o las transformaciones necesarias para lograr los resultados. Sobre esta ventana actúa el módulo de modificaciones y análisis.

Esta ventana es doble porque está constituida en dos hojas. La primera visualiza los valores de las variables que se han determinado o ventana de datos. La segunda hoja o ventana de variables es en donde se describe variable a variable; su nombre completo; propiedades de naturaleza; formato; propiedades de visualización; así como otras propiedades acerca de cómo se identifica la información perdida, errónea o faltante; y, en el caso de variables categóricas, las etiquetas de las categorías de la misma si es que la codificación es numérica.

Cualquier acción que se ejecute mediante instrucciones presentes en la barra de herramientas de la ventana de datos genera, en primer lugar, unas instrucciones en sintaxis SPSS, la cual puede visualizarse en una nueva ventana (ventana de sintaxis). Si, en vez de la acción **Aceptar** —que aparece en las ventanas de cada acción—, se utiliza la opción **Pegar**, la ejecución de las instrucciones se realiza desde esta nueva ventana.

En segundo lugar, genera un archivo de resultados, el cual es gestionado por el módulo de resultados, que permite su traducción en diversos formatos con el fin de facilitar su exportación. Por último, el módulo de resultados, mediante el módulo de comunicación, posibilita almacenar y exportar toda la información generada por el módulo de modificación y análisis.

2.2 Módulo de comunicación. Lectura y traducción de archivos de datos

Este módulo tiene herramientas de lectura de archivos que ya están en lenguaje SPSS y que se caracterizan por la extensión del fichero **Nombre.sav.** Este tipo de archivos ya traducidos son de lectura inmediata. Y una vez leídos permiten ya trabajar con ellos.

No obstante, el SPSS tiene utilidades para leer bases de datos en muy diferentes formatos, lo que permite la importación y traducción de los mismos. El formato más frecuente es aquel en el que la información ha sido introducida

en Excel. La sola detección de la extensión ***xls** hace que el programa utilice un algoritmo de traducción, pudiendo importar no solo lo escrito en las hojas de Excel, sino también la naturaleza de la variable, si es cadena, letras o caracteres alfanuméricos, numérica tanto si está expresada en números naturales como reales y también fechas, si bien para este último tipo de variable debe de estar previamente definida como tal en el archivo de origen.

El resultado de esta importación es la creación de un archivo de datos llamada fila activa y puede almacenarse con la extensión ***sav.** Es decir, el módulo de comunicación tendrá las opciones de abrir ficheros, buscando en el directorio o carpeta adecuado, activando submódulos o funciones de lectura dependiendo de la extensión que acompañe al archivo.

Por ejemplo, si el archivo de datos posee la extensión ***dat**, se activa un módulo de importación de un fichero en ASCII. De esta forma, no solo se importa la estructura y la matriz de datos sino también la información de las variables que se encuentran en ese archivo, si bien no siempre es totalmente automático y con frecuencia se debe aportar información adicional para lograr la traducción del archivo.

Una vez leído el archivo, se abre una pantalla que se conoce como la matriz de datos, o *file* activo, la cual ofrece información de los datos adquiridos, **ventana de datos**, también incluye la información acerca de cada variable y cómo esta se define.

Es decir, número de caracteres que se visualizan, número de decimales, formato de la fecha en las variables tiempo, la naturaleza numérica o alfanumérica de la misma y en el caso de variable numérica, si esta es escalar, nominal u ordinal. En la figura 2.2 se muestra el proceso de lectura por ventanas cuyo resultado es el archivo: **escolares_zona_1y 2.sav.**

Dado que la extensión es de naturaleza.sav, el programa, al activar la lectura con las instrucciones archivo Abrir datos, genera un *file* activo. Para ello, abra el programa SPSS y en la ventana emergente, lleve a cabo la acción **Abrir ► Datos** y busque con el explorador la carpeta y el nombre del archivo que pretende abrir.

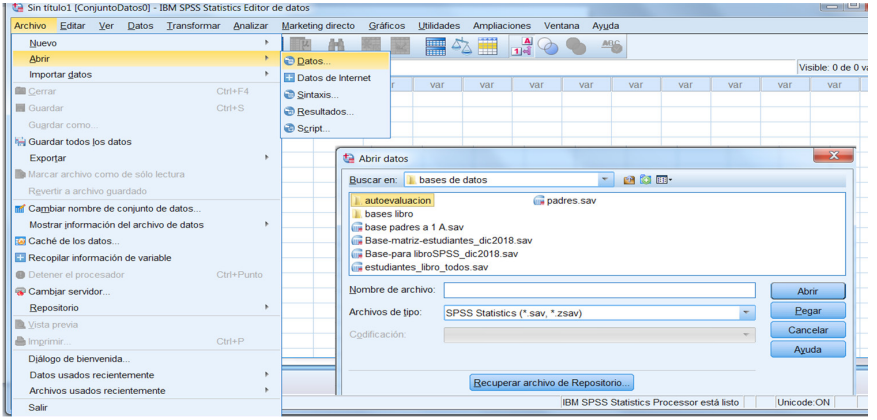


Figura 2.2. Proceso de lectura de un fichero en formato SPSS, extensión.sav

El procedimiento lógico que el lector acostumbra a realizar en el entorno Windows sería aceptar pulsando la tecla correspondiente de la subventana; sin embargo, desde el principio, le aconsejamos que utilice la opción **Pegar**.

2.3 Módulo de sintaxis

Con la opción indicada**, Pegar**, el programa activa una ventana y genera un nuevo archivo, de extensión **.sps**, en el cual va insertando todas las acciones efectuadas por el analista con el fin de poder verificarlas, guardarlas y ejecutar las mismas nuevamente cuando fuese necesario. En esta nueva ventana denominada editor de sintaxis, aparecerá la instrucción correspondiente al hecho de apretar la tecla en aceptar si así lo hubiese hecho:

```
GET
FILE='E:\Libro SPSS Ecuador\bases de datos\estudiantes_zonas_1y2.sav.
DATASET NAME ConjuntoDatos1 WINDOW=FRONT.
```

Recuerde que este proceso no se realiza si acciona la tecla Aceptar. Si analiza el texto generado de sintaxis, observará que está constituido por dos instrucciones, las cuales siempre van delimitadas por un punto final.

La primera **GET FILE** es una instrucción de mandato, selecciona ese archivo o *file*, el cual se encuentra situado en la unidad E: en la carpeta Libro SPSS Ecuador, subcarpeta bases de datos y nombre estudiantes_zonas_1y2.sav. Obviamente, el lector, cuando ejecute esta acción, deberá cargar el archivo desde el directorio que tenga configurado en su computador.

La segunda **DATA SET NAME** indica el nombre temporal que adjudica a la base de datos o *file* activa, ya que puede tener más de una simultáneamente, y que no presentará un nombre reconocible hasta que no se haya grabado el archivo.

Comentario:

Recomendamos al estudiante que, al menos hasta que no tenga una familiaridad con el programa, no abra más de una base de datos simultáneamente.

Un aspecto que el lector debe entender desde el inicio es que las instrucciones que aparecen en el editor de sintaxis solo son expresiones correspondientes a las acciones que ha realizado el analista, pero que aún no se han llevado a cabo. Estas deben ejecutarse apretando el símbolo  de la barra de herramientas de la ventana, una vez seleccionadas las instrucciones que desee ejecutar.

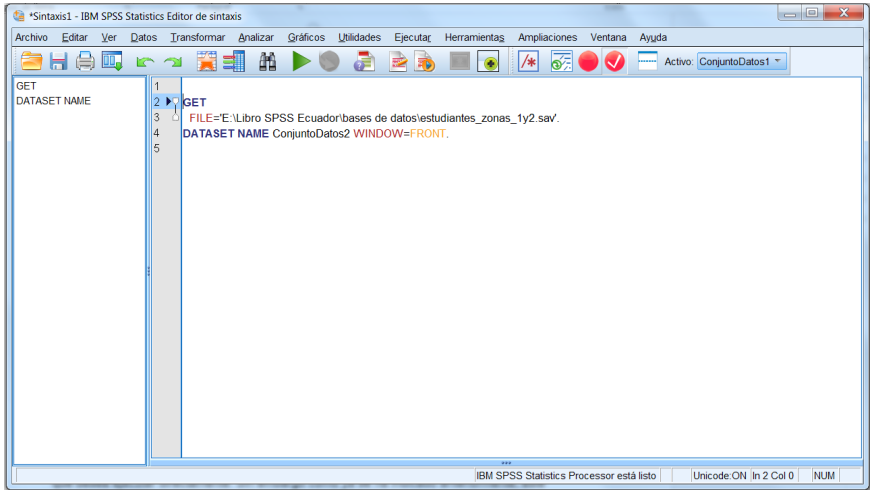


Figura 2.3. Apertura de un archivo de datos mediante sintaxis

El analista avezado puede abrir directamente una nueva ventana de sintaxis desde el módulo de comunicación, barra de herramientas de la ventana que se abre al iniciar el SPSS, en la pestaña **Archivo ► nuevo ►sintaxis** y en la nueva ventana escribir la sintaxis de las acciones que desea ejecutar directamente. Sin embargo, como ya se ha indicado anteriormente, este curso no pretende ser una introducción al lenguaje SPSS, si no a utilizarlo racionalmente sin necesidad de su aprendizaje previo. No obstante, y para que el lector conozca las reglas básicas del lenguaje de programación, estas se describen en el siguiente apartado.

2.4 Normas generales de la sintaxis

Como en la mayoría de los lenguajes de programación, las instrucciones de SPSS acostumbran a ser abreviaturas o expresiones de interpretación obvia en inglés. Todas las instrucciones pueden complementarse con subinstrucciones opcionales, tales como el tipo de subanálisis que se debe realizar, el tipo de estadísticos que se debe calcular, la presentación de los resultados, la repetición del análisis en otro conjunto de variables o de relación entre ellas, entre otros.

A través del índice del sistema de ayuda —presente en la barra de herramientas de todas las ventanas— indicado con el símbolo **Ayuda**, se accede a la sintaxis de las instrucciones SPSS (*command syntax*) y a una somera descripción de su significado.

Existen normas de presentación de las mismas que facilitan la síntesis en la explicación. Así, en cualquier instrucción:

- Los paréntesis, apóstrofes y caracteres de repetición, /, deben escribirse obligatoriamente.
- Cuando en una instrucción está escrito algo entre corchetes **[ ]**, indica que explicitar ese contenido es opcional; por lo tanto, su presencia o ausencia no impide la ejecución de la instrucción general. Su utilización será necesaria o no en función de la acción concreta que se quiera efectuar.
- Las llaves **{ }**, indican que las opciones que se describen entre ellas son electivas y pueden escogerse, en ocasiones, más de una. Cuando vea que existe un doble asterisco, es un indicador de que es la opción

por defecto de todas ellas. Constará con la marca de doble asterisco aquella o aquellas que el sistema opta por defecto, es decir, opción que se realizará si el usuario no indica nada al respecto.

- Las opciones expresadas a continuación de una barra inclinada /, implican que su contenido se puede repetir diverso número de veces o bien que es una subinstrucción que se diferencia de la misma opción cuando no va precedida por dicho símbolo.
- Todas las instrucciones deben finalizar con un punto, en cuya ausencia el compilador encadenará con la siguiente instrucción y, por lo tanto, indicará error de sintaxis.

Independientemente de que la instrucción siguiente se explicará con detalle en otro capítulo, queremos que el lector identifique los signos de sintaxis presentes en la siguiente instrucción:

```
DATASET ACTIVATE ConjuntoDatos1.
FREQUENCIES VARIABLES=talla
  /STATISTICS=STDDEV MINIMUM MAXIMUM MEAN MEDIAN
  /HISTOGRAM NORMAL
  /ORDER=ANALYSIS.
```

En primer lugar, observe que el punto solo puede aparecer como final de toda la instrucción. Así esta frase de sintaxis consta de dos instrucciones.

- La primera refiere al nombre interno que atribuye a la base de datos en la que se ejecutará la siguiente frase.[1]
- La segunda consta de cuatro líneas, que finalizan en un punto al final de la última línea.

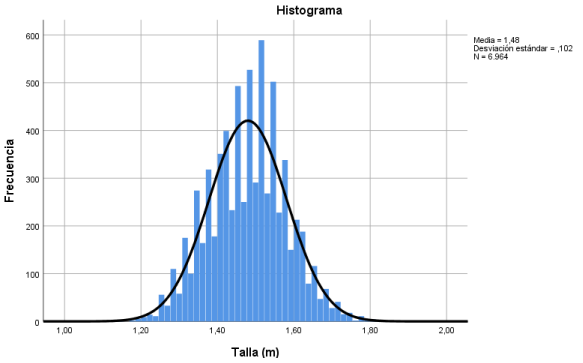
La primera palabra FREQUENCIES indica el nombre que se le da a la acción de determinar las frecuencias de aparición de los valores de la variable **talla** con la cual se pretende, además, ejecutar tres subinstrucciones indicadas por el símbolo /.

La primera señala que estadísticos se quieren calcular. En este caso, dado que la variable es de naturaleza escalar o continua, indica que se quiere conocer el valor de la desviación estándar, el valor máximo, el mínimo, la media y la mediana.

1 Recuerde que el programa puede tener abiertas diversas bases de datos o *datasets*.

Es responsabilidad del analista si requiere erróneamente esta información para una variable que sea categórica o por ejemplo de naturaleza no numérica. Asimismo, por la misma razón de la naturaleza continua de la variable, solicita en la segunda subinstrucción que se cree el histograma, superponiendo en el gráfico resultante una campana de Gauss centrada en la misma media y con la misma desviación estándar de los datos analizados. Ya veremos en su momento que esa acción difícilmente nos puede indicar acerca de la normalidad de la variable y existen técnicas más adecuadas que basarse en un gráfico.

Por último, la tercera subinstrucción indica que la descripción de la tabla de frecuencias se efectúe desde el valor más bajo de la variable hasta el más alto. Además de la tabla de frecuencias, si copia estas frases de sintaxis en la ventana de sintaxis y activa la acción obtendrá:



Estadísticos		
Talla (m)		
N	Válido	6964
	Perdidos	0
Media		1,4798
Mediana		1,4800
Desv. Desviación		,10166
Mínimo		1,15
Máximo		1,92

Figura 2.4. Resultados obtenidos mediante la sintaxis analizada en lo referente a las dos subinstrucciones */STATISTICS* y */ HISTOGRAM NORMAL*

2.5 Módulo de modificación y análisis

El resultado de abrir el archivo **escolares_zona_1y2.sav.**, ya sea directamente por la ventana emergente al iniciar el programa, **Archivo ► Abrir ► Datos** (ver figura 2.3), o por la ejecución derivada de la acción que consta en la ventana de sintaxis al haber efectuado la acción **Pegar** (figura 2.5).

En la figura 2.5, en la parte superior de la ventana, puede observarse la barra de herramientas —presente en todas las ventanas que se pueden abrir—. En la que, en cualquier momento, el analista puede comunicarse y activar el módulo de comunicación, pestaña **Archivo,** o las acciones que incluye este módulo de modificación y análisis, así como otras acciones que se irán detallando a lo largo del curso.

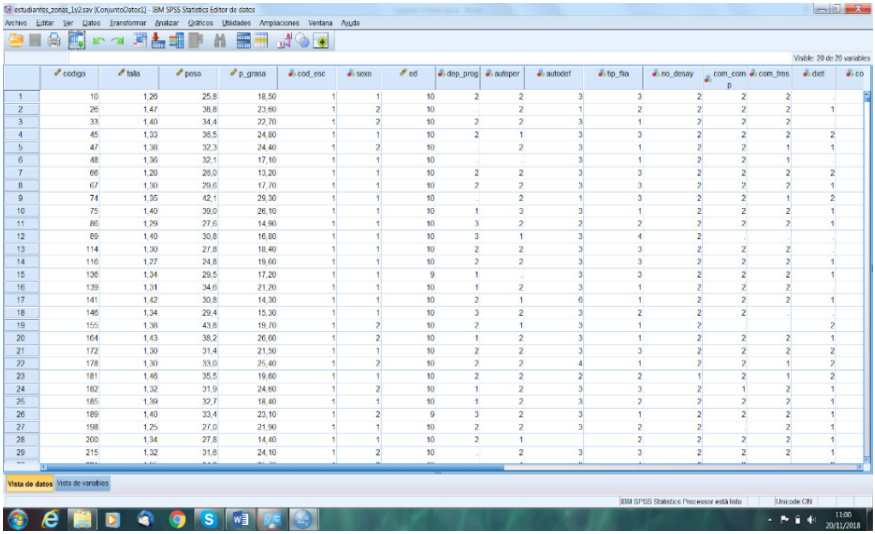
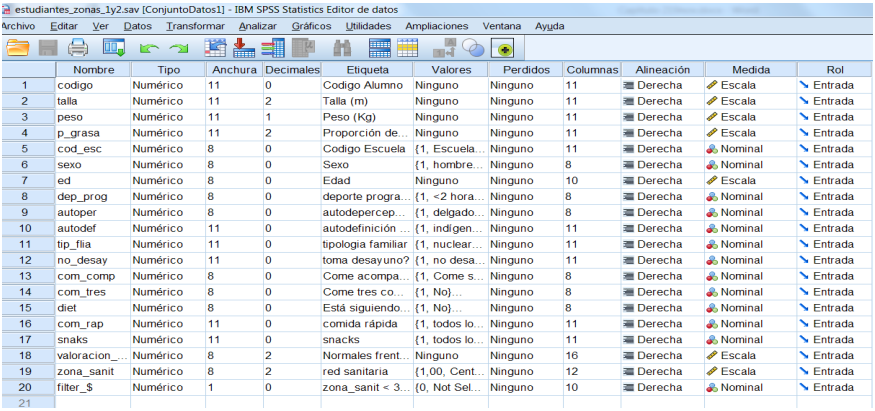


Figura 2.5. Ventana de visualización de los datos tal y como se han traducido por el módulo de comunicación

También, en la parte inferior izquierda, se puede acceder a la información de las variables en la pestaña **Vista de variables,** ventana que se muestra en la figura 2.6.


estudiantes_zonas_1y2.sav [ConjuntoDatos1] - IBM SPSS Statistics Editor de datos

Archivo Editar Ver Datos Transformar Analizar Gráficos Utilidades Ampliaciones Ventana Ayuda

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	codigo	Numérico	11	0	Codigo Alumno	Ninguno	Ninguno	11	Derecha	Escala	Entrada
2	talla	Numérico	11	2	Talla (m)	Ninguno	Ninguno	11	Derecha	Escala	Entrada
3	peso	Numérico	11	1	Peso (Kg)	Ninguno	Ninguno	11	Derecha	Escala	Entrada
4	p_grasa	Numérico	11	2	Proporción de...	Ninguno	Ninguno	11	Derecha	Escala	Entrada
5	cod_esc	Numérico	8	0	Codigo Escuela	{1, Escuela...	Ninguno	11	Derecha	Nominal	Entrada
6	sexo	Numérico	8	0	Sexo	{1, hombre...	Ninguno	8	Derecha	Nominal	Entrada
7	ed	Numérico	8	0	Edad	Ninguno	Ninguno	10	Derecha	Escala	Entrada
8	dep_prog	Numérico	8	0	deporte progra...	{1, <2 hora...	Ninguno	8	Derecha	Nominal	Entrada
9	autoper	Numérico	8	0	autodepercep...	{1, delgado...	Ninguno	8	Derecha	Nominal	Entrada
10	autodef	Numérico	11	0	autodefinición ...	{1, indígen...	Ninguno	11	Derecha	Nominal	Entrada
11	tip_flia	Numérico	11	0	tipologia familiar	{1, nuclear...	Ninguno	11	Derecha	Nominal	Entrada
12	no_desay	Numérico	11	0	toma desayuno?	{1, no desa...	Ninguno	11	Derecha	Nominal	Entrada
13	com_comp	Numérico	8	0	Come acompa...	{1, Come s...	Ninguno	8	Derecha	Nominal	Entrada
14	com_tres	Numérico	8	0	Come tres co...	{1, No}...	Ninguno	8	Derecha	Nominal	Entrada
15	diet	Numérico	8	0	Está siguiendo...	{1, No}...	Ninguno	8	Derecha	Nominal	Entrada
16	com_rap	Numérico	11	0	comida rápida	{1, todos lo...	Ninguno	11	Derecha	Nominal	Entrada
17	snaks	Numérico	11	0	snacks	{1, todos lo...	Ninguno	11	Derecha	Nominal	Entrada
18	valoracion_...	Numérico	8	2	Normales frent...	Ninguno	Ninguno	16	Derecha	Escala	Entrada
19	zona_sanit	Numérico	8	2	red sanitaria	{1,00, Cent...	Ninguno	12	Derecha	Escala	Entrada
20	filter_$	Numérico	1	0	zona_sanit < 3...	{0, Not Sel...	Ninguno	10	Derecha	Nominal	Entrada
21											

Figura 2.6 Ventana de vista de variables.

2.6 Módulo de resultados

Cualquier acción que se realice con los datos da lugar a un resultado disponible en la ventana de **Resultados**, en la que, además de los mismos, consta el proceso de análisis realizado.

Por ejemplo, como veremos más adelante, se hubiese deseado conocer la descripción de la variable sexo de estos estudiantes, mediante las instrucciones de ventana **Analizar ►Estadísticos descriptivos ►Frecuencias ►** y en la ventana de variables seleccionar **sexo**, al ejecutar la orden que se desea ejecutar

```
FREQUENCIES VARIABLES=sexo
  /ORDER=ANALYSIS.
```

Generada mediante la acción **Pegar**, que consta en el editor de sintaxis y presionando el símbolo ▶ de la barra de herramientas de la ventana la sintaxis asociada. O bien directamente mediante la acción de **Aceptar** de la ventana de **Análisis**, se genera el archivo de resultados en una nueva ventana, tal y como se muestra en la figura 2.7.

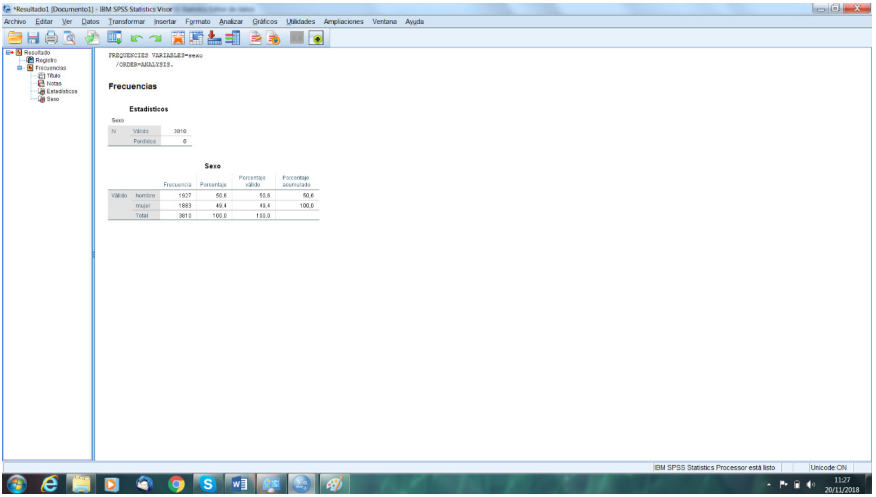


Figura 2.7 Ejemplo de ventana de resultados

Este archivo, a su vez, puede editarse al copiarlo en una carpeta del usuario, le añade automáticamente la extensión .spv, o exportarlo a otro formato utilizando la opción que se encuentra en la pestaña con símbolo ▶ de la barra de herramientas de la ventana. Por último, la ventana de datos con todas las modificaciones efectuadas durante el análisis puede grabarse en la carpeta destinada por el analista. Esta información puede grabarse entera o seleccionando las variables tal y como se puede ver en la figura 2.8.

Esta ventana se abre en la de datos al seleccionar Archivo de la barra de herramientas mediante las acciones **Archivo ► Guardar como.** Obviamente, si en vez de aceptar, se oprime la tecla de pegar, la sintaxis que se genera es:

```
SAVE OUTFILE='E:\Libro SPSS Ecuador\bases de
datos\estudiantes_zonas_1y2.sav'
  /COMPRESSED.
```

La cual, como en casos anteriores, debe seleccionarse y activar mediante el símbolo ▶.

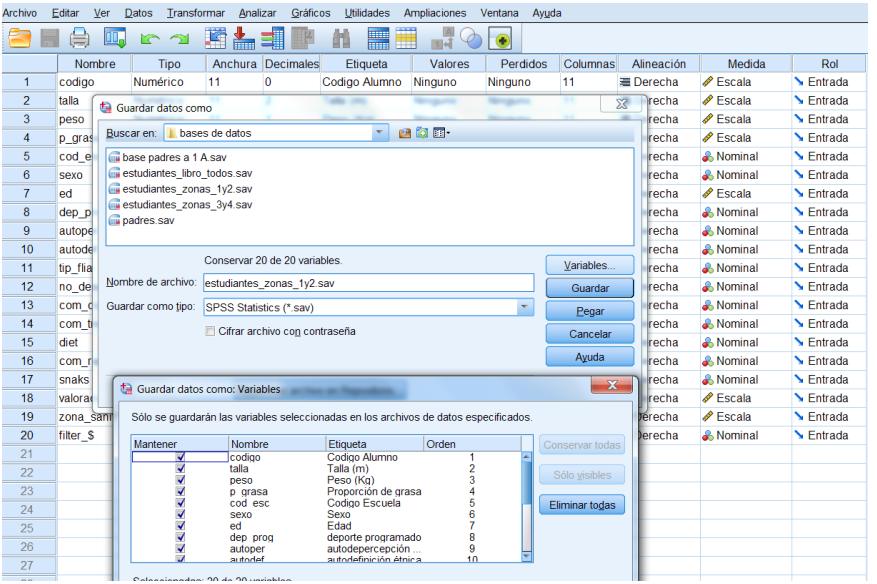


Figura 2.8 Ejemplo de ventana que se genera al guardar un archivo de datos en una carpeta del usuario

El lector debe tener mucho orden a la hora de guardar archivos para que, además de la fecha de creación, en el nombre quede constancia del contenido, ya que, en un análisis de un cierto volumen, el número de archivos que pueden llegar a generarse es muy elevado.

Ejercicio 2.1

Siga los pasos que se han explicado en este capítulo con el archivo estudiantes_zonas_1y2 .sav.

Especialmente le recomendamos que adquiera habilidad en el uso de la opción **Pegar** y su activación. Guarde los archivos generados y observe en qué carpeta se han grabado.

Repita el ejercicio hasta que logre que los tres archivos se encuentren en el mismo directorio.

Capítulo tres
Obtención de datos en SPSS

3.1 Lectura y traducción de archivos de datos

Tal y como se comentó en el capítulo anterior, el programa SPSS tiene un módulo de comunicación, el cual permite, entre otras acciones, la lectura de archivos caracterizados por la extensión **.sav**, es decir, archivos de lectura directa. Además, es capaz de leer archivos generados en otros formatos, especialmente en formato **.xls**[2], también puede leer formatos como Dbase, **.dbd**; .**dat**; **.txt**; o **.csv**.

No obstante, dada la popularidad del uso de la hoja Excel para la adquisición de datos, nos vamos a limitar a este formato. En cualquier caso, la opción **Abrir** que consta en la barra de herramientas de la ventana de datos ofrece múltiples opciones para acceder a otros tipos de formatos, tal y como se observa en la Figura 3.1.

2 Generados por el programa Excel.

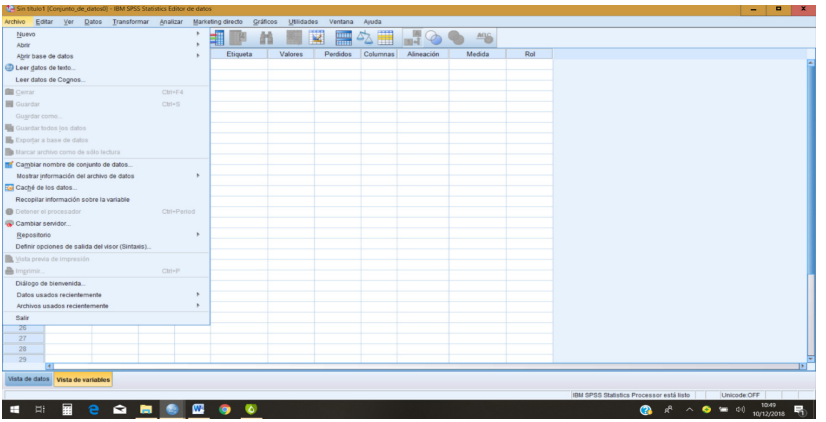


Figura 3.1 Ventana y opciones de abrir archivos

En la figura, se observa que esta opción, en caso de duda, también permite conocer el contenido de archivos en formato **.sav** sin abrirlos. Por ejemplo, ejecute la pestaña **Archivo ► mostrar información del archivo de datos ► archivo externo ►** y busque con el explorador el archivo de datos del que desee visualizar su información, tanto de variables como de las etiquetas de las mismas y de las categorías de las variables categóricas.

SYSFILE INFO 'C:\Users\usuario\Desktop\Libro SPSS Ecuador diciembre\bases de datos\estudiantes_zonas_1y2.sav'.

Así mismo, esta pestaña da opción para renombrar archivos ya almacenados o revisar archivos de datos de otra naturaleza usados recientemente, entre otras opciones.

Ejercicio 3.1

Usando las opciones de visualización del contenido de un fichero de datos, ejecute esas acciones para conocer el contenido de la información que contiene el archivo estudiantes_zonas_3y4.sav.

Logre la sintaxis que se corresponde a esas acciones.

3.2 Lectura y traducción de un archivo en Excel

En la carpeta de datos que se administró para el aprendizaje del programa, podrá observar que hay un archivo en formato Excel, con la extensión .**xlsx**, **padres.xlsx**.

Esta extensión es la que Windows asigna a una hoja en formato Excel 2007 a 2010. Si utiliza esta versión, abra el archivo con Excel y verá la información que contiene. Para ello active las pestañas **Archivo ► Abrir ► Datos** y busque con el explorador el archivo en la carpeta donde haya copiado los archivos de datos del curso, modificando en la pestaña de selección del tipo de archivo, que por defecto es .sav a .xls o .xlsx. Corresponde a los datos de los padres que han sido estudiados en los alumnos de las escuelas municipales de Quito.

Observe que el programa le muestra una ventana de datos traducidos al sistema SPSS y que ha mantenido las etiquetas de las variables que constaban en la hoja Excel.

Ejercicio 3.2

Compare la información que ha obtenido al abrir con SPSS ese archivo de datos en Excel y la que contenía el original, abriendo con Excel el mismo archivo.

Comentario:

En este momento es necesario que revise cómo se definieron las variables originalmente y si esas definiciones corresponden con sus intereses de análisis, ya que con frecuencia los creadores de una hoja Excel, mantienen por defecto la naturaleza de cadena, o formato alfanumérico, para facilitar la digitación. Debe, como se indicará más adelante, adecuar los formatos de cadena a formato numéricos o de fecha que serán los que utilizará durante el análisis o transformar las variables nominales, como sexo (masculino y femenino), a una variable nominal numérica sexo, en dos valores 1 y 2 y adjudicar una etiqueta, masculino o femenino, según corresponda.

La razón es que, el manejo de variables alfanuméricas es más complejo y con facilidad lleva a errores que posteriormente notaremos al efectuar el control de calidad de los datos.

3.3 Grabar un archivo de trabajo

3.3.1 Grabar en formato SPSS

El mismo módulo de comunicación, utilizando las pestañas desplegables de la ventana de datos permite grabar los archivos de trabajo, ya sean de datos como de resultados o de sintaxis. Asimismo, este módulo faculta la exportación de los archivos de datos a otros sistemas de análisis.

Estas pestañas se muestran en cualquiera de las ventanas abiertas, independientemente de su naturaleza, datos, resultados o sintaxis. El archivo guardado tendrá la extensión, **.sav**, **.spv** o **.sps** respectivamente. De esta manera, el archivo abierto con la acción **Archivo ►Abrir ► Datos ►padres.xls** será guardado con el nombre **padres.sav.**

Tenga mucho cuidado a la hora de guardar dichos archivos con la información en el nombre que los haga fácilmente reconocibles en un uso futuro, ya que si no les cambia el nombre puede sobreponer la información de un archivo anterior ya utilizado. Es recomendable pues que simultáneamente lleve un diario de trabajo en el que vaya indicando los nombres de los archivos, su contenido y la fecha de creación para evitar, especialmente, pérdidas de información.

3.3.2 Grabar en formato de otros programas

Con frecuencia, el analista debe grabar la información para ser analizada con otros programas, ya sean de soporte diferente de base de datos como en otros lenguajes correspondientes a diferentes paquetes de análisis estadístico.

La opción más sencilla es utilizar la pestaña desplegable **Archivo** seguida de la opción **Guardar como**. Observará que, con la ventana guardar como **tipo** puede guardar sus datos en Excel, Dbase, ASCII en diversas versiones, así como ya preparados para su análisis en SAS y Stata, además de la que por defecto marca como .sav, .sps o .spv; según la naturaleza del archivo de origen que pretende guardar datos ya interpretados o fichero activo, instrucciones de programa o fichero de resultados.

Un caso especial lo constituyen los archivos de resultados, los cuales pueden exportarse a un formato Word con el fin de facilitar su lectura, sin necesidad de abrir el programa, utilizando la pestaña de la barra de herramientas de la ventana de resultados  . El resultado de esta acción

produce un archivo en Word editable posteriormente, así como la exportación de gráficos a Power Point.

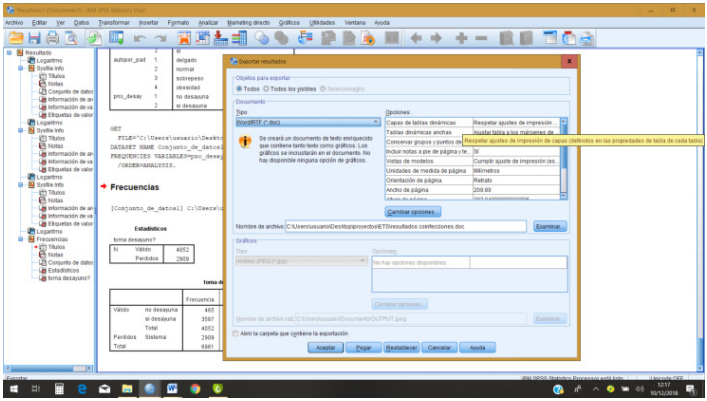


Figura 3.2 Sub ventana que se activa presionando sobre el quinto icono de la barra de herramientas de la ventana de resultados.

3.4 Fusión de archivos

Con frecuencia el analista se encuentra en la situación en la que la información repartida en diferentes archivos y debe reunirla en uno solo. Ya sea añadiendo casos y variables al primer archivo, tenemos dos situaciones que se activan con las pestañas **Datos ► Fusionar archivos.** Observará que se abren dos opciones, la primera permite añadir casos a un archivo ya abierto y la segunda añadir más variables a los casos.

3.5 Añadir casos

Los pasos adecuados son los siguientes:

1. Tener un archivo de datos activo. Sea por ejemplo el archivo de datos correspondientes a los estudiantes de las zonas 1 y 2, **estudiantes_zonas_1y2.sav**.

2. Aplicar, desde la barra de herramientas de la ventana de datos, **Datos ► Fusionar archivos ►añadir casos**. Con esta acción se le abrirá una ventana en la que le muestra si el archivo que tiene la información de los casos está ya abierto seleccionado o bien debe buscarlo externamente con el *browser* de la misma ventana.

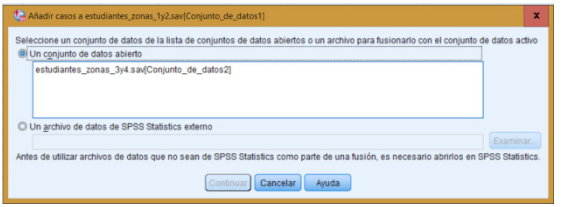


Figura 3.3 Ventana que se activa al ejecutar Datos ► Fusionar archivos

Una vez marcado el archivo que se desea —en este caso **estudiantes_zonas_3y4.sav**— al seleccionar la tecla Continuar, se abrirá una nueva ventana, que a su lado derecho indica las variables que estarán en el archivo; suma así, como las variables desemparejadas, es decir que se encuentran en un archivo, pero no en el otro. Estas variables no constarán en la suma ya que tendrían información faltante en el resultado final.

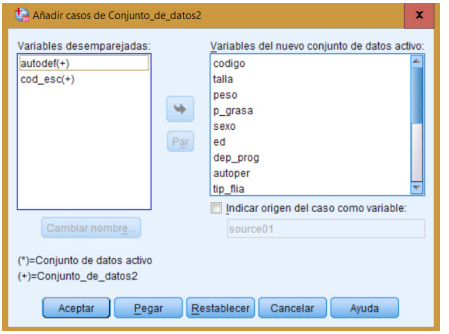


Figura 3.4 Ventana orientativa del archivo resultante al activar Continuar en la ventana anterior.

Puede observar que le ofrecen dos opciones. La primera que, en el archivo resultante, se añada una variable opcional, **indicar el origen del caso como variable**, con lo que siempre tendrá la información a cuál los archivos

pertenecen. La segunda es cambiar el nombre a las variables desemparejadas. Esta opción resuelve el problema que se presenta cuando una variable tiene nombres diferentes en cada uno de los archivos. Lógicamente aparecerían dos variables desemparejadas, por tener diferente nombre y aquí podría cambiarlos por uno común con lo que al estar ya aparejadas pasarían a formar como variable del nuevo archivo.

Con el fin de detectar el origen de las variables, estas llevan una marca (*) o (+) según pertenezcan al primer o segundo archivo abierto. Así, en el ejemplo que se muestra en la figura 3.4, las dos variables desemparejadas, marcadas con (+) pertenecen al archivo **estudiantes_zonas_3y4.sav**.

Comentario:

Si las variables están definidas en naturaleza diferente, por ejemplo, en un archivo Nominal y en el otro como Escalar, el resultado mantiene la naturaleza del archivo activo o abierto en primer lugar.

Si el tipo de variable es diferente, por ejemplo, en uno de tipo cadena o alfanumérico y en el otro tipo numérico, en la figura 3.4 aparecerán como si fuesen dos variables diferentes aunque tengan el mismo nombre y por lo tanto desaparejadas. Debe ajustar los tipos para que sean iguales de forma previa a la suma de casos. Debe asegurarse de que, especialmente si las variables son categóricas, tengan igual codificación, es decir los valores de las categorías. Por ejemplo si en el primer archivo la variable sexo tiene como valores 1 = hombre y 2 = mujer esta misma codificación debe ser la que la variable sexo tenga en el segundo archivo.

Por estas razones se recomienda que ejecute **Archivo ► mostrar información del archivo de datos ► archivo externo ►** de los archivos que quiere unir con el fin de asegurarse que tienen la misma información y que esta está estructurada de la misma forma.

La sintaxis asociada a estas acciones es:

```
DATASET ACTIVATE Conjunto_de_datos1.
ADD FILES /FILE=*
  /FILE='Conjunto_de_datos2'
  /RENAME (autodef cod_esc=d0 d1)
  /IN=zonas
  /DROP=d0 d1.
VARIABLE LABELS zonas
 'Case source is Conjunto_de_datos2'.
```

```
EXECUTE.
SAVE OUTFILE='C:\Users\usuario\Desktop\Libro SPSS Ecuador diciembre\
bases de datos\bases libro\estudiantes_total.sav'
 /COMPRESSED.
```

Ejercicio 3.3

Sume los casos correspondientes a la zona 3 y 4 a los de las zonas 1 y 2.

Previamente asegúrese, mediante la acción necesaria, de que ambos tienen la misma información y sepa de antemano qué variables estarán desaparejadas.

Indique en una nueva variable el origen de los datos y a continuación guarde el resultado en un archivo con el nombre **estudiantes_total.sav**.

Visualice la sintaxis utilizando la opción Pegar y posteriormente ejecútela.

Una vez realizada estas acciones, cerciórese mediante SYSFILE INFO 'C:\Users\usuario\Desktop\Libro SPSS Ecuador diciembre\bases de datos\estudiantes_total.sav'. o por ventanas del contenido del resultado de sumar los casos.

3.6 Añadir variables

A menudo la información de cada caso está repartida en archivos diferentes y nos interesa el que, en un solo archivo, se encuentren todas las variables que se pretenden analizar. Por ejemplo, en el caso de querer adjuntar la información obtenida por medio de los padres de los alumnos a cada uno de ellos, información que consta en el archivo **padres.sav**, obtenido por la lectura del archivo en Excel y guardado posteriormente. En los dos archivos, evidentemente, **estudiantes_total.sav** y **padres.sav** debe existir un índice que permita asignar de forma inequívoca la información de los padres al estudiante correspondiente. Este indicador debe recibir el mismo nombre en estos archivos de código.

El proceso requerido es **Datos ► Fusionar archivos ►añadir variables**, acción con la que se abre la ventana descrita en la figura 3.3. Sin embargo, el procedimiento solo es posible si previamente el índice o los índices identificadores están ordenados en el mismo sentido en los dos archivos. Para ello debe ejecutarse en cada archivo la acción **Datos ► Ordenar casos,** acción que abre una ventana que requiere la información en base a qué variable y si se ordena de forma ascendente o descendente.

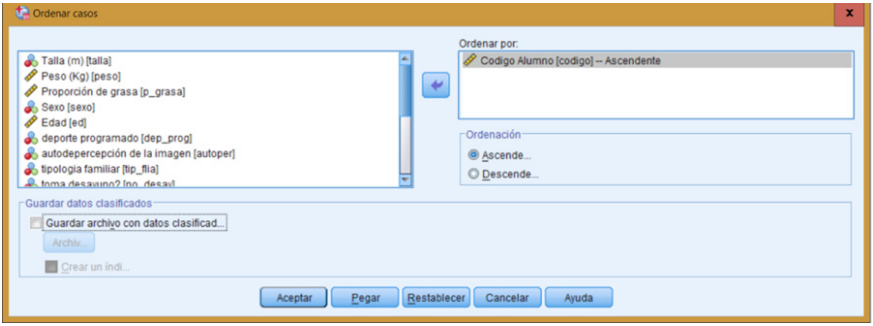


Figura 3.5 Ventana que aparece con la acción Datos ► Ordenar casos

Acción que se realiza con los dos ficheros como puede observarse en la expresión de sintaxis correspondiente:

```
DATASET ACTIVATE Conjunto_de_datos3.
SORT CASES BY codigo(A).
DATASET ACTIVATE Conjunto_de_datos4.
SORT CASES BY codigo(A).
```

El programa da un nombre secuencial a los archivos de datos que va abriendo, por lo que debe tener muy claro cuál es cada uno. En este caso, Conjunto_de_datos3 es el archivo de estudiantes y Conjunto_de_datos4 el de los padres. Este orden puede observarse en las ventanas de datos de ambos archivos y debe tener en cuenta que, si no los guarda, los mismos no quedarán registrados.

Una vez están ordenados por las mismas variables y en el mismo sentido **Ascendente o Descendente**, es el momento de repetir la acción **Datos ► Fusionar archivos ►añadir variables** y, si se ha empezado a trabajar con la base de estudiantes, se marca como archivo que tiene la otra información y se **Continúa** (figura 3.3). El resultado es una nueva ventana que muestra la siguiente información:

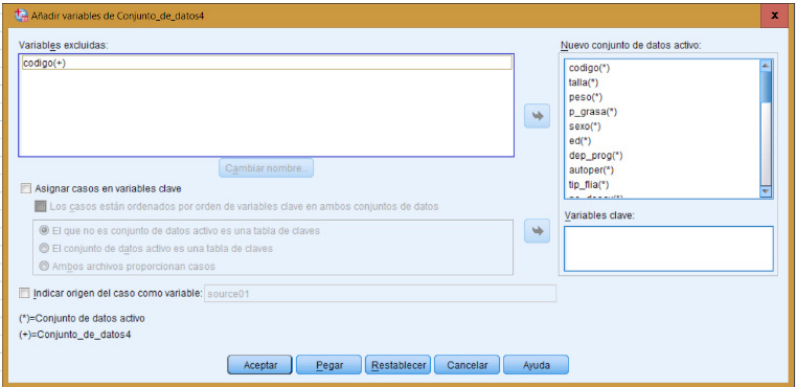


Figura 3.6 Ventana de fusión de suma de variables

En la ventana de la derecha constan las variables resultado de la suma de variables y que estarán en el nuevo archivo. Puede eliminar variables marcándolas y, con el cursor, desplazarlas a la ventana, más reducida de variables excluidas.

Observará que, de entrada, se ha excluido la variable **código** debido a que está en los dos archivos de origen de las variables. Esta variable es precisamente el índice por el cual se añade la información a cada caso, es decir, a cada estudiante. También da la opción, si el índice tiene nombre diferente en cada base, de cambiar su nombre en este momento.

A continuación, se marca **Asignar casos en variable clave,** seleccionar también la variable índice, código en este ejemplo, y la transporta con la flecha inferior a la ventana de la derecha inferior **variables clave,** indicando que los datos ya están ordenados en los dos archivos.

Comentario:

Variables clave está en plural porque es frecuente tener dos o más índices de referencia. Por ejemplo, suponga que hubiésemos tenido el indicador de escuela y hubiésemos ordenado por código dentro de cada escuela. Tendríamos dos índices los cuales pasaríamos en el mismo orden en que se han utilizado. Es decir, si primero se hubiese ordenado por escuela y luego por un código dentro de cada escuela, se pasaría primero el índice escuela y después el de código.

Figura 3.7 Asignación de la variable clave

Se le ofrecen tres opciones.

Las dos primeras suponen que uno de los archivos es una tabla, en la cual el indicador, la variable clave, busca la información en el segundo archivo. Esta acción se puede realizar en dos sentidos: desde el archivo no activo —que en este caso sería **Padres.sav**— o a partir del archivo activo **estudiantes_total.sav**, creado en pasos anteriores, y mediante este indicador buscaría la información del padre correspondiente.

Una tercera opción es que los dos archivos aportan casos con lo que se podría fundir en ambos, aunque los indicadores no coincidiesen. Esta acción no es muy recomendable. En este caso, lo lógico es que estudiantes, es decir, el archivo activo, sea la tabla con las claves que hay que buscar en el segundo, llamado no activo, que contiene la información de sus padres. La acción de Pegar nos ofrecería la sintaxis siguiente:

```
MATCH FILES /TABLE=*
  /FILE='Conjunto_de_datos4'
  /BY codigo.
EXECUTE.
```

Comentario:

Recomendamos al estudiante que, al menos hasta que no tenga una familiaridad con el programa, no abra más de una base de datos simultáneamente.

Una vez realizada la fusión, guarde y grabe el archivo resultante. Con el nombre **Estudio.sav**.

Ejercicio 3.4

Añada las variables que informan los padres a cada estudiante y pegue la sintaxis correspondiente al proceso.

Investigue la información de las variables y su tipología del fichero resultante.

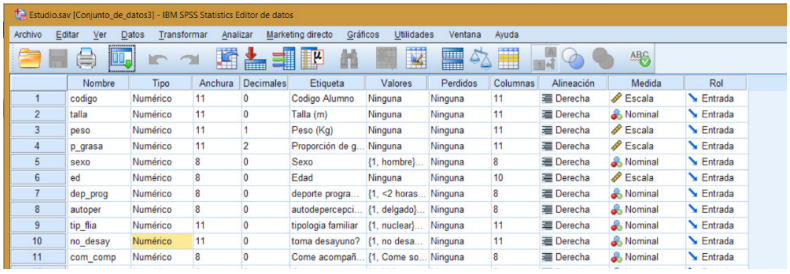
Capítulo cuatro

Definición, creación, naturaleza y etiquetas de las variables

4.1 Variables. Definiciones

En toda base de datos, es crucial la definición con gran rigor de las características de las variables que la configuran. No puede abordarse un análisis sin definir la naturaleza de cada variable, es decir, su tipo, el formato en que está expresada, así como la forma de presentación, los valores que significan desconocimiento del valor o que no han sido medidos, y cómo debe ser considerada la medida.

Todos estos aspectos, así como la codificación de los valores si la variable es categórica, se encuentran descritos en la ventana de variables de la base de datos .sav.



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	codigo	Numérico	11	0	Codigo Alumno	Ninguna	Ninguna	11	Derecha	Escala	Entrada
2	talla	Numérico	11	0	Talla (m)	Ninguna	Ninguna	11	Derecha	Nominal	Entrada
3	peso	Numérico	11	1	Peso (Kg)	Ninguna	Ninguna	11	Derecha	Escala	Entrada
4	p_grasa	Numérico	11	2	Proporción de g...	Ninguna	Ninguna	11	Derecha	Escala	Entrada
5	sexo	Numérico	8	0	Sexo	{1, hombre}...	Ninguna	8	Derecha	Nominal	Entrada
6	ed	Numérico	8	0	Edad	Ninguna	Ninguna	10	Derecha	Escala	Entrada
7	dep_prog	Numérico	8	0	deporte progra...	{1, <2 horas...	Ninguna	8	Derecha	Nominal	Entrada
8	autoper	Numérico	8	0	autodepercepci...	{1, delgado}...	Ninguna	8	Derecha	Nominal	Entrada
9	tip_flia	Numérico	11	0	tipologia familiar	{1, nuclear}...	Ninguna	11	Derecha	Nominal	Entrada
10	no_desay	Numérico	11	0	toma desayuno?	{1, no desa...	Ninguna	11	Derecha	Nominal	Entrada
11	com_comp	Numérico	8	0	Come acompañ...	{1, Come so...	Ninguna	8	Derecha	Nominal	Entrada

Figura 4.1. Aspecto de la ventana de vista de variables y sus características

Como puede observarse en la figura 4.1., en esta ventana se muestran 28 variables que conforman el archivo **Estudio.sav.** Las propiedades que las caracterizan se encuentran en la parte superior, bajo de la barra de herramientas, y son las siguientes.

4.1.1 Nombre

Texto con el cual se reconoce a la variable. Acostumbra a ser un texto nemotécnico que nos recuerde el contenido de la misma. Generalmente no tiene mayor longitud que 10 caracteres y es el nombre que, a lo largo del estudio, será utilizado en los análisis.

4.1.2 Tipo

Indica si la información está expresada en números, caracteres alfanuméricos o cadena, *string* en las versiones en inglés, o es una fecha. Para ver los tipos de variables puede presionar sobre cualquier valor de esta segunda columna y verá que se despliega una pestaña indicando los posibles tipos de variables, así como el ancho y el número de decimales con los que se expresa los valores. Estas dos últimas características puede también modificarlas presionando en cualquier valor de esas columnas, Anchura y Decimales.

4.1.3 Etiqueta de la variable, variable label

Es el nombre completo de la variable y que constará en los resultados de los análisis. Puede tener hasta 64 caracteres, si bien se recomienda que no tenga más de 10, pues, si no, posteriormente distorsiona la presentación de los resultados.

Se utiliza cuando el Nombre de la variable usado en la primera columna es un código o no es claro, pues emplean textos nemotécnicos. Evidentemente si el nombre ya es aclaratorio de la variable, puede prescindirse de la etiqueta. Por ejemplo, si el nombre de la variable es Sexo, no es necesario explicitarlo en la etiqueta, lo mismo ocurriría con la edad, a no ser que se quisiese especificar que la edad está expresada en meses. La definición por sintaxis sería:

```
VARIABLE LABELS P_GRASA 'PROPORCIÓN DE GRASA CORPORAL'.
```

Es decir, se indica la etiqueta de la variable con nombre p-grasa entre comillas simples inglesas. Como toda expresión de sintaxis debe finalizar con un punto, signo de fin de sentencia en SPSS.

4.1.4 Etiquetas de los valores, value labels

Si la variable es categórica pero expresada en tipo numérico, es imprescindible añadir las etiquetas de los valores numéricos. Por ejemplo, observe la variable Sexo. Si mira los valores en la ventana de datos, verá que los mismos en las categorías son 1 y 2. Con el fin de evitar equívocos u olvidos de cómo se codificó, en valores se añade 1 'hombre' 2 'mujer'. En sintaxis, se escribiría:

```
VALUE LABELS SEXO 1 'HOMBRE' 2 'MUJER'.
```

En el archivo, queda registrado si lo escribe en mayúsculas o minúsculas. Esta acción también puede realizarla directamente sobre la ventada de variables, presionando en la casilla correspondiente a la columna de valores de la variable sexo.

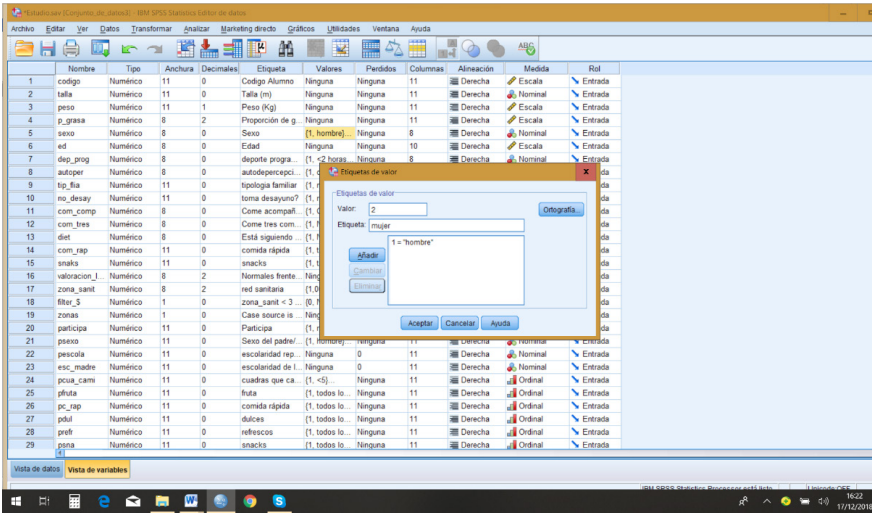


Figura 4.2 Adición por ventanas de etiquetas de valores a una variable

Al ejecutar por ventanas debe saber que no hay forma de que quede constancia en sintaxis, lo cual no es muy adecuado porque otra persona puede interpretar las etiquetas de forma diferente. Al hacerlo por sintaxis es mucho más seguro, tanto cuando se trabaja por primera vez un archivo, así como al haber adquirido experiencia. Obviamente, las variables numéricas escalares no poseen etiquetas.

4.1.5 Valores perdidos, missing values

En esta columna, constan aquellos valores de la variable que se corresponden a falta de información, *missing* de sistema, o bien aquellos valores que se ignoran en el análisis. Ya sea porque son valores erróneos o porque se quieren ignorar en cierto análisis, *missing* de usuario. Por ejemplo, consideremos que, en este estudio por la edad definida de los estudiantes no pueden tener edades superiores a 19 años.

Si presionamos la casilla Perdidos de la variable edad (**ed**) vemos que no hay definido ningún valor perdido. Podríamos definir un valor de edad entre 19 y 25 como valores perdidos, o específicamente definir 88 como tal si así lo hubiésemos codificado por común acuerdo, para aquellos valores en los que el estudiante no responde a su edad.

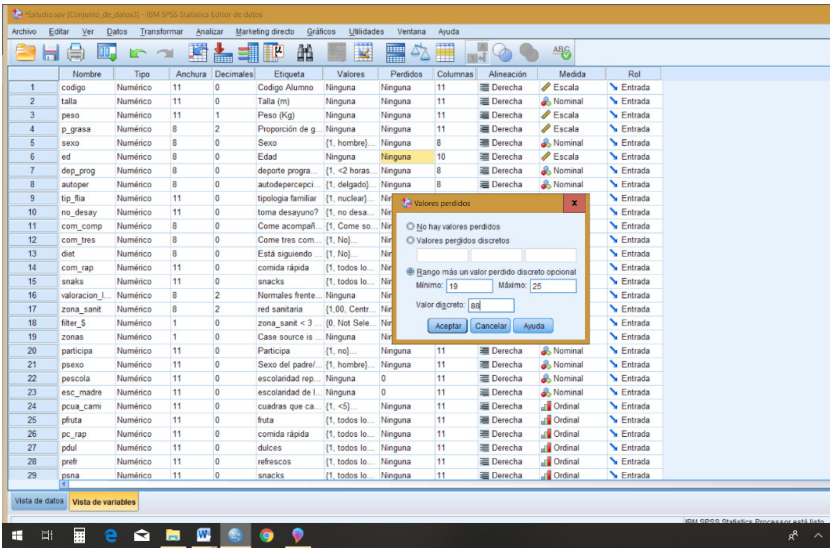


Figura 4.3 Definir por ventanas los valores faltos de información o *missing* de usuario

Si el valor *missing* es un valor discreto, como 88, es recomendable añadir a las etiquetas de valores al mismo y ponerle la etiqueta de *missing* de usuario. Si en la ventana de datos no hay información aparece un punto, ".", lo cual indica *missing* de sistema y no debe definirse. La expresión de sintaxis que define los *missing* es:

```
MISSING VALUES ED (88) (19 TO 25).
```

4.1.6 Columnas

Se refiere a con cuántas columnas se quiere visualizar los valores en la ventana de datos. No modifica en absoluto el valor de la variable, si no solo su visualización. Si el número de columnas a visualizar es inferior al ancho definido en la columna Anchura, en la ventana de datos aparecerá como asteriscos.

4.1.7 Alineación

También relacionado a la visualización de los datos de esa variable en la ventana de datos, si alineados a la derecha o a la izquierda o centrados. No repercute en los análisis posteriores.

4.1.8 Medida

Indica la naturaleza de la variable escalar, nominal u ordinal. Puede modificarse presionando sobre la casilla correspondiente.

4.1.9 Rol

Campo que se utiliza en algunos programas muy definidos para indicar si la variable es independiente, dependiente o se usa para segmentar. No se emplea con frecuencia ya que en módulo de análisis siempre se pregunta cuál es la variable dependiente o independiente y ese concepto puede cambiar en cada análisis o ser indefinido.

Todas estas características pueden definirse mediante la pestaña **Datos ► Definir propiedades de las variables**, seleccionando cada variable e introduciendo las características sin que estas acciones puedan incorporarse a la sintaxis.

4.2 Definición de valores desconocidos, *missing values*

En un estudio, es frecuente que en alguno de los casos estudiados no exista la información de alguna variable. Esto es lo que, en la nomenclatura del programa, se conoce por un valor *missing*. Existen dos tipos de valores *missing* en la matriz de datos.

4.2.1 System missing

Son aquellos valores en los que la información se ha dejado en blanco al introducir los datos y, en ese caso, en la variable, el valor está en blanco si la variable está definida como tipo cadena o bien con un punto decimal si la variable se definió como tipo numérico.

Esta particularidad debe considerarse por el hecho de que, si se desconoce el valor de la variable, cualquier transformación efectuada con este tipo de información debe tenerse en cuenta de forma que, posteriormente, no se confunda con un valor correcto de la variable, principalmente en caso de variables categóricas o de efectuar divisiones por cero si no se declara como tal.

Por ejemplo, una variable como sexo, si en la matriz de datos no se declara *missing* y esta se define como categórica, el valor de la variable en blanco será considerado como valor válido, desvirtuando todos los resultados de su descripción. Esto ocurre porque, en una variable categórica alfanumérica, un blanco es un valor alfanumérico como otro cualquiera. Debe definir, como se explica más adelante, que ese valor en blanco es un valor sin información o *missing* del sistema.

4.2.2 Missing *de usuario*

Con frecuencia, si de una variable nominal, pero expresada numéricamente, se desconoce su verdadero valor, existe la tendencia de codificarla con un valor imposible como 9, 99, 999, valores que deben definirse también como *missing*.

Si la variable es escalar como la edad, se codifica con 999 cuando se desconoce o bien 888 para expresar que el individuo no respondió a la pregunta o no lo sabe. Estos valores también pueden ser definidos como *missing*, ya que, en caso contrario, serán tomados como valores reales y lógicamente desvirtuarán posteriormente propiedades de la variable como su valor de media o desviación estándar. En otras ocasiones se pueden definir rangos de valores medidos, codificados, pero imposibles.

Por ejemplo, suponga que se ha pedido en el cuestionario responder la edad de la madre de los estudiantes y que, observando los valores, existen edades superiores a los 70 años en cierto número de casos. En esta situación se podrá definir un rango de valores que se deben considerar *missing* y que los define usted como usuario.

Esta misma acción puede utilizarse con precaución si, en un momento determinado, quisiera considerar solo las edades de madres inferiores a 50 años. Sin embargo, podría ser que esta exclusión fuese momentánea y posteriormente considerar todos los valores. Tendría que deshacer esa definición de *missing* de usurario con el fin de no perder la información en el análisis.

Estas definiciones puede llevarlas a cabo en la ventana de Vista de variables presionando en la casilla Missing o Perdidos de la variable de interés —por ejemplo, edad—; se abrirá la siguiente ventana.

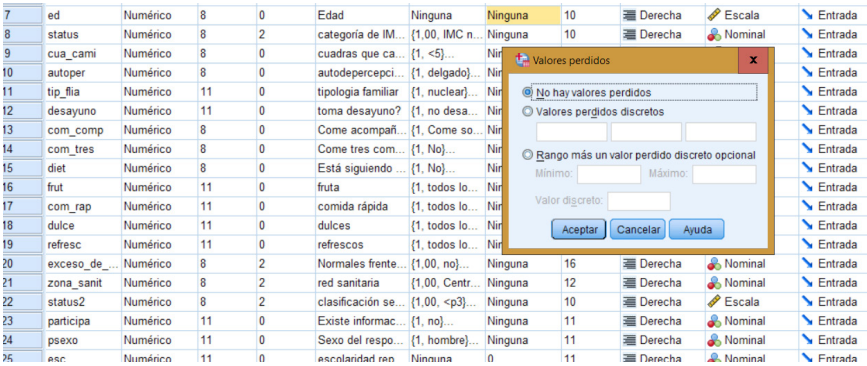


Figura 4.4 Selección y definición de valores perdidos definidos por el usuario

En esta Figura, al presionar sobre la casilla correspondiente a la columna Perdidos en la variable ed se abre la ventana en la que indica que no hay definidos por el usuario valores perdidos y da la opción de definir hasta tres valores discretos de edad o bien a los valores contenidos en un rango más un valor discreto.

Ejercicio 4.1

Ejecute la instrucción una vez abierto el archivo **Estudio.sav**.

FREQUENCIES VARIABLES=ed/STATISTICS=MINIMUM MAXIMUM MEAN /ORDER=ANALYSIS.

Observará que se muestra el valor máximo y mínimo así como la edad.

- ¿Existe algún estudiante sin definir la edad?
- ¿Observa algún valor erróneo? Resuelva la situación definiendo ese valor como perdido por el usuario, y vuelva a ejecutar la sintaxis.

Comentario:

La situación que ha encontrado en el ejercicio anterior se da con mucha frecuencia y es debido fundamentalmente a un error de entrada de datos.

En ningún caso debe suponer el verdadero valor, así que, después de definir ese valor como perdido por el usuario, debe contactar con el responsable de creación

de la base de datos y solicitar, preferentemente por escrito, que revisen ese caso y le brinden el verdadero valor, para cambiarlo en la ventana de datos.

Solo en el caso en que el error viniese del propio encuestador mantendrá el valor de edad como perdido por el usuario, sin eliminar el caso, ya que el resto de la información del mismo tiene utilidad para otros análisis y perdería información en caso de eliminarlo.

4.3 Creación y modificación de variables

En múltiples ocasiones, el investigador debe crear nuevas variables a partir de las que ha obtenido del cuestionario, así como recodificar en categorías variables escalares o agrupar categorías en variables nominales. También puede generar otras variables diferentes condicionadas al valor de otras existentes.

Todas estas acciones se encuentran a su disposición en la pestaña Transformar de la ventana de datos. Al desplegar la pestaña **Transformar**, se le ofrecen diversas posibilidades. Vamos a proceder con la descripción de las más frecuentes.

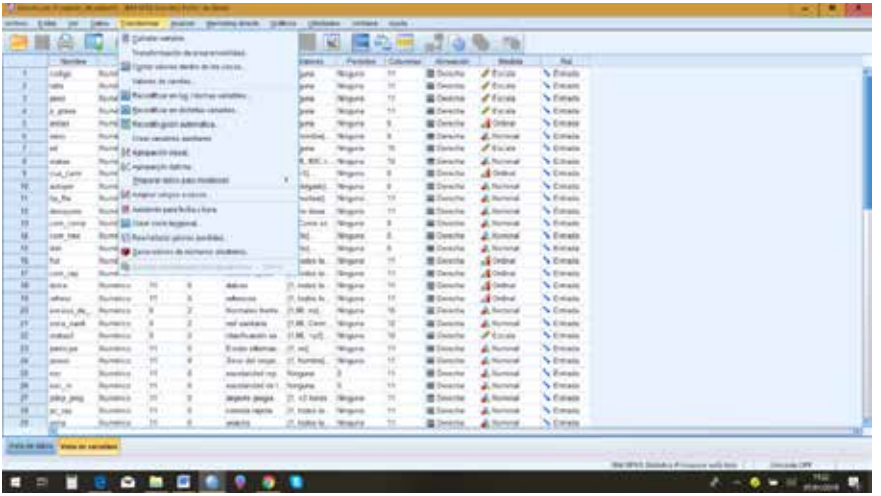


Figura 4.5 Opciones que se incluyen en la pestaña Transformar

4.3.1 Calcular variable

Mediante esta opción, el usuario tiene diferentes posibilidades. Crear una variable nueva con valor constante para todos los casos. Aparentemente no debería tener interés porque no sería una variable sino una constante y, por lo tanto, sin interés para el análisis de variabilidad, pero se acostumbra a usar esta opción para posteriormente cambiar el valor en función de otras variables. El esquema general se muestra en la figura 4.6, la cual se obtiene al ejecutar las acciones **Transformar ▸ Calcular variable**.

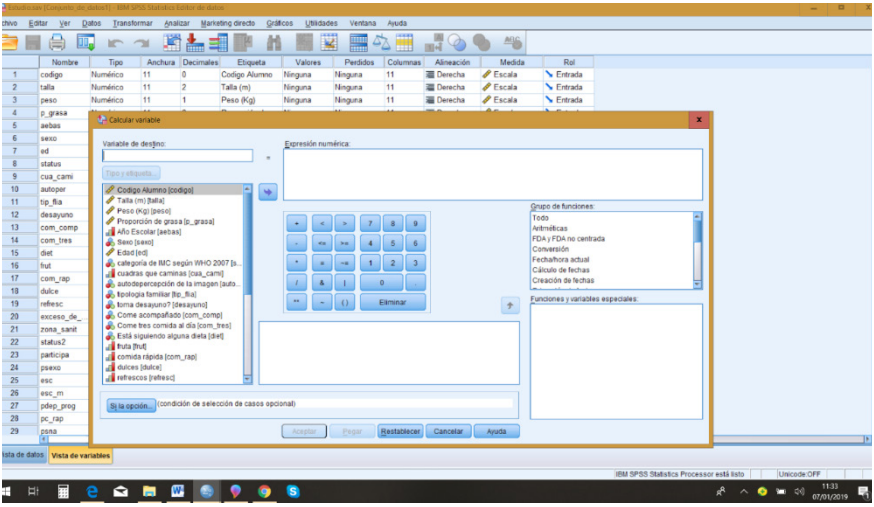


Figura 4.6 Subventana de Calcular variable

Lo primero que puede observar es que hay un espacio para nombrar a la variable nueva o variable destino. A su derecha, está escrito el símbolo de igualdad y a su derecha la expresión de tipo numérico que empleará para generar esta nueva variable. El nombre no puede contener espacios en blanco o utilizar signos de puntuación. En caso de querer que el nombre sea la composición de dos palabras, únalas con guion bajo. Por ejemplo, edad padres no será un nombre aceptado, pero si, edad_padres.

También se puede ver que se abre una ventana con el nombre de todas las variables que constan en el archivo, con el símbolo correspondiente a si son escalares o nominales o variables de tipo fecha.

Seleccionando con el cursor qué variables existentes desea usar para definir la nueva variable destino, las puede trasladar a la ventana de Expresión numérica y, mediante una operación matemática o una función, unirlas a otra variable de las existentes y crearla al aceptar la operación. Además, de las cuatro reglas aritméticas (suma, resta, multiplicación y división) que se muestran en el teclado en la parte derecha, existe una gran variedad de funciones de transformación que, en diversos ejemplos, veremos su utilidad.

Ejercicio 4.2

En los estudios que analizan el sobrepeso y obesidad en sujetos de gran variabilidad en sexo y en edad, el peso por sí solo no es un elemento que indique casi nada del estatus de obesidad de los individuos, ya que, como el peso incluye la masa ósea, además de la muscular y la grasa, entre otros, la edad en fases de crecimiento introduce grandes cambios en el componente óseo. Por esta razón, se define un indicador que relaciona el peso con la talla del individuo, indicador que se conoce como Índice Másico Corporal (IMC).

Se define como el peso del individuo dividido para la talla expresada en metros al cuadrado. Es más práctico realizar esta creación de IMC a partir de los datos, peso y talla existentes en la base de datos que calcularlo para cada individuo y entrar la variable previamente como una más.

Así, utilizando **Transformar ► Calcular variable** cree la variable IMC para todos los estudiantes participantes. Compruebe mediante la opción **Pegar** que la sintaxis que corresponde a su acción es

```
DATASET ACTIVATE Conjunto_de_datos1.
COMPUTE IMC=peso / (talla) ** 2.
EXECUTE.
```

Realice el análisis de frecuencias de la nueva variable IMC mediante **Analizar ► Estadísticos descriptivos ► Frecuencias** y seleccione las opciones **Estadísticos** el **mínimo** y el **máximo**. Interprete los resultados.

En resumen, la instrucción genérica es:

```
COMPUTE nombre variable = expresión.
```

En la que debe especificarse la expresión que regula la creación de la variable.

***Expresión*:** indica cualquier operación que involucre constantes o variables del fichero mediante los operadores descritos en la tabla 4.1. Como

se ha indicado anteriormente, existen una serie de funciones transformadoras disponible para los usuarios. Algunas de estas funciones, que la experiencia nos indica que son las más utilizadas, se describen a continuación.

Tabla 4.1 Operadores de transformación

OPERADORES		DESCRIPCIÓN
Aritméticos		
+		Suma
-		Resta
*		Producto
/		Cociente
**		Exponente
Relacionales		
EQ (Equal)	=	Igual
NE (Non Equal)	<> o ~=	Diferente o No es igual
LT (Less Than)	<	Menor que
LE (Less than or Equal to)	<=	Menor o igual que
GT (Greater Than)	>	Mayor que
GE (Greater than or Equal to)	>=	Mayor o igual que
Lógicos		
AND	&	Se cumplen las dos condiciones
OR	|	Alguna condición es cierta
NOT	~	Condición falsa o excluyente

Las funciones transformadoras más usuales más allá de las descritas con los operadores de la tabla 4.1, incluyen un conjunto de transformaciones pre programadas que permiten la manipulación de los resultados numéricos, así como la creación de variables a partir de otras de naturaleza cadena, fechas,

o que generan propiedades estadísticas sencillas. Los operadores relacionales se utilizan para imponer condiciones a la transformación. En la tabla 4.2 se muestran ejemplos de las funciones de uso más frecuente.

Comentario:

Al dar el nombre a la nueva variable debe procurar que no coincida con el nombre de alguna variable ya existente, ya que la ejecución de la acción reemplazaría la información existente.

Tabla 4.2 Funciones transformadoras de uso más frecuente

Función	Resultado	Descripción	Ejemplos Valor,V, resultado de la acción
ABS (expr_num)	Numérico	Determina el valor absoluto de expr_num, el cual será un valor numérico.	V= - 3 ; ABS(V)=3
EXP (expr_num)	Numérico	Valor de la exponencial de expr_num. Atención: si expr_num es muy grande, los resultados pueden exceder la capacidad máxima de la memoria.	V= 1; EXP(V)=2,7183
LN (expr_num)	Numérico	Calcula el logaritmo en base e de la expresión numérica expr_num.	V= 2,7183; LN(V)=1
LG10 (expr_num)	Numérico	Efectúa el logaritmo en base 10 de la expr_num.	V= 10; LG10(V)=1
SQRT (expr_num)	Numérico	Función que determina la raíz cuadrada positiva del número.	V= 9; SQRT(V)=3
TRUNC (expr_num)	Numérico	Devuelve la parte entera del valor de expr_num.	V= 7,86; TRUNC(V)=7
LAG (variable,ncasos)	Numérico o alfanumérico	Devuelve el valor de la variable del caso que está situado n casos antes en el fichero. Atención: para los n casos primeros del fichero, el resultado es *missing* de sistema (si V es variable numérica) o espacios en blanco (si V es variable alfanumérica). Por defecto n casos=1.	V LAG(V) LAG(V,2) 6 8 6 1 8 6 2 1 8 5 3 1
CONCAT (expr_alf,expr_alf[,...])	Alfanumérico	Genera una cadena, que es la concatenación de todos los argumentos expr_alf indicados.	V1 V2 CONCAT(V1,V2) a b ab
INDEX (cadena,'subcadena')	Numérico	Crea un indicador entero según la posición del carácter inicial la subcadena buscada en la cadena analizada. Solo muestra la primera aparición, es decir, si la subcadena está otras veces lo ignora. Retorna 0 si la subcadena no aparece en la cadena.	V1 INDEX (V1,'+') -+- 2 +-+ 1 ++- 1 +++ 1 --- 0 -++ 2

LTRIM (expr_alf)	Alfanumérico	Suprime de la expr_alf los espacios en blanco en los caracteres de la izquierda. Devuelve el resultado sin ellos.	V1 LTRIM(V1) b-+- -+- b significa espacio en blanco
RTRIM (expr_alf)	Alfanumérico	Suprime de la expr_alf los blancos al final de la cadena y devuelve el resultado sin ellos.	V1 RTRIM(V1) NMNb NMN b significa espacio en blanco
SUBSTR (expr_alf,pos,long)	Alfanumérico	Crea una variable alfanumérica con los long caracteres que se encuentran a partir de la posición pos de l'expr_alf .	V1 SUBSTR(V1,4,3) Abcdefgh def
CTIME.DAYS (valortiempo)	Numérico	Para cada una de les fechas incluidas valor tiempo calcula los días que transcurridos desde el 15 de octubre de 1582. Luego efectúa les operaciones indicadas y retorna el número de días resultantes. Útil para registrar los días entre dos fechas.	V1=21-12-2000; V2=10-12-2000; CTIME.DAYS(V1-V2)=11
DATE.DMY (día,mes,año)	Fecha	Retorna la fecha especificada día, mes y año, datos que deben existir en tres variables diferenciadas. Así, coloca en una sola variable una fecha que estaba expresada en tres variables distintas. Para visualizar correctamente la nueva variable, debe asignarla previamente un formato DATE.	VD=18; VM=6; VA=1974; DATE.DMY(VD,VM,VA)=18-6-1974
DATE.YRDAY (año,num_día)	Fecha	Rehace la fecha correspondiente al año y número de día del año existente en dos variables previamente definidas. También debe asignar con anterioridad a la nueva variable un formato de tipo DATE.	VD=27; VA=2002; DATE.YRDAY(VA,VD)=27-01-2002
YRMODA (año,mes,día)	Numérico	Calcula el número de días desde el 15 de octubre de 1582 hasta la fecha representada por los argumentos año, mes y día.	VD=16; VM=10; VA=1582; YRMODA(VA,VM,VD)=2.
NMISS (variable[,...])	Numérico	Recuenta cuantos missings de sistema y usuario existen entre las variables descritas en el argumento.	V1 V2 V3 NMISS(V1,V2,V3) 10 , 55 1

Uno de los aspectos más complejos que encuentra el estudiante que inicia en SPSS es la generación de variables tipo fecha.

Comentario:

Las variables fecha, date, son de naturaleza numérica y son siempre el tiempo transcurrido desde el día que se efectuó el cambio de calendario por el sistema gregoriano (el 15 de octubre de 1582). En esa fecha se "perdieron" 15 días lo cual motivó la existencia de diversos calendarios en diferentes países, como por ejemplo, todos los países en los que la religión cristiana seguía el rito ortodoxo ruso.

En la revolución soviética de octubre en nuestro calendario ocurrió realmente el 7 de noviembre o el día de Navidad se sigue celebrando en el día 7 de enero en estos países.

Internacionalmente se sigue el calendario gregoriano a todos los demás efectos, si bien múltiples entornos culturales mantienen su calendario nacional cultural, como el calendario en los países musulmanes que empiezan a contar desde el 10 de julio del año 622 o año en que Mahoma inició la hégira o huida de la Meca y con la particularidad de que, al ser sus meses lunares, cada 33 años en ese sistema de calendario se corresponden a 32 gregorianos. En la actualidad, el año 2019 gregoriano corresponde con el año 1439 musulmán, También los pueblos de China tienen otro calendario interno aunque oficialmente se usa el gregoriano desde 1911. Otros calendarios se utilizan en diferentes culturas como la hebrea, en la que, además de utilizar meses lunares, su origen es el bíblico de creación del mundo, por lo que está en el año 5777.

Una forma sencilla y que difícilmente lleva a errores consiste en ejecutar la siguiente sintaxis:

```
STRING fecha (Ann).
COMPUTE fecha = 'fecha en el formato que se pretende expresar'
EXECUTE.
```

Por ejemplo:

```
STRING fecha (A11).
COMPUTE fecha = "21.12.2018".
EXECUTE.
```

En primer lugar, en esta expresión, que debe ejecutarse desde una ventana de sintaxis, se define una variable como cadena o *string*. Entre paréntesis se indica que el formato es alfanumérico A, y que se le reservan 11 espacios. Debe siempre indicarse un número de espacios de forma que luego la fecha tenga igual o menos caracteres.

Por ejemplo, si la fecha se expresa con la modalidad día, mes, año, puede tener diferentes expresiones como dd.mm.aaaa que son 10 caracteres o dd.mm. aa que son ocho. Así mismo existen otras modalidades y debe considerarlo a la hora de definir la variable como *string*.

A continuación, se efectúa la transformación COMPUTE, asignando la fecha que, como se ha definido como cadena debe asignarse entre comillas, siempre que se asigna un valor alfanumérico. En este caso, el 21 de diciembre

del año 2018. Es muy importante tener presente siempre el criterio para definir la fecha, ya que existen múltiples formatos que posteriormente son complejos de manipular si no se corresponden con el formato pensado.

Al ejecutar estas instrucciones observará que, en la última columna de la ventana de datos, aparece la variable fecha y que, para todos los individuos, ese valor es 21.12.2018; sin embargo, si mira el tipo de variable en la ventana de vista de variables, observará que la variable fecha está definida como cadena. Debe cambiar el tipo seleccionando una vez presionada esa casilla, la opción fecha y seleccionar, entre todas las que se le ofrecen, aquella que usted piensa utilizar, es decir, **dd.mm.aaaa**. Una vez convertida la variable a fecha, puede cambiar el formato a la forma de expresión de fecha que le sea más cómoda, por ejemplo aaaa.mmm.dd.

Si seleccionase otro formato de fecha no se correspondería con la estructura de variable cadena creada y, en la variable fecha, aparecería un punto decimal, es decir, valor perdido por el sistema. Por esta razón, es muy importante decidir previamente qué forma de expresión de la variable quiere utilizar.

Otra forma de definir una fecha sería utilizar la expresión de transformación DATE.DMY (día, mes, año) y ejecutar. El resultado es una variable numérica que indica el tiempo en formato SPSS, es decir, el tiempo transcurrido desde el 15 de octubre de 1582. Una vez obtenida, modificando el Tipo en la ventana de vista de variables, seleccionando fecha, podrá fijar la forma de expresión de la variable, por ejemplo dd.mm.aaa.

Ejercicio 4.3

Una vez haya creado dos variables que, por ejemplo, correspondiesen la fecha de la encuesta y a la fecha actual, mediante la opción **Transformar ► Calcular variable**, determine el tiempo transcurrido desde la fecha de encuesta y la fecha actual.

Pegue y analice la sintaxis creada al pegar en la pantalla.

En la figura adjunta, se muestra el procedimiento por ventanas.

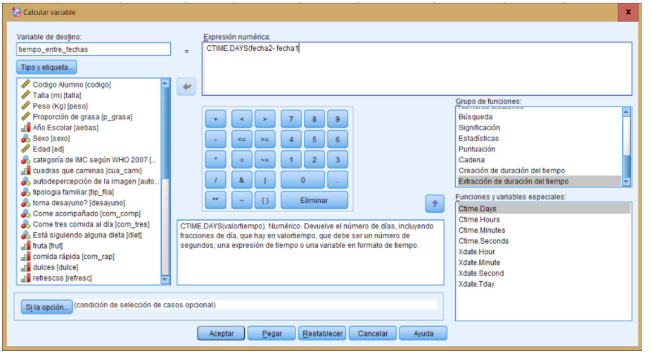


Figura 4.7 Subventana de calcular variable en la que se muestra el grupo de Funciones para determinar el tiempo transcurrido entre dos fechas

4.3.2 Calcular una variable de forma condicionada

Con frecuencia, el analista debe crear una nueva variable o modificar una ya existente en función de los valores que tiene otra variable, la cual condicionará el nuevo valor. Para ello se utilizan las funciones relacionales.

Existen diversas formas de efectuar el proceso, a partir de las pestañas de la barra de Herramientas, ► **Transformar** ► **Calcular variable** y proceder de la misma forma que en apartado crear una variable. La diferencia está en que a continuación, en la subventana, observará un recuadro a la izquierda en la parte inferior que dice **Si la opción**.

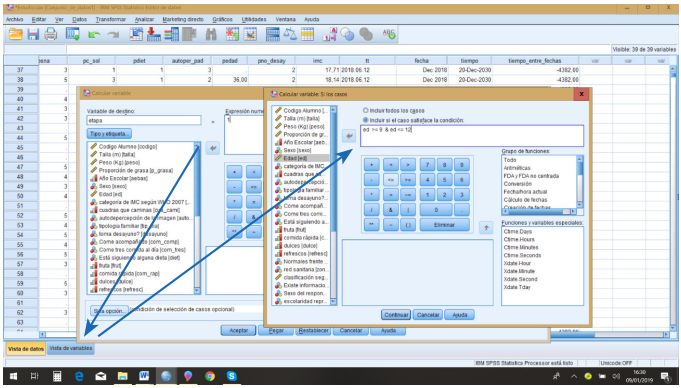


Figura 4.8 Ejemplo de asignación del valor 1 a una nueva variable de nombre etapa si la opción de edad está comprendida entre las edades 9 y 12 (Edad es la variable condicionante)

En la figura 4.8, se muestra el recorrido de creación de una variable nueva de nombre etapa con valor igual a 1 si los estudiantes están entre nueve y 12 años, ambos incluidos. La sintaxis asociada es:

```
IF  (ed  >= 9  & ed <= 12) etapa=1.
VARIABLE LABELS  etapa 'etapa escolar'.
EXECUTE.
```

El resultado equivale a la sintaxis:

```
IF  (ed  >12  ) etapa=2.
VARIABLE LABELS  etapa 'etapa escolar'.
EXECUTE.
```

Las etiquetas de la variable se obtienen presionando la pestaña Tipo y etiqueta al definir el nombre de la variable. Repetimos el proceso, pero definiendo etiqueta = 2 para, cambiando la opción de selección por edad mayor que 12. Con lo que la sintaxis correspondiente a las dos acciones efectuadas por ventana sería:

```
IF  (ed  >= 9  & ed <= 12) etapa=1.
EXECUTE.
IF  (ed  >12) etapa=2.
VARIABLE LABELS  etapa 'etapa escolar'.
Value labels etapa 1 'primaria' 2 'secundaria'.
EXECUTE.
```

Sintaxis a la que hemos añadido manualmente en la ventana de sintaxis la expresión de las etiquetas o *labels* de los valores 1 y 2 de la variable etapa.

Comentario:

Si la variable ya existía, el nuevo valor sustituye al que constaba en la base de datos.

Si no, se efectúan todas las transformaciones dependientes de la variable condicionante. La nueva variable presentará valores missing, punto decimal, o solo cambiaría los definidos por las condiciones si esta ya existiese.

Es preferible siempre crear la transformación en una variable nueva con el fin de no perder la información original.

4.3.3 DO IF

Cuando las transformaciones que se deben realizar dependen de una variable con un número elevado de categorías, o bien la condición depende de la combinación de categorías de dos o más variables, es recomendable efectuar las transformaciones por sintaxis mediante la instrucción DO IF, cuya estructura general es:

```
DO IF [()expresión lógica[)].
Instrucción de transformación.
[ELSE IF[()expresión lógica [)]].
Instrucción de transformación.
[ELSE IF[()expresión lógica [)]].
[ELSE IF ...
[ELSE IF ...
[ELSE] Este último [ELSE] quiere decir “y en cualquier otro caso”
Instrucción de transformación.
END IF.
```

En el caso indicado anteriormente, la sintaxis sería, recordando las expresiones lógicas y de relación que se describen en la tabla 4.1:

```
DO IF (ed GE 9 and ed LE 12).
COMPUTE etapa =1.
ELSE.
COMPUTE etapa =2.
END IF.
```

La interpretación sería crea la etapa = 1 si la edad está comprendida entre 9 y 12 años y en cualquier otro caso etapa = 2.

Comentario:

Este ejemplo debe realizarse una vez comprobado que no existen otros valores de edad fuera del rango 9-17, ya que, si no, una edad de, por ejemplo, 7 años, la clasificaría como de etapa = 2.

Debe estar muy seguro en la utilización de la instrucción ELSE en cualquier otro caso, para no realizar asignaciones no deseadas.

Veamos otro ejemplo. Supongamos que deseamos definir el nivel de desnutrición de estos estudiantes. La desnutrición la definimos según los criterios de la WHO de 2007[3] los cuales se basan en clasificar como desnutrido a aquella persona cuyo IMC está por debajo del percentil 5, pero dependiendo de la edad y sexo de cada estudiante. Dichos valores se describen en la tabla 4.3.

Tabla 4.3 Valores del percentil 5 de IMC para hombres y mujeres en función de la edad

	Edades en años								
SEXO	9	10	11	12	13	14	15	16	17
Masculino	13,7	14,0	14,4	14,8	15,3	15,9	16,4	16,9	17,3
Femenino	13,4	13,8	14,2	14,8	15,3	15,8	16,2	16,5	16,6

Ejercicio 4.4

Utilizando la expresión DO IF defina si el estudiante está desnutrido según los criterios de la tabla 4.3.

Amplíe la sintaxis de forma que la nueva variable desnutrido tenga dos posibles categorías: 1 con etiqueta NO y 0 con etiqueta SÍ.

4.3.4 COUNT

Es muy frecuente que, en un cuestionario, se efectúe un conjunto de preguntas que buscan investigar un mismo concepto en las que la respuesta sea categórica. Es importante conocer a cuántas de las preguntas la respuesta ha sido la misma. Ese estudio —que si bien podría resolverse mediante un DO IF que tuviese en cuenta todas las variables— puede resolverse con facilidad con la instrucción COUNT, ya que cuenta el número de veces que aparece un valor o un rango de valores en un conjunto de variables. Además, COUNT puede también efectuarse por ventanas.

Así, por ejemplo, en el archivo **Estudio.sav**, se encuentra la información acerca de si el estudiante es usuario de comida rápida, si consume fruta, dulces o refrescos habitualmente. Como podrá observar en la ventana de vista de variables, nos interesa conocer si en su dieta habitual el individuo nunca

3 World Health Organization. Child growth standards. WHO Anthro Survey Analyser. Disponible en: https://www.who.int/tools/child-growth-standards/software

consume fruta = 5, si es usuario de comida rápida, com_rap = 1 Todos los días, ingiere dulces, dulces = 1 Todos los días, y refrescos =1, es decir todos los días.

Evidentemente el comportamiento simultáneo puede considerarse como un factor de riesgo para presentar obesidad o sobrepeso, por lo que interesa conocer cuántos de estos comportamientos inadecuados realiza diariamente, y crear una variable que podemos denominar riesgos. El resultado, como evalúa la respuesta de riesgo en cuatro variables, oscilará entre 0 y 4, es decir, desde que no realiza una conducta de riesgo, riesgos = 0 a que practica cotidianamente las cuatro conductas de riesgo, riesgos = 4. Para ello veamos cómo se calcula a través de ventana, efectúe ►**Transformar** ► **Contar valores dentro de los casos**... con lo cual se abrirá:

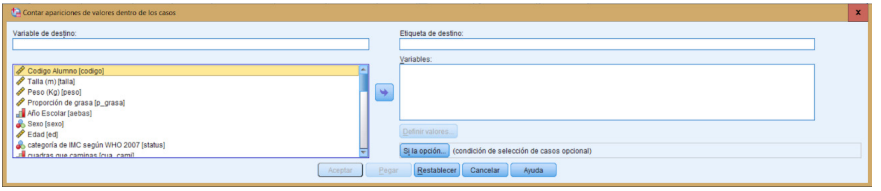


Figura 4.9 Ventana para Contar apariciones de valores dentro de los casos

En primer lugar, debe definir la variable destino, en este caso riesgos, indicando a la derecha la etiqueta de la variable creada o de destino. A continuación, irá seleccionando las variables que quiere analizar y, cada vez que seleccione una de ellas apretando la tecla de definir valores, podrá indicar cuál es el valor que se debe considerar en cada variable.

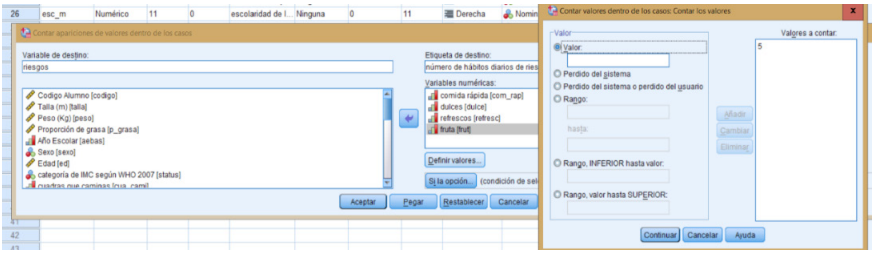


Figura 4.10 Asignación del valor 5 de interés para la variable frut

Cuando haya introducido las variables, la ventana resultante dará como sintaxis resultante.

```
COUNT riesgos=com_rap dulce refresc frut(5).
VARIABLE LABELS  riesgos 'número de hábitos diarios de riesgo'.
EXECUTE.
```

Puede ver que como la categoría de riesgo cambia en fruta y es la última que ha introducido, el valor (5) lo asigna a las anteriores variables, lo cual es incorrecto, por lo que manualmente en la ventana de sintaxis, se debe corregir a:

```
COUNT riesgos=com_rap dulce refresc(1) frut(5).
VARIABLE LABELS  riesgos 'número de hábitos diarios de riesgo'.
EXECUTE.
```

Es decir, por ventana, solo puede definir valores comunes a todas las variables. Si no lo hace así, siempre el último valor definido en la ventana será el utilizado. Para ver el resultado, efectúe el análisis de frecuencias de la variable creada y el resultado expresado con la siguiente tabla.

Tabla 4.4 Descripción de la variable riesgos

Número de comportamientos diarios de riesgo		Frecuen.	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	,00	3893	55,9	55,9	55,9
	1,00	2079	29,9	29,9	85,8
	2,00	801	11,5	11,5	97,3
	3,00	184	2,6	2,6	99,9
	4,00	7	,1	,1	100,0
	Total	6964	100,0	100,0	

Como puede observar en la tabla anterior, el 55,9 % de los estudiantes no tienen comportamientos de riesgo de forma cotidiana. Por otro lado, un 14,2 % de los estudiantes declaran dos o más hábitos de riesgo, (11,5 + 2,6 + 0,1) tal y como se han definido.

Ejercicio 4.5

Consideremos que también el hecho de comer solos, no realizar tres comidas al día y no desayunar sean así mismo conductas de riesgo.

Recalcule la variable riesgos añadiendo la información de estas tres nuevas variables.

Realice el ejercicio por sintaxis y describa la variable comparando con la tabla 4.3.

Como habrá podido comprobar, es mucho más seguro efectuar el conteo por sintaxis y que puede definir incluso variable a variable el valor que quiere incluir en el conteo.

4.3.5 Recodificar una variable

Los valores que presenta una variable, sea escalar o nominal, se pueden reagrupar dándoles una nueva codificación que responda mejor a los intereses del análisis. Esta redificación podría llevar a cabo, sin duda, a través de una serie de COMPUTE condicionados, pero en la barra de herramientas existe, en la pestaña de Transformar una forma directa de recodificación, ya sea en la misma variable o en otra creada exprofeso con el fin de mantener la información inicial.

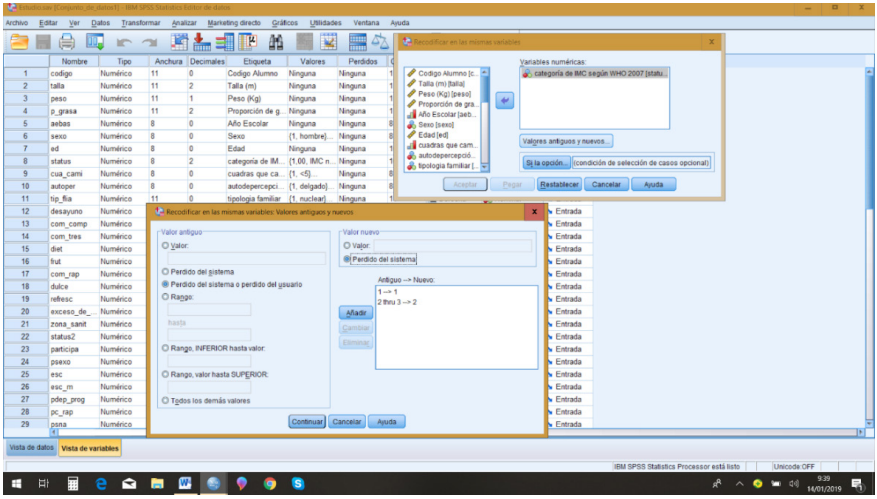


Figura 4.11 Ventanas y proceso de recodificación del status de los estudiantes en función de su definición como normal, sobrepeso u obeso en dos categorías, 1 y 2, con etiquetas 1 normal y 2 exceso de peso

La expresión de sintaxis una vez efectuada la acción de Pegar es:

```
RECODE status_valoración_IMC (1=1) (MISSING=SYSMIS) (2 thru 3=2).
EXECUTE.
```

En este caso se perdería la información original de las tres categorías, por lo que sería más adecuado recodificar en una nueva variable. Para ello efectúe los siguientes pasos: **Transformar ► Recodificar en distintas variables.**

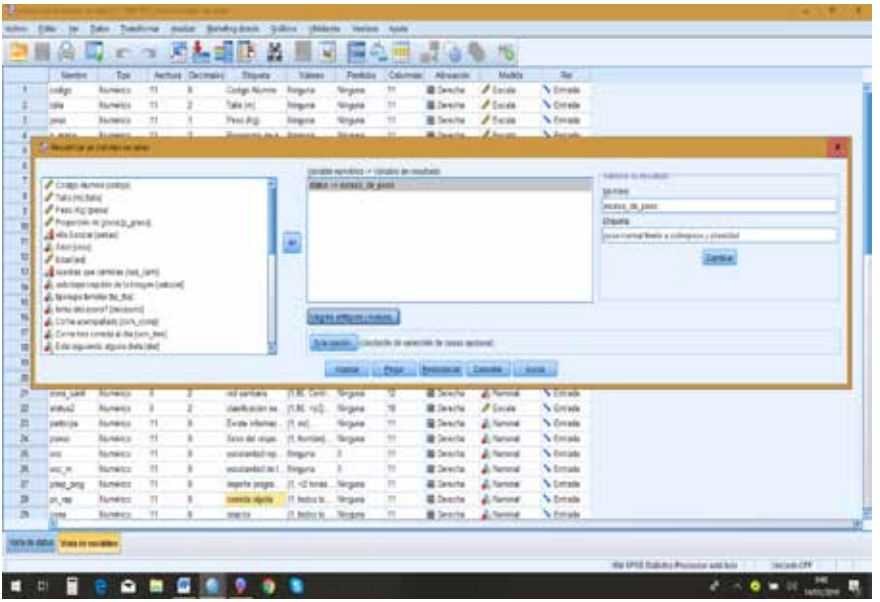


Figura 4.12 Ventanas y proceso de recodificación con el resultado en una variable nueva

La sintaxis asociada es:

```
RECODE status (1=0) (MISSING=SYSMIS) (2 thru 3=1) INTO exceso_de_peso.
EXECUTE.
```

Observará que el único cambio es que al final de la instrucción RECODE consta INTO nombre de la nueva variable. Como se aprecia en la ventana de transformación, también pueden asignarse valores si se cumple una condición externa (Si la opción).

En el caso en que desee crear una variable con la recodificación de valores de una variable escalar, observe que, en la pestaña Valores antiguos y nuevos, existe la posibilidad de codificar por rangos de valores de la escalar. Es pues una acción muy similar al DO IF que hemos visto anteriormente.

4.3.6 Recodificación automática

Esta opción es de gran utilidad en caso de querer recodificar una variable cadena con múltiples opciones.

Comentario:

Imagine que dispusiésemos de los diagnósticos de un conjunto de historiales médicos y que los mismos estuviesen expresados en una variable tipo cadena y quisiéramos recodificar en otra variable pero numérica. La opción recodificación automática asigna un valor numérico a los valores cadena de la variable de estudio manteniendo el nombre original como etiqueta del valor numérico asignado.

Evidentemente se podría llevar a cabo con un IF o un DO IF, pero es bastante tedioso debido a que las cadenas no siempre poseen campos bien definidos y con frecuencia existen caracteres en blanco que no pueden ignorarse pero que formalmente son valores diferentes si se trabaja en string o cadena.

La variable **psexo**, Sexo del responsable, ha sido introducida como una variable cadena que tiene una categoría igual a “hombre”, que tiene un espacio en blanco antes de la h, sería diferente de “hombre”, sin espacio en blanco o de “Hombre” con mayúscula. Por esa razón es preferible utilizar esta recodificación automática que da como resultado, por ejemplo, 1 “hombre”, 2 “Hombre” y 3 “ hombre”. Una vez realizada la autorrecodificación, esta nueva variable se puede recodificar agrupando todos aquellos valores equivalentes, como en este caso asignar los valores 1, 2, 3 a un solo valor 1 con etiqueta “hombre”.

Igual podría suceder con las opciones “mujer ”, “ mujer”, “Mujer” o “femenino”.

La variable sexo, definida como cadena atendría siete categorías y una vez autorrecodificado y asignados los valores 1 a 3 como 1 “hombre” y 4 a 7 como 2 “mujer” tendríamos la variable sexo en las dos categorías lógicas.

Ejercicio 4.6

Analice con la instrucción de frecuencias, **Analizar ► Estadísticos descriptivos ► Frecuencias** la variable **psexo**, sexo de los responsables del estudiante.

- ¿Cuántas categorías aparecen?
- ¿Cómo puede recodificar esta variable en solo dos categorías en la nueva variable **sexo_padres**?

Observe que los datos en blanco de psexo son considerados como opción en blanco pero no como valores perdidos. Autorrecodifique en una variable de nombre sexo_padres, aprovechando la autorrecodificación para definir los blancos como valores perdidos y anote los resultados para posteriormente recodificar en esta misma variable.

Capítulo cinco

Selección de casos, segmentación del archivo y agregación de casos

5.1 Introducción

En el análisis estadístico, las variables de estudio, especialmente las dependientes o variables respuesta al diseño, presentan una variabilidad cuyo conocimiento y explicación son el motivo fundamental de cualquier estudio de investigación.

Así, no solo hay que saber describir esa variabilidad si no también explorar la fuente de variación de la propiedad que se estudia. Una de las fases de cualquier estudio es analizar si esa variabilidad es la misma para todos los individuos estudiados, si no también si existen diferencias entre individuos de características distintas, las cuales están definidas en las variables del diseño o variables independientes. Es oportuno el poder efectuar ciertos análisis con un grupo de casos, para lo cual hay que seleccionarlos o bien generar el análisis por estratos definidos por una o varias de las variables dependientes y aprender a segmentar el archivo en función de las mismas.

Otra situación, no tan frecuente, es la consistente en seleccionar una muestra aleatoria de los datos ya sea como descripción orientadora cuando el archivo es de gran volumen de casos o bien como muestra de validación posterior a la obtención de modelos estadísticos obtenidos con el resto de los datos. A continuación, se describen los dos tipos de acciones correspondientes a las dos situaciones indicadas.

5.2 Selección de casos. Selección mediante una condición

La selección de casos con características comunes puede llevarse a cabo de forma temporal —manteniendo todo el archivo completo— o generando un nuevo archivo con solo aquellos casos que cumplan las condiciones de selección. Ambos procedimientos se obtienen a través de las acciones por ventana **Datos ► Seleccionar casos.**

Como se observa en la figura 5.1, podemos trabajar con todos los datos o con aquellos que satisfacen una condición y el resultado de la acción del criterio de selección sea descartar (temporalmente) los datos no seleccionados, copiarlos en un nuevo archivo o bien eliminarlos definitivamente.

El criterio de selección, **Si la opción**, abre una subventana en la que, de forma similar, al transformar una variable de forma condicionada, ofrece la posibilidad de escoger la variable o variables que generarán el criterio y los valores de las mismas.

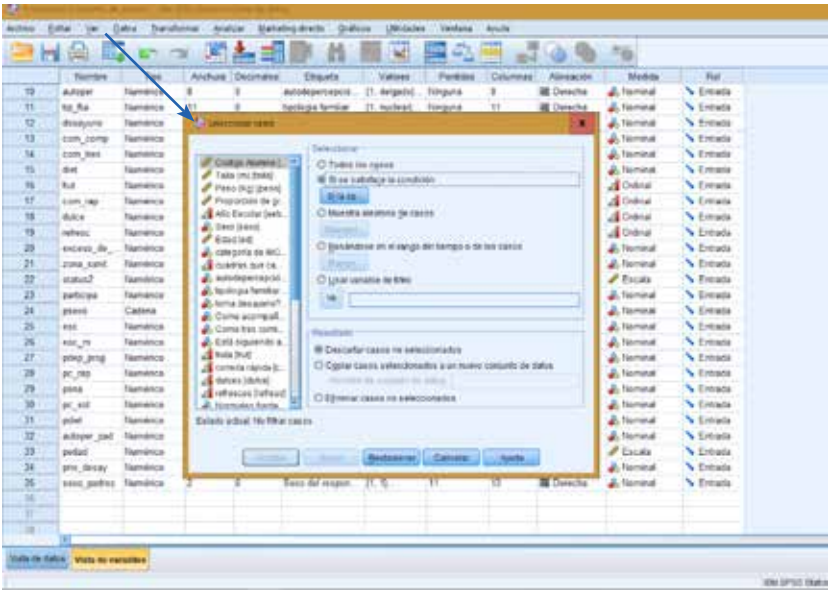


Figura 5.1 Ventana resultante para la selección de casos

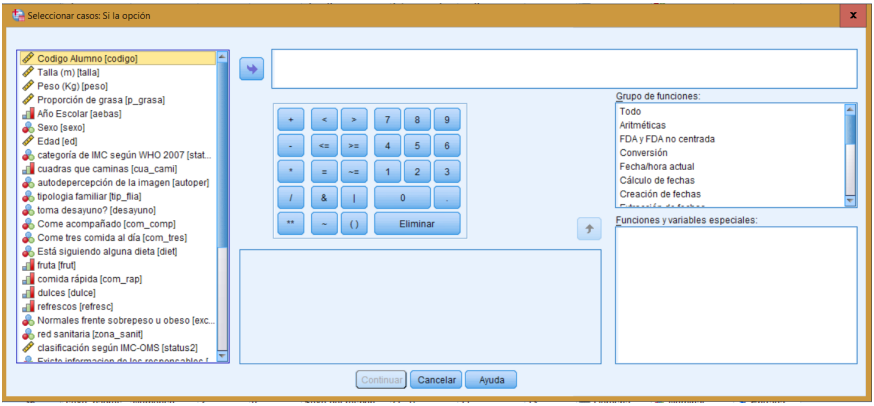


Figura 5.2 Subventana de la acción **Datos ►Seleccionar casos ►Si la op...**

El criterio de selección puede ser un valor o valores de una o varias variables o bien una combinación lógica o matemática de las mismas. La sintaxis asociada es:

```
SELECT IF [()condición lógica[])].
```

La cual debe estar precedida por la instrucción TEMPORARY si la selección se desea que sea temporal y se deshaga al finalizar la acción que se pretende realizar con los casos seleccionados, si bien se neutraliza inmediatamente después de la primera acción.

En el caso en que no se seleccione TEMPORARY en sintaxis o Descartar casos no seleccionados, el procedimiento tiene el inconveniente de que, si la condición se expresa equivocadamente, no es posible rectificar sin volver a abrir el archivo que se tenía activo.

Ejercicio 5.1

Partiendo del archivo **estudio.sav**, seleccione temporalmente por sintaxis, aquellos estudiantes que pertenecen a la **zona sanitaria = 1**.

Observe el efecto de esta acción en la ventana vista de datos.

Repita la acción, pero, en este caso, creae un archivo de datos que contenga exclusivamente a estos estudiantes. Indique la sintaxis generada por la acción **Pegar**.

5.3 Selección por muestreo aleatorio

En el otro supuesto descrito, es decir, cuando el subgrupo se quiera generar con la condición de ser una muestra aleatoria del conjunto de datos disponible, la instrucción es la siguiente:

SAMPLE {método de selección}

En la que el método de selección especifica el tamaño y criterio de selección de la muestra aleatoria utilizado:

- **Proporción**: permite especificar la proporción de casos que deseamos seleccionar del fichero activo. Por ejemplo, si queremos configurar una muestra que represente el 20% del total de casos, la instrucción sería: SAMPLE 0.2.

- **n FROM m**: indica que se seleccionan **n** de los primeros **m** casos del fichero activo. Por ejemplo, si nos interesase seleccionar 35 de los 150 primeros casos del fichero la instrucción sería: SAMPLE 35 FROM 150.

Debe conocer que el programa efectúa la selección a través de una rutina pseudoaleatoria que se inicia con una semilla de aleatorización concreta —en el programa se utiliza el valor por defecto 2 000 000—, la cual se repite cada vez que se ejecuta una nueva sesión del programa. Es decir que si, al volver a abrir el programa selecciona una muestra de las mismas características, el resultado siempre será el mismo. Esta semilla se puede alterar mediante la instrucción, SET SEED=N, en la que N debe ser un entero inferior a 2 000 000 000.

En la práctica, el conocimiento del valor de la semilla solo presenta utilidad si nos interesa repetir exactamente una misma selección de casos, o por el contrario, si se desean muestras diferentes, en cada selección, debe modificar la semilla, al iniciar la sesión. La acción por ventanas se muestra en la figura 5.3 una vez ejecutado **Datos ► Selecciona Casos** en la barra de herramientas de la ventana **Vista de datos** o en la **Vista de variables**.

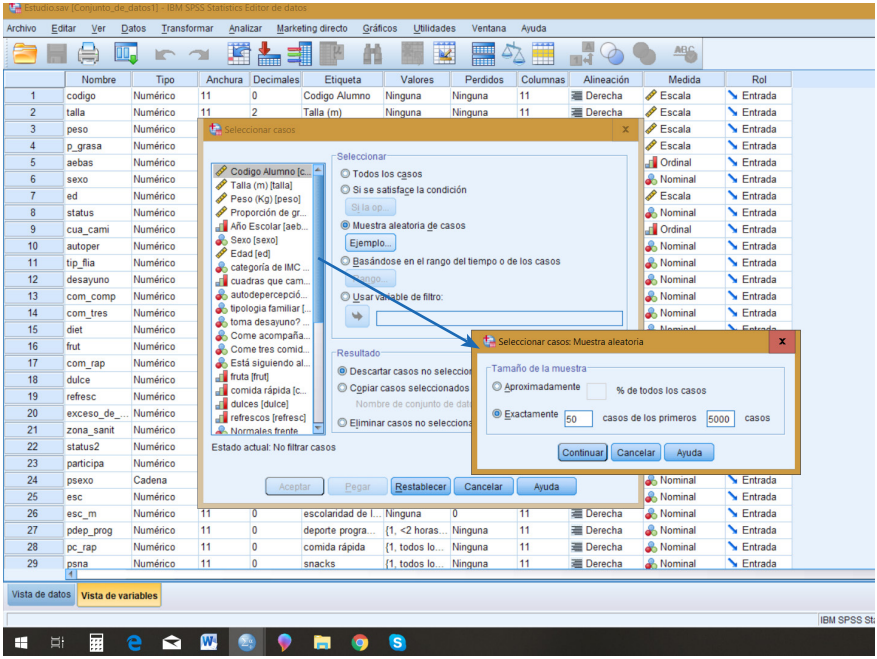


Figura 5.3 Subventana de selección de casos correspondiente a activar ► **Muestra aleatoria de casos**

El resto de las instrucciones es el mismo, que la selección de casos sea permanente o no y si se quiere crear un fichero con los datos.

Comentario:

Cuando selecciona casos de forma temporal se crea una nueva variable que por defecto el programa denomina filtro. Esta variable puede ser utilizada en otra ocasión si pretende seleccionar de nuevo la misma muestra activando la opción **Usar variable de filtro** indicando en la ventana adjunta a la opción el nombre del filtro que se pretende utilizar.

Si desea crear dos filtros o más, no puede encadenarlos ya que solo se ejecutaría el último. Es preferible en ese caso definir un filtro compuesto en la operación lógica que debe definir en la ventana que se muestra en la figura 5.2.

Ejercicio 5.2

Ejecute la siguiente sintaxis:

```
SELECT IF (sexo = 1 & ed = 9).
EXECUTE.
DATASET ACTIVATE  Conjunto_de_datos1.
```

Con esta acción crea un archivo nuevo de datos que solo contiene a los niños varones de 9 años que participan en el estudio. Verá que, en la barra inferior del SPSS, le indicará que existe un archivo nuevo que se llama Sin título (niños de 9 años).

Describa en este nuevo archivo las acciones **Transformar ▶ Analizar ▶ Estadísticos descriptivos ▶ Frecuencias** para la variable **sexo**.

El resultado es que hay 66 niños de esta edad.

A continuación, vaya al archivo **Estudio** y seleccione los casos que cumplen esa misma condición pero sin guardar los seleccionados en un archivo nuevo.NObservará en la matriz de datos que no han desaparecido casos, pero que se ha creado una variable que se denomina **filter_$** con valores cero y verá que el número de registro está marcado con una línea inclinada o uno en función de si el caso cumple la condición de selección. Si efectúa el análisis de frecuencias de esta variable, verá que también el número seleccionado es 66.

Vuelva **Datos ▶ Seleccionar casos** y utilice la opción **usar variable de filtro** y efectúe la descripción de frecuencias. Observará que el filtro se mantiene hasta que no lo anule, aunque haya usado la instrucción USE ALL que equivale a **Selección de datos** opción **Todos los casos.**

El esquema de funcionamiento es el siguiente:

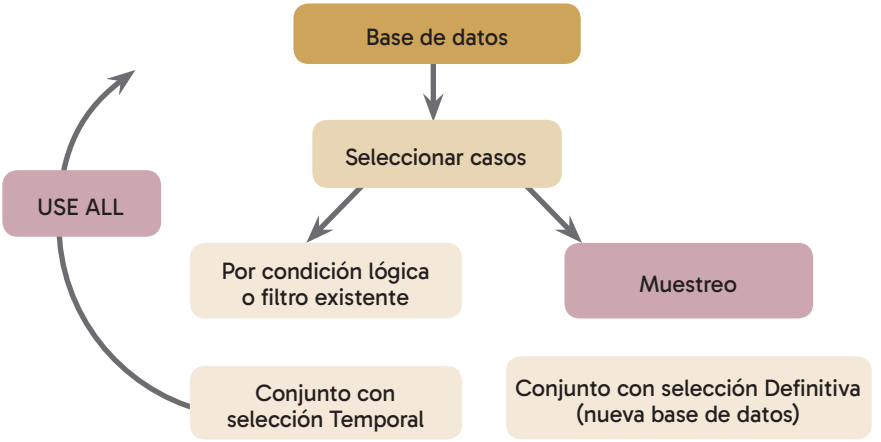


Figura 5.1 Selección por muestreo aleatorio

5.4 Segmentación de un archivo

Imagine que desea efectuar un análisis separando diversos estratos. Estratos llamamos a un subconjunto de datos que tienen en común alguna propiedad que, a su vez, puede ser simple o compleja, resultante de la intersección de diferentes propiedades.

En el ejemplo que nos ocupa, imagine que nos interesa efectuar el análisis por sexo y por área sanitaria. Cada estrato estará definido por el valor de la variable sexo y por el valor del área sanitaria. Es decir, tendremos ocho estratos, ya que existen cuatro áreas sanitarias, Centro, Norte, Sur y Periférica.

Evidentemente, un análisis puntual se podría efectuar mediante la selección de casos de cada estrato, intersección de cada categoría de sexo con categoría de Centro y efectuar el análisis. No obstante, sobre todo si el número de estratos es elevado, el número de acciones de selección de casos es también elevado.

Existe la opción de segmentar el archivo de forma que, con una única opción, el archivo se estructura según la definición de los estratos y se puede efectuar una única vez la sintaxis del análisis. Los resultados estarán ordenados según el orden de los estratos.

Para lograr esta segmentación, el archivo debe de estar ordenado en función de las variables que generan dichos estratos, en este caso por Sexo y por Centro, si el interés está en visualizar las diferencias de centro dentro de cada sexo, o bien por centro y analizar las diferencias de resultados según el sexo del estudiante en cada centro. La opción para ordenar los casos del archivo es **Datos ▸ Ordenar Casos**.

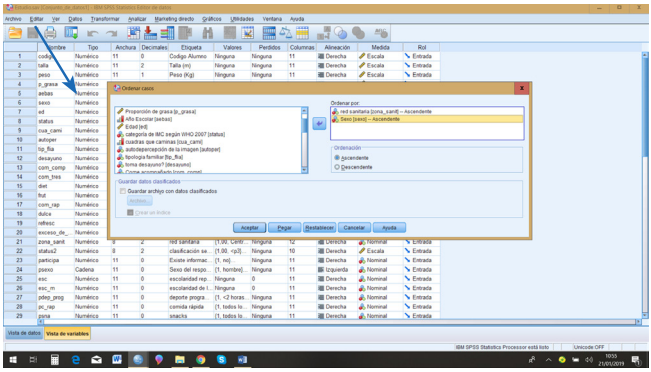


Figura 5.4 Ventana que muestra la acción de ordenar los casos del archivo por Zona Sanitaria y Sexo

La expresión de sintaxis que realizaría esta acción es:

```
SORT CASES BY zona_sanit(A) sexo(A).
```

Cuyo resultado no es el mismo si la instrucción hubiese sido:

```
SORT CASES BY sexo(A) zona_sanit(A).
```

Comentario:

El orden de las variables por las cuales se ordena genera la jerarquía de anidamiento de los estratos, por lo que debe cuidar su definición.

Una vez ordenados, puede llevar a cabo la segmentación, si bien la instrucción que vemos a continuación ordenará los casos en el caso de que usted no hubiese ordenado. La segmentación de un archivo en estratos se logra mediante **Datos ►Segmentar Archivo**.

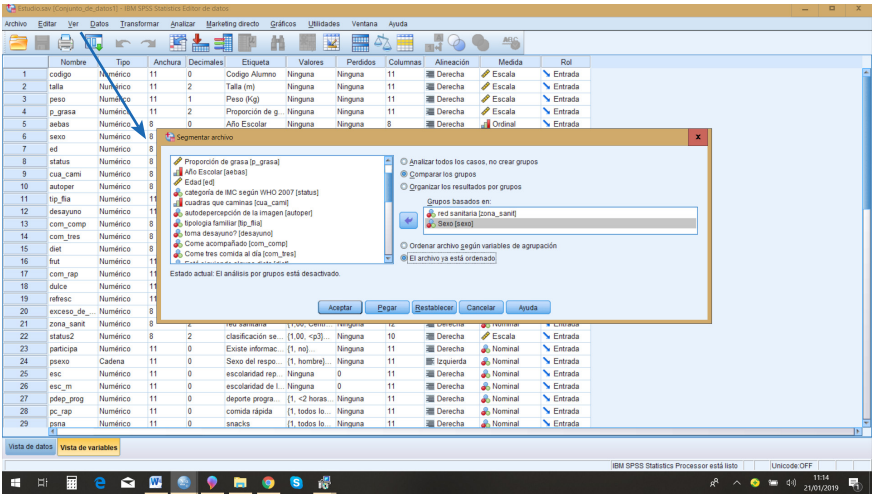


Figura 5.5 Ventana que muestra la petición de segmentar el archivo de trabajo Estudio.sav

Como puede observar en la figura 5.5, la segmentación del archivo se efectuará de forma jerárquica por Zona Sanitaria y dentro de cada zona por la variable Sexo. Asimismo, se indica que el archivo ya ha sido ordenado con

ese criterio y que los resultados del análisis los exprese de forma comparativa y no para cada estrato o grupo por separado. La instrucción de sintaxis que se asocia a esta orden es:

```
SPLIT FILE LAYERED BY zona_sanit sexo.
```

Una vez ejecutada la acción, todos los análisis que se soliciten se efectuarán con este criterio organizativo, hasta que anule la segmentación con la instrucción **Analizar todos los casos, no crear grupos** de la misma ventana que se muestra en la figura anterior. La instrucción por sintaxis sería:

```
SPLIT FILE OFF.
```

Por ejemplo, si se desease conocer la frecuencia de la variable desayuna, cuyas opciones son no desayuna, sí desayuna, se segmentaría el archivo por zona sanitaria y a continuación la instrucción **Analizar ► Estadísticos descriptivos ► Frecuencias** con la variable Desayuno

El resultado como puede ver en la ventana de Resultados tal y como se muestra en la tabla 5.1. La instrucción de sintaxis una vez deshecha la segmentación es:

```
SPLIT FILE LAYERED BY zona_sanit sexo.
FREQUENCIES VARIABLES=desayuno
  /ORDER=ANALYSIS.
SPLIT FILE OFF.
```

Mientras no se indica la instrucción Split File Off los análisis seguirían la estructura estratificada.

Ejercicio 5.3

Vuelva a efectuar el análisis, pero cambiando el orden de las variables de agrupación, es decir, visualizando las frecuencias por zona sanitaria para cada sexo. En este caso, la jerarquía de agrupación es Sexo, Zona sanitaria.

Pegue las instrucciones de sintaxis y observe las diferencias al definir los estratos por sexo y zona sanitaria con la que se muestra en la tabla 5.1.

Tabla 5.1 ¿Toma desayuno? Resultados según los estratos zona sanitaria y sexo

Red sanitaria	Sexo			Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Centro	Hombre	Válido	no desayuna	1	16,7	16,7	16,7
			sí desayuna	1	100,0	100,0	100,0
			Total	6	100,0	100,0	
		Perdidos	Sistema	1	100,0		
	Mujer	Válido	no desayuna	1	50,0	50,0	50,0
			sí desayuna	1	100,0	100,0	100,0
			Total	2	100,0	100,0	
		Perdidos	Sistema	1	10,0		
		Total		10	100,0		
Sur	Hombre	Válido	no desayuna	1	50,0	50,0	50,0
			sí desayuna	1	100,0	100,0	100,0
			Total	2	100,0	100,0	
		Perdidos	Sistema	1	20,0		
		Total		5	100,0		
	Mujer	Válido	no desayuna	1	33,3	33,3	33,3
			sí desayuna	1	100,0	100,0	100,0
			Total	3	100,0	100,0	
		Perdidos	Sistema	1	25,0		
		Total		4	100,0		
Periférica	.	Perdidos	Sistema	3	100,0		
	Hombre	Válido	no desayuna	1	50,0	50,0	50,0
			sí desayuna	1	100,0	100,0	100,0
			Total	2	100,0	100,0	
		Perdidos	Sistema	1	100,0		
		Total		7	100,0		
	Mujer	Válido	no desayuna	6	19,4	19,4	19,4
			sí desayuna	25	80,6	80,6	100,0
			Total	31	100,0	100,0	

		Perdidos	Sistema	1	50,0		
		Total		2	100,0		
Norte	Hombre	Válido	no desayuna	1	33,3	33,3	33,3
			sí desayuna	5	100,0	100,0	100,0
			Total	3	100,0	100,0	
		Perdidos	Sistema	1	16,7		
		Total		6	100,0		
	Mujer	Válido	no desayuna	1	100,0	100,0	100,0
			sí desayuna	2	100,0	100,0	100,0
			Total	2	100,0	100,0	
		Perdidos	Sistema	1	33,3		
		Total		3	100,0		

5.5 Agregación de datos

Como puede observar en la tabla 5.1, la variable desayuno está descrita por dos variables, pero esta tabla no puede manipularse para análisis posteriores ya que no es un nuevo archivo de datos. En diversas ocasiones, es muy útil disponer de un archivo con la información agregada en esos estratos y en cada estrato el resumen de diversas variables.

Por ejemplo, imagine que quisiésemos disponer de la siguiente información en cada estrato: edad media de los estudiantes que componen el estrato, que proporción están llevando a cabo una dieta y qué proporción tiene sobrepeso u obesidad. Además de saber cuántos estudiantes hay en cada estrato.

El resultado deseamos sea en formato .sav, es decir en forma de archivo de datos y no de tabla. En primer lugar, el archivo debe de estar ordenado por la jerarquía del estrato. En segundo lugar, debemos tener bien claras las propiedades que deseamos agregar, es decir, qué variables y qué características de las mismas. La acción se efectúa mediante **Datos ► Agregar datos**.

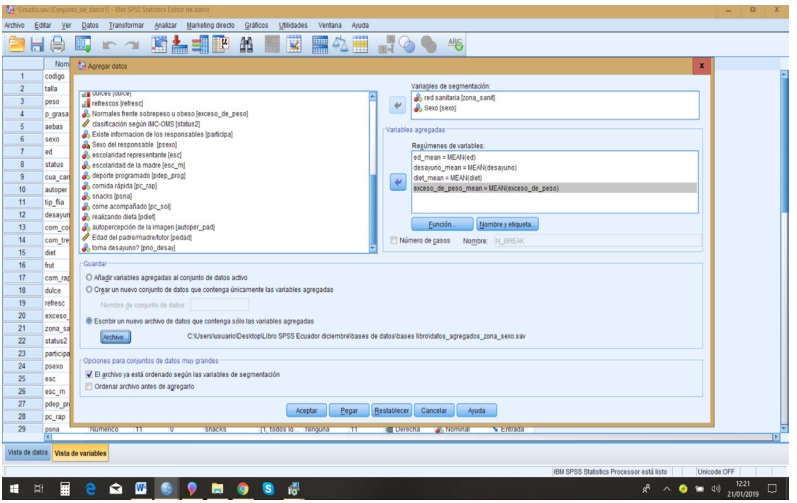


Figura 5.6 Ventana de agregación de datos por zona sanitaria y sexo de las variables edad, desayuno, dieta y normales frente a exceso de peso

Como puede observarse en la figura, se indican las variables que definen la jerarquía de grupos o estratos, en función de Zona Sanitaria y Sexo. En la ventana de la derecha se muestran las variables cuya información se pretende resumir.

Por defecto la información, como las variables son de tipo numérico, indica que agregará el valor promedio (MEAN), por lo que las nuevas variables en el archivo agregado se denominan **nombre de la variable_mean** para notar que el valor que se agrega es la media de la variable cuyo nombre se indica. Además, existe la posibilidad de cambiar la etiqueta de la nueva variable con la opción **Nombre y etiqueta.** Esta alternativa es recomendable con el fin de aclarar la naturaleza de la nueva variable.

Por otro lado, se le señala que el archivo está ya ordenado en sentido jerárquico que se exprea en las variables de agrupación, y que el resultado se guardará en un archivo llamado **datos_agregados_zona_sexo**.

Observará que además se puede marcar que recoja el número de casos que hay en cada estrato, dato que, por defecto, denomina N_BREAK y que puede cambiar de nombre. Si se fija, existe debajo de la segunda ventana una tecla que indica **FUNCIÓN**. Marcando esta tecla después de remarcar cada variable a agregar puede cambiar la información que requiere de esa variable.

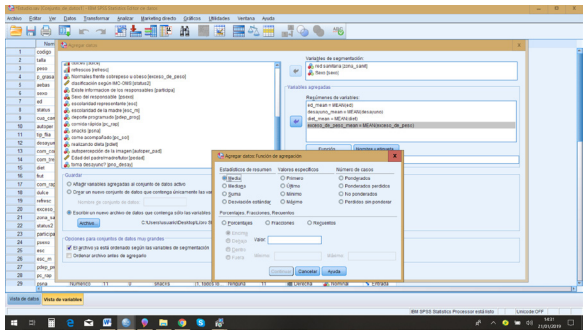


Figura 5.7 Ventana de las diferentes funciones que pueden utilizarse para agregar información

Como se explicita en la figura 5.7, si la variable es escalar, puede agregarse información muy diversa como la media, desviación estándar, mediana, suma de datos, el valor máximo, mínimo, así como indicar que solo guarde el primer valor que encuentre de la variable, o el último.

Si la variable es de tipo nominal, además puede agregar tanto el porcentaje de casos, o el número de casos (Recuentos) que se encuentran en un rango de valores numéricos de dicha variable. Si de una misma variable quiere agregar más de una característica puede volver a introducir la misma variable y seleccionar otra función.

Así, en el ejemplo enunciado, se guardaría, por ejemplo, de la edad, la media, el porcentaje de casos que no desayunan, el de los que hacen dieta y el número de alumnos que tienen exceso de peso, es decir, que la variable exceso de peso es igual a 1. En la siguiente figura se describe la ventana final. OBJ

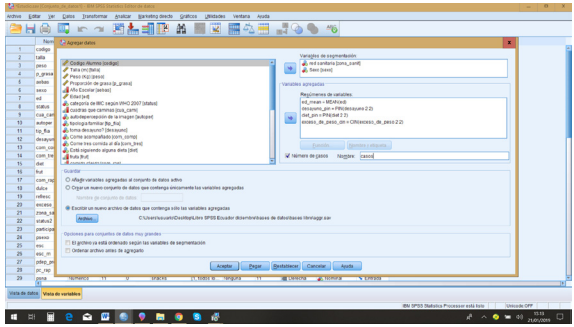


Figura 5.8 Visualización de la pantalla definitiva de la agregación

La sintaxis correspondiente es

```
AGGREGATE
  /OUTFILE='C:\Users\usuario\Desktop\Libro SPSS Ecuador diciembre\bases
de datos\bases '+
    'libro\aggr.sav'
  /BREAK=zona_sanit sexo
  /ed_mean=MEAN(ed)
  /desayuno_pin=PIN(desayuno 1 1)
  /diet_pin=PIN(diet 2 2)
  /exceso_de_peso_cin=CIN(exceso_de_peso 2 2)
  /N_BREAK=N.
```

En esta sintaxis puede verse las acciones correspondientes a:

- Crear una base de datos con el resultado, en un archivo denominado **aggr.sav'**.
- La creación de los estratos que describirán la agregación de la información requerida en cada zona sanitaria y en cada una de ellas por sexo.
- La información deseada es la media de la variable edad, el porcentaje de estudiantes que no toman desayuno, es decir, entre 1 y 1, que es la forma de decir igual a 1, el porcentaje de los estudiantes que hacen dieta o porcentaje entre 2 y 2 y casos con exceso de peso, entre 2 y 2.
- El número de estudiantes en cada estrato o casos.

Ejercicio 5.4

Ejecute la sintaxis arriba indicada. (*Copy* y *paste* en una ventana de sintaxis).

Abra el fichero **aggr.sav**, generado al activar la instrucción de agregar, y mediante la acción **Transformar ▶ Calcular variable** determine la prevalencia, es decir, el porcentaje de alumnos con exceso de peso en cada estrato.

¿En qué estrato es mayor la prevalencia de exceso de peso y de alumnos que van a la escuela sin desayunar?

¿Cuántos alumnos del estudio están realizando dieta?

(**Frecuencias ▶ Estadísticos ▶ suma**).

Capítulo seis

Control de calidad. Análisis descriptivo

6.1 Introducción

Al llegar a este capítulo, el lector debe haber adquirido la suficiencia en el manejo de las acciones básicas del programa. Por lo que es recomendable que haya realizado todos los ejercicios y comprobar en el último capítulo que su resolución ha sido satisfactoria.

Con el fin de unificar los resultados obtenidos por los alumnos, se recomienda utilizar para los próximos capítulos el archivo "Estudio con definiciones agregadas.sav" ubicado en el repositorio.

A partir de este momento, los capítulos restantes se dedicarán —más que a aumentar el conocimiento de técnicas— a su aplicación, al menos en las acciones más básicas. Esta obra no pretende ser un libro de introducción a las técnicas de análisis mediante la Bioestadística, por lo que se supone que el lector las conoce, si bien los dos capítulos que siguen pueden servirle para reforzar los conceptos de la estadística descriptiva básica como de la bivariada y las pruebas de contraste más comunes. O bien, busca ser la motivación para que el lector avance en sus estudios de Bioestadística.

Bajo estas premisas, en este capítulo abordaremos, en primer lugar, un aspecto fundamental cual es el control de calidad de los datos que se encuentran en la base de datos. A continuación, se describirán con cierto detalle las acciones necesarias para llevarla a cabo.

6.2 Control de calidad de la base de datos

Es común que la base de datos haya sido creada por múltiples personas, desde la definición de las variables y sus categorías si es el caso, la creación de nuevas variables, su recodificación, la definición de valores perdidos, fuera de rango, o, por errores de tecleo o de adquisición de datos, constan valores erróneos.

Un analista estadístico, o de una base de datos, nunca debe suponer que en todos esos procesos no se han cometido errores o que se ha revisado la información antes de introducirla. Por todas estas razones, debemos estudiar si todos los datos disponibles son, al menos aparentemente, correctos. Debemos evaluar cuál es la calidad de la información recogida en el estudio, mediante alguna técnica descriptiva que permita el control de la misma.

La ausencia de esta acción puede producir, a lo largo del análisis, múltiples problemas que no pueden resolverse más que corrigiendo esos posibles errores y deficiencias, lo cual es mucho más costoso que efectuarlo antes de iniciar cualquier tipo de análisis. Las diferentes etapas que componen el control de calidad de los datos de la base definitiva son:

6.2.1 Control de valores perdidos o missing

Al visualizar la matriz de datos, podrá observar que, en diversas variables —según su tipo *string* o numérica— presentan falta de información en algunos casos, ya sea como espacio en blanco si la variable es cadena o con un punto decimal o una coma si es numérica.

La primera acción de control consiste en evaluar el porcentaje de información faltante o *missing* para cada variable, definiendo previamente los espacios en blanco como valores perdidos en la vista de variables.

No existe una norma de consenso para evaluar qué porcentaje de valores *missing* ponen en dificultad el análisis de los datos, si bien generalmente puede decirse que se acepta hasta un 5 % de los datos de cada variable.

Hay que considerar que los *missing*, al combinar variables, pueden llevar a perder una gran parte de la información y por esa razón hay que cuantificar su presencia. Una manera rápida de saber los valores perdidos de cada variable consiste en ejecutar la sintaxis

```
Frequencies var= XXX.
```

Esta sintaxis, evidentemente, da como resultado tantas tablas de frecuencia como variables haya especificado en la instrucción, sin embargo, una vez ejecutada, fíjese solo en la primera información, que es precisamente el número de casos en cada variable sin información.

Ejercicio 6.1

Siguiendo con el archivo **Estudio.sav**, explore, mediante la acción **Analizar ► Estadísticos descriptivos ► Frecuencias,** qué variables del archivo poseen más del 5 de valores perdidos.

Compruebe previamente que todas las variables tienen definidos correctamente los valores considerados *missing*.

6.2.2 Missing *de usuario*

La descripción de las variables categóricas que se efectúa mediante la instrucción FREQUENCIES permite, además, detectar cuatro tipos de errores que se corresponden a:

6.2.3 Valores imposibles

Valores que no tienen sentido para la variable estudiada. Por ejemplo, una edad del jefe de familia igual a 953 años, valores negativos en el peso o talla del estudiante, etc.

6.2.4 Valores fuera de rango

Los valores fuera de rango son aquellos que, siendo posibles en general, no lo son en nuestra investigación, debido a los criterios de exclusión definidos en el protocolo de investigación. En concreto, si la variable es escalar, la detección de errores puede realizarse mediante la sentencia:

```
DESCRIPTIVES nombre de las variables escalares
  /STATISTICS=MIN MAX.
```

Expresión que se logra al pegar la información de la ventana **Analizar ► Estadísticos descriptivos ► Descriptivos** y desplazar a la subventana las variables que desee analizar, indicando en las **Opciones** que indique los valores máximos y mínimos.

En este estudio, por el rango de edades de los estudiantes de las escuelas incluidas en el estudio, se decidió estudiar los estudiantes entre 9 y 17 años, por lo que cualquier edad fuera de ese rango, si bien es posible, debe de ser definida como valor no válido o *missing* de usuario o de sistema.

Por ejemplo, en nuestra base debemos estar seguros de que no hay valores inferiores a 9 años ni superiores a 17 en la edad de los niños (ya que el rango de valores válidos está, por definición de objetivo del estudio, entre 9 y 17 años). Su definición como *missing* sigue la instrucción:

```
MISSING VALUES Variable (valores fuera del rango).
```

Aunque también puede efectuarse recodificando la variable:

```
RECODE Variable (valor,valor2,...valorn = SYSMIS).
```

6.2.5 Incumplimiento de ceros estructurales

Los ceros estructurales son aquellas situaciones generadas al relacionar dos variables categóricas, en las cuales, forzosamente, hay casillas o situaciones en las que no puede observarse ningún caso por definición. Por ejemplo, si tenemos la variable sexo y la variable uso de anticonceptivos orales, el cruce de ambas genera situaciones (celdas de una tabla bidimensional) como Hombre/Sí toma anticonceptivos orales en la que la frecuencia observada de casos debe ser siempre cero. Pues bien, debemos examinar que, efectivamente, en situaciones de este tipo, no encontramos frecuencias mayores que cero. En un siguiente apartado, se describirá la técnica adecuada para detectarlos.

6.2.6 Variables alfanuméricas

Las variables creadas en este formato merecen ser examinadas con mucha atención. Con frecuencia, se abusa de este tipo de variables para evitar aviso de error en la entrada de datos, lo cual es una ventaja de ahorro de tiempo para los digitadores en la creación de la base de datos. Este comportamiento trae diferentes problemas para el analista, debido a que, en este formato, una falta de información generalmente queda reflejada como un espacio en blanco, lo cual no es detectado como valor perdido al ser los caracteres en blanco un valor determinado en alfanumérico, debiendo efectuarse un análisis muy exhaustivo de este tipo de variables.

Una forma de detectar los valores perdidos consiste en la acción **Transformar ► Recodificación automática ►crear una nueva variable** que asigna un valor numérico a cada posible valor alfanumérico y detectar qué valores están en blanco o con caracteres no deseados en los valores.

Una vez autorrecodificada la variable, la variable creada, de naturaleza numérica, se puede recodificar y asignar el número que corresponde a respuesta en blanco a un valor *missing*, así como recodificar aquellos valores equivalentes pero que difieren en el texto alfanumérico. Cuide el hecho de que al recodificar los valores de la variable se correspondan con los recodificados.

6.2.7 Recuperación de la información perdida

En todo este proceso, el lector habrá detectado qué valores son desconocidos o que valores son erróneos. Las acciones que se deben seguir consisten, en primer lugar, en asignar a cada error el número identificativo del caso y efectuar una consulta, mediante documento escrito, a la persona responsable de la custodia de los cuestionarios para que compruebe el verdadero valor o que confirme que la información no consta o que la que escrita es verdadera.

Esta acción jamás se puede sustituir por asignar un valor que puede o no parecer lógico al analista, y solo puede efectuar un cambio una vez que reciba la respuesta por parte del custodio o del Investigador Principal. El resultado puede ser que los errores sean debidos al tecleo, esto es, introducir un número en vez de otro en la entrada informatizada, o, en otros casos, un carácter alfanumérico por otro o bien la adición de un carácter involuntariamente. En estas situaciones, las correcciones se efectúan sustituyendo el valor erróneo por el valor correcto.

En otras situaciones en las que se confirme que el error ya proviene de la propia recogida manual de datos, solo queda la medición de nuevo de la variable, aunque esta opción es muchas veces inviable por la imposibilidad de repetir la medición o porque el número de casos es tan elevado que representaría casi la repetición del estudio. Lo cual puede ser imposible por el aumento de costes y tiempo o porque la muestra no está ya disponible.

En una segunda situación, en la que se considerase que subsanar del error no es relevante, o no es viable, simplemente se convierte el valor erróneo en valor *missing* de sistema.

Comentario:

En todo caso merece la pena mencionar que en esta fase debe producirse la interacción entre analista e investigador. El primero no puede modificar valores que son, o cree, erróneos sin consultarlo con el segundo. Para el segundo, este proceso debe enseñarle la importancia del diseño previo de una hoja de recogida de la información, así como el de una base de datos con las condiciones necesarias para reducir, al mínimo, la posibilidad de errores.

En definitiva, es aconsejable que la colaboración entre analista e investigador no se inicie en este punto, si no en el momento en que se diseña el estudio.

Una vez localizado el valor erróneo, la localización de los códigos en los que se encuentra el error se realiza mediante la sintaxis

```
COMPUTE filtro=(nombre de la variable= valor erróneo).
FILTER BY filtro .
EXECUTE .
FREQ codigo .
```

Teniendo en cuenta que, si la variable es alfanumérica, el valor erróneo debe estar entre comillas.

Ejercicio 6.2

Efectúe el control de calidad de las variables talla en metros, peso en kg, edad y sexo de los padres.

Cuando sea necesario, redacte una petición de rectificación de errores indicando el caso o código a que hace referencia cada error y realice la corrección.

En el caso de la variable **psexo**, como es alfanumérica, observará que hay una gran cantidad de valores en blanco, es decir sin información, debido a que los padres negaron dar la información o no consta en el cuestionario. Pase esta información a perdidos por el sistema. En cuanto al resto de valores observará que los diferentes valores de sexo se deben a escribir con mayúscula o minúscula, a dejar un espacio en blanco antes de especificar, ya que hay valores numéricos pero que por la naturaleza de la variable son caracteres alfanuméricos.

Por ejemplo, se puede dar un caso en que cambia la escritura, si se hace con mayúscula o minúscula o si se coloca mujer o femenino o si se realiza con espacios al inicio o al final, el programa lo toma como si fuese una categoría distinta de la variable como se aprecia en la siguiente tabla de resultados:

Tabla 6.1 Sexo del responsable, dada la naturaleza inicial de la variable como cadena alfanumérica

Sexo del responsable					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido		2975	42,7	42,7	42,7
	HOMBRE	1	,0	,0	42,7
	femenino	1	,0	,0	42,7
	mujer	1	,0	,0	42,8
	hombre	1189	17,1	17,1	59,8
	mujer	2790	40,1	40,1	99,9
	FEMENIO	1	,0	,0	99,9
	HOMBRE	3	,0	,0	100,0
	MUJER	3	,0	,0	100,0
	Total	6964	100,0	100,0	

Esta situación es muy frecuente si el digitador no tiene las precauciones pertinentes o porque hay más de un digitador.

```
COMPUTE  psexo=LTRIM (UPCAS (psexo) ) .
EXECUTE .
```

Esta instrucción que corresponde a una función de transformación presente en esa ventana y que elimina los caracteres en blanco de la variable alfanumérica que existen a la izquierda del valor (LTRIM), una vez pasados todos los caracteres a mayúsculas (UPCAS) con tal de que todas empiecen en la misma columna.

Posteriormente se recodifica la variable creada para lo cual lo más seguro, dado lo complejo y fácil de cometer errores, es, en primer lugar, autorrecodificar, **Transformar ► Recodificación automática** de la variable

psexo, en una nueva variable que podría ser **sexo_responsables**, con lo cual se obtiene la equivalencia numérica, pasando los valores de la variable alfanumérica como etiquetas de estos nuevos valores. Esta información se encuentra en la ventana de resultados. Una vez convertida en numérica, la recodificación es simple.

6.3 Descripción univariada de variables

A lo largo de este libro se han utilizado recursos del programa para poder conocer el contenido de las variables en el archivo de datos, como acciones que se pueden realizar, pero sin entrar en detalles de dichas acciones. En este apartado se profundiza en las instrucciones para llevar a cabo una descripción más completa y analizar todas las opciones que las integran.

En primer lugar, en lo referente a la descripción de cada variable, estadística univariada, las acciones son más o menos adecuadas en función de si la variable es categórica, nominal o escalar discreta o bien se trata de una variable escalar continua. Esta distinción no siempre es clara y depende el tipo de análisis del interés del analista o del investigador como veremos en cada una de ellas. Así pues, debemos considerar la naturaleza de la variable y como está expresada en el archivo para poder planificar correctamente su descripción.

En una primera fase, podemos describir las variables categóricas, incluyendo en este apartado tanto las de naturaleza nominal u ordinal, como las cuantitativas discretas o resultado de un conteo con pocas categorías. En segundo lugar, describiremos las variables de tipo cuantitativo continuas o discretas cuantitativas de rangos elevados. Todas las opciones harán referencia a la pestaña **Analizar** de la barra de Herramientas, en la subpestaña **Estadísticos descriptivos**.

6.3.1 Descripción de variables categóricas

Este tipo de variables se caracterizan generalmente por tener un número reducido de opciones de respuesta y su mejor descripción es la distribución de frecuencias o número de casos que se observa en cada posible valor de la variable.

Para ello, la opción más utilizada es la instrucción Frecuencias. La acción se logra mediante la consecución de **Analizar ► Estadísticos descriptivos**

► **Frecuencias**. Como podrá observar al activar la acción, a la izquierda verá todas las variables incluidas en el archivo con el símbolo que indica si la variable es nominal, nominal alfabética, ordinal o escalar.

Seleccione con el cursor las variables categóricas que desee analizar y marque la casilla inferior izquierda ■ **Mostrar Tabla de frecuencias**. Al ejecutar la instrucción que ha obtenido en la ventana de Sintaxis, tras Pegar en vez de Aceptar, tendrá la distribución de los valores de frecuencia, el porcentaje referido al total de datos del archivo, así como el porcentaje referido a los casos con valores válidos, es decir excluyendo los casos con valores *missing*, ya sean de sistema o definidos por el propio usuario. Por ejemplo, si efectuamos la acción para la variable Comida rápida, **com_rap**, el resultado se muestra a continuación en la tabla 6.2.

La sintaxis correspondiente a esta acción es:

```
FREQUENCIES VARIABLES=com_rap
  /ORDER=ANALYSIS.
```

Tabla 6.2 Comida rápida, que corresponde al resultado de ejecutar la instrucción

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	todos los días	474	6,8	7,5	7,5
	tres o más veces/ semana, pero no todos los días	848	12,2	13,4	20,9
	2 veces/semana	1346	19,3	21,3	42,2
	una vez/semana	1859	26,7	29,4	71,7
	nunca	1788	25,7	28,3	100,0
	Total	6315	90,7	100,0	
Perdidos	Sistema	649	9,3		
Total		6964	100,0		

El subcomando /ORDER = ANALYSIS implica que los resultados se expresen en el mismo orden en que ha sido codificada la variable, tal y como puede observar en la **Vista de variables**, en la pestaña **Valores**. Es decir, puede

obtener una presentación diferente, como que la tabla la escriba ordenando las categorías en función de la frecuencia, ya sea ascendente o descendente. Por defecto, la presentación siempre es en el orden en que se han codificado. Para cambiar el formato de presentación, en la ventana de Frecuencias, active la pestaña **Formato**.

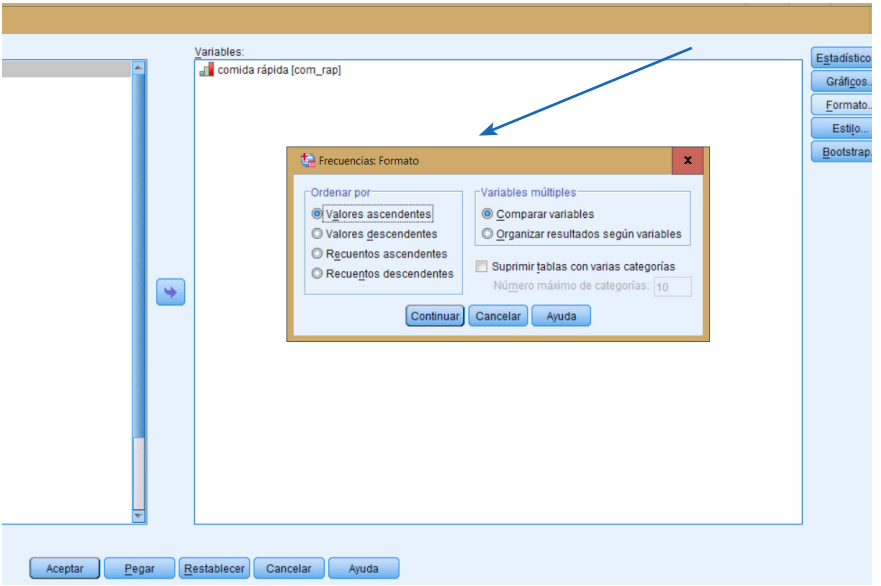


Figura 6.1 Opción de cambiar el formato de presentación

Si, en la ventana que se muestra en la figura 6.1, escogiese la opción recuentos descendentes, la sintaxis indicaría el cambio.

```
FREQUENCIES VARIABLES=com_rap
  /FORMAT=DFREQ
  /ORDER=ANALYSIS.
```

Y el resultado sería:

Tabla 6.3 Comida rápida en orden de frecuencias observadas

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	una vez/semana	1859	26,7	29,4	29,4
	nunca	1788	25,7	28,3	57,8
	2 veces/semana	1346	19,3	21,3	79,1
	tres o más veces/ semana, pero no todos los días	848	12,2	13,4	92,5
	todos los días	474	6,8	7,5	100,0
	Total	6315	90,7	100,0	
Perdidos	Sistema	649	9,3		
Total		6964	100,0		

Como puede observar, las frecuencias no han cambiado, pero es una opción muy útil a la hora de redactar los resultados, ya que, con mucha facilidad, puede expresar la opción u opciones más frecuentes o también las menos frecuente. Observe la diferencia entre la columna de Porcentaje y de Porcentaje válido. La primera se refiere a todos los casos del estudio (6964), mientras que el Porcentaje válido se refiere a los casos de los que hay información, es decir, 6315.

En el caso en que la variable sea ordinal, la columna Porcentaje acumulado nos indica el porcentaje en forma acumulada; en este caso, podríamos decir que un 57,8 % de los casos consumen comida rápida nunca o una vez a la semana. Si la variable es de tipo nominal, aveces no tiene sentido la acumulación, más que si el formato es de orden ascendente o descendente.

Como en todas las descripciones de variable categórica, la descripción puede realizarse de forma gráfica, para lo que debe seleccionar en la ventana de frecuencias la pestaña **Gráficos,** en la cual se muestra el tipo de gráfico que puede realizar y seleccionar si quiere representar las frecuencias o los porcentajes.

Comentario:
En una variable categórica NUNCA seleccione el Histograma que solo tiene sentido en variables escalares continuas.

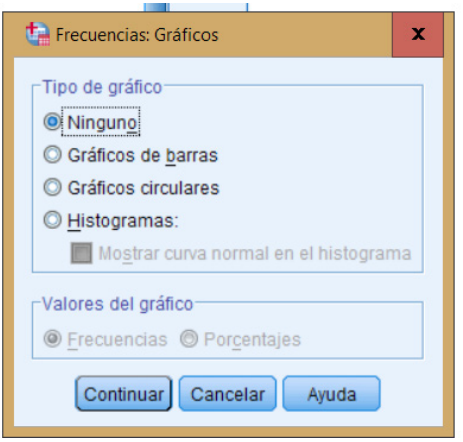


Figura 6.2 Selección de gráficos

La sintaxis para el gráfico de barras es:

```
FREQUENCIES VARIABLES=com_rap
  /BARCHART FREQ
  /FORMAT=DFREQ
  /ORDER=ANALYSIS.
```

O bien de gráficos circulares (PIECHART) es:

```
FREQUENCIES VARIABLES=com_rap
  /PIECHART PERCENT
  /FORMAT=DFREQ
  /ORDER=ANALYSIS.
```

En este caso de gráfico circular es más informativo expresarlo en porcentajes:

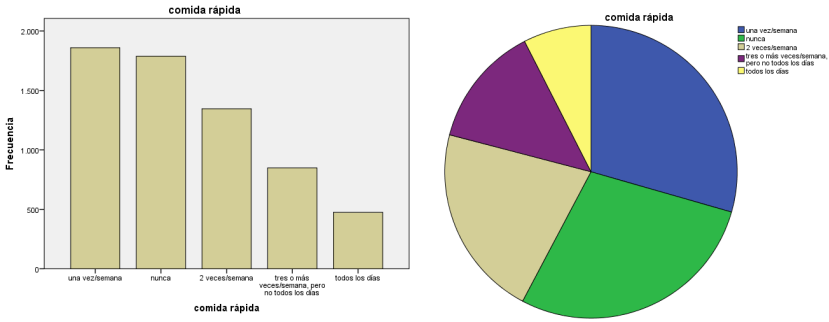


Figura 6.3 Gráficos de barras y circular de la variable Com_rap

En la instrucción Frecuencias, existen más opciones, pero solo son utilizables si la variable es escalar y serán descritas en ese apartado.

6.3.2 Descripción de variables escalares continuas o discretas cuantitativas

Para este tipo de variables, también puede utilizarse la instrucción FRECUENCIAS si bien con opciones diferentes. Por ejemplo, si quisiéramos describir la variable peso, una vez activada la acción **Analizar ► Estadísticos descriptivos ► Frecuencias** e indicado que la variable a analizar es la talla en metros, lo primero que debe efectuar es anular la opción Mostrar tablas de frecuencias, ya que en caso contrario obtendría la lista de tallas de los 6964 estudiantes.

Una vez anulada, deberá introducir qué información de carácter descriptivo requiere, común a todas las variables de tipo escalar cuantitativo. Al presionar la pestaña Estadísticos, se le muestran todas las opciones que constituyen una descripción completa de este tipo de variable:

- Medidas de tendencia central, como media y mediana y moda.
- Medidas de dispersión: desviación estándar y varianza, así como la distribución de los percentiles o percentiles específicos, cuartiles, deciles o valores concretos.
- Medidas de simetría o Sesgo y la forma de la distribución o Curtosis.
- Valor máximo y valor mínimo, así como el rango.

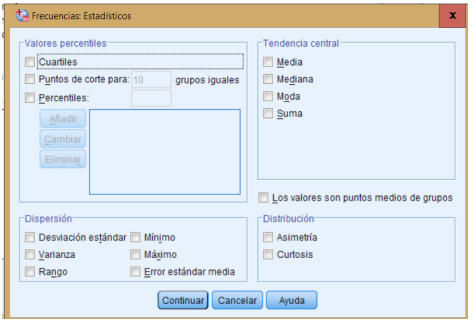


Figura 6.4 Estadísticos muestrales disponibles

Si analiza simultáneamente varias variables de esta naturaleza, los estadísticos seleccionados los calculará para todas las variables.

Comentario:

Reflexione ahora por qué se le indicó en el análisis de variables categóricas el no utilizar la pestaña de Estadísticos

En este punto, recomendamos al lector que repase en cualquier manual de estadística básica la definición de estos estadísticos y su interpretación. Solo debe tener en cuenta que, en el caso del estadístico de forma o curtosis, el SPSS, al calcularlo, le resta el valor de 3 que es el valor de esta propiedad en una distribución normal.

Asimismo, también puede obtener una representación gráfica con las opciones que se le ofrecen en la pestaña Gráficos escogiendo la opción Histograma con o sin superposición de una línea que le da información solo gráfica de como los datos se ajustan a este tipo de distribución.

Comentario:

La mayor parte de pruebas estadísticas se basan en que los datos siguen una distribución normal o campana de Gauss, pero para ello existen pruebas estadísticas que analizan si es o no esa distribución y es mejor que fiarse de una impresión visual del gráfico.

Por otro lado, si una variable sigue una distribución normal o semejante, debe cumplir con que la media y la mediana son iguales, que la asimetría o sesgo es cero y que la curtosis no se separa del valor 3 o del 0 si sigue el criterio del SPSS.

En este tipo de acción, utilizando la variable peso, la sintaxis correspondiente es:

```
FREQUENCIES VARIABLES=peso
  /FORMAT=NOTABLE
  /NTILES=4
  /STATISTICS=STDDEV VARIANCE MINIMUM MAXIMUM MEAN MEDIAN
MODE SKEWNESS  KURTOSIS
  /HISTOGRAM NORMAL
  /ORDER=ANALYSIS.
```

Lo cual corresponde a analizar la variable sin escribir la tabla de frecuencias (NOTABLE), calculando los cuartiles, la desviación estándar, la varianza, máximo, mínimo, media, moda sesgo o asimetría (SKEWNESS) y curtosis. Como gráfico, se solicita dibuje el histograma y sobrepuesta la campana de Gauss con esa misma media y desviación estándar. Los resultados son:

Tabla 6.4 Resultados de la descripción completa de la variable *peso*

Estadísticos Peso (kg)			
N	Válido		6964
	Perdido		0
Media			43,848
Mediana			43,200
Moda			44,5
Desviación estándar			10,5911
Varianza			112,170
Asimetría			,677
Curtosis			1,073
Mínimo			19,7
Máximo			114,7
Percentiles		25	36,000
		50	43,200
		75	50,300

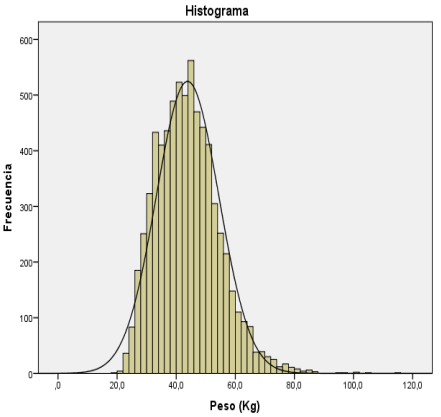


Figura 6.5 Histograma de la variable peso

Como puede leer en los resultados, media y mediana prácticamente coinciden, y si bien la asimetría de signo positivo, pequeño, indica una tendencia a valores por encima de la media y la curtosis, positiva (1,073) implica una curtosis real de 4,073, hacen a esta distribución que asumamos que es ligeramente diferente de la normal. Por otro lado, nos indica que no hay valores perdidos, si bien el valor máximo 114,7 merecería una revisión.

Ejercicio 6.3

Describa las variables **IMC** y **p_grasa** y comente la posible normalidad de las distribuciones.

Analice los valores erróneos y defínalos como ***missing*** de usuario. Repita el análisis e interprete los cambios al eliminar valores erróneos.

Otra opción más simple de describir una variable escalar es mediante la función DESCRIPTIVOS, la cual se obtiene mediante la acción Analizar ► Estadísticos descriptivos ► Descriptivos. Esta acción, si se utilizase con las variables del ejemplo, tendría como expresión de sintaxis:

```
DESCRIPTIVES VARIABLES=p_grasa IMC
  /STATISTICS=MEAN STDDEV MIN MAX.
```

Como puede observar en la figura 6.5, esta instrucción ofrece mucha menos información que la descrita anteriormente de FRECUENCIA.

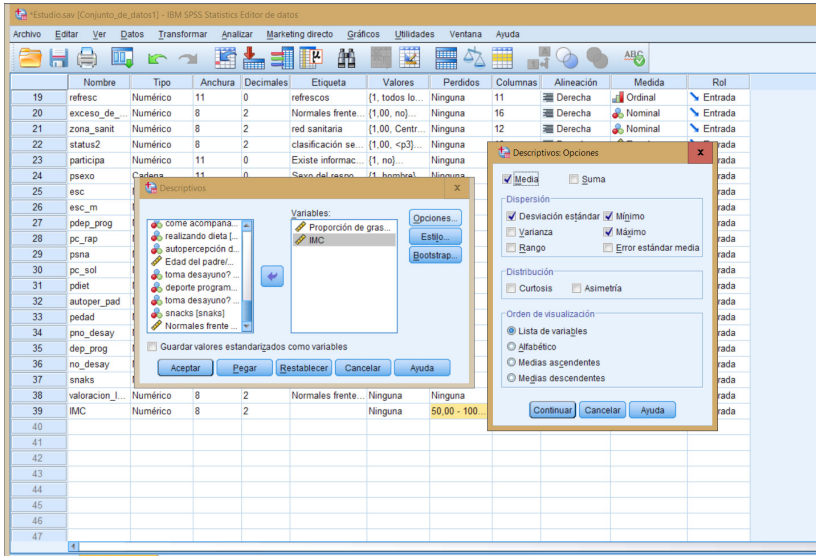


Figura 6.6 Ventana de la acción Descriptivos con las opciones de los estadísticos muestrales

Comentario:

Evidentemente, la información dada por Descriptivos es muy limitada, por lo que siempre puede utilizar Frecuencias, teniendo en cuenta que la variable es escalar o nominal o alfanumérica. En estos dos últimos casos, evite solicitar estadísticos y escoja el gráfico pertinente.

6.4 Descripción bivariada

La descripción conjunta de dos variables debe considerar la naturaleza de estas. En esta sección, se describen algunas opciones de análisis.

6.4.1 Variable cuantitativa según una variable categórica

En los dos ejemplos de descripción utilizados, puede preguntarse, ¿será diferente la distribución según el sexo del estudiante? La resolución de esta pregunta puede abordarse de dos formas.

La primera, segmentando el archivo **Datos ► Segmentar el archivo**, SPLITT FILE, por la variable sexo y ejecutar el análisis de frecuencias. La segunda, utilizando la acción Analizar ►Estadísticos descriptivos ► Explorar.

Con esta acción se describe una variable cuantitativa que, en la ventana, se identifica como Dependiente en función de los valores de otra variable, categórica, que, en la ventana, se reconoce como factor. Las opciones de descripción son muy variadas como puede ver en las pestañas de la ventana.

Al activar la pestaña Estadísticos, verá que aparentemente dice Descriptivos, intervalos de confianza y percentiles. Por defecto, los descriptivos ofrecen una amplia gama de indicadores a los que se añade respecto a la instrucción de frecuencias, los intervalos de confianza, así como otros parámetros y la posibilidad de listar los valores atípicos por separarse muchas desviaciones estándar de la media.

También ofrece la posibilidad de etiquetar los casos mediante una etiqueta que conste en otra variable. Por defecto, los etiquetará, los casos más relevantes, con el número de registro dentro del fichero. En la pestaña de Gráficos, se presentan diversas opciones. Le recomendamos que solicite la información eliminando la opción gráfica de tallo y hojas. Por otro lado, tiene la opción de analizar, mediante el test de Levene si los grupos poseen igual varianza en la variable descrita. La sintaxis generada al pegar es:

```
EXAMINE VARIABLES=IMC BY sexo
  /PLOT BOXPLOT HISTOGRAM
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

En este caso, se ha solicitado un histograma de IMC para cada sexo, así como el intervalo de confianza del 95 % para los indicadores muestrales. También se solicita el gráfico del histograma de cada subgrupo y el *boxplot*.

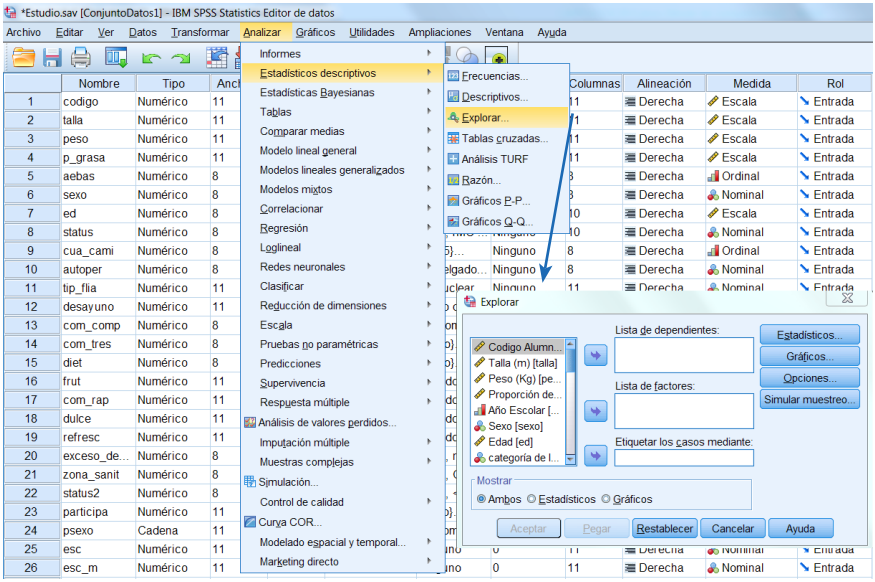


Figura 6.7 Ventana de análisis por la acción **Analizar ► Estadísticos descriptivos ► Explorar**

Una vez abierta y seleccionada la variable dependiente que quiere explorar en función de la independiente o factor, en las pestañas Estadísticos y Gráficos, puede seleccionar la información que requiere. Le recomendamos que use las opciones estadísticas que le ofrece el programa y, en gráficos, no use la de tallo y hojas y, si no sabe interpretar, no utilice las pruebas de normalidad.

Si, como variable dependiente, desplaza de la lista de variables el IMC y, en factores, desplaza la variable sexo, el resultado que obtiene se describe en primer lugar en la tabla 6.5 en la cual se indican todos los valores de los estadísticos muestrales correspondientes a cada sexo.

Si ha solicitado que se represente el histograma, el resultado se muestra en la figura 6.8 así como en la figura 6.9 se describen los diagramas de caja o *boxplot*. El resultado más importante se muestra en la siguiente tabla:

Tabla 6.5 Descriptivos de la variable IMC por sexo

	Sexo			Estadístico	Error estándar
IMC	hombre	Media		19,5875	,05486
		95 % de intervalo de confianza para la media	Límite inferior	19,4799	
			Límite superior	19,6950	
		Media recortada al 5 %		19,4067	
		Mediana		19,0448	
		Varianza		9,790	
		Desviación estándar		3,12890	
		Mínimo		10,24	
		Máximo		38,44	
		Rango		28,19	
		Rango intercuartil		3,89	
		Asimetría		,983	,043
		Curtosis		1,539	,086
	mujer	Media		19,9643	,05236
		95 % de intervalo de confianza para la media	Límite inferior	19,8616	
			Límite superior	20,0669	
		Media recortada al 5 %		19,8143	
		Mediana		19,5628	
		Varianza		10,156	
		Desviación estándar		3,18679	
		Mínimo		10,53	
		Máximo		40,69	
		Rango		30,16	
		Rango intercuartil		4,07	
		Asimetría		,818	,040
		Curtosis		1,351	,080

La información es muy extensa. Muestra la media aritmética, con su intervalo de confianza del 95 %, lo cual nos permite, en este caso, al comparar entre los dos sexos, que los intervalos no se solapan. Luego, en principio y a falta de más pruebas, podemos decir que tienen valores de IMC en media diferentes. Luego, el IMC en mujeres es mayor que en los hombres, lo cual también se confirma por el valor de las medianas. En los dos subgrupos, la simetría indica tendencia a valores por encima de la media, es decir cola a la derecha, y también la forma de la distribución con más valores extremos que en la distribución normal, ya que la curtosis, una vez sumado el valor de 3 nos da valores de 4,539 y 4,351.

Comentario:

Como puede informase en cualquier manual de estadística básica, las distribuciones se clasifican según el valor de la curtosis en platicúrticas, valores inferiores a 3, en las que hay mucha dispersión pero no valores extremos o a más de 3 desviaciones estándar, mesocúrticas en las que la proporción de valores extremos como en la distribución normal no superan el 1 %, y las leptocúrticas en las que hay una proporción apreciable de valores extremos y su curtosis es superior a 3.

Obviamente no puede juzgarse de forma estricta por el número obtenido, ya que es un valor muestral, sino que, al valor encontrado, debe sumar y restar el doble del valor de su error estándar y entonces decidir. Por ejemplo, una curtosis de 2,5 con un error estándar de 0,6 implicaría que el intervalo de confianza aproximado de ese valor de curtosis puede oscilar entre 3,7 y 1,3 por lo que incluye el valor 3 y por tanto sería mesocúrtica. Los histogramas obtenidos son:

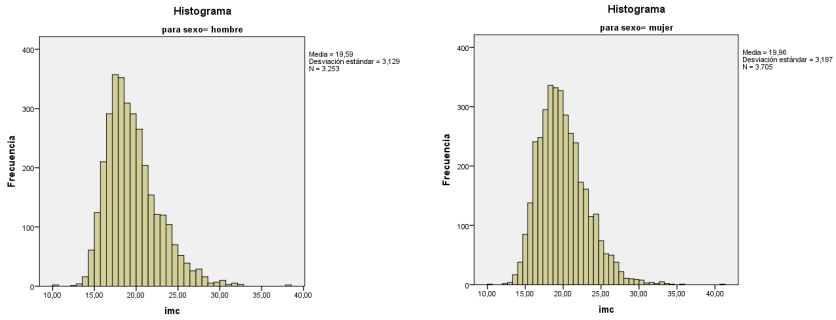


Figura 6.8 Representación de los histogramas de la opción Explorar

En los cuales puede observarse la asimetría y la presencia de colas, lo cual aleja estas distribuciones de la normalidad. Un gráfico característico de este análisis descriptivo es el *boxplot*.

6.4.2 Boxplot

Es una representación bivariada de una variable continua. En las cajas (*box*), se dibuja la porción de datos que se encuentran en el rango Intercuartílico, es decir, entre el percentil 25 y el percentil 75.

Desde los extremos de la caja nace una línea que se corta por otra de forma perpendicular, llamada bigote, cuyos extremos se calculan multiplicando por 1,5 la longitud de la caja, desde la parte superior y desde la parte inferior. Desde esos puntos, se define como valor atípico, representado por el símbolo O, aquellos valores que se denominan atípicos. Finalmente, en los más lejanos y con un * se nos señalan los casos que están a más de 3 longitudes, o sea los extremos. Los cuales siempre ameritan revisión de los datos por ser además valores influyentes en la media, ya que esta propiedad es muy sensible a la presencia de valores extremos.

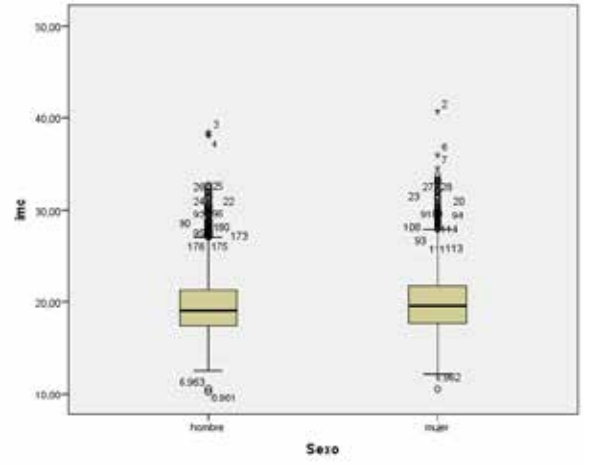


Figura 6.9 Gráfico conjunto de la distribución de valores. Diagrama de cajas.

Para localizarlos con facilidad, además del símbolo *, a su lado consta su número de registro, orden dentro del archivo, el cual no debe confundirse con el código que se haya dado por el investigador.

Ejercicio 6.4

Efectúe la descriptiva bivariada de la proporción de grasa según el sexo de los estudiantes.

Tenga en cuenta que si, por algún motivo, ha ordenado el archivo por cualquier criterio, el número identificativo que obtiene en el gráfico cambia y se refiere a la situación en el archivo después del nuevo orden introducido.

Compare el resultado obtenido si en la pestaña de Gráficos selecciona variables o factores juntos.

Indique el código de los alumnos con valores extremos.

6.4.3 Descripción conjunta de dos variables categóricas: tablas de contingencia

En la mayoría de los análisis de cuestionarios, la presencia de variables categóricas es muy frecuente y su descripción conjunta permite ver cómo se distribuye una de las variables en función de las categorías de la otra. Es importante también considerar si esas distribuciones son diferentes o no y la probabilidad de que difieran, más allá de que, al ser muestras, siempre diferirán por azar o muestreo.

Generalmente, en estas variables, hay una categoría que resume el objetivo del estudio, como es en nuestro caso el exceso de peso. Por lo que es necesario preguntarse si el exceso de peso depende del sexo, de la edad, del hecho de desayunar o no, etc.

La presentación de ese análisis se conoce como la presentación de las tablas de contingencia, cruzadas o *crosstabs* en el programa SPSS. El cruce de dos o más variables produce una tabla en la que las celdas de la misma se corresponden con la intersección de las categorías de las variables que se analizan.

La acción para crearlas es **Analizar ► Estadísticos descriptivos ►Tablas cruzadas**. En la ventana emergente, observará que le solicita qué variable quiere situar en las columnas y cuál en las filas. Esto es en realidad intrascendente, pero, por convención, es preferible colocar en las columnas la variable de interés, si existe, y en las filas, la variable explicativa.

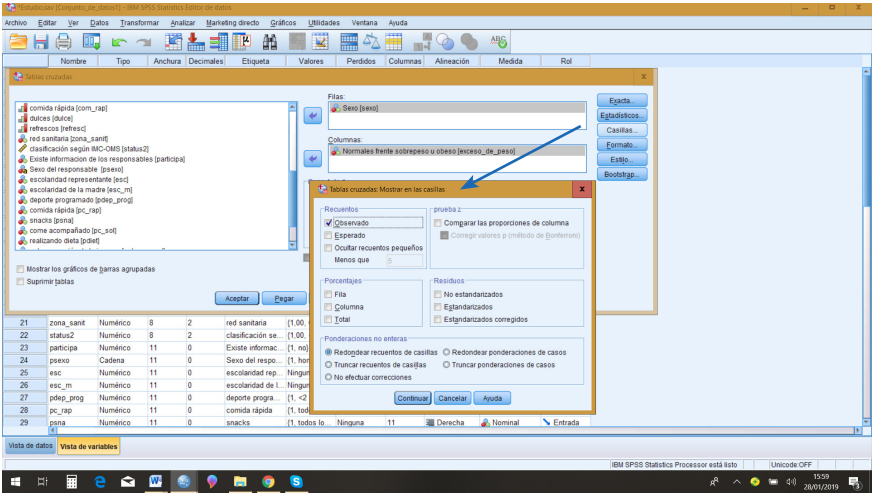


Figura 6.10 Ventana de tablas cruzadas creada por la variable exceso de peso en función del sexo

Como evidencia la figura anterior, al seleccionar la pestaña Casillas, puede escoger entre que escriba la frecuencia de cada casilla observada, así como si solicita el porcentaje de cada casilla respecto al total de su Fila o bien respecto al total de Columna en la tabla o respecto al Total. Puede también solicitar los tres porcentajes simultáneamente.

Si desea conocer si las distribuciones difieren más allá de lo que podría explicarse por el muestreo, en la pestaña Estadísticos, se le ofrecen diversos contrastes, siendo el más frecuente el valor de Chi-Cuadrado.

Si existen pocos casos o bien quiere ser algo más riguroso, en la pestaña Exactos, tiene la posibilidad de solicitar ese contraste estadístico sin aproximaciones, como es el test de Chi Cuadrado. La sintaxis generada por esta acción es:

```
CROSSTABS
  /TABLES=sexo BY exceso_de_peso
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT ROW
  /COUNT ROUND CELL.
```

Instrucción que indica que usted solicita una tabla cruzada entre la variable sexo en fila y la variable exceso de peso en columna, calculando el estadístico de contraste Chi Cuadrado y que se escriban el número de casos (COUNT) en cada casilla y en ellas el porcentaje respecto al total de la fila de la casilla (ROW). El resultado sería, en primer lugar, un resumen de la tabla.

Tabla 6.6 Resumen de procesamiento de casos

	Casos					
	Válido		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Sexo * Normales frente sobrepeso u obeso	6961	100,0 %	3	0,0 %	6964	100,0 %

Es decir, hay tres casos perdidos por falta de información en una o en las dos variables. En segundo lugar, la tabla cruzada es:

Tabla 6.7 Sexo*Normales frente sobrepeso u obeso tabulación cruzada

			Normales frente sobrepeso u obeso		Total
			no	sí	
Sexo	hombre	Recuento	2274	979	3253
		% dentro de Sexo	69,9 %	30,1 %	100,0 %
	mujer	Recuento	2834	874	3708
		% dentro de Sexo	76,4 %	23,6 %	100,0 %
Total		Recuento	5108	1853	6961
		% dentro de Sexo	73,4 %	26,6 %	100,0 %

Lo cual nos indica que hay 979 hombres con exceso de peso que es el 30,1 % de los hombres del estudio y 23,6 % de mujeres en la misma situación es decir 874 de las 3708 mujeres del estudio.

Comentario:

Tenga mucho cuidado al interpretar los porcentajes de si estos los ha solicitado respecto al total de cada fila o a otros totales, ya que esto lleva frecuentemente a confusión.

Por último, si el lector ha solicitado el estadístico de contraste de independencia, la información se completa con la siguiente tabla. Como se aprecia, el SPSS arroja cinco pruebas según la distribución Chi cuadrado, a excepción de la prueba exacta de Fisher. El programa SPSS calcula de forma asintótica (aproximada) el valor de la probabilidad del estadístico, el cual si es inferior al 0,05 se interpreta como que se rechaza la hipótesis nula de independencia o de igualdad de proporciones o de porcentajes. Este valor se encuentra en la columna etiquetada como **Sig. Asintótica.**

Tabla 6.8 Pruebas de independencia según la distribución Chi cuadrado

Pruebas de chi cuadrado					
	Valor	gl	Sig. asintótica (2 caras)	Significación exacta (2 caras)	Significación exacta (1 cara)
Chi cuadrado de Pearson	37,764[a]	1	,000		
Corrección de continuidad[b]	37,431	1	,000		
Razón de verosimilitud	37,709	1	,000		
Prueba exacta de Fisher				,000	,000
Asociación lineal por lineal	37,759	1	,000		
N de casos válidos	6961				

a. 0 casillas (0,0 %) han esperado un recuento menor que 5. El recuento mínimo esperado es 865,94.

b. Solo se ha calculado para una tabla 2x2.

En esta tabla se indica que el estadístico de contraste Chi cuadrado es 37,764, lo cual implica que la probabilidad de que las dos distribuciones de exceso de peso, para hombres y mujeres sea igual es inferior a 0,001; por lo tanto, se concluye que es un valor muy bajo como para decir que son iguales

y que sean debidas al muestreo o al azar. Ese valor denominado significación es conocido vulgarmente como el valor P.

En esta tabla, se muestran diversas pruebas de contraste, ya que el valor de Chi cuadrado es aproximado pues en realidad fue creado para analizar las varianzas de distribuciones continuas, por lo que el programa aporta una corrección, Corrección de continuidad porque, en la variable estudiada, las frecuencias no son de tal naturaleza. Vea que a pie de página nota el número de casillas con valores esperados menores de cinco casos, que es la condición de aplicabilidad de este test en el estudio de frecuencias.

Asimismo, se ofrece el contraste basado en la razón de verosimilitudes, *Likelyhood Ratio* o LR, equivalente al test de *Chi cuadrado*, pero con menos restricciones de aplicabilidad. En caso de haberlo solicitado en la pestaña Exacta se le ofrece el test de contraste exacto de Fisher.

Los valores esperados, que también puede solicitarlos en la pestaña Casillas, representan la frecuencia esperada si ambas variables fuesen independientes y su cálculo es Total de fila * Total de columna / Total de la tabla, lo cual equivale a la probabilidad de la casilla —probabilidad de fila por probabilidad de columna— multiplicado por el total de casos de la tabla.

Piense que si la pregunta es si las dos distribuciones son iguales, es equivalente a decir que el exceso de peso es independiente del hecho de ser hombre o mujer. Si la significación es menor al 5 % el resultado puede expresarse como que el exceso de peso está asociado a la variable sexo. En el siguiente apartado vemos la forma de calcular la fuerza de la asociación.

6.4.4 Riesgo

La fuerza de asociación es una medida de cuánto son dependientes las dos variables y, por cuestiones históricas que se derivan de los estudios de casos y controles, la nomenclatura, inadecuada formalmente, que utiliza el programa es el riesgo. Estos índices pueden obtenerse en la pestaña Estadísticos solicitando explícitamente la opción Riesgo.

Así hemos visto que el exceso de peso es mayor en hombres que en mujeres, por lo que podemos de alguna forma decir que los hombres tienen mayor riesgo de sufrir exceso de peso. Todas estas maneras de expresar los resultados, como veremos al final, son bastante equivalentes.

Si hubiésemos solicitado esta opción, habríamos obtenido una tabla de resultados adicional, en la cual se ofrecen tres medidas del riesgo con sus intervalos de confianza.

La interpretación de si ese riesgo existe se logra observando si, en el intervalo de confianza, se incluye la unidad, es decir, tienen el mismo riesgo o bien no incluye la unidad si existe el riesgo. Que dichos valores sean mayores o menores de la unidad depende de cómo se han codificado las variables y, por lo tanto, qué filas compara o bien qué columnas relaciona.

Siempre tiene que coincidir que no aparezca la unidad en los intervalos de confianza con el hecho de que la relación entre las variables sea significativa (significancia <0,05).

Tabla 6.9 Estimación de riesgo

	Valor	Intervalo de confianza de 95 %	
		Inferior	Superior
Odds ratio para Sexo (hombre / mujer)	,716	,644	,797
Para cohorte Normales frente sobrepeso u obeso = no	,915	,889	,941
Para cohorte Normales frente sobrepeso u obeso = sí	1,277	1,181	1,381
N de casos válidos	6961		

La interpretación de estos resultados puede ser muy confusa y debe analizarse con mucha precaución. Esto se debe a que hay diferentes diseños de estudio cuyo resultado final puede expresarse como una tabla bidimensional y que el SPSS no considera qué tipo de diseño se está analizando. Por esta razón, existe una gran diversidad de formas para expresar el Riesgo o asociación entre dos variables categóricas.

Verá en la tabla 6.9 de resultados que, si bien el exceso de peso es mayor en hombres, un 27,7 % mayor en ellos que en las mujeres —exceso de peso en hombres / exceso de peso en mujeres = 1,277—, en la tabla de riesgos, se denomina a ese número para cohorte Normales frente a sobrepeso u obeso = sí, con lo cual es confuso.

Le recomendamos previamente conocer cuál es ese resultado para seleccionar en la tabla de riesgos y poder dar el intervalo de confianza de ese número que, en el caso de un estudio transversal, se denominaría razón de

probabilidades o de prevalencia, y nunca riesgo. El número que da la tabla depende de cómo se ha codificado el sí y el no, ya que, si se intercambian, el valor es el inverso.

6.4.5 Razón de Odds

Este caso es igual de engañoso. La *odds* es el cociente de probabilidad de que ocurra una propiedad; en este caso, el exceso de peso, en una categoría respecto a la probabilidad de que no ocurra.

Así, en este resultado, la *odds* de exceso de peso en hombres es de 0,301/0,699= 0,4306; mientras que en las mujeres es 0,236/0,764= 0,3089. Es decir, nos indica cuánto es más probable o menos, como en este caso, el fenómeno que estudiamos en cada sexo. Su cociente conocido como *Odds ratio*, OR, razón de ventajas, razón de oportunidades o de momios, nos indica la fuerza de la asociación. En este caso, la propiedad vale 1,394.

Este valor no quiere decir que el exceso de peso es 39,4 % más probable en hombres que en mujeres, si no que las *odds* tienen esa relación y por lo tanto la *odds* en hombres es mayor que en mujeres. Lo que implica que el fenómeno de exceso de peso también lo es, pero 1,277 veces como indicaba la razón de proporciones o prevalencias, no como indica la OR, la cual, por definición, siempre es mayor que la razón de proporciones o prevalencias.

También en este caso el orden de codificación de las variables cambia el valor, si observa el valor que aporta el programa OR = 0,716 es justamente el inverso del calculado, porque el criterio del programa implica una codificación contraria, es decir que siempre compara la primera columna con la segunda o la primera fila con la segunda.

Estas características deben hacerle muy cauto a la hora de dar los resultados, pero nunca se equivocará si siempre se refiere a la fila que tenga una prevalencia menor, poniendo en el numerador la de mayor prevalencia.

6.4.6 Datos dependientes o apareados

En múltiples ocasiones, se dispone de información acerca de una variable que se determina en dos tiempos, situaciones o personas diferentes, pero referidas a los mismos sujetos. Su descripción conjunta se realiza también mediante la acción Tablas cruzadas, si bien la asociación se caracteriza de forma diferente.

Así, una vez determinada que la distribución de frecuencias no se rige por el muestreo o azar, es decir, la significancia es menor de 0,05. Los análisis

se refieren a la concordancia entre las dos mediciones, es decir, los datos que aparecerían en la diagonal de la tabla, así como la forma de discrepancia.

En el primer caso, la concordancia se podría determinar de forma ingenua, dividiendo cuántos valores se encuentran en la diagonal, es decir, las dos mediciones coinciden, dividiendo la suma de las casillas de la diagonal por el total de casos en los que se ha determinado dos veces la propiedad. Sin embargo, no hay que olvidar que solo por azar, al determinar dichas variables equivalentes los resultados pueden concordar en el caso de que las dos variables fuesen independientes entre sí.

La corrección de ese efecto aleatorio se lleva a cabo con la medida del índice Kappa, el cual distrae el efecto de coincidencia aleatoria al total de casos observados que son coincidentes. Pongamos un ejemplo y es el hecho de si los estudiantes declaran hacer dieta frente a la declaración de sus padres de si sus hijos hacen o no dieta. La sintxis es:

```
CROSSTABS
  /TABLES=diet BY pdiet
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ KAPPA MCNEMAR
  /CELLS=COUNT EXPECTED TOTAL
  /COUNT ROUND CELL.
```

Que se corresponde al hecho de **Pegar** la información solicitada por ventana.

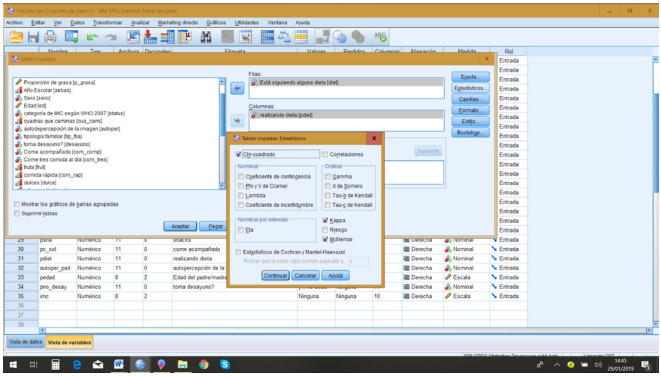


Figura 6.11 Ventana para determinar la concordancia (*kappa index*) y forma de la discordancia (*Mc Nemar test*) para analizar la respuesta a si los estudiantes hacen dieta según su declaración o la declarada por sus padres

En este caso, la tabla 6.10 que se obtiene se interpreta de la siguiente forma:

Tabla 6.10 Está siguiendo alguna dieta*realizando dieta tabulación cruzada

			Realizando dieta		Total
			no	sí	
Está siguiendo alguna dieta	No	Recuento	2915	260	3175
		Recuento esperado	2840,5	334,5	3175,0
		% del total	74,7	6,7	81,4
	Sí	Recuento	575	151	726
		Recuento esperado	649,5	76,5	726,0
		% del total	14,7	3,9	18,6
Total		Recuento	3490	411	3901
		Recuento esperado	3490,0	411,0	3901,0
		% del total	89,5	10,5	100,0

De los casos analizados, 2915 estudiantes coinciden con la opinión de sus padres en que no están haciendo dieta, y 151 en que sí están haciendo dieta. Parecería, pues, que un 78,6 % coinciden; no obstante, si se analizan los casos esperados en el supuesto en que dichas opiniones no se corresponden más que a una coincidencia al azar el número de casos coincidentes atribuibles al azar es muy elevado (2917) lo cual nos hace pensar que el nivel de concordancia más allá de lo esperado por respuestas al azar no es tan alto.

Por otro lado, hay un 6,7 % de casos en los que los estudiantes dicen que no hacen dieta y, sin embargo, los responsables dicen que sí. Por el contrario, un 14,7 % de los estudiantes indican que sí hacen dieta y los responsables que no. ¿Son estas discrepancias simétricas? ¿Tienden a decir sí los estudiantes y no sus padres más que no y los padres sí? Estas dos propiedades se describen en la tabla 6.11 que acompaña a la de contingencia cuando se ha solicitado el cálculo de Kappa y McNemar.

Tabla 6.11 Pruebas de chi cuadrado

	Valor	gl	Sig. asintótica (2 caras)	Significación exacta (2 caras)	Significación exacta (1 cara)
Chi cuadrado de Pearson	99,681a	1	,000		
Corrección de continuidadb	98,348	1	,000		
Razón de verosimilitud	85,197	1	,000		
Prueba exacta de Fisher				,000	,000
Asociación lineal por lineal	99,656	1	,000		
Prueba de McNemar				,000c	
N de casos válidos	3901				

a. 0 casillas (0,0 %) han esperado un recuento menor que 5. El recuento mínimo esperado es 76,49.

b. Solo se ha calculado para una tabla 2x2

c. Distribución binomial utilizada.

Tabla 6.12 Medidas simétricas

		Valor	Error estándar asintóticoa	Aprox. Sb	Aprox. Sig.
Medida de acuerdo	Kappa	,151	,019	9,984	,000
N de casos válidos		3901			

a. No se supone la hipótesis nula.

b. Utilización del error estándar asintótico que asume la hipótesis nula.

Observamos en primer lugar que la asociación no es por azar, pues el valor de Chi cuadrado, nos indica que la significancia es menor que el 0,05 %. En la misma tabla, el test de McNemar señala que la discrepancia no es simétrica en una significancia muy alta (<0,001) y el test Kappa también indica una concordancia significativa si bien el orden de magnitudes de la concordancia no es elevado (Kappa= 0,151). En resumen, la concordancia es pobre y la discrepancia no es simétrica.

Comentario:

La valoración del índice Kappa no se efectúa por pruebas de probabilidad si no que existen tablas que indican su valoración, siendo una concordancia pobre si el índice Kappa de Cohen es <0,2; débil si el valor está entre 0,21 y 0,40; moderada entre 0,41 y 0,60; buena entre 0,61 y 0,80 y muy buena entre 0,8 y 1.

Ejercicio 6.5

Analice la tabla cruzada desayunar y exceso de peso.

Solicite en la pestaña Casillas, la información: Observados, Esperados, Fila, Columna y Total, así como en Estadísticos los valores de Chi cuadrado y Riesgo.

Responda a las siguientes preguntas:

1.- ¿Qué proporción de los que no toman desayuno presentan exceso de peso?

2.- ¿Qué proporción de los que tienen exceso de peso no toman desayuno?

3.- ¿Qué porcentaje de estudiantes no toman desayuno y tienen exceso de peso?

4.- ¿Qué razón de prevalencia de exceso de peso existe entre los dos grupos?

5.- ¿Cuál es el valor de la OR?

6.- ¿Puede decirse que estas diferencias pueden ser debidas al muestreo?

En resumen:

- Coloque como variable columna la respuesta.
- Solicite la proporción en Filas (Columnas si la respuesta la colocó como fila).
- Observe si la relación es estadísticamente significativa, significancia menor a 0,05.
- Si y solo si el resultado es significativo analice los índices de fuerza de la asociación.

6.5 Descripción de los resultados en forma gráfica

Si bien el programa SPSS no es un programa diseñado para realizar gráficos de calidad, tiene diversas opciones, pero de bastante complejidad y solo sirven para un informe técnico, por lo que son de difícil publicación en un artículo.

Sin embargo, a título de mostrar tendencias y complementar la descripción, existe una opción más sencilla y es activando **Gráficos ► Cuadros de diálogo antiguos**. Es inmediata la obtención de figuras que reflejan los

resultados en forma de histogramas, diagramas de dispersión, o una pirámide poblacional, diagramas de barras en tres dimensiones o dibujar las barras de errores estándar en variables continuas.

Como ejemplo, veremos la generación del gráfico de la pirámide poblacional de los casos estudiados, es decir, la gráfica que representa las frecuencias de cada edad para cada sexo. La sintaxis correspondiente sería:

```
XGRAPH CHART=[HISTOBAR] BY ed[s] BY sexo[c]
 /COORDINATE SPLIT=YES
 /BIN START=AUTO SIZE=AUTO
 /TITLES
  TITLE='Pirámide poblacional sexo edad'.
```

Instrucción que solicita la gráfica de la pirámide sexo / edad utilizando los valores por defecto de la opción gráfica. El resultado se muestra a continuación, asegurando previamente la ausencia de valores erróneos en la edad.

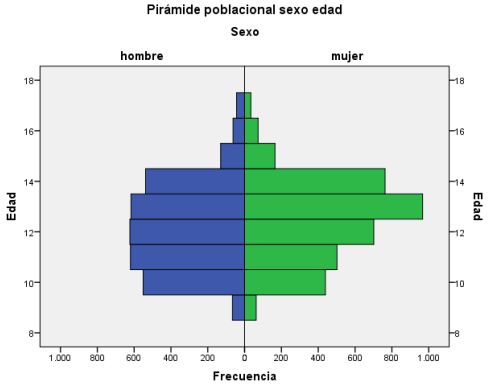


Figura 6.12 Representación gráfica de la distribución de casos en función de la edad y el sexo de los estudiantes

La figura refleja en realidad el sistema de muestreo exhaustivo de los alumnos por cursos lectivos, por lo que la distribución muestra la menor presencia de alumnos con edad de 9 y de 17 años.

Otro ejemplo de gráfico útil para mostrar tendencias podría ser la representación de la dispersión conjunta de los valores de talla y peso.

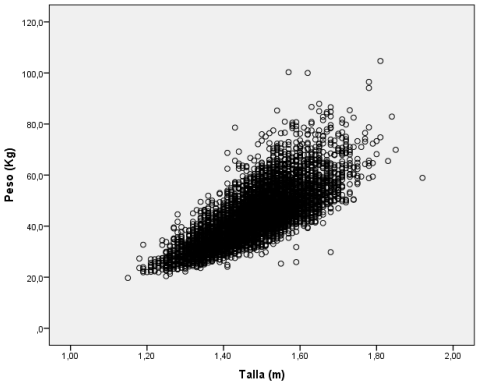


Figura 6.13 Gráfico que muestra la relación, caso a caso, entre las dos variables

Nota. Es evidente la tendencia a aumentar el peso a medida que aumenta la talla.

Supongamos que deseamos representar el valor de IMC en función de la edad y sexo, pero no los valores sino el valor medio de IMC con su intervalo de confianza.

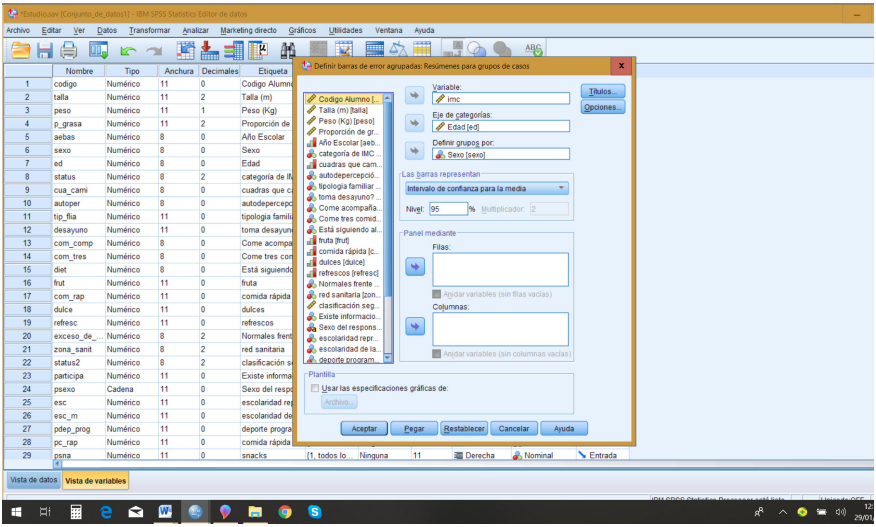


Figura 6.14 Ventana de definición del gráfico de medias de IMC por sexo y edad con sus intervalos de confianza

El gráfico resultante, cuya sintaxis es

```
GRAPH
  /ERRORBAR(CI 95)=imc BY ed BY sexo.
```

se muestra en la siguiente figura:

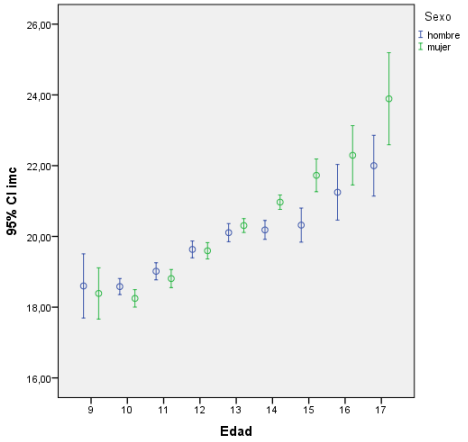


Figura 6.15 Representación gráfica de la media el IMC en función de la edad, para cada sexo

Un último ejemplo sería representar la talla y el peso conjuntamente con la edad. Para ello, la acción sería **Gráficos ► Cuadro de diálogo antiguos ► Dispersión ► Dispersión 3D ► Definir**.

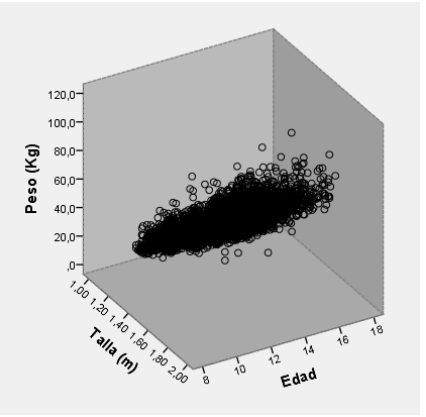


Figura 6.16 Representación tridimensional

En la ventana de resultados donde se encuentra el gráfico, al marcar dos veces sobre la figura permite la edición del gráfico, así como rotar la figura para obtener la visualización que refleje mejor, a criterio del investigador, la relación entre las tres variables. Como ha podido observar, no son gráficos de calidad para incluir en el texto de un artículo, aunque sí que se puede copiar y pegar la figura en un texto de Word. Dejamos al lector la exploración del resto de opciones que se presentan en la pestaña **Gráficos ► Cuadros de diálogo antiguos**.

Ejercicio 6.6

Lleve a cabo una descripción completa de la variable talla en metros, indicando qué valores declara como missing de usuario.

Relacione la talla y el peso describiendo por el hecho de desayunar o no, según el sexo de los responsables de los alumnos.

Describa el valor de IMC en función de si tienen exceso de peso o no y relacione esta última propiedad en función del sexo y si desayuna o no.

Responda a la siguiente pregunta: ¿cómo abordaría la coincidencia de respuesta entre padres e hijos sobre el hecho de consumir comida rápida? ¿Concuerdan? ¿Cómo discrepan?

Realice un gráfico de las medias de talla en función de la edad y el sexo, incluyendo en el gráfico los intervalos de confianza.

Capítulo siete

Contrastes estadísticos más comunes

7.1 Introducción

El contraste estadístico responde a la forma práctica de llevar a cabo el esquema de Falsación de Popper. En resumen, consiste en no mantener una hipótesis porque se haya observado repetidamente, si no colocar circunstancias experimentales que permitan poner en crisis o duda los resultados conocidos. Es decir, el objetivo de un estudio es mostrar que las diferentes medidas que se obtienen al realizar un experimento se consideren que son diferentes debido al azar, situación que se conoce como hipótesis nula, hasta que la probabilidad de que se produzcan las diferencias observadas sea lo suficientemente pequeña como para rechazar esa hipótesis y por lo tanto proponer una alternativa que asigne esa posible diferencia a un factor que ha creado los grupos que se comparan.

Entendemos que el lector está familiarizado con los términos básicos de la Estadística Inferencial y que, si no es así, no continúe en este capítulo hasta que los adquiera.

Comentario:

Efectuar análisis estadísticos tan simples como los que a continuación se describen, así como otros más complejos que ameritan un profundo conocimiento estadístico, no pueden realizarse, aunque físicamente el programa le permita hacerlo.

Nunca lleve a cabo un contraste si no conoce su base teórica, fijándose exclusivamente en si es significativo o no, y que no se corresponda con las hipótesis de su proyecto de investigación. La probabilidad de extraer conclusiones erróneas y espurias es mucho más elevada de la que uno supone ingenuamente.

Piense que si generalmente se acepta la probabilidad de concluir que se rechaza la hipótesis nula igual al 5 %, el efectuar repetidamente **n** contrastes para ver si alguno es significativo, la probabilidad de equivocarse al menos una vez es de $(1 - 0{,}95^n)$ lo cual produce un riesgo de error criterio. Por ejemplo, si realiza 10 contrastes al azar, la probabilidad de que al menos uno sea debido al azar es del 40,1 %, muy elevado si solo se buscan, sin criterios, resultados significativos.

7.2 Comparación de dos porcentajes o dos distribuciones de frecuencias

Este contraste se refiere a la comparación de dos distribuciones de frecuencias de una misma variable generadas por los distintos valores o categorías de una variable categórica. Hemos visto cómo la distribución del exceso de peso o de la variable desayunar cambia al considerar las variables en función de si los estudiantes son niños o niñas. La representación se efectúa a través de las tablas de contingencia, aspecto ya descrito anteriormente en este capítulo.

La hipótesis nula es que la distribución de la variable que nos interesa, en este caso el exceso de peso o el tomar desayuno, no cambia en la población, si no que las diferencias que observamos son debidas al azar. Otra forma de plantearlo es que la proporción de exceso de peso o de alumnos que desayunan es igual en niños que en niñas. Por último, la hipótesis que resume los dos planteamientos es que exceso de peso o desayunar es independiente del sexo.

En el capítulo anterior, habrá podido observar que, en esquema de tablas cruzadas o de contingencia, se aporta el p valor, o significancia, y que, cuando este es inferior al 5 %, se rechaza la hipótesis nula, es decir el exceso de peso es dependiente del sexo del estudiante o también el hecho de desayunar.

Ejercicio 7.1

Analice y describa las hipótesis nulas que se asocian a la descripción de la relación entre el hecho de desayunar y la ausencia de exceso de peso, en función del sexo del estudiante.

Para ello, lleve a cabo la acción **Analizar ▸ Estadísticos descriptivos ▸ tablas cruzadas** y, en filas, coloque la variable desayuno; en columnas, la variable exceso de peso y, en la ventana de capa, la variable sexo, indicando en las casillas el porcentaje de fila. No olvide solicitar el cálculo de estadísticos y los indicadores de riesgo.

Interprete los resultados.

7.3 Comparación de dos medias muestrales

Tal y como pudimos ver en el capítulo anterior con la instrucción Explorar, al cruzar una variable continua con una categórica, se generan tantas distribuciones de la variable continua como categorías tiene la variable categórica. La pregunta más frecuente es ¿serán las medias iguales?

Es decir, tenemos una hipótesis nula que implica que las medias de la variable continua de las distribuciones generadas son en población iguales y que las diferencias que podemos observar son solo debidas al muestreo, es decir al azar. El principio de análisis se basa en considerar si las medias varían más que la varianza aleatoria que se observa dentro de los grupos generados por la variable categórica.

Si es así y la diferencia es suficientemente grande, el estadístico de contraste nos permite calcular la probabilidad de que esa diferencia sea debida al azar o no, es decir, si la diferencia es significativa. En estos casos existen dos grandes grupos de estadísticos de contraste, la t-Student de Gosset o la F de Fisher-Snedecor, en función de que el número de medias que se quieren analizar sea dos o más respectivamente. Sin embargo, su principio de análisis, como veremos a continuación, es el mismo.

Comentario:

Volvemos a insistir encarecidamente que, si desconoce estos términos o no tiene las bases estadísticas suficientes, recurra a un texto de Estadística básica.

7.4 Comparación de dos medias pertenecientes a dos muestras independientes

Realice la siguiente acción: **Analizar ►Estadísticos descriptivos ►Explorar** y obtenga la distribución de la variable IMC para cada sexo. Observará que el valor de las dos medias es diferente y que también las varianzas muestrales difieren.

El test de comparación de medias de muestras independientes más conocido es el llamado test de t de igualdad de medias poblacionales o coloquialmente t- de Student. Para realizarlo ejecute **Analizar ►Comparar medias ► Prueba T para muestras independientes** y en la ventana que se abre indique IMC en la ventana **Variables de prueba** y debajo en la otra ventana inferior, **variable de agrupación,** la variable categórica que genera las dos muestras, en este caso el sexo.

Inmediatamente debajo le pide que defina qué dos categorías tiene esta variable. En este caso, 1 y 2. Recuerde que las variables categóricas están generalmente codificadas de forma numérica y que hombre y mujer son meras etiquetas de dichos números.

En las opciones, por defecto, le indicará el intervalo de confianza de las dos medias que quiere analizar, con un nivel de confianza del 95 %, nivel que usted puede modificar. La sintaxis generada al aplicar Pegar en vez de Aceptar es:

```
T-TEST GROUPS=sexo(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=imc
  /CRITERIA=CI(.95).
```

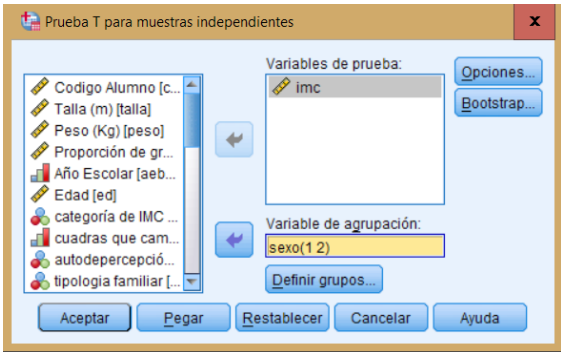


Figura 7.1 Ventana de definición de la variable continua que se analiza y la variable categórica y sus dos categorías

Al ejecutar dicha sintaxis o bien al apretar Aceptar el resultado es:

Tabla 7.1 Estadísticas de grupo

	Sexo	N	Media	Desviación estándar	Media de error estándar
IMC	hombre	3253	19,5875	3,12890	,05486
	mujer	3705	19,9643	3,19968	,05255

Puede comprobar que tanto los valores de las medias como de las desviaciones y Media del error estándar coinciden con las obtenidas mediante la acción Explorar. Asimismo, la información del contraste es la que se indica a continuación.

La construcción del estadístico de contraste t depende del hecho de si los dos grupos tienen o no varianzas iguales, y eso lo indica el test de Levene. Si no se puede rechazar que las dos varianzas sean iguales **(sig)** mayor de 0,05 el estadístico correcto es el de la fila superior —tal y como es en este caso— ya que la sig. del test de Levene es 0,225. El valor t = 4,963 nos indica que las medias varían 4,963 veces más que la varianza promedio dentro de las categorías de sexo. Esta diferencia, –0,376 es estadísticamente significativa, si bien es discutible si es una diferencia relevante antropológicamente. La significación (bilateral) obtenida no es 0,000 si no menor que 0,001 y, como solo reserva tres espacios para decimales, trunca el valor. Si desea ver el valor real de la significación, sitúese encima del ,000 y apriete dos veces en cursor y el indicará el verdadero valor, que es 7,1124E-7, es decir 0,0000007112.

Tabla 7.2 Pruebas de muestras independientes

		IMC	
		Se asumen varianzas iguales	No se asumen varianzas iguales
Prueba de Levene de calidad de varianzas	F	1,471	
	Sig.	0,225	
Prueba t para la igualdad de medias	T	–4,963	–4,969
	gl	6956	6869,97

	Sig. (bilate-ral)		0	0
	Diferencia de medias		−0,37679	−0,37679
	Diferencia de error estándar		0,07592	0,07583
	95 % de intervalo de confianza de la diferencia	**Inferior**	−0,52563	−0,52545
		Superior	−0,22796	−0,22814

Comentario:

Tenga siempre la precaución de distinguir estadísticamente significativo con resultado relevante.

En el ejemplo la pequeña diferencia de medias de IMC es significativa porque el análisis es con muchos datos y cualquier diferencia, por pequeña que sea, no es debida al azar si no al hecho de que las muestras son muy grandes, pero no necesariamente significan algo.

El programa SPSS se caracteriza por su capacidad de realizar el mismo análisis de diferentes formas. Una acción equivalente al análisis anterior es la acción **Analizar ► Comparar medias ► Medias**.

Verá que, en este caso, la diferencia es que puede analizar la media de un número de variables continuas y categóricas mayor, no una a una. Además, en las **opciones,** le ofrece la descripción de las variables continuas para cada categoría de las variables de agrupación y que estas pueden tener más de dos categorías. Asimismo, el contraste de hipótesis que realiza se basa en el análisis de la Varianza mediante el test de F Fisher, el cual tiene el mismo principio que el estadístico t. Es decir, compara la variabilidad de las medias, sean dos o más, mediante ese criterio.

Active el programa Medias e introduzca la variable continua IMC y en la lista de independientes sexo y zona sanitaria. Como observará no le pide información de las categorías, sino que asume todas las que encuentre siendo la sintaxis:

```
MEANS TABLES=imc BY sexo zona_sanit
  /CELLS=MEAN COUNT STDDEV
  /STATISTICS ANOVA.
```

En este caso la información que se le ofrece es muy parecida a la del test de t, a excepción del tipo de contraste:

Tabla 7.3 Informe previo del programa *Means*

Sexo	Media	N	Desviación estándar
hombre	19,5875	3253	3,12890
mujer	19,9643	3705	3,18679
Total	19,7881	6958	3,16522

Tabla 7.4 Tabla de ANOVA que compara la media del IMC entre hombres y mujeres

Tabla de ANOVA

			Suma de cuadrados	gl	Media cuadrática	F	Sig.
IMC * Sexo	Entre grupos	(Combinado)	245,920	1	245,920	24,630	,000
	Dentro de grupos		69 453,598	6956	9,985		
	Total		69 699,518	6957			

En esta tabla pude apreciar que el test se realiza comparando la variabilidad o varianza de las medias entre los grupos creados por la variable categórica con la variabilidad dentro de los grupos o varianza aleatoria.

En este caso el cociente o valor de F nos indica que las medias varían 24,6 veces más que la variación de la variable IMC dentro de cada categoría y eso es altamente significativo.

Comentario:

Ambos contrastes de t y F son equivalentes, hasta el punto de que el estadístico t obtenido anteriormente 4,963 es exactamente la raíz cuadrada del estadístico F obtenido 24,6. La razón es que cuando se comparan dos grupos, pero solo en esta circunstancia, la construcción de los estadísticos t y F siguen esa relación.

Entonces, ¿cuál se debe realizar? Son equivalentes, la diferencia es sutil, y es que el test de t permite comparar de forma diferente, unilateral o bilateral mientras que el test de F siempre es bilateral. Es decir, el test de t puede comparar las diferencias en un sentido, si las diferencias, son mayores o menores que la hipótesis nula y el test de F solo analiza si existen diferencias sea cual sea su sentido. Habitualmente si son dos muestras se prefiere el test de t. Si hay más de dos grupos, siempre el test de F.

Ejercicio 7.2

Analice la hipótesis nula de que las medias de la proporción de grasa corporal no dependen del exceso de peso. Efectúe el análisis mediante el **test de t** y mediante la opción **medias.** Compruebe la relación entre el valor de los estadísticos t y F.

7.5 Comparación de dos medias de muestras dependientes

Se definen muestras dependientes cuando una misma medición se lleva a cabo en los mismos individuos en tiempo diferente, ya sea antes o después de un tratamiento, o a lo largo del tiempo para ver si existe evolución. En cada momento de medición, se genera una muestra de datos, por lo que si la medición se realiza en dos momentos tendremos dos muestras de datos dependientes ya que se mide a los mismos individuos.

El test se realiza mediante la acción **Analizar ►Comparar medias ►Prueba de T de muestras dependientes**. En el ejemplo que estamos utilizando, no hay ningún tipo de variable de este tipo, pero si analiza la información de la ventana correspondiente verá la figura 7.2.

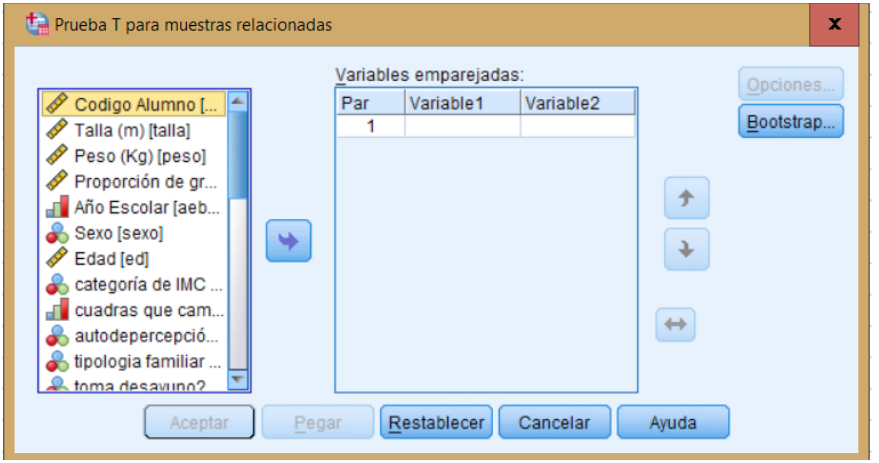


Figura 7.2 Ventana de definición de las dos variables que se aparejan, es decir la variable antes y la variable después.

La interpretación es similar al test de t pero debe interpretarse como si el cambio observado promedio se diferencia de cero o no.

7.6 Comparación de más de dos medias muestrales

Aunque, en el apartado anterior, se ha descrito cómo analizar si existen diferencias significativas entre diversas medias a través de la instrucción **Medias**, el programa SPSS ofrece una opción más potente para analizarlas. Es exactamente el mismo mecanismo, pero esta nueva opción permite dos ventajas adicionales.

En primer lugar, diseñar los contrastes *a priori* y, en el caso más clásico, no solo valorar la significación de las diferencias entre medias —es decir, hay o no diferencias—, sino que, una vez que se muestra que esa diferencia es significativa, cuáles de esas medias son diferentes unas de otras. Valoraremos esta segunda opción.

7.7 Análisis de la varianza. ANOVA

Tal y como ya se ha indicado, el análisis de varianza compara la varianza entre las medias generadas por las diferentes categorías en que subdivide a la variable continua, con la varianza promedio que se observa dentro de cada categoría, la cual se atribuye al azar. Cuando solo se considera una variable categórica, o factor, se acostumbra a denominar ANOVA de un factor. Podría analizarse varios factores simultáneamente, pero para ello se requieren conocimientos más específicos de Estadística que superan el objetivo de este libro.

Evidentemente se espera que la primera varianza, entre las medias, llamada también varianza entre grupos (media cuadrática en SPSS), sea mayor que la varianza dentro de esos grupos (media cuadrática intragrupos). Por tanto, su cociente debe superar la unidad que se correspondería con la Hipótesis nula, todas las medias en la población son iguales y por lo tanto varían igual que lo hacen las medidas individuales dentro de cada grupo.

A partir de qué valor de ese cociente se supone que ya es tan grande que no es debido al azar, o más estrictamente, cuándo la significación del

cociente, en función de los grados de libertad de ambas varianzas, es menor que 0,05. Así, al accionar **Analizar ► Comparación de medias ► ANOVA** de un factor, se abre la siguiente ventana:

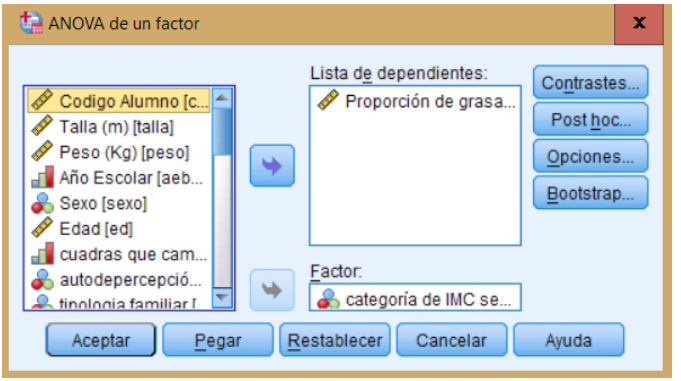


Figura 7.3 Ventana de información requerida para ejecutar un ANOVA de un factor

Además de pasar el nombre de las variables continuas (dependientes), en este ejemplo, proporción de grasa; y categórica, factor, en este caso, la categoría del IMC; podría definir qué medias se quieren comparar *a priori* en la pestaña contrastes.

Lo habitual es solicitar los contrastes "post_hoc", es decir, los contrastes entre pares de medias *a posteriori*, una vez que se sepa si el análisis es o no significativo. El número de contrastes *post hoc* es muy variado, siendo los más frecuentes el de Scheffé si los grupos tienen diferente número de datos o Tuckey en caso contrario. En esta instrucción ANOVA puede solicitar adicionalmente aspectos descriptivos y contraste de si las varianzas son homogéneas o no, así como un gráfico de las medias de los grupos.

La sintaxis del análisis más frecuente se indica a continuación, incluyendo la opción del test de Levene que analiza la homogeneidad de varianzas de cada grupo, así como el test de Sheffé para detectar que grupos se diferencias estadísticamente:

```
ONEWAY p_grasa BY status
  /STATISTICS HOMOGENEITY
  /MISSING ANALYSIS
  /POSTHOC=SCHEFFE ALPHA(0.05).
```

El resultado de ejecutar estas instrucciones es:

Tabla 7.5 Prueba de homogeneidad de varianzas. Proporción de grasas

Estadístico de Levene	df1	df2	Sig.
18,616	2	6958	,000

Esta información nos indica que se rechaza la homogeneidad de varianzas en los grupos generados. Este aspecto es importante porque el análisis parte del supuesto de homocedasticidad o igualdad de varianzas entre las subpoblaciones estudiadas, ya que, en el caso en que posteriormente no se detecte diferencia entre las medias, puede deberse precisamente a que pertenecen a grupos con dispersiones tan diferentes que la dispersión de un grupo puede llegar a ocultar las diferencias entre las medias. Solo debe considerarse cuando no se rechaza la hipótesis nula de igualdad de medias.

A continuación, se efectúa el contraste ya explicado en el apartado **Medias**.

Tabla 7.6 ANOVA. Proporción de grasa

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Entre grupos	131 339,515	2	65669,758	2012,884	,000
Dentro de grupos	227 002,688	6958	32,625		
Total	358 342,204	6960			

Podemos decir que la varianza entre las medias es 2012,88 veces mayor que la varianza dentro de los grupos, lo cual puede darse por el azar con una probabilidad menor a 0,001. En definitiva, rechazamos que sean iguales o, lo que es lo mismo, la media de proporción de grasa depende del estatus de IMC de forma muy significativa. La pregunta es la siguiente, ¿entre qué grupos, normal, sobrepeso y obeso, hay diferencias significativas? La respuesta, ya que hemos visto que hay diferencias, se describe en la tercera tabla.

Esa tabla describe las comparaciones posibles entre las tres medias y podemos observar que las tres medias son diferentes entre sí. También se describen los intervalos de confianza de las diferencias entre las medias, en las que solo hay que interpretar que el signo de la diferencia solo depende del orden en que el programa las compara. Evidentemente, una diferencia no significativa se vería también en el intervalo de confianza, ya que incluiría una diferencia 0.

Tabla 7.7 Pruebas *post hoc*

Comparaciones múltiples

Variable dependiente: proporción de grasa						
Scheffé						
(I) categoría de IMC según WHO 2007	**(J) categoría de IMC según WHO 2007**	**Diferencia de medias (I-J)**	**Error estándar**	**Sig.**	**95 % de intervalo de confianza**	
					Límite inferior	**Límite superior**
IMC normal	IMC sobrepeso	–7,88781*	,17727	,000	–8,3218	–7,4538
	IMAC obesidad	–13,08224*	,25633	,000	–13,7098	–12,4547
IMC sobrepeso	IMC normal	7,88781*	,17727	,000	7,4538	8,3218
	IMAC obesidad	–5,19443*	,29044	,000	–5,9055	–4,4834
IMAC obesidad	IMC normal	13,08224*	,25633	,000	12,4547	13,7098
	IMC sobrepeso	5,19443*	,29044	,000	4,4834	5,9055

*. La diferencia de medias es significativa en el nivel 0,05.

Subconjuntos homogéneos

Proporción de grasa

Scheffé[a,b] categoría de IMC según WHO 2007	N	Subconjunto para alfa = 0,05		
		1	2	3
IMC normal	5108	19,9707		
IMC sobrepeso	1303		27,8585	
IMAC obesidad	550			33,0529
Sig.		1,000	1,000	1,000

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

a. Utiliza el tamaño de la muestra de la media armónica = 1078,589.

b. Los tamaños de grupo no son iguales. Se utiliza la media armónica de los tamaños de grupo. Los niveles de error de tipo I no están garantizados.

El test de F se conoce por ser un test robusto ¿Qué quiere decir test robusto? Pues que resiste bastante bien el incumplimiento de las condiciones de aplicabilidad, es decir que la variable continua sigue una distribución normal y que los grupos son homocedásticos. En realidad, robusto quiere decir que, si el test detecta diferencias es que existen, aunque la variable, porcentaje de grasa, en este caso, no cumpla con esas condiciones. Solo debe preocuparse si no observa diferencias y, en este caso, se detectan.

Ejercicio 7.3

Lleve a cabo el análisis completo de la dependencia entre el porcentaje corporal de grasa y ser usuario de comida rápida.

Interprete los resultados.

7.8 Análisis de la dependencia lineal entre dos variables continuas

Cuando se desea analizar si dos variables continuas son dependientes y si esta dependencia es lineal o no, existen dos formas complementarias. La primera de ellas es el análisis de la covarianza estandarizada entre las dos covariables y que resume la fuerza de asociación entre las mismas.

7.9 Correlación

Nos cuantifica la relación que ya se pudo observar mediante el diagrama de dispersión explicado en el Capítulo 6. La cuantificación se efectúa mediante la acción **Analizar ► Correlaciones ► Bivariadas**.

Figura 7.4 Ventana de correlaciones entre pares de variables

En la ventana interna, se desplazan las variables de las que se pretende analizar si tienen o no dependencia lineal y, en coeficientes de correlación, se selecciona Pearson o Spearman en función de si las variables son normales o no, o si son variables ordinales resultado de aplicar una escala de puntuación. Como ocurre a menudo, la tecla de opciones permite obtener datos de la descripción de cada variable, las cuales generalmente ya se han descrito.

La sintaxis de esta acción representada en la figura sería:

```
CORRELATIONS
  /VARIABLES=peso talla p_grasa imc
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

Y al ejecutar el resultado se muestra en la siguiente tabla:

Tabla 7.8 Correlaciones entre variables continuas

Correlaciones					
		Peso (kg)	**Talla (m)**	**Proporción de grasa**	**IMC**
Peso (kg)	Correlación de Pearson	1	,768**	,480**	,848**
	Sig. (bilateral)		,000	,000	,000
	N	6964	6964	6961	6964
Talla (m)	Correlación de Pearson	,768**	1	,000	,325**
	Sig. (bilateral)	,000		,980	,000
	N	6964	6964	6961	6964
Proporción de grasa	Correlación de Pearson	,480**	,000	1	,732**
	Sig. (bilateral)	,000	,980		,000
	N	6961	6961	6961	6961
IMC	Correlación de Pearson	,848**	,325**	,732**	1
	Sig. (bilateral)	,000	,000	,000	
	N	6964	6964	6961	6964

**. La correlación es significativa en el nivel 0,01 (2 colas).

En este resultado, puede observarse que el IMC y el peso junto con la proporción de grasa son las variables más directamente relacionadas, ya que su coeficiente de correlación de Pearson es muy significativo ($p<0,001$), siendo mayor la relación entre IMC y Peso.

También se evidencia como la talla se relaciona, lógicamente con el peso, de forma significativa y en una dependencia directamente proporcional, r cercano a la unidad, mientras que su relación con la proporción de grasa es nula.

Comentario:

¿Cómo evaluar la magnitud del coeficiente de correlación?

El coeficiente de correlación de Pearson oscila entre los valores –1 y 1, siendo, –1 una dependencia matemática perfecta, pero de forma inversamente proporcional; cero si no existe relación lineal ninguna —lo cual no implica falta de relación de otro tipo—; y 1 si la relación es matemáticamente perfecta directamente proporcional.

Sin embargo, el estadístico tiene la siguiente propiedad y es que r^2 nos da la información en tanto por uno de la varianza que en una variable describe la varianza de la otra. Es decir, el r=0,847 entre las variables IMC y peso se debe interpretar como que las variaciones de peso producen una variación del 0,717 en la variación del IMC, o como se expresa generalmente, las variaciones de peso explican un 71,7 % de las variaciones en IMC y viceversa.

En la mayoría de casos, sin embargo, interesa conocer ¿cuánto varía la media de una variable definida como dependiente para cada cambio de una unidad en la definida como independiente? La respuesta a esta pregunta se responde efectuando un análisis de regresión lineal.

7.10 Regresión lineal entre dos variables continuas

En este caso, debe de existir una definición clara de cuál es la variable dependiente y cuál la independiente. Si consideramos la variable IMC y peso, claramente la variable dependiente es el valor de IMC, mientras que el peso es la variable independiente. Para llevar a cabo el estudio, efectúe la acción **Analizar ► Regresiones ► Lineales** en la ventana que se abre como se muestra en la figura 7.5.

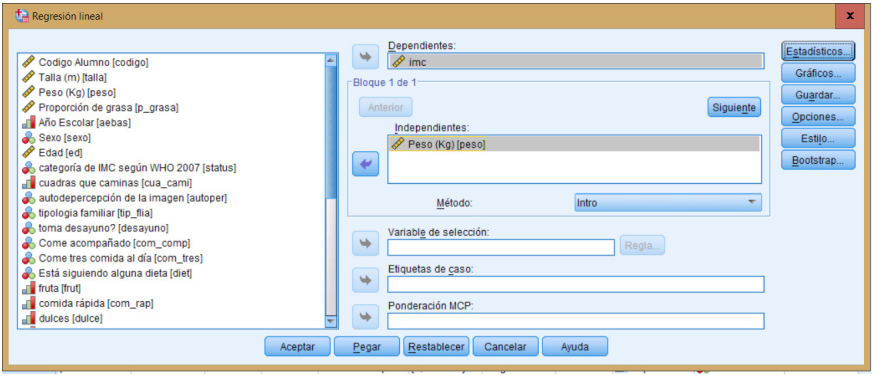


Figura 7.5 Ventana correspondiente al análisis de regresión lineal

Se introduce la variable que se pretende explicar o dependiente y la independiente en su ventana inferior. En una primera fase y sin los conocimientos necesarios de Estadística básicos, solo introduzca en la pestaña Estadísticos que nos aporte los intervalos de confianza de las estimaciones de los coeficientes del modelo lineal así como una valoración del ajuste. Todas las demás opciones se refieren en especial a modelos con varias variables independientes.

La sintaxis asociada a esta figura al accionar **Pegar** es:

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT imc
  /METHOD=ENTER peso.
```

Y el resultado se muestra a continuación:

Tabla 7.9 Resumen del modelo de regresión lineal entre las variables IMC y Peso

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,848a	,719	,719	1,68232

Como puede comprobar, los valores de R y R cuadrado son los mismos que ya habíamos conocido y que, en Regresión, se conoce como característica de ajuste del modelo. En la siguiente tabla, podrá comprender mejor el significado de R cuadrado.

Tabla 7.10 Resultados ANOVA que evalúa las diferencias entre la media del IMC con la media del Peso

Modelo		Suma de cuadrados	Gl	Media cuadrática	F	Sig.
1	Regresión	50 351,040	1	50 351,040	17 790,701	,000b
	Residuo	19 703,773	6962	2,830		
	Total	70 054,813	6963			

a. Variable dependiente: IMC.
b. Predictores: (Constante), Peso (kg).

En esta tabla, se realiza un análisis de la varianza explicada por el modelo frente a la residual o que no puede explicar. Compruebe y verá que el valor de R^2 es el cociente entre lo que la tabla denomina Suma de cuadrados, la dispersión, entre Regresión y el Total.

La dispersión dividida por los grados de libertad es lo que se conoce como varianza, aunque el SPSS insiste en denominarla Media Cuadrática. Así, el modelo explica 17 667,97 veces que más, en términos de varianza que lo que queda por explicar o Residual. Este valor es claramente significativo o diferente de 1, que se correspondería con la hipótesis nula; es decir, no existe una relación lineal. El modelo que nos detalla esta relación se obtiene en la tabla a continuación:

Tabla 7.11 Coeficientes del modelo de regresión lineal. IMC = B * Peso + Constante

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	95,0 % intervalo de confianza para B	
		B	Error estándar	Beta			Límite inferior	Límite superior
1	(Constante)	8,657	,086		100,822	,000	8,489	8,826
	Peso (kg)	,254	,002	,848	133,382	,000	,250	,258

a. Variable dependiente: IMC

La interpretación es la siguiente: por cada kilogramo de peso, en media, el IMC aumenta 0,254 unidades, corregido por una constante para que se cumpla que el modelo, que es una línea recta, pase por la media del IMC y la media del peso de todos los datos.

Ejercicio 7.4

Determine la relación entre porcentaje de grasa e IMC, explicitando la bondad del ajuste y el cambio de IMC por unidad porcentual de la proporción de grasa corporal

Explicite la sintaxis utilizada.

Evidentemente, podríamos describir muchos más tipos de análisis, pero, como se ha venido insistiendo a lo largo del libro, esta obra no pretende ser un libro de Estadística y, sin sus conocimientos, no tiene sentido ampliar el número de análisis.

Capítulo ocho

Aplicaciones en estudios de tipo observacional

8.1 Presentación

En el presente capítulo, se muestran algunas aplicaciones de uso frecuente en estudios epidemiológicos y de salud pública que pueden ser de utilidad en el desarrollo de análisis que, con cierta frecuencia, se llevan a cabo en los estudios observacionales, en los que el investigador no interviene en el resultado de ninguna variable respuesta concreta.

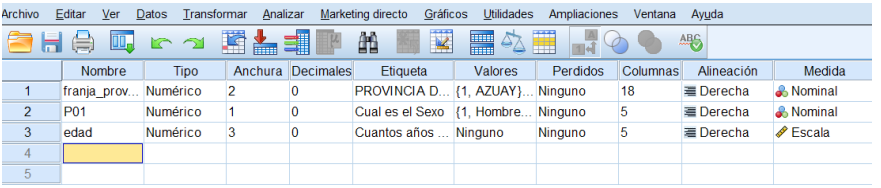
8.2 Pirámide demográfica poblacional

En un estudio, al determinar una propiedad, sea un problema de salud, con frecuencia se desea comparar entre poblaciones diferentes o bien entre subpoblaciones de una misma población. Como el estudiante ya debe conocer, cualquier problema de salud viene determinado, generalmente, por dos características fundamentales, el sexo de los participantes y su edad.

Es pues de obligado conocimiento la estructura poblacional de la población que se estudia y de las subpoblaciones que se quieran comparar. En este apartado, reflexionaremos acerca del significado de la llamada pirámide poblacional y las diferentes formas de obtenerla y representarla, así como generar un archivo que pueda ser utilizado para cálculos posteriores.

Usaremos para ello el archivo generado por el Instituto de Estadística y Censos de Ecuador (INEC), con el censo realizado en el año 2010. El archivo, con gran cantidad de información, se limita a efectos de este ejemplo, a las variables, provincia, sexo y edad de los participantes. El archivo se encuentra en el repositorio que se le ha indicado al inicio del curso y tiene como nombre **CPV2010_2.sav.**

Una vez importado, revise la definición de las variables. Podrá observar que las tres variables son numéricas, si bien en el archivo original del INEC están definidas como cadenas, ya que ya han sido transformadas a numéricas mediante la instrucción AUTORECODE una vez que se importó originalmente el archivo Excel. Así, la provincia es numérica, y se han indicado las etiquetas de dicha codificación, al igual que la variable sexo.



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida
1	franja_prov...	Numérico	2	0	PROVINCIA D...	{1, AZUAY}...	Ninguno	18	Derecha	Nominal
2	P01	Numérico	1	0	Cual es el Sexo	{1, Hombre...	Ninguno	5	Derecha	Nominal
3	edad	Numérico	3	0	Cuantos años ...	Ninguno	Ninguno	5	Derecha	Escala
4										
5										

Figura 8.1 Muestra de la ventana de variables

Una vez comprobada la naturaleza de las variables, que deben ser categóricas las dos primeras y escalar la tercera, ejecute el siguiente paso **Gráficos ► Cuadros de diálogo antiguos ► Pirámide poblacional**, en la barra de herramientas. La ventana que obtendrá es:

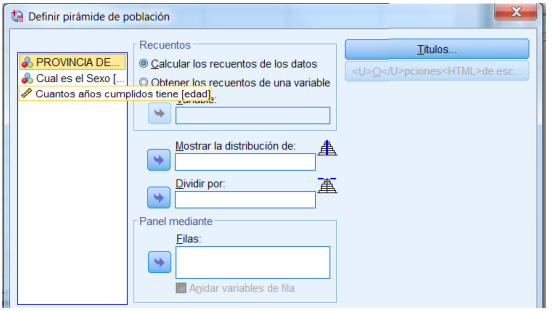


Figura 8.2 Ventana de la obtención de la pirámide poblaciona con datos totales

En la que debe introducir la variable edad en el eje de ordenadas y la variable sexo en el de abcisas. La sintaxis asociada es:

```
DATASET ACTIVATE ConjuntoDatos1.
XGRAPH CHART=[HISTOBAR] BY edad[s] BY P01[c]
  /COORDINATE SPLIT=YES
  /BIN START=AUTO SIZE=AUTO.
```

La cual, al ejecutarla, le da como resultado:

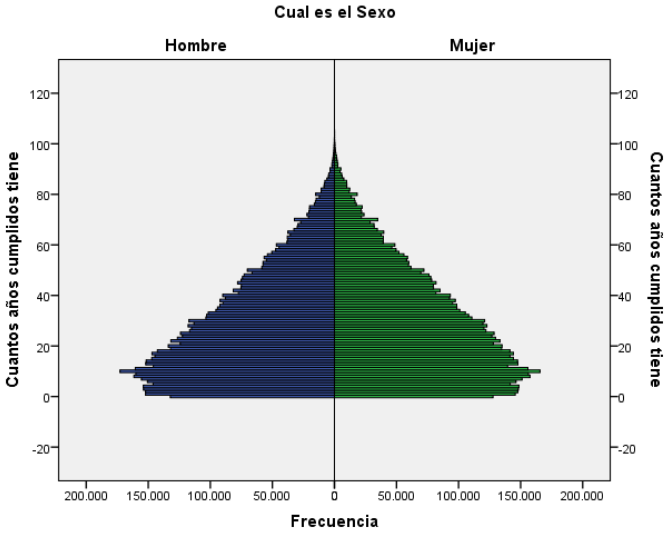


Figura 8.3 Representación de la pirámide con la edad sin agrupar

Una pirámide de este estilo es característica de una población en crecimiento; sin embargo, generalmente no se representa la edad sino franjas de edad, agrupando la edad según algún criterio. Por ejemplo, podemos agrupar en menores de 1 año, de 1 a 4 años, de 5 a 9 años y así sucesivamente.

El sistema de agrupación puede efectuarse de diversas formas, por ejemplo, utilizando la opción **transformar ► recodificar en otra variable**, con el fin de no perder la información original, o bien mediante el siguiente mecanismo de sintaxis:

```
COMPUTE franja_edad=1.
EXECUTE.
IF  (edad >= 1) franja_edad=trunc(edad / 5) + 2.
if (edad ge 90) franja_edad =20.
EXECUTE.
Value labels franja_edad  1 'menores d e1 año', 2 ' 1 a 4  ' , 3 ' 5 a 9', 4' 10 a 14
' , 5 '15 a a 19', 6 ' 20 a 24', 7 ' 25 a 29' , 8 ' 30 a 34', 9 ' 35 a 39', 10 '40 a 44 ', 11
'45 a 49', 12 '50 a 54',
                                13 '55 a 59' ,14 '60 a 64', 15 ' 65 a 69', 16 '70 a 74', 17 '75
a 79',18 '80 a 84',19 '85 a 89', 20 '90 o más'.
EXECUTE.
```

De esta forma, después de definir la franja de edad igual a 1, se cambia por el valor truncado de la edad dividida por cinco, si se cumple la condición de que la edad es igual o superior a 1. De forma que, al resultado, se le sume dos, ya que la franja 1 está ya definida, y de edades de 1 a 4 el resultado del cociente truncado es cero y es la franja 2. En la última franja se han agrupado los mayores de 89 años.

Ahora podemos volver a visualizar la pirámide de edad agrupada por franjas mediante la sintaxis:

```
XGRAPH CHART=([COUNT][BAR]) BY franja_edad[c] BY P01[c]
 /COORDINATE SPLIT=YES.
```

Con el siguiente resultado:

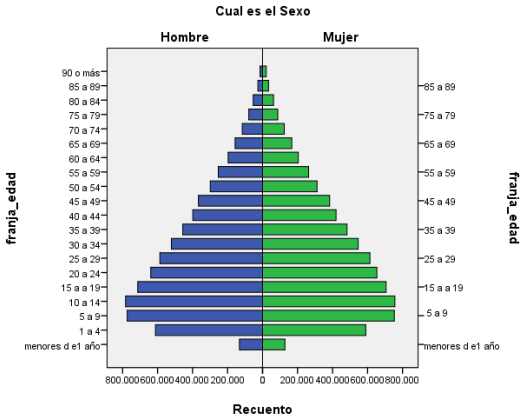


Figura 8.4 Representación de la pirámide poblacional con edades agrupadas

La agrupación de edades, evidentemente, depende del objetivo del estudio y es realmente arbitraria.

La verdadera pirámide poblacional, no obstante no es la figura que la representa sino el conjunto de frecuencias de cada franja. Esto puede obtenerse a través de **Analizar ►Estadísticos descriptivos► Tablas cruzadas,** colocando en filas franja_edad y en columna el sexo.

Tabla 8.1 Pirámide o distribución de frecuencias agrupadas.

Tabla cruzada franja_edad*Cual es el Sexo				
Recuento				
		Cual es el Sexo		**Total**
		Hombre	**Mujer**	
franja_edad	**menores de1 año**	132 183	127 774	259 957
	1 a 4	612 122	590 198	1 202 320
	5 a 9	773 890	752 916	1 526 806
	10 a 14	782 977	756 365	1 539 342
	15 a 19	713 548	705 989	1 419 537
	20 a 24	639 140	652 986	1 292 126
	25 a 29	586 950	613 614	1 200 564
	30 a 34	520 891	546 398	1 067 289
	35 a 39	456 202	482 524	938 726
	40 a 44	399 230	419 772	819 002
	45 a 49	366 448	383 693	750 141
	50 a 54	298 728	311 404	610 132
	55 a 59	253 106	262 787	515 893
	60 a 64	196 414	204 345	400 759
	65 a 69	156 804	167 013	323 817
	70 a 74	116 203	123 888	240 091
	75 a 79	78 602	86 616	165 218
	80 a 84	53 157	62 395	115 552
	85 a 89	26 734	34 001	60 735
	90 o más	14 354	21 138	35 492
Total		7 177 683	7 305 816	14 483 499

Esta distribución de frecuencias es la verdadera pirámide poblacional, la cual también puede tener como representación gráfica la que ofrece el análisis de tablas cruzadas.

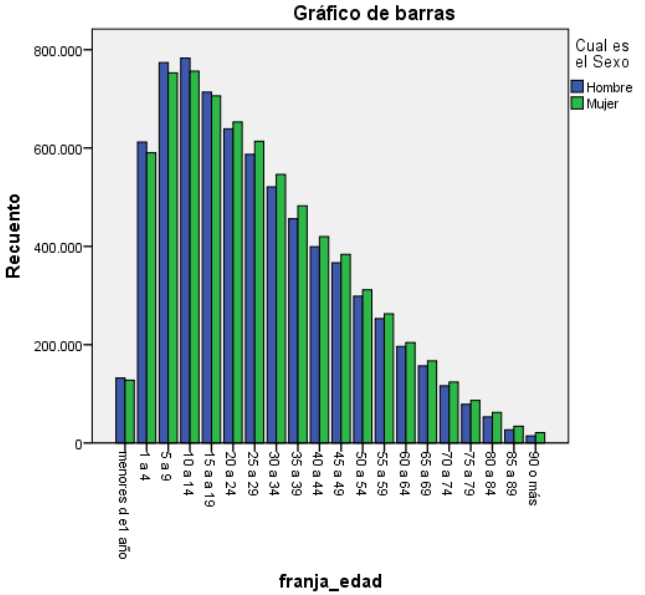


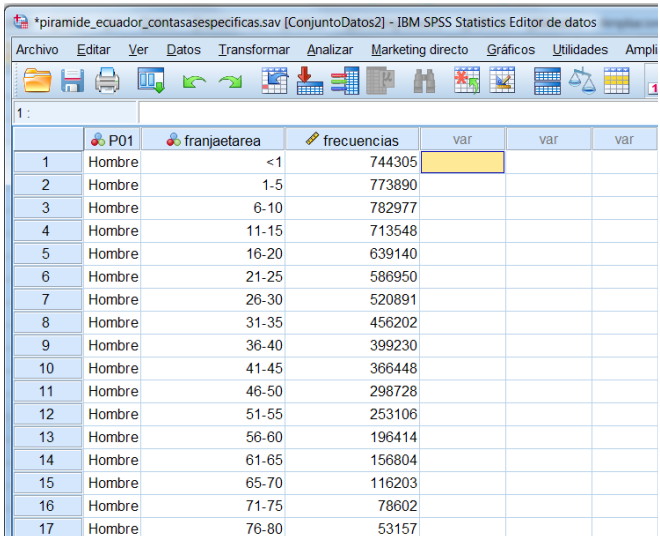
Figura 8.5 Representación de la pirámide de sexo/franja de edad obtenida en una tabla de contingencia.

Es decir, la pirámide no es la figura, sino las frecuencias que constan en las tablas. Con frecuencia, la pirámide poblacional no se utiliza solo como representación de la estructura demográfica sino que la distribución de frecuencias obtenidas se emplea para ciertas operaciones, como calcular y estandarizar tasas, por lo que la figura o la tabla de contingencia no puede utilizarse directamente, sino que debe generarse un nuevo archivo en el que las variables, **sexo**, **franja_etárea** y frecuencia en cada franja pueda ser utilizada para cálculos posteriores.

Para ello generaremos el archivo agregando los datos por sexo y franja de edad y en cada combinación acumularemos el número de casos. La sintaxis en este caso es, si no ha ordenado previamente el archivo por las variables sexo y franja_edad:

```
SORT CASES BY P01 franja_edad.
AGGREGATE
 /OUTFILE='C:\Users\1001278\Desktop\MSP COvid19\taller tasas\piraamideEcuador.sav'
 /PRESORTED
 /BREAK=P01 franja_edad
 /N_BREAK=N.
```

Creándose el resultado en un arcivo de nombre pirámide Ecuador.sav.



	P01	franjaetarea	frecuencias	var	var	var
1	Hombre	<1	744305			
2	Hombre	1-5	773890			
3	Hombre	6-10	782977			
4	Hombre	11-15	713548			
5	Hombre	16-20	639140			
6	Hombre	21-25	586950			
7	Hombre	26-30	520891			
8	Hombre	31-35	456202			
9	Hombre	36-40	399230			
10	Hombre	41-45	366448			
11	Hombre	46-50	298728			
12	Hombre	51-55	253106			
13	Hombre	56-60	196414			
14	Hombre	61-65	156804			
15	Hombre	65-70	116203			
16	Hombre	71-75	78602			
17	Hombre	76-80	53157			

Figura 8.6 Pirámide poblacional en el archivo activo, resultado de la agregación

Ahora que tenemos agregadas las frecuencias en cada franja de edad y por sexo, podemos representar de nuevo la pirámide con la siguiente instrucción:

```
DATASET ACTIVATE ConjuntoDatos2.
XGRAPH CHART=(frecuencias [SUM] [BAR]) BY franjaetarea[c] BY P01[c]
 /COORDINATE SPLIT=YES.
```

Observará que la figura obtenida es la misma que la que se representa en la figura 8.4.

A partir de esta pirámide poblacional, añadiendo los valores de una variable de interés, en una nueva columna, se pueden calcular tasas específicas, tasa cruda y, como se verá en el siguiente ejemplo, estandarizar la tasa para su comparación con la tasa de otra población de diferente estructura poblacional.

8.3 Estudio de proporciones en estudios no dependientes del tiempo: cálculo de tasas y su estandarización

En salud y en otros campos científicos, como la sociología, psicología, educación, nutrición, etc., los estudios descriptivos se llevan a cabo mediante la aplicación de una encuesta en la que el resultado, generalmente, se expresa como una proporción o porcentaje de los encuestados que presentan una característica que es el objetivo del estudio.

Si esta característica es un problema de salud, a la proporción se le da el nombre de prevalencia, si bien, por extensión, se utiliza esa denominación para la variable objeto de estudio. En general, por practicidad de protocolo y también por la disponibilidad de los recursos, se analiza esa variable en el momento en que se encuesta a las personas que componen la muestra y no es objetivo del estudio ver cómo esta propiedad evoluciona en el tiempo.

Esta manera de actuar es característica de los llamados estudios transversales o *cross sectional study* en la literatura científica. Son estudios que, en el ámbito de la salud, se circunscriben preferentemente en el estudio de enfermedades crónicas, y no en el de enfermedades agudas, las cuales se caracterizan por desaparecer en el tiempo con cierta velocidad.

En este caso, la variable que puede interesarnos no sería la prevalencia sino la incidencia acumulada en un cierto intervalo de tiempo. Ambas variables, si bien son diferentes conceptualmente, pueden ser analizadas de la misma forma, ya que la distribución de probabilidad aproximada que las regula es la misma, la distribución binomial, la cual analiza la probabilidad de que esté presente en el momento de hacer el estudio un número determinado de veces. O bien aparezca un fenómeno en un lapsus de tiempo de estudio, que, aunque en puridad la describe mejor una distribución de Poisson, puede mostrarse como una suma de fenómenos de este tipo y aproximarse suficientemente bien por una distribución binomial.

Para ejemplarizar el uso del programa en el cálculo de estas propiedades, prevalencia e incidencia acumulada, utilizaremos el estudio de trabajadores de un hospital a los cuales se efectuará un análisis de las ausencias laborales debidas a enfermedad.

En este estudio participan 2643 trabajadores, los cuales han sido seguidos desde el mes de enero de 2000 al mes de diciembre de 2002. En el archivo **consultas.sav**, se encuentran los datos de las bajas por enfermedad observadas.

Estas bajas laborales por enfermedad están definidas de la siguiente manera: han de tener una duración mayor a los dos días y serán diagnosticadas por los servicios de salud ocupacional del centro, ya sea directamente o aceptando el parte médico que las justificó de forma externa.

Como se aprecia en la tabla 8.2, las bajas observadas han sido generadas en un 79 % por trabajadores de sexo femenino. Sin embargo, de esta descripción, no pueden sacarse conclusiones, ya que se desconoce el número de mujeres que forman la plantilla de trabajadores y tampoco se conoce el número de bajas que se producen en cada trabajador

Tabla 8.2 Distribución de bajas laborales en función del sexo del trabajador

		Sexo			
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	F	10 021	79,0	79,0	79,0
	M	2656	21,0	21,0	100,0
	Total	12 677	100,0	100,0	

Es decir, si el interés está en analizar las bajas por trabajador, la mejor opción es agregar por inscr, que es el número de registro utilizado para mantener la confidencialidad de los trabajadores. Para ello utilizamos la siguiente sintaxis:

```
DATASET ACTIVATE ConjuntoDatos1.
SORT CASES BY inscr(A) f_consul(A).
AGGREGATE
  /OUTFILE='C:\Users\1001278\OneDrive\librosps\por_trabajador.sav'
  /BREAK=inscr
```

```
/bajas_sum=SUM(bajas)
/sexo_first=FIRST(sexo)
/edad5_first=FIRST(edad5)
/hta_sum=SUM(hta)
/asma_sum=SUM(asma)
/N_BREAK=N.
```

Lo efectuado ha sido: acumulamos, agregamos, a cada trabajador, su código identificativo, el sexo, grupo de edad al inicio del estudio, número de consultas en las que se la concedido la baja laboral, y número de episodios de hipertensión arterial y de asma a lo largo de los años analizados. Ahora sí podemos describir la composición de hombres y mujeres.

Tabla 8.3 Trabajadores que han sufrido baja por sexo

		Sexo			
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	F	1942	73,5	73,5	73,5
	M	701	26,5	26,5	100,0
	Total	2643	100,0	100,0	

El número de bajas producido por estos trabadores es de 7855 bajas:

Tabla 8.4 Bajas acumuladas

Estadístico		
bajas_sum		
N	Válido	1958
	Perdidos	685
Asimetría		1,985
Error estándar de asimetria		,055
Mínimo		,00
Máximo		28,00
Suma		7855,00

Con la siguiente distribución:

Tabla 8.5 Descriptiva univariada de la variable Suma de las bajas laborales

Estadísticos		
bajas_sum		
N	**Válido**	1958
	Perdidos	685
Media		4,0117
Mediana		3,0000
Desviación estándar		4,19277
Asimetría		1,985
Error estándar de asimetria		,055
Mínimo		,00
Máximo		28,00

Decida en primer lugar si la ausencia de consulta la considera baja o no, ya que, según la decisión que tome, las prevalencias e incidencias cambiarán. Es realmente opinable, pero, en un estudio, son definiciones previas al análisis.

Tabla 8.6 Descriptiva de la variable suma de bajas en los años de estudio

bajas_sum					
		Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
Válido	**,00**	175	6,6	8,9	8,9
	1,00	464	17,6	23,7	32,6
	2,00	307	11,6	15,7	48,3
	3,00	234	8,9	12,0	60,3
	4,00	193	7,3	9,9	70,1
	5,00	119	4,5	6,1	76,2
	6,00	94	3,6	4,8	81,0
	7,00	61	2,3	3,1	84,1
	8,00	62	2,3	3,2	87,3
	9,00	50	1,9	2,6	89,8

	10,00	40	1,5	2,0	91,9
	11,00	23	,9	1,2	93,1
	12,00	33	1,2	1,7	94,7
	13,00	22	,8	1,1	95,9
	14,00	14	,5	,7	96,6
	15,00	15	,6	,8	97,3
	16,00	11	,4	,6	97,9
	17,00	7	,3	,4	98,3
	18,00	6	,2	,3	98,6
	19,00	4	,2	,2	98,8
	20,00	6	,2	,3	99,1
	21,00	2	,1	,1	99,2
	22,00	6	,2	,3	99,5
	23,00	4	,2	,2	99,7
	24,00	3	,1	,2	99,8
	25,00	2	,1	,1	99,9
	28,00	1	,0	,1	100,0
	Total	1958	74,1	100,0	
Perdidos	**Sistema**	685	25,9		
Total		2643	100,0		

Así, puede verse que el promedio de bajas por trabajador que usó el servicio de salud ocupacional, que es el que debe conceder la situación de baja por enfermedad, es de 3 bajas de mediana en los años estudiados, con un mínimo de 0. Es decir, no logró esa condición, y un máximo de 28. Es pues una variable muy sesgada positiva, como puede observarse en la Figura 8.7.

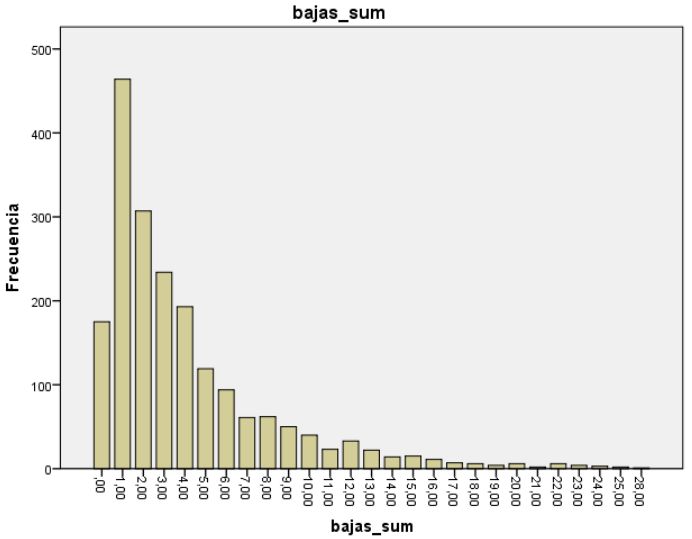


Figura 8.7 Diagrama de barras que representa el número de bajas producidas en el período de estudio

De los 2643 trabajadores, 685 no fueron a consulta médica, por lo que no consta ninguna baja. De los 1958 restantes, en 175, no les fue otorgada la baja. Es decir, 1783 tuvieron alguna baja, una o más. Para poder tener esta información, creamos una nueva variable que sea, tuvo_bajas, sí (1) o no (2). Utilice las herramientas de recodificación ya explicadas anteriormente.

Tabla 8.6 Frecuencia de trabajadores con al menos una baja

		tuvo_bajas			
		Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
Válido	**No**	175	6,6	8,9	8,9
	Sí	1783	67,5	91,1	100,0
	Total	1958	74,1	100,0	

De esta forma podemos incluir como no tuvo bajas a los trabajadores que no utilizaron la consulta del servicio de salud ocupacional, sin arrastrar los valores perdidos de la suma de bajas, ya que, en realidad, no fueron al servicio de salud. Es ahora cuando podemos calcular las tasas de haber sufrido al menos una baja laboral. Tasa cruda global: 67,5 %, Intervalo de confianza, $IC_{95\,\%}$= 65,7 % - 69,3 %.

Tabla 8.7 Tasas específicas por sexo

Tabla cruzada Sexo*tuvo_bajas

			tuvo_bajas		Total
			no	sí	
Sexo	F	Recuento	534	1408	1942
		% dentro de Sexo	27,5	72,5	100,0
	M	Recuento	326	375	701
		% dentro de Sexo	46,5	53,5	100,0
Total		Recuento	860	1783	2643
		% dentro de Sexo	**32,5**	67,5	100,0

La tasa específica para mujeres es 72,5 %, $IC_{95\%}$= 70,5 % - 74,5 %; y, para los hombres, 53,5 %, $IC_{95\%}$= 49,8 % - 57,2 %. Según este resultado, las mujeres sufren más bajas que los hombres, y la diferencia, dado que los límites del intervalo de confianza no se solapan, es estadísticamente significativa.

Comentario:

Antes de afirmar que las mujeres tienen una peor salud que los hombres, deberíamos tener en cuenta muchos más factores, como si la edad es la misma, el tipo de trabajo que realizan en el hospital, el turno de trabajo y otras características.

Para su información, los problemas de salud asociados al embarazo y el parto no están considerados como bajas laborales debidas a enfermedad.

A continuación, analizamos esta tasa por grupos de edad.

Tabla 8.8 Tasas específicas por grupo de edad

Tabla cruzada Edad de cinco en cinco años*tuvo_bajas					
			tuvo_bajas		Total
			no	sí	
Edad de cinco en cinco años	15-19	Recuento	14	**19**	33
		% dentro de Edad de cinco en cinco años	42,4	**57,6**	100,0
	20-24	Recuento	70	**102**	172
		% dentro de Edad de cinco en cinco años	40,7	**59,3**	100,0
	25-29	Recuento	130	**238**	368
		% dentro de Edad de cinco en cinco años	35,3	**64,7**	100,0
	30-34	Recuento	134	**336**	470
		% dentro de Edad de cinco en cinco años	28,5	**71,5**	100,0
	35-39	Recuento	163	**376**	539
		% dentro de Edad de cinco en cinco años	30,2	**69,8**	100,0
	40-44	Recuento	145	**313**	458
		% dentro de Edad de cinco en cinco años	31,7	**68,3**	100,0
	45-49	Recuento	93	**189**	282
		% dentro de Edad de cinco en cinco años	33,0	**67,0**	100,0
	50-54	Recuento	57	**93**	150
		% dentro de Edad de cinco en cinco años	38,0	**62,0**	100,0
	55-59	Recuento	24	**42**	66
		% dentro de Edad de cinco en cinco años	36,4	**63,6**	100,0
	Mayor de 60	Recuento	6	**21**	27
		% dentro de Edad de cinco en cinco años	22,2	**77,8**	100,0
Total		Recuento	836	**1729**	2565
		% dentro de Edad de cinco en cinco años	32,5	**67,5**	100,0

Tal y como era de esperar, las tasas aumentan con la edad. Analicemos la pirámide poblacional. Para ello volvemos a agregar el archivo de los trabajadores, no el de consultas por sexo y edad, y el número de bajas de cada franja sexo edad.

```
SORT CASES BY sexo_first(A) edad5_first(A).
SORT CASES BY sexo_first(A) edad5_first(A).
AGGREGATE
/OUTFILE='C:\Users\1001278\OneDrive\librosps\Agregado_bajas_pirámide.sav'
/BREAK=sexo_first edad5_first
/tuvo_bajas_sum=SUM(tuvo_bajas)
/N_BREAK=N.
```

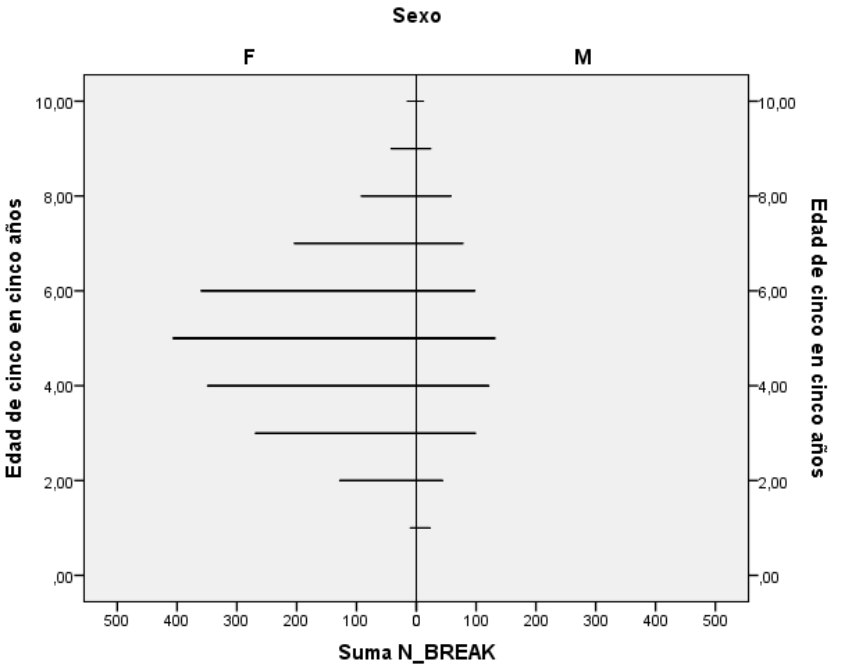
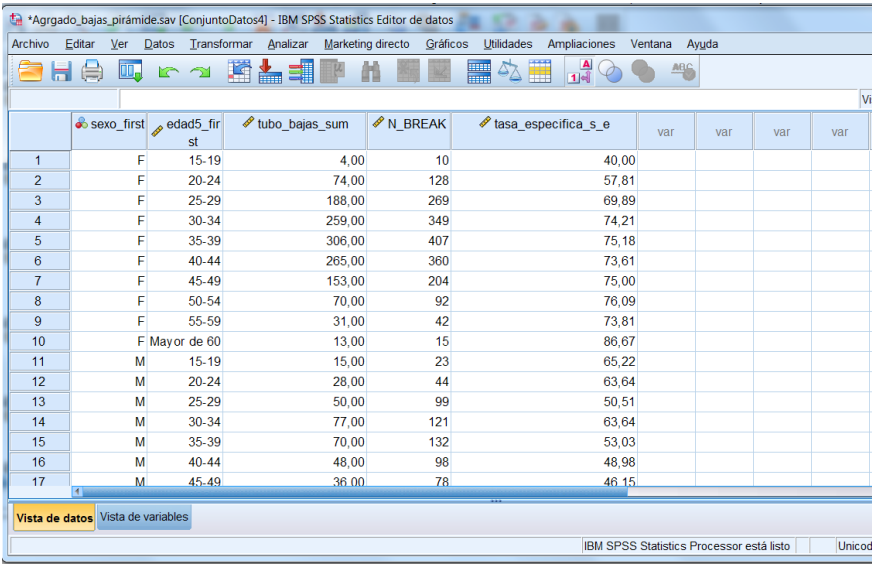


Figura 8.8 Representación gráfica de la pirámide demográfica de los trabajadores

Como ya explicamos en el apartado anterior, la verdadera pirámide es el archivo resultante de la última agregación.



	sexo_first	edad5_first	tubo_bajas_sum	N_BREAK	tasa_especifica_s_e	var	var	var	var
1	F	15-19	4,00	10	40,00				
2	F	20-24	74,00	128	57,81				
3	F	25-29	188,00	269	69,89				
4	F	30-34	259,00	349	74,21				
5	F	35-39	306,00	407	75,18				
6	F	40-44	265,00	360	73,61				
7	F	45-49	153,00	204	75,00				
8	F	50-54	70,00	92	76,09				
9	F	55-59	31,00	42	73,81				
10	F	Mayor de 60	13,00	15	86,67				
11	M	15-19	15,00	23	65,22				
12	M	20-24	28,00	44	63,64				
13	M	25-29	50,00	99	50,51				
14	M	30-34	77,00	121	63,64				
15	M	35-39	70,00	132	53,03				
16	M	40-44	48,00	98	48,98				
17	M	45-49	36,00	78	46,15				

Figura 8.9 Archivo de la pirámide demográfica de los trabajadores con las tasas específicas por sexo y edad (Tuvo_bajas/N_break)*100

Previamente, hemos efectuado este proceso para todos los trabajadores agregando por grupo de edad en el archivo anterior.

```
SORT CASES BY edad5_first(A).
AGGREGATE
  /OUTFILE='C:\Users\1001278\OneDrive\librosps\piramide_global_per_edad.sav'
  /BREAK=edad5_first
  /trabajadores_con_bajas=SUM(tuvo_bajas_sum)
  /N_trabajadores_franja_edad=SUM(N_BREAK).
```

Con lo cual se obtiene la pirámide de edades de todos los trabajadores, mujeres y hombres.

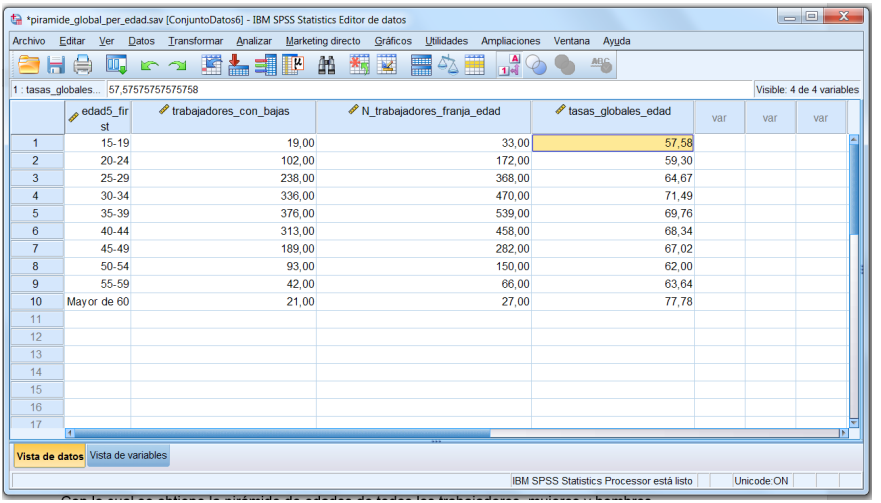


Figura 8.10 Pirámide demográfica de edades de todos los trabajadores indistintamente del sexo

Nota. Se ha incluido el cálculo de la tasa por grupo de edad.

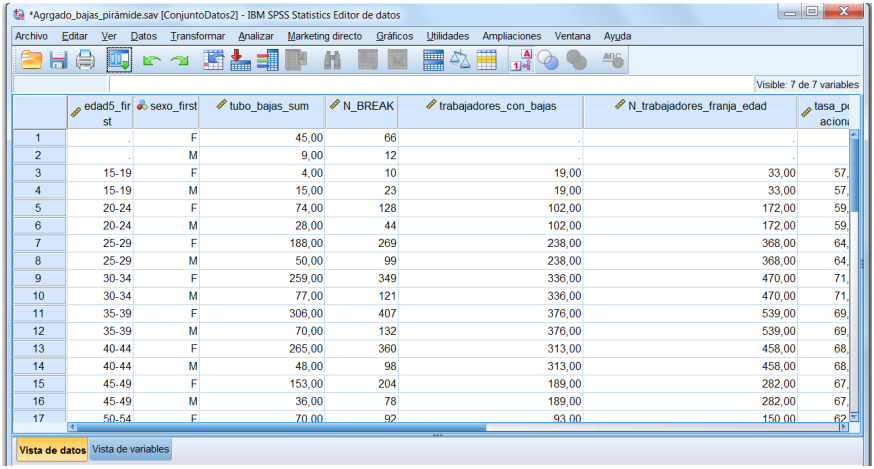
¿Cómo podemos comparar los resultados obtenidos para hombres y mujeres?

Efectuando una estandarización indirecta. Es decir, si hombres y mujeres tuviesen la misma tasa que el conjunto, multiplicando el número de mujeres u hombres de una franja etaria del archivo de la pirámide sexo edad obtenida anteriormente, por la tasa de esa edad para todos los trabajadores obtendríamos los casos que deberían haber tenido baja si el sexo no influyese.

Para ello añadimos la variable tasa específica de edad de todo el conjunto, tasas_globales_edad, al archivo de la pirámide de sexo edad que ya teníamos anteriormente, **Agregado_bajas_pirámide.sav**. Para ello, previamente este archivo lo ordenamos solo por edad y después fusionamos con el archivo **piramide_global_per_edad.sav**.

```
DATASET ACTIVATE ConjuntoDatos2.
SORT CASES BY edad5_first(A).
STAR JOIN
```

```
/SELECT t0.sexo_first, t0.tubo_bajas_sum, t0.N_BREAK, t1.trabajadores_con_
bajas,
  t1.N_trabajadores_franja_edad, t1.tasa_poblacional
/FROM * AS t0
/JOIN 'ConjuntoDatos1' AS t1
 ON t0.edad5_first=t1.edad5_first
/OUTFILE FILE=*.
```



	edad5_first	sexo_first	tubo_bajas_sum	N_BREAK	trabajadores_con_bajas	N_trabajadores_franja_edad	tasa_poblacional
1	.	F	45,00	66	.	.	
2	.	M	9,00	12	.	.	
3	15-19	F	4,00	10	19,00	33,00	57
4	15-19	M	15,00	23	19,00	33,00	57
5	20-24	F	74,00	128	102,00	172,00	59
6	20-24	M	28,00	44	102,00	172,00	59
7	25-29	F	188,00	269	238,00	368,00	64
8	25-29	M	50,00	99	238,00	368,00	64
9	30-34	F	259,00	349	336,00	470,00	71
10	30-34	M	77,00	121	336,00	470,00	71
11	35-39	F	306,00	407	376,00	539,00	69
12	35-39	M	70,00	132	376,00	539,00	69
13	40-44	F	265,00	360	313,00	458,00	68
14	40-44	M	48,00	98	313,00	458,00	68
15	45-49	F	153,00	204	189,00	282,00	67
16	45-49	M	36,00	78	189,00	282,00	67
17	50-54	F	70,00	92	93,00	150,00	62

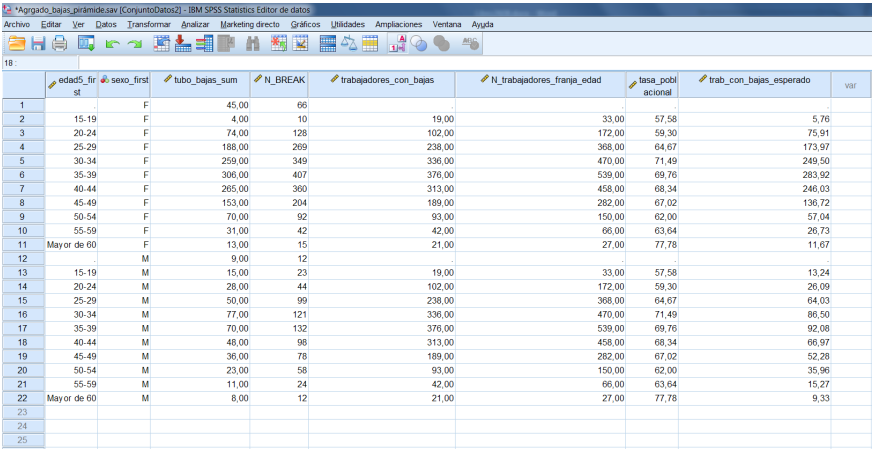
Figura 8.11 Archivo resultante después de añadir las tasas de todos los trabajadores al archivo de la pirámide sexo edad

Puede observarse que hay un cierto número de casos en los que, al no constar la edad, no se pudieron agregar previamente y, por esa razón, no poseemos sus bajas reales. El siguiente paso es calcular en cada franja de edad, para mujeres y para hombres, el número de bajas esperado si la incidencia de bajas fuese indistinta en ambos sexos.

Para ello calculamos el número de casos que deberían haber aparecido si la población fuese homogénea.

```
SORT CASES BY sexo_first(A) edad5_first(A).
COMPUTE trab_con_bajas_esperado=(N_BREAK*tasa_poblacional)/100.
EXECUTE.
```

Si observa ahora el archivo obtenido, podrá ver, en la última columna, cuántas personas deberían haber sufrido bajas laborales por enfermedad en mujeres y hombres si el problema les afectase por igual. La hipótesis es, pues, que la tasa de aparición de baja en la franja de edad 15-19 en mujeres, debería ser la misma en hombres e igual a la que se calculó para el conjunto global de los trabajadores.


*Agrgado_bajas_pirámide.sav [ConjuntoDatos2] - IBM SPSS Statistics Editor de datos

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

18 :

	edad5_first	sexo_first	tubo_bajas_sum	N_BREAK	trabajadores_con_bajas	N_trabajadores_franja_edad	tasa_poblacional	trab_con_bajas_esperado	var
1	.	F	45,00	66	.	.	.	.	
2	15-19	F	4,00	10	19,00	33,00	57,58	5,76	
3	20-24	F	74,00	128	102,00	172,00	59,30	75,91	
4	25-29	F	188,00	269	238,00	368,00	64,67	173,97	
5	30-34	F	259,00	349	336,00	470,00	71,49	249,50	
6	35-39	F	306,00	407	376,00	539,00	69,76	283,92	
7	40-44	F	265,00	360	313,00	458,00	68,34	246,03	
8	45-49	F	153,00	204	189,00	282,00	67,02	136,72	
9	50-54	F	70,00	92	93,00	150,00	62,00	57,04	
10	55-59	F	31,00	42	42,00	66,00	63,64	26,73	
11	Mayor de 60	F	13,00	15	21,00	27,00	77,78	11,67	
12	.	M	9,00	12	.	.	.	.	
13	15-19	M	15,00	23	19,00	33,00	57,58	13,24	
14	20-24	M	28,00	44	102,00	172,00	59,30	26,09	
15	25-29	M	50,00	99	238,00	368,00	64,67	64,03	
16	30-34	M	77,00	121	336,00	470,00	71,49	86,50	
17	35-39	M	70,00	132	376,00	539,00	69,76	92,08	
18	40-44	M	48,00	98	313,00	458,00	68,34	66,97	
19	45-49	M	36,00	78	189,00	282,00	67,02	52,28	
20	50-54	M	23,00	58	93,00	150,00	62,00	35,96	
21	55-59	M	11,00	24	42,00	66,00	63,64	15,27	
22	Mayor de 60	M	8,00	12	21,00	27,00	77,78	9,33	
23									
24									
25									

Figura 8.12 Archivo una vez calculadas las personas con baja esperadas

En este momento podemos calcular la tasa estandarizada de forma indirecta para hombres y mujeres, definiéndolas como la suma, en cada sexo, de personas afectadas por bajas esperadas dividido para el número de trabajadores de su sexo.

```
SORT CASES  BY sexo_first.
SPLIT FILE LAYERED BY sexo_first.
FREQUENCIES VARIABLES=tuvo_bajas_sum trab_con_bajas_esperado N_
BREAK
/FORMAT=NOTABLE
/STATISTICS=SUM
/ORDER=ANALYSIS.
```

Lo cual da como resultado:

Tabla 8.9 Suma de casos de bajas observades y esperadas

Estadísticos					
Sexo			tuvo_bajas_sum	trab_con_bajas_esperado	N_BREAK
F	N	Válido	11	10	11
		Perdidos	0	1	0
	Suma		1408,00	1267,24	1942
M	N	Válido	11	10	11
		Perdidos	0	1	0
	Suma		375,00	461,76	701

Es decir, en un total de 1942 mujeres, la suma de ellas con bajas registradas es de 1408. La tasa cruda es por lo tanto (1408/1942)*100 = 72,5 % y se habrían esperado en la hipótesis de homogeneidad entre hombre y mujeres (1267/1942)*100 = 65,2 %.

Una forma de expresar el resultado es el cociente entre la tasa observda y la tasa estandarizada o esperada. A este cociente se le denomina RME = 72,5 / 65,2 = 1,11 o en otras palabras, la aparición de bajas en las mujeres es un 11 % superior que lo esperado si no hubiesen diferencias entre ambos sexos. Se analiza restando al RME la unidad y multiplicando por 100.

En los hombres, sin embargo, la tasa observada es de (375/701)*100 = 53,5 %, mientras que la estandarizada o esperada sería de 65,8 %. Es decir, la RME = 0,81 nos indica que los hombres presentan procesos de baja laboral por enfermedad en un 19 % menos de lo esperado.

8.4 Análisis de proporciones en estudios dependientes del tiempo: cálculo de tasas de densidad de incidencia

En los estudios en el que el tiempo de aparición de un fenómeno, incidencia, nos interese, se busca conocer la velocidad de aparición o densidad de incidencia. En este ejemplo valoraremos la densidad de incidencia de las lesiones osteomusculares, limitándonos a la aparición del primer evento desde el inicio de la observación.

Evidentemente se trata de un ejemplo teórico, ya que no se pretende analizar los factores que influyen en la presencia del problema de salud en toda su dimensión. Nos limitaremos al cálculo de esta densidad de incidencia en función del sexo del trabajador y el tipo de trabajo que realiza en el hospital en donde se llevó a cabo el estudio.

Los datos iniciales se encuentran en el archivo **consultas.sav.** El proceso de cálculo de las tasas de densidad de incidencia se efectúa con los siguientes pasos.

1. Crear una variable que nos indique si ha sido diagnosticado (1 = sí; 0 = no) con este problema.
2. Crear un archivo agregado de los trabajadores en los que conste si existe diagnóstico de lesión osteomuscular y, en concreto, el primer diagnóstico de la misma y la fecha de su diagnóstico.
3. Fijar la fecha del primer diagnóstico. En el caso en que no haya sido diagnosticado esta fecha es igual al a del fin del seguimiento del estudio.

También debe constar el sexo, la edad al inicio del estudio y el grupo de actividad laboral que lleva a cabo en la institución (Grup_4). Una vez obtenido este archivo, se calcula el tiempo transcurrido desde la fecha de inicio del estudio hasta le fecha de este primer diagnóstico, que, si no ha ocurrido, es la de final del estudio, tal y como se ha indicado.

¿Qué es la densidad de incidencia? Es una medida de la velocidad de aparición del suceso, por lo que se calcula como la suma de casos diagnosticados dividido por la suma de tiempos trascurridos hasta su aparicición, propiedad que se conoce técnicamente como **personas tiempo,** por una deficiente traducción del inglés, ya que lo correcto sería tiempo de observación de las personas.

El cociente es la densidad de incidencia, propiedad que depende de las unidades en que ha sido calculado el tiempo. En general este tiempo se calcula en días, pero, posteriormente, por 365 días, nos ofrece la densidad de incidencia por trabajador y por año. O multiplicando por 30, sería la densidad de incidencia por trabajador y mes de trabajo.

Evidentemente puede expresarse por 100 trabajadores, multiplicando por 100 o por la potencia de 10 que sea más manejable. La precisión con que se

conocen los tiempos, las fechas, influye en el cálculo. Así, si solo, conociésemos el año, la precisión es menor que si la fecha se expresase como meses y año.

Por otro lado, aunque se tenga en cuanta otra característica, la densidad de incidencia calculada de esta forma está infravalorada, ya que se está suponiendo que el trabajador está constantemente trabajando y expuesto el riesgo de lesión osteomuscular los 365 días del año.

Obviamente esto no es real, ya que, como mínimo, los trabajadores realizan semanas de 40 horas y no de 7*24 horas, y además disfrutan de permisos, vacaciones, tiempos de ausencia por otras enfermedades, etc. Es decir, cuanto más se perfile el tiempo en que está trabajando en una exposición, más alta será la densidad de incidencia acercándose mejor al verdadero valor o riesgo de padecer la enfermedad.

La sintaxis de ejecución del ejercicio es la siguiente:

```
GET
  FILE='C:\Users\1001278\OneDrive\librosps\consultas.sav'.
DATASET NAME ConjuntoDatos3 WINDOW=FRONT.
DATASET ACTIVATE ConjuntoDatos2.
COMPUTE diag_osteomuscular=0.
EXECUTE.
IF  (gra_grup = 13) diag_osteomuscular=1.
EXECUTE.
Value labels diag_osteomuscular 1'si' 0 'n0'.
Execute.
```

Tabla 8.10 Proporción de consultes con baja laboral por diagnósticos osteomusculares

diag_osteomuscular					
		Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
Válido	**no**	10 913	86,1	86,1	86,1
	sí	1764	13,9	13,9	100,0
	Total	12 677	100,0	100,0	

Se han efectuado 1764 consultas en las que se ha diagnosticado una lesión osteomuscular; sin embargo, este diagnóstico no es el del número de trabajadores sino el de consultas, ya que dichos trabajadores pueden presentar múltiples diagnósticos.

Por esta razón, debemos crear un archivo de trabajadores que hayan sido diagnosticados sin importarnos cuántas veces, considerando la fecha en que fue diagnosticado por primera vez. Para ello, lo primero que efectuamos es una ordenación de casos de forma que el primer registro sea el ser diagnosticado en alguna consulta, en caso de que así fuese.

```
SORT CASES BY inscr(A) diag_osteomuscular(D) f_consul(A).
```

Observe que el orden por diag_osteomuscular es descendente. Con este proceder, los registros que tengan un diagnóstico de lesión osteomuscular serán los primeros de aquellos trabajadores que, poseyendo varios diagnósticos, puedan tener más de uno osteomuscular y al haber ordenado por fechas, el primer registro corresponderá a la fecha en que fue diagnosticado por primera vez.

Posteriormente, a aquellos trabajadores sin este diagnóstico, diag_osteomuscular = 0, deberemos atribuirles la fecha diagnóstica igual a 31 de diciembre de 2002, fin del seguimiento. Una vez convenientemente ordenado, agregamos el archivo por trabajador, al cual denominaremos trabajadores con la abreviatura LER que corresponde a Lesión por esfuerzo repetitivo.

```
AGGREGATE
  /OUTFILE='C:\Users\1001278\OneDrive\librosps\trabajadores con ler.sav'
  /BREAK=inscr
  /diag_osteomuscular_first 'tiene lesion ostreomuscula'=FIRST(diag_osteo-
  muscular)
  /f_consul_first 'fecha diagnostico osteomuscular'=FIRST(f_consul)
  /sexo_first 'sexo'=FIRST(sexo)
  /eda_an_first 'edad'=FIRST(eda_an)
  /cargo4_first 'ocupación'=FIRST(cargo4)
  /fecha3_first 'fecha fin seguimiento'=FIRST(fecha3)
  /numero_de_consultas=N.
```

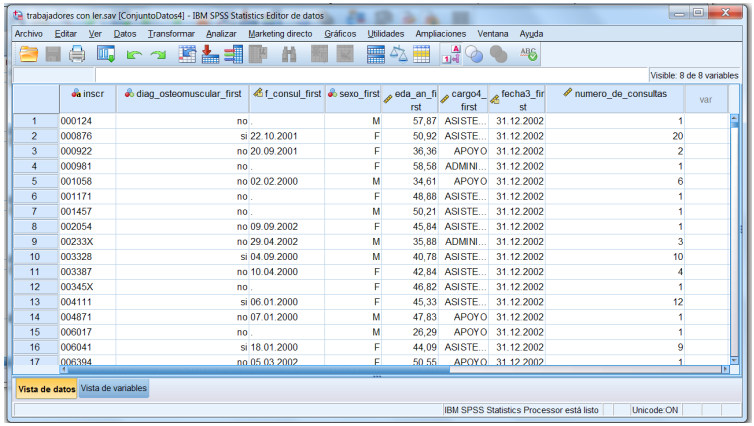


Figura 8.13 Visión parcial del archivo agregado en la anterior instrucción

Como se observa, la edad tiene decimales debido a que, en su inicio, se calculó a partir de la edad de inicio del estudio y la fecha de nacimiento, por lo que la transformaremos a edad civil, es decir, años cumplidos mediante la instrucción:

```
DATASET ACTIVATE ConjuntoDatos4.
COMPUTE edad_civil=trunc(eda_an_first).
EXECUTE.
```

Función que redondea los datos de edad con decimales al entero anterior.

En este momento debemos incluir la fecha de final de estudio en la fecha diagnóstica en aquellos casos no diagnosticados por la enfermedad. Para ello llevamos a cabo la siguiente transformación:

```
IF  (diag_osteomuscular_first = 0) f_consul_first=fecha3_first.
EXECUTE.
```

Y creamos la fecha de inicio de la observación, 01 enero 2000, para todos los casos.

```
COMPUTE fecha_inicio=DATE.DMY(01,01,2000).
EXECUTE.
```

Si observa en la ventana de datos, verá que dicha fecha es una expresión numérica a la que debe cambiar el formato en la ventana de variables, en la columna TIPO, por un formato fecha, dd.mm.aaaa.

En este momento ya podemos determinar la propiedad personas tiempo, la cual, como se ha dicho anteriormente, es la suma de tiempos de observación hasta que se produce el suceso de estudio. Es decir, el primer diagnóstico de lesión osteomuscular más todo el tiempo de seguimiento de aquellos trabajadores que no han sido diagnosticados. El tiempo transcurrido desde el inicio al suceso final —primer diagnóstico o no diagnóstico— se determina mediante la transformación.

```
COMPUTEpersonas_tiempo=CTIME.DAYS(f_consul_first)-CTIME.DAYS(fe-
cha_inicio).
EXECUTE.
```

Expresión que nos indica el tiempo de cada trabajador en días.

Ahora ya podemos calcular la densidad de incidencia como número de casos observados dividido por la suma de los tiempos de observación. Para ello ejecutamos la siguiente instrucción:

```
FREQUENCIES VARIABLES=diag_osteomuscular_first  personas_tiempo
/STATISTICS=SUM
/ORDER=ANALYSIS.
```

Que da como resultado:

Tabla 8.11 Número acumulado de trabajadores con la primer baja laboral por diagnósticos osteomusculares

Estadísticos			
		tiene lesión ostreomuscular	personas_tiempo
N	Válido	2643	2643
	Perdidos	0	0
Suma		723,00	2 444 166,00

Por lo que, al dividir ambas sumas, el resultado es la densidad de incidencia de lesiones (primera lesión) de 2,958e-4 lesiones por trabajador

y día. Es más claro si lo expresamos en un lapsus de tiempo de un año y, en ese caso, al multiplicar por 365, la densidad de incidencia es 0,1079 lesiones por trabajador al año.

Si lo expresamos cada 100 trabajadores, el resultado es 10,79 lesiones cada 100 trabajadores por año. A partir de ahora, el estudio podría continuar, calculando dicha tasa para hombres y mujeres. El procedimiento es sencillo, ya que solo precisa de segmentar el archivo por sexo y repetir el cálculo de la suma:

```
SORT CASES  BY sexo_first.
SPLIT FILE LAYERED BY sexo_first.
FREQUENCIES VARIABLES=diag_osteomuscular_first personas_tiempo
  /STATISTICS=SUM
  /ORDER=ANALYSIS.
```

Con el resultado siguiente y al efectuar el cociente de las sumas:

Tabla 8.12 Número acumulado de trabajadores con la primer baja laboral por diagnósticos osteomusculares, por sexo

			Estadísticos	
Sexo			tiene lesión ostreomuscular	personas_tiempo
F	N	Válido	1942	1942
		Perdidos	0	0
	Suma		577,00	1 771 453,00
M	N	Válido	701	701
		Perdidos	0	0
	Suma		146,00	672 713,00

La tasa de densidad de incidencia para el sexo femenino es 11,89 diagnosticadas por 100 trabajadoras por año, mientras que, en los de sexo masculino, es 7,91 bajas por 100 trabajadores por año.

Como puede comprender, en un trabajo de investigación, se buscarían factores que expliquen estas diferencias. La primera y más evidente es el tipo de actividad en el hospital, para lo cual analizamos el tipo de trabajo mediante

```
SPLIT FILE OFF.
SORT CASES  BY cargo4_first.
SPLIT FILE LAYERED BY cargo4_first.
FREQUENCIES VARIABLES=diag_osteomuscular_first personas_tiempo
  /FORMAT=NOTABLE
  /STATISTICS=SUM
  /ORDER=ANALYSIS.
```

Tabla 8.13 Número acumulado de trabajadores con la primera baja laboral por diagnósticos osteomusculares por tipo de trabajo

Ocupación			tiene lesión ostreomuscular	personas_tiempo
ASISTENCIAL	N	Válido	1450	1450
		Perdidos	0	0
		Suma	367,00	1 360 744,00
ADMINISTRATIVO	N	Válido	481	481
		Perdidos	0	0
		Suma	129,00	448 151,00
APOYO	N	Válido	712	712
		Perdidos	0	0
		Suma	227,00	635 271,00

En el trabajo asistencial, la tasa es 9,84 diagnósticos por 100 trabajadores al año sin distinción del sexo; 10,5, en los trabajadores administrativos; y 13,04, en los trabajadores de apoyo (trabajadores que descargan en almacén, cocina, celadores, etc.).

Este resultado incita a buscar otros factores como, dentro del trabajo asistencial, el tipo de trabajo: medicina, enfermería, auxiliar de enfermería y demás.

Comentario:

A estas alturas del curso, el alumno ya podrá comprender que un ejemplo académico, como el que hemos visto, requiere de mucha más información y hacerse preguntas como: ¿Influye la edad? ¿Influye la historia previa de presencia de lesiones osteomusculares? ¿Es importante la antigüedad en el trabajo? ¿El estado civil? ¿El turno horario?

¿Son comparables las tasas si la distribución de edades difiere en cada tipo de trabajo y la composición de sexos en los trabajadores?

¿Sería necesario estandarizar tal y como se efectuó en el apartado anterior?

Son preguntas que reflejan la complejidad de analizar unos resultados de un proyecto de investigación, pero en estos momentos el alumno ya debe estar capacitado no solo para contestarlas, sino también para resolverlas.

Capítulo nueve
Ejercicio de autoevaluación

9.1 Introducción

Este capítulo tiene como objetivo principal que el lector pueda comprobar hasta qué punto ha adquirido los conocimientos de gestión de datos y análisis de los mismos. Además de detectar aquellos puntos en los que ha logrado habilidades suficientes para el manejo del programa SPSS, así como los puntos del libro que debe revisar. Para ello, se presenta un único ejercicio que se debe efectuar desde el principio, partiendo de la base de datos que se corresponde a un análisis y reproducir todos los pasos necesarios que le permitan generar los resultados, ya publicados.

En este ejercicio, el estudiante debe llevar a cabo los procesos sin instrucciones intermedias y solo conociendo los objetivos del estudio que se le presenta y la información recogida en el mismo —disponible en las bases de datos que acompañan el ejercicio— para poder mostrar su capacidad de llevar a cabo un análisis de los resultados de un proyecto.

La base de datos ha sido simplificada de forma que no se deduzca de la misma el realizar análisis que van más allá de las técnicas aprendidas en este libro. El resultado del ejercicio debe llevarse a cabo mediante la redacción de un informe que dé repuestas las preguntas que se le indican en el texto del ejercicio, incluyendo la sintaxis que ha utilizado, en un archivo Word.

9.2 El estudio

En noviembre del año 2011, se declaró un brote de rabia selvática en las comunidades indígenas que residen en la frontera amazónica de Ecuador con Perú, en la provincia de Morona Santiago. El brote produjo una mortalidad del 15 % de los niños de estas comunidades.

Las características del estudio, objetivos y resultados se muestran en el siguiente artículo de acceso libre:

Artículo 1

Romero-Sandoval, N.; Parra, C.; Gallegos, G.; Guanopatín, A.; Campaña, M. F.; Haro, M.; Calapaqui, S.; Moreta, C.; Viteri, F.; Feijoo-Cid, M.; & Martin, M. (2013). "Haematophagous bat bites in Ecuadorian Amazon: characterisation and implications for sylvatic rabies prevention". *Public Health Action*, 3(1); 85-89(5). https://doi.org/10.5588/pha.12.0070.

Adicionalmente se efectuó otro estudio acerca de las características asociadas a las mordeduras de los murciélagos hematófagos en la misma zona amazónica, en comunidades pertenecientes a las provincias limítrofes de Morona Santiago y Pastaza, en 3518 habitantes, cuyo planteamiento y resultado se encuentra en el artículo dos.

Artículo 2

Romero-Sandoval, N., Escobar, N., Utzet, M., Feijoo-Cid, M., & Martin, M. (2014). "Sylvatic rabies and the perception of vampire bat activity in communities in the Ecuadorian Amazon". *Cadernos de saude publica*, 30, 669-674. https://doi.org/10.1590/0102-311X00070413

9.3 El ejercicio

Los datos se encuentran en los archivos que se incluyen en el **Repositorio bases_grupo2 (Capítulo 1).** El archivo correspondiente al primer artículo indicado se denomina **vivienda_articulo_1.sav** mientras que las encuestas efectuadas y cuyo análisis se publicó en el segundo artículo se denominan **morona.sav** y **pastaza.sav**.

Ejercicio 9.1

- Lea los dos artículos y describa los objetivos de cada uno de ellos.
- Analice la información de las viviendas encuestadas, realizando, en primer lugar, el control de calidad de las mismas, valorando la pérdida de información, y resuelva los errores de tecleo del nombre de la comunidad mediante la acción de recodificación automática y recodificar.
- Reproduzca el análisis descriptivo que se encuentra en la tabla 1 del primer artículo y explore las posibles relaciones binarias entre las características de las viviendas y la percepción del aumento de mordeduras tanto en animales como en los habitantes.
- Genere el archivo que contenga tanto a los habitantes de Morona como de Pastaza.
- Reproduzca los resultados que aparecen en las tres primeras columnas de la tabla 2 del artículo 2. Valore la fuerza de asociación entre las variables que describe dicha tabla.
- Compare los resultados del punto 3 y punto 4 y concluya en su análisis en el informe.
- El informe global no puede superar las cinco páginas en Word, el espaciado que se debe utilizar es de 1,15 y letra Arial 10, incluyendo las tablas o figuras que usted crea necesarias.

Capítulo diez

Resultados de los ejercicios

10.1 Ejercicios del Capítulo 2

Ejercicio 2.1

Siga los pasos que se han explicado en este capítulo con el archivo **estudiantes_zonas_1y2 .sav**.

Especialmente le recomendamos que adquiera habilidad en el uso de la opción **Pegar** y su activación. Guarde los archivos generados y observe en qué carpeta se han grabado.

Repita el ejercicio hasta que logre que los tres archivos se encuentren en el mismo directorio.

Como ejemplo, una de las acciones que debe realizar es obtener una distribución de frecuencias de la variable edad, para ello se coloca en la viñeta analizar, se escoge estadísticos descriptivos y frecuencias.

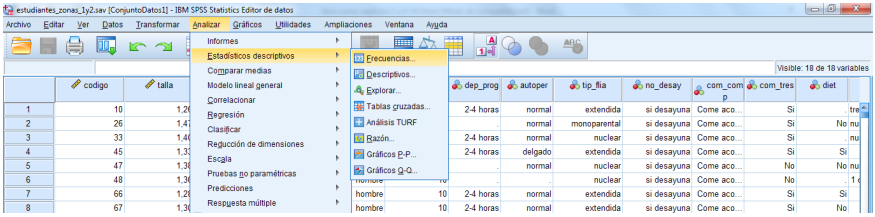


Figura 10.1

La pantalla que se despliega es la siguiente, en donde colocamos la variable requerida mediante la acción Pegar.

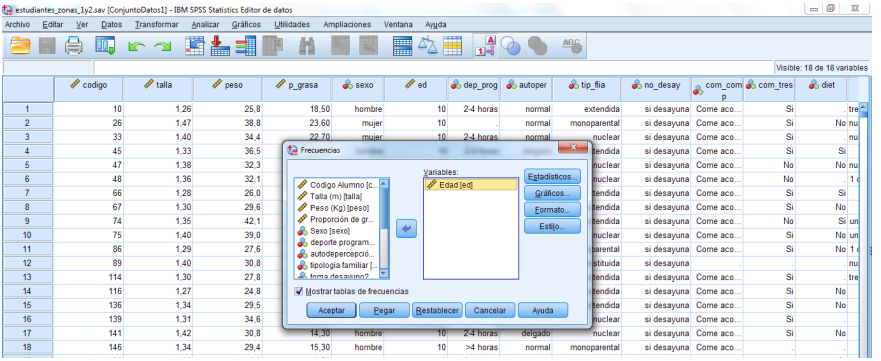


Figura 10.2

Se despliega la siguiente pantalla con la siguiente sintaxis:

```
DATASET ACTIVATE ConjuntoDatos1.
FREQUENCIES VARIABLES=ed
  /ORDER=ANALYSIS.
```

Donde hay que presionar el símbolo ▶ de la barra de herramientas para obtener la pantalla con los resultados que se verifican en la siguiente tabla:

Tabla 10.1 Edad

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	9	45	1,2	1,2	1,2
	10	571	15,0	15,0	16,2
	11	584	15,3	15,3	31,5
	12	862	22,6	22,6	54,1
	13	912	23,9	23,9	78,1
	14	744	19,5	19,5	97,6
	15	71	1,9	1,9	99,4
	16	21	,6	,6	100,0
	Total	3810	100,0	100,0	

Si se repite el ejercicio con tres acciones diferentes, se obtendrá, por ejemplo, una sintaxis como la siguiente:

```
FREQUENCIES VARIABLES=ed
  /ORDER=ANALYSIS.
FREQUENCIES VARIABLES=sexo
  /ORDER=ANALYSIS.
DESCRIPTIVES VARIABLES=talla
  /STATISTICS=MEAN STDDEV MIN MAX.
```

10.2 Ejercicios del Capítulo 3

Ejercicio 3.1

Usando las opciones de visualización del contenido de un fichero de datos, ejecute esas acciones para conocer el contenido de la información que contiene el archivo **estudiantes_zonas_3y4.sav**.

Logre la sintaxis que se corresponde a esas acciones.

Al realizar lo solicitado, **Archivo ► mostrar información del archivo de datos ► archivo externo ►, en el archivo estudiantes_zonas_3y4.sav.**, se obtendrá la siguiente pantalla:

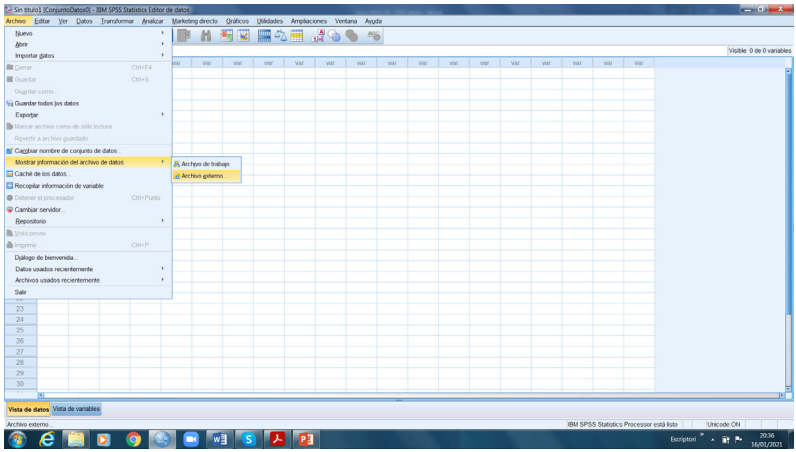


Figura 10.3

Y una vez analizada la información, con Archivos ►**guardar como**►**Guardar como tipo** ►**Excel**.

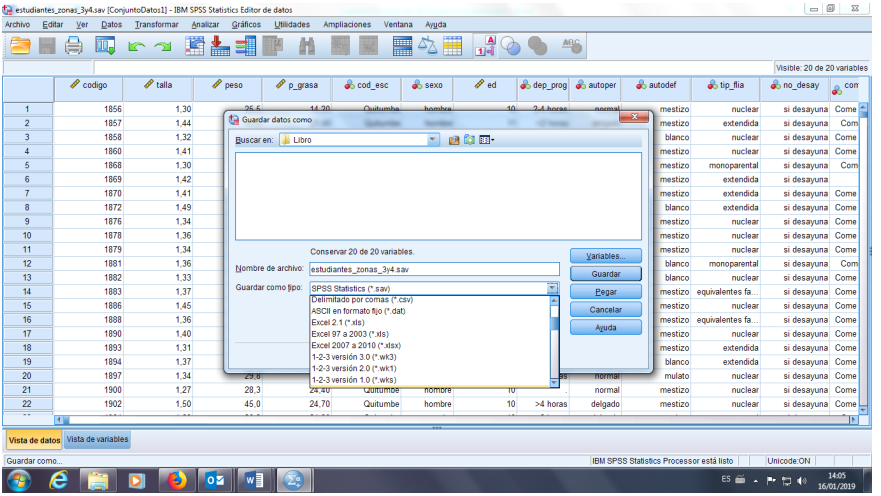


Figura 10.4

Luego ir a Pegar; la sintaxis correspondiente es:

```
SAVE TRANSLATE OUTFILE='C:\Users\diramosfl\Desktop\Investigacion\Libro\
estudiantes_zonas_3y4.xlsx'
  /TYPE=XLS
  /VERSION=12
  /MAP
  /FIELDNAMES VALUE=NAMES
  /CELLS=VALUES.
```

Al Ejecutar, se crea el archivo en Excel.

Ejercicio 3.2

Compare la información que ha obtenido al abrir con SPSS ese archivo de datos en Excel y la que contenía el original, abriendo con Excel el mismo archivo.

Al abrir el archivo de Excel guardado en el ejercicio anterior obtendrá:

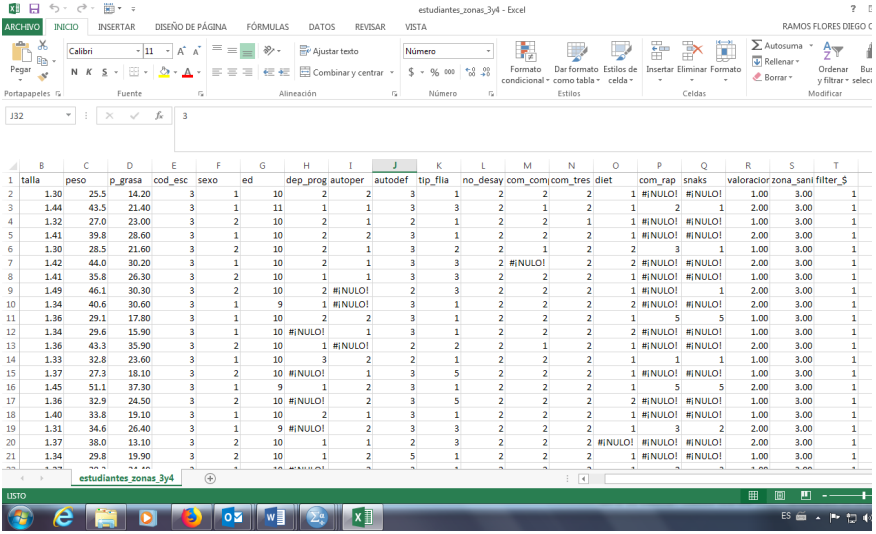


Figura 10.5

Si realizamos por la segunda opción, en la pantalla de sintaxis se obtiene:

```
GET DATA
  /TYPE=XLSX
  /FILE='C:\Users\diramosfl\Desktop\Investigacion\Libro\estudiantes_zo-
  nas_3y4.xlsx'
  /SHEET=name 'estudiantes_zonas_3y4'
  /CELLRANGE=FULL
  /READNAMES=ON
  /DATATYPEMIN PERCENTAGE=95.0
  /HIDDEN IGNORE=YES.
EXECUTE.
DATASET NAME ConjuntoDatos3 WINDOW=FRONT.
```

Al ejecutarla la base de datos en el SPSS obtenida es:

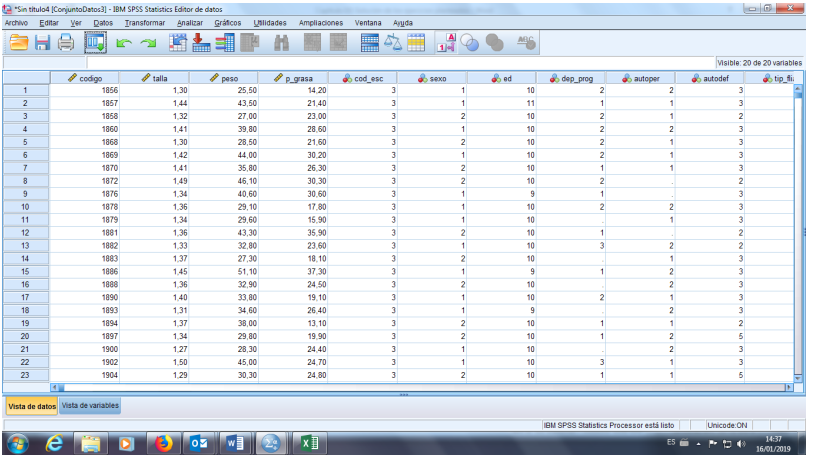


Figura 10.6

Ejercicio 3.3

Sume los casos correspondientes a la zona 3 y 4 a los de las zonas 1 y 2.

Previamente asegúrese, mediante la acción necesaria, de que ambos tienen la misma información y sepa de antemano qué variables estarán desaparejadas.

Indique en una nueva variable el origen de los datos y a continuación guarde el resultado en un archivo con el nombre **estudiantes_total.sav**.

Visualice la sintaxis utilizando la opción **Pegar** y posteriormente ejecútela.

Una vez realizada estas acciones, cerciórese mediante:

SYSFILE INFO 'C:\Users\usuario\Desktop\Libro SPSS Ecuador diciembre\bases de datos\estudiantes_total.sav'.

O por ventanas del contenido del resultado de sumar los casos.

Al realizar la operación pedida la ventana de sintaxis es:

```
DATASET ACTIVATE ConjuntoDatos1.
ADD FILES /FILE=*
  /FILE='ConjuntoDatos2'
  /RENAME (autodef cod_esc=d0 d1)
  /DROP=d0 d1.
EXECUTE.
```

Y la ventana de resultados es:

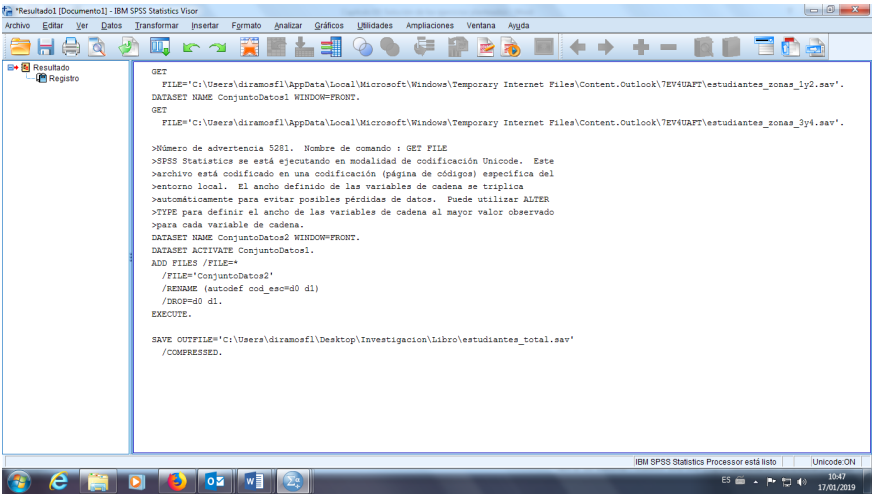


Figura 10.7 Ventana del SPSS para la sintaxis “sumar los casos”

Ejercicio 3.4

Añada las variables que informan los padres a cada estudiante y pegue la sintaxis correspondiente al proceso.

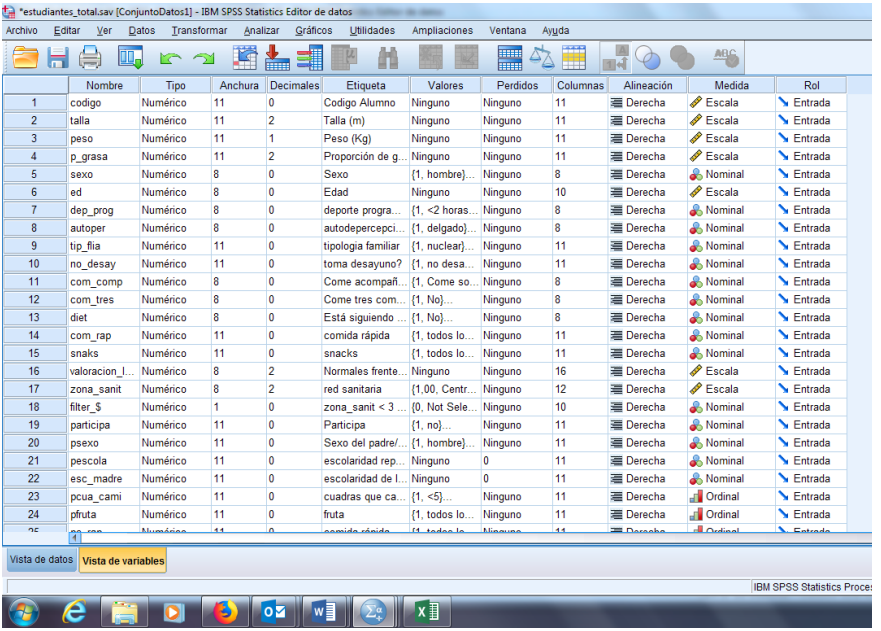
Investigue la información de las variables y su tipología del fichero resultante.

Al realizar lo pedido, la ventana de sintaxis es:

```
DATASET ACTIVATE ConjuntoDatos1.
SORT CASES BY codigo(A).
DATASET ACTIVATE ConjuntoDatos5.
SORT CASES BY codigo(A).
DATASET ACTIVATE ConjuntoDatos1.
STAR JOIN
   /SELECT t0.talla, t0.peso, t0.p_grasa, t0.sexo, t0.ed, t0.dep_prog, t0.autoper,
 t0.tip_flia,
        t0.no_desay, t0.com_comp, t0.com_tres, t0.diet, t0.com_rap, t0.snaks,
t0.valoracion_IMC,
```

```
        t0.zona_sanit, t0.filter_$, t1.participa, t1.psexo, t1.pescola, t1.esc_ma-
dre, t1.pcua_cami,
        t1.pfruta, t1.pc_rap, t1.pdul, t1.prefr, t1.psna, t1.pc_sol, t1.pdiet, t1.au-
toper_pad, t1.pedad,
        t1.pdesay
    /FROM * AS t0
    /JOIN 'ConjuntoDatos5' AS t1
      ON t0.codigo=t1.codigo
    /OUTFILE FILE=*.
```

Al ejecutar esta sintaxis, en la ventana de resultados, se verifica que ya se ha realizado exitosamente la acción, la cual puede verificarse mirando la ventana de variables en donde se han aumentado las mismas, y grabe el archivo resultante con el nombre Estudio.sav.



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	codigo	Numérico	11	0	Codigo Alumno	Ninguno	Ninguno	11	Derecha	Escala	Entrada
2	talla	Numérico	11	2	Talla (m)	Ninguno	Ninguno	11	Derecha	Escala	Entrada
3	peso	Numérico	11	1	Peso (Kg)	Ninguno	Ninguno	11	Derecha	Escala	Entrada
4	p_grasa	Numérico	11	2	Proporción de g...	Ninguno	Ninguno	11	Derecha	Escala	Entrada
5	sexo	Numérico	8	0	Sexo	{1, hombre}...	Ninguno	8	Derecha	Nominal	Entrada
6	ed	Numérico	8	0	Edad	Ninguno	Ninguno	10	Derecha	Escala	Entrada
7	dep_prog	Numérico	8	0	deporte progra...	{1, <2 horas...	Ninguno	8	Derecha	Nominal	Entrada
8	autoper	Numérico	8	0	autodepercepci...	{1, delgado}...	Ninguno	8	Derecha	Nominal	Entrada
9	tip_flia	Numérico	11	0	tipologia familiar	{1, nuclear}...	Ninguno	11	Derecha	Nominal	Entrada
10	no_desay	Numérico	11	0	toma desayuno?	{1, no desa...	Ninguno	11	Derecha	Nominal	Entrada
11	com_comp	Numérico	8	0	Come acompañ...	{1, Come so...	Ninguno	8	Derecha	Nominal	Entrada
12	com_tres	Numérico	8	0	Come tres com...	{1, No}...	Ninguno	8	Derecha	Nominal	Entrada
13	diet	Numérico	8	0	Está siguiendo ...	{1, No}...	Ninguno	8	Derecha	Nominal	Entrada
14	com_rap	Numérico	11	0	comida rápida	{1, todos lo...	Ninguno	11	Derecha	Nominal	Entrada
15	snaks	Numérico	11	0	snacks	{1, todos lo...	Ninguno	11	Derecha	Nominal	Entrada
16	valoracion_l...	Numérico	8	2	Normales frente...	Ninguno	Ninguno	16	Derecha	Escala	Entrada
17	zona_sanit	Numérico	8	2	red sanitaria	{1,00, Centr...	Ninguno	12	Derecha	Escala	Entrada
18	filter_$	Numérico	1	0	zona_sanit < 3 ...	{0, Not Sele...	Ninguno	10	Derecha	Nominal	Entrada
19	participa	Numérico	11	0	Participa	{1, no}...	Ninguno	11	Derecha	Nominal	Entrada
20	psexo	Numérico	11	0	Sexo del padre/...	{1, hombre}...	Ninguno	11	Derecha	Nominal	Entrada
21	pescola	Numérico	11	0	escolaridad rep...	Ninguno	0	11	Derecha	Nominal	Entrada
22	esc_madre	Numérico	11	0	escolaridad de l...	Ninguno	0	11	Derecha	Nominal	Entrada
23	pcua_cami	Numérico	11	0	cuadras que ca...	{1, <5}...	Ninguno	11	Derecha	Ordinal	Entrada
24	pfruta	Numérico	11	0	fruta	{1, todos lo...	Ninguno	11	Derecha	Ordinal	Entrada

Figura 10.8 Ventana de variables del archivo resultante Estudio.sav.

10.3 Ejercicios del Capítulo 4

Ejercicio 4.1

Ejecute la instrucción una vez abierto el archivo **Estudio.sav**.

```
FREQUENCIES VARIABLES=ed
  /STATISTICS=MINIMUM MAXIMUM MEAN
  /ORDER=ANALYSIS.
```

Observará que se muestra el valor máximo y mínimo así como la edad.

- ¿Existe algún estudiante sin definir la edad?
- ¿Observa algún valor erróneo? Resuelva la situación definiendo ese valor como perdido por el usuario, y vuelva a ejecutar la sintaxis.

Para realizar lo solicitado por ventanas, hay que ir a la pestaña **Analizar ► Estadísticos descriptivos ► Frecuencias,** donde se escoge la variable edad.

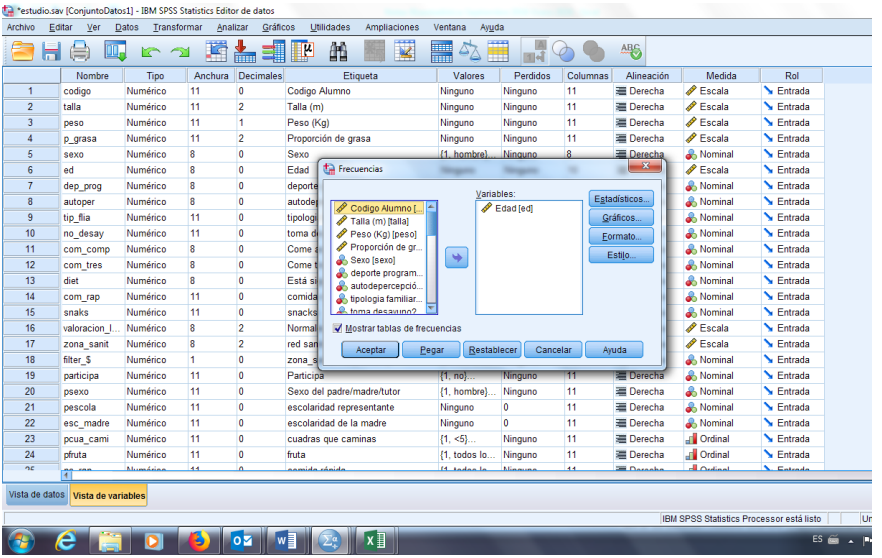


Figura 10.9 Pasos para describir la variable edad

Luego ingrese a Estadísticos y se abre una subventana donde se escoge lo que se desea calcular.

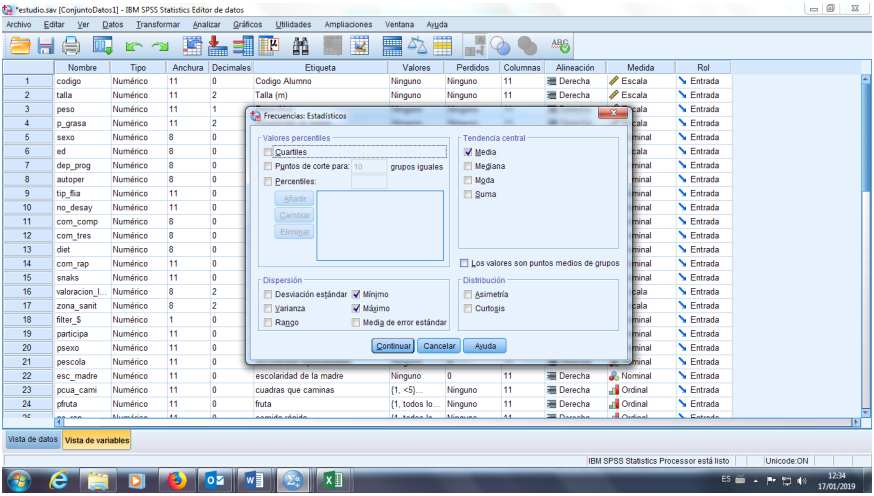


Figura 10.10 Estadísticos para describir la variable edad (variable discreta)

Posteriormente se pone continuar y, por último, **Pegar** y la ventana de sintaxis es la sugerida:

```
FREQUENCIES VARIABLES=ed
  /STATISTICS=MINIMUM MAXIMUM MEAN
  /ORDER=ANALYSIS.
```

En donde colocamos **Ejecutar** y la ventana de resultados muestra la tabla:

Tabla 10.2 Estadísticos

		Edad
N	**Válido**	6964
	Perdidos	0
Media		12,38
Mínimo		9
Máximo		17

En ella se visualiza que el promedio de edad entre los estudiantes es de 12,38 años con un valor máximo de 17 y un valor mínimo de 9 años. No existe ningún estudiante sin definir la edad.

Ejercicio 4.2

En los estudios que analizan el sobrepeso y obesidad en sujetos de gran variabilidad en sexo y en edad, el peso por sí solo no es un elemento que indique casi nada del estatus de obesidad de los individuos, ya que, como el peso incluye la masa ósea, además de la muscular y la grasa, entre otros, la edad en fases de crecimiento introduce grandes cambios en el componente óseo. Por esta razón, se define un indicador que relaciona el peso con la talla del individuo, indicador que se conoce como Índice Másico Corporal (IMC).

Se define como el peso del individuo dividido para la talla expresada en metros al cuadrado. Es más práctico realizar esta creación de IMC a partir de los datos, peso y talla existentes en la base de datos que calcularlo para cada individuo y entrar la variable previamente como una más.

Así, utilizando **Transformar ► Calcular variable** cree la variable IMC para todos los estudiantes participantes. Compruebe mediante la opción **Pegar** que la sintaxis que corresponde a su acción es

```
DATASET ACTIVATE Conjunto_de_datos1.
COMPUTE IMC=peso / (talla) ** 2.
EXECUTE.
```

Realice el análisis de frecuencias de la nueva variable IMC mediante **Analizar ► Estadísticos descriptivos ► Frecuencias** y seleccione las opciones **Estadísticos** el **mínimo** y el **máximo**. Interprete los resultados.

¿Cómo lo resolvería?

Al realizar lo sugerido, se obtiene la pantalla con la sintaxis:

```
COMPUTE IMC=peso / ( talla)**2 .
EXECUTE.
FREQUENCIES VARIABLES=IMC
  /STATISTICS=MINIMUM MAXIMUM MEAN
  /ORDER=ANALYSIS.
```

Y al ejecutarlo, se verificar en la ventana vista de variables la creación de esta nueva variable, que se encuentra al final:

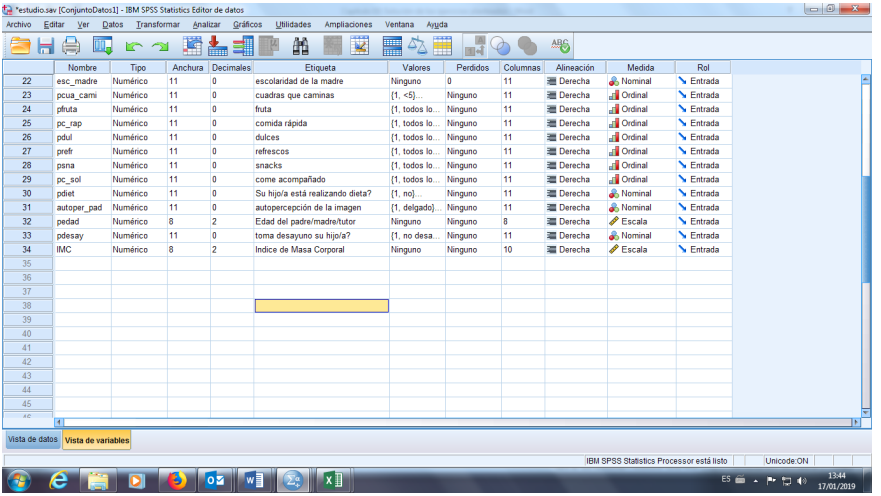


Figura 10.11 Constatación en vista de variables de la variable IMC

Los resultados del análisis estadístico pedido son:

Tabla 10.3 Estadísticos

Índice de Masa Corporal		
N	**Válido**	6961
	Perdidos	3
Media		19,7876
Mínimo		10,24
Máximo		40,69

Donde se puede apreciar que el valor promedio del IMC es 19,79 considerando un valor máximo de 40,69 y un valor mínimo de 10,24, todo en kg/m^2.

Ejercicio 4.3

Una vez haya creado dos variables que, por ejemplo, correspondiesen la fecha de la encuesta y a la fecha actual, mediante la opción **Transformar ► Calcular variable**, determine el tiempo transcurrido desde la fecha de encuesta y la fecha actual.

Pegue y analice la sintaxis creada al pegar en la pantalla.

En la figura adjunta, se muestra el procedimiento por ventanas.

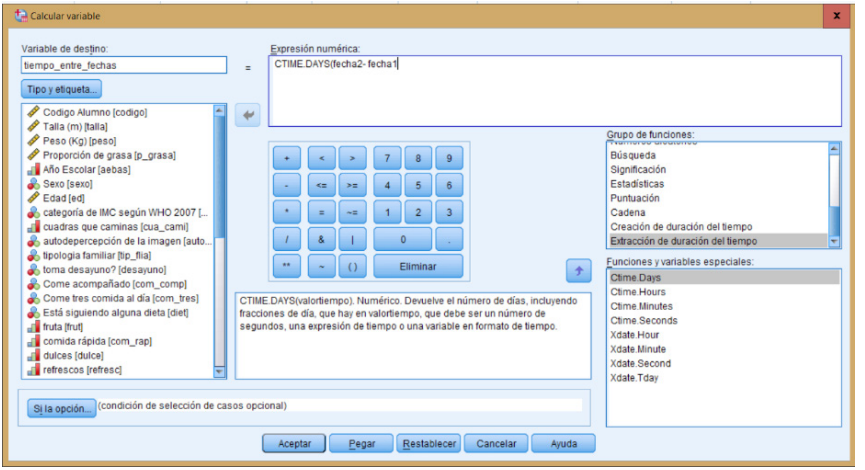


Figura 10.12

Pegue y analice la sintaxis creada al pegar en la pantalla. Se crea como fecha actual 18 de enero de 2019 y como fecha de la encuesta el 23 de junio de 2011 y luego se encuentra el tiempo transcurrido en días. En la ventana de sintaxis se tiene:

```
DATASET ACTIVATE ConjuntoDatos1.
STRING  Fecha_actual (A11).
COMPUTE Fecha_actual=”18-01-2019”.
EXECUTE.
STRING  Fecha_encuesta (A11).
COMPUTE Fecha_encuesta=”23-06-2011”.
EXECUTE.
COMPUTE rango_fechas=CTIME.DAYS(Fecha_actual-Fecha_encuesta).
EXECUTE.
```

Una vez ejecutado, se puede verificar en la vista de datos las variables creadas:

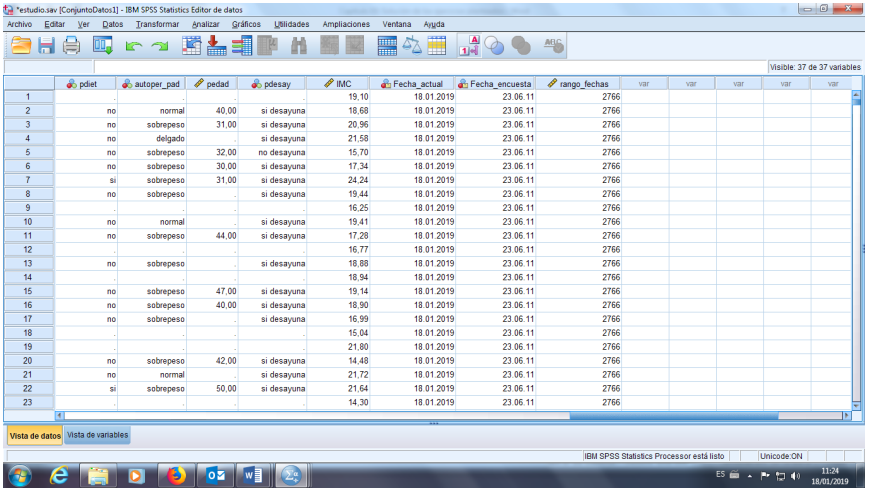


Figura 10.13

Ejercicio 4.4

Utilizando la expresión DO IF defina si el estudiante está desnutrido según los criterios de la tabla 4.3.

Amplíe la sintaxis de forma que la nueva variable desnutrido tenga dos posibles categorías: 1 con etiqueta No y 0 con etiqueta Sí.

Para realizar esto, se puede crear una nueva variable llamada desnutrido donde se colocará las condiciones de ser hombre o mujer de una determinada edad con un IMC mayor o menor que el valor de la tabla dada. Por lo tanto, la sintaxis hay que realizarla para cada una de estas condiciones. Una parte de la ventana de sintaxis será:

```
DO IF ( IMC &amp;gt; 13.7 &amp;amp; sexo = 1 &amp;amp; ed = 9).
COMPUTE DESNUTRIDO=1.
ELSE IF ( IMC &amp;gt; 14.0 &amp;amp; sexo = 1 &amp;amp; ed = 10).
COMPUTE DESNUTRIDO=1.
ELSE IF ( IMC &amp;gt; 14.4 &amp;amp; sexo = 1 &amp;amp; ed = 11).
COMPUTE DESNUTRIDO=1.
```

```
ELSE IF ( IMC &gt; 14.8 &amp; sexo = 1 &amp; ed = 12).
COMPUTE DESNUTRIDO=1.
ELSE IF ( IMC &gt; 15.3 &amp; sexo = 1 &amp; ed = 13).
COMPUTE DESNUTRIDO=1.
ELSE IF ( IMC &gt; 15.9 &amp; sexo = 1 &amp; ed = 14).
COMPUTE DESNUTRIDO=1.
ELSE IF ( IMC &gt; 16.4 &amp; sexo = 1 &amp; ed = 15).
COMPUTE DESNUTRIDO=1.
ELSE IF ( IMC &gt; 16.9 &amp; sexo = 1 &amp; ed = 16).
COMPUTE DESNUTRIDO=1.
ELSE IF ( IMC &gt; 13.4 &amp; sexo = 2 &amp; ed = 9).
COMPUTE DESNUTRIDO=1.
ELSE IF ( IMC &gt; 13.8 &amp; sexo = 2 &amp; ed = 10).
COMPUTE DESNUTRIDO=1.
ELSE IF ( IMC &gt; 14.2 &amp; sexo = 2 &amp; ed = 11).
COMPUTE DESNUTRIDO=1.
ELSE IF ( IMC &gt; 14.8 &amp; sexo = 2 &amp; ed = 12).
COMPUTE DESNUTRIDO=1.
ELSE IF ( IMC &gt; 15.3 &amp; sexo = 2 &amp; ed = 13).
COMPUTE DESNUTRIDO=1.
ELSE IF ( IMC &gt; 15.8 &amp; sexo = 2 &amp; ed = 14).
COMPUTE DESNUTRIDO=1.
ELSE IF ( IMC &gt; 16.2 &amp; sexo = 2 &amp; ed = 15).
COMPUTE DESNUTRIDO=1.
ELSE IF ( IMC &gt; 16.5 &amp; sexo = 2 &amp; ed = 16).
COMPUTE DESNUTRIDO=1.
ELSE IF ( IMC &gt; 16.6 &amp; sexo = 2 &amp; ed = 17).
COMPUTE desnutrido=1.
ELSE.
COMPUTE desnutrido=0.
END IF.
```

Para la segunda parte, solo hay que aumentar en la ventana de sintaxis:

```
VARIABLE LABELS  desnutrido ‘Condición de Desnutrido’.
Value labels desnutrido 0 ‘Si’ 1 ‘No’.
EXECUTE.
```

Para comprobar estos resultados en la ventana de VISTA DE DATOS, en las columnas, se habrá creado una nueva variable con el nombre "desnutrido" en la cual se puede ver un sí o un no de acuerdo a las condiciones dadas.

Ejercicio 4.5

Consideremos que también el hecho de comer solos, no realizar tres comidas al día y no desayunar sean así mismo conductas de riesgo.

Recalcule la variable riesgos añadiendo la información de estas tres nuevas variables.

Realice el ejercicio por sintaxis y describa la variable comparando con la tabla 4.3.

Al realizar lo solicitado, todos tienen valor de etiqueta = 1, por lo que no existe ningún inconveniente y la sintaxis es:

```
COUNT riesgo=com_comp com_tres no_desay(1).
VARIABLE LABELS  riesgo 'numero de conductas de riesgo'.
EXECUTE.
FREQUENCIES VARIABLES=riesgo
  /ORDER=ANALYSIS.
```

Las frecuencias respectivas de la nueva variable de riesgo son:

Tabla 10.4 Número de conductas de riesgo

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	,00	4882	70,1	70,1	70,1
	1,00	1585	22,8	22,8	92,9
	2,00	409	5,9	5,9	98,7
	3,00	88	1,3	1,3	100,0
	Total	6964	100,0	100,0	

Con lo que se podría concluir que existe el 70,1 % de los estudiantes que no tienen ninguna de estas conductas de riesgo y que el 7,2 % tienen por lo menos dos conductas de riesgo.

Ejercicio 4.6

Analice con la instrucción de frecuencias, **Analizar ► Estadísticos descriptivos ► Frecuencias** la variable **psexo**, sexo de los responsables del estudiante.

- ¿Cuántas categorías aparecen?
- ¿Cómo puede recodificar esta variable en solo dos categorías en la nueva variable **sexo_padres**?

Observe que los datos en blanco de psexo son considerados como opción en blanco pero no como valores perdidos. Autorrecodifique en una variable de nombre sexo_padres, aprovechando la autorrecodificación para definir los blancos como valores perdidos y anote los resultados para posteriormente recodificar en esta misma variable.

Al realizar lo sugerido, los datos en blanco salen como perdidos por el sistema como se aprecia en la siguiente tabla:

Tabla 10.5 Sexo del padre/madre/tutor

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
	hombre	1191	17,1	29,9	29,9
Válido	mujer	2796	40,1	70,1	100,0
	Total	3987	57,3	100,0	
Perdidos	Sistema	2977	42,7		
Total		6964	100,0		

Para recodificar la variable psexo en la nueva variable con nombre sexo_padres por medio de la autorecodificación, colocando que esta se va a realizar por valor superior con la siguiente sintaxis:

```
AUTORECODE VARIABLES=psexo
  /INTO sexo_padres
  /DESCENDING
  /PRINT.
```

Al realizar esta acción, la variable se recodifica llenando los espacios en blanco con el número 3, como se puede verificar en la ventana de resultados:

Old Value	New Value	Value Label
2	1	mujer
1	2	hombre
Blanco	3	

10.4 Ejercicios del Capítulo 5

Ejercicio 5.1

Partiendo del archivo **estudio.sav**, seleccione temporalmente por sintaxis, aquellos estudiantes que pertenecen a la **zona sanitaria = 1**.

Observe el efecto de esta acción en la ventana vista de datos.

Repita la acción, pero, en este caso, creae un archivo de datos que contenga exclusivamente a estos estudiantes. Indique la sintaxis generada por la acción **Pegar**.

Al realizar esta instrucción la ventana de sintaxis es:

```
DATASET ACTIVATE ConjuntoDatos1.
USE ALL.
COMPUTE filter_$=(zona_sanit = 1).
VARIABLE LABELS filter_$ 'zona_sanit = 1 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
TEMPORARY.
EXECUTE.
```

Inmediatamente se crea una nueva variable filter_$ en donde, si cumple con la condición pedida, en este caso zona_sanit = 1, en la nueva variable se obtiene Selected y en todos los otros casos Not Selected, como se aprecia en la vista de datos:

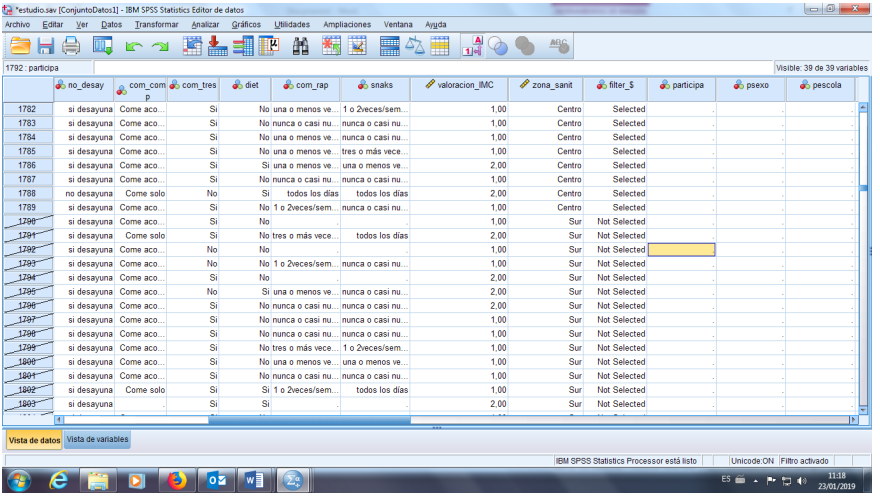


Figura 10.14 Vista de casos resultado de la instrucción Selección de casos

La segunda acción se podría realizar creando una nueva variable donde se considere solo a los alumnos que cumplan con la condición zona_sanit = 1, en cuyo caso la sintaxis podría ser:

```
DATASET ACTIVATE ConjuntoDatos1.
COMPUTE Zona_1=zona_sanit = 1.
EXECUTE.
```

Ejercicio 5.2

Ejecute la siguiente sintaxis:

```
SELECT IF (sexo = 1 & ed = 9).
EXECUTE.
DATASET ACTIVATE  Conjunto_de_datos1.
```

Con esta acción crea un archivo nuevo de datos que solo contiene a los niños varones de 9 años que participan en el estudio. Verá que, en la barra inferior del SPSS, le indicará que existe un archivo nuevo que se llama Sin título (niños de 9 años).

Describa en este nuevo archivo las acciones **Transformar ▶ Analizar ▶ Estadísticos descriptivos ▶ Frecuencias** para la variable **sexo**.

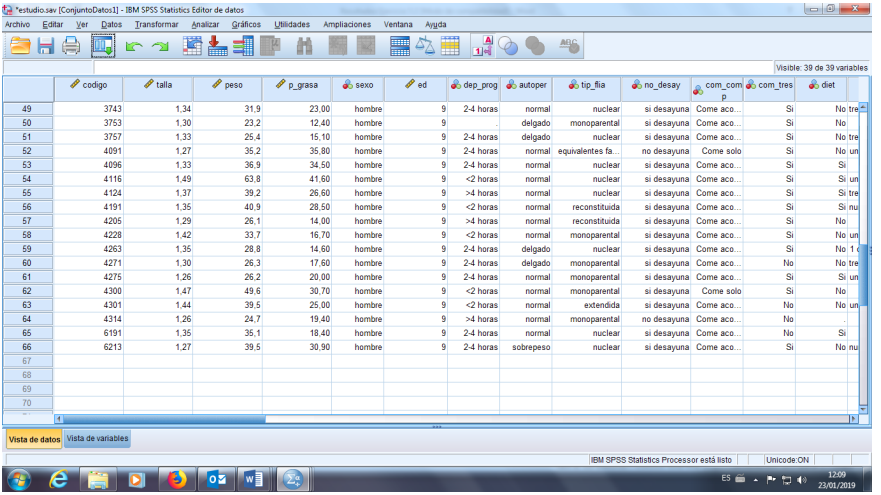
El resultado es que hay 66 niños de esta edad.

A continuación, vaya al archivo **Estudio.sav** y seleccione los casos que cumplen esa misma condición pero sin guardar los seleccionados en un archivo nuevo.

Observará en la matriz de datos que no han desaparecido casos, pero que se ha creado una variable que se denomina **filter_$** con valores cero y verá que el número de registro está marcado con una línea inclinada o uno en función de si el caso cumple la condición de selección. Si efectúa el análisis de frecuencias de esta variable, verá que también el número seleccionado es 66.

Vuelva **Datos ▸ Seleccionar casos** y utilice la opción **usar variable de filtro** y efectúe la descripción de frecuencias. Observará que el filtro se mantiene hasta que no lo anule, aunque haya usado la instrucción USE ALL que equivale a **Selección de datos** opción **Todos los casos.**

Se crea un nuevo archivo con solo los 66 casos que cumple con las condiciones dadas:



	codigo	talla	peso	p_grasa	sexo	ed	dep_prog	autoper	tip_flia	no_desay	com_comp	com_tres	diet	
49	3743	1,34	31,9	23,00	hombre	9	2-4 horas	normal	nuclear	si desayuna	Come aco...	Si	No	tre
50	3753	1,30	23,2	12,40	hombre	9		delgado	monoparental	si desayuna	Come aco...	Si	No	
51	3757	1,33	25,4	15,10	hombre	9	2-4 horas	delgado	nuclear	si desayuna	Come aco...	Si	No	tre
52	4091	1,27	35,2	35,80	hombre	9	2-4 horas	normal	equivalentes fa...	no desayuna	Come solo	Si	No	un
53	4096	1,33	36,9	34,50	hombre	9	2-4 horas	normal	nuclear	si desayuna	Come aco...	Si	Si	
54	4116	1,49	63,8	41,60	hombre	9	<2 horas	normal	nuclear	si desayuna	Come aco...	Si	Si	un
55	4124	1,37	39,2	26,60	hombre	9	>4 horas	normal	nuclear	si desayuna	Come aco...	Si	Si	tre
56	4191	1,35	40,9	28,50	hombre	9	<2 horas	normal	reconstituida	si desayuna	Come aco...	Si	Si	nu
57	4205	1,29	26,1	14,00	hombre	9	>4 horas	normal	reconstituida	si desayuna	Come aco...	Si	No	
58	4228	1,42	33,7	16,70	hombre	9	<2 horas	normal	monoparental	si desayuna	Come aco...	Si	No	un
59	4263	1,35	28,8	14,60	hombre	9	2-4 horas	delgado	nuclear	si desayuna	Come aco...	Si	No	1
60	4271	1,30	26,3	17,60	hombre	9	2-4 horas	delgado	monoparental	si desayuna	Come aco...	No	No	tre
61	4275	1,26	26,2	20,00	hombre	9	2-4 horas	normal	monoparental	si desayuna	Come aco...	Si	Si	un
62	4300	1,47	49,6	30,70	hombre	9	<2 horas	normal	monoparental	si desayuna	Come solo	Si	No	
63	4301	1,44	39,5	25,00	hombre	9	<2 horas	normal	extendida	si desayuna	Come aco...	No	No	un
64	4314	1,26	24,7	19,40	hombre	9	>4 horas	normal	monoparental	no desayuna	Come aco...	No		
65	6191	1,35	35,1	18,40	hombre	9	2-4 horas	normal	nuclear	si desayuna	Come aco...	No	Si	
66	6213	1,27	39,5	30,90	hombre	9	2-4 horas	sobrepeso	nuclear	si desayuna	Come aco...	Si	No	nu
67														
68														
69														
70														

Figura 10.15 Nuevo archivo con los casos que cumplen las condiciones

Y, al obtener frecuencias, se tienen los siguientes resultados, 66 válidos y 0 perdidos:

Tabla 10.6 Sexo

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	hombre	66	100,0	100,0	100,0

La sintaxis para realizar la selección de los casos sin guardar en otro archivo es:

```
DATASET ACTIVATE ConjuntoDatos2.
USE ALL.
COMPUTE filter_$=(sexo = 1 & ed = 9).
VARIABLE LABELS filter_$ 'sexo = 1 & ed = 9 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
TEMPORARY.
EXECUTE.
```

Creándose esta nueva variable **filter**_$, en donde los casos no seleccionados se encuentran con una línea inclinada.

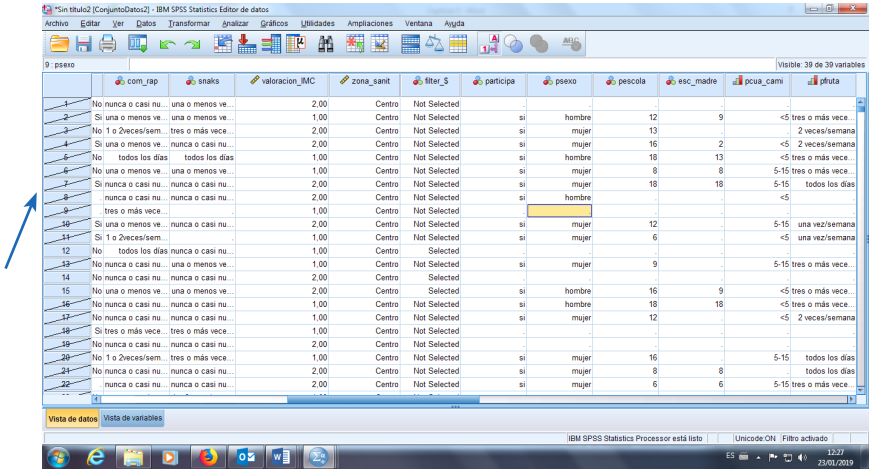


Figura 10.16

Al obtener una distribución de frecuencias usando esta nueva variable de filtro con la sintaxis:

```
USE ALL.
FILTER BY filter_$.
EXECUTE.
FREQUENCIES VARIABLES=filter_$
  /ORDER=ANALYSIS.
```

Donde se obtendrá los mismos resultados que en el caso anterior con 66 válidos y 0 perdidos:

Tabla 10.7 sexo = 1 & ed = 9 (FILTER)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Selected	66	100,0	100,0	100,0

Ejercicio 5.3

Vuelva a efectuar el análisis, pero cambiando el orden de las variables de agrupación, es decir, visualizando las frecuencias por zona sanitaria para cada sexo. En este caso, la jerarquía de agrupación es Sexo, Zona sanitaria.

Pegue las instrucciones de sintaxis y observe las diferencias al definir los estratos por sexo y zona sanitaria con la que se muestra en la tabla 5.1.

La sintaxis correspondiente al ejercicio es:

```
SORT CASES BY sexo(A) zona_sanit(A).
SORT CASES  BY sexo zona_sanit.
SPLIT FILE LAYERED BY sexo zona_sanit.
FREQUENCIES VARIABLES=no_desay
  /ORDER=ANALYSIS.
```

Y la tabla correspondiente para si toma o no desayuno es:

Tabla 10.8 ¿Toma desayuno?

zona sanitaria				Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Centro	hombre	Válido	no desayuna	23	2,2	2,2	2,2
			sí desayuna	1012	97,7	97,8	100,0
			Total	1035	99,9	100,0	
		Perdidos	Sistema	1	0,1		
		Total		1036	100,0		
	mujer	Válido	no desayuna	55	4,9	4,9	4,9
			sí desayuna	1076	95,1	95,1	100,0
			Total	1131	99,9	100,0	
		Perdidos	Sistema	1	0,1		
		Total		1132	100.0		
Norte	hombre	Válido	no desayuna	33	3,7	3,7	3,7
			sí desayuna	857	96,2	96,3	100,0
			Total	890	99,9	100,0	
		Perdidos	Sistema	1	0,1		
		Total		891	100,0		
	mujer	Válido	no desayuna	55	7,3	7,3	7,3
			sí desayuna	695	92,5	92,7	100,0
			Total	750	99,9	100,0	
		Perdidos	Sistema	1	0,1		
		Total		751	100,0		

Sur	hombre		no desayuna	38	4,2	4,3	4,3
		Válido	sí desayuna	855	95,4	95,7	100,0
			Total	893	99,7	100,0	
		Perdidos	Sistema	3	0,3		
		Total		896	100,0		
	mujer		no desayuna	29	2,9	2,9	2,9
		Válido	sí desayuna	968	97,0	97,1	100,0
			Total	997	99,9	100,0	
		Perdidos	Sistema	1	0,1		
		Total		998	100,0		
Periférica	hombre	Perdidos	Sistema	3	100,0		
			no desayuna	26	6,0	6,1	6,1
		Válido	sí desayuna	397	92,3	93,9	100,0
			Total	423	98,4	100,0	
		Perdidos	Sistema	7	1,6		
		Total		430	100,0		
	mujer		no desayuna	113	13,7	13,8	13,8
		Válido	sí desayuna	707	85,5	86,2	100,0
			Total	820	99,2	100,0	
		Perdidos	Sistema	7	0,8		
		Total		827	100.0		

Se encuentra tomado en cuenta primero la variable sexo y, en segundo lugar, la red sanitaria.

Ejercicio 5.4

Ejecute la sintaxis arriba indicada. (*Copy* y *paste* en una ventana de sintaxis).

Abra el fichero **aggr.sav**, generado al activar la instrucción de agregar, y mediante la acción **Transformar ▸ Calcular variable** determine la prevalencia, es decir, el porcentaje de alumnos con exceso de peso en cada estrato.

¿En qué estrato es mayor la prevalencia de exceso de peso y de alumnos que van a la escuela sin desayunar?

¿Cuántos alumnos del estudio están realizando dieta?

(**Frecuencias ▸ Estadísticos ▸ suma**).

Al realizar lo solicitado mediante la sintaxis dada al abrir el nuevo archivo y aplicar la sintaxis:

```
COMPUTE porcentaje_exceso_peso=(exceso_de_peso_cin / N_BREAK).
EXECUTE.
```

Se obtiene, en la ventana vista de datos, lo siguiente:

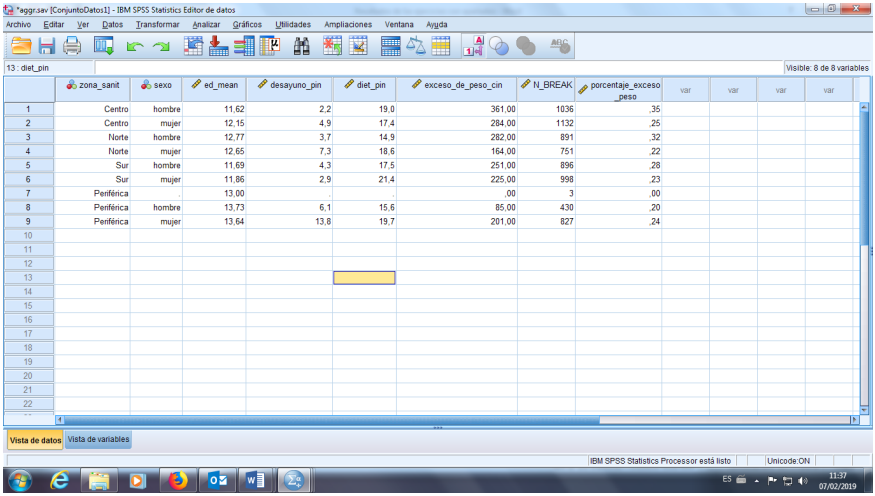


Figura 10.17 Archivo resultante de aplicar la instrucción “Compute” o “Calcular”

No se necesita ningún análisis adicional y, solo a partir de esta pantalla, se puede contestar las preguntas solicitadas:

En el sector centro, los hombres tienen mayor prevalencia de exceso de peso con un valor del 35 %. Las mujeres del sector periférico tienen mayor prevalencia de no desayunar con un valor del 13,8 %. 1219 estudiantes se encuentran realizando una dieta. Para responder esta pregunta, se realiza una distribución de frecuencia mediante la sintaxis:

```
FREQUENCIES VARIABLES=diet
  /STATISTICS=SUM
  /ORDER=ANALYSIS.
```

Y se obtiene en resultados la tabla:

Tabla 10.9 Está siguiendo alguna dieta

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
	No	5476	78,6	81,8	81,8
Válido	Sí	1219	17,5	18,2	100,0
	Total	6695	96,1	100,0	
Perdidos	Sistema	269	3,9		
Total		6964	100,0		

10.5 Ejercicios del Capítulo 6

Ejercicio 6.1

Explore mediante la acción **Analizar ▶ Estadísticos descriptivos ▶ Frecuencias** qué variables del archivo poseen más del 5 % de valores perdidos. Compruebe previamente que todas las variables tienen definidos correctamente los valores considerados *missing*.

Al realizar esta operación, la sintaxis correspondiente es:

```
FREQUENCIES VARIABLES=talla peso p_grasa sexo ed dep_prog autoper
tip_flia no_desay com_comp  com_tres diet com_rap snaks valoracion_IMC
```

```
zona_sanit filter_$ participa psexo pescola esc_madre  pcua_cami pfruta
pc_rap pdul prefr psna pc_sol pdiet autoper_pad pedad pdesay IMC Fecha_ac-
tual  Fecha_encuesta rango_fechas desnutrido riesgo
 /ORDER=ANALYSIS.
```

Y, en resultado, solo nos interesa la primera tabla, donde se resumen los válidos y los perdidos.

		Talla (m)	Peso (kg)	Proporción de grasa	Sexo	Edad	deporte programado	autodepercepción de la imagen	tipologia familiar
N	Válido	6964	6964	6961	6961	6964	6299	6545	6887
	Perdidos	0	0	3	3	0	665	419	77

		toma desayuno?	Come acompañado	Come tres comidas al día	Está siguiendo alguna dieta	Comida rápida	snacks	Normales frente sobrepeso u obeso	red sanitaria
N	Válido	6939	6854	6840	6695	6315	6486	6964	6964
	Perdidos	25	110	124	269	649	478	0	0

		Participa	Sexo del padre/ madre/tutor	escolaridad representante	Escolaridad de la madre	Cuadras que caminas	fruta	comida rápida	dulces
N	Válido	4064	3987	3793	1918	3499	3897	3861	3861
	Perdidos	2900	2977	3171	5046	3465	3067	3103	3103

		Refrescos	*snacks*	come acompañado	Su hijo/a está realizando dieta?	Auto-percepción de la imagen	Edad del padre/ madre/ tutor	toma desayuno su hijo/a?	IMC
N	Válido	3899	3846	3710	4055	3876	3188	4055	6964
	Perdidos	3065	3118	3254	2909	3088	3776	2909	0

		Fecha_ actual	Fecha_ encuesta	rango_fechas	desnutrido	Numero de conductas de riesgo
N	Válido	6964	6964	6964	6964	6964
	Perdidos	0	0	0	0	0

Las variables que tienen más del 5 % de valores perdidos son: deporte programado, autopercepción de la imagen, comida rápida, *snacks*, participa, sexo del tutor, escolaridad del tutor, escolaridad de la madre, cuadras que camina, fruta, comida rápida según tutor, dulces, refrescos, *snaks*, come acompañado, hijo realiza dieta, edad tutor, hijo desayuna.

Ejercicio 6.2

Efectúe el control de calidad de las variables talla en metros, peso en kg, edad y sexo de los padres.

Cuando sea necesario, redacte una petición de rectificación de errores indicando el caso o código a que hace referencia cada error y realice la corrección.

Para la talla, peso y la edad, se comprueba revisando el máximo y mínimo, con la sintaxis:

```
DESCRIPTIVES VARIABLES=talla peso ed
  /STATISTICS=MIN MAX.
```

Y se obtiene la siguiente tabla:

Tabla 10.10 Estadísticos descriptivos

	N	Mínimo	Máximo
Talla (m)	6964	1,15	1,92
Peso (kg)	6964	19,7	114,7
Edad	6964	9	17
N válido (por lista)	6964		

Como en la psexo, que corresponde al sexo del tutor, existen muchos perdidos, se puede realizar la rectificación y, para comprobarla, debe realizar una distribución de frecuencias con la siguiente sintaxis:

```
COMPUTE  psexo=LTRIM (UPCAS (psexo) ) .
EXECUTE .
FREQUENCIES VARIABLES=psexo
  /ORDER=ANALYSIS.
```

Tabla 10.11 Estadísticos

		Sexo del padre/madre/tutor
N	Válido	6964
	Perdidos	0

Sexo del padre/madre/tutor					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido		2977	42,7	42,7	42,7
hombre	1191	17,1	17,1	59,9	
mujer	2796	40,1	40,1	100,0	
Total	6964	100,0	100,0		

Ejercicio 6.3

Describa las variables **IMC** y **p_grasa** y comente la posible normalidad de las distribuciones.

Analice los valores erróneos y defínalos como ***missing*** de usuario. Repita el análisis e interprete los cambios al eliminar valores erróneos.

Se puede obtener lo pedido para las dos variables simultáneamente con la sintaxis:

```
FREQUENCIES VARIABLES=IMC p_grasa
  /FORMAT=NOTABLE
  /STATISTICS=STDDEV MINIMUM MAXIMUM MEAN MEDIAN MODE SKEW-
  NESS SESKEW KURTOSIS SEKURT
  /HISTOGRAM NORMAL
  /ORDER=ANALYSIS.
```

Y se obtienen los siguientes resultados:

Tabla 10.12 Estadísticos

		Índice de Masa Corporal	Proporción de grasa
N	Válido	6964	6961
	Perdidos	0	3
Media		19,7904	22,4808
Mediana		19,3236	22,5000
Moda		19,34	26,40
Desviación estándar		3,17191	7,17537
Asimetría		,908	,209
Curtosis		1,507	–,162
Mínimo		10,24	1,70
Máximo		40,69	57,70

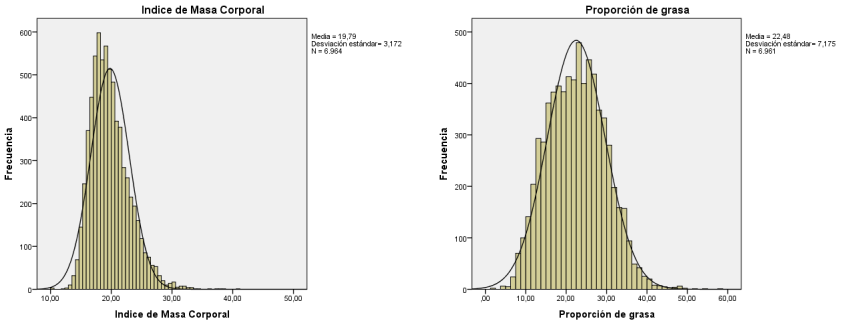


Figura 10.18

Para el IMC, si analizamos los valores de media y mediana, son muy parecidos con una ligera asimetría positiva y una curtosis de 4,5; esto quiere decir que hay más datos concentrados alrededor de la media. Por lo tanto, llama la atención el valor máximo, el cual debería ser revisado, pero con estos valores se podría considerar ligeramente diferente a la normal.

Para la proporción de grasa, los valores tendencia central son mucho más similares, por lo tanto con un valor de asimetría más bajo, con un valor de curtosis más cercano a 3, por lo que es mucho más cercana a una distribución normal. El valor máximo de esta variable si es muy importante revisarlo ya que es bastante elevado.

En el IMC, no existen valores perdidos y, en la proporción de grasa, existen muy pocos valores perdidos, por lo que los resultados sin ellos serán muy similares.

Ejercicio 6.4

Efectúe la descriptiva bivariada de la proporción de grasa según el sexo de los estudiantes.

Tenga en cuenta que si, por algún motivo, ha ordenado el archivo por cualquier criterio, el número identificativo que obtiene en el gráfico cambia y se refiere a la situación en el archivo después del nuevo orden introducido.

Compare el resultado obtenido si en la pestaña de Gráficos selecciona variables o factores juntos.

Indique el código de los alumnos con valores extremos.

Los resultados deseados se obtienen mediante la sintaxis:

```
EXAMINE VARIABLES=p_grasa BY sexo
  /PLOT BOXPLOT HISTOGRAM
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

Y estos resultados son un resumen de casos:

Tabla 10.13 Resumen de procesamiento de casos

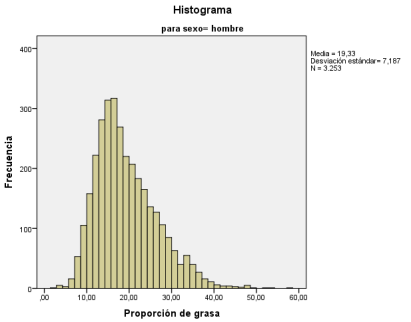
		Casos					
	Sexo	Válido		Perdidos		Total	
		N	Porcentaje	N	Porcentaje	N	Porcentaje
Proporción de grasa	hombre	3253	100,	0	0,0	3253	100,0
	mujer	3705	99,9	3	0,1	3708	100,0

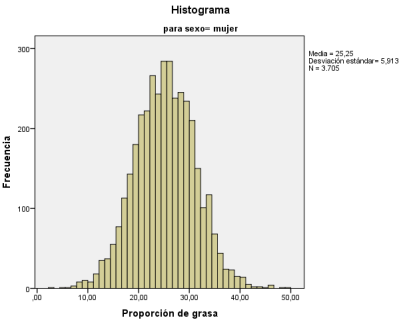
Y luego la información de los estadísticos:

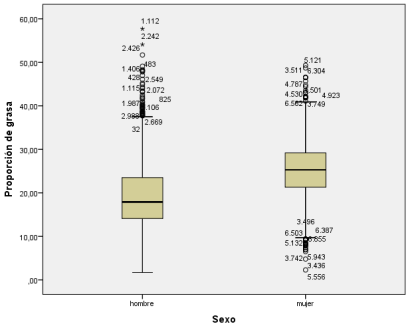
Tabla 10.14 Descriptivos

	Sexo			Estadístico	Error estándar
Proporción de grasa	Hombre	Media		19,3264	,12602
		95% de intervalo de confianza para la media	Límite inferior	19,0793	
			Límite superior	19,5735	
		Media recortada al 5 %		18,9477	
		Mediana		17,9000	
		Varianza		51,658	
		Desviación estándar		7,18732	
		Mínimo		1,70	
		Máximo		57,70	
		Rango		56,00	
		Rango intercuartil		9,40	
		Asimetría		,872	,043
		Curtosis		,959	,086
	Mujer	Media		25,2509	,09714
		95 % de intervalo de confianza para la media	Límite inferior	25,0605	
			Límite superior	25,4414	
		Media recortada al 5 %		25,2481	
		Mediana		25,3000	
		Varianza		34,962	
		Desviación estándar		5,91284	
		Mínimo		2,28	
		Máximo		49,30	
		Rango		47,02	
		Rango intercuartil		7,95	
		Asimetría		,029	,040
		Curtosis		,212	,080

Por último, los gráficos solicitados, los histogramas y el diagrama de caja:







Al colocar con variable o valores juntos, la sintaxis cambia a:

```
EXAMINE VARIABLES=p_grasa BY sexo
  /PLOT BOXPLOT HISTOGRAM
  /COMPARE VARIABLES
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

Y en este caso, los resultados obtenidos son los mismos.

Los códigos con valores extremos son, en hombres, 1112, 2242 y, en mujeres no existen valores extremos.

Ejercicio 6.5

Analice la tabla cruzada desayunar y exceso de peso.

Solicite en la pestaña Casillas, la información: Observados, Esperados, Fila, Columna y Total, así como en Estadísticos los valores de Chi cuadrado y Riesgo.

Responda a las siguientes preguntas:

1.- ¿Qué proporción de los que no toman desayuno presentan exceso de peso?

2.- ¿Qué proporción de los que tienen exceso de peso no toman desayuno?

3.- ¿Qué porcentaje de estudiantes no toman desayuno y tienen exceso de peso?

4.- ¿Qué razón de prevalencia de exceso de peso existe entre los dos grupos?

5.- ¿Cuál es el valor de la OR?

6.- ¿Puede decirse que estas diferencias pueden ser debidas al muestreo?

Para analizar la tabla cruzada entre las variables desayunar y exceso de peso se realiza la siguiente sintaxis:

```
CROSSTABS
  /TABLES=no_desay BY valoracion_IMC
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ RISK
  /CELLS=COUNT EXPECTED ROW COLUMN TOTAL
  /COUNT ROUND CELL.
```

Y se obtienen los siguientes resultados:

Tabla 10.15 Resumen de procesamiento de casos

	Casos					
	Válido		perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Toma desayuno? * Normales frente sobrepeso u obeso	6939	99,6	25	0,4	6964	100,0

Tabla 10.16 Tabla cruzada ¿toma desayuno? *Normales frente sobrepeso u obeso

			Normales frente sobrepeso u obeso	
			Normal	**Con sobrepeso**
toma desayuno?	no desayuna	Recuento	246	126
		Recuento esperado	272,9	99,1
		% dentro de toma desayuno?	66,1	33,9
		% dentro de Normales frente sobrepeso u obeso	4,8	6,8
		% del total	3,5	1,8
	sí desayuna	Recuento	4845	1722
		Recuento esperado	4818,1	1748,9
		% dentro de toma desayuno?	73,8	26,2
		% dentro de Normales frente sobrepeso u obeso	95,2	93,2
		% del total	69,8	24,8
Total		Recuento	5091	1848
		Recuento esperado	5091,0	1848,0
		% dentro de toma desayuno?	73,4	26,6
		% dentro de normales frente sobrepeso u obeso	100,0	100,0
		% del total	73,4	26,6

Tabla 10.17 Tabla cruzada ¿toma desayuno? *Normales frente sobrepeso u obeso

			Total
toma desayuno?	no desayuna	Recuento	372
		Recuento esperado	372,0
		% dentro de toma desayuno?	100,0
		% dentro de Normales frente sobrepeso u obeso	5,4
		% del total	5,4
	sí desayuna	Recuento	6567
		Recuento esperado	6567,0
		% dentro de toma desayuno?	100,0
		% dentro de Normales frente sobrepeso u obeso	94,6
		% del total	94,6
Total		Recuento	6939
		Recuento esperado	6939,0
		% dentro de toma desayuno?	100,0
		% dentro de Normales frente sobrepeso u obeso	100,0
		% del total	100,0

Tabla 10.18 Pruebas de chi cuadrado

	Valor	df	Significación asintótica (bilateral)	Significación exacta (bilateral)	Significación exacta (unilateral)
Chi cuadrado de Pearson	10,542[a]	1	,001		
Corrección de continuidad[b]	10,154	1	,001		
Razón de verosimilitud	10,050	1	,002		
Prueba exacta de Fisher				,002	,001
Asociación lineal por lineal	10,540	1	,001		
N de casos válidos	6939				

a. Ninguna casilla (0,0 %) ha esperado un recuento menor que 5. El recuento mínimo esperado es 99,07.

b. Solo se ha calculado para una tabla 2x2.

Tabla 10.19 Estimación de riesgo

	Valor	Intervalo de confianza de 95 %	
		Inferior	Superior
Razón de ventajas para ¿toma desayuno? (no desayuna / sí desayuna)	,694	,556	,866
Para cohorte Normales frente sobrepeso u obeso = Normal	,896	,832	,965
Para cohorte Normales frente sobrepeso u obeso = Con Sobrepeso	1,292	1,114	1,497
N de casos válidos	6939		

A partir de estos resultados, se pueden responder a las preguntas solicitadas:

1. El 33,9 % de estudiantes de un total de 372 que no toman desayuno tienen exceso de peso.
2. El 6,8% de estudiantes de un total de 1648 que tienen sobrepeso no toman desayuno.
3. De forma global, se tiene que el 1,8 % del total de 6939 del estudio son alumnos que no toman desayuno y presentan un exceso de peso.
4. Existe una prevalencia, de los estudiantes que tienen exceso de peso, del 29.2 % mayor de los que no desayunan respecto a los que desayunan.
5. El valor de OR es de 1,441.
6. No, ya que $p < 0{,}005$, por lo que el exceso de peso está asociado al hecho de tomar o no desayuno.

Ejercicio 6.6

Lleve a cabo una descripción completa de la variable talla en metros, indicando qué valores declara como missing de usuario.

Relacione la talla y el peso describiendo por el hecho de desayunar o no, según el sexo de los responsables de los alumnos.

Describa el valor de IMC en función de si tienen exceso de peso o no y relacione esta última propiedad en función del sexo y si desayuna o no.

Responda a la siguiente pregunta: ¿cómo abordaría la coincidencia de respuesta entre padres e hijos sobre el hecho de consumir comida rápida? ¿Concuerdan? ¿Cómo discrepan?

Realice un gráfico de las medias de talla en función de la edad y el sexo, incluyendo en el gráfico los intervalos de confianza.

Para realizar un análisis descriptivo de la variable talla en metros, se realiza la siguiente sintaxis:

```
FREQUENCIES VARIABLES=talla
  /NTILES=4
  /STATISTICS=STDDEV MINIMUM MAXIMUM SEMEAN MEAN MEDIAN MODE
SUM SKEWNESS SESKEW KURTOSIS SEKURT
  /HISTOGRAM NORMAL
  /ORDER=ANALYSIS.
```

Y se obtiene:

Tabla 10.20 Estadísticos

Talla (m)		
N	Válido	6964
	Perdidos	0
Media		1,4798
Mediana		1,4800
Moda		1,51
Desviación estándar		,10166
Asimetría		,023
Curtosis		-,116
Mínimo		1,15
Máximo		1,92
Percentiles	25	1,4100
	50	1,4800
	75	1,5500

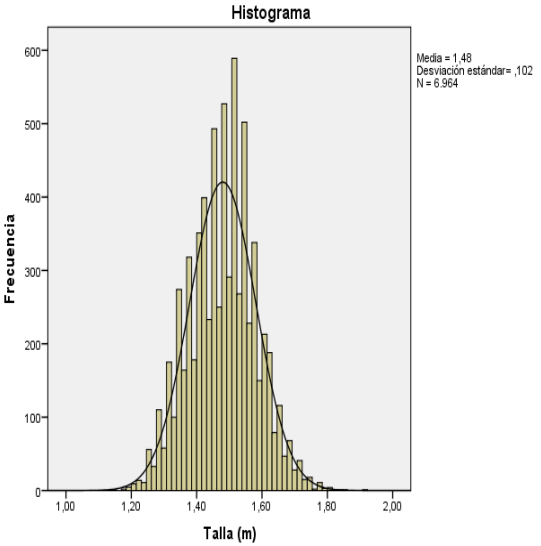


Figura 10.20

Donde se tiene valores como que la media de la estatura es 1,47 m con una desviación estándar de 0,1 m y la mediana 1,48 m, valores muy similares. Además, existe un valor pequeño de una asimetría positiva y un valor de curtosis muy cercano a 3, por lo que la distribución es ligeramente diferente a la Normal. No existe ningún valor perdido; sin embargo, podría revisarse el valor máximo.

Para relacionar la talla y el peso descritos por el hecho de desayunar o no, según el sexo de los responsables de los alumnos, se realiza la siguiente sintaxis:

```
EXAMINE VARIABLES=talla peso BY no_desay
/ID=psexo
/PLOT BOXPLOT HISTOGRAM
/COMPARE GROUPS
/STATISTICS DESCRIPTIVES
/CINTERVAL 95
/MISSING LISTWISE
/NOTOTAL.
```

Se obtienen los siguientes cuadros y gráficos:

Tabla 10.21 ¿Toma desayuno?

Descriptivos					
	toma desayuno?			**Estadístico**	**Error estándar**
Talla (m)	no desayuna	Media		1,5091	,00464
		95 % de intervalo de confianza para la media	Límite inferior	1,4999	
			Límite superior	1,5182	
		Mediana		1,5100	
		Varianza		,008	
		Desviación estándar		,08956	
		Mínimo		1,26	
		Máximo		1,84	
		Rango		,58	
		Rango intercuartil		,10	
		Asimetría		–,031	,126
		Curtosis		,731	,252
	sí desayuna	Media		1,4781	,00126
		95 % de intervalo de confianza para la media	Límite inferior	1,4756	
			Límite superior	1,4806	
		Mediana		1,4800	
		Varianza		,010	
		Desviación estándar		,10204	
		Mínimo		1,15	
		Máximo		1,92	
		Rango		,77	
		Rango intercuartil		,14	
		Asimetría		,034	,030
		Curtosis		–,145	,060
Peso (kg)	no desayuna	Media		49,262	,5656
		95 % de intervalo de confianza para la media	Límite inferior	48,150	
			Límite superior	50,374	
		Mediana		48,600	
		Varianza		118,985	
		Desviación estándar		10,9080	
		Mínimo		24,7	
		Máximo		114,7	
		Rango		90,0	
		Rango intercuartil		12,6	
		Asimetría		1,095	,126
		Curtosis		4,729	,252
	sí desayuna	Media		43,533	,1294
		95 % de intervalo de confianza para la media	Límite inferior	43,280	
			Límite superior	43,787	
		Mediana		42,800	
		Varianza		109,968	
		Desviación estándar		10,4866	
		Mínimo		19,7	
		Máximo		100,3	
		Rango		80,6	
		Rango intercuartil		14,2	
		Asimetría		,655	,030
		Curtosis		,805	,060

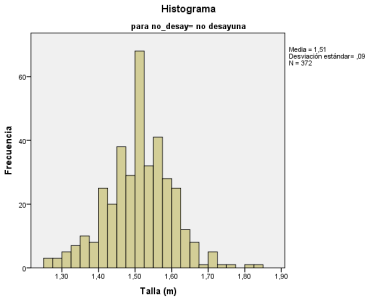


Figura 10.21 Talla (m) – Histogramas

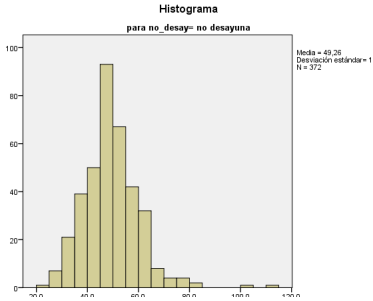


Figura 10.22 Peso (kg) – Histogramas

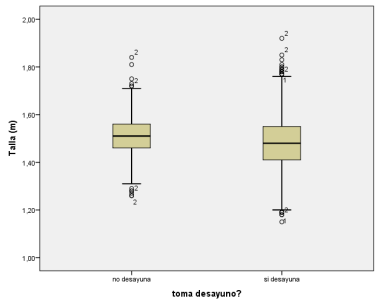


Figura 10.23 Diagramas de caja

Para describir el valor de IMC en función de si tienen exceso de peso o no, se realiza la siguiente sintaxis:

```
EXAMINE VARIABLES=IMC BY valoracion_IMC
  /PLOT BOXPLOT HISTOGRAM
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

Se obtienen los siguientes resultados:

Tabla 10.22

Indice de Masa Corporal	Normal	Media		18,4262	,02814
		95% de intervalo de confianza para la media	Límite inferior	18,3710	
			Límite superior	18,4814	
		Media recortada al 5%		18,4081	
		Mediana		18,3325	
		Varianza		4,048	
		Desviación estándar		2,01198	
		Mínimo		10,24	
		Máximo		24,65	
		Rango		14,40	
		Rango intercuartil		2,83	
		Asimetría		,131	,034
		Curtosis		-,215	,068
	Con Sobrepeso	Media		23,5532	,06300
		95% de intervalo de confianza para la media	Límite inferior	23,4296	
			Límite superior	23,6767	
		Media recortada al 5%		23,3822	
		Mediana		23,2735	
		Varianza		7,355	
		Desviación estándar		2,71201	
		Mínimo		18,33	
		Máximo		40,69	
		Rango		22,36	
		Rango intercuartil		3,40	
		Asimetría		1,122	,057
		Curtosis		2,713	,114

Con sus gráficos respectivos:

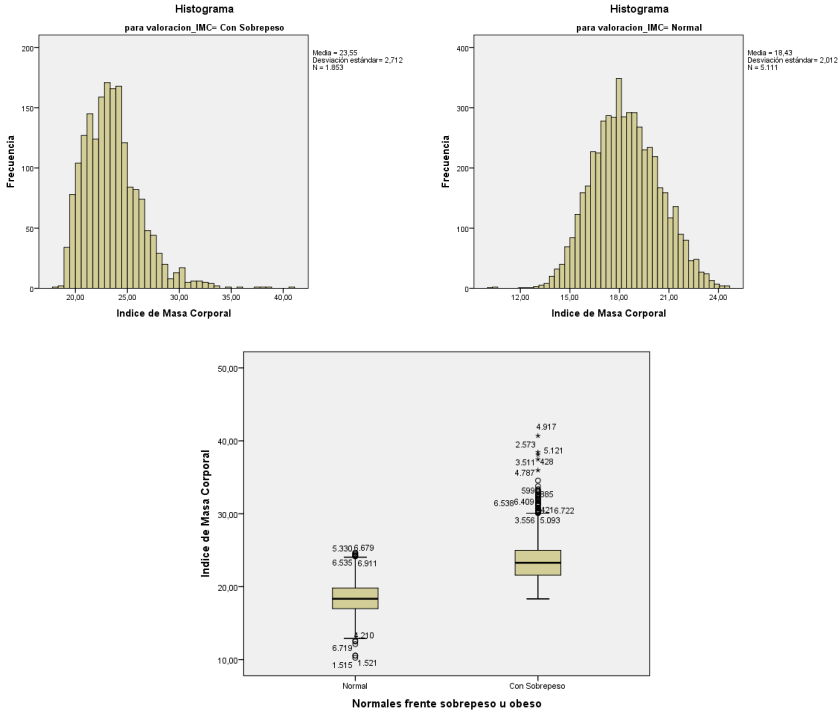


Figura 10.24

Y para relacionar si tienen exceso de peso o no en función del sexo y si desayuna o no, se realiza la siguiente sintaxis:

```
CROSSTABS
  /TABLES=valoracion_IMC BY sexo no_desay
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT ROW
  /COUNT ROUND CELL.
```

Tabla 10.23 Normales frente sobrepeso u obeso * Sexo

Tabla cruzada

			Sexo		
			hombre	**mujer**	**Total**
Normales frente sobrepeso u obeso	Normal	Recuento	2274	2834	5108
		% dentro de Normales frente sobrepeso u obeso	44,5	55,5	100,0
	Con sobrepeso	Recuento	979	874	1853
		% dentro de Normales frente sobrepeso u obeso	52,8	47,2	100,0
Total		Recuento	3253	3708	6961
		% dentro de Normales frente sobrepeso u obeso	46,7	53,3	100,0

Tabla 10.25 Normales frente sobrepeso u obeso * toma desayuno?

Tabla cruzada

			toma desayuno?		
			no desayuna	**sí desayuna**	**Total**
Normales frente sobrepeso u obeso	Normal	Recuento	246	4845	5091
		% dentro de Normales frente sobrepeso u obeso	4,8	95,2	100,0
	Con sobrepeso	Recuento	126	1722	1848
		% dentro de Normales frente sobrepeso u obeso	6,8	93,2	100,0
Total		Recuento	372	6567	6939
		% dentro de Normales frente sobrepeso u obeso	5,4	94,6	100,0

Tabla 10.26 Pruebas de chi cuadrado

	Valor	df	Significación asintótica (bilateral)	Significación exacta (bilateral)	Significación exacta (unilateral)
Chi cuadrado de Pearson	10,542[a]	1	,001		
Corrección de continuidad[b]	10,154	1	,001		
Razón de verosimilitud	10,050	1	,002		
Prueba exacta de Fisher				,002	,001
Asociación lineal por lineal	10,540	1	,001		
N de casos válidos	6939				

a. Ninguna casilla (0,0 %) ha esperado un recuento menor que 5. El recuento mínimo esperado es 99,07.

b. Solo se ha calculado para una tabla 2x2.

Se puede apreciar que existe una dependencia estadísticamente significativa entre las variables de estudio.

Para responder a las preguntas:

- ¿Cómo abordaría la coincidencia de respuesta entre padres e hijos sobre el hecho de consumir comida rápida?
- ¿Concuerdan?
- ¿Cómo discrepan?

Se realiza la siguiente sintaxis para obtener una tabla cruzada entre estas dos variables conjuntamente con la prueba kappa index y Mc Nemar test:

```
CROSSTABS
  /TABLES=com_rap BY pc_rap
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ KAPPA MCNEMAR
  /CELLS=COUNT EXPECTED ROW
  /COUNT ROUND CELL.
```

Con lo que se tienen los siguientes resultados:

Tabla cruzada comida rápida*comida rápida

			comida rápida					
			todos los días	tres o más veces/ semana, pero no todos los días	2 veces/semana	una vez/semana	nunca	Total
comida rápida	todos los días	Recuento	23	23	42	89	63	240
		Recuento esperado	6,8	11,7	39,0	103,0	79,5	240,0
		% dentro de comida rápida	9,6 %	9,6 %	17,5 %	37,1 %	26,3 %	100,0 %
	tres o más veces/ semana, pero no todos los días	Recuento	24	36	103	175	105	443
		Recuento esperado	12,6	21,7	71,9	190,1	146,7	443,0
		% dentro de comida rápida	5,4 %	8,1 %	23,3 %	39,5 %	23,7 %	100,0 %
	1 o 2veces/ semana	Recuento	18	51	187	335	160	751
		Recuento esperado	21,3	36,7	121,9	322,3	248,7	751,0
		% dentro de comida rápida	2,4 %	6,8 %	24,9 %	44,6 %	21,3 %	100,0 %
	una o menos veces/ semana	Recuento	20	45	153	556	303	1077
		Recuento esperado	30,6	52,6	174,9	462,3	356,7	1077,0
		% dentro de comida rápida	1,9 %	4,2 %	14,2 %	51,6 %	28,1 %	100,0 %
	nunca o casi nunca	Recuento	16	19	93	373	548	1049
		Recuento esperado	29,8	51,3	170,3	450,2	347,4	1049,0
		% dentro de comida rápida	1,5 %	1,8 %	8,9 %	35,6 %	52,2 %	100,0 %
Total		Recuento	101	174	578	1528	1179	3560
		Recuento esperado	101,0	174,0	578,0	1528,0	1179,0	3560,0
		% dentro de comida rápida	2,8 %	4,9 %	16,2 %	42,9 %	33,1 %	100,0 %

Tabla 10.27 Pruebas de chi cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi cuadrado de Pearson	399,650[a]	16	,000
Razón de verosimilitud	375,896	16	,000
Asociación lineal por lineal	243,802	1	,000
Prueba de McNemar-Bowker	328,153	10	,000
N de casos válidos	3560		

a. Ningua casilla (0,0 %) ha esperado un recuento menor que 5. El recuento mínimo esperado es 6,81.

Tabla 10.28 Medidas simétricas

		Valor	Error estándar asintóticoa	T aproximada	Significación aproximada
Medida de acuerdo	Kappa	,150	,010	15,745	,000
N de casos válidos		3560			

a. No se presupone la hipótesis nula.
b. Utilización del error estándar asintótico que presupone la hipótesis nula.

Lo que significa que tanto las respuestas de los hijos como la de los padres concuerdan en su mayoría.

El gráfico solicitado es el siguiente, que se realiza con su respectiva sintaxis:

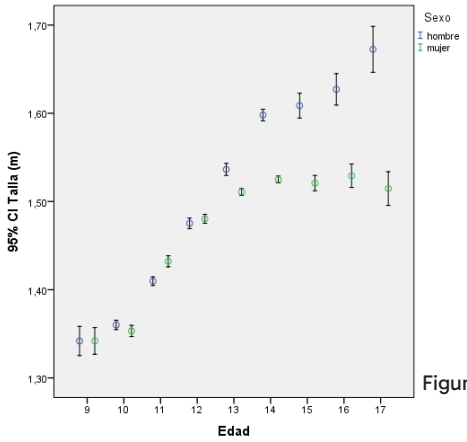


Figura 10.25

```
GRAPH
  /ERRORBAR(CI 95)=talla BY ed BY sexo.
```

10.6 Ejercicios del Capítulo 7

Ejercicio 7.1

Analice y describa las hipótesis nulas que se asocian a la descripción de la relación entre el hecho de desayunar y la ausencia de exceso de peso, en función del sexo del estudiante.

Para ello, lleve a cabo la acción **Analizar ▸ Estadísticos descriptivos ▸ tablas cruzadas** y, en filas, coloque la variable desayuno; en columnas, la variable exceso de peso y, en la ventana de capa, la variable sexo, indicando en las casillas el porcentaje de fila. No olvide solicitar el cálculo de estadísticos y los indicadores de riesgo.

Interprete los resultados.

La hipótesis nula para este caso se podría escribir: "El hecho de que un estudiante tenga un sobrepeso no depende de si toma desayuno", tanto en hombres como en mujeres. Esto se realiza mediante la sintaxis:

```
CROSSTABS
  /TABLES=no_desay BY valoracion_IMC BY sexo
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ RISK
  /CELLS=COUNT ROW
  /COUNT ROUND CELL.
```

De donde se desprenden las siguientes tablas de resultados:

Tabla 10.29 Tabla cruzada ¿toma desayuno? *Normales frente sobrepeso u obeso *Sexo

Sexo				Normales frente sobrepeso u obeso		Total
				Normal	Con sobrepeso	
Hombre	toma desayuno?	no desayuna	Recuento	77	43	120
			% dentro de toma desayuno?	64,2	35,8	100,0
		sí desayuna	Recuento	2187	934	3121
			% dentro de toma desayuno?	70,1	29,9	100,0
	Total		Recuento	2264	977	3241
			% dentro de toma desayuno?	69,9	30,1	100,0

Mujer	toma desayuno?	no desayuna	Recuento	169	83	252
			% dentro de toma desayuno?	67,1	32,9	100,0
		sí desayuna	Recuento	2658	788	3446
			% dentro de toma desayuno?	77,1	22,9	100,0
	Total		Recuento	2827	871	3698
			% dentro de toma desayuno?	76,4	23,6	100,0
Total	toma desayuno?	no desayuna	Recuento	246	126	372
			% dentro de toma desayuno?	66,1	33,9	100,0
		sí desayuna	Recuento	4845	1722	6567
			% dentro de toma desayuno?	73,8	26,2	100,0
	Total		Recuento	091	1848	6939
			% dentro de toma desayuno?	73,4	26,6	100,0

Tabla 10.30 Pruebas de chi cuadrado

Sexo		Valor	df	Significación asintótica (bilateral)	Significación exacta (bilateral)	Significación exacta (unilateral)
Hombre	Chi cuadrado de Pearson	1,915[c]	1	,166		
	Corrección de continuidad[b]	1,645	1	,200		
	Razón de verosimilitud	1,856	1	,173		
	Prueba exacta de Fisher				,187	,101
	Asociación lineal por lineal	1,914	1	,166		
	N de casos válidos	3241				
Mujer	Chi cuadrado de Pearson	13,224[d]	1	,000		
	Corrección de continuidad[b]	12,670	1	,000		
	Razón de verosimilitud	12,307	1	,000		
	Prueba exacta de Fisher				,001	,000
	Asociación lineal por lineal	13,220	1	,000		
	N de casos válidos	3698				

Total	Chi cuadrado de Pearson	10,542[a]	1	,001		
	Corrección de continuidad[b]	10,154	1	,001		
	Razón de verosimilitud	10,050	1	,002		
	Prueba exacta de Fisher				,002	,001
	Asociación lineal por lineal	10,540	1	,001		
	N de casos válidos	6939				

a. Ninguna casilla (0,0 %) ha esperado un recuento menor que 5. El recuento mínimo esperado es 99,07.

b. Solo se ha calculado para una tabla 2x2.

c. Ninguna casilla (0,0 %) ha esperado un recuento menor que 5. El recuento mínimo esperado es 36,17.

d. Ninguna casilla (0,0 %) esperó un recuento menor que 5. El recuento mínimo esperado es 59,35.

Tabla 10.31 Estimación de riesgo

Sexo		Valor	Intervalo de confianza de 95 %	
			Inferior	Superior
Hombre	Razón de ventajas para toma desayuno? (no desayuna / sí desayuna)	,765	,523	1,119
	Para cohorte Normales frente sobrepeso u obeso = Normal	,916	,800	1,049
	Para cohorte Normales frente sobrepeso u obeso = Con sobrepeso	1,197	,937	1,530
	N de casos válidos	3241		
Mujer	Razón de ventajas para toma desayuno? (no desayuna / sí desayuna)	,604	,459	,794
	Para cohorte Normales frente sobrepeso u obeso = Normal	,869	,796	,950
	Para cohorte Normales frente sobrepeso u obeso = Con sobrepeso	1,440	1,195	1,736
	N de casos válidos	3698		

Total	Razón de ventajas para toma desayuno? (no desayuna / sí desayuna)	,694	,556	,866
	Para cohorte Normales frente sobrepeso u obeso = Normal	,896	,832	,965
	Para cohorte Normales frente sobrepeso u obeso = Con sobrepeso	1,292	1,114	1,497
	N de casos válidos	6939		

De acuerdo con estos resultados, en hombres, se aceptaría la hipótesis nula, lo que significaría que las diferencias son debidas al azar. Y en las mujeres, todo lo contrario. De acuerdo con la prueba chi cuadrado, se rechazaría la hipótesis nula, es decir, que es estadísticamente significativo que el sobrepeso en mujeres dependa del tomar desayuno.

Ejercicio 7.2

Analice la hipótesis nula de que las medias de la proporción de grasa corporal no dependen del exceso de peso. Efectúe el análisis mediante el **test de t** y mediante la opción **medias.** Compruebe la relación entre el valor de los estadísticos t y F.

Para realizar el análisis entre las medias de la proporción de grasa corporal mediante exceso de peso se realiza la siguiente sintaxis:

```
T-TEST GROUPS=valoracion_IMC(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=p_grasa
  /CRITERIA=CI(.95).

MEANS TABLES=p_grasa BY valoracion_IMC
  /CELLS=MEAN COUNT STDDEV
  /STATISTICS ANOVA.
```

Tanto para la prueba t como para la prueba f, de donde se desprenden los siguientes resultados:

Prueba T

Tabla 10.32 Estadísticas de grupo

	Normales frente sobrepeso u obeso	N	Media	Desviación estándar	Media de error estándar
Proporción de grasa	Normal	5108	19,9707	5,80159	,08117
	Con sobrepeso	1853	29,4003	5,94903	,13820

Tabla 10.33

		Prueba de muestras independientes								
		Prueba de Levene de igualdad de varianzas		prueba t para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	95 % de intervalo de confianza de la diferencia	
									Inferior	Superior
Proporción de grasa	Se asumen varianzas iguales	6,218	,013	−59,528	6959	,000	−9,42960	,15841	−9,74013	−9,11907
	No se asumen varianzas iguales			−58,833	3211,711	,000	−9,42960	,16028	−9,74386	−9,11535

Medias

Tabla 10.34 Informe

Proporción de grasa			
Normales frente sobrepeso u obeso	**Media**	**N**	**Desviación estándar**
Normal	19,9707	5108	5,80159
Con sobrepeso	29,4003	1853	5,94903
Total	22,4808	6961	7,17537

Tabla 10.34 Medidas de asociación

	Eta	Eta cuadrada
Proporción de grasa * Normales frente sobrepeso u obeso	,581	,337

En la prueba *t*, las varianzas se pueden considerar iguales, por lo tanto, el valor de t es 59,528. En la prueba de medias, el valor de f es 3543,54 que resulta ser el cuadrado del valor de t. De las dos pruebas, se desprende que las dos variables están relacionadas.

Ejercicio 7.3

Lleve a cabo el análisis completo de la dependencia entre el porcentaje corporal de grasa y ser usuario de comida rápida.

Interprete los resultados.

Para realizar dicho análisis, se ejecuta la sintaxis:

```
ONEWAY p_grasa BY com_rap
 /STATISTICS HOMOGENEITY
  /MISSING ANALYSIS
  /POSTHOC=SCHEFFE ALPHA(0.05).
```

De donde se despliegan los siguientes resultados:

Tabla 10.35 Prueba de homogeneidad de varianzas

Proporción de grasa			
Estadístico de Levene	**gl1**	**gl2**	**Sig.**
6,218	1	6959	,013

Tabla 10.36 *ANOVA*

Proporción de grasa					
	Suma de cuadrados	**gl**	**Media cuadrática**	**F**	**Sig.**
Entre grupos	1888,351	4	472,088	9,181	,000
Dentro de grupos	324 360,069	6308	51,420		
Total	326 248,420	6312			

Tabla 10. 37 Comparaciones múltiples

Variable dependiente:

Scheffe

(I) comida rápida		Diferencia de medias (I-J)	Error estándar	Sig.	Intervalo de confianza al 95 %	
					Límite inferior	Límite superior
todos los días	tres o más veces/ semana, pero no todos los días	–0,44456	0,41124	0,883	–1,7116	0,8225
	1 o 2veces/semana	–1,04848	0,38303	0,112	–2,2287	0,1317
	una o menos veces/ semana	–1,34279*	0,36897	0,010	–2,4797	–0,2059
	nunca o casi nunca	–1,80486*	0,37048	0,000	–2,9464	–0,6634
tres o más veces/ semana, pero no todos los días	todos los días	0,44456	0,41124	0,883	–0,8225	1,7116
	1 o 2veces/semana	–0,60392	0,31443	0,450	–1,5727	0,3649
	una o menos veces/ semana	–0,89823	0,29715	0,058	–1,8138	0,0173
	nunca o casi nunca	–1,36029*	0,29902	0,000	–2,2816	–0,4390
1 o 2veces/ semana	todos los días	1,04848	0,38303	0,112	–0,1317	2,2287
	tres o más veces/ semana, pero no todos los días	0,60392	0,31443	0,450	–0,3649	1,5727
	una o menos veces/ semana	–0,29431	0,25669	0,859	–1,0852	0,4966
	nunca o casi nunca	–0,75637	0,25885	0,074	–1,5539	0,0412
una o menos veces/ semana	todos los días	1,34279*	0,36897	0,010	0,2059	2,4797
	tres o más veces/ semana, pero no todos los días	0,89823	0,29715	0,058	–0,0173	1,8138
	1 o 2veces/semana	0,29431	0,25669	0,859	–0,4966	1,0852
	nunca o casi nunca	–0,46206	0,23756	0,436	–1,1940	0,2699
nunca o casi nunca	todos los días	1,80486*	0,37048	0,000	0,6634	2,9464
	tres o más veces/ semana, pero no todos los días	1,36029*	0,29902	0,000	0,4390	2,2816
	1 o 2veces/semana	0,75637	0,25885	0,074	–0,0412	1,5539
	una o menos veces/ semana	0,46206	0,23756	0,436	–0,2699	1,1940

*. La diferencia de medias es significativa en el nivel 0,05.

Tabla 10.38 Proporción de grasa

Scheffe[a,b]

comida rápida	N	Subconjunto para alfa = 0,05		
		1	2	3
todos los días	474	21,3209		
tres o más veces/semana, pero no todos los días	848	21,7654	21,7654	
1 o 2 veces/semana	1345		22,3694	22,3694
una o menos veces/semana	1859		22,6637	22,6637
nunca o casi nunca	1787			23,1257
Sig.		,759	,106	,247

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

a. Utiliza el tamaño de la muestra de la media armónica = 974,666.

b. Los tamaños de grupo no son iguales. Se utiliza la media armónica de los tamaños de grupo. Los niveles de error de tipo I no están garantizados.

En primer lugar, el test de Levène nos indica que no se puede garantizar la homogeneidad de varianzas; es decir, existe heterocedasticidad. Este incumplimiento de las condiciones de aplicabilidad del análisis de varianza no es preocupante, ya que, al analizar el cociente entre la varianza explicada por las medias es superior a la varianza residual, F= 9,181, valor que se asocia a un valor de p <0,001. Este resultado nos lleva a rechazar la hipótesis de igualdad de medias en la proporción de grasa en función de las categorías en las que se divide la variable comida rápida.

Una vez detectado que se puede rechazar la igualdad, el test de Scheffé analiza entre que categorías de la variable comida rápida existen diferencias en la proporción de grasa corporal.

Ejercicio 7.4

Determine la relación entre porcentaje de grasa e IMC, explicitando la bondad del ajuste y el cambio de IMC por unidad porcentual de la proporción de grasa corporal

Explicite la sintaxis utilizada.

Para realizar el análisis de la regresión pedida se realiza la sintaxis:

```
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS CI(95) R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT IMC
 /METHOD=ENTER p_grasa.
```

Se obtienen las siguientes tablas en resultados:

Tabla 10.39 Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,732[a]	,535	,535	2,16167

a. Predictores: (Constante), Proporción de grasa

Tabla 10.40 *ANOVA*[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	37 477,870	1	37 477,870	8020,398	,000b
	Residuo	32 518,148	6959	4,673		
	Total	69 996,018	6960			

a. Variable dependiente: Índice de Masa Corporal.
b. Predictores: (constante), Proporción de grasa.

Tabla 10.41 Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	95,0 % intervalo de confianza para B	
		B	Error estándar	Beta			Límite inferior	Límite superior
1	(Constante)	12,521	,085		146,938	,000	12,354	12,688
	Proporción de grasa	,323	,004	,732	89,557	,000	,316	,330

a. Variable dependiente: Índice de Masa Corporal.

El valor de R cuadrado no es tan alto por lo que la dependencia lineal es moderada. El modelo considera que por cada unidad de aumento en el porcentaje de masa corporal, el IMC aumenta 0,323 unidades, previamente corregido por un valor constante de 12,52.

10.7 Ejercicio del Capítulo 9

Ejercicio único

- Lea los dos artículos y describa los objetivos de cada uno de ellos.
- Analice la información de las viviendas encuestadas, realizando en primer lugar el control de calidad de las mismas, valorando la pérdida de información y resuelva los errores de tecleo del nombre de la comunidad mediante la acción de recodificación automática y recodificar.
- Reproduzca el análisis descriptivo que se encuentra en la tabla 1 del primer artículo y explore las posibles relaciones binarias entre las características de las viviendas y la percepción del aumento de mordeduras, tanto en animales como en los habitantes.
- Genere el archivo que contenga a los habitantes de Morona y de Pastaza.
- Reproduzca los resultados que aparecen en las tres primeras columnas de la tabla 2 del segundo artículo. Valore la fuerza de asociación entre las variables que describe dicha tabla.
- Compare los resultados de los puntos tres y cuatro; y concluya su análisis en el informe.

Ejercicio único

- Los objetivos de los artículos son:
- Determinar si existe un incremento en la mortalidad de seres humanos por la picadura de murciélagos hematófagos en la región amazónica de Ecuador.
- Buscar factores de riesgo para las mordeduras de estos vampiros. Estos podrían ser: fallas en las viviendas que permitan el ingreso de los vampiros, dormir sin protección (toldos), dormir en el suelo o en hamacas y demás; para su futura prevención.
- Buscar los grupos más vulnerables a la mordedura de vampiros, estos podrían ser edad, o presencia de animales en el lugar de su vivienda que podrían atraer a los vampiros.

Para analizar la información de las viviendas, se realiza, en sintaxis, un pedido de las frecuencias para valorar posibles errores. Para ello, use la sintaxis:

```
FREQUENCIES VARIABLES=AV Casa_abert Casa_luz Casa_ilumi PIC CPAbo
CPAeq CPApo CPAav CPApg tivs  tiempovi cir_salud compa_salud mum12
mccm scc aam mbo meq mpo mav mgp qaa vaa acmm cacmh cacma nco-
muna1
  /ORDER=ANALYSIS.
```

En la cual se puede apreciar que no existe en ninguna variable *missing* por el sistema; sin embargo, en algunas variables sí existe la categoría perdidos que son casos en los que, en la variable anterior, su respuesta fue negativa.

Para evaluar la información de la variable “nombre de la comunidad”, se realizó la siguiente sintaxis:

```
AUTORECODE VARIABLES=ncomuna
  /INTO ncomuna1
  /PRINT.
RECODE ncomuna1 (8=1) (17=2) (18=2).
EXECUTE.
FREQUENCIES VARIABLES=ncomuna1
  /ORDER=ANALYSIS.
```

Ya que se detectó ciertos nombres repetidos, se recodifica esta variable en una nueva para no perder las respuestas originales; por lo tanto, al pedir las frecuencias, se tiene:

Tabla 10.42 Nombre de la comunidad

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	macuma	43	11,3	11,3	11,3
	wanpuik	16	4,2	4,2	15,5
	achuntz	43	11,3	11,3	26,8
	amazonas	47	12,3	12,3	39,1
	arutan	10	2,6	2,6	41,7
	ipiak	26	6,8	6,8	48,6
	kuamar	20	5,2	5,2	53,8
	paantin	13	3,4	3,4	57,2
	pumpuentsa	22	5,8	5,8	63,0
	shimkiatan	32	8,4	8,4	71,4
	surik nuevo	12	3,1	3,1	74,5
	tarimiat	9	2,4	2,4	76,9
	timias	9	2,4	2,4	79,3
	tunants	28	7,3	7,3	86,6
	tuutinentza	50	13,1	13,1	99,7
	Wanpuik	1	,3	,3	100,0
	Total	381	100,0	100,0	

Para realizar el análisis y reproducir la tabla 1 se realiza la siguiente sintaxis:

```
CROSSTABS
  /TABLES=Casa_abert Casa_luz Casa_ilumi PIC BY cacmh cacma
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT ROW
  /COUNT ROUND CELL.
```

En cuanto a la percepción de mordeduras a humanos con las características de la vivienda, se tiene la siguiente tabla:

Tabla 10.43 TABLA 1

Variable		Percepción de mordedura a humanos		
		Sí (%)	No (%)	Total (%)
Aberturas en la casa	Sí	114 (35,4)	208 (64,6)	322 (84,5)
	No	20 (33,9)	39 (66,1)	59 (15,5)
Energía proporcionada por	Empresa eléctrica	5 (20,0)	20 (80,0)	25 (6,6)
	Planta eléctrica	27 (33,3)	54 (66,7)	81 (21,3)
	Paneles solares	33 (45,8)	39 (54,2)	72 (28.9)
	Vela, candil, mechero, gas	50 (30,7)	113 (69,3)	163 (42,8)
	Otros	8 (80,0)	2 (20,0)	10 (2,6)
	No tiene luz	11 (36,7)	19 (63,3)	30 (7,9)
Iluminación de la casa por la noche	Sí	30 (47,6)	33 (52,4)	63 (16,5)
	No	104 (32,7)	214 (67,3)	318 (83,5)
Protección de entrada a insectos	Sí	19 (30,6)	43 (69,4)	62 (16,3)
	No	115 (36,1)	204 (63,9)	319 (83,7)
Presencia de animales bovinos	Sí	28 (28,9)	69 (71,1)	97 (25,5)
	No	106 (37,3)	178 (62,7)	284 (74,5)

En cuanto a la percepción de mordeduras a animales con las características de la vivienda, se tiene la siguiente tabla:

Tabla 10.44 TABLA 2

Variable		Percepción de mordedura a animales		
		Sí (%)	No (%)	Total (%)
Aberturas en la casa	Sí	216 (67,1)	106 (32,9)	322 (84,5)
	No	35 (59,3)	24 (40,7)	59 (15,5)
Energía proporcionada por	Empresa eléctrica	16 (64,0)	9 (36,0)	25 (6,6)

	Planta eléctrica	53 (65,4)	28 (34,6)	81 (21,3)
	Paneles solares	42 (58,3)	30 (41,7)	72 (28.9)
	Vela, candil, mechero, gas	114 (69,9)	49 (30,1)	163 (42,8)
	Otros	8 (80,0)	2 (20,0)	10 (2,6)
	No tiene luz	18 (60,0)	12 (40,0)	30 (7,9)
Iluminación de las casa en la noche	Sí	41 (65,1)	22 (34,9)	63 (16,5)
	No	210 (66,0)	108 (34,0)	318 (83,5)
Protección de entrada a insectos	Sí	41 (66,1)	21 (33,9)	62 (16,3)
	No	251 (65,9)	130 (34,1)	319 (83,7)
Presencia de animales bovinos	Sí	70 (72,1)	27 (27,9)	97 (25,5)
	No	181 (63,7)	103 (36,3)	284 (74,5)

Para fusionar los dos archivos hay que considerar que el nombre de la variable como la anchura que se encuentra definidos en cada uno de los archivos sea la misma, para lograr una un archivo con todas las variables, con la siguiente sintaxis:

```
DATASET ACTIVATE ConjuntoDatos1.
ADD FILES /FILE=*
  /FILE='ConjuntoDatos2'.
EXECUTE.
```

Una vez fusionados los dos archivos, para realizar un análisis de asociación, se requieren ciertas tablas cruzadas con las variables solicitadas. Para ello se aplica la siguiente sintaxis:

```
CROSSTABS
  /TABLES=mordidoultimo BY CPAbo Casa_ilumi Casa_abert ipmt ldnd sexo
 edad2r
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ RISK
  /CELLS=COUNT ROW COLUMN TOTAL
  /COUNT ROUND CELL.
```

Los resultados de esta información solicitada se pueden resumir en la siguiente tabla:

Tabla 10.45

Variable	LR (df)	*P*	OR (95 %IC)
Edad	167,99 (2)	< 0,001	
Sexo	1,274 (1)	0,259	1,04 (0,97; 1,11)
Iluminación de las casa en la noche	0,16 (1)	0,899	0,99 (0,82; 1,19)
Uso de toldo	1,786 (1)	0,182	0,95 (0,88; 1,03)
Aberturas en la casa	14,446 (1)	< 0,001	0,69 (0,57; 0,85)
Presencia de animales bovinos	12,846 (1)	< 0,001	1,06 (1,03; 1,09)
Lugar donde duerme	64,341 (1)	< 0,001	0,84 (0,80; 0,88)

De acuerdo con los resultados de esta tabla, se concluye que las mordeduras tienen asociación estadísticamente significativa con variables como la edad, presencia de aberturas en las estructuras de los hogares, la presencia de animales bovinos y el lugar donde duerme.

Con el objetivo de poder comparar en la siguiente tabla, se resume los resultados de las tablas cruzadas anteriormente obtenidas:

Tabla 10.46

Variable		Mordedura a animales		
		Sí (%)	No (%)	Total (%)
Sexo	Masculino	509 (21,2)	1891 (78,8)	2400 (50,0)
	Femenino	477 (19,9)	1921 (80,1)	2398 (50,0)
Edad codificada	Hasta de 12 años	581 (29,0)	1424 (71,0)	2005 (44,1)
	De 13 a 19 años	101 (17,4)	480 (82,6)	581 (12,8)
	Mayores de 19 años	247 (12,6)	1709 (87,4)	1956 (43,1)

Presencia de animales bovinos	Sí	141 (16,2)	731 (83,8)	872 (18,5)
	No	827 (21,5)	3024 (78,5)	3581 (81,5)
Iluminación de las casa en la noche	Sí	122 (20,3)	479 (79,7)	601 (12,7)
	No	846 (20,5)	3276 (79,5)	4122 (87,3)
Uso de protección para dormir (toldo)	Sí	565 (21,2)	2094 (78,8)	2659 (55,4)
	No	421 (19,7)	1718 (80,3)	2139 (44,6)
Aberturas en la casa	Sí	866 (21,4)	3185 (78,6)	4051 (85,8)
	No	102 (15,2)	570 (84,8)	672 (14,2)
Lugar habitual de dormir	Cama	651 (17,9)	2996 (82,1)	3647 (76,0)
	Hamaca o suelo	335 (29,1)	815 (70,9)	1150 (24,0)

Capítulo once
Material complementario

11.1 Presentación

En este capítulo, se presentan las bases de datos según los capítulos.

Bases de datos y su información

Se indican los archivos con su nombre, la información del contenido y los capítulos en que se utilzan. No se indican los archivos generados por el estudiante a lo largo del curso.

11.2 Capítulos 1 a 3

estudiantes_zonas_1y2.sav

Información de variable

Variable	Posición	Etiqueta	Nivel de medición	Rol	Ancho de columna	Alineación	Formato de impresión	Formato de grabación
Codigo	1	Codigo Alumno	Escala	Entrada	11	Derecha	F11	F11
Talla	2	Talla (m)	Escala	Entrada	11	Derecha	F11.2	F11.2
Peso	3	Peso (Kg)	Escala	Entrada	11	Derecha	F11.1	F11.1
p_grasa	4	Proporción de grasa	Escala	Entrada	11	Derecha	F11.2	F11.2
Sexo	5	Sexo	Nominal	Entrada	8	Derecha	F8	F8
Ed	6	Edad	Escala	Entrada	10	Derecha	F8	F8
dep_prog	7	deporte programado	Nominal	Entrada	8	Derecha	F8	F8
Autoper	8	autodepercepción de la imagen	Nominal	Entrada	8	Derecha	F8	F8
tip_flia	9	tipologia familiar	Nominal	Entrada	11	Derecha	F11	F11
no_desay	10	toma desayuno?	Nominal	Entrada	11	Derecha	F11	F11
com_comp	11	Come acompañado	Nominal	Entrada	8	Derecha	F8	F8
com_tres	12	Come tres comida al día	Nominal	Entrada	8	Derecha	F8	F8
Diet	13	Está siguiendo alguna dieta	Nominal	Entrada	8	Derecha	F8	F8
com_rap	14	comida rápida	Nominal	Entrada	11	Derecha	F11	F11
Snaks	15	snacks	Nominal	Entrada	11	Derecha	F11	F11
valoracion_IMC	16	Normales frente sobrepeso u obeso	Escala	Entrada	16	Derecha	F8.2	F8.2
zona_sanit	17	red sanitaria	Escala	Entrada	12	Derecha	F8.2	F8.2
filter_$	18	zona_sanit < 3 (FILTER)	Nominal	Entrada	10	Derecha	F1	F1

Variables en el archivo de trabajo

estudiantes_zonas_3y4.sav

Información de variable								
Variable	Posición	Etiqueta	Nivel de medición	Rol	Ancho de columna	Alineación	Formato de impresión	Formato de grabación
Codigo	1	Codigo Alumno	Escala	Entrada	11	Derecha	F11	F11
Talla	2	Talla (m)	Escala	Entrada	11	Derecha	F11.2	F11.2
Peso	3	Peso (Kg)	Escala	Entrada	11	Derecha	F11.1	F11.1
p_grasa	4	Proporción de grasa	Escala	Entrada	11	Derecha	F11.2	F11.2
cod_esc	5	Codigo Escuela	Nominal	Entrada	11	Derecha	F8	F8
Sexo	6	Sexo	Nominal	Entrada	8	Derecha	F8	F8
Ed	7	Edad	Escala	Entrada	10	Derecha	F8	F8
dep_prog	8	deporte programado	Nominal	Entrada	8	Derecha	F8	F8
Autoper	9	autodepercepción de la imagen	Nominal	Entrada	8	Derecha	F8	F8
Autodef	10	autodefinición étnica	Nominal	Entrada	11	Derecha	F11	F11
tip_flia	11	tipologia familiar	Nominal	Entrada	11	Derecha	F11	F11
no_desay	12	toma desayuno?	Nominal	Entrada	11	Derecha	F11	F11
com_comp	13	Come acompañado	Nominal	Entrada	8	Derecha	F8	F8
com_tres	14	Come tres comida al día	Nominal	Entrada	8	Derecha	F8	F8
Diet	15	Está siguiendo alguna dieta	Nominal	Entrada	8	Derecha	F8	F8
com_rap	16	comida rápida	Nominal	Entrada	11	Derecha	F11	F11
Snaks	17	snacks	Nominal	Entrada	11	Derecha	F11	F11
valoracion_IMC	18	Normales frente sobrepeso u obeso	Escala	Entrada	16	Derecha	F8.2	F8.2
zona_sanit	19	red sanitaria	Escala	Entrada	12	Derecha	F8.2	F8.2
filter_$	20	zona_sanit > 2 (FILTER)	Nominal	Entrada	10	Derecha	F1	F1

Variables en el archivo de Trabajo

Padres.xls

```
Data written to C:\Users\1001278\AppData\Local\Temp\padres.xlsx.
16 variables and 4153 cases written to range: SPSS.
Variable: codigo            Type: Number   Width:  11   Dec: 0
Variable: participa         Type: Number   Width:  11   Dec: 0
Variable: psexo             Type: Number   Width:  11   Dec: 0
Variable: pescola           Type: Number   Width:  11   Dec: 0
Variable: esc_madre         Type: Number   Width:  11   Dec: 0
Variable: pcua_cami         Type: Number   Width:  11   Dec: 0
Variable: pfruta            Type: Number   Width:  11   Dec: 0
Variable: pc_rap            Type: Number   Width:  11   Dec: 0
Variable: pdul              Type: Number   Width:  11   Dec: 0
Variable: prefr             Type: Number   Width:  11   Dec: 0
Variable: psna              Type: Number   Width:  11   Dec: 0
Variable: pc_sol            Type: Number   Width:  11   Dec: 0
Variable: pdiet             Type: Number   Width:  11   Dec: 0
Variable: autoper_pad       Type: Number   Width:  11   Dec: 0
Variable: pedad             Type: Number   Width:   8   Dec: 2
Variable: pdesay            Type: Number   Width:  11   Dec: 0
```

Padres.sav

Información de variable									
Variable	Posición	Etiqueta	Nivel de medición	Rol	Ancho de columna	Alineación	Formato de impresión	Formato de grabación	Valores perdidos
Codigo	1	Codigo Alumno	Escala	Entrada	11	Derecha	F11	F11	
Participa	2	Participa	Nominal	Entrada	11	Derecha	F11	F11	
Psexo	3	Sexo del padre/ madre/tutor	Nominal	Entrada	11	Derecha	F11	F11	
Pescola	4	escolaridad representante	Nominal	Entrada	11	Derecha	F11	F11	0
esc_madre	5	escolaridad de la madre	Nominal	Entrada	11	Derecha	F11	F11	0
pcua_cami	6	cuadras que caminas	Ordinal	Entrada	11	Derecha	F11	F11	
Pfruta	7	fruta	Ordinal	Entrada	11	Derecha	F11	F11	
pc_rap	8	comida rápida	Ordinal	Entrada	11	Derecha	F11	F11	
Pdul	9	dulces	Ordinal	Entrada	11	Derecha	F11	F11	
Prefr	10	refrescos	Ordinal	Entrada	11	Derecha	F11	F11	
Psna	11	snacks	Ordinal	Entrada	11	Derecha	F11	F11	
pc_sol	12	come acompañado	Ordinal	Entrada	11	Derecha	F11	F11	
Pdiet	13	Su hijo/a está realizando dieta?	Nominal	Entrada	11	Derecha	F11	F11	
autoper_pad	14	autopercepción de la imagen	Nominal	Entrada	11	Derecha	F11	F11	
Pedad	15	Edad del padre/ madre/tutor	Escala	Entrada	8	Derecha	F8.2	F8.2	
Pdesay	16	toma desayuno su hijo/a?	Nominal	Entrada	11	Derecha	F11	F11	
Variables en el archivo de trabajo									

11.3 Capítulo 4 a 7

Estudio.sav

Información de variable									
Variable	Posición	Etiqueta	Nivel de medición	Rol	Ancho de columna	Alineación	Formato de impresión	Formato de grabación	Valores perdidos
Codigo	1	Codigo Alumno	Escala	Entrada	11	Derecha	F11	F11	
Talla	2	Talla (m)	Escala	Entrada	11	Derecha	F11.2	F11.2	2,00 a 300,00
Peso	3	Peso (Kg)	Escala	Entrada	11	Derecha	F11.1	F11.1	De 105,0 a 700,0, y ,0
p_grasa	4	Proporción de grasa	Escala	Entrada	11	Derecha	F11.2	F11.2	
Aebas	5	Año Escolar	Ordinal	Entrada	8	Derecha	F8	F8	
Sexo	6	Sexo	Nominal	Entrada	8	Derecha	F8	F8	
Ed	7	Edad	Escala	Entrada	10	Derecha	F8	F8	
Status	8	categoría de IMC según WHO 2007	Nominal	Entrada	10	Derecha	F8.2	F8.2	
cua_cami	9	cuadras que caminas	Ordinal	Entrada	8	Derecha	F8	F8	
Autoper	10	autodepercepción de la imagen	Nominal	Entrada	8	Derecha	F8	F8	
tip_flia	11	tipologia familiar	Nominal	Entrada	11	Derecha	F11	F11	
Desayuno	12	toma desayuno?	Nominal	Entrada	11	Derecha	F11	F11	
com_comp	13	Come acompañado	Nominal	Entrada	8	Derecha	F8	F8	
com_tres	14	Come tres comida al día	Nominal	Entrada	8	Derecha	F8	F8	

Diet	15	Está siguiendo alguna dieta	Nominal	Entrada	8	Derecha	F8	F8	
Frut	16	fruta	Ordinal	Entrada	11	Derecha	F11	F11	
com_rap	17	comida rápida	Ordinal	Entrada	11	Derecha	F11	F11	
Dulce	18	dulces	Ordinal	Entrada	11	Derecha	F11	F11	
Refresc	19	refrescos	Ordinal	Entrada	11	Derecha	F11	F11	
exceso_ de_peso	20	Normales frente sobrepeso u obeso	Nominal	Entrada	16	Derecha	F8.2	F8.2	
zona_sanit	21	red sanitaria	Nominal	Entrada	12	Derecha	F8.2	F8.2	
status2	22	clasificación según IMC-OMS	Escala	Entrada	10	Derecha	F8.2	F8.2	
Participa	23	Existe informacion de los responsables	Nominal	Entrada	11	Derecha	F11	F11	
Psexo	24	Sexo del responsable	Nominal	Entrada	11	Izquierda	A11	A11	
Esc	25	escolaridad representante	Nominal	Entrada	11	Derecha	F11	F11	0
esc_m	26	escolaridad de la madre	Nominal	Entrada	11	Derecha	F11	F11	0
pdep_prog	27	deporte programado	Nominal	Entrada	11	Derecha	F11	F11	
pc_rap	28	comida rápida	Nominal	Entrada	11	Derecha	F11	F11	
Psna	29	snacks	Nominal	Entrada	11	Derecha	F11	F11	
pc_sol	30	come acompañado	Nominal	Entrada	11	Derecha	F11	F11	
Pdiet	31	realizando dieta	Nominal	Entrada	11	Derecha	F11	F11	
autoper_ pad	32	autopercepción de la imagen	Nominal	Entrada	11	Derecha	F11	F11	
Pedad	33	Edad del padre/ madre/tutor	Escala	Entrada	8	Derecha	F8.2	F8.2	
pno_desay	34	toma desayuno?	Nominal	Entrada	11	Derecha	F11	F11	
Imc	35	<ninguno>	Escala	Entrada	10	Derecha	F8.2	F8.2	
Variables en el archivo de trabajo									

11.4 Capítulo 8

CPV2010_Poblacion_2.sav

Dada la extensión de los datos censales que se indican en este archivo, solo los datos de la provincia, edad y sexo de los habitantes censados en el año 2010.

Información de variable

Variable	Posición	Etiqueta	Nivel de medición	Rol	Ancho de columna	Alineación	Formato de impresión	Formato de grabación
I01	1	PROVINCIA	Escala	Entrada	5	Derecha	F2	F2
P01	2	Cual es el Sexo	Nominal	Entrada	5	Derecha	F1	F1
Edad	3	Cuantos años cumplidos tiene	Escala	Entrada	5	Derecha	F3	F3

Variables en el archivo de trabajo

Valores de variable		
Valor		Etiqueta
I01	1	Azuay
	2	Bolívar
	3	Cañar
	4	Carchi
	5	Cotopaxi
	6	Chimborazo
	7	El Oro
	8	Esmeraldas
	9	Guayas
	10	Imbabura
	11	Loja
	12	Los Ríos
	13	Manabí
	14	Morona Santiago

Valores de variable		
Valor		Etiqueta
	15	Napo
	16	Pastaza
	17	Pichincha
	18	Tungurahua
	19	Zamora Chinchipe
	20	Galápagos
	21	Sucumbios
	22	Orellana
	23	Santo Domingo de los Tsáchilas
	24	Santa Elena
	90	Zonas No Delimitadas
P01	1	Hombre
	2	Mujer

consultas.sav

Información de variable								
Variable	Posición	Etiqueta	Nivel de medición	Rol	Ancho de columna	Alineación	Formato de impresión	Formato de grabación
Inscr	1	INSCRIPCION	Nominal	Entrada	8	Izquierda	A6	A6
nombre	2	Nombre del trabajador enmascarado	Nominal	Entrada	8	Izquierda	A8	A8
f_consul	3	fecha de consulta	Escala	Entrada	11	Izquierda	EDATE10	EDATE10
bajas	4	sufrió baja en esa consulta?	Escala	Entrada	8	Izquierda	F8	F8
fecha_ba	5	fecha de baja	Escala	Entrada	17	Izquierda	EDATE10	EDATE10
dias	6	dias de baja	Escala	Entrada	8	Izquierda	F8	F8
gra_grup	7	Grandes grupos CIE-10	Escala	Entrada	30	Derecha	F8	F8
consulta	8	consulto o no?	Escala	Entrada	8	Derecha	F1	F1
cargo4	9	Tipo de trabajo que realiza	Escala	Entrada	7	Derecha	F8	F8
esc3	10	Escolaridad	Escala	Entrada	20	Derecha	F8	F8
sexo	11	Sexo	Nominal	Entrada	8	Derecha	F1	F1
cbo	12	Clasificación Brasileira de Ocupaciones	Escala	Entrada	16	Derecha	F8	F8
turno	13	Turno laboral	Escala	Entrada	8	Derecha	F8	F8
e_civ	14	Estado civil	Escala	Entrada	19	Derecha	F8	F8
ant4	15	Antigüedad en el Hospital	Escala	Entrada	8	Derecha	F8.2	F8.2
edad3	16	Edad en tres categorías	Escala	Entrada	8	Derecha	F8.2	F8.2
vinculo	17	Vínculo contractual	Nominal	Entrada	8	Derecha	F2	F2

eda_an	18	Edad al principio del estudio	Escala	Entrada	8	Derecha	F8.2	F8.2
edad5	19	Edad de cinco en cinco años	Escala	Entrada	8	Derecha	F8.2	F8.2
hta	20	Hipertensión arterial	Escala	Entrada	8	Derecha	F8.2	F8.2
asma	21	Asmático	Escala	Entrada	8	Derecha	F8.2	F8.2
ano_cons	22	<ninguno>	Escala	Entrada	8	Derecha	F8.2	F8.2
fecha2	23	<ninguno>	Escala	Entrada	8	Derecha	EDATE10	EDATE10
fecha3	24	<ninguno>	Escala	Entrada	8	Derecha	EDATE10	EDATE10
filter_$	25	anybaja = 2002 (FILTER)	Escala	Entrada	8	Derecha	F1	F1
anybaja	26	<ninguno>	Escala	Entrada	8	Derecha	F8.2	F8.2

Variables en el archivo de trabajo

11.5 Capítulo 9

viviendas_articulo_1.sav

Información de variable									
Variable	Posición	Etiqueta	Nivel de medición	Rol	Ancho de columna	Alineación	Formato de impresión	Formato de grabación	Valores perdidos
n_cuest	1	Numero de la Encuesta	Nominal	Entrada	10	Derecha	F8	F8	
f_cuest	2	Fecha de la Encuesta	Nominal	Entrada	12	Derecha	EDATE10	EDATE10	
c_cuest	3	Codigo del Encuestador	Nominal	Entrada	10	Derecha	F8	F8	
Ncomuna	4	Nombre de la comunidad	Nominal	Entrada	13	Izquierda	A11	A11	
cod_ comuna	5	Codigo de la comunidad	Nominal	Entrada	8	Derecha	F8	F8	

cod_ vivienda	6	Codigo de la vivienda	Nominal	Entrada	14	Derecha	F8	F8
AV	7	Apariencia de la vivienda (por observaci?n)	Nominal	Entrada	10	Derecha	F8	F8
Casa_ abert	8	Su casa tiene aberturas en el techo, paredes, piso o en las uniones entre estos?	Escala	Entrada	8	Derecha	F8	F8
Casa_luz	9	La luz que tiene en la casa s proporcionada por	Escala	Entrada	8	Derecha	F8	F8
Casa_ilumi	10	La casa permanece iluminada toda la noche	Escala	Entrada	8	Derecha	F8	F8
PIC	11	Existe alguna proteccion para evitar la entrada de insectos y animales a la casa?	Nominal	Entrada	10	Derecha	F8	F8
CPAbo	12	Verificar y cuantificar la presencia de animales bovino (vaca, toro)	Nominal	Entrada	10	Derecha	F8	F8
Cantbo	13	Cantidad de Bobinos	Escala	Entrada	8	Derecha	F8	F8
CPAeq	14	Verificar y cuantificar la presencia de animales equino (caballo, burro)	Nominal	Entrada	10	Derecha	F8	F8
Canteq	15	Cantidad de equinos	Escala	Entrada	8	Derecha	F8	F8
CPApo	16	Verificar y cuantificar la presencia de animales porcino	Nominal	Entrada	10	Derecha	F8	F8
Cantpo	17	Cantidad de porcinos	Escala	Entrada	8	Derecha	F8	F8
CPAav	18	Verificar y cuantificar la presencia de aves	Nominal	Entrada	10	Derecha	F8	F8
Cantav	19	Cantidad de aves	Escala	Entrada	8	Derecha	F8	F8
CPApg	20	Verificar y cuantificar la presencia de perros y gatos	Escala	Entrada	8	Derecha	F8	F8
Cantpg	21	Cantidad de perros y gatos	Escala	Entrada	8	Derecha	F8	F8
CPVC	22	Cuantas personas viven en la casa?	Nominal	Entrada	10	Derecha	F8	F8
Tivs	23	Para ir a la Unidad de Salud que usted más utiliza, cuál es la vía habitual que usa	Nominal	Entrada	10	Derecha	F8	F8
Tiempovi	24	¿Cuánto tiempo ocupa en el viaje a la Unidad de Salud que indicó (en horas cerradas)	Nominal	Entrada	10	Derecha	F8	F8
cir_salud	25	En qué circunstancias acude a la unidad de salud?	Escala	Entrada	8	Derecha	F8	F8

compa_ salud	26	Con quién acude a la unidad de salud	Escala	Entrada	8	Derecha	F8	F8	
mum12	27	Cuántas personas han muerto en los últimos 12 meses luego de la mordedura de murciélago en su comunidad	Nominal	Entrada	10	Derecha	F8	F8	99
Mccm	28	Conoce algún caso de muerte por mordedura de murciélago en alguna comunidad cercana?	Nominal	Entrada	10	Derecha	F8	F8	99
Scc	29	Si conoce, cuántas ?	Nominal	Entrada	10	Derecha	F8	F8	99
Aam	30	¿Han sido atacados sus animales por los murciélagos	Nominal	Entrada	10	Derecha	F8	F8	0
Mbo	31	mordedura animales bovino (vaca, toro)	Nominal	Entrada	18	Derecha	F8	F8	0
Meq	32	mordedura (caballo, burro)	Nominal	Entrada	10	Derecha	F8	F8	0
Mpo	33	mordedura porcino	Nominal	Entrada	10	Derecha	F8	F8	0
Mav	34	mordedura de aves	Nominal	Entrada	10	Derecha	F8	F8	0
Mgp	35	mordedura perros y gatos	Escala	Entrada	8	Derecha	F8	F8	0
Qaa	36	Que ocurrió con los animales atacados	Nominal	Entrada	10	Derecha	F8	F8	0
Vaa	37	Le han visitado para vacunar a sus animales: (en el caso de tener ganado bovino o equino)	Nominal	Entrada	10	Derecha	F8	F8	0
Acmm	38	Usted cree que ha aumentado la cantidad de murciélagos en su comunidad?	Nominal	Entrada	10	Derecha	F8	F8	
Cacmh	39	Usted cree que en su comunidad ha aumentado las mordeduras o ataques de murciélagos: a los humanos	Nominal	Entrada	10	Derecha	F8	F8	
Cacma	40	Usted cree que en su comunidad ha aumentado las mordeduras o ataques de murciélagos: a los animales	Nominal	Entrada	10	Derecha	F8	F8	

Variables en el archivo de trabajo


morona.sav


pastaza.sav

Los dos archivos contienen la misma información y tienen la misma estructura.

Información de variable									
Variable	Posición	Etiqueta	Nivel de medición	Rol	Ancho de columna	Alineación	Formato de impresión	Formato de grabación	Valores perdidos
cod_vivienda	1	Còdigo de vivienda	Escala	Entrada	8	Derecha	F8	F8	
n_cuest	2	Numero de encuesta	Nominal	Entrada	10	Derecha	F8	F8	
f_cuest	3	Fecha de la encuesta	Escala	Entrada	12	Derecha	DATE11	DATE11	
c_mfamilia	4	Codigo del miembro de la Familia	Nominal	Entrada	12	Derecha	F8	F8	
Sexo	5	Sexo	Nominal	Entrada	10	Derecha	F8	F8	
f_nam	6	Fecha de Nacimiento	Nominal	Entrada	12	Derecha	EDATE10	EDATE10	
lnacimiento	7	Lugar de Nacimiento	Nominal	Entrada	26	Izquierda	A18	A18	
lresidencia	8	Lugar de Residencia	Nominal	Entrada	26	Izquierda	A18	A18	
salecon	9	Usted sale de su comunidad algún tiempo en otra comunidad	Nominal	Entrada	10	Derecha	F8	F8	
plengua	10	Lengua Natal	Nominal	Entrada	10	Izquierda	A10	A10	
slengua	11	Segunda lengua	Nominal	Entrada	10	Izquierda	A10	A10	
tvc	12	Tiempo de vivir en su comunidad	Nominal	Entrada	10	Derecha	F8	F8	
nacionalidad	13	Nacionalidad Indigena del Encuestado	Nominal	Entrada	14	Izquierda	A8	A8	
uaea	14	Ultimo año escolar aprobado	Nominal	Entrada	10	Derecha	F8	F8	
ocupacionqd	15	Quehaceres Domesticos	Nominal	Entrada	13	Derecha	F8	F8	
ocupaciona	16	Agricultor o recolector	Nominal	Entrada	12	Derecha	F8	F8	
ocupaciong	17	Ganadero	Nominal	Entrada	12	Derecha	F8	F8	
ocupacione	18	Estudiante	Nominal	Entrada	8	Derecha	F8	F8	

ocupacionm	19	Minería	Nominal	Entrada	8	Derecha	F8	F8	99
ocupacionc	20	Cazador o Pescador	Nominal	Entrada	12	Derecha	F8	F8	
ocupacionT	21	Talador de Arboles	Nominal	Entrada	12	Derecha	F8	F8	
rol	22	Rol en la comunidad	Nominal	Entrada	14	Derecha	F8	F8	
rol_otro	23	Rol de la comunidad (otro)	Nominal	Entrada	13	Izquierda	A46	A46	
ldnd	24	Lugar donde normalmente duerme	Nominal	Entrada	10	Derecha	F8	F8	
ipmt	25	Con la intención de protegerse de los murciélagos, usa toldo?	Nominal	Entrada	10	Derecha	F8	F8	
ipmr	26	Con la intención de protegerse de los murciélagos, usa repelente?	Nominal	Entrada	10	Derecha	F8	F8	
ipmo	27	Con la intención de protegerse de los murciélagos, otros (especifique)?	Nominal	Entrada	10	Derecha	F8	F8	
ipmoespc	28	Con la intención de protegerse de los murciélagos, especifique el otro	Nominal	Entrada	8	Izquierda	A11	A11	
cvam	29	Cuántas veces ha sido agredido por un murciélago en los últimos 12 meses	Nominal	Entrada	10	Derecha	F8	F8	
um	30	Si usted ha sido mordido por un murciélago, cuando fue la última mordedura?	Nominal	Entrada	10	Derecha	F8	F8	
mord_lug	31	Le ha mordido en la misma herida el murciélago	Nominal	Entrada	8	Derecha	F8.2	F8.2	
Cant_mord_lug	32	Cuántas mordeduras en el mismo sitio	Escala	Entrada	8	Derecha	F8.2	F8.2	
doum	33	Donde ocurrió la última mordedura:	Nominal	Entrada	10	Derecha	F8	F8	
qecuo	34	Que estaba haciendo cuando ocurrió?	Nominal	Entrada	10	Derecha	F8	F8	
lcc	35	Lugar del cuerpo donde ocurrió la última mordedura, cabeza y cuello	Nominal	Entrada	10	Derecha	F8	F8	

lcm	36	Lugar del cuerpo donde ocurrió la última mordedura, manos	Nominal	Entrada	10	Derecha	F8	F8	
lcp	37	Lugar del cuerpo donde ocurrió la última mordedura, pies	Nominal	Entrada	10	Derecha	F8	F8	
lco	38	Lugar del cuerpo donde ocurrió la última mordedura, otros (especifique)	Nominal	Entrada	10	Derecha	F8	F8	
lcoespc	39	Especifique el otro lugar de la última mordedura	Nominal	Entrada	8	Izquierda	A11	A11	
pcc	40	Presencia de mordeudra reciente, cabeza y cuello	Nominal	Entrada	10	Derecha	F8	F8	
pcm	41	Presencia de mordeudra reciente, manos	Nominal	Entrada	10	Derecha	F8	F8	
pcp	42	Presencia de mordeudra reciente, pies	Nominal	Entrada	10	Derecha	F8	F8	
pco	43	Presencia de mordeudra reciente, otros (especifique)	Nominal	Entrada	10	Derecha	F8	F8	
pcoespc	44	Especifique el lugar de la mordedura reciente	Nominal	Entrada	8	Izquierda	A7	A7	
phl	45	Que fue lo primero que hizo en el lugar de la herida de la última mordedura	Nominal	Entrada	23	Derecha	F8	F8	
phlespc	46	Especifique el remedio natural o medicamento	Nominal	Entrada	14	Izquierda	A15	A15	
lh	47	Si se lavó la herida ¿dónde lo hizo?	Nominal	Entrada	10	Derecha	F8	F8	
moment_heri	48	En qué momento lo hizo	Nominal	Entrada	8	Derecha	F8.2	F8.2	
pa	49	¿Dónde acudió primero para la atención?	Nominal	Entrada	10	Derecha	F8	F8	
mce	50	¿Cree usted que la mordedura de murciélagos puede causar enfermedades	Nominal	Entrada	10	Derecha	F8	F8	99
qepc	51	Que enfermedades puede causar las mordeduras de murcielagos	Nominal	Entrada	8	Derecha	F8	F8	

vva	52	Le han visitado para vacunarle contra la rabia antes de ser mordido: (vacuna pre exposición)	Nominal	Entrada	10	Derecha	F8	F8	
avr	53	Cuántas dosis de vacuna ha recibido (numero de pinchazos)	Nominal	Entrada	10	Derecha	F8	F8	0
edad	54	Edad	Escala	Entrada	10	Derecha	F8	F8	
mordedura	55	mordedura de murciélago	Nominal	Entrada	11	Derecha	F8	F8	
Casa_abert	56	Su casa tiene aberturas en el techo, paredes, piso o en las uniones entre estos?	Nominal	Entrada	8	Derecha	F8	F8	
Casa_ilumi	57	La casa permanece iluminada toda la noche	Nominal	Entrada	8	Derecha	F8	F8	
CPAbo	58	Verificar y cuantificar la presencia de animales bovino (vaca, toro)	Nominal	Entrada	10	Derecha	F8	F8	
provincia	59	Case source is H:\ bases_moronamiguel. sav	Nominal	Entrada	11	Derecha	F1	F1	
edadr	60	Edad recodificada	Ordinal	Entrada	10	Derecha	F8.2	F8.2	
vive_en	61	Lugar de Residencia	Nominal	Entrada	9	Derecha	F3	F3	
indios	62	Nacionalidad Indigena del Encuestado	Nominal	Entrada	8	Derecha	F2	F2	
edad2r	63	edad recodificada 2	Ordinal	Entrada	10	Derecha	F8.2	F8.2	
donde_ mordio	64	<ninguno>	Nominal	Entrada	14	Derecha	F8.2	F8.2	
filter_$	65	mordedura = 1 (FILTER)	Escala	Entrada	10	Derecha	F1	F1	
mordidoultimo	66	<ninguno>	Escala	Entrada	15	Derecha	F8.2	F8.2	
prot	67	<ninguno>	Nominal	Entrada	10	Derecha	F8.2	F8.2	

Variables en los archivos de trabajo

Valores de variable		
Valor		Etiqueta
c_mfamilia	1	Jefe de familia
	2	Miembro de familia
Sexo	1	Femenimo
	2	Masculino
	99	Perdidos
Salecon	1	SI
	2	NO
	99	Perdidos
Ocupacionqd	1	SI
	2	No
	99	Perdidos
Ocupaciona	1	SI
	2	No
	99	Perdidos
Ocupaciong	1	SI
	2	No
	99	Perdidos
Ocupacione	1	Si
	2	No
Ocupacionm	1	Si
	2	No
	99a	Perdidos
ocupacions	1	SI
	2	NO
	99	Perdidos
Ocupaciont	1	SI
	2	NO
	99	Perdidos
Rol	1	Jefe de la comunidad
	2	Sindico
	3	Shaman
	4	religioso
	5	otro (especificar)
	6	Ninguno
	99	Perdidos

Ldnd	1	cama
	2	hamaca o suelo
Ipmt	1	SI
	2	NO
	99	Perdidos
Ipmr	1	SI
	2	NO
	99	Perdidos
Ipmo	1	SI
	2	NO
	99	Perdidos
mord_lug	1,00	Si
	2,00	No
Doum	1	Dentro de la casa
	2	Alrededor de la casa
	3	Fuera mientras trabaja
	99	Perdidos
Qecuo	1	Durmiendo
	2	Despierto
	3	Trabajando
	99	Perdidos
Lcc	1	SI
	2	NO
	99	Perdidos
Lcm	1	SI
	2	NO
	99	Perdidos
Lcp	1	SI
	2	NO
	99	Perdidos
Lco	1	Si
	2	No
Pcc	1	SI
	2	NO
	99	Perdidos

Pcm	1	SI
	2	NO
	99	Perdidos
Pcp	1	SI
	2	NO
	99	Perdidos
Pco	1	Si
	2	No
Phl	1	No hizo nada
	2	Se lavo solo con agua
	3	Se lavo solo con agua caliente
	4	Se lavó con agua y jabón
	5	Se lavó con agua caliente y jabón
	6	Colocó algún remedio natural o medicamentos
	99	Perdidos
Lh	1	Rio
	2	casa y comunidad
	3	Le lavaron en la unidad de salud
	99	Perdidos
moment_heri	1,00	Apenas se dio cuenta
	2,00	Cuando pasó por una fuente de agua
	3,00	Cuando le tocó bañarse
Pa	1	Curandero o Shaman
	2	Medico
	3	No fue con ninguno
	99	Perdidos
Mce	1	SI
	2	NO
	99a	Perdidos
Qepc	1	Enfermedades que curan los médicos
	2	Males que curan los shamanes
	3	los dos
vva	1	SI
	2	NO
	99	Perdidos

mordedura	0	No mrdedura
	1	Si mordedura
Casa_abert	1	Si
	2	No
Casa_ilumi	1	Si
	2	No
CPAbo	1	SI
	2	NO
	99	Perdidos
provincia	1	Morona
	2	Pastaza
edadr	1,00	0-4
	2,00	5-9
	3,00	10-14
	4,00	14-19
	5,00	20
vive_en	1	ijinti
	2	10 de agosto
	3	24 de mayo
	4	achuar
	5	achuntz
	6	amazonas
	7	ankuash
	8	arutam
	9	arutan
	10	baños
	11	centro yuu
	12	consuelo
	13	copataza
	14	cuchentza
	15	cueva de los tayos
	16	cueva de los toyos
	17	cueva delos tayos
	18	chapintza
	19	charuzi
	20	chchirota
	21	chichirota

	22	chiriap
	23	chiwitayo
	24	chuwitayo
	25	don bosco
	26	guarani
	27	imeyoa
	28	ipiak
	29	kajekai
	30	kajekay
	31	kapawi
	32	karama
	33	kawa
	34	kemkuim
	35	kiim
	36	kintiuk
	37	kuakash
	38	kuamar
	39	kuankua
vive_en	40	kumay
	41	kunkuk
	42	kusute
	43	kusutka
	44	kusutkau
	45	kutsukau
	46	macuma
	47	makucham
	48	makuzar
	49	mamayak
	50	mashient
	51	masurash
	52	mutints
	53	nankay
	54	nayontz
	55	nayumentza
	56	numbaimi
	57	paantin
	58	payashnia

	59	pumpuentsa
	60	samikim
	61	san alfonso
	62	san antonio
	63	san carlos
	64	san francisco
	65	san pedro
	66	san rafael
	67	san ramon
	68	santiak
	69	shaimi
	70	shakap
	71	shakay
	72	sharamentza
	73	shinkiatam
	74	shiram
	75	sucre
	76	surik nuevo
vive_en	77	suritiak nunka
	78	suwa
	79	taisha
	80	tamiriat
	81	timiantzu
	82	timias
	83	tinkias
	84	tinshi
	85	tres marias
	86	tumbaim
	87	tunants
	88	tuutinentza
	89	unt pastaza
	90	uyuimi
	91	wampuik
	92	wasakentza
	93	wasusentza
	94	wayusentza
	95	wiririma

vive_en	96	wisui
	97	yajintz
	98	yampis
	99	Yavintz
	100	Yukaip
	101	Yumiantza
Indios	1	Achuar
	2	Shuar
	3	Otros
edad2r	1,00	0-12
	2,00	13-19
	3,00	>19
donde_mordio	1,00	cabeza y cuello
	2,00	manos
	3,00	pies

a. Valor perdido

Este libro está pensado para facilitar el aprendizaje del programa de análisis estadístico SPSS, de forma profesional. El proceso se lleva a cabo desde la creación o importación de la base de datos, su control de calidad y el desarrollo de los análisis estadísticos.

Los datos corresponden a diversos casos reales de estudios llevados a cabo en Ecuador, como parte de las actividades de colaboración en el grupo de investigación iberoamericano Grup´s de Recerca d´Amèrica y Àfrica Llatines-GRAAL, por sus siglas en catalán (Grupo de Investigación de América y África latinas).

El formato del libro se complementa con diversos ejercicios que ejemplarizan el proceso de análisis y finalmente con un ejercicio de autoevaluación con el fin de que el estudiante ponga a prueba las habilidades adquiridas. Este es un libro que permite su utilización en cursos de autoaprendizaje o como texto guía para un curso presencial, a desarrollar en las clases y está especialmente indicado para estudiantes de cuarto nivel en aspectos de Metodología de la Investigación con énfasis en Ciencias de la Salud y de la Vida.

UIDE
Powered by Arizona State University®

UAB
Universitat Autònoma de Barcelona

ISBN: 978-9942-682-01-7